International Series in
Operations Research & Management Science

H. A. Eiselt
Vladimir Marianov *Editors*

Applications of Location Analysis

Springer

International Series in Operations Research & Management Science

Volume 232

More information about this series at http://www.springer.com/series/6161

H. A. Eiselt • Vladimir Marianov
Editors

Applications of Location Analysis

Springer

Editors
H. A. Eiselt
Faculty of Business Administration
University of New Brunswick
Fredericton
New Brunswick
Canada

Vladimir Marianov
Department of Electrical Engineering
Pontificia Universidad Católica de Chile
Macul
Chile

ISSN 0884-8289 ISSN 2214-7934 (electronic)
International Series in Operations Research & Management Science
ISBN 978-3-319-20281-5 ISBN 978-3-319-20282-2 (eBook)
DOI 10.1007/978-3-319-20282-2

Library of Congress Control Number: 2015948175

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media (www.springer.com)

Preface

This book is designed to highlight some of the applications of location analysis in a variety of different fields. In some sense, it is a companion piece of the "Foundations of Location Analysis" volume that was published in 2011. While the latter book focused on the problems that can be solved, this book describes scenarios, in which location techniques have been used to solve actual problems.

It is well known that the techniques that are usually described in books and taught in many location courses—medians, centers, covering problems—are but prototypes that may be applied to real situations when suitably modified. In addition, many location problems comprise multiple objectives, which makes tools from multicriteria decision making an attractive choice. As was stated elsewhere, the fact that location problems are typically found on the strategic level, means that more often than not, simple single-objective optimization models may very well be a good starting point, but have to be supplemented by other tools that allow the inclusion of more sophisticated features, such as nonlinear choice rules and multiple objectives.

This book features spotlights on some important applications of location models and the different tools used in the decision-making process. There are three main classes of applications: applications by businesses, those that deal with public services, and applications that deal with law enforcement and first responders. Overall, it is noticeable that the public sector appears to be the main user of location models.

The first application in the private sector is described by Başar et al. The authors describe the location of bank branches in Istanbul, Turkey. The tool is a mixed-integer programming problem, whose objective is to maximize the net profit, based on the expected transaction volume at the branches as well as branch opening and closing penalties. One of the main issues here is a description of the banking behavior of individual and commercial users.

In the second chapter, Bhadury et al. discuss the potential location of a logistics park (a.k.a. freight village) in the southeastern part of North Carolina. The authors delineate a four-step procedure that incorporates the main phases of the location process: choice of the general region, identification of suitable sites within the chosen region, assessment of the sites that were identified in the previous step, and a

comparison of the sites. In the end, the situation had changed, and the logistics park was not built and a new process started. While disappointing, this is actually not uncommon: the case of the sewage treatment plant in Halifax, Canada, was, at least in that respect, similar.

In the third chapter, Gunn focuses on sustainable forest management. His main objective is the maximization of discounted profit. Again, the tool are mixed-integer optimization formulations. Important features include a concentration of the entire supply chain that comprises the forest, sawmills, pulpmills, and different products delivered to different customers. The constructions of roads to bring equipment to the parcels that are to be harvested is an important cost factor. Also important are neighborhood restrictions that ensure that the planners avoid denuding large swaths of land, which would foster erosion.

Chapter 4 by Arnolds and Nickel looks at layout problems in a hospital. It is well known that layout problems are much more difficult than location problems, as the former have to respect the shape of the facilities (e.g., rooms), whereas location problem essentially locate points in some space. The objective in this piece is to locate wards, walk-in clinics, and operating rooms so as to minimize travel distances and times. It is well-known that quadratic assignment problems, one of the cornerstones of layout problems, are already among the most difficult problems in integer programming. Any additional model features that make the formulations more realistic will further increase the degree of difficulty, so that heuristics have to be applied.

The first chapter dealing with public services is Chap. 5. Here, Church et al. deal with the design of habitats for wildlife in California under special consideration of the spotted owl. The tool of choice is a mixed-integer programming formulation of an anti-covering model. In general, the problem of designing a habitat is a difficult one, in part due to the different requirements of individual species: for instance, while mountain lions will thrive in large, contiguous areas, bald eagles will prefer a large number of small protected zones. If preserves for multiple species are to be designed, suitable compromises will have to be found.

Chapter 6 deal with the control of forest fires. Church et al. set up a multiobjective mixed-integer optimization model that includes fuel removal, the minimization of damage in the transition region, known as wildland-urban interface, the minimization of the variation in year-to-year workload of the crews, and the maximization of adjacency in some period and subsequent periods. The main concern is scheduling, and the authors set up a decision support system for that purpose.

Chapter 7 by Geetla et al. discusses the location of intelligent sensors along highways to enable emergency vehicles to respond in a timely fashion. The authors devise an explicit-implicit model, which is succeeded by an optimization that includes single and double coverage. It is noteworthy that the model maximizes coverage from nodes as well as coverage from paths for an area in a part of Buffalo, New York. A variety of sensitivity analyses establishes the robustness of the system.

In Chap. 8, Verter and Zhang describe the problem of locating breast cancer screening centers in Montreal, Quebec, Canada. The authors formulate a mixed-integer programming problem, whose objective is the maximization of the level

of participation in the program. Again, proximity and patient behavior are main features of the model.

Chapter 9 by Yezer and Gillula is rooted in economic theory. The paper performs an economic analysis for the locations of post offices. Among other concerns, included in the model are revenue, the number of windows, where service is offered, and the size of the facility. The objective is the maximization of net revenue per unit area of the facility. The authors' model allows the analysis of an existing retail network and permits decisions regarding the closures of some of the facilities.

In Chap. 10, Giesen et al. investigate the location of schools in a region of Brazil. The objective function minimizes the sum of transportation, operating, and penalty costs. Complicating features include the existence of different types of schools (public and charger schools), which operate under different rules (children must attend the school to which they are assigned by a central authority, or the child's parents have the choice), and there may be a choice between multigrade schools and single-grade schools. Communities, politicians, unions, and others have views that will have to be included for a successful implementation. A significant increase in the efficiency is possible when the system is adopted. One of the key recommendations for a successful implementation includes a smooth transition from the present solution to the optimized solution.

The third part of this book deals with applications of enforcement and first responders. Chapter 11 by Murray discusses the well-known problem of locating fire stations. The objective function for the problem in a city in California minimizes the sum of fixed and operating costs, while constraints ensure that a certain percentage of the demand is actually satisfied. The author's model is a generalization of the well-known maximum capture problem. After the recommendations were made to the authorities, the city in question merged with another population agglomeration, which, in conjunction with a general economic downturn, resulted in the fact that the system was not implemented. The lessons of the process, though, remain for future users.

In Chap. 12, Gentili and Mirchandani describe a model for the location of sensors for travel time information. These sensors can be used to detect changes in traffic patterns to avoid congestion as well as plan future traffic networks. The application maximizes the distance that can be monitored by the sensors. The model is applied in an area of Texas. Tradeoff curves are determined for the location of additional sensors.

In Chap. 13, Bucarey et al. consider the problem of police districting. Their model is a largely extended p-median formulation that locates centers of districts, so as to minimize a linear convex combination of three objective functions, *viz.*, the sum of center-customer distances, a penalty for odd shapes of the district, and prevention demand. Based on the complexity of this mixed integer programming problem, the authors design a location—allocation heuristic to solve the problem. The model is applied to Chile, resulting in major benefits as compared to the present solution.

In Chap. 14, Marianov analyzes the location of jails in Chile. A mixed-integer optimization problem is formulated, which minimizes the costs of opening new

facilities, the costs of expanding existing facilities, a penalty cost for inmates' transportation, and the last term is another penalty cost of overpopulation in the jail. The author outlines some of the major difficulties with this subject, particularly changes in demographics, changes in the law, and major trends, such as the increase in drug-related offenses.

Chapter 15 by Pelot et al. deals with the location of Coast Guard vessels off the Canadian Atlantic Coast, in order to be able to efficiently respond to incidents, whose severity is measured in different categories. The model also uses different workload capacities, and the different versions are variations on max cover problems.

In the last chapter of the book, Bell surveys military applications of location analysis throughout different periods of recent history. The type of models used in the military spans a wide range from weapons positioning, the selection of training sites, hospital locations, locations for search and rescue aircraft, the location of spare aircraft engines, the closure of military bases, and many others.

Finally, it is our great pleasure to express our sincere thanks to our authors and their cooperation, as well as to Camille Price, Matthew Amboy, Christine Crigler, and Neil Levine. Without their assistance and encouragement, the task of putting this book together would have been a lot less pleasant. Many thanks to all of them.

H. A. Eiselt
Vladimir Marianov

Contents

Contributors

Amin Akbari Department of Industrial Engineering, Dalhousie University, Halifax, NS, Canada

Ines Arnolds Discrete Optimization and Logistics at the IOR, Karlsruhe Institute of Technology, Karlsruhe, Germany

Klaus Barber US Forest Service Region 5, Vallejo, CA, USA

Ayfer Başar Industrial Engineering Department, Istanbul Technical University, Maçka Campus, İstanbul, Turkey

Enrique Bassaletti Department of Criminal Analysis, Carabineros de Chile, Santiago, Chile

Rajan Batta Department of Industrial and Systems Engineering, University at Buffalo (SUNY), Buffalo, NY, USA

John E. Bell Department of Marketing & Supply Chain Management, University of Tennessee, Knoxville, TN, USA

Joyendu Bhadury Bryan School of Business and Economics, University of North Carolina at Greensboro, Greensboro, NC, USA

Alan Blatt Center for Transportation Injury Research, CUBRC, Buffalo, NY, USA

Burçin Bozkaya Sabanci School of Management, Sabanci University, Tuzla, İstanbul, Turkey

Victor Bucarey Department of Industrial Engineering, Universidad de Chile, Santiago, Chile

Mark L. Burkey School of Business and Economics, North Carolina A & T State University, Greensboro, NC, USA

Richard L. Church University of California, Santa Barbara, CA, USA

H. A. Eiselt Faculty of Business Administration, University of New Brunswick, Fredericton, NB, Canada

Marie Flanigan Center for Transportation Injury Research, CUBRC, Buffalo, NY, USA

Tejswaroop Geetla Department of Industrial and Systems Engineering, University at Buffalo (SUNY), Buffalo, NY, USA

Monica Gentili Department of Mathematics, University of Salerno, Fisciano, Salerno, Italy

Ross A. Gerrard USDA Forest Service, PSW Research Station, Davis, CA, USA

Ricardo Giesen Department of Transport Engineering and Logistics, Pontificia Universidad Católica de Chile, Santiago, Chile

James W. Gillula IHS Economics, Washinton, D.C., USA

Stanley E. Griffis Department of Supply Chain Management, Michigan State University, East Lansing, MI, USA

Eldon Gunn Department of Industrial Engineering, Dalhousie University, Halifax, NS, Canada

Özgür Kabak Industrial Engineering Department, Istanbul Technical University, Maçka Campus, İstanbul, Turkey

Li Li Department of Industrial Engineering, Dalhousie University, Halifax, NS, Canada

Vladimir Marianov Department of Electrical Engineering, Pontificia Universidad Católica de Chile, Santiago, Chile

Kevin Majka Center for Transportation Injury Research, CUBRC, Buffalo, NY, USA

Richard Middleton Los Alamos National Laboratory, Los Alamos, NM, USA

Pitu B. Mirchandani School of Computing, Informatics and Decision Systems Engineering, Arizona State University, Tempe, AZ, USA

Alan T. Murray Center for Spatial Analytics and Geocomputation, College of Computing and Informatics, School of Public Health, Drexel University, Philadelphia, PA, USA

Matthew R. Niblett University of California, Santa Barbara, CA, USA

Stefan Nickel Discrete Optimization and Logistics at the IOR, Karlsruhe Institute of Technology, Karlsruhe, Germany

Jesse O'Hanley Kent Business School, Kent University, Canterbury, UK

Paulo Rocha E Oliveira IESE Business School, University of Navarra, Pamplona, Spain

Fernando Ordóñez Department of Industrial Engineering, Universidad de Chile, Santiago, Chile

Ronald Pelot Department of Industrial Engineering, Dalhousie University, Halifax, NS, Canada

Y. İlker Topçu Industrial Engineering Department, Istanbul Technical University, Maçka Campus, İstanbul, Turkey

Samuel P Troy Bryan School of Business and Economics, University of North Carolina at Greensboro, Greensboro, NC, USA

Vedat Verter Desautels Faculty of Management, McGill University, Montreal, QC, Canada

Anthony M. Yezer George Washington University, Washinton, D.C., USA

Yue Zhang College of Business and Innovation, University of Toledo, Toledo, OH, USA

Chapter 1
Location Analysis in Practice

H. A. Eiselt, Vladimir Marianov and Joyendu Bhadury

1.1 Location Problems and Its Features

Location choices are as old as mankind. Individuals have always chosen their place of residence, starting with the appropriate cave, which would have to be in reasonable proximity to places in which food, and later work, could be found). Municipalities located central places to conduct business, such as the Greek *agora* or the Roman *forum* (or arenas for gladiator fights, for that matter), while leaders of countries located places for their administration, their armies, &, eventually, their burial sites (e.g., pyramids).

An interesting look at one of these early situations is provided by Wilamowsky et al. (1995), who demonstrated that Joshua's (of Bible fame) choice of locations for three cities for "unintentional murderers" were actually optimal sites. Along similar lines, ReVelle and Rosing (2000) showed Emperor Constantine the Great's choices for the locations of his armies in the fourth century AD. More specifically, there were eight regions of the Roman Empire, in which four existing field armies had to be positioned. The authors of the study used the concept of covering models to locate the field armies, as to require the smallest number of steps of relocation, in case an attack on any of the regions ensued. Actually, the emperor's choice was not optimal with respect to ReVelle and Rosing's proximity criterion; it appears that stationing two field armies in Rome was more of a political than strategic decision. Even in

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

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

H. A. Eiselt, V. Marianov (eds.), *Applications of Location Analysis,* International Series in Operations Research & Management Science 232, DOI 10.1007/978-3-319-20282-2_1

literature we can find location problems, albeit hidden somewhat. A good example is Lewis Carroll's (1880–5) *Tangled Tales*, where in *Knot 2: Eligible apartments*, the author has one of his protagonists state: "'One day-room and three bedrooms,' said Balbus, as they returned to the hotel. 'We will take as our day-room the one that gives us the least walking to do to get to it."' In location analytic terms, we are looking for the location of a median (the concept will be defined below).

Since that time, a large number of contributions have described possible, realistic, real, and implemented location models. The contribution by Eiselt (1992) surveys some of the applications of the 1980s as a lead-in to a special issue of applications in location analysis. A survey that investigates and summarizes contributions related to agriculture is found in Lucas and Chhajed (2004). In addition to the many facilities that have been located throughout the decades, e.g., warehouses, blood banks, fire stations (a good taxonomy is provided by Başar et al. 2012), ambulances, etc., (many of these real location problems since 1990 are surveyed in Table 1.1 in this paper), there are some unusual location problems. Among them are the optimal location of disinfection stations for water supply networks (Tryby et al. 2002) and the prime locations for street robbers (and, presumably, cops, who will try to foil attempts of the robbers to ply their craft), see Bernasco et al. (2013).

This discussion leads us immediately to the fact that many real location decisions will not rely on a single quantitative criterion, such as distances, but rely on other, often, intangible criteria. In personal location decisions (e.g., choosing a place to live), this could be the perceived beauty of the area, the perceived quality of the neighborhood including the type of school district, and many others. In public location decisions, public support/opposition, environmental impact, and similar concerns are often found. Private location problems, on the other hand, are typically simpler, as profit or cost are the main concerns. These criteria are easily quantifiable and fit very well into a mathematical model.

This chapter will analyze some of the foundations of location models, and determined classes of applications and their main features. In order to do so, we will very loosely and generally model a location problem, in which we have a metric space, in which customers (we use the term here in the widest possible sense), at least in the short to medium run, occupy fixed locations or move along known paths, and in which a decision maker will locate facilities based on known or perceived interactions between the facilities and the customers. These interactions will typically be transportation, either from a facility to a customer or vice versa.

Weber (1909) was among the first to look at actual transportation problems from a mathematical point of view. More specifically, the appendix of his book written by Georg Pick, minimizes transportation costs among three customers and one facility, which are assumed to be proportional to Euclidean distances between the points. Equivalently, this can be seen as finding the point in a triangle, such that the sum of distances between the new point and the vertices of the triangle are as small as possible. This problem was first posed by Fermat (formulated before 1640), solved by Torricelli in 1645, and for which a well-known iterative algorithm was described by Weiszfeld (1937). Since Weber's seminal work, geographers and economists have debated the location of facilities (mostly firms in the early days). It is interesting to

note that while many facilities that deal with the first sector of the economy (i.e., extraction, such as agriculture and mining) are located close to the source, many facilities of the third sector of the economy (the service sector) are located close to the customers. (We wish to note at this point that this assertion holds true for brick-and-mortar facilities, we do not include online firms or facilities in our discussion). There are many reasons for this phenomenon: Consider a firm in the first sector, e.g., a gold mine. Typically, we may expect about one ounce of gold in three tons of ore. Clearly, given the very low output-to-input ratio, nobody would transport the ore to sites close to the customer to extract the gold at that location. On the other hand, facilities in the service sector have extensive dealings with customers, so that their locations are chosen quite naturally in close proximity to customers.

Consider now a typical scenario, in which multiple facilities of one type exist. In such a case, there are two issues to be addressed: on the one hand, does the firm or the customer choose the facility he is dealing with? The second issue concerns the facility customer interaction. In particular, we distinguish the case in which the firm is responsible for the interaction (e.g., the transport), and the case in which this interaction is done by the customer. This leaves us with four cases. In the first class of models, the firm chooses the facility the customer is served from and the firm also deals with the interaction. Typical examples of this case are trucking, waste collection, ambulances, and many others. Models with this feature are often referred to as *allocation* (as the firms allocate facilities to their customers) or *shipping* models. On the other hand, consider models in which customers choose the facility they are dealing with and are also responsible for the interaction with the chosen facility. Models in this category include retail stores, libraries, ATM machines, and other, service-related facilities. Models of this type are frequently called *customer choice* models or *shopping* models. Much less are "mixed models", in which one agent is choosing the facility, while the other is responsible for the interaction. Examples for the firm (in the widest sense, typically a government agency) chooses the facility customers are to patronize, while customers are responsible for the interaction. Scenarios of this type include polling stations, public schools, and similar facilities. Finally, there are rather few cases, in which customers choose the facility, while the firm takes care of the interaction. At first glance, it appears that a customer's choice of the branch of a chain of retail, e.g., furniture, stores falls into that category, as the store may then ultimately take care of the interaction with the customer by delivering the goods. While this is true, the customer cannot choose (and has no interest in choosing) the warehouse, from which the furniture is delivered to him. This is then again a standard allocation/shipping model.

Another possible way to distinguish between location models was suggested by ReVelle et al. (1970), see also ReVelle and Eiselt (2005), where the authors distinguish between the location of public and private facilities. In some cases, such a distinction will not reveal any insight. Consider, for instance, the location of a library, typically a public facility. The city or whichever government department is responsible for the planning, will try to make the library accessible to as many people as possible, so that the average facility—customer (or, in this case, potential patron) is as short as possible. On the other hand, a firm that plans the location

of a regional distribution center will locate the center in such a way that the sum of distances from the center to the customers is minimized, as this can be seen as a proxy for cost. (Note that this argument is not as straightforward as it appears: first, it assumes that special trips are made to the customers rather than milk runs to serve multiple customers on a single trip, which would result in much more difficult location-routing problems, and secondly that there is an established cost per ton-mile for trucking, such as $ 1.38 as illustrated in The Trucker's Report 2015). In other words, the decision maker's behavior will be the same (locating the facility in close proximity to the customers), while the underlying motives differ: accessibility for the library planner, and cost minimization for the trucking firm. On the other hand, the public vs. private distinction may be quire relevant in other aspects: typically, public decision makers face many stakeholders in their decision making and, as a result, many conflicting objectives and criteria are to be included. On the other hand, private decision makers can much more easily agree on (profit maximizing or cost minimizing) objectives, which are much easier to incorporate in optimization models given the relative ease with which they can be quantified.

As outlined above, location models deal with the spatial separation and the interaction of customers and facilities. The separation of facilities and customers is expressed in terms of their distance (shortest path in networks and some Minkowski distance in the plane are the most frequently used measures). These distance can then appear in the objective function of a mathematical optimization problem as is the case in typical median or center problems, or they appear in the constraints, as is the case in covering models. Naturally, mixed forms are possible and have been used. In addition to the usual models in which proximity of facilities to customers is desired (they minimize a function of the distances, where customers "pull" a facility towards themselves, see, e.g., Eiselt and Laporte 1995) or undesirable (where the customers "push" the facility away from their location, in case the facility is deemed undesirable), there are models in which only the distances between facilities are relevant. Typical examples are dispersion and defender models; see, e.g., Daskin (2008).

1.2 A Locator's Toolkit

This section will delineate some tools locators have at their disposal. We will first survey generic formulations of a number of the standard location models. We then discuss a few of the approaches to multiobjective optimization, and finally, we will review some of the main techniques in multicriteria decision making.

In order to facilitate our discussion, we first distinguish between two main categories of location problems: discrete problems and continuous problems. Whereas there exists only a finite number of potential locations in discrete problems, there is an infinite number of location to choose from in continuous problems. If a discrete problem is formulated on a network, the finite set of potential locations is often equal to (or a subset of) the set of nodes. This is, however, not a necessary requirement.

The *location variables*, i.e., the variables that deal with where to locate a facility in the given space, in discrete spaces are simply defined as zero-one variables: for each potential facility location, a variable is created that assumes the value of one, if we decide to locate at that site, and zero otherwise. Such an approach is obviously not feasible in continuous problems, in which an infinite number of potential locations exists. In such cases, in which a facility can either be located in the plane or, in more general cases, in any d-dimensional space, or at any point on a network, the facility location can then be determined by the coordinates of the facility location in space, or the distance of the location from any existing point. To simplify our exposition, we will describe the main location models by using discrete location problems.

In order to do so, first define a vector of *location variables* $y = (y_1, y_2, \ldots, y_n)$ that indicate whether or not a facility will be located at the potential locations. In addition to the location variables y_j, we need to refer to customers $i = 1, \ldots, m$ and, at least whenever multiple facilities are to be located, *allocation variables* $\mathbf{X} = (x_{ij})$, where $x_{ij} = 1$, if customer i is served by facility j (or, more precisely, if the customers at site i are served by a facility at site j). Furthermore, define $\mathbf{w} = (w_i)$ as a vector of weights (which, for instance, denote the number of customers at customer site i), $\mathbf{D} = (d_{ij})$ as the aligned of distances between customers i and potential facility sites j, and $\mathbf{f} = (f_j)$ the fixed costs of a facility, if it were to be constructed at site j. We also need $\mathbf{g}(\mathbf{X}, \mathbf{y})$ as a set of constraints and S as the set of feasible solutions. We can then easily describe some of the main types of location problems.

First, consider the *simple plant location problem* (*SPLP*) and its close cousin, the *capacitated plant location problem* (*CPLP*). The main idea is to locate a number of facilities (its number is determined endogenously), so as to minimize the sum of facility location and allocation, i.e., transportation, costs. Typically, the constraints will ensure that all customers receive service, and that customers can only be served from existing facilities. The closely related capacitated version of the problem adds realism to the problem by associating a capacity to each facility. (Actually, it would also be possible to consider a finite number of possible capacities to different versions of facilities, each of which would then have its own cost). While the introduction of capacities appears to be a minor step in the sense that it just adds a few constraints to the problem, the consequences are much more severe. While the optimizer will automatically assign each customer to its closest facility (by virtue of the minimization objective function), this may no longer be possible if capacities are introduced. This also gives rise to two different versions of the capacitated problem: the single-source assumption, where all customers at one site are served by the same facility, and the multiple source version, in which this is not the case. In some applications, the single-source problem is more realistic, for instance in waste collection, in which the collected garbage anywhere in one town is hauled to the same landfill or transfer station.

The simple plant location problem and capacitated plant location problems can be formulated as follows:

$$\begin{aligned} SPLP/CPLP\text{: Min } & \mathbf{wDX} + \mathbf{fy} \\ \text{s.t. } & \mathbf{g}(\mathbf{X}, \mathbf{y}) \in S \\ & \mathbf{X} \in \{0, 1\}^{n \times n} \\ & \mathbf{y} \in \{0, 1\}^n. \end{aligned}$$

This problem is actually a general version of a number of other well-known problems. The best-known such problem is the *p-median problem* (*p-MP*). Loosely speaking, it requires us to locate a given number of p facilities, so that, given that each customer is served by its closest facility, the total transportation costs are minimized. As the *SPLP* and *CPLP* before it, this model assumes that each service requires a special trip. It can be formulated as

$$\begin{aligned} p - MP\text{: Min } & \mathbf{wDX} \\ \text{s.t. } & \mathbf{ey} = p \\ & \hat{\mathbf{g}}(\mathbf{X}, \mathbf{y}) \in S \\ & \mathbf{X} \in \{0, 1\}^{n \times n} \\ & \mathbf{y} \in \{0, 1\}^n, \end{aligned}$$

where $\mathbf{e} = (1, 1, \ldots, 1)$ is the vector of ones and $\hat{\text{g}}$ refers to the appropriate constraints. As a matter of fact, the *p-MP* can be derived from the *SPLP* by setting $\mathbf{f} = \mathbf{0}$ and introducing the additional constraint $\mathbf{ey} = p$. In other words, while the number of facilities in the *SPLP*/*CPLP* is endogenous, it is exogenous in the *p-MP*.

The *anti p-median* (also referred to as *maxian*) is a problem very similar to the p-median, except that it attempts to maximize the average distance between facilities and customers. It was designed to be used for hazardous or noxious facilities, which customers would like to push away as much as possible from their own location. However, formulating *p-MP* as an anti-median problem takes more than simply replacing the "Min" by a "Max" function. The reason is that the "Min" function automatically allocated each customer to its closest facility. Replacing it by a "Max" objective will result in customers automatically being allocated to their farthest facilities. However, what the planner is trying to do is to push the closest, not the farthest, facility as far away from himself as it is the closest facility that is polluting his location most. In order to formulate this, we will need additional $O(n^2)$ constraints that will guarantee the appropriate allocation.

Consider now the *location set covering problem* (*LSCP*). Its purpose is to minimize the sum of facilities that are to be located while ensuring that each customer has a facility within a predetermined service distance from himself. Many such covering problems are found in the context of the location of emergency facilities, such as fire halls, hospitals, police stations, and others. The *LSCP* can be seen as a special case of the *SPLP* by deleting the allocation variables $\mathbf{X}$ and setting the fixed setup costs of all potential locations equal to each other, i.e., $\mathbf{f} = \mathbf{e}$ The problem can then be written as

$$\begin{aligned} &\textit{LSCP}\text{: Min } \mathbf{ey} \\ &\text{s.t. } \mathbf{g}'(\mathbf{y}) \in S \\ &\mathbf{y} \in \{0,1\}^n. \end{aligned}$$

In many applications, covering each customer is not feasible, especially in situations, in which customers are very much spread out in the given space. In such case, which gave rise to *the max cover problem* (*MCP*); see, e.g., Church and ReVelle (1974). In this problem, the decision maker specifies p, the number of facilities to be located and, given that each customer node has a weight associated with it, the objective is then to maximize the total weight that can be covered within a given distance from the facilities. The problem can be formulated as follows:

$$\begin{aligned} \textit{MCP}\text{: Max } &\mathbf{wz} \\ \text{s.t. } &\mathbf{ey} = p \\ &\mathbf{g}(\mathbf{y}, \mathbf{z}) \in S \\ &\mathbf{y}, \mathbf{z} \in \{0,1\}^n, \end{aligned}$$

where $\mathbf{z}$ is a vector of zero-one variables whose i-th component equals1, if customer i is within a prescribed distance from its closest facility, and 0 if it is not.

There are other problems such as *center problems* (in which the objective is to locate facilities, so that the longest customer—facility distance is as short as possible) and *anti-center problems* (in which the shortest customer—facility distance is as long as possible), but very few of them have been applied in practice. One of the problems associated with these problems is their unique focus on the worst case. Potentially, objectives of this type can be applied as secondary objectives in cases in which the worst case is to be avoided if at all possible, such as in hazmat transportation or similar instances.

Many practical location problems have more than a single objective. In general, we will distinguish between *multiobjective (linear) optimization problems* (*MOLP*), which are typically mixed-integer optimization problems, *and multicriteria decision-making problems* (*MCDM*), which feature a small number of decision alternatives, whose consequences are measured on a number of different criteria. Among multiobjective optimization problems, *vector optimization* is a prominent approach. First described by Zeleny (1974), vector optimization problems are designed to generate nondominated solutions, i.e., solutions which cannot be improved upon on all objectives while remaining in the feasible set. Since the generation of all nondominated solutions is not practical, techniques have been devised to approximate the set of nondominated solutions. Cohon (1978) described a number of techniques for that purpose. Prominent among them are the weighting method and the constraint method. The weighting method associates positive weights with each of the objectives (typically explained as measuring tradeoffs between units of the objectives) and then constructs a composite objective as the sum of the weighted individual objectives. The resulting single-objective problem can then be solved by the pertinent methods. The constraint method chooses one

objective to remain part of the model as an objective, while all other objectives are transformed into constraints by using upper or lower limits on them. These limits can, and usually are, parametrically modified in a sequence of sensitivity analyses. While the constraint method will always find an extreme point of the original feasible set, the weighting method may not.

Another approach to multiobjective optimization problems is *goal programming*. Based on Charnes et al. (1955) and Charnes and Cooper (1961), Ignizio (1982) was one of the main proponents of the approach. Its basic premise is as follows. Constraints are rarely as rigid in real life as they are in a mathematical optimization problem. For instance, a budget typically requires that the amount of money spent does not exceed the amount available. Inserted in an optimization problem, such a constraint is absolute and cannot be violated. In practice, though, it is, of course, possible to borrow money, albeit at a cost (interest). This feature is captured in goal programming by using so-called deviational variables d_k^+ and d_k^- for positive and negative deviations from a set target value of t_k, respectively. The variables d_k^+ and d_k^- are commonly interpreted as "overachievement" and "underachievement," respectively. As an illustration, consider a covering problem, in which it is desired to cover each customer at least once. Each customer, who is not covered, causes a penalty of, say, 5, while each customer, who is covered more than once receives a penalty of 1 for each unit of excess coverage. The reason for the latter are "equity" considerations, which, at least in public decision making, are often considered rather important. As above in the *MCP*, define a zero-one variable y_i, which is equal to 1, if a customer at site i is covered and 0, if this is not the case. In addition, let the vector $\mathbf{a}_{i\bullet} = (a_{ij})$ equal 1 in its j-th component, if customers at site i can be covered by a facility at site j. Then $\mathbf{a}_{i\bullet}\mathbf{y}$ equals the number of times customers at site i are covered, and we can formulate the goal constraint

$$\mathbf{a}_{i\bullet}\mathbf{y} + d_k^- - d_k^+ = 1,$$

where the subscript of the deviational variables simply indicates that this is the k-th goal. As usual, we assume that the deviational variables have to satisfy the nonnegativity constraints. Hence, if customer i is not covered, $d^- = 1$ and $d^+ = 0$, indicating that there is a coverage deficit of 1. On the other hand, if customer i is covered twice, $d^+ = 1$ and $d^- = 0$, indicating that there is an excess coverage of 1. In case customer i is covered exactly once, both deviational variables will assume the same value. Consider the penalties of 5 and 1 for overachievement and underachievement of the target value, respectively, the contribution to the objective function by this goal constraint is then

$$\text{Min} \ldots + 5w_i d^- + 1w_i d^+ \ldots$$

This part of the objective will penalize each solution that does not cover site i and its w_i customers by 5 for each customer, while each customer generates one unit of penalty for each time he is covered more than once. Furthermore, in case the target of single coverage is satisfied, both deviational variables will not only be equal, but

will equal 0, as any other solution with $d^+ = d^-$ will have a higher value of the objective function in the minimization function.

Finally, goal programming problems come in a variety of versions. A popular, albeit somewhat problematic, version features preemptive priorities, where a higher-ranking goal is considered "infinitely more important" than a goal on the next lower level, so that no finite tradeoff exists. Another important issue concerns the commensurabilities, as deviational variables from different constraints are added in the objective. Care must be taken that they are transformed to similar units.

The last multiobjective optimization technique reviewed here is *fuzzy programming*. Fuzzy logic was pioneered by Bellman and Zadeh (1970). It deals with the vagueness of an achievement, such as "this machine will cost a lot more than $ 250,000." In order to deal with the problem, a fuzzy membership function is set up that assigns a degree of membership to each achievement or outcome, for which a lower and an upper bound are specified by the decision maker. The value of this membership function is a number between 0 and 1; it assumes a value of 1, if the (maximization) objective achieves the upper bound or is even higher, it assumes a value of 0, if the objective falls short of the lower bound, and is a value in between for achievements between lower and upper bounds. A maximin function with the membership function values as achievements is then the single objective, which can then be solved with the pertinent methods.

Consider now multicriteria decision making problem, in which a decision maker faces a "payoff" matrix $\mathbf{C} = (c_{ij})$, so that c_{ij} denotes the achievement of solution i on criterion j. For simplicity, it is often assumed that all criteria are of the "maximization" variety, i.e., higher values indicate better solutions. In the context of location models, the decisions (rows) refer to the potential locations for the (single) facility. An interesting reference that illustrates the use of multicriteria decision analysis to location problems is Larichev and Olson (2001). The book includes models that site facilities to dispose of hazmat, pipeline locations, and the selection of municipal solid waste systems.

A simple *generic method* for multicriteria decision making problems is as follows. Suppose that the decision maker were able to specify weights λ for all criteria ℓ, then aggregate achievements could be determined by calculating weighted average achievements for all decision, which then allows the decision maker to choose the decision with the highest average achievement. Some refinements are easily incorporated: a decision maker could delete decisions whose achievements on specific crucial criteria fall short of a prescribed threshold, or achievements beyond some upper bound will no longer considered to be relevant as contributing towards the aggregated average. Clearly, any version of this technique will have to be followed by extensive sensitivity analyses.

The next class of techniques to deal with multicriteria decision problems are so-called *reference point methods*. One technique in this class is *TOPSIS*, and it was first described by Hwang and Yoon (1981). The idea of these techniques is to specify a norm or target value and then measure how far the individual decisions are from this absolute value. In addition to a metric to measure the distance, this requires the specification of tradeoffs between criteria. Note the similarity of this approach and

goal programming and its target values. A number of interesting approaches in this class are described by Eiselt and Marianov (2014).

In contrast to reference point methods, which compare decisions and their outcomes to exogenously given standards, *data envelopment analysis* (*DEA*) compares decisions with each other and determines whether or not they are efficient. In order to do so, analysts first have to group the input factors (such as costs, required manpower, and others) as well as the output factors of the locations (employment, pollution, and others) together. Then a set of simple linear programming problems will determine whether or not a decision (we will consider one decision at a time) is efficient compared to the other decisions. Note that this approach only demonstrates relative efficiency, as a decision may be the best of the lot, while it may, in the grand scheme of things, still be quite terrible and unacceptable.

Preference cones are an interesting and relatively easy to incorporate concept in multicriteria decision making. The idea is that a decision i dominates a decision ℓ, if the aggregated outcome of the *i*-th decision with any set of nonnegative, finite weights (as in the generic method described above) is no less than the aggregated outcome of the· ℓ-th decision for the same weights. The advantage of this method is that it requires no numerical input from the decision maker. However, it may very well turn out that no dominances exist, so that no reduction of the system is possible. We are not aware of any applications in location analysis that have applied this concept.

Another approach is taken by *outranking methods*. In essence, they apply a weak dominance concept and enable a persuasive visualization. Outranking methods were first suggested by Benayoun et al. (1966), with Roy's (1971) *ELECTRE* methods (many other versions exist) providing the first workable concept. The basic technique requires that the decision maker specify weights for each of the criteria that indicate its relative importance. The technique then compares each pair of decisions and determines a concordance aligned and a discordance aligned on the basis of the weights. More specifically, the concordance of one decision over another equals the sum of all weights, in which the former decision is preferred over the latter. The discordance aligned can be constructed according to similar lines. A graph, in which each node represents a decision then has an arc leading from a node i to a node ℓ if the concordance of decision i over decision ℓ exceeds a prespecified threshold *and* the discordance of that relation does not exceed a prespecified threshold.

Based on earlier work by Brans in the early 1980s, Brans and Vincke's (1985) *PROMETHEE* method takes a different route. On each criterion separately, the method determines the differences of the achievements for each pair of decisions. This difference is then translated into a preference by way of a preference function. These preference functions could be step functions, piecewise linear, or nonlinear functions, not unlike decay functions in covering models (see, e.g., Church and Roberts 1983, or Berman and Krass 2002). Given the weights of the criteria specified by the decision maker, these preferences are then aggregated. The aligned of these aggregated preferences is then the basis, on which the graphs for

the different versions of *PROMETHEE* are constructed. Karagiannidis and Moussiopoulos (1997) and Hokkanen and Salminen (1997) are among the authors who have suggested to apply outranking methods to location problems.

Finally, techniques that allow inconsistent evaluations by decision makers have become quite popular. Most prominent among them is the *analytic hierarchy process* (*AHP*), or it successor, the analytic network process *ANP*). These techniques are based on the seminal work by Saaty (1980). There is a large number of studies that have suggested using the method for actual applications; see, e.g., Aragonés-Beltrán et al. (2010) and Tavares et al. (2011). The technique requires decision makers to specify their degree of preference between each pair of decisions on all criteria separately. While reciprocity must be satisfied asserting that, if decision i is, say, considered three times as good as decision ℓ, this necessitates that decision ℓ is considered only $^1/_3$ as good as decision i, transitivity does not (i.e., the decision maker may determine that decision i is twice as good as decision ℓ which, in turn, is three times as good as decision r, while claiming that decision i is five times as good as decision r on the same criterion). After this substantial input, the method will provide the decision maker with two outcomes: first, it will determine the degree of consistency of the decision maker's input (which means, that if it falls short of an agreed-upon threshold, the decision maker will have to rethink his evaluations), send secondly, the technique will provide a final ranking of the decisions under consideration. There has been much discussion concerning the theoretical underpinnings concerning the method, and some alternatives have been suggested, such as the *geometric mean method*. Surveys of this and many other methods can be found in Olson (1996) and in Eiselt and Sandblom (2004).

1.3 Location Applications in the Literature

This section will summarize applications of location analysis that have been reported in the literature. While there are many contributions that use real data to test their models and methods, not all of them are actual applications. For the purpose of this paper, we would like to distinguish between *realistic applications* and *real applications*. Realistic applications are those papers, whose main focus is on the description of a new model or method, which is subsequently tested on data that have been gleaned from real life. On the other hand, a real application's main focus is on the actual application. Clearly, the distinction between realistic and real applications is fuzzy, which may explain why we may have not included some papers others may deem true applications.

In our discussion, it will be beneficial to distinguish between the firm's view and the (collective) customers' view. More specifically, while in both cases it is the firm (or government agency or whoever the planner may be) that locates the facility in question, the two distinct angles of view emphasize who the ultimately beneficiary will be. Models that espouse the firm's view typically include a simple cost minimization or profit maximization objective. This does, of course, not mean that the

models themselves are necessarily simple: they may (and often do) include complex functions that attempt to model customer behavior. Modeling the customers' view is much more contentious, as a single model will have to somehow express the collective view of customers. As an example, consider the (re-) location of a hospital. Here, a planner (typically not a representative of the customers themselves) will set up objectives for the customers' benefit. In doing so, he will usually consider average travel times of the patients to the hospital, and try to locate that facility, so that as many potential patients are within a specific travel time from the hospital. Such a planner may also wish to include in the model some measure of "equity," which is almost always ill-defined and, in order to find a quantifiable proxy, replaced by a measure of equality.

We would like to point out that there is no clear relation between the firm's and the customers' view on the one hand and the shipping and shopping models on the other: in case of models that take the firm's view, firms can locate trucking terminals in the context of a delivery (i.e., shipping) model, or may locate retail outlets (a shopping model). Similarly, in case of the customers' view, planners may locate ambulances (a shipping model, as dispatchers will decide which ambulance to use and they will be responsible for the transport), or they may locate libraries (a shopping model).

Before we go into any further detail, we would like to offer a few thoughts on the scale of the decision-making process. Two major tools from the vast toolkit of location planners are mathematical optimization problems and methods in multicriteria decision making with multiobjective models occupying a middle ground between the two. More often than not, simple, single-objective optimization models are used on the macro scale in the first stage of a decision-making process. The solution may identify a specific point, which decision makers will usually take as an indication of the general area, in which the facility in question should be located. Subsequently, the process zooms in on this area and at this point, more things are asked of the location, which is accomplished by considering only a smaller number of available locations, and additional criteria. This is what multicriteria approaches do.

Models that use the firm's view may, in addition to the usual cost minimization or profit maximization objective, include features that relate to customer behavior. For instance, when locating a gas station, planners often look not at customer location, but at typical trips made by customers, as it has been observed that customers usually fill up their cars on their way to work or while they are on trips for other purposes. Similar behavior has been observed for other small purchases and the use of child care. Once such a criterion has been determined, planners may include a flow capturing (or flow interception) feature in their model. For details on these approaches, see, e.g., Hodgson (1981).

Another feature often included in models that use the firm's view is routing. The main reason is that the usual median models assume that each facility—customer interaction requires a separate trip. Once this is no longer true, routing features may be incorporated in the model. Clearly, this is a serious complicating factor: p-median problems as well as even (for practical purposes) simple routing problems such as

standard vehicle routing are **NP**-hard by themselves, so that in most cases a heuristic method for the solution will be required. For a survey, readers are referred to Nagy and Salhi (2007). In addition to the many existing heuristics and metaheuristics, ranging from simulated annealing to tabu search, neighborhood search, genetic algorithms, and many others (for a survey, see, e.g., Richards 2000) Nagy and Salhi (2007) provide a number of heuristic methods, which are tailor-made for location-routing problems. They distinguish between *sequential methods* (which solve the location problem first, followed by a routing heuristic), *iterative techniques* (which shuttle between location and routing components—a sequential method with a feedback loop), *clustering-based methods* (which subdivide the set of customers into clusters, find a route within each cluster, and then determine the facility locations), and hierarchical techniques.

The most frequently used feature included in models that use the customers' point of view is accessibility and "equity." A good example is the work by Burkey et al. (2012), whose hospital model uses the average time for a customer to reach a hospital as one criterion, the second is the coverage of as many potential patients within a half-hour travel time (this can be considered providing equitable service), and also the Gini index (Gini 1912 as well as Ceriani and Verme 2012), which uses the Lorenz curve to measure the degree of equality of a solution.

In addition to these typical concerns included in individual models, there are also industry-specific features that may be included. Typical examples include the access to sources of water in case of fire stations, energy and materials recovery in case of the location of facilities in solid waste management, penalties for overcrowding in the planning of jails, proximity to highways, ports or airports in case distribution centers or manufacturing companies are to be located, to sources of pure water for beer breweries.

Table 1.1 below summarizes some of the applications of location analysis that were published in the last 15 years. The table concentrates on the main features of the model, the objectives, the solution approach, the type of facility, the country it is located in, and the journal in which the piece was published. The exact references are found at the end of this section. To save space in the table, we are using some abbreviations: *EJOR* refers to the *European Journal of Operational Research*, *C & OR* stands for *Computers & Operations Research*, *JORS* denotes the *Journal of the Operational Research Society* and *ITOR* is the *International Journal of Operational Research*.

Table 1.1 Summary of location applications 1990–2015

Year of application	Country of application	Facilities to be located	Methodology	Objectives	Authors of study	Journal
1990	U.S.	Telemarketing centers	Integer programming	Min communication, labor, & real estate costs	Spencer III et al.	*Interfaces*
1990	Arizona	Emergency medical vehicles	Nonlinear integer programming model	Max the expected number of calls that can dealt with within 8 min	Goldberg et al.	*EJOR*
1990	U.S.	Service terminal	Analytic hierarchy process	Operating costs, availability of staff, ease of service, convenience for customers, etc.	Hegde, Tadikamalla	*EJOR*
1992	Australia	Maintenance deports (roads)	Max cover	Distance	Rose et al.	*EJOR*
1993	Connecticut	Closure of a fire station	Spatial waiting line model	Retain first due travel time, service	Swersey et al.	*Interfaces*
1994	Kentucky	Ambulances	Expected covering, simulation	Max coverage within 10 min	Repede, Bernardo	*EJOR*
1994	Northern Ireland	Ambulances	*p*-median	Distances/response time	McAleer, Naqvi	*EJOR*
1995	Turkey	Brewery	Integer programming	Min transport costs & inventory costs	Köksalan et al.	*EJOR*
1995	Connecticut	Vehicle emissions testing stations	Set covering, queuing, simulation	Stay within service constraints (distance & waiting time)	Swersey, Thakur	*Management Science*
1996	U.S.	Multimodal rail plants	Integer programming	Min logistics costs	Miller et al.	*Location Science*
1996	U.S.	Blood collection & distribution facilities	Integer programming with different scenarios	Min transportation costs	Jacobs et al.	*Interfaces*
1996	Israel	Hospital	Nonlinear & integer programming, analytic hierarchy process	Min transportation costs, penalty for overcrowding, AHP with population diversity, employment & service	Mehrez, Sinuany-Stern, Arad-Geva, Binyamin	*JORS*

Table 1.1 (continued)

Year of application	Country of application	Facilities to be located	Methodology	Objectives	Authors of study	Journal
1997	India	Propane bottling plants	Integer programing	Min the sum of facility costs & transportation costs	Sankaran, Raghavan	*Interfaces*
1998	Dubai	Fire stations	Set covering & 11-criterion integer goal programming	Min cost, time, min service overlap, min water problems, &c.	Badri et al.	*EJOR*
1998	Worldwide	Army installations	Dynamic integer programming	Min allocation, closing, & other costs	Dell	*Interfaces*
1999	Portugal	Landfills & transfer stations	Integer programming	Min sum of transportation costs & location setup costs	Antunes	*Interfaces*
1999	Turkey	Malt processing plants	Integer programming	Min discounted transportation & facility opening costs	Köksalan, Süral	*Interfaces*
2000	U.S.	Manufacturing/distribution plant	Vector optimization	Max profit, min access time, max local incentives	Melachrinoudis, Min	*EJOR*
2001	Hong Kong	Satellite depots	Integer programming	Min installation & transportation costs	Cheung et al.	*Interfaces*
2002	Georgia & Quebec	Preventive health care facilities	Integer programming	Maximize the number of participants in preventive health care	Verter, Lapierre	*Annals of OR*
2002	Colorado	Fire station	Graphical aggregation	Proximity, soft constraints	Hewitt	*Interfaces*
2002	Finland	Landfill	*MCDM* (stochastic acceptability analysis)	Environmental, cost, property values, land use, etc.	Lahdelma et al.	*EJOR*
2002	Worldwide	Repair-parts warehouse	Analytic network process	Costs, taxation, risk, labor, delivery time, etc.	Sarkis, Sundarraj	*EJOR*
2004	Illinois	Automatic meter reading machines	Integer programming	Location set covering model	Gavirneni et al.	*Interfaces*
2004	North Carolina	Ambulances	Integer programming	Max expected covering	Tavakoli, Lightner (2004)	*C & OR*

Table 1.1 (continued)

Year of application	Country of application	Facilities to be located	Methodology	Objectives	Authors of study	Journal
2005	Florida	Disaster recovery centers	Heuristics	min avg distance, min max distance, min number of centers to be opened, max the prob that a center is available after disaster	Dekle et al.	*Interfaces*
2005	Chile	Jails	Integer programming	Location costs, expansion costs, penalties for overpopulation, transportation for shuttling inmates between jails	Marianov, Fresard	*JORS*
2006	Italy	Organ transplant centers	Integer programming	Transportation costs plus penalty term for inequality between regions	Bruni et al.	*Health Care Mg'mt Sci*
2007	Portugal	Disposal facilty for treated wood	Integer programming	*p*-median	Gomes et al.	*Envir Mod & Software*
2007	Spain	Incinerators	Tabu search	Location-routing	*Caballero* et al.	*EJOR*
2008	U.S.	Military units to be closed	Integer programming	Min net present value of (various types of) costs	*Dell* et al.	*Interfaces*
2008	Greece	Different solid waste facilities	Lexico minimax	Min greenhouse effect, min disposal to landfill, max energy recovery, max material recovery, min total cost.	Erkut et al.	*EJOR*
2008	Ohio	Park & ride facilities	Noninferior solutions & tradeoff curves.	Max demand covered, min avg travel time, max the number of existing facilities that are used.	Farhan, Murray	*C & OR*

Table 1.1 (continued)

Year of application	Country of application	Facilities to be located	Methodology	Objectives	Authors of study	Journal
2009	Florida	Hydrogen refueling stations	Integer programming with different scenarios	*p*-facility flow capturing model	Kuby et al.	*Int J of Hydrogen Energy*
2009	British Columbia	Location of clinical services	Dynamic integer programming	Min linear convex combination of patient travel distances & & deviation of solution from existing solution	Santibáñez et al.	*Interfaces*
2009	./.	Diopter strength	Integer programming	*p*-median, *p*-center, covering	Francis	*Interfaces*
2009	Texas	Priests	Integer programming	Min traveling, max coverage	Butler et al.	*Interfaces*
2010	Greece	Landfills for electronic waste	*ELECTRE* III	Cost, accessibility, public acceptance	Achillas et al.	*Waste Management*
2011	Portugal	Charging stations for electric vehicles	Integer programming	Max demand coverage—number of stations located	Frade et al.	*Transp. Res. Record*
2011	Worldwide	CARE warehouses	Integer programming	Min average response time	Duran et al.	*Interfaces*
2011	Turkey	Fire stations	Multiperiod integer programming	Max cover problem	Çatay	*OR Insight*
2012	China	Bank branches	Integer programming	Max cover	Wang et al.	*Interfaces*
2012	Spain	Resort	*MCDM* Borda's voting/consent method	Water quality, existing facilities, environment, etc.	Crecente et al.	*Landscape and Urban Planning*

Table 1.1 (continued)

Year of application	Country of application	Facilities to be located	Methodology	Objectives	Authors of study	Journal
2012	Chile	Schools	Integer programming	Opening closure, operating, transportation, & tuition costs	Araya et al.	*ITOR*
2013	Portugal	Recycling plant	Cost-benefit analysis, no optimization	Costs	Coelho, deBrito	*J of Cleaner Production*
2013	Turkey	Fire stations	Integer programming	Covering problems	Aktas et al.	*Interfaces*
2013	Serbia	Landfill	Fuzzy analytic hierarchy process, *Vikor*	Hydrogeological, meteorological, spatial, socio-political, legal, economic criteria	Milosevic, Naunovic	*Waste Management & Research*
2013	Sierra Leone	Landfill	Analytic hierarchy process, weighted linear combination	Water, topography, slope, access, cost of land, transportation costs	Gbanie et al.	*Applied Geography*
2014	Columbia	Bioethanol processing plant	Integer programming	Max total benefits	Duarte, Sarache, Costa	*Energy*
2014	Pakistan	Warehouses	Integer programming	min (all sorts of) costs	Brahimi, Khan	*4OR*
2014	Italy	Schools	Integer programming	*p*-median	Bruno, Genovese, Piccolo, Sterle	*Procedia–Soc and Behav Sci*
2014	Brazil	Storage hubs for soybeans	Integer programming	min costs	Milanez, deAzevedo et al.	*Simp Brasil de Pesquisa Oper*

Table 1.1 (continued)

Year of application	Country of application	Facilities to be located	Methodology	Objectives	Authors of study	Journal
2014	South Korea	Battery exchange stations	Integer programming	Modified p-median with max distance constraints	Ko, Shim	*Int J of Sustainable Transportation*
2015	Turkey	Freight villages	Integer programming	Min transport cost	Aksoy, Özyörük	*Appl. Math Modeling*
2015	China	Watchtowers for forest fire monitoring	Integer programming	Covering problems	Bao et al.	*Fire Safety Journal*

1.4 Summary & Outlook

This paper has described some of the main approaches to location analysis under special consideration of different types of applications, followed by a survey of descriptions of actual applications of location analysis. As expected, the main tools are various types of integer programming and techniques from multicriteria decision-making. The summary also reveals that location models are applied all across the globe. In addition to the classics: finding optimal locations for landfills, fire stations, ambulances, and bank branches, there are a number of new trends. It is hardly surprising that many of these are fueled by technology: (mobile) vehicle inspection stations (see the piece by Geetla *et al*. in this volume), charging stations for electrical vehicles (or those that use hydrogen fuel), cell phone towers, wifi hot spots, toll stations for electronic road pricing, and similar facilities. What is entirely missing are applications for non-physical location problems, such as models that position employees in positions in a firm (finding the best fit), locating brands that locate close to existing customer tastes, and other "facilities." The groundwork for such location models has been made a long time ago (Niemi and Weisberg 1976 for political models, and Eiselt and Marianov 2008, for employee positioning), but the calibration of data is probably one of the key difficulties for these models. One of the very positive approaches that exemplify "thinking outside of the box" is presented by Swersey et al. (1993), who, in their quest to keep public services at an acceptable level, while saving money, considered dual training of first responders as firefighters and paramedics. This integration of services appears very promising.

Still, the survey also demonstrates that modeling is pretty much still at the "ad hoc" level. In other words, there is not generally accepted "fire station location model" that functions as transferable technology and is, with appropriate modifications for the local situation, usable in many different places. It would not free modelers from the task of putting together a problem description for their specific situation, but it would allow them to skip the basic steps, which are already done, and skip to the fine tuning right away. This has not yet happened, but, as the saying goes, hope springs eternal.

References

Aragonés-Beltrán P, Pastor-Ferrando JP, García-García F, Pascual-Agulló A (2010) An analytic network process approach for siting a municipal solid waste plant in the metropolitan area of Valencia (Spain). J Environ Manage 91:1071–1086

Başar A, Çatay B, Ünlüyurt T (2012) A taxonomy for emergency service station location problem. Optim Lett 6:1147–1160

Bellman RE, Zadeh LA (1970) Decision making in a fuzzy environment. Manage Sci 17:141–164

Benayoun R, Roy B, Sussman B (1966) ELECTRE: Une méthode pour guider le choix en presence de points de vue multiples. SEMA Note 49

Berman O, Krass D (2002) The generalized maximal covering location problem. Comput Oper Res 29:563–581

Bernasco W, Block R, Ruiter S (2013) Go where the money is: modeling street robbers' location choices. J Econ Geogr 13:119–143
Brans JP, Vincke Ph (1985) A preference ranking organization method. Manage Sci 31:647–656
Burkey ML, Bhadury J, Eiselt HA (2012) A location-based comparison of health care services in four U.S. states with efficiency and equity. Socio-Econ Plan Sci 46:157–163
Carroll L (1880–5) A tangled tale. Knot two: eligible apartments. https://ebooks.adelaide.edu.au/c/carroll/lewis/tangled/knot2.html. Accessed 29 March 2015
Ceriani L, Verme P (2012) The origins of the Gini index: extracts from Variabilità e Mutuabilità (1912) by Corrado Gini. J Income Inequal 10(3):421–443
Charnes A, Cooper WW (1961) Management models and industrial applications of linear programming. Wiley, New York
Charnes A, Cooper WW, Ferguson RO (1955) Optimal estimation of executive compensation by linear programming. Manage Sci 1:138–151
Church RL, ReVelle CS (1974) The maximal covering location problem. Pap Reg Sci Assoc 32:101–118
Church R, Roberts K (1983) Generalized coverage models and public facility location. Pap Reg Sci Assoc 53:117–35
Cohon J (1978) Multiobjective programming and language. Academic Press, New York. Reissued in 2004 by Dover Publications, Mineola, NY
Daskin MS (2008) What you should know about location modeling. Nav Res Logist 55(4):283–294
Eiselt HA (1992) Location modeling in practice. Am J Math Manag Sci 12(1):3–18
Eiselt HA, Laporte G (1995) Objectives in location problems. In: Drezner Z (ed) Facility location: a survey of applications and methods. Springer-Verlag, New York, pp 151–180
Eiselt HA, Marianov V (2008) Employee positioning and workload allocation. Comput Oper Res 35(2): 513–524
Eiselt HA, Marianov V (2014) Multicriteria decision making under uncertainty: a visual approach. Int Trans Oper Res 21:525–540
Eiselt HA, Sandblom C-L (2004) Decision analysis, location models, and scheduling problems. Springer-Verlag, Berlin
Gini C (1912) Variabilità e mutuabilità. Contributo allo studio delle distribuzioni e delle relazioni statistiche. C. Cuppini, Bologna
Hodgson MJ (1981) The location of public facilities intermediate to the journey to work. Eur J Oper Res 6:199–204
Hokkanen J, Salminen P (1997) Locating a waste treatment facility by multicriteria analysis. J Multi-Criteria Decis Anal 6:175–184
Hwang CL, Yoon K (1981) Multiple attribute decision making: methods and applications. Springer-Verlag, New York
Ignizio JP (1982) Linear programming in single- & multiple-objective systems. Prentice-Hall, Englewood Cliffs
Karagiannidis A, Moussiopoulos N (1997) Application of ELECTRE III for the integrated management of municipal solid wastes in the Greater Athens Area. Eur J Oper Res 97:439–449
Larichev OI, Olson DL (2001) Multiple criteria analysis in strategic siting problems. Kluwer Academic Publishers, Boston
Lucas MT, Chhajed D (2004) Applications of location analysis in agriculture: a survey. J Oper Res Soc 55(6):561–578
Nagy G, Salhi S (2007) Location-routing: issues, models, and methods. Eur J Oper Res 177(2):649–672
Niemi RG, Weisberg HF (eds) (1976) Controversies in American voting behavior. WH Freeman and Co., San Francisco
Olson DL (1996) Decision aids for selection problems. Springer-Verlag, New York
ReVelle CS, Eiselt HA (2005) Location analysis: a synthesis and survey. Eur J Oper Res 165(1): 1–19
ReVelle C, Marks D, Liebman JC (1970) An analysis of private and public sector location models. Manage Sci 16(11):692–707

Richards E (2000) Heuristic algorithms. In: Eiselt HA, Sandblom C-L (eds) Integer programming and network models. Springer-Verlag, Berlin, pp 229–258

Rosing KE, ReVelle CS (2000) Defendens imperium romanum: a classical problem in military strategy. Am Math Mon 107(7):585–594

Roy B (1971) Problems and methods with multiple objective functions. Math Progr 1:280–283

Saaty TL (1980) The analytic hierarchy process. Mc Graw-Hill, New York

Swersey AJ, Goldring L, Geyer ED (1993) Improving fire department productivity: merging fire and emergency medical units in new haven. Interfaces 23(1):109–129

Tavakoli A, Lightner C (2004) Implementing a mathematical model for locating EMS vehicles in Fayetteville, NC. Comput Oper Res 31:1549–1563

Tavares G, Zsigraiová Z, Semiao V (2011) Multi-criteria GIS-based siting of an incineration plant for municipal solid waste. Waste Manage (Oxford) 31:1960–1972

The Trucker's Report (2015) The real cost of trucking in the United States. http://www.thetruckersreport.com/infographics/cost-of-trucking/. Accessed 26 March 2015

Tryby ME, Boccelli DL, Uber JB Rossman LA (2002) Facility location model for booster disinfection of water supply networks. J Water Resour Plann Manage 128(5):322–333

Weber A (1909) Über den standort der industrien. Mohr Verlag, Tübingen (Theory of the location of industries). The University of Chicago Press, Chicago

Weiszfeld E (1937) Sur le point pour lequel la somme des distances de *n* points donnés est minimum. Tohoku Math J 43:355–386

Wilamowsky Y, Epstein S, Dickman B (1995) How the oldest recorded multiple facility location problem was solved. Locat Sci 3(1):55–60

Zeleny M (1974) Linear multiobjective programming. Lecture notes in economics and mathematical systems, vol 95. Springer-Verlag, Berlin

Part I
Business

Chapter 2
Location Analysis in Banking: A New Methodology and Application For a Turkish Bank

Ayfer Başar, Özgür Kabak, Y. İlker Topçu and Burçin Bozkaya

2.1 Introduction

In recent years, technology has improved and distribution channels such as credit cards, telephone-internet banking, Automated Teller Machines (*ATM*s) etc. have become alternative opportunities for reaching services of banks. However, banks generally gain new customers and develop customers' loyalty at their branches. Since branches are the indispensable contact points between the banks and their customers, no bank can easily avoid opening new branches or reorganizing the locations of current ones. According to the current statistics of The Banks Association of Turkey, the number of total bank branches has increased by 5.35 % from 10,450 to 11,009 in the last year. (The Banks Association of Turkey 2014). According to the statistics of Retail Banker International, JPMorgan & Chase opened 89 new branches in June 2013, which increased number of its branches from 5,608 to 5,697. In the same period, *BB&T* increased the number of its branches from 1,775 to 1,851 (Retail Banker International 2013). This shows that, due to the effects of increase in total population, population per bank branch and individual earnings, banks try to increase the number of their branches by locating them in the right places. Therefore, the branch location problem is a fundamental topic for banks in reaching their strategic goals.

B. Bozkaya (✉)
Sabanci School of Management, Sabanci University, Tuzla, 34956 İstanbul, Turkey
e-mail: bbozkaya@sabanciuniv.edu

A. Başar · Ö. Kabak · Y. İlker Topçu
Industrial Engineering Department, Istanbul Technical University,
Maçka Campus, 34357 İstanbul, Turkey
e-mail: ayferbasar@gmail.com

Ö. Kabak
e-mail: kabak@itu.edu.tr

Y. İlker Topçu
e-mail: topcuil@itu.edu.tr

H. A. Eiselt, V. Marianov (eds.), *Applications of Location Analysis,* International Series in Operations Research & Management Science 232, DOI 10.1007/978-3-319-20282-2_2

Although location problems for different types of places such as fire stations, schools, post offices, etc. are frequently studied in the literature, the problem of selecting the best places for new bank branches is much less often visited. Also, it is seen that the bank branch location problem is studied mostly with a multiple criteria decision making (*MCDM*) approach. The mathematical programming models in the literature, on the other hand, concentrate on modeling and solving a pre-defined problem. They do not provide any information on how to specify the decision criteria nor do they consider the multiple criteria nature of the problem.

Zhao et al. (2004) state that the criteria that are taken into consideration while opening a new branch may change according to the banks' strategies. Banks can focus on the places where they or their competitors have no branches, often areas involving industrial and commercial activities, organized industrial zones, shopping centers, collective housing areas, touristic regions, universities etc. in order to open new branches. Economic development level, population and demographic characteristics, distribution network, latent customers and their proximity to the potential markets, physical location, credit and deposit potential of the candidate regions may be important indicators for the branch location problem depending on the bank's strategies. Meanwhile, banks prefer to open various types of branches (individual, commercial, corporate, private, entrepreneur, etc.) due to the different characteristics of their target customers to minimize their costs and maximize their business process efficiency. Thus, the importance and the effect of the characteristics of potential locations can differ depending on the branch type. For example, in regions where commercial and industrial activities are high and big investments are made, commercial and corporate branches are generally located, while places where population and/or level of collective housing is high are selected for individual banking. Also, branches providing private banking services are mostly located where average household income is high.

This chapter presents an integrated decision support methodology for bank branch location problems in order to find relevant criteria and their importance weights to be used in a new mathematical programming modeling framework. In our methodology, we first select a set of criteria with the help of a detailed literature review and expert judgments. The criteria affecting the selection of optimum location for bank branches may differ due to the banks' strategies, customer profile and characteristics of the region. According to a survey of the literature, transaction volume of potential sites is one of the most important performance criteria of bank branches (Manandhar and Tang 2002; Cook et al. 2004; Camanho and Dyson 2005; Portela and Thanassoulis 2007). Therefore in this study we intend to discover the criteria affecting transaction volume based on a detailed literature survey and experts' opinions. In addition to the transaction volume, penalty of opening branches close to each other, cost of opening a new and closing an existing one are also considered as the other main criteria depending on the experts' judgments. Secondly, we identify importance of these criteria for different types of bank branches based again on expert opinions using pairwise comparisons. Furthermore, we develop a novel mathematical programming model, which helps banks to plan the location of their branches for different types of branches. Since the model is **NP**-hard and an optimal

solution cannot be found for large problems, we propose a tabu search approach and compare our results to those of *CPLEX* 12.2. We find that tabu search gives better and faster solutions than *CPLEX* (based on limited run times). Finally, we apply the proposed methodology in locating branches of a Turkish bank in Istanbul.

The remainder of this chapter is organized as follows: in Sect. 2.2, we present the studies in the literature about general location planning and bank branch location; in Sect. 2.3, we provide the details of our proposed methodology; in Sect. 2.4, we describe its application in Turkey. Finally, we provide conclusions and further research directions in Sect. 2.5.

2.2 Literature Review

Problems about finding the most appropriate place are important especially for strategic planning in both public and private institutions. In the literature, there are lots of studies involving location problems of different types such as for fire brigades, emergency medical services, schools, post offices, police patrols, military services, etc. Başar et al. (2012) classified the site selection problems in different ways: Being deterministic or stochastic according to the model structure; being static or dynamic, etc. Arabani and Farahani (2012) presented a review of studies in different location problems; essentially, coverage based and *p*-median models are the most common and well-known location models in the literature.

2.2.1 Coverage Based and p-median Models

The very basic deterministic model is the set covering problem (*SCP*) described by Toregas et al. (1971). The objective of *SCP* is to find the minimum number of potential sites covering all demand points. *SCP* tries to cover all demand points by at least one site without taking into account the population of demand points. Hwang (2002) described the stochastic set covering model and Baron et al. (2009) proposed *SCP* with stochastic demand. Murray et al. (2010) proposed a location set covering model, assuming that each demand area can be covered not only by one facility but also by two or more so that each facility covers a percentage of demand. Another common coverage based model is maximal coverage location problem (*MCLP*) proposed by Church and ReVelle (1974). The aim of *MCLP* is to maximize the population or the number of the demand points covered by the potential sites. Daskin (1983) proposed the maximal expected covering location problem with an equal busy probability assigned to all potential sites. Berman and Krass (2002) proposed a generalized *MCLP* by using the distance to the nearest potential site as a non-increasing function. Both *SCP* and *MCLP* depend on single coverage of demand points. This means that, if a server is busy due to serving a demand

point, other demand points covered by this server will no longer be covered. Therefore, backup coverage models have been proposed to avoid this situation. Daskin and Stern (1981) proposed the modified maximal covering location problem (*MMCLP*) with a second objective of maximizing the demand points covered multiple times. Curtin et al. (2010) and Erdemir et al. (2010) have also studied backup coverage problems. Since backup coverage problems are based on multiple coverage of demand points in a single travel time or distance restriction, models using double coverage with different time or distance standards were required in the literature. Therefore, Gendreau et al. (2000) introduced the double standard problem (*DSP*) maximizing the demand covered multiple times using two different travel time restrictions. The objective of *DSP* is to maximize the demand covered at least twice in the shorter travel time limit. The constraints include a set covering requirement of all demand points within the longer travel time and a given proportion of the population to be covered within the shorter travel time limit. Başar et al. (2011) extended the double coverage problem.

Another common type of model used for location problems is *p*-median studied by Kuehn and Hamburger (1960), Hakimi (1964) and Manne (1964). In *p*-median problems, the average total weighted access time is minimized while *p* service facilities are located. ReVelle and Swain (1970) solved the *p*-median model by using linear programming and branch-bound algorithm. Since *p*-median problems are **NP**-Hard, they cannot be solved in polynomial time. Therefore, many heuristics and exact methods have been proposed to solve the problem such as variable neighborhood search (Hansen et al. 2009) and simulated annealing (Brimberg and Drezner 2013), etc.

The *p*-median model and its extended versions in the literature all have an assumption that customers make dedicated visits to the bank to and from their home/work location. While this assumption may not hold for many types of location problems, it holds to a reasonable extent in the case of locating bank branches. Based on experience and observations, we find that many customers of certain type or age (e.g. housewives or home-office working people, retirees, part-time workers) who live in residential as well as commercial zones make special trips to a bank branch. Such a trip is usually unrelated to any other tasks that may precede or follow the bank visit, as the main purpose of the bank visit is to perform a task (other than money withdrawal) that cannot be completed online or via other channels (phone, *ATM*, mobile, etc.). There is, of course, still a possibility that such customers may combine the visit with other stops nearby the branch location.

2.2.2 Bank Branch Location Literature

The literature on finding the best place for bank branches can be classified into three groups: Statistical models, *MCDM* models and mathematical programming models.

2.2.2.1 Statistical Models

In most studies, different statistical techniques and criteria are discussed for the decision-making problem of finding the best places for bank branches. Clawson (1974) proposed stepwise linear regression to solve the bank branch location problem, setting realistic performance standards for different potential sites, and specifying remedial actions for poorly performing branches from a sample of 26 branches. Boufounou (1995) used regression analysis and conducted some statistical tests to determine the importance and significance level of related criteria for a Greek bank to evaluate existing branches' performance, assess performance of potential sites and place branches in more efficient locations. Ravallion and Wodon (2000) also used regression analysis to explain the relationship between demographic, economic, employment indicators and the location of Bangladesh's Grameen Bank's branches. They found that the bank branch location problem is mainly affected by the economic indicators. On the other hand, industrial characteristics and banking products served to these industries by the branches are also important factors. They discussed that the bank branch location decision was influenced by the potential gain from switching to more profitable non-farm activities in rural areas. All these factors and their relative influences on the location decision process, resulting from the regression modeling approach, could allow an analyst to pin down the most relevant set of criteria, which can subsequently be used to locate new facilities in future analyses.

2.2.2.2 MCDM Models

Since they consider multiple criteria, *MCDM* models are very common for bank branch location problems. As stated by Carlsson and Fuller (1996), *MCDM* is a fundamental field of study dealing with decision making including multiple criteria, attributes and objectives. The main objective of the *MCDM* methods is to help decision makers solve complicated decision making problems systematically. There are four major groups of methods in *MCDM*: (*i*) the outranking, (*ii*) the value and utility theory based, (*iii*) the multiple objective programming, (*iv*) group decision and negotiation theory based methods. As a consequence of the developments in fuzzy set theory, fuzzy *MDCM* are also commonly discussed in the literature. Chen and Hwang (1993) made distinctions between fuzzy ranking methods and multiple attribute decision making methods all included in the groups (*i*)-(*iv*).

Fuzzy set theory is used very frequently for the bank branch location problem, because of the uncertainties in the comparisons of the criteria and the alternatives. For instance, Min (1989) proposed fuzzy goal programming method to find the most appropriate places for commercial bank branches in Ohio. In order to select one among six cities in South Eastern Anatolia for opening a new branch, Cinar (2009) used fuzzy analytical hierarchy process (*AHP*) to find the importance of the related criteria and *TOPSIS* to rank the cities. Rahgan and Mirzazadeh (2012) used fuzzy

AHP to specify the importance of criteria and evidential reasoning to order the alternatives. As a result of the improvements in information technologies, Morrison and O'Brien (2001) used geographical information systems (*GIS*) and a spatial interaction model (Huff 1963) in four stages: The probability of customers visiting a branch is estimated, the expected distribution of customers is determined, the expected number of transactions at a given branch is computed and the impact of removing one or more branches from the network is analyzed.

2.2.2.3 Mathematical Programming Models

Although the mathematical programming literature related to facility location problems is rich, studies specific to bank branch location problems are scarce. Min and Melachrinoudis (2001) proposed a three-level hierarchical location-allocation model for bank branches by considering risk and uncertainty, where a chance constrained goal programming model was developed using forecasting techniques. The methodology was applied in a state in the USA. Miliotis et al. (2002) introduced a two-step methodology in which firstly the minimum number of branches to meet the coverage requirement was found and then locations of branches to maximize coverage was determined. They used *GIS* and applied their methodology to a bank in Greece. Wang et al. (2002) considered the problem of locating *ATMs,* internet mirror sites, or other immobile (permanent location) service systems of limited service capacity. They model these service facilities as simple *M*/*M*/1 queuing systems and solve the model using three different heuristic approaches. Wang et al. (2003) developed a mathematical model for the branch location in Amherst, New York. They improved greedy interchange, tabu search, and Lagrangian Relaxation to apply this **NP**-Hard model for the branch location in New York by using 270 generated problems. Unlike the *p*-median model, the model consists of a budget constraint related to opening and closing branches. Zhang and Rushton (2008) proposed a model maximizing total benefit with the budget constraints related to opening branches and the waiting time of the customers. They used a genetic algorithm to solve their proposed problem. Alexandris and Giannikos (2010) proposed a new model for the *MCLP* and illustrated the applicability of the proposed model by means of a case study concerning the location of bank branches. The aim of the model is to maximize the total population covered by the selected branches. Xia (2010) formulated a mathematical programming model considering operations and rental costs, demand and distance between branches by proposing a hybrid nested partitioning algorithm to solve the large-scale problem.

As a result of the literature review, there is no integrated methodology that combines problem structuring phase (i.e., as in *MCDM* models) and mathematical programming models for the bank branch location problem. Therefore this study proposes a novel integrated methodology for not only identifying the criteria and their importance but also finding the exact locations of bank branches through a mathematical programming model.

2.3 Proposed Methodology

The basic aim of this study is to solve the bank branch location problem using an integrated methodology that combines the problem structuring phase and mathematical programming model. To this aim, firstly related criteria are listed based on a literature survey and expert opinions. Secondly, the importance of each criterion is identified depending on the pairwise comparison method conducted with experts working in the banking sector. Furthermore, a new mathematical model is developed to determine the specific locations of branches. Since the model is **NP**-Hard, a tabu search approach is developed to find (near-) optimal solutions for the big problems and the results are compared to those of the *CPLEX* 12.2. The methodology is applied to the real life practice of a Turkish national bank branches' location problem in Istanbul.

2.3.1 Identifying the Criteria

Banks prefer providing service with different types of branches having different customers such as individual, commercial, private, corporate, entrepreneur etc. to decrease their costs and increase the efficiency of business processes. Therefore, they need to open different types of branches. According to expert opinion, finding the best locations especially for individual, entrepreneur, corporate and commercial bank branches is an important topic for banks.

Individual branches can serve busy and standard transactions to every kind of customer. Entrepreneur branches are generally located in the industrial zones, trading estates and agricultural spheres. Corporate branches make high contribution to the gross national product with their predominantly agricultural customers and are located in the regions where there is no need to open a commercial branch. Commercial branches' customers generally deal with industrial, business and agricultural activities and also make the highest contribution to the gross national product. Entrepreneur, corporate and commercial branches generally do not offer busy and standard service to their customers. On the other hand, customer segmentation rules of the banks may affect customer profile and the location regions' specification for these branches.

Firstly, we analyze the criteria used in the literature to identify the criteria for bank branch location problem. We see that there are lots of criteria taken into consideration while opening or closing a bank branch. Clawson (1974) found that population of middle age, average per capita domestic income and home ownership rate are the most important factors. Meidan (1983) pointed out that commercial potential, population, state of employment in the candidate regions and the location of competitor bank branches were fundamental aspects to be chosen. Boufounou (1995) indicated that total population in terms of gender and age, average household size, population growth rate, domestic per capita income, number of firms

in terms of sector and location of competitor banks were important criteria for bank branch performance in Greece. Min (1989) proposed demographic characteristics, ease of access and business operations of candidate regions as important criteria for bank branch location problems in Ohio. Pastor (1994), Kaufman and Mote (1994) stated that economic and demographic variables had an important effect on location of bank branches in Chicago. Ravallion and Wodon (2000) showed that demographic characteristics, employment and economic indicators of the candidate locations were the most important factors for bank branch location problems in Bangladesh. Abbasi (2003) developed a decision support system for bank branch location problems in Jordan and used population, income level, cultural characteristics, number of firms, capital, growth potential and number of competitors as variables in their proposed model. Zhao et al. (2004) emphasized that financial indicators, demographic characteristics, customer segmentation, position of competitor banks and means of access of the potential points had to be considered for bank branch location. They indicated the main and sub criteria for location of bank branches in Sydney as follows:

- Service variables (number of people older than 18, population growth rate, average annual household income, employee rate, number of small firms, number of competitor branches),
- Location variables (number of small firms, competitor banks, and employed population within 200 m distance to the potential location; number of bank branches within the context of application; shopping centers within 500 m distance to the potential location; transportation index; closeness to the public transportation).

Cinar (2009) identified 5 main criteria, including 21 sub criteria, affecting a bank's mission and strategy as follows:

- Demographic (total population, annual population growth and urbanization rate),
- Socioeconomic (average household size, literacy rate, rate of higher education graduate, average per capita domestic income, employee and employer rate),
- Banking indicators (number of banks and branches, deposit and credits per bank branch-person),
- Employment rate in accordance with the sectors (agriculture, building, manufacturing, and service sectors),
- Trade potential (number of firms and industrial zones).

Rahgan and Mirzazadeh (2012) introduced a hierarchical model for bank branch location problems using cost, demography, banking, geographical conditions and accessibility as main criteria. Criteria used in the literature for bank branch location problems are given in Table 2.1.

As one of the aims of our study is to find the related criteria for the banking sector, expert opinion is also considered. By the help of one-on-one interviews, necessary information is gathered through an unstructured questionnaire. Experts are asked to indicate their opinion on both how banks decide where to locate new branches in current conditions and the effective strategies they employ. Interviews

Table 2.1 Criteria used in the literature

Criteria	Paper
Population	Clawson 1974; Olsen and Lord 1979; Doyle et al. 1981; Meidan 1983; Boufounou 1995; Abbasi 2003; Zhao et al. 2004; Cinar 2009
Average per capita domestic income	Clawson 1974; Boufounou 1995; Cinar 2009
Home ownership rate	Clawson 1974; Olsen and Lord 1979
Commercial potential	Meidan 1983; Cinar 2009
Location of competitor bank branches	Doyle et al. 1981; Meidan 1983; Boufounou 1995, Abbasi 2003; Zhao et al. 2004; Cinar 2009
State of employment	Olsen and Lord 1979; Doyle et al. 1981; Meidan 1983; Zhao et al. 2004; Cinar 2009
Demographic characteristics	Olsen and Lord 1979; Doyle et al. 1981; Min 1989; Kaufman and Mote 1994; Ravallion and Wodon 2000; Rahgan and Mirzazadeh 2012
Ease of access	Doyle et al. 1981; Min 1989; Zhao et al. 2004; Rahgan and Mirzazadeh 2012
Business operations	Min 1989
Average household size	Boufounou 1995; Cinar 2009
Population growth rate	Boufounou 1995; Abbasi 2003; Zhao et al. 2004; Cinar 2009
Number of firms	Doyle et al. 1981; Boufounou 1995; Abbasi 2003; Zhao et al. 2004; Cinar 2009
Income level	Abbasi 2003; Zhao et al. 2004
Total deposit	Abbasi 2003
Cultural characteristics	Abbasi 2003
Literacy rate	Cinar 2009
Deposit-credits per bank branch-person	Cinar 2009

are conducted with five experts who work in the Turkish banking and information technology sector as senior executives.

All the experts indicate that average transaction volume is the most effective factor for Turkish bank branch profit. This result is also supported by the literature (Jablonsky et al. 2004; Camanho and Dyson 2005; Portela and Thanassoulis 2007). There are different parameters affecting banks' daily transaction volume such as payment of wages, education fees, bills etc., banking transactions of which have to be done at a specified date. Similarly, annual transaction volume of the sites can be misleading due to the long periods involved. Thus, based on the experts' judgment, average monthly transaction volume of potential sites is identified as one of the main criteria for deciding the location of bank branches in Turkey. Transaction volume is defined so as to include all types of banking operations such as credits, deposits, bonds, cheques, money transfers etc.

Table 2.2 Criteria used for the estimation of average monthly transaction volume

Criteria	Proposed sub criteria	Related criteria in the literature
Number of potential customers	Daytime population	Population (Clawson 1974; Olsen and Lord 1979; Doyle et al. 1981; Meidan 1983; Boufounou 1995; Abbasi 2003; Zhao et al. 2004; Cinar and Ahiska 2010)
Socioeconomic situation	Education level, number of houses and summer houses	Literacy rate and education level (Cinar 2009)
Social potential	Number of education and entertainment places, parks, hospitals	Social potential (Abbasi 2003)
Commercial potential	Number of private institutions, shopping centers, car parks, financial institutions, restaurants and car services	Commercial potential (Meidan 1983; Cinar 2009)
Growth potential	Population growth rate	Population growth rate (Boufounou 1995; Abbasi 2003; Zhao et al. 2004; Cinar 2009)
Ease of access	Ease of travel, accessibility and proximity to public transport	Ease of access (Doyle et al. 1981; Min 1989; Zhao et al. 2004; Rahgan and Mirzazadeh 2012)
Competition	Number of competitors' bank branches	Number of banks and branches, position of competitor banks (Meidan 1983; Boufounou 1995; Abbasi 2003; Zhao et al. 2004; Cinar 2009)
Financial situation	Average household income	Home ownership rate (Clawson 1974); income level (Abbasi 2003; Zhao et al. 2004); employee and employer rate (Meidan 1983; Zhao et al. 2004; Cinar 2009)

Since forecasting the average monthly transaction volume accurately is almost impossible, different criteria and sub criteria are identified depending on both the literature survey and expert opinion. These are listed in Table 2.2 along with the related sub-criteria. According to the literature review and experts' judgment, "population growth rate," "number of competitor bank branches" and "average household income" are determined as measures of growth potential, competition and financial situation, respectively.

Since the selected criteria cover them, a number of important criteria mentioned in the literature survey are not directly used in this study. For instance, "employee-employer rate" mentioned in Zhao et al. (2004) and "average household income" represented by Cinar (2009). Credit/deposit per bank branch/person are the banks'

financial data, therefore it is not possible to directly use these data because of security and confidentiality rules of banks; thus, "average household income" is used instead.

Since data related to daytime population are not available, number of potential customers for all candidate places is estimated by different criteria affecting daytime crowdedness of the regions which are total population, number of work and education places, financial institutions, and hospitals. Similarly, proximity to public transport is not readily available; hence total distance to the neighboring districts, calculated on the transportation network, is used as a cost measure for ease of access. Also, after analyzing correlation between criteria values among 763 districts in Istanbul; number of houses, parks and restaurants are eliminated since there exists high correlation between these criteria and total population, number of car parks and financial institutions, respectively.

Moreover, according to the experts' judgments, the distance between all the potential points is another main criterion, especially to avoid opening multiple branches close to each other. Thus, we determine that the opening of new branches in close proximity should be penalized. Moreover, experts indicate that costs of opening a new branch as well as closing an existing one are other important criteria for bank branch location problems. Finally, the selected criteria of the proposed model are presented in Fig. 2.1.

2.3.2 Specifying the Importance of the Criteria

As analyzed in Sect. 2.3.1, bank branch location problems can be modeled as a hierarchy containing the decision goal, which is selecting the best place, and related criteria affected by different factors. Since it is difficult to use an objective weighting method such as entropy method, regression analysis and the *CRITIC* (*C*riteria *I*mportance *T*hrough *I*ntercriteria *C*orrelation) method, because of data confidentiality in the banking sector, a subjective weighting technique is required to find the importance of the criteria. Miller (1956) argues that the brain of an average human can simultaneously process, differentiate and deal with at most seven factors, with this limit decreasing to five for some, while increasing to nine for others, and people are generally more consistent when they do pairwise comparisons than when they just assign relative importance to the criteria. Thus when different approaches such as rating, point allocation, ratio, raking, pairwise comparison and tradeoff are analyzed, pairwise comparison is found to be the most efficient approach since it focuses on only two alternatives at each time (Malczewski 1999). Also since decision makers focus on finding the relative importance of two criteria without being affected by external factors and generally have deep knowledge with which to compare the criteria, pairwise comparison generally gives more accurate results compared to the other weighting methods (Badri 2001). Pairwise comparison is also the main approach of *AHP* which is one of the most common techniques in decision making literature as originally proposed by Saaty (1980). The *AHP* has

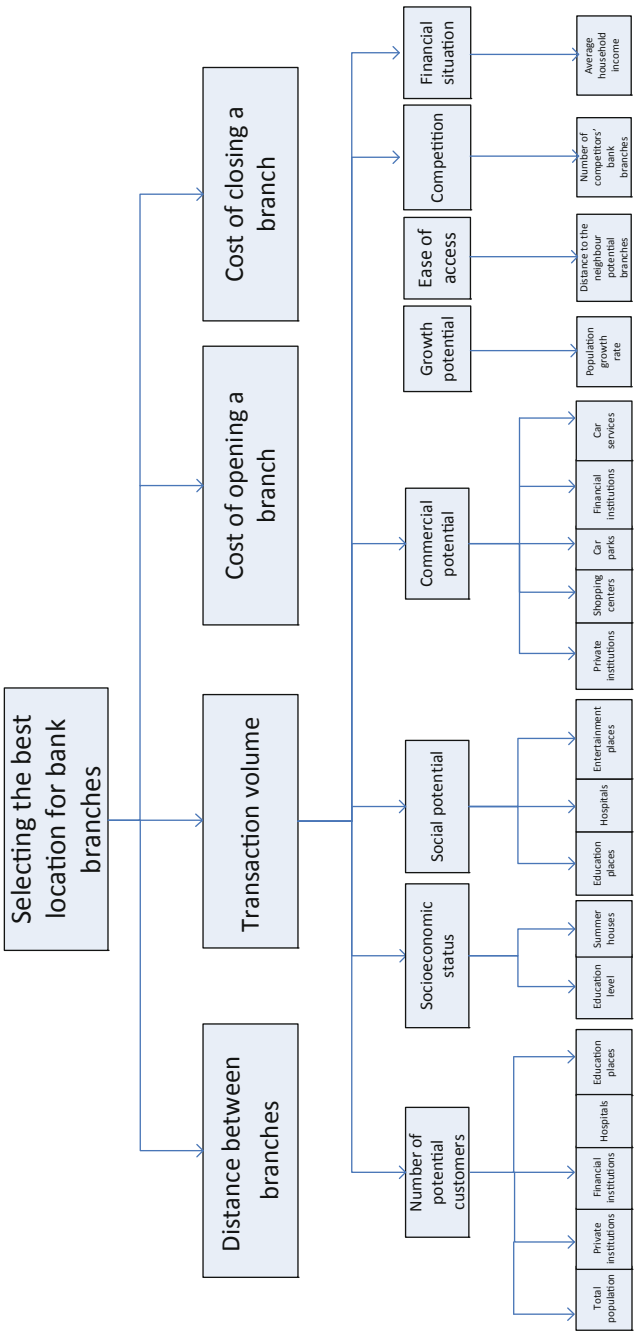


Fig. 2.1 Selected criteria of the proposed model

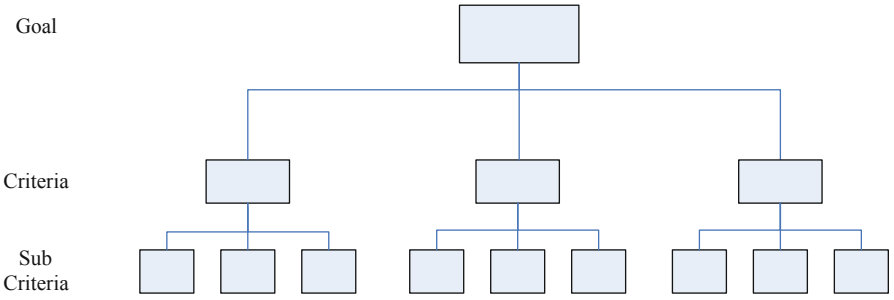


Fig. 2.2 A typical decision hierarchy

Table 2.3 Pairwise comparison scale used in the study

Intensity of importance	Definition
1	Equal importance
3	Moderate importance of one over another
5	Strong or essential importance
7	Very strong or demonstrated importance
9	Extreme importance
2, 4, 6, 8	Intermediate values

been applied in a variety of studies including bank branch location problems (e.g. Chen et al. 2001; Sato 2004; Ishizaka and Lusti 2004; Ahsan and Bartlema 2004; Tzeng et al. 2002; Aras et al. 2004; Wu et al. 2007; Fernandez and Ruiz 2009). Thus, pairwise comparison is determined as an effective method for establishing the importance of main and sub criteria by making a series of expert judgments. The method is applied by following these steps:

Step 1 The decision hierarchy, including a goal, criteria and sub criteria (if exist), is prepared to make judgments on the elements of the hierarchy in pairs with respect to the parent element. A typical decision hierarchy is illustrated in Fig. 2.2.

Step 2 A pairwise comparison matrix **A** is built to compute the importance of different criteria and sub criteria. Each entry (a_{ij}) of the matrix **A** shows the importance of the criterion i compared to the criterion j.

The relative weights of the criteria are determined according to the scale of 1–9 corresponding to the linguistic comparisons, which are described in Table 2.3 (Saaty 1990).

Step 3 After building matrix **A**, relative importance weights of the criteria $\mathbf{w} = (w_1, w_2, \ldots, w_n)$ are found by solving $\mathbf{Aw} = n\boldsymbol{w}$, which is called the principal right eigenvector of **A** (Saaty 1990).

2.3.3 The Proposed Mathematical Programming Model

As stated in the literature review, there are only a few mathematical models proposed for bank branch location problem. Experts indicate that a maximization model calculated by the help of criteria in parallel with banks' mission and strategy is more useful than a model using to minimize distance or time between demand points and branches. The number of each type of branch to open is definite due to the budget constraint, which is generally planned annually. Moreover, opening a new branch and closing an existing one both have associated costs and some of existing branches cannot be closed due to company strategy, according to experts' opinion. Thus, a new mathematical programming model to locate bank branches in Turkey using criteria defined in Sect. 2.3.1. is defined as follows. (Table 2.4)

The average monthly transaction volume h_{ij} can be forecasted according to the criteria identified in Sect. 2.3.1 (See Fig. 2.1 and Table 2.2), and the weights specified in Sect. 2.3.2 (Table 2.5) as follows:

$$h_{ij} = w_{1j}NPC_i + w_{2j}SS_i + w_{3j}SP_i + w_{4j}CMP_i + w_{5j} GP_i + w_{6j}EA_i w_{7j}C_i + w_{8j}FS_i \tag{2.1}$$

where w_{yj} ($y = 1,\ldots, 8, j \in J$) is the weight of criterion y for location type j as defined in Table 2.5.

In Eq. 1, NPC_i, SS_i, SP_i, and CMP_i have sub-criteria according to the criteria hierarchy given in Fig. 2.1. Therefore these are calculated as follows (see Table 2.11 for the notation):

$$NPC_i = w_{9j}TP_i + w_{10j}PI_i + w_{11j}FI_i + w_{12j}H_i + w_{13j}EP \tag{2.2}$$

$$SS_i = w_{14j}EL_i + w_{15j}NS_i, \tag{2.3}$$

$$SP_i = w_{16j}EP_i + w_{17j}H_i + w_{18j}ENP_i, \tag{2.4}$$

$$CMP_i = w_{19j}PI_i + w_{20j}SC_i + w_{21j}CP_i + w_{22j}FI_i + w_{23j}CS_i \tag{2.5}$$

where w_{yj} ($y = 9,\ldots, 23, j \in J$) is the weight of criterion y for location type j as defined in Tables 2.6, 2.8 and 2.9.

$$Z\,(\max) \sum_{i\in I}\sum_{j\in J} c_{1j}h_{ij}x_{ij} - \sum_{i\in I}\sum_{m\in I, i\neq m}\sum_{j\in J} c_{2j}\lambda_{imj} - \sum_{i\notin K}\sum_{j\in J} c_{3j}x_{ij} - \sum_{j\in J} c_{4j}\sum_{i\in K}(1 - x_{ij})$$

$$\text{s.t.} \sum_{i\in I} x_{ij} = P_j \quad \forall j \in J \tag{2.6}$$

$$x_{lj} = 1 \quad \forall\, \ell \in L,\ j \in J \tag{2.7}$$

$$\lambda_{imj} \geq k_{im}(x_{ij} + x_{mj} - 1) \quad \forall i,\ m \in I,\ i \neq m,\ j \in J \tag{2.8}$$

$$x_{ij} \in \{0,\ 1\} \quad \forall i \in I,\ j \in J \tag{2.9}$$

$$\lambda_{imj} \geq 0 \quad \forall i, m \in I, j \in J \tag{2.10}$$

Table 2.4 Notation used in the model

Notation	Definition
Sets	
I	Set of potential sites to open bank branches
J	Set of types of branches (individual, entrepreneur, corporate and commercial)
K	Set of sites of current branches
L	Set of sites of the branches that cannot be closed by strategy
Indices	
i	Potential site to open bank branch
j	Type of branch (individual, entrepreneur, corporate and commercial)
Parameters	
h_{ij}	Average monthly transaction volume of the potential site i, when branch type j is located at site i
NPC_i	Number of potential customers at the potential site i
SS_i	Socio-economic situation of the potential site i
SP_i	Social potential of the potential site i
CMP_i	Commercial potential of the potential site i
GP_i	Growth potential of the potential site i
EA_i	Ease of access in the potential site i
C_i	Competition of the potential site i
FS_i	Financial situation of the potential site i
TP_i	Total population of the potential site i
PI_i	Private institutions of the potential site i
FI_i	Institutions of the potential site i
H_i	Hospitals of the potential site i
EL_i	Education level of the potential site i
NS_i	Number of summer houses of the potential site i
EP_i	Education places of the potential site i
ENP_i	Entertainment places of the potential site i
SC_i	Shopping centers of the potential site i
CP_i	Car parks of the potential site i
CS_i	Car services of the potential site i
d_{im}	Distance between the potential sites of i and m (in meters)
k_{im}	Parameter depending on distance between i and m (takes value between 0 and 1)
P_j	Maximum number of branches of different types
S	Threshold (in meters) to consider the penalty for opening branches near each other
c_{1j}	Benefit coefficient (as percentage) of transaction volume for opening branch type j
c_{2j}	Penalty coefficient (as percentage) for opening branches type j near each other
c_{3j}	Coefficient (as percentage) for opening a branch type j
c_{4j}	Penalty coefficient (as percentage) for closing a branch type j

Table 2.4 (continued)

Notation	Definition
Decision Variables	
λ_{imj}	Penalty of opening j type branches in potential nearby sites i and m
x_{ij}	Binary variable which is 1 if a type j branch is located at site i, and 0 otherwise

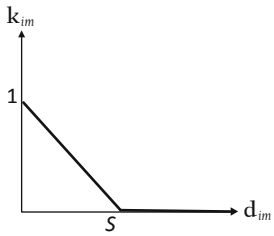


Fig. 2.3 The relationship between distance and parameter based on the threshold value

The objective function maximizes total net profit as the difference between total benefits of expected average monthly transaction volume minus the penalty of opening branches nearby to each other and cost of opening new and closing existing branches. Constraint (2.6) shows how many branches are to be opened for each type given the budget. Constraint (2.7) avoids closing branches that cannot be closed due to company strategy (in our case, branches opened in the last 3 years, according to the expert opinion). Constraint (2.8) indicates penalty of opening branches near each other according to the threshold values and distances. λ_{imj} is the penalty of opening type j branches in potential nearby sites i and m, taking a value between 0 and 1. It is designed to be 1 if the distance between branches i and m is zero, and 0 if the distance is equal to or more than a threshold value S, linearly decreasing in between. Obviously, if a branch is not opened in either site i or site m, λ_{imj} will be zero. To guarantee λ_{imj} gets values as designed, a parameter namely k_{im} is defined. k_{im} is formulated based on distance d_{im} between potential sites i and m as given in Fig. 2.3. λ_{imj} is equal to k_{im} if branches are opened in both sites i and m, and 0 otherwise.

Constraints (2.9) and (2.10) show that decision variables of bank branch location are binary and penalties of opening branches near each other are nonnegative, respectively. The model permits both opening new and closing existent branches at the same time. Also, while multiple branches of the same type cannot be opened in the same locations, different types of branches are allowed to be opened in the same candidate location.

In order to solve any bank branch location problem using the proposed model, the related data components have to be compatible with one another. Therefore, we convert the data on education level of people living around the potential site to numerical data using a four-point scale, one for the people who are unable to read and write, two for primary school, three for high school, four for university graduates. Furthermore, we normalize the values for all the criteria. To this aim, we first specify S based on the experts' judgment and the geographical structure of the potential sites. Thus, after collecting the distances d_{im}, using the transportation network and

the help of a *GIS*, between the best-served potential sites, we calculate the k_{im} values. Since d_{im} is based on the transportation network instead of the straight-line distance between i and m, d_{im} may differ from d_{mi} for all $i \neq m$. Also, for the districts having an existing branch, the place where the branch is located is considered as the reference point. Thus, the location of the branches that will continue to serve will remain unchanged. In order to provide concurrency, these data have to be normalized as benefit attributes, since the bigger k_{im} values are, the more penalty will be imposed to the potential site. Furthermore, the average monthly transaction volume of the potential site is forecasted in accordance with the hierarchy defined in Sect. 2.3.1. This means that, after normalizing all the data and using the importance weights found by pairwise comparisons, the value of the related criteria is specified beginning from the bottommost and moving up through the hierarchy. At this point, the direction of the effect must also be considered. Therefore, while the "number of competitors' bank branches" and "distance to the neighbor potential branches" have to be normalized as a cost attribute, all the other criteria are evaluated as benefit attributes. Thus, each term in the objective function can take a value between 0 and 1.

It can be proven that proposed mathematical model is **NP**-hard by considering a special case of the problem where $L = \emptyset$ and $\lambda_{imj} = 0$ such that there are no branches that cannot be closed due to strategy and the distance between the all potential locations is bigger than S. In this case, the problem reduces to *MCLM*, which is known to be **NP**-hard (Marianov and ReVelle 1995; Berman and Krass 2002). Therefore, optimal solution of the problem cannot be easily found for large problem sizes.

2.3.4 Proposed Tabu Search Algorithm

Banks generally plan to open branches at the city or metropolitan level and prefer big cities to locate new branches. Therefore, it is not possible to solve the location problem of bank branches in large cities having lots of candidate points. Accordingly, efficient heuristic and meta heuristic approaches are required. Youssef et al. (2001) and Arostegui et al. (2006) experimentally pointed out that tabu search gives better results than genetic algorithms or simulated annealing for location problems. Also, Başar et al. (2011) discussed that tabu search is a very efficient approach for a mathematical model similar to the proposed one in our context. Therefore, we propose a tabu search approach to solve the mathematical model in Sect. 2.3.3.

Tabu search is a metaheuristic originally developed by Glover (1977), which uses an initial solution, searches the neighbors of the incumbent solution at each iteration to reach better solutions and improve the best solution obtained. It keeps a tabu list to avoid the repetition of the same solutions, by forbidding a certain set of moves. In principle, a tabu move can still be accepted if it gives a better best known solution. Due to its efficiency and effectiveness, tabu search is commonly used in the

literature for different optimization problems (Osman and Kelly 1996; Pardalos and Resende 2002; Ribeiro and Hansen 2002).

2.3.4.1 Initialization

We use three initialization methods in our implementation of tabu search to observe their effects on the overall solution quality: (*i*) random method, (*ii*) linear programming, relaxation-based method and (*iii*) criteria-based method. Since both feasibility conditions (constraints 2.6 and 2.7) are satisfied by opening P_j branches for each type and not closing the branches in the set of L, all initialization methods are feasible.

In the random method, after satisfying the feasibility conditions, locations of remaining branches for each type are determined randomly. The initial solution thus found is expected to be of low quality since the method is not rule-based. In the linear programming relaxation method, x_{ij} decision variables, which cause the model to be **NP**-hard are relaxed and the resulting linear program is solved. The candidate point as many as P_j for each branch type with the biggest x_{ij} values are selected to locate the branches. This initial solution is expected to be of high quality since the x_{ij} values are considered. In the criteria-based method, after satisfying feasibility conditions, locations of remaining branches for each branch type are determined according to the best value of the main criterion with the highest importance. In Table 2.4, we find that transaction volume is the most important main criterion for all types of branches. Therefore, tabu search selects candidate points with the highest transaction volume to open the remaining branches.

2.3.4.2 Tabu search procedure

We propose a neighborhood search structure in which one branch is closed without violating the feasibility (which is not in the set L) and a new one is opened (without causing the same location to have multiple branches of the same type) simultaneously for each type. Since initial solutions used are feasible, the proposed tabu search also maintain the feasibility with respect to constraint (2.6) and (2.7). The tabu search algorithm searches all such possible moves for four types of branches and determines the candidate solution as the one providing the largest increase in the objective function value at each iteration. An alternative approach would be to open multiple branches and close the same number of branches without violating feasibility. However, it is clear that this will increase the computational time significantly because of searching high number of neighbors. Also, it is not ensured that such a move structure will lead to better results, so we choose not to implement it.

The tabu search algorithm keeps track of a tabu list to avoid cycling of solutions. It does this by forbidding the same moves for a number of iterations called the tabu list size. Since tabu list size affects solution quality meaningfully, it has to be determined properly according to the problem structure and size. Although, there are

studies that use static tabu list size (Malek et al. 1989; Glover 1989), there are also works including examples of dynamic tabu list size (Zhang et al. 2003; Grabowski and Wodecki 2004). The tabu search procedure identifies a move and then it checks to see if the move is in the tabu list or not. If the move is tabu and the aspiration criterion is satisfied by leading to a solution better than best obtained so far, the corresponding solution is accepted for the next iteration; otherwise more neighbors are reinvestigated.

Our definition of a tabu move is such that we keep the closed and opened branch in the tabu list for four types of branches separately. This means that the exact station pair to be opened and closed simultaneously is forbidden for a specific number of iterations. It is clear that using two separate tabu lists for closed and opened branches for each type will cause the algorithm to be stuck when the number of candidate regions is not high. So we choose not to implement this alternative, either.

Generally, tabu search may get trapped at a local optimum solution after a while when it can no longer find a better solution than the best-so-far solution. Therefore, we propose using a diversification strategy to overcome this problem. If tabu search cannot improve the best-so-far solution after k_2 consecutive iterations, it closes a branch (which is not in the set L) and opens another branch (without causing the same location to have multiple branches of the same type) randomly for four types without violating feasibility. Başar et al. (2011) experimented also with randomly closing and opening multiple stations for the location problem of emergency service stations, which, however, has led to lower performance compared to randomly closing and opening a single station.

Also, the current solution may not improve for k_1 consecutive iterations. In such a situation, another solution providing the least decrease in the current objective function value is selected instead of the one giving the highest increase. Lastly, a stopping criterion is determined to end the algorithm at the end of k_3 iterations.

2.3.4.3 Experimental Study

Forty problem instances are generated randomly in order to calibrate the parameters, 5 each with 50, 100, 200, 300, 400, 500, 750 and 1,000 candidate locations. The number of branches to open (P_j) is assumed to be 10 %, 4 %, 2 %, and 1 % of the total number of candidate locations for individual, entrepreneur, corporate and commercial branches, respectively, depending on the relation between the number of sub-districts and branches in the cities of Turkey. Also, the number of branches to open is rounded down in case it is found fractional (e.g. since the number of commercial branches for a town with 50 potential locations will be $50(0.01) = 0.5$, it will be used as 0, thus no commercial branch will be opened in this area. On the other hand, we assume that banks do not prefer to decrease the number of their current branches. Therefore, P_j is set equal to at least the number of current branches for each type. Moreover, the locations of current branches and the ones which cannot be closed due to company strategy (that were opened in the last 3 year) are generated randomly depending on the problem size. The proposed tabu search procedure is

coded in Microsoft *Visual* C++ and executed on 2.3 *GHz* Intel Pentium with 4 *GB* of *RAM*. All problem instances are solved using *OPL* Studio 6.3 with *ILOG CPLEX* 12.2 on the same processor.

We conducted an extensive experimental study to set the parameters of tabu search, thus we determined the best value of k_1 as 8 and k_2 as 12. Also, we observe that bigger tabu list sizes give better results for the instances with more candidate locations. Therefore, tabu list size is decided to be dynamic by using 5, 7, 8, 10, 13, 15, 16 and 19 for the problems with 50, 100, 200, 300, 400, 500, 750 and 1,000, respectively. On the other hand, 5,000 and 10,000 iterations are used as k_3 values to assess the performance of tabu search with respect to the high number of iterations.

The experimental study shows that tabu search gives very good results compared to the solutions found by *CPLEX*, independent from the initialization approach and it outperforms *CPLEX* in terms of both the objective function value and computation time for most problems.

2.4 Locating Bank Branches of a Turkish Bank in Istanbul

The proposed tabu search algorithm is applied to branch location problem of a Turkish National Bank (for confidentiality reasons, the name of the bank cannot be provided) in Istanbul based on the real data collected by the help of Turkish Statistical Institute and The Banks Association of Turkey. *GIS* is used to determine the candidate location set for the branches. There are 763 sub-districts in Istanbul, which are all potential sites for a new branch location. The location of the bank's current branches, their types and opening dates are obtained for use in the methodology. The coordinates of bank's current branches are considered in order not to change the location of the working branch if it is in the solution. Otherwise, coordinates of the central points of the districts are taken into account due to the assumption that these places are on the most visited roads. Distances of candidate branches to each other are calculated based on the transportation network using the *GIS*. Thus, the penalty of opening the same type of branches near each other and ease of access of all candidate locations are determined.

Currently, the bank has 152 individual, 18 entrepreneur, 12 corporate and 3 commercial branches in Istanbul. Also, 16 individual, 18 entrepreneur, 10 corporate and 1 commercial branches cannot be closed due to company strategy since they were opened within the last 3 years. According to the experts' opinion and structure of the sub-districts in Istanbul, S is specified as 1,000 m.

Through one-on-one interviews that involve a series of pairwise comparisons for four different types of branch, we sought the help of 10 experts working as senior executives in the Turkish banking sector in exploring the hierarchy and evaluating all the criteria and sub-criteria shown in Fig. 2.1. As a result, six pairwise comparison matrices are prepared for each branch type: One for four main criteria in the first level of the hierarchy, one for eight sub-criteria in the second level of the hierarchy to find their importance on transaction volume, and four for the lowest level of the

hierarchy to establish the importance of the sub-criteria for the number of potential customers, socio economic situation, social and commercial potential. After collecting the evaluations of each expert based on pairwise comparisons, the importance of the main and sub-criteria are identified using Eq. (2). As a result, the importance weights given in Tables 2.5–2.10 are calculated.

As seen in Table 2.5, transaction volume is the most important main criterion for all branch types. According to the results summarized in Table 2.6, the most important criterion affecting transaction volume is the number of potential customers

Table 2.5 Weights of main criteria

Criteria	Notation	Individual	Entrepreneur	Corporate	Commercial
Transaction volume	c_{1j}	*0.51*	*0.56*	*0.51*	*0.52*
Distance between branches	c_{2j}	*0.32*	*0.27*	*0.30*	*0.30*
Cost of opening a new branch	c_{3j}	*0.10*	*0.09*	*0.12*	*0.12*
Cost of closing a branch	c_{4j}	*0.07*	*0.08*	*0.07*	*0.06*

Table 2.6 Weights of criteria for the transaction volume

Criteria	Notation	Individual	Entrepreneur	Corporate	Commercial
Number of potential customers	w_{1j}	*0.24*	*0.14*	*0.07*	*0.05*
Socioeconomic situation	w_{2j}	*0.06*	*0.06*	*0.06*	*0.07*
Social potential	w_{3j}	*0.08*	*0.08*	*0.07*	*0.06*
Commercial potential	w_{4j}	*0.16*	*0.18*	*0.39*	*0.44*
Competition	w_{5j}	*0.12*	*0.14*	*0.12*	*0.10*
Financial situation	w_{6j}	*0.16*	*0.22*	*0.18*	*0.15*
Ease of access	w_{7j}	*0.10*	*0.09*	*0.05*	*0.05*
Growth potential	w_{8j}	*0.08*	*0.09*	*0.06*	*0.08*

Table 2.7 Weights of criteria for the number of potential customers

Criteria	Notation	Individual	Entrepreneur	Corporate	Commercial
Total population	w_{9j}	*0.45*	*0.36*	*0.15*	*0.12*
Private institutions	w_{10j}	*0.24*	*0.29*	*0.33*	*0.42*
Financial institutions	w_{11j}	*0.19*	*0.23*	*0.36*	*0.35*
Education places	w_{12j}	*0.07*	*0.07*	*0.09*	*0.08*
Hospitals	w_{13j}	*0.05*	*0.05*	*0.07*	*0.03*

Table 2.8 Weights of subcriteria for socio-economic status

Sub criteria	Notation	Individual	Entrepreneur	Corporate	Commercial
Education level	w_{14j}	*0.68*	*0.71*	*0.68*	*0.62*
Summer houses	w_{15j}	*0.32*	*0.29*	*0.32*	*0.38*

Table 2.9 Weights of subcriteria for social potential

Sub criteria	Notation	Individual	Entrepreneur	Corporate	Commercial
Education places	w_{16j}	*0.51*	*0.51*	*0.33*	*0.30*
Hospitals	w_{17j}	*0.31*	*0.31*	*0.39*	*0.38*
Entertainment	w_{18j}	*0.18*	*0.18*	*0.28*	*0.32*

Table 2.10 Weights of subcriteria for commercial potential

Sub criteria	Notation	Individual	Entrepreneur	Corporate	Commercial
Private institutions	w_{19j}	*0.35*	*0.37*	*0.41*	*0.41*
Shopping centers	w_{20j}	*0.25*	*0.17*	*0.16*	*0.11*
Car parks	w_{21j}	*0.08*	*0.06*	*0.06*	*0.05*
Financial institutions	w_{22j}	*0.27*	*0.34*	*0.33*	*0.39*
Car services	w_{23j}	*0.05*	*0.06*	*0.04*	*0.04*

for individual, financial situation for entrepreneur; but commercial potential for corporate and commercial branches. From the perspective of socioeconomic status, education level is the most fundamental sub-criterion for all types of branches. As a social potential indicator; number of education places in the potential sites has the highest weight for individual and entrepreneur branches, while the number of hospitals is the most important criterion for corporate and commercial branches. Moreover, the number of private institutions has the highest importance for all types of branches as a commercial potential indicator. Table 2.11 summarizes final importance weights of all sub-criteria depending on the hierarchical structure and by branch type.

According to the weights in Table 2.5, the total distance from a place chosen for a branch to the other potential locations has almost equal importance as transaction volume. This shows that banks would be ill-advised to open branches very near to each other. Since the distance between potential locations cannot be merged to the decision model, a mathematical programming model is proposed to solve the problem.

Next, the pairwise comparison results found in above are used for the benefit-cost parameters in the objective function and for the determination of transaction volume. As it is proven in Sect. 2.3.3 and shown in Sect. 2.3.4.3, the proposed mathematical model is **NP**-Hard. Therefore, an optimal solution cannot be found for aforementioned bank branch location problem in Istanbul, which has 763 subdistricts and hence the proposed tabu search is applied.

The bank management wants to increase the number of individual branches in Istanbul to 161, entrepreneur branches to 20 and corporate branches to 13 by opening 9 new individual, 2 new entrepreneur branches and 1 new corporate branch. The best values of the tabu search parameters obtained in our experimental study are used as $k_1 = 8$, $k_2 = 12$ and tabu list size = 16 (as with the random instance of

Table 2.11 Final weights of criteria for transaction volume

Sub Criterion	Individual	Entrepreneur	Corporate	Commercial
Total population	0.11	0.05	0.01	0.01
Education level	0.04	0.05	0.04	0.04
Number of summer houses	0.02	0.02	0.02	0.03
Education places	0.05	0.05	0.03	0.02
Hospitals	0.04	0.04	0.03	0.03
Entertainment places	0.01	0.01	0.02	0.02
Private institutions	0.11	0.11	0.18	0.21
Shopping centers	0.04	0.03	0.06	0.05
Car park	0.01	0.01	0.02	0.02
Financial institutions	0.09	0.09	0.15	0.19
Car services	0.01	0.01	0.02	0.02
Number of competitor bank branches	0.12	0.14	0.12	0.10
Average household income	0.16	0.22	0.18	0.15
Ease of access	0.10	0.08	0.05	0.05
Growth potential	0.09	0.09	0.07	0.06

750 potential locations). The linear programming relaxation method is selected as the initialization approach due to its high performance in comparison to the random and criteria-based methods. Finally, the results of the tabu search algorithm with $k_3 = 5{,}000$ and 10,000 are shown in Table 2.12 (where TS refers to tabu search).

The solution found with maximum run time set to 10 h by *CPLEX* is used as the benchmark for the solution obtained by tabu search. The deviation (gap) in Table 2.12 is calculated as (*TS*/*CPLEX* solution)—1. Since we try to maximize the objective function value, the positive gap values mean that tabu search gives better results than *CPLEX*. As seen in Table 2.12, the tabu search solution is 0.032 % higher than *CPLEX* solution and it stays the same after $k_3 = 5{,}000$ iterations until we reach $k_3 = 10{,}000$ iterations. Also, it is observed that current 12 individual branches are closed and 21 new branches are opened, current 5 entrepreneur branches are closed and 7 new branches are opened, current 2 corporate branches are closed and 3 new branches are opened instead at the end of $k_3 = 10{,}000$ iterations. The number of commercial branches in Istanbul (3) stays the same while the location of a branch is changed. The results of the location problem of a Turkish National Bank's branches are summarized in Table 2.13.

The changes in the locations of this Turkish National Bank's branches in Istanbul given in Table 2.13 are illustrated in Fig. 2.4.

In Fig. 2.4, the dark and light grey colored boxes labeled 1, 2, 3 and 4 symbolize the locations of individual, entrepreneur, corporate and commercial branches that are closed and opened, respectively. Also, labels (1, 2, 3, 4) of closed (opened) branches are written with white (black) font inside the dark (light) grey boxes. The dark grey colored box labeled (1, 2) illustrates both individual and entrepreneur

Table 2.12 Results for the location problem of a Turkish National Bank in Istanbul

Number of current branches according to their types respectively	CPLEX (OFV)	CPLEX Time (s.)	Initial Solution 2 (OFV)	Initial Solution 2 Time (s.)	Initial Solution 2 Gap (%)	TS (Time) 5000 iterations	TS Gap (%) 5000 iterations	TS Time (s.) 10,000 iterations	TS Gap (%) 10,000 iterations
161,20,13,3	*38.871*	36,000	32.836	327.55	− 15.53	3,968.13	*0.032*	7,786.96	*0.032*

Table 2.13 Change in the locations of a Turkish National Bank's branches in Istanbul

	Branch type	Tabu search (District—Sub-district)
Closed Branches	Individual	Bahçelievler-Siyavuşpaşa; Bahçelievler-Çobançeşme; Bakırköy-Zeytinlik; Beylikdüzü-Barış; Beyoğlu-Tomtom; Eyüp-Karadeniz; Fatih-Saraç İshak; Güngören-Güven; Kadıköy-Eğitim; Kağıthane-Ortabayır; Küçükçekmece-Fatih; Küçükçekmece-Tevfikbey
	Entrepreneur	Bağcılar-Evren; Başakşehir-Ziya Gökalp; Beylikdüzü-Barış; Şişli-Kaptanpaşa; Tuzla-Aydınlı
	Corporate	Bayrampaşa-Muratpaşa; Güngören-Mehmet Nesih Özmen
	Commercial	Bağcılar-Evren
Opened Branches	Individual	Küçükçekmece-Halkalı Merkez; Üsküdar-Kandilli; Üsküdar-Salacak; Maltepe-Zümrütevler; Şişli-Esentepe; Ataşehir-Barbaros; Bakırköy-Zuhuratbaba; Şişli-Fulya; Ümraniye-İstiklal; Pendik-Yenişehir; Bahçelievler-Hürriyet; Esenler-Kazım Karabekir; Bakırköy-Sakız Ağacı; Ataşehir-Fetih; Sarıyer-Bahçeköy Yeni; Sultanbeyli-Hasanpaşa; Çekmeköy-Aydınlar; Kartal-Yalı; Gaziosmanpaşa-Sultançiftliği; Ataşehir-Mimar Sinan; Zeytinburnu-Telsiz
	Entrepreneur	Beşiktaş-Akat; Sarıyer-İstinye; Şişli-Teşvikiye, Kadıköy-Suadiye; Avcılar-Denizköşkler; Pendik-Yenişehir; Kartal-Cevizli
	Corporate	Şişli-Esentepe; Beşiktaş-Bebek; Kadıköy-Caddebostan
	Commercial	Ümraniye-Yukarı Dudullu

branches' closure in the sub-district Beylikdüzü-Barış. Similarly, the dark grey colored box labeled (2, 4) illustrates both entrepreneur and commercial branches' closure in the sub-district Bağcılar-Evren. As it is stated in Sect. 2.3.3, while multiple branches of the same type cannot be opened in the same locations; different types of branches are allowed to be opened in the same candidate location. Thus, the light grey colored box labeled (1, 2) illustrates both individual and entrepreneur branches' opening in the sub-district Pendik-Yenişehir while (1, 3) illustrates individual and corporate branches' opening in Şişli-Esentepe. As it can be seen in Fig. 2.4, closed branches are mostly located in the western (European) side of the

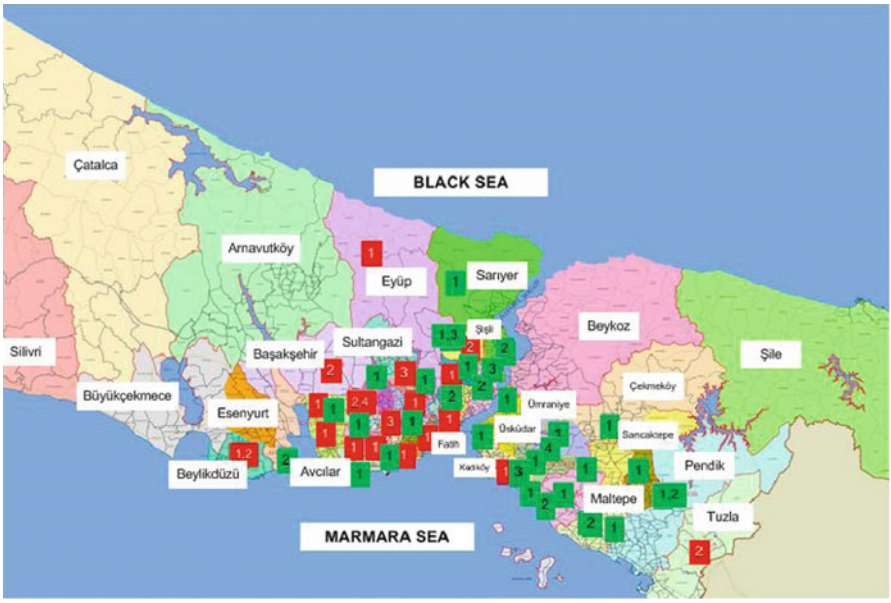


Fig. 2.4 Solution for the branch location problem of a Turkish National Bank's in Istanbu

city, since a considerable part of the current branches are located here. It is observed that closed branches are the ones located in the sub-districts having multiple branches or near to the current branches of same type. On the other hand, locations with high transaction volume or those with no surrounding nearby branches are selected as new branch locations. Moreover, it is easily seen that the service network is enlarged by closing branches which are very close to each other and opening new ones in the easily accessible locations instead. The results are also approved by the experts of the related bank and other experts who contributed in identifying the criteria and specifying the importance of the criteria.

In order to test the validity of the results and evaluate the robustness of the solution, a sensitivity analysis is conducted by selecting four main criteria, namely transaction volume, distance between branches, cost of opening a new branch and cost of closing a branch. Their importance weights are varied by $\pm 10\,\%$ and $\pm 20\,\%$. The results of the criteria sensitivity obtained at the end of $k_3 = 10{,}000$ iterations are summarized in Table 2.14.

As seen in Table 2.14, the solution does not alter for the changes in four main criteria in the interval of $+10\,\%$ and $-10\,\%$. If the weights obtained via expert judgments are changed by $20\,\%$ for the most important criteria, which are transaction volume and distance between branches, only the results for a few scenarios change slightly. Thus, we find the results of the method to be robust.

Table 2.14 Sensitivity analysis of the solution in Istanbul obtained by tabu search

	Change ratio (%)	Change
Original solution	–	Table 2.13
Transaction volume	+ 20	Do not close individual branches in Bakırköy-Zeytinlik and Güngören-Güven sub-districts; do not open individual branches in Pendik-Yenişehir and Sultanbeyli-Hasanpaşa sub-districts. Do not close the entrepreneur branch in Şişli-Kaptanpaşa sub-district and do not open the entrepreneur branch in Avcılar-Denizköşkler sub-district
	+ 10	Same as the original solution
	− 10	Same as the original solution
	− 20	Do not close individual branches in Beylikdüzü-Barış, Beyoğlu-Tomtom and Kağıthane-Ortabayır sub-districts; do not open individual branches in Şişli-Esentepe, Bakırköy-Zuhuratbaba, Esenler-Kazım Karabekir and Kartal-Yalı sub-districts, select Bağcılar-Demirkapı sub-district to open a new individual branch. Do not close the entrepreneur branch in Beylikdüzü-Barış sub-district and do not open the entrepreneur branch in Avcılar-Denizköşkler sub-district
Distance between branches	+ 20	Do not close individual branches in Beylikdüzü-Barış and Kağıthane-Ortabayır sub-districts. Do not open individual branches in Şişli-Esentepe and Esenler-Kazım Karabekir sub-districts
	+ 10	Same as the original solution
	− 10	Same as the original solution
	− 20	Do not close the individual branch in Bakırköy-Zeytinlik sub-district and do not open the individual branch in Sultanbeyli-Hasanpaşa sub-district
Cost of opening a new branch	+ 20	Same as the original solution
	+ 10	Same as the original solution
	− 10	Same as the original solution
	− 20	Same as the original solution
Cost of closing a branch	+ 20	Same as the original solution
	+ 10	Same as the original solution
	− 10	Same as the original solution
	− 20	Same as the original solution

An important limitation of our study lies with the trouble of finding real data for some criteria such as daytime population for number of potential customers. Moreover, since we do not have real transaction volume data or banking outputs such as volume of deposits or credits of the mentioned bank, we were unable to use an objective weighting method and confirm the accuracy of the importance weights obtained via pairwise comparison. On the other hand, we can see that the results of the model can easily be applied in real life directly, since we have considered the most important criteria for the branch location problem. Also, we can use the proposed mathematical model for any bank since it considers both opening new and closing inefficient branches at the same time. Furthermore, opening a new branch in the most exact place is suggested by penalizing the opening of same-type branches near one another, taking ease of access into consideration and selecting the most frequently visited places as candidate locations by the help of GIS.

2.5 Conclusion and Future Research

In this study, we presented an integrated methodology to make decisions for the location of bank branches. As we stated in the literature review, *MCDM* is a common technique employed, and there are not many studies using mathematical models to solve this problem. However, both *MCDM* and mathematical modeling approaches have some shortcomings. *MCDM* deals mainly with analyzing decision problems and evaluating the alternatives/criteria based on decision makers' preferences. It is clear that *MCDM* lacks the capability of an optimization aspect. On the other hand, mathematical models have limited capability of considering expert opinion. By combining these two techniques, we benefit from the strengths and discard the shortcomings of both methods. To this aim, we determine the most relevant criteria and obtain their importance weights via extensive literature review and expert opinion. We also develop a new mathematical model to determine the exact locations of the branches. Our model permits both the opening and closing of branches at the same time, which shows that it can be used not only for deciding the best locations for branches, but also for measuring the performance of existing branches. Since the model is **NP**-Hard and an optimal solution for the problems with large number of candidate locations might not be so easily found using an exact method, we develop and propose a tabu search algorithm. We experimentally show that tabu search gives better solutions than *CPLEX* in terms of both the objective function value and computational time. We apply our proposed methodology to the branch location problem in Istanbul for one of the biggest banks in Turkey. We find the results to be robust and validate them by both sensitivity analysis and expert opinion.

Future studies may include the presentation of a multi-period mathematical model to support banks' long-term strategies, application of a different metaheuristic to compare with the tabu search results and researching the applicability of the proposed methodology in similar location problems (e.g., supermarkets, restaurants etc.)

References

Abbasi GY (2003) A decision support system for bank location selection. Int J Comput Appl Technol 16:202–210

Ahsan MK, Bartlema J (2004) Monitoring healthcare performance by analytic hierarchy process: a developing-country perspective. Int Trans Oper Res 11:465–478

Alexandris G, Giannikos I (2010) A new model for maximal coverage exploiting GIS capabilities. Eur J Oper Res 202:328–338

Arabani AB, Farahani RZ (2012) Facility location dynamics: an overview of classifications and applications. Comput Ind Eng 62:408–420

Aras H, Erdogmuş S, Koç E (2004) Multi-criteria selection for a wind observation station location using analytic hierarchy process. Renewable Energy 29(8):1383–1392

Arostegui MA, Kadipasaoglu SN, Khumawala BM (2006) An empirical comparison of tabu search, simulated annealing, and genetic algorithms for facilities location problems. Int J Product Econ 103(2):742–754

Badri MA (2001) A combined AHP—GP model for quality control systems. Int J Product Econ 72:27–40

Banking and sector information (2014) http://www.tbb.org.tr/tr/banka-ve-sektor-bilgileri/banka-bilgileri/subeler/65. Accessed 23 June 2014

Baron O, Berman O, Kim S, Krass D (2009) Ensuring feasibility in location problems with stochastic demands and congestion. IIE Trans 41:467–481

Başar A, Çatay B, Ünlüyurt T (2011) A multi-period double coverage approach for locating the emergency medical service stations in Istanbul. J Oper Res Soc 62(4):627–637

Başar A, Çatay B, Ünlüyurt T (2012) A taxonomy for emergency service station location problem. Optim Lett 6(6):1147–1160

Berman O, Krass D (2002) The generalized maximal covering location problem. Comp Oper Res 29:563–581

Boufounou PV (1995) Evaluating bank branch location and performance: a case study. Eur J Oper Res 87:389–402

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

Camanho AS, Dyson RG (2005) Cost efficiency measurement with price uncertainty: a DEA application to bank branch assessments. Eur J Oper Res 161:432–446

Carlsson C, Fuller R (1996) Fuzzy multiple criteria decision making: recent developments. Fuzzy Sets Syst 78:139–153

Chen SJ, Hwang CL (1993) Fuzzy multiple attribute decision-making, methods and applications. Lecture notes in Economics and Mathematical systems 375. Springer, Heidelberg

Chen H, Barna BA, Rogers TN, Shonnard DR (2001) A screening methodology for improved solvent selection using economic and environmental assessments. Clean Prod Processes 3: 290–302

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

Cinar N (2009) A decision support model for bank branch location selection. World Acad Sci Eng Technol 60:126–131

Cinar N, Ahiska SS (2010) A decision support model for bank branch location selection. In YAEM 2010. Proceedings of the 2010 International Conference on Industrial Engineering and Operations Management, Sabancı University, Istanbul, June 30-July 2, (CD-ROM)

Clawson CJ (1974) Fitting branch locations, performance standards, and marketing strategies to local conditions. J Marketing 38:8–14

Cook WD, Seiford LM, Zhu J (2004) Models for performance benchmarking: measuring the effect of e-business activities on banking performance. Omega 32:313–322

Curtin KM, Hayslett-McCall K, Qiu F (2010) Determining optimal police patrol areas with maximal covering and backup covering location models. Netw Spat Econ 10(1):125–145

Daskin M (1983) A maximum expected covering location model: formulation, properties and heuristic solution. Transp Sci 17:48–70
Daskin MS, Stern EH (1981) A hierarchical objective set covering model for emergency medical service vehicle deployment. Transp Sci 15:137–152
Doyle P, Fenwick I, Savage GP (1981) A model for evaluating branch location and performance. J Bank Res 12:90–95
Erdemir ET, Batta R, Spielman S, Rogerson PA, Blatt A, Flanigan M (2010) Joint ground and air emergency medical services coverage models: a greedy heuristic solution approach. Eur J Oper Res 207:736–749
Fernandez I, Ruiz MC (2009) Descriptive model and evaluation system to locate sustainable industrial areas. J Cleaner Prod 17(1):87–100
Gendreau M, Laporte G, Semet F (2000) A dynamic model and parallel tabu search heuristic for real time ambulance relocation. Parallel Comput 27:1641–1653
Glover F (1977) Heuristics for integer programming using surrogate constraints. Decis Sci 8: 156–166
Glover F (1989) Tabu search—part I. ORSA J Comput 1(3):190–206
Grabowski J, Wodecki M (2004) A very fast tabu search algorithm for the permutation flow shop problem with makespan criterion. Comput Oper Res 31:1891–1909
Hakimi S (1964) Optimum locations of switching centres and the absolute centres and medians of a graph. Oper Res 12:450–459
Hansen P, Brimberg J, Urosevic D, Mladenovic N (2009) Solving large *p*-median clustering problems by primal-dual variable neighborhood search. Data Min Knowl Disc 19:351–375
Huff D (1963) A probabilistic analysis of shopping center trade areas. Land Econ 39:81–90
Hwang H (2002) Design of supply-chain logistics system considering service level. Comput Industrial Eng 43:283–297
Ishizaka A, Lusti M (2004) An expert module to improve the consistency of AHP matrices. Int Trans Oper Res 11:97–105
Jablonsky J, Fiala P, Smirlis Y, Despotis DK (2004) DEA with interval data: an illustration using the evaluation of branches of a Czech bank. Central Eur J Oper Res 12:323–337
Kaufman G, Mote R (1994) A review from the Federal Reserve Bank of Chicago. Federal Reserve Bank of Chicago, Chicago
Kuehn A, Hamburger M (1960) A heuristic program for locating warehouses. Manage Sci 9: 643–666
Malczewski J (1999) GIS and multicriteria decision making. Wiley, New York
Malek M, Guruswamy M, Pandya M, Owens H (1989) Serial and parallel simulated annealing and tabu search algorithms for the traveling salesman problem. Ann Oper Res 21:59–84
Manandhar R, Tang JCS (2002) The evaluation of bank branch performance using data envelopment analysis: a framework. J High Technol Manage Res 13:1–17
Manne A (1964) Plant location under economies of scale. decentralization and computation. Manage Sci 11:213–235
Marianov V, ReVelle CS (1995) Facility location. Springer, Berlin
Meidan A (1983) Distribution of bank services and branch location. Int J Phys Distrib Managerial Manage 13(3):5–18
Miliotis P, Dimopoulou M, Giannikos I (2002) A hierarchial location model for locating bank branches in a competitive environment. Int Trans Oper Res 9:549–565
Miller GA (1956) The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychol Rev 63:81–97
Min H (1989) A model based decision support system for locating banks. Inform Manag 17: 207–215
Min H, Melachrinoudis E (2001) The three-hierarchical location-allocation of banking facilities with risk and uncertainty. Int Trans Oper Res 8:381–401
Morrison PS, O'Brien R (2001) Bank branch closures in New Zealand: the application of a spacial interaction model. Appl Geogr 21:301–330

Murray AT, Tong D, Kim K (2010) Enhancing classic coverage location models. Int Reg Sci Rev 33(2):115–133
Olsen LM, Lord JD (1979) Market area characteristics and branch bank performance. J Bank Res Summer 10:102–110
Osman IH, Kelly JP (1996) Meta-heuristics: theory and applications. Kluwer Academic Publishers, Boston
Pardalos PM, Resende MGC (2002) Handbook of applied optimization. Oxford University Press, New York
Pastor JT (1994) Bicriterion programs and managerial location decisions: application to the banking sector. J Oper Res Soc 45(12):1351–1362
Portela MCAS, Thanassoulis E (2007) Comparative efficiency analysis of Portuguese bank branches. Eur J Oper Res 177:1275–1288
Rahgan SH, Mirzazadeh A (2012) A new method in the location problem using fuzzy evidential reasoning. Eng Techno 4(22):4636–4645
Ravallion M, Wodon Q (2000) Banking on the poor? Branch location and nonfarm rural development in Bangladesh. Rev Devel Econ 4:121–139
Retail Banker International (2013) US Branch numbers fall for fourth year running (2014). https://dscqm8cqg6d5o.cloudfront.net/uploads/articles/pdfs/mnetisgnefhmblsrdclkablzye_rbioct13issue694usbranches.pdf. Accessed 23 June 2014
ReVelle CS, Swain R (1970) Central facilities location. Geogra Anal 2:30–42
Ribeiro CC, Hansen P (2002) Essays and surveys in metaheuristics. Kluwer Academic Publishers, Norwell
Saaty TL (1980) The analytic hierarchy process. McGraw-Hill Inc, New York
Saaty TL (1990) How to make a decision: the analytic hierarchy process. Eur J Oper Res 48:9–26
Sato Y (2004) Comparison between multiple-choice and analytic hierarchy process: measuring human perception. Int Trans in Ope Res 11:77–86
The Banks Association of Turkey (2014) Available at http://www.tbb.org.tr/tr/banka-ve-sektor-bilgileri/banka-bilgileri/subeler/65. Accessed 12 Apr 2015
Toregas CR, Swain R, ReVelle CS, Bergman L (1971) The location of emergency service facilities. Oper Res 19:1363–1373
Tzeng GH, Teng MH, Chen JJ, Opricovic S (2002) Multi criteria selection for a restaurant location in Taipei. Int J Hosp Manage 21(2):171–187
Wang Q, Batta R, Rump CM (2002) Algorithms for a facility location problem with stochastic customer demand and immobile servers. Ann Oper Res 111:17–34
Wang Q, Batta R, Bhadury J, Rump CM (2003) Budget constrained location problem with opening and closing of facilities. Comput Oper Res 30:2047–2069
Wu CR, Lin CT, Chen HC (2007) Optimal selection of location for Taiwanese hospitals to ensure a competitive advantage by using the analytic hierarchy process and sensitivity analysis. Build Environ 42(3):1431–1444
Xia L, Yin W, Dong J, Wu T, Xie M, Zhao Y (2010) Hybrid nested partitions algorithm for banking facility location problems. IEEE Transautomation Sci Eng 7(3):654–658
Youssef H, Sait SM, Adiche H (2001) Evolutionary algorithms, simulated annealing and tabu search: a comparative study. Eng Appl Artif Intell 14:167–181
Zhang L, Rushton G (2008) Optimizing the size and locations of facilities in competitive multi-site service systems. Comput Oper Res 35:327–338
Zhang G, Habenicht W, Spieß WEL (2003) Improving the structure of deep frozen and chilled food chain with tabu search procedure. J Food Eng 60(1):67–79
Zhao L, Garner B, Parolin B (2004) Branch bank closures in Sydney: a geographical perspective and analysis. Proceedings of the 12th International Conference on Geoformatics, Sweden. June 7–9:541–548

Chapter 3
Location Modeling for Logistics Parks

Joyendu Bhadury, Mark L. Burkey and Samuel P Troy

3.1 Background

Location theory is an extensive field with hundreds of published papers drawing upon disciplines as diverse as operations research, operations management, regional science, geography, economics, computer science and mathematics. Some recent and well-known references that survey the field are Hamacher and Drezner (2002), Current et al. (2004), Revelle and Eiselt (2005) and Eiselt and Marianov (2011). Yet, the discipline suffers from a paucity of publications that involve real-life applications, especially case studies. This paucity is noted in Current et al. (2004), which begins with the statement "Much of the literature on facility location modeling has not been directed to specific applications (that is, case studies)" and pointedly reinforced in Revelle and Eiselt (2005, p. 15) who state "...when it comes to applications, there appears to be a significant deficit, at least as compared to other, similar, fields." This is of particular concern since substantial inefficiencies can occur if facilities are located sub-optimally—for example, Rushton (1988, p. 101), illustrates this fact in an application involving the location of health clinics in India. Despite this paucity however, there are a few applications of facilities location models documented in the published literature from a variety of different disciplines and countries, as first noted in Eiselt (1992). For example, Rushton (1988) contains references on the use of location-allocation models in siting public-service facilities such as health clinics and schools, in the developing world.

J. Bhadury (✉) · S. P Troy
Bryan School of Business and Economics, University of North Carolina at Greensboro, Greensboro, NC 27402, USA
e-mail: joy_bhadury@uncg.edu

S. P Troy
e-mail: sptroy@uncg.edu

M. L. Burkey
School of Business and Economics, North Carolina A & T State University, Greensboro, NC 27411, USA
e-mail: burkeym@ncat.edu

H. A. Eiselt, V. Marianov (eds.), *Applications of Location Analysis,* International Series in Operations Research & Management Science 232, DOI 10.1007/978-3-319-20282-2_3

This work extends the extant literature on location related case studies based on a project related to the location of a Logistics park in the southeastern region of North Carolina, USA. While a portion of this work draws upon that project, the primary objective of this case study is more than to simply to describe the location modeling process used in that application. Instead, the main objective is to distill lessons learned that can help guide future academic research in location theory. To that end, this case study begins with solving the problem at hand using a theoretical approach so that a contrast may be drawn with how the modeling was done in practice. In turn, that contrast forms the basis of the final section on conclusions and lessons learned.

The remaining case study is organized as follows. Sect. 3.2 gives the necessary background to the case study by providing the contextual framework of the project and an introduction to *NCSE*, detailing the socio-economic and infrastructural information that informed the research. The next two sections are devoted to modeling the location of a Logistics Park in *NCSE*. While Sect. 3.3 details how a "theoretical" modeling approach would have worked, Sect. 3.4 describes the actual approach used when conducting the study within a structured framework, referred to by its acronym, *SIRC*. Sect. 3.5 ends this work by describing the follow-up to the Seven Portals Study, lessons learned and recommendations for location scientists involved in modelling real-life location/site selection problems.

3.1.1 Background: Logistics Parks, Seven Portals Study and NC Southeast

A "Logistics Park" is defined as a "development concept in which distribution centers typically seen in a suburban area are built in a park like setting, created by landscaping" (Gardner, Kansas Official Website 2015). It is usually populated by warehouses, distribution centers and logistics-related companies/offices. In almost all cases, it is also an intermodal facility where truck trailers and containers are transferred between trucks and the railroad. The concept of Logistics Park has been applied and used in various international settings and may be referred to by other names such as Freight Village, Güterverkehrszentrum, Interporto etc. Many such facilities exist throughout the world and a few examples include Burlington Northern Logistics Park (Illinois, USA), large cargo airports such as Hong Kong and Memphis, Pinghu Logistics Park in Shenzen, China and Schipol Logistics Park in the Netherlands. In the last decade, academic research has begun to emerge in location theory around logistics parks; for instance, Lee et al. (2001) and Lee and Yang (2003) discuss models and strategies for the location of such facilities, whereas El Amrani (2007) investigates the impact of international Logistics Parks on supply chains of multinationals. To the best of our knowledge, however, this case study is the first that is based on an actual application related to Logistics Park location.

As to the history of why the state of North Carolina undertook planning for locating Logistics Parks, one needs to begin with the state-wide strategic plan for logistics (List et al. 2008). In order to implement the same, the North Carolina State

Legislature established the Governor's Logistics Task Force (*GLTF*) on December 7, 2009 and assigned this group seven tasks (see *GLTF*-Final Report 2012) which included the following:

- "Conduct a thorough inventory and evaluation of existing public and private transportation and commerce assets, including ports, inland ports, airports, highways, railroads, major distribution centers, and business and industrial parks.
- Report on the current system for moving goods and people, including the condition of the system, its overall performance, and its safety.
- Project future needs for the state's multi-modal transportation system and explore challenges and opportunities in meeting those needs.
- Explore innovative ideas in transportation and economic development that can help support the state's logistics capacity, including public private partnerships."

The final report from *GLTF* (*GLTF*-Final Report 2012) details the accomplishment of all its objectives, including the four listed above. In order to implement its charge, *GLTF* undertook some key studies, one of them being the "Seven Portals Study" (List and Foyle 2011), which was administered by the North Carolina Department of Transportation (*NCDOT*). The purpose of that study was to "describe ways in which North Carolina's transportation infrastructure investments can help with economic development and the creation of jobs" (List and Foyle 2011). The overall goal was investigating potential Logistics Parks throughout the state and identifying what infrastructure improvements are needed to support such a facility at that location. It was made clear that the study would not recommend specific sites above others but merely present facts about each to assist the North Carolina State Legislature make the final decisions. The final report of the Seven Portals Study by List and Foyle (2011) was based on smaller regional reports each of which focused on one specific region of North Carolina. This case study focuses on one such region, namely, North Carolina's Southeastern Region (*NCSE*) (see Fig. 3.1) and is drawn, in part, from the final report submitted for the same (Bhadury and Troy 2011).

For purposes of regional economic development, the North Carolina Department of Commerce (http://www.nccommerce.com/) has divided the state into seven distinct geographical regions, each with an established "Regional Partnership" agency to foster a collective approach to regional economic development; see http://www.thrivenc.com/ for a complete listing. North Carolina's Southeast Region, abbreviated as *NCSE*, (see Figs. 3.1 and 3.2) is one of these seven. The region is composed of 14 counties stretching from the Atlantic Ocean to Sandhills region of North Carolina. The region's 500,000 strong workforce, is employed in a diverse cross-section of workers engaged in agriculture, wood products, manufacturing, wholesale trades (i.e., distribution), construction, healthcare, government and the professions. Economically, *NCSE* is substantially diverse. Along with a strong agricultural, wood products and food processing sector supported by companies such as Smithfield Foods, International Paper Company and Campbell Soup, it is also home to high tech companies like GE Nuclear Energy, Pharmaceutical Product Development Company (*PPD*) and Corning.

Fig. 3.1 Relative position of *NCSE* in *Eastern* USA

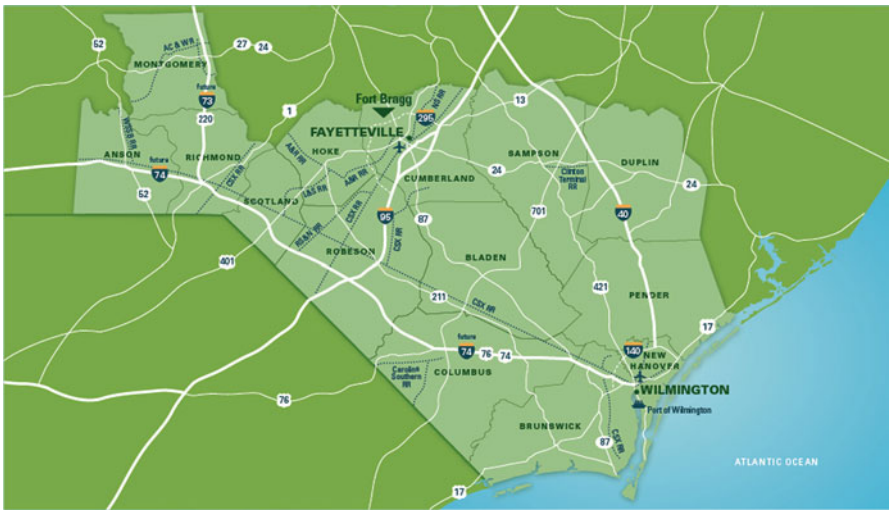


Fig. 3.2 North Carolina's *south east* region. (Source: http://www.ncse.org/)

Two sectors are of outstanding importance to overall economy of *NCSE* and deserve a special mention: military (Fort Bragg) and shipping (Port of Wilmington). As for the military sector, it is centered on Fort Bragg near Fayetteville. Established in 1918, Fort Bragg is now the largest US Army base by population in USA and

includes 10 % of the US Army's active component forces. The base covers 161,047 acres and the military population comprises almost 60,000 officers and enlisted men and women with an additional 21,000 civilian workers employed directly by the army base. The *BRAC* (Base Realignment and Closure) process undertaken since 2005 by the US Department of Defense has contributed significantly to the growth of Fort Bragg. For example, European-based forces have been relocated to Fort Bragg. Additionally, several aircrafts are now located at Fort Bragg to form an Air Force Reserve/active duty associate unit as is Air Force Reserve Command operations and maintenance. Finally, *BRAC* has resulted in two major headquarters, Army Forces Command and Army Reserve Command to be located at Fort Bragg. With such a significant presence, it is no surprise that Fort Bragg and thus, the military sector, is a major economic driver in *NCSE*. The current annual payroll of Fort Bragg is estimated at $ 3.5 billion which generates a direct and indirect annual impact of approximately $ 10.9 billion in the immediate region (*NCSE* Regional Economic Profile 2013).

As for Port of Wilmington, it is North Carolina's sole container shipping port. The port is owned and operated by the North Carolina State Ports Authority (http://www.ncports.com) and offers terminal facilities serving container, bulk and break-bulk operations. The 42 ft navigational channel is complemented on the land side by available modern transit and warehouse facilities, state-of-the-art Panamax container cranes and support equipment, nine berths with 6768 ft. of wharf frontage and the latest in cargo management technology (North Carolina State Ports Authority 2015). Railroad service is provided by *CSX* Transportation which has daily service for boxcar, tanker and general cargo services. In order to facilitate the position of the port as a hub of international trade, it is a designated Foreign Trade Zone that is administered by North Carolina Department of Commerce. The port itself hosts almost 1 million sq. ft. of storage. The port has an annual traffic of over 2.1 million containers, which makes it smaller compared to the neighboring East Coast ports of Charleston, Savanah and Norfolk. Nonetheless, it provides container and bulk shipping to/from most world markets. Ocean carriers that call on the Port of Wilmington include major international shipping lines such as "K" Line America, Cosco Container Lines, Hanjin Shipping Company, Maersk Line and National Shipping Company of Saudi Arabia. As for the economic impact of the port and the shipping industry overall, Findley et al. (2014) estimate that this port has an annual economic impact of $ 12.9 billion and along with a smaller bulk port in Morehead City, supports 76,000 direct and indirect jobs.

Notwithstanding the above, the *NCSE* is a predominantly rural part of North Carolina, which suffers from significant unemployment and poverty issues. Figure 3.3 reveals that the region has few towns/counties that have a population above 1000. Also, as evident from Table 3.1, the region had an overall unemployment rate of 6.46 % as of December 2014 (above the state rate of 5.5 %) with 11 out of the 14 counties having a worse than state average unemployment rate. In particular, Scotland County, with a 10 % unemployment rate, was the 2nd highest in the state and has historically been in a similar position for a long while. With regards to family income, the picture is better balanced. As of end of fiscal year 2013-14, the average

Fig. 3.3 *NCSE*—towns over 1000, over 10,000 labeled

Table 3.1 *NCSE* counties at a glance. (Source: http://quickfacts.census.gov/, http://www.ncse.org/ and https://desncc.com/deshome)

County	Population (2014)	Median family income 2013-14 (US $)	Unemployment in Dec 2014 (%)
Anson county	26,948	34,659	6.6
Bladen county	35,190	42,473	8.1
Brunswick county	113,235	54,172	6.1
Columbus county	54,899	41,957	7.3
Cumberland county	344,167	55,675	6.3
Duplin county	58,710	33,172	6
Hoke county	46,265	49,400	5.7
Montgomery county	27,798	32,946	5.8
New Hanover county	196,320	65,219	5
Pender county	54,546	53,225	5.5
Richmond county	46,893	44,189	7
Robeson county	132,092	41,304	7.7
Sampson county	65,513	46,947	5.1
Scotland county	37,059	49,852	10.1
NCSE region (weighted average of 14 counties)		*51,091*	*6.46*
NC (overall)	*9,848,060*	*46,334*	*5.5*

Table 3.2 *NCSE's* employment profile (3rd quarter, 2012). (Source: http://www.ncse.org/)

Industry	Number of people employed in *NCSE* by the industry	Average weekly wages in the industry (in $)
Agriculture, forestry, fishing & hunting	6075	477.74
Mining	232	843.94
Utilities	2192	1576.91
Construction	15,764	713.37
Manufacturing	37,270	868.16
Wholesale trade	10,029	898.68
Retail trade	49,036	453.30
Transportation & warehousing	9910	786.35
Information	5068	841.25
Finance & insurance	7633	913.92
Real estate & rental & leasing	5210	648.31
Professional & technical services	14,642	1139.03
Management of companies & enterprises	2463	830.73
Administrative & waste services	19,126	513.31
Educational services	33,587	727.48
Health care & social assistance	61,993	760.62
Arts, entertainment & recreation	5557	333.21
Accommodation & food services	41,116	270.90
Other services except public administration	8360	483.12
Public administration	31,981	866.52

median family income was $ 51,091, which is higher than the state-wide average of $ 46,334. The $ 51,091 figure was buoyed primarily by the two urban centers in the region, Wilmington and Fayetteville, and the economic sectors around them. Overall, half of the region had a median family income below the state average. The distribution of the employment in this region across the various industry sectors and the median weekly wages in each are displayed in Table 3.2.

The region also has a strong logistical infrastructure for all modes of transportation. With regards to roadways, the bulk of this region is contained within a triangle of three major US interstates—Interstate 95, Interstate 40 and Interstate 73/74 and more importantly, the *NCDOT* has targeted these highways for improvement in the next decade. These three collectively make the region accessible to over half of the population of USA within one day's driving. Wilmington hosts North Carolina's only container shipping port, namely, the Port of Wilmington. The railroads in the region include both of USA's major freight carriers: *CSX* and Norfolk-Southern as well as several short-haul local lines such as Aberdeen & Rockfish, Carolina-Southern and Clinton Terminal Railroad. Finally, air service is provided at two primary airports: Wilmington and Fayetteville that offer connections via major American carriers such as Delta, US Airways, and American Eagle.

The regional economic development agency, known as North Carolina's Southeast (http://www.ncse.org/) has identified the following nine as the targeted industry clusters for future economic development of the region: advanced textiles, aviation and aerospace, biotechnology, building products, defense, distribution and logistics, energy, food processing and agri-industry, metalworking. As evident, critical to each of these is the presence of an efficient logistical infrastructure in the region including state of the art multi-modal Logistics Park with expeditious connection to the Port of Wilmington. This formed the basis for their desire to locate one in their region.

3.2 A Theoretical Approach to Determine the "Optimal" Location

In order to provide a contrast with how location decisions for Logistics Parks are made in practice as opposed to theory, we will begin with modeling the problem of locating a logistics park in *NCSE* as a traditional 1-median problem in location theory and analyze the results.

As is typical for any location model that seeks to determine the optimal location of a service facility, we have to begin by modeling the "consumers" of the services that would be provided by the logistics park. For that purpose, we represent *NCSE* as a collection of demand points on the two-dimensional Euclidean plane. With that, assume that there are N demand points located in *NCSE*, each of which is referenced as demand point i, where $i = 1,2,\ldots,N$. We assume here that each demand point represents an appropriately defined cluster of businesses or citizens of *NCSE* that are potential users of the services provided by a logistics park. It must be noted here that such aggregations essentially convert what is a location problem with continuous demand into one with discrete demand. In turn, that generates agglomeration errors that are well studied in location theory, see Francis et al. (2009) for a survey. Nonetheless, for purposes of illustration we will proceed with this demand agglomeration.

Further, for each demand point i above, let (a_i,b_i) represent the abscissa and the ordinate respectively and let w_i represent the total demand for the services provided by the logistics park. Additionally, let $X=(a_X,b_X)$ represent any arbitrary point on this same plane. Then, the l_p distance between X and demand point is i, denoted by $d_p(X, i)$, is given by the following formula:

$$d_p(X,i) = \left[|a_x - a_i|^p + |b_X - b_i|^p\right]^{\frac{1}{p}} \tag{3.1}$$

Perhaps the most commonly used version of (1) above is when $p=2$, i.e., L_2 distance, given by

$$d_2(X,i) = \sqrt{(a_X - a_i)^2 + (b_X - b_i)^2} \tag{3.2}$$

As is well known, the l_2 distance formula (2) shown above is also known as the Euclidean, or straight-line distance, which measures the length of the line segment

connecting X and demand point i. This metric assumes that travel takes place along straight lines, and without barriers.

Another distance function that is used, albeit less commonly is the squared Euclidean distance given by

$$d^2(X, i) = (a_X - a_i)^2 + (b_X - b_i)^2 \tag{3.3}$$

The squared Euclidean distance metric (3) assumes that the disutility of travel increases quadratically. For example, the disutility of commuting is more costly for the 2nd mile than for the 1st mile, and more costly for the 10th mile than for the 9th. There are two reasons behind 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. It must also be mentioned that this metric is widely used by practitioners in selecting locations for distribution centers and warehouses.

Another commonly used distance metric is the l_1 metric when $p = 1$, also known as the Manhattan metric or the rectilinear distance metric. In this case, the distance between X and demand point i is given by:

$$d_1(X, i) = |a_X - a_i| + |b_X - b_i| \tag{3.4}$$

Two good sources of information for different kinds of l_p metrics and their implications in practice are Brimberg and Love (1995) and Burkey et al. (2011). As shown in Brimberg et al. (1994) and Fernández et al. (2002), variants of the l_p norm with intermediate values of p between 1 and 2 best predict road distances in real-world applications. This is why the analytical results in this section include those for $p = 1.5$.

Notwithstanding the definition of distance, the objective of locating a Logistics Park would obviously be to be as proximate to the users as possible. To that end, for any given l_p-metric, the 1-median of these of demand points is given by a point X^*, with the property that

$$\sum_{i=1}^{N} w_i d_p(X^*, i) \leq \sum_{i=1}^{N} w_i d_p(X, i) \text{ for all points } X \text{ in } NCSE \tag{3.5}$$

In other words, a logistics park located at X^* would be guaranteed to have the minimum possible total weighted average distance to all the customer demand points $i = 1, 2, \ldots, N$, where the weight of each point i is w_i and represents the total demand at that point. It is useful to note that when $p = 2$, i.e., in the case of the Euclidean metric, the 1-median is also referred to as the Fermat-Weber point, a classical problem in geometry going back to the seventeenth century.

As for computing the 1-median for a given set of points, the best known algorithm is given in Weiszfeld (1937), which has recently been translated into English (Weiszfeld and Plastria 2009). Numerous improvements to the original Weiszfeld algorithm have been proposed over time, see Plastria (2011) for a comprehensive update on them. Finally, we note that in the special case where the distance metric

used is the Squared Euclidean Metric given in (3), the 1-median is referred to as the *center of mass* or *center of gravity*—we will use the former term. Determining the center of mass is straightforward, as it involves taking the weighted average of the coordinates of the demand points. However, it is important that one uses a rectangular projection of abscissa and ordinate coordinates rather than longitude and latitude (which are spherical coordinates).

With the backdrop above, it is clear that an analytical approach to locating a logistics park in *NCSE* would essentially determine the 1-median as the optimal location. The next questions then are: how do we define the appropriate clusters to aggregate demand and what distance metric (i.e., the value of p) do we use in the distance calculation formula in determination of the 1-median. Given that many such answers are possible, we provide below an analysis of how different solutions for the 1-median problem (5) may be obtained based on different answers to each of the two questions above. We refrain from prescribing one solution over another, but do present a rationale for why they differ from each other.

The first issue that needs to be resolved is a determination of where people live or work in *NCSE*, since that information is the foundation of all location modelling. There are many choices for the demand points. Two broad categories of choices are (*i*) choice of demander, and (*ii*) choice of scale. The simplest choice to use for the demanders would be to use several specific points if there are only a few relevant demand points, e.g., specific distribution centers or factories. Another common option (for small area studies) is to use the number of people at the location of their residences as the intensity and location of demand points. However, in the current case, it might make more sense to use the location (and size) of businesses in formulating the demand points.

Whether one uses locations of people or businesses, the coarsest practical scale for *NCSE* would be to use the 14 counties as the demand points. The drawback to doing this is that one must assume a single point to represent many demand points that are spread out over a large area, probably non-uniformly. When using people to represent the demanders, the finest practical scale is to use census blocks or block groups[1]. The block is the smallest unit of data released by the US Census Bureau. Typically, only population counts are released at this level, and at roughly 29,000 blocks per county in North Carolina[2], might be unnecessarily fine. Table 3.3 gives a breakdown of several intermediate scales that might be chosen, using North Carolina's numbers for reference.

When using businesses as the demand points, the finest practical scale for most analyses is to use business data aggregated at the zip code level. A convenient source of such data in the US is the economic census from the *BLS*. For each zip code, data are available on how many firms there are, how many total employees work in these firms, and the total payroll.

[1] Though some work on a small scale, such as for a single city or county could use individual addresses as demand nodes e.g., Qabaja and Bikdash (2014).

[2] https://www.census.gov/geo/maps-data/data/tallies/tractblock.html.

Table 3.3 Various scales for data for North Carolina

Type	Number	Average size (mi^2)
County	100	486.19
Zip code (or *ZCTA*)	1048	46.39
Tract	2195	22.15
Block group	6155	7.90
Block	288,987	0.17

In what follows, we will compare several different solutions for finding the 1-median of *NCSE*. As mentioned above, the key difference in each solution is a different answer to how the demand is being calculated and what distance metric is being used. As for distance metrics, we will show the solutions for four different distance measures: $p = 1$ (l_1 or Manhattan metric, given in (4)), $p = 1.5$, $p = 2$ (l_2 or Euclidean metric given in (2)) and the squared Euclidean metric described in (3). As for estimating the demand points $i = 1,2,\ldots, N$, we will also compare using two different weights: the number of people living in each zip code[3] in the 2010 census of the population and how many employees were recorded working at businesses in each zip code in the 2012 economic census. Because codes often cross county boundaries, zip codes were selected if their centroids fell within the set of counties under study (Fig. 3.4).

Fig. 3.4 Different solutions to the 1-median problem for *NCSE*

[3] Or *ZCTA* (Zip Code Tabulation Area), an approximation of zip code boundaries using census data.

The results from the computation of the 1-median as given in (5) are shown in Fig. 3.4 and the endings *EMP* and *POP* are for points using employment and population as the demand weights, respectively. Here, l_1, $l_{1.5}$, and l_2 represent the distance metric used, and *MEAN* denotes the center of mass.

These results lend themselves to some interesting observations. First, note that all of the solutions result in locations in northern Bladen County, whose county seat is Elizabethtown with a population of around 4000 people.

Second, we see that all of the measures using employees as the weights are to the southeast of their counterparts weighted on population, which may be explained as follows. There are only two major cities in this region of 14 counties: Fayetteville/Fort Bragg, and Wilmington. Of the two, Fayetteville/Fort Bragg has a much higher population represented in its zip codes (256,233 vs. 186,140), but the number of employees counted in the economic census is about the same (78,155 vs. 77,164). Despite this, the observed shift to the southeast is likely to be due to the fact that while Fort Bragg and surrounding *ZIP* codes have a large population, relatively few of those employees, who are primarily employed in the military sector, would be measured by the economic census[4]. Therefore, one can expect areas with large military bases, universities, or other categories of employer not considered to be "industries" in the economic census are likely to see sizeable variation in results based on whether population or employees are chosen as weights.

A third interesting observation is that the centers of mass (labeled *MEANPOP* and *MEANEMP* in Fig. 3.4) are to the southeast of the 1-median obtained using l_1, $l_{1.5}$, and l_2 distance measures. The difference comes from the underlying objective of the two types of measures. While with the l_1, $l_{1.5}$, and l_2 we minimize the sum of the weighted distances, the center of mass minimizes the sum of squared distance. This means that points farther away will have a larger influence on the center of mass, because the square of a large distance becomes much larger than the sum of many medium-sized distances. Thus the distant lying sparsely populated regions of *NCSE* have tended to "pull" the center of mass in their direction.

Finally, note from Fig. 3.4 that all the solutions to the 1-median problem for l_1, $l_{1.5}$, and l_2 metrics in *NCSE* occur in an area that is largely agricultural land, and not in the middle of cities or towns. Also, several state highways (seen labeled with numbered ovals in Fig. 3.4) run nearby many of these locations, also with fairly convenient access to federal and interstate highways. This makes these sites well-suited for the location of a logistics park. Despite the same, Sect. 3.4 will make evident that none of the initial candidate sites selected in the actual implementation of the project were even proximate to the location indicated in Fig. 3.4. This discrepancy, in turn, is behind one of the key conclusion of this chapter that analytical results are, at best, only a small part of location modelling in practice. Other factors, be they social, economic or political, are usually much more important in influencing final outcomes.

[4] The economic census covers businesses of all sizes, but excludes most government-owned industries. Also excluded are schools (primary through university, including private schools), agriculture, and religious organizations (see http://www.census.gov/econ/census/help/naics_other_classification_systems/codes_not_covered.html for a complete list).

3.3 Location in Practice: Site Selection Process and Analysis

This section presents a synopsis of the actual approach taken in modelling the site selection process for a Logistics Park in *NCSE* and is based in large part on Bhadury and Troy (2011). The process followed may be structured formally within a framework that we introduce here and refer to by its acronym *SIRC*, which stands for the following steps that need to be executed in their prescribed order.

*Step 1 S*ituational analysis of the region where the logistics park will be located.

*Step 2 I*nitial selection of candidate sites for locating the logistics park in the region.

*Step 3 R*eadiness assessment for each of the candidate sites selected in Step 2 above. Such assessment should include current infrastructure, desired infrastructure for peak performance of the logistics park and the gap between the two.

*Step 4 C*ompetitive summary of the candidate sites, stating strengths and weaknesses of each. This step is obviously based on the data collected in the prior three steps.

While the *SIRC* framework is general enough to be used for any real-life location modeling problem, the remaining portion of this section is devoted to illustrating its implementation in the context of the current application.

3.3.1 Step 1 of SIRC: S*ituational Analysis*

The application of this first step of *SIRC* comprised of collecting social and economic information about *NCSE* that is relevant to a public works project such as the location of a logistics park. Such information comprised of population, employment, labor wages and their trends. Information was also gleaned about the industry clusters that were targeted for development by the regional economic development agency. Besides studying published reports, 14 interviews were held with 34 selected industry leaders, economic development officials, public officials and senior management from the existing logistical facilities such as Port of Wilmington, Wilmington International Airport, Fayetteville Airport, Laurinburg-Maxton Airport and International Logistics Park to name a few. The objective of these interviews was to collect background information and assess the current and anticipated logistical needs of the region and their views on the Logistics Park in *NCSE*. The results from this first step of Situational Assessment forms the basis of most of the information presented in Sect. 3.2 above.

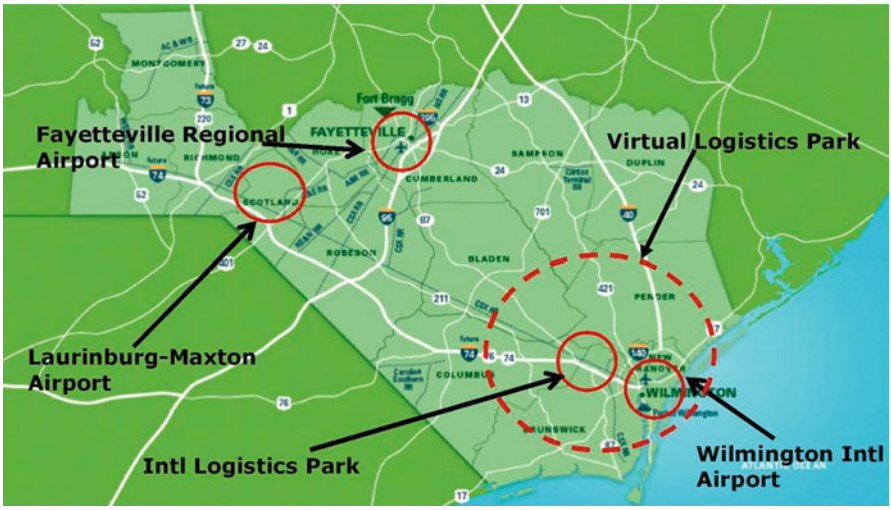


Fig. 3.5 Initial candidate sites

3.3.2 Step 2 of SIRC: I*nitial Selection of Candidate Sites*

The next step in the location modelling process is to select initial candidate sites for location. In this case, this was done on the basis of the data collected in the first step and an analysis of all available sites across the *NCSE*. The criteria used for the selection were sites proximate to highway, rail, and air access and having at least 400–500 acres of developable land to accommodate the needs of the companies that were expected to locate in the logistics park. Additionally, the opinions and ideas of the local economic and political leaders, as gleaned from the interviews, weighed heavily on the decision making process. Based on all the above, four sites were finally selected as candidate sites for locating a logistics park[5] and as depicted in Fig. 3.5, they were as follows: Laurinburg-Maxton Airport (*MEB*); Fayetteville Regional Airport (*FAY*); International Logistics Park of NC (*ILP*) and Wilmington International Airport (*ILM*). Additionally, as a result of numerous interviews and research conducted to evaluate the four designated sites, the research team added a fifth possible site, a "virtual logistics park" for the Greater Wilmington Area. A brief overview of each site is presented below.

[5] An analytical approach to the modelling would have selected all sites in *NCSE* that matched the criteria being used, consider this entire set as initial candidates for location and determine the optimal location of the logistics park by solving the for the 1-median of this resulting discrete location problem. Nonetheless, other factors (preferences of interviewees, need for economic development in economically depressed counties, need to promote sites that were already developed etc.) prevailed and contributed to only the above four sites from this set being selected as candidate sites in Step 2 of the *SIRC* process.

1. *Laurinburg-Maxton Airport (MEB):* The Laurinburg-Maxton Airport (see http://www.lmairport.com) is a small airport that is located in a rural region that borders Scotland and Robeson counties and is operated by the Laurinburg-Maxton Airport Commission that reports to the governing bodies of the cities of Laurinburg and Maxton. No commercial flights operate at *MEB* and its chief current use is for military training. Among the features of this small airport are a 6500-foot lighted runway, equipped with high-intensity lights, new *LED* taxiway edge lights, signage and pavement marking and a new and full *ILS* (instrument landing system). It also serves as the last landing site for many planes that are targeted to be decommissioned. *MEB* offers significant opportunity for development as a result of the availability of developable land, adequate and improving roads, and ready access to most infrastructural needs. In addition, as a direct result of its location, *MEB* can serve as a link to the Port of Charleston making it an attractive site as a distribution center for both North and South Carolina as well. However, the airport currently lacks a strategic plan and facilities need major costly improvement and while there are immediate plans and funds available to improve two of the three runways, the improvements do not fill all the needs and requirements to make *MEB* truly competitive. In addition, the area has access to national railroads operated by *CSX* but to take full advantage of these resources costly rail connections are needed. Finally, while most infrastructure is available at the *MEB* site, much of it needs substantial upgrades and retrofitting.
2. *Fayetteville Regional Airport (FAY):* Fayetteville Regional Airport (http://www.flyfay.com/) is located in Cumberland County and is operated by an airport director for the City of Fayetteville. The primary advantage of this as a possible site for a logistics park is proximity to potential demand given the industrial presence around Fayetteville and the growing military sector in Fort Bragg. The site is geographically well located with regards to US interstates. It has access to *CSX*, Norfolk Southern and short line railroads; while these connections are not on the airport site, they are all located within the City of Fayetteville. The area is also in close proximity to major interstate highways, the Port of Wilmington, Research Triangle Park, recreational facilities and developable land including shovel ready sites. In addition, and perhaps most significantly, it is located near an expanding military facility of Fort Bragg.
3. *International Logistics Park (ILP):* The International Logistics Park of North Carolina (http://brunswickedc.com/sites-buildings/available-sites/international-logistics-park-of-nc) is built on the Columbus/Brunswick County line on US 74/76 just 15 miles from the Port of Wilmington. This is important because the Port of Wilmington promotes the "At Port" site location model whereby facilities (e.g., warehouses, distribution centers) located within 20 miles of the port enjoy the same state, regional, local and port tax incentives as those located within the port premises and a lower rate from trucking companies. *ILP* is a joint venture between Brunswick and Columbus Counties, and has an undeveloped 1100 acre park that hopes to capitalize upon the "at port" site location promoted by the Port of Wilmington. The most significant attribute of *ILP* is the vast amount of developable land that includes shovel ready sites where utilities, gas, water electricity and sewer are readily available. Further, the existing roads and

planned road projects make *ILP* accessible, especially to the Port of Wilmington and Wilmington International Airport. Although *ILP* lacks a rail connection, located directly across the street is the site of another large industrial park, the Mid-Atlantic Logistics Center, that has *CSX* rail access.

4. *Wilmington International Airport (ILM)***:** The Wilmington International Airport (http://www.flyilm.com/) is located in New Hanover County and serves southeast North Carolina. The airport, located off I-40 and I-140, is operated by the New Hanover County Airport Authority. Although underutilized, it is a full-service airport that offers commercial, cargo and general aviation facilities and a state of the art Federal Inspection Services (*FIS*) facility. Runways are adequate and expandable to meet demand. Shovel ready industrial sites that meet the Port of Wilmington's "at port" location criteria are available at the airport and larger tracks of developable land are located nearby in Pender, Brunswick and Columbus Counties. The existing basic infrastructure (communications, water, sewer and power) at *ILM* may be classified as being of average quality. The airport also has at-site rail service with an indirect link to the Port of Wilmington. In addition, the high quality of life in the area provides the ability to attract highly skilled workers. However, a weakness for *ILM*, like the nearby International Logistics Park, is its risk exposure to the future viability of the Port of Wilmington.
5. *A "Virtual Logistics Park" for the Greater Wilmington Area:* The final assessment of the research team was an innovative idea that, to the best of knowledge of the authors, has never before been mentioned in the location modeling literature. This was to create a "virtual logistics park" from the Greater Wilmington area that comprises of all logistics assets of the region and is coordinated by a central organization to work synergistically as a unified and coherent institution. These assets include: Port of Wilmington, *ILM*, *ILP* discussed above but also two additional industrial parks in this region, namely, Mid-Atlantic Logistics Center and the Pender Commerce Park all of which are within 30 miles of each other. There are numerous reasons in support for this idea. First, geographical proximity puts all of these existing industrial parks within the requirement to be considered "at port" with regards to the Port of Wilmington. Second, the "virtual logistics park" collectively has all modes of transportation—shipping (Port of Wilmington); air freight (*ILM*), rail (*CSX* rail access at *ILM*, Port of Wilmington as well as Mid-Atlantic Logistics Center) and abundant trucking services. Additionally, developable land parcels of all sizes, many shovel-ready, are found in the metro area and/or *ILP* and the Greater Wilmington area has excellent utilities infrastructure (communications, water, sewer and power) to accommodate growth. Therefore, all that would be needed to create a "virtual logistics park" would be some minor infrastructure needs and for the area's political and economic leaders to work together and cooperatively in creating an umbrella organization to administer this park as one coherent entity. And it is this last condition that engenders the most important challenge involved with the "virtual logistics park," namely, the practical difficulties involved in initiating such an entity that brings all these disparate organizations under one umbrella and thereafter, getting them to work cooperatively towards common regional goals rather than individual ones.

3.3.3 *Step 3 of* SIRC: R*eadiness Assessment of Candidate Sites*

This step is the heart of the location modeling process espoused by the *SIRC* framework. As implied in the title, this step was devoted to conducting an in-depth assessment of each of the five candidate sites mentioned in Step 2 above and on the basis of the same, develop a comprehensive summary of the site's logistical assets, shortcomings and an estimate of what improvements would be needed to locate a regional logistics park at that site and time and cost estimates of the same. After reviewing existing practitioner literature and speaking to local economic developers, it became clear that for each site, the research team would have to focus on all infrastructure within a 10 and 60 miles radius since those would be the most heavily impacted areas if a logistics park was located at the site. In order to assess the local workforce available to work at companies expected to operate within the logistics park, existing commuter data for the state indicated that examining the workforce qualifications within a 30 min drive time from each site would provide the best information. Finally, it was important to study and assess the facilities available at the site itself. As an example of this portion of the location modeling, see Fig. 3.6 for an example of the 60 miles radius area examined for the *FAY* site. Also, Fig. 3.7 shows the 30 mile radius that comprises the "virtual logistics park."

After examining the areas mentioned above, three tables were produced for each site. The first table, labeled the "assessment matrix" presented an overall summary of the readiness of each site. The second table, labeled "site summary" presented detailed information on existing infrastructure such as roads, airports etc. at each site and the improvements to them that would be needed in order to locate a logistics park at the site. The third and final table was based on the first two and summarized the strengths, weaknesses and recommendations for each site. See Tables 3.4, 3.5 and 3.6 for examples of what was produced for the *FAY* site.

This was followed by a narrative explaining the basis for the information presented in the three tables. For example, the narrative accompanying the assessment of the *FAY* site was as follows (see Bhadury and Troy 2011):

"The Fayetteville Regional Airport (*FAY*) is located in Cumberland County and is operated by an Airport Director for the City of Fayetteville. Cumberland County is a Tier 2 county and has an unemployment rate of 9.2 % (November 2010) and is home to an increasingly important military base, Ft. Bragg. ... "Proximity" is the key word needed to understand the Fayetteville Regional Airport area and its potential as a 'logistics village.' Fayetteville is located half-way between Miami and Maine. It is on I-95 and 30 min away from I-40. Although not at the airport, the area has rail access on *CSX* and *NS* as well as a local short line railway. It is near the Port of Wilmington, Research Triangle Park and recreational facilities at the ocean and near the Pinehurst resorts. In addition, and perhaps most significantly, it is in very near the expanding Ft. Bragg. Airport facilities are in good condition with a 6500 ft. runway that could be expanded to 8500 ft. The airport has several parcels of land available for development and adjoining land in private hands could be developed as well. Furthermore, additional industrial land is available in the immediate area. Some of

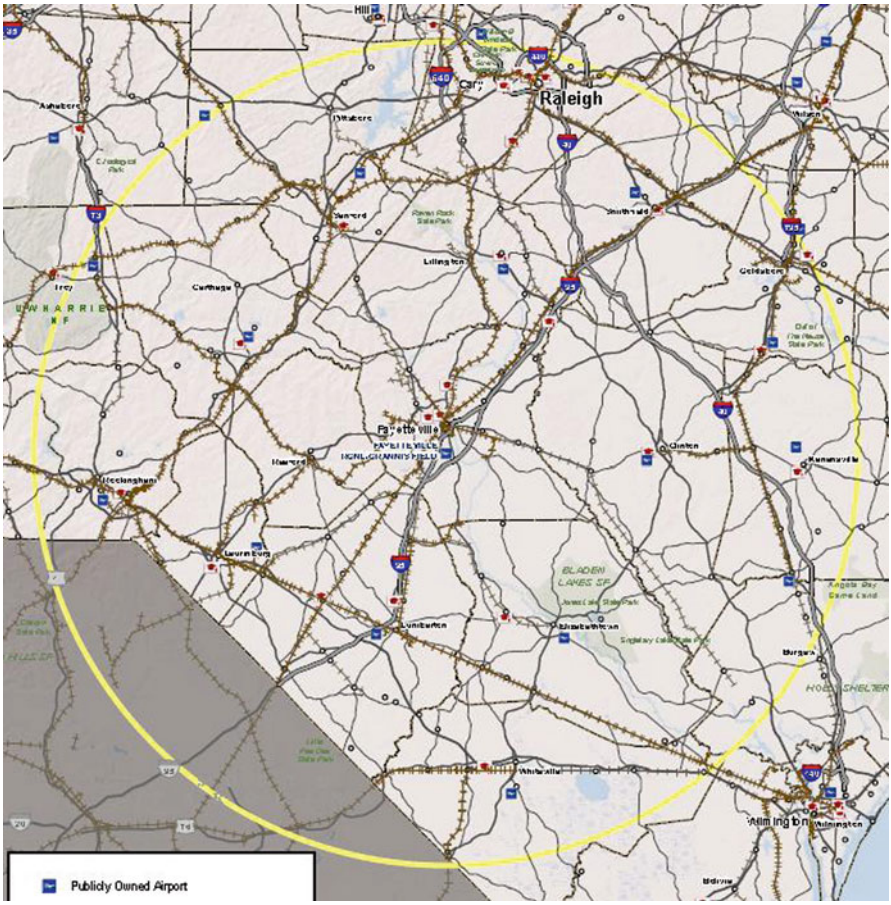


Fig. 3.6 60 mile radius for the *FAY* site. (Source: http://www.ncdot.gov/doh/preconstruct/tpb/research/download/2010-34-3southeastregionreport.pdf)

the land is 'shovel ready.' Access from the airport to I-95 is excellent. To reach its potential as a 'logistics village,' *FAY* would have to develop cargo facilities. It is currently considered a 'truck' market. Water and sewer issues at the airport would need to be addressed and should include a pumping station for sewer service. To help make some of this happen, the Airport Plan and the Fayetteville Area Plan should be coordinated to maximize the economic development benefits of the airport."

Given the novelty of the idea of creating a "virtual logistics park", see Table 3.7 and Table 3.8 for the assessment matrix and for the strengths, weaknesses and recommendations regarding the same.

Based on Table 3.7 and Table 3.8, the narrative assessment for the "virtual logistics park" was presented as below (Bhadury and Troy 2011).

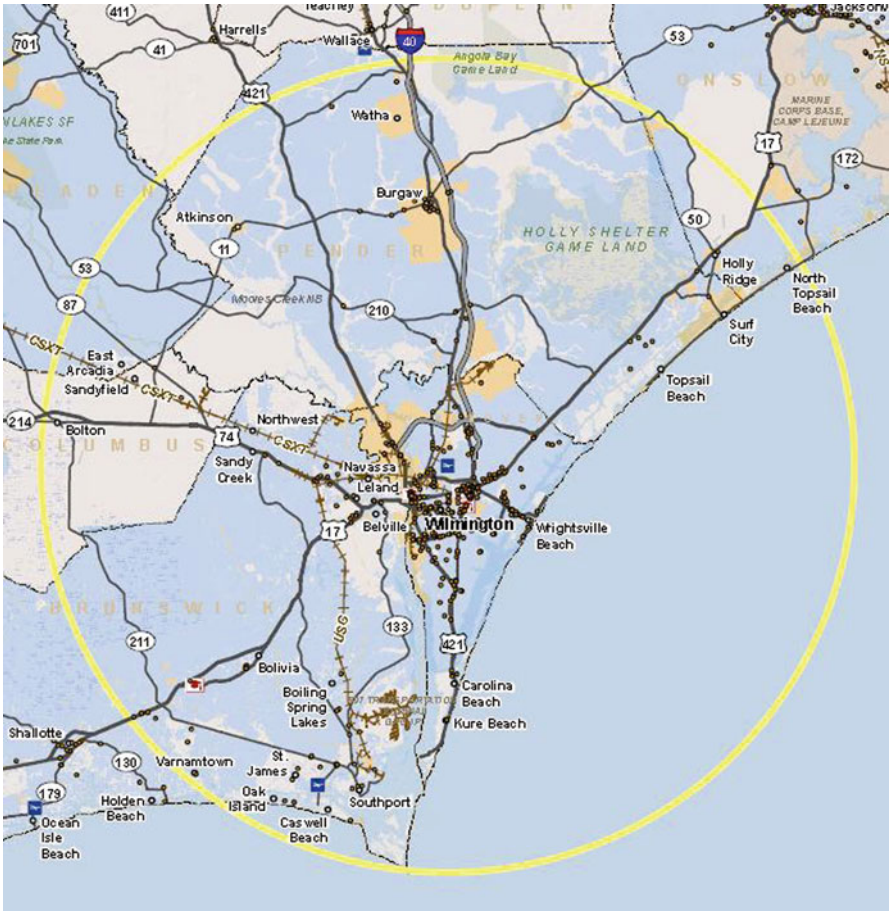


Fig. 3.7 30 mile radius for the "virtual logistics park". (Source: http://www.ncdot.gov/doh/preconstruct/tpb/research/download/2010-34-3southeastregionreport.pdf)

". . .the Greater Wilmington Area is strategically positioned to capitalize upon the area's logistics assets and establish a 'virtual logistics park.' A 'virtual logistics park' for the Greater Wilmington Area would not only incorporate two sites under consideration in this study, the International Logistics Park and Wilmington International Airport but also the Port of Wilmington, Pender Commerce Park and the Mid-Atlantic Logistics Center. In fact, the 'virtual logistics park' would include all of the logistics assets within a 30 mile radius of Wilmington International Airport. While all components of a 'virtual logistics park' are not located in one spot in the metro area, they are all in close enough proximity to each other to satisfy the 'at port' logistics model criteria for the Port of Wilmington. This is a significant factor that will serve as a strong foundation for developing logistics assets in the Greater Wilmington Area.

Table 3.4 Assessment matrix for *FAY* Site. (Source: http://www.ncdot.gov/doh/preconstruct/tpb/research/download/2010-34-3southeastregionreport.pdf)

Measure	Fayetteville regional airport (*FAY*) area
Facility type[a]	Commercial and general aviation
Target industries	Defense/homeland security; distribution; aerospace; general manufacturing
Speed of implementation	5–10 years
Infrastructure—transport	
Highway	Very good
Rail	Not at site but good connections in metro area
Air	
Existing runways	Two-only one *FAA* supported Lengths: 7712' & 4800' ILS Category 1 capability
Can runways be extended?	Yes, at a cost
Maritime (ship/barge)	92 miles to port of wilmington; fayetteville is accessible by barge via cape fear but not currently operational
Infrastructure—other	
communications	Very good
Power	Adequate
Water	Improvements needed
Sewer	Improvements needed-pumping station
Land available (acreage)	Some existing sites available at airport, improvements needed for other sites
Labor force	
Availability	Yes, including skilled labor
Education programs to support local industry	Yes
Specialty criteria[b]	Fayetteville economic development seeking *FTZ*; no cargo facilities

[a] *Facility* type major business practice at this location (intermodal facility, warehouse, distribution center, light manufacturing, and so forth)
[b] *Specialty criteria* foreign trade zone, customs, high-security, and so forth

The 'virtual logistics park' should be established, developed and coordinated by area leaders through the creation of a central 'facilitation' board. . . . The combined logistics assets of a 'virtual logistics park' far outweigh the value of Wilmington International Airport and the International Logistics Park when evaluated separately. . . . Most significantly, for the short-term, the 'virtual logistics park' is essentially ready to begin operation without additional major investments as the infrastructure already exists to support this effort. In fact, only two things are immediately needed to support the creation of a 'virtual logistics park,' the establishment of a coordinating board and a lighted intersection at the entrance to *ILP* and the Mid-Atlantic Logistics Park."

Table 3.5 Site summary for *FAY. (*Source: http://www.ncdot.gov/doh/preconstruct/tpb/research/download/2010-34-3southeastregionreport.pdf)

Facility	Current status	Improvements needed
Roads	Adequate plus	See roads recommendations
Airports	Central location, halfway between Maine and Miami Moderate commercial and general aviation facility (no cargo facility) One FAA supported runway which can only handle narrow body aircraft. Cash flow positive but limits growth (debt free)	Runway/taxiway upgrades needed (see water/sewer) Cargo facility required Additional sources for capital improvements
Railroads	Inadequate at *FAY* but good metro access; *FAY* 5–8 miles from main line	Spur required if cargo issue is resolved; costly
Ports	92 miles to port of Wilmington with adequate highway	
Shovel ready sites	Moderate airport sites available in NW corner with additional private tracts available	Water and sewer improvements needed for development of existing property
Workforce	Better than adequate skilled labor & technical labor available (military discharges around 6000 per year)	Needs to be promoted for recruitment of industry to area
Transport servicing	Trucking facilities are plentiful	
Communication	Good to very good	
Power	Adequate, owned by city	
Water/sewer Water/sewer (continued)	Adequate, presuming "no growth"	Pumping station at *FAY* for additional sewer capacity No sewer facilities on Southside of airport Taxiways need remediation for water problems Overall improvements needed for future growth
Specialty criteria	Applying for *FTZ*, location undetermined	

3.3.4 Step 4 of SIRC: C*omparative Summary of Candidate Sites*

This is the final step in the *SIRC* framework and is made necessary by the fact that in most real-life location problems, the decision makers do not ask for a single prescriptive recommendation, opting instead to be presented with alternate solutions with rationale for each so that they can make the final decision. Therefore, after completion of the individual assessments for each of the five candidate sites

Table 3.6 Strengths, weaknesses, and Recommendations for *FAY* Site. (Source: http://www.ncdot.gov/doh/preconstruct/tpb/research/download/2010-34-3southeastregionreport.pdf)

Strengths	Weaknesses	Recommendations
Proximity (highways, Ft. Bragg, recreation, port, *RTP*, etc.) Ft. Bragg expansion will stimulate airport growth (commercial/industrial, general aviation, and military related traffic) Military outsourcing leads to airport growth potential (hangers, runways, land sites) Better than adequate skilled and technical labor available (around 6000 military discharged per year) Roads are better than adequate in all directions	No cargo facility No at site rail connection Port 92 miles away Only one *FAA* supported runway that can handle narrow body aircraft Additional capital sources are not currently available No updated fayetteville area plan and airport plan not tied to economic development driven issues Water and sewer issues need to be addressed Limited number of large parcels of land for growth Expansion of runways will be expensive	See that city and economic development proactively recruit a champion, military is a good candidate to spur growth and provide dollars for logistics village at airport. Update Fayetteville area plan and coordinate with airport plan in order to promote economic development at *FAY* Correct water and sewer issues to maximize growth potential

in Sect. 3.4.3 above, a comparative summary was drawn up for all the five sites. This summary ensured that it did not recommend one site over another. Instead, it attempted to compare the relative strengths and weaknesses of each site so that the ultimate decisions makers (*NCDOT* and North Carolina State Legislature) could choose the final location for the logistics park.

The first part of the comparative summary was to enlist the common characteristics of all five sites studied. Some examples were as follows (Bhadury and Troy 2011):

- "The area shares an excellent network of existing roads including interstates, limited access highways and four-lane highways. Also, completion of planned road construction for improvement of the regional highway infrastructure will further enhance area transportation and provide a stronger foundation for a distribution and logistics based cluster for *NCSE*.
- While not excellent, the region generally has good access to railroads, *CSX*, Norfolk Southern and numerous short line connections. Improvements such as the establishment of an intermodal rail/truck facility in or near the *NCSE* region that has a direct rail link from the Port of Wilmington and the reestablishment of some preserved rail corridors could improve the attractiveness of the area to industry especially for supply chain based companies.
- The region has a strong agricultural and food processing base. Additionally, the growing military presence in the Fayetteville area and the Port of Wilmington are regional assets and are critical components for economic development in *NCSE*."

Thereafter, an added narrative was devoted to comparing the site specific resources and infrastructure needs. Specific topics addressed were: highway infrastructure,

Table 3.7 Assessment matrix for the "virtual logistics park" for the greater Wilmington area. (Source: http://www.ncdot.gov/doh/preconstruct/tpb/research/download/2010-34-3southeastregionreport.pdf)

Measure	Virtual logistics park
Facility type[a]	Airport, trucking, rail, port, logistics and distribution, warehousing center, and manufacturing
Target industries	See *ILM*, *ILP* and *NCSEs* targeted clusters
Speed of implementation	Immediate
Infrastructure—transport	
Highway	Good to very good and improving (see "cape fear commutes 2035: transportation plan")
Rail	Available with limitations
Air	*ILM*
Existing runways	Two at *ILM* Lengths: 8000' & 7004' ILS category 1 capability
Can runways be extended?	Yes, at a cost
Maritime (ship/barge)	Port of Wilmington
Infrastructure—other	
Communications	Good to excellent
Power	Adequate
Water	Adequate, but secondary water source needed. Overall area has future water concerns
Sewer	Septic and traditional available
Land available (acreage)	Excellent- numerous shovel ready of all sizes
Labor force	
Availability	Overall excellent- some skills training may be needed
Education programs to support local industry	Yes
Specialty criteria[b]	ILM has full-service customs *FIS* facility; area meets "at port" logistics model; inactive *FTZ* at port could be reactivated and sub-zones established as needed at other "virtual logistics village" sites

[a] *Facility type* major business practice at this location (intermodal facility, warehouse, distribution center, light manufacturing, and so forth)
[b] *Specialty Criteria* foreign trade zone, customs, high-security, and so forth

railroads, air connections, utilities (including broadband), availability of developable land, labor force availability and specialty criteria (such as the need to establish free trade zones at sites other than the Port of Wilmington).

Two examples of such narratives from the report (Bhadury and Troy 2011) are presented below. The first is for the highway infrastructure at these sites and the second focusses on the amount of developable land available.

Table 3.8 Strengths, weaknesses, and recommendations: "virtual logistics park" for the greater Wilmington area. (Source: http://www.ncdot.gov/doh/preconstruct/tpb/research/download/2010-34-3southeastregionreport.pdf)

	Strengths	Weaknesses	Recommendations
"Virtual logistics village" for the greater Wilmington area	All transportation modes Adequate infrastructure Shovel ready sites International trade (Port of Wilmington) Major industrial parks including: international logistics park, Mid-Atlantic logistics center, pender commerce park	Absence of strong coordination Highway and rail access ("the last mile") Risk exposure to future viability of the Port of Wilmington	Enhance coordination with a "facilitation" group (e.g., similar to Aerotropolis Leadership Board for PTI) Capitalize on "At Port" logistics model Improve highway and rail access Continue to invest in maritime activities

"Highways: In general, all five sites have good road access and planned improvements will only make access easier. When I-74 is complete to the Piedmont Triad and beyond, *NCSE* access to the Midwest will improve as truck traffic will be able to bypass the congested Raleigh area. This is especially true for the International Logistics Park (*ILP*), the Wilmington International Airport (*ILM),* the 'virtual logistics village' and the Laurinburg-Maxton Airport (*MEB*) sites. For these same sites, The Monroe Bypass, along with the completion of I-74 to Rockingham, and completion of I-140 in the Wilmington area will provide better access to the Charlotte market and beyond. Key to the near term success of the *ILP* is a full lighted intersection at the entrance to the Mid-Atlantic Logistics Center and *ILP*. In the future the state should also consider building a full interchange at the intersection of Highway 87 and US 74 that would provide links via access roads to not only *ILP* but the Mid-Atlantic Logistics Center and allow the removal of the interim lighted intersection. While the Port of Wilmington by itself is not a targeted site, it is part of the 'virtual logistics village' and this site will benefit significantly from all road projects and will help the Port to implement its 'at-port' logistics model and become more globally competitive. Planned road projects for the Fayetteville area will slightly enhance traffic around the Fayetteville Regional Airport (*FAY*); however, overall, *FAY* already has excellent immediate access to I-95 and access to I-40 is just 30 min away via I-95."

"Developable land: each of the five possible sites studied for a 'logistics village' has land available for development. Similar to the Global TransPark, *MEB* has land available to meet almost any need. Some of that land is already shovel ready and located in a certified industrial site fronting the future I-74. The *ILP* has 1100 acres available and, immediately across the street from *ILP*, the Mid-Atlantic Logistics Center has an additional 1100 acres with rail access. *FAY* has some parcels immediately available but infrastructural improvements would be needed to build on other existing parcels. Additional land is also available near the airport. *ILM* has around 150 acres available at site and other possible sites adjoin the airport. If the 'virtual

Table 3.9 Comparative summary for all candidate sites

Site	Strengths	Weaknesses	Needs
Laurinburg-Maxton airport (*MEB*)	Significant capacity for expansion Proximity to Ft. bragg and camp McCall	Facilities require significant costly improvements Lack of strategic plan for airport	Needs champion (private/government)
Fayetteville airport (*FAY*)	"Proximity" Ft. bragg expansion Adequate labor available	Lack of capital sources for infrastructure improvement Disconnect between airport plan and economic development	Proactively recruit a champion (military?) Update Fayetteville area plan and coordinate with airport plan Correct water and sewer issues
International Logistics Park (*ILP*)	Abundance of shovel ready sites Meets "at port" criteria Tier 1 (economically disadvantaged) county status makes it eligible for tax incentive support from state and local governments	Risk exposure to future viability of the Port of Wilmington	Planned road projects (full intersection) Cooperate with Mid-Atlantic logistics center Leverage regional logistics assets
Wilmington international airport (*ILM*)	Modern full-service (*FIS*) airport Shovel ready industr*ial* sites Quality of life	Risk exposure to future viability of Port of Wilmington	Leverage regional logistics assets Utilize professional skills of local retirees
"Virtual logistics park" in the greater Wilmington area	All the strengths of regional logistics assets including *ILP* and *ILM* above	Risk exposure to future viability of Port of Wilmington	Form a coordinating board that can leverage regional logistics assets in creating this "virtual" organization

logistics village' is considered significant, developable land is available at *ILP*, the Mid-Atlantic Logistics Center, and Pender Commerce Park and at or adjacent to *ILM*. A large percentage of this land is shovel ready and meets the criteria for the Port of Wilmington's 'at port' logistics model."

The complete comparative summary is available in Bhadury and Troy (2011); key points from that are captured in Table 3.9.

3.4 Epilogue and Conclusions

This final and most important section of the case study will first describe the epilogue of events after the completion of the Seven Portals Study. Thereafter, conclusions are presented on the lessons learned from the study as well as its implications for future research in location theory.

After the completion of the research, the key findings were presented in early 2011 to Governor's Logistics Task Force (*GLTF*) in the form of the final report Bhadury and Troy (2011). That report became a part of the final report submitted by the lead research team of the Seven Portals Study project, see List and Foyle (2011). In turn, List and Foyle (2011) became an integral part of the report submitted by *GLTF*, namely, the *GLTF* Final Report (2012). However, in 2012, all recommendations from prior studies undertaken by *NCDOT*, including the Seven Portals Study, were placed on hold, and *NCDOT* announced that it would substantially change the way transportation infrastructure in North Carolina was financed. Subsequently, *NCDOT* announced a new method for evaluating infrastructure projects such as the ones recommended in the Seven Portals Study. This new method is referred to as *Strategic Transportation Investments* and a description of the same is available at http://www.ncdot.gov/strategictransportationinvestments/. As of the end of 2014, no logistics park had been located by *NCDOT* anywhere in the state directly as a result of the Seven Portals Study. Nonetheless, smaller recommendations from the study were implemented; for example, cold storage facilities were added at the Port of Wilmington in order to facilitate exports of agricultural products.

There are numerous lessons to be learned by location theorists from an actual application such as the one described in this chapter. Three key lessons are as follows:

1. *Most Problems Involve Location of a Single Facility:* In most cases in real life, site location problems consider the location of only one facility. In case multiple facilities are involved, the decision-makers usually divide the area to be served into smaller regions and focus primarily on locating one facility in each of them, much like the original Seven Portals Study was broken up into smaller, regional ones, each involving the location of one logistics park in a specific region of the state.
2. *Most Problems Have Multiple, often Conflicting, Objectives:* In real life location problems, decision-makers have multiple objectives, some of which are not even quantifiable. This makes it almost impossible to determine an optimal location. In fact, the fuzziness of some of the objectives, as for example, the objective to stimulate economic development through location of logistics parks in the case of the Seven Portals Study, make it questionable if an optimal location can even be defined that is acceptable to all decision makers. For example, whereas minimizing travel distance might attract the location of a facility towards the demand centers (usually high population urban centers), governments locating public facilities also have a tendency to consider sites in high unemployment areas (usually sparsely populated rural areas) so as to stimulate economic development. Therefore it is important that in the analysis, researchers stick to evaluating strengths and weaknesses of candidate sites with regards to the various objectives of the decision-makers rather than being prescriptive and recommending one particular site.

3. *Rigorous Scientific Analysis is a Small Part of The Modeling Process*[6]: When locating public facilities, especially large and expensive facilities such as logistics parks, theoretical analysis such as the one presented in Sect. 3.3, is at best a small part of the actual location modeling process. In fact, as is evident from this chapter, the 1-median locations determined by the analytical process of Sect. 3.3 and as presented in Fig. 3.4 are not even proximate to any of the initial candidate sites selected by the research team (Fig. 3.5) on the basis of the Situational Analysis step of *SIRC*. This exemplifies that site selection of public facilities in practice mostly involves factors other than analytical ones. Public facilities such as logistics parks are viewed by the citizenry as job creators as well as a nuisance (causing congestion, pollution etc.). As a result, public perception, as reflected directly by the citizenry, as well as through their elected leaders and/or special interest groups and the clout that these have in the decision-making process have a far larger bearing on the modeling process in practice than does the mathematical analysis presented in Sect. 3.3.

The import of lessons 2 and 3 is an important guideline for location theorists who are called upon to perform location modelling of public facilities. The guideline is that in Step 1 of the *SIRC* framework (namely, situational analysis), researchers must first identify all stakeholders involved and make sure that all necessary background information is collected about them. Such background research should involve a study of the socio-economic profile of the region as well as their future trends, as we presented in Sect. 3.2. In addition to this in-depth study of the region being considered for location, it is also important for researchers to get to know the important social, economic and political figures in the region and interview them to find out their expectations from the facility being located as well as the process that ought to be used in determining the location. That is why the research team conducted 14 interviews with 34 different people in *NCSE* as a part of completing the Situational Analysis for this project. Not only does this create the information that forms the bedrock of the subsequent steps (Steps 2–4 of the *SIRC* framework), but also lends credibility to the work being done which, in turn, lays the groundwork for the eventual acceptance by stakeholders of the findings and conclusions of the research project.

These lessons presented above also lead to two important recommendations about future research in location theory and we conclude this case study with the same. The first of these points to the importance of single-facility location models. While multi-facility location models are theoretically much richer than single facility location models, more research needs to be focused on the latter. The primary source of complexity in real life location models comes not from having to locate multiple facilities but from having to locate one with numerous, often conflicting and non-quantifiable, objectives. Thus, it is our recommendation that more work

[6] A member of the research team succinctly stated this principle as " *Location in practice is 80 % politics and 20 % science*".

needs to be done on multi-objective, single facility location models where some of the objectives are allowed to be fuzzy and/or qualitative. By nature, such models will need to combine quantitative analysis with qualitative research and should add to the richness of extant literature in location theory.

The second observation about future academic research stresses the importance of exploring alternate optimal and near-optimal solutions in location models. As mentioned above under lessons learned, it is difficult to define "optimality" with a high degree of precision for most real life location problems. As a result, decision-makers look not for a single prescriptive recommendation but a range of alternate recommendations that do "well enough" on most objectives. In the language of optimization, this implies that research in location theory should focus on efficient ways to generate alternate optimal locations as well as near-optimal solutions to complex, multi-objective problems. The goal of location modeling in practice is to then present the best of these alternate and near-optimal solutions to the decision-makers along with strengths and weaknesses of each to enable them to make final decisions.

Acknowledgment This work draws, in part, upon the experiences of the first and the third author in the implementation of the Seven Portals Study project. That project was supported by *NCDOT*, which is gratefully acknowledged.

References

Bhadury J, Troy SP (2011) Seven portals study—an investigation of economic development in North Carolina through logistics villages (SouthEast Region Report). http://www.ncdot.gov/doh/preconstruct/tpb/research/download/2010-34-3southeastregionreport.pdf. Accessed 10 April 2015

Brimberg J, Dowling PD, Love RF (1994) The weighted one-two norm distance model: empirical validation and confidence interval estimation. Location Science, 2:91–100

Brimberg J, Love RF (1995) Estimating distances. In: Drezner Z (ed) Facility location. Springer, New York, pp 9–31

Burkey ML, Bhadury J, Eiselt HA (2011) Voronoi diagrams and their uses. In: Eiselt HA, Marianov MV (eds) Foundations of location analysis. Springer, New York, pp 445–470

Current J, Daskin M, Schilling D (2004) Discrete network location models. In: Drezner Z, Hamacher HW (eds) Facility locations: applications and theory. Springer, Berlin, pp 81–118

Eiselt HA (1992) Location modeling in practice. Am J Math Manage Sci 12(1):3–18

Eiselt HA, Marianov V (2011) Foundations of location analysis. (Vol. 155). Springer Science & Business Media, Berlin

El Amrani A (2007) The impact of international logistics parks on global supply chains. (Doctoral dissertation, Massachusetts Institute of Technology)

Fernández J, Fernández P, Pelegrin B (2002) Estimating actual distances by norm functions: comparison between the k, p, θ-norm and the *i*b1, b2, θ-norm and a study about the selection of the data set. Comput Oper Res 29:609–623

Findley DJ, Small JD, Tran W, Heller A, Bert SA, Searcy SE, Hall WW (2014) Economic contribution of the North Carolina ports. http://www.ncports.com/elements/media/files/economic-contribution-north-carolina-ports.pdf. Accessed 10 April 2015

Francis RL, Lowe TJ, Rayco MB, Tamir A (2009) Aggregation error for location models: survey and analysis. Ann Oper Res 167(1):171–208

Gardner, Kansas Official Website (2015) What is a logistic park. http://www.gardnerkansas.gov/images/uploads/Administration/Intermodal/WhatisaLogisticsPark.pdf. Accessed 10 April 2015

Governor's Logistics Task Force, Final Report (2012) http://www.ncdot.gov/download/business/committees/logistics/GovernorsReport_Jun2012.pdf. Accessed 10 April 2015

Hamacher HW, Drezner Z (eds) (2002) Facility location: applications and theory. Springer Science & Business Media, Berlin

Lee H, Yang HM (2003) Strategies for a global logistics and economic hub: Incheon international airport. J Air Transp Manage 9(2):113–121

Lee KL, Huang WC, Kuo MS, Lin SC (2001) Competitiveness model of international distri-park using the virtual value chain analysis. J East Asia Soc for Transp Stud 4(4):313–325

List GF, Foyle RS (2011) Seven portals study: an investigation of how economic development can be encouraged in North Carolina through infrastructure investment. https://apps.dot.state.nc.us/Projects/Research/ProjectInfo.aspx?ID=2761 Accessed 10 April 2015

List GF, Foyle RS, Canipe H, Cameron J, Stromberg E (2008) Statewide logistics plan for North Carolina. http://www.ncdot.gov/download/business/committees/logistics/StatewideLogisticsPlan_080513.pdf. Accessed 10 April 2015

North Carolina's Southeast regional economic development partnership (2013). www.ncse.org. Accessed 10 April 2015

North Carolina State Ports Authority (2015) Port of Wilmington. http://www.ncports.com/elements/media/files/port-wilmington-fact-sheet.pdf. Accessed 10 April 2015

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

Qabaja H, Bikdash M (2014) Identification of closest safe places and exit routes during evacuation from GIS Data. ASE@360 Open Scientific Digital Library. http://ase360.org/handle/123456789/141. Accessed 10 April 2015

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

Rushton G (1988) The roepke lecture in economic geography location theory, location-allocation models, and service development planning in the third world. Econ Geogr 64(2):97–120

Weiszfeld E (1937) Sur le point pour lequel la somme des distances de n points donnés est minimum. Tohoku Mathematical J 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(1):7–41

Chapter 4
An Introduction to Industrial Forestry from a Location Perspective

Eldon Gunn

4.1 Introduction

Forestry is the natural resource industry most dominated by spatial considerations. Forests are an extensive resource typically spread out over millions of hectares. In this paper we focus on the industrial aspects of forestry, and some location issues that flow from these. This is in contrast to a focus on location analysis and then asking about its applications in forestry.

We are conscious of the distinction made by Revelle and Eiselt (2005) that location analysis involves the siting of facilities in a given space with the facilities being "small relative to the space in which they are sited and interaction between the facilities may or may not occur," while in the case of layout problems, the facilities are "fairly large relative to the space in which they are positioned, and interaction between facilities is the norm." In a strict sense, some readers may feel that this paper does not deal with location problems since forestry problems that we discuss do not necessarily fit neatly into this framework. In some of the cases discussed, such as the location of processing facilities, the fit within this framework is obvious, but in other cases, the facilities are not always obvious. For example, forested stands where harvest occur or harvest regions which serve as a source of wood flow, can be seen as facilities with the siting seen as the choice of stands to harvest in a given period. These are location problems in the sense that the facilities are a small proportion of the total area in any one period and the facilities do not interact in terms of the flow of material from one to another. The flows from the "harvest facilities" occur to the various customer plants that use the forest products produced from harvesting. Typically the metrics are total profits for delivering wood products.

E. Gunn (✉)
Department of Industrial Engineering, Dalhousie University,
Halifax, B3H 4R2, NS, Canada
e-mail: eldon.gunn@dal.ca

H. A. Eiselt, V. Marianov (eds.), *Applications of Location Analysis,* International Series in Operations Research & Management Science 232, DOI 10.1007/978-3-319-20282-2_4

In this review, we propose to touch on at six interrelated areas where location issues. These include:

1. Long term location of harvest by species, log type and age class in response to strategic decisions on ecosystem management
2. Location of processing facilities relative to harvest
3. Joint location of harvest and facilities
4. Shorter term location of harvest w.r.t adjacency and road building
5. Shorter term location of facilities (camps)
6. Operational location of harvest equipment

Our goal is to illustrate the hierarchy of inter-related location problems and indicate some of the accomplishments in these areas. These problems can be seen can be seen as arising from the long-term, medium term, and short-term, design of the forest products supply chain. However, it is worth emphasizing that this is more than a supply chain, it is what many of us referred to as a value chain, in that, many different value streams emanate from the forest.

4.2 Some Terms

Since some of the terms may not be obvious to the non-specialist, we begin with a few terms that make up the spatial issues of the problem.

Stand A relatively small area over which the forest can be considered as homogenous with respect to features such as species distribution, age, site quality, and costs of harvest . In a *GIS*, a stand is often equated to a polygon.

Ecodistrict Part of the larger ecosystem classification of the landscape into ecozones, ecoregions, ecodistricts and ecosites. An ecodistrict is a connected landbase that has similar landform, soil and vegetation. In my home province of NS, the 4.5 million ha of forested land are divided into 38 ecodistricts.

Watersheds An area of land that drains to a common river mouth. Here we are thinking of fairly significant river systems. Again in Nova Scotia, the land base can be partitioned into 46 watersheds[1].

Timbershed A region from which logs are harvested. The purpose of a timbershed is to provide a location that can be used as a point to estimate transportation costs (see Gunn 2009).

Sawmills Mills that process round logs into semi-finished lumber. The main feature here is that the process of sawing produces multiple lumber products as well as chips, sawdust, shavings and bark. Typically the lumber will be less than 45 %

[1] https://www.novascotia.ca/nse/water.strategy/docs/WaterStrategy_NSWatershedMap.pdf.

of the input volume, with chips as high as 40 %. The percentages are dependent on the lumber products being produced, and on the size of logs being processed. Products are distinguished by species (fir, spruce, pine, birch, maple, oak). Mills are often classified as softwood and hardwood mills depending on the species most commonly sawn.

Pulpmills/biorefineries Here we mean a more general concept of mills that are aimed at processing the fibres in the wood as opposed to solid wood. Some aim at thermo-mechanical processing where heat and grinding reduced the wood to a pulp. Others deal with the wood chemistry and attempt to break out the cellulose, hemi-cellulose and lignin components of the wood. Typical is the kraft process with a primary aim at producing cellulose but more recently new processes are aimed at high end chemical processing of the lignin and hemi-cellulose components. These have come to be referred to as biorefineries. Pulpmills/biorefineries are usually based on wood chips from sawmills and other sources. If they receive round wood, the bark byproduct typically is for energy generation either onsite or elsewhere.

Energy Facilities Energy facilities may be relatively small scale, typically heat or combined heat and power. Energy facilities may be located concurrently with sawmills, pulpmills, or they may be located independently in large consumers, such as hospitals or universities. Energy facilities consume woodchips, bark, shavings. In some cases, of standalone large-scale energy facilities, the inputs may need to be preprocessed in the form of pellets, either green or torrified.

4.3 The Sustainable Forest Management Setting

Although this paper is focused on industrial force management, one also has to be cognizant of the fact that all forestry operates under commitments to sustainable forest management. Internationally, one has the Montréal Process (2014), which has been signed by more than 20 countries. In Canada, we operate under Canadian Council of Forest Ministers (2003) Criteria and Indicators (*CCFM C&I*) for sustainable forest management. The main criteria under the *CCFM* are the following: (1) Biodiversity, (2) Ecosystem condition and productivity, (3) Soil and water, (4) Global cycles, (5) Multiple economic benefits, and (6) Societies responsibility. Biodiversity is often discussed in terms of preservation of habitats for endangered species, location of old forest conditions, but also, importantly, in terms of distributions of representative habitats across the landscape. Because ecosystems are represented as distinct spatial entities, the representativeness of species and age distributions on an ecodistrict is one important measure of ecosystem condition. Watersheds, by their very nature, depend of the location of major lakes and rivers. The last two criteria (5–6) involve supply to industries and affects on communities, both aboriginal communities and small rural communities and again require

explicit recognition of relative location of the forest activities and the location of both industrial sites and communities. Perhaps only the global cycles criterion (4), with its main focus on the carbon cycle can be thought as aspatial. Yet, even here the location of fire, insect out-breaks, and forest harvesting play significant roles and the net contribution of the green house gases will depend on transportation and the relative location of harvesting processes, and wood consuming facilities.

In most of the situations discussed this paper, we will focus on the economic issues, but, the criteria for sustainable forest management, will always be lurking in the background. It is true that there are opportunities for direct locational analysis in sustainable forest management. This often arises in issues of location of protected and special management areas and are dealt with elsewhere in this book (Chap. 6). In looking at industrial forest management, we will often take the position that some agency has adopted certain strategies to deal with the criteria and indicators of sustainable forest management and that these strategies impose locational constraints on what we can do logistically.

4.4 Strategic, Tactical and Operational Planning

Within forestry there has long been a focus on distinguishing to strategic, tactical, and operational management problems. We discuss these briefly below. For a good example of the forestry planning hierarchy from an information perspective see Marques (2011).

Strategic level decisions are those that affect the resources available to the organization (Anthony 1965). Here we will discuss three closely related decision problems. The first is the problem of how much to harvest, where, and when, to make the most effective use of the forest's growth capabilities. Often, such decisions have been made with no regard to the location to which the harvest of material will be sent. More recently, models which take account the location and capacity of the wood using facilities are being discussed. The second problem involves the design of the facilities making up the overall supply chain. This problem has usually focused on the location of facilities, particularly intermediate processing facilities, and the flow of materials between them. What has been missing is a focus on the appropriate capacity and location of the larger scale facilities. This is important because large-scale facilities such as pulpmills are subject to enormous economies of scale in terms of both capital and labor. However, there are also diseconomies in terms of transportation and materials acquisition cost both from the forest directly and from intermediate facilities. The third problem involves the joint design of both forest harvesting and the facilities making up the forest supply chain. This recognizes the dual facts that appropriate location of the right types of facilities can facilitate much improved forest harvesting strategies, and that the appropriate forest harvesting strategies can significantly improve the economics of investment in new capital facilities.

Tactical level problems involve how to best exploit the resources provided through strategic decisions. In most organizations, tactical level problems typically have something like 1–5 year time horizons. Forestry has been unusual in that its tactical problems have often been posed over time horizon's ranging up to 20 years. Typically these problems involve location of harvest and the decisions on road building to provide access to the harvest. The dominant feature in these decision problems involves issues of harvest of stands and their relative locations to each other and to the road network. One of the problems in this area that has attracted significant research attention is the adjacency problem in which we try to prevent harvesting too many adjacent stands that will create overly large openings in the forest.

Operational problems are those that occur at a relatively short time frame and typically involve minimization. We will focus only on the operational problems that occur within the forest since these tend to have the most location oriented aspects. These include problems of location of landings and equipment, problems of location of skid roads, and problems of crew scheduling and movements between harvest locations. A problem that has recently come up that involves the location and sizing of camps for the forestry work crews. These problems arise in the northern regions of Canada where sparse settlement and long distances may make staying in regional towns prohibitive.

4.5 Previous Reviews

This review is not meant to be comprehensive. Thus the reader should refer to Church et al. (1998) for additional material particularly from a location perspective. Bare et al. (1984) is an early review focussing on location is Gibson and Rodenburg (1975). Martell et al. 1998 updates this work with an emphasis on the complex hierarchy of forestry decisions. Baskent and Keles (2005) gives a focus on spatial forest planning. Melo et al. (2009) reviews facility location issues associated with supply chains. Carlsson et al. (2009) review the issues of of supply chain planning in the pulp and paper industries. Chauhan et al. (2009) look at the planning of the multi-commodity supply network where many products emanate from the forest. D'Amours et al. (2008) reviews operational research applications to supply chain planning in the forest industry. For the serious reader, Davis et al. (2001) is a fundamental work on forest management issues.

4.6 Strategic Harvest Planning

The first of our location models we will examine is the strategic harvest planning model (see Gunn 2007, 2009 for a longer discussion). This is the problem of deciding where and how much to harvest from various forest types over time subject to

requirements of a reasonably constant supply of wood. It arose from an initial concern (Ware and Clutter 1971) that if forest stands were harvested at their time of peak output (economic or biological) then this would result in a very irregular supply that could not be used by the industry. This is the longest standing forest location problem, but it had its origin largely in an aspatial context. As will be evident below, this class of models are not strategic in the sense of computing strategy. The sustainable forest management strategic decisions define both the constraints and variables of the problem and are given a priori. Thus these are models for assessment of strategy, not choosing strategy.

The early paper by Ware and Clutter (1971) did focus on "cutting units" with revenues based upon "net stumpage values." This does imply a location, but the main emphasis has been on defining cutting units as homogenous units in terms of future growth capabilities. When these types of models were used early on, there was a tendency to combine stands into "timber classes" (Navon 1971) or "macro-stands" with a corresponding loss of location identity. The initial modelling (Model I) was based around the concept of assigning a cutting unit to a prescription where a prescription was an entire plan for harvesting, planting, thinning, harvesting, etc over the entire planning horizon. With many cutting units and many possible prescriptions, models rapidly grow. This was one reason for the amalgamation of stands with similar growth characteristics into a much smaller number of cutting units with a subsequent loss of identity. Commercial software that came into widespread usage in the 1990s was based on the Model II ideas of Johnson and Scheurman (1977) which automatically generates prescriptions through an acyclic network model. The nodes of the network correspond to points at which the forest is regenerated and the arcs correspond to decisions of what is to be done between regeneration points. Stand identity is lost as all stands that regenerate at a point in time merge. These models (both Model I and Model II) came to be widely used to develop the concept of "annual allowable cut" without much thought about where this cut was coming from or where it was going (Church 2007; Gunn 2007, 2009). These models have been widely used is a variety of government and industrial analyses.

In Gunn (2009) we formulated a model that has a more spatial orientation than the conventional. This model was developed into a case study in Martin (2013) Thus this model is not typical of those in common use in government and industry. To see the type of model we have in mind, we use the following notation, taken from Gunn (2009):
Sets:

I:	the set of stands
P_i:	the set of allowable prescriptions for stand i
Y:	the set of yields under strategic consideration
Y^w:	the set of wood products ($Y^w \subset Y$)
$R(\ell)$:	the set of regions applicable for yield type$\cdot\ \ell$. If $\ell \in Y^w$, then $R(\ell)$ is a timbershed for wood product ℓ
M:	the set of mills (forest products demand centres)
$S(m)$:	the set of demand segments for mill m

Data:

A_i: area of stand i
$LY_{\ell rt}$: lower limit on yield ℓ in region r in period t
$UY_{\ell rt}$: upper limit on yield ℓ in region r in period t
LD_{mt}: lower limit on demand in mill m in period t
$d_{smt}^{\max}$: upper limit on demand segment s for mill m in period t
y_{iklrt}: yield of type ℓ produced in region r in period t if prescription k is used for stand i
$\partial_{\ell m}$: conversion coefficient of log type ℓ to meeting the demands at mill m
p_{smt}: demand curve price for segment s at mill m in period t (discounted to time 0)
C_{ik}: silvicultural costs for managing stand i with prescription k (all costs discounted back to time 0)
$\tau_{\ell rmt}$: unit transportation costs for logs of type ℓ from region r to mill m in period t (discounted to time 0)

Variables:

X_{ik}: area of stand i managed using prescription k
$Y_{\ell rt}$: yield of type ℓ produced in region r in period t
$z_{\ell rmt}$:: transportation of logs of type ℓ from region r to mill m in period t
D_{smt}: amount of demand segment s supplied to mill m in period t

Then D_{smt} allow us to define demand curves at each mill m using the price endogenous ideas described in Lebow et al. (2003). Since the model is aimed at strategic assessment, it is worth noting that these demand curves here also need to be defined in strategic terms. The main objective is not to predict pricing, but rather to link supply chain mill capacity to the forest strategy. This requires an ability to represent the fact that some forest products may command premium prices p_{smt} and the extent of the market assumed to exist for these products. There may be limited ability to use certain log types (species, size) at high prices and the representation of the low price section of the demand curve can useful to represent disposal. The mill subscript m allows representation of both product type and location. In a strategic model we might expect these to be broadly defined, with several facilities that require the same type of wood product and located in the same general region aggregated into a single mill type.

$$Max/Min \sum_{t}\sum_{m}\sum_{s\in S(m)} p_{smt}D_{smt} - \sum_{i=1}^{I}\sum_{k=1}^{P_i} c_{ik}X_{ik} - \sum_{t}\sum_{m}\sum_{\ell}\sum_{r} \tau_{\ell rmt}z_{\ell rmt} \tag{4.1}$$

$$\text{s.t.} \sum_{k=1}^{P_i} X_{ik} = A_i,\ i \in I \tag{4.2}$$

$$\sum_{k=1}^{P_i} y_{ik\ell rt} X_{ik} = Y_{\ell rt},\ \ell \in L, r \in R(\ell), t \in T \tag{4.3}$$

$$LY_{\ell rt} \leq Y_{\ell rt} \leq UY_{\ell rt},\ \ell \in L, r \in R(\ell), t \in T \tag{4.4}$$

$$\sum_{m \in M} z_{\ell rmt} = Y_{\ell rt},\ \ell \in L_w, r \in R(\ell), t \in T \tag{4.5}$$

$$\sum_{\ell} \sum_{r} \partial_{\ell m} z_{\ell rmt} = \sum_{s \in S(m)} D_{smt},\ m \in M, t \in T \tag{4.6}$$

$$\sum_{s \in S(m)} D_{smt} \geq LD_{mt},\ m \in M, t \in T \tag{4.7}$$

$$0 \leq D_{smt} \leq d_{smt}^{\max},\ m \in M, s \in S(m), t \in T \tag{4.8}$$

The objective (4.1) is to maximize the net present value of all cash flows. These include the discounted revenues associated with the delivery of logs to markets, the discounted harvesting and silvicultural costs associated that result from applying silvicultural prescriptions to stands and the discounted transportation costs of delivering harvested log types from regions to mills. Equation (4.2) requires that the areas treated by all prescriptions equal the stand area. Equation (4.3) computes the yield of type l in each period in each region r where l is relevant. Equation (4.4) constrains this yield to lie within bounds chosen by the user over time. For those yields that correspond to wood products, Eq. (4.5) requires that the sum of the amount transported from the region to mill centres equals the amount produced in each period. Lumber products are converted to market demands via conversion coefficients in (4.6) and the total of those products delivered to market m must equal the amount sold in that market summed over the price ranges in that market. Market minimum demand is defined by the user in (4.7) and the amount that can be sold at each price is limited in (4.8).

Let us point out a few differences between the model (4.1–4.8) and conventional practice. Equations (4.5–4.8) and the corresponding terms in the objective function are not conventional. Most uses of the strategic harvest modelling focus on the land base and the production of timber. Constraints (4.2–4.3) are conventional although the most popular commercial software uses a Model II formulation instead of the Model I formulation in (4.2). The yield steering constraints (4.4) is one of the main differences. What is common instead of (4.4) is some sort of flow constraints for timber yields. Non-declining yield asserts that $Y_{\ell rt} \leq Y_{\ell r(t+1)}$; level flow $Y_{\ell rt} = Y_{\ell r(t+1)}$ and many practioners allow some percentage variation from either level flow or non-decline. Although the mathematical difference may be minor, philosophically there is quite a difference between non-declining yield and yield steering (4.4). The first is a normative model, the second is a strategic choice of the decision makers. This becomes more pronounced when we start to consider more than just timber yields over one or a small number of regions. The indicators to model model biodiversity (e.g., habitat), ecosystem condition (e.g., age

class distribution), watershed condition (percentage forest cover) are very complex (Gunn 2005) and the idea of non-decline or constant makes little sense. Note also the definition of yields in terms of regions for that yield. This is necessary to discuss the *SFM* criteria. This regional representation of yields implies a partitioning of the landscape where each element of the partition represents a unique combination of the underlying spatial regions (eg a stand is in a particular combination of *GIS* layers-ecodistrict, watershed, species habitat, economic region, etc.). One consequence of this is that Model II formulations rapidly grow in size as we attempt to account for this spatial complexity because we cannot logically merge stands that lie in different elements of the partition.

As indicated above, a second aspect to note is the explicit representation of where harvests occur in terms of the transportation of the harvested material to the mills where it can be used (Eqs. 4.5–4.8). More conventional models value the forest products produced by harvest at the stand and do not take into account that the demands for these products occur at specific locations with limited capacity to absorb these products. This can lead to an unrealistic assessment of overall system economics.

The third aspect is the objective function (4.1). The use of net present value is itself problematic for strategic planning. It is quite possible for an *NPV* objective to have positive cash flows in early years, but reducing, possibly to negative cash flows in subsequent years. This is obviously a problem for any real economic systems. Church and Daugherty (1999) has pointed out that other objectives may address the issues of intergenerational equity and may produce more sustainable economics. However, a somewhat surprising phenomenon found in many jurisdictions, is that the objective is purely maximization of volume harvested. In this viewpoint, the role of the forest is to produce wood, possibly subject to ecological constraints and it does not really matter where or what type of wood is produced. It is up to the industry to decide how to use the wood.

Martin (2013) and Martin et al. (2014a) shows that this form of model is capable of dealing with an extensive spatial representations of the indicators of sustainable forest management without blowing up as is the case when more conventional model II representations are used. In Martin (2013) and Martin et al. (2014b), a case study on Crown Lands (provincial lands) in central Nova Scotia shows that taking into account the location of facilities and the transportation issues can lead to substantial changes in the level and timing of harvest of various types of timber.

4.7 Strategic Location of Mills, Processing and Transport Facilities

If the role of the strategic harvesting model is to decide how much wood is to be produced, of what type and where, a natural converse is where should this wood go to make the maximum economic contribution. This raises the question of what types of mills we should have, how large should they be and where should they be located? There are several factors that make these questions very important. The first

involves economies of scale. Sawmills and pulpmills, like most process industries, are subject to substantial economies of scale in both capital cost and labor cost. This suggests a few big mills. A second is transportation cost. Wood coming from the forest is about 50 % moisture; moisture that will eventually have to be removed in processing. Moving wet wood long distances is expensive suggesting the need for many small mills. A third is integration. Forest harvesting produces wood with a broad variety of species, dimensions and quality attributes. Small material is either sent directly to pulpmills or to some intermediate facility where it is debarked and chipped. Larger material goes to sawmills where, depending on species and log quality, it can be debarked, and sawn into a variety of types of lumber. However the sawing process produces a substantial percentage of chips, sawdust, planer mill shavings and bark. A volume breakdown of 40 % lumber, 45 % chips, 8 % sawdust, 7 % shavings and 10 % bark would not be unrealistic[2]. The chips, sawdust, shavings and bark will then go to pulpmills and/or energy facilities.

Thus, at the overall forest system level, there is an overall capacitated flow network design problem to be solved. In Gunn (2012) and Gunn and Martell (2015) a mathematical model like the following is presented for this design problem.
Sets:

T: set of time periods $(1 \ldots T)$
D: the set of districts
L: the set of facility locations
C: the set of capacity types
F: the set of wood flow types
M: the set of markets
$P(m)$: the set of price ranges for market m.

Data:

A^{Res}_{fdt}: resource availability of wood type $f \in F$ in district $d \in D$ in time period $t \in T$.
ρ_{mpt}: unit revenue at price level p in market m in period t
$o_{\ell ct}$: operating cost per unit production location ℓ, capacity type c in period t
$\tau^{W}_{df\,\ell ct}$: transportation cost from district d of type f wood to location ℓ which has capacity type c in period t
$\tau^{F}_{\ell_1 f \ell ct}$: transportation cost from location ℓ_1, of type f wood to location ℓ which has capacity type c in period t
$\tau^{M}_{mf\ell ct}$: transportation cost to market m, of type f wood from location ℓ which has capacity type c in period t
$\gamma^{+}_{fc}, \gamma^{-}_{fc}$: maximum, minimum percentage of wood flow type f in operating levels of capacity type c
β_{fc}: conversion coefficients for input wood flow type f to operating of capacity type c

[2] The reason for these percentages summing to 110 % is that, in Canada, tree stem volumes are usually measured inside bark.

α_{fc}: amount of output wood flow type f for unit operating level of capacity type c
MA^{+}_{mpt}: maximum demand at price level p for market m in period t
MA^{-}_{mpt}: minimum demand at price level p for market m in period t

Variables:

$MREV_t$: total market revenues in period t
$OpCost_t$: total operating cost in period t
TranCost_t: total transportation cost in period t
$X^{Cap}_{\ell ct}$: capacity of type c installed in location ℓ at beginning of period t
$CAP_{\ell ct}$: capacity of type c at location ℓ available throughout period t
$WF_{fd\ell ct}$: wood flow type f (typically logs) from district d to location ℓ which has capacity type c in period t
$CF_{f\ell_1\ell ct}$: wood flow type f (typically intermediate products) from location ℓ_1 to location ℓ which has capacity type c in period t
$MF_{mf\ell t}$: wood flow type f (typically market products) from location ℓ to market m in period t
$OL_{\ell ct}$: operating level of capacity type c at location ℓ in period t
M_{mpt}: market sales at market m and price level p in period t

The model can then be written as:

$$\text{Max} \sum_t (i+1)^{-t} \left[MREV_t - CapCost_t - OpCost_t - TranCost_t \right] \tag{4.9}$$

$$\text{s.t.} CapCost_t = \sum_{\ell, f, t} C\left(X^{Cap}_{\ell ct}\right) \;\; t = 1, \ldots, T \tag{4.10}$$

$$MREV_t = \sum_{m,p} \rho_{mpt} M_{mpt} \;\; t = 1, \ldots, T \tag{4.11}$$

$$OpCost_t = \sum_{\ell,c,t} o_{\ell ct} OL_{\ell ct} \;\; t = 1, \ldots, T \tag{4.12}$$

$$\begin{aligned} TranCost_t = &\sum_{f,d} \sum_{\ell,c,t} \tau^{W}_{fd\ell ct} WF_{fd\ell ct} + \sum_{f,\ell_1} \sum_{\ell,c,t} \tau^{F}_{f\ell_1\ell ct} CF_{f\ell_1\ell ct} \\ &+ \sum_{m,f} \sum_{\ell,t} \tau^{M}_{mf\ell ct} MF_{mf\ell t} t = 1, \ldots, T \end{aligned} \tag{4.13}$$

$$Cap_{\ell ct} = X^{Cap}_{\ell ct} + Cap_{\ell c,t-1}, \ell \in L, c \in C, t = 1, \ldots, T \text{(capital additions)} \tag{4.14}$$

$$\sum_{\ell,c} WF_{fd\ell ct} \leq A^{\text{Res}}_{fdt}, \;\; f \in F, \; d \in D, \; t = 1, \ldots, T \text{(wood availability)} \tag{4.15}$$

$$OL_{\ell ct} \leq Cap_{\ell ct}, \;\; \ell \in L, \; c \in C, \; t = 1, \ldots, T \text{(capacity limitation)} \tag{4.16}$$

$$\left\{\begin{array}{l}\gamma_{fc}^{-} OL_{\ell ct} \leq \sum\limits_{d} WF_{fd\ell ct} \leq \gamma_{fc}^{+} OL_{\ell ct} \\ \gamma_{fc}^{-} OL_{\ell ct} \leq \sum\limits_{\ell_1} CF_{f\ell_1 \ell ct} \leq \gamma_{fc}^{+} OL_{\ell ct}\end{array}\right\} \ell \in L,\ c \in C,\ t = 1, \ldots, T$$

receipe inputs (4.17)

$$\left\{\begin{array}{l}\sum\limits_{d} \beta_{fc} WF_{fd\ell ct} + \sum\limits_{\ell_1} \beta_{fc} CF_{f\ell_1 \ell ct} = OL_{\ell ct} \\ \sum\limits_{c} \alpha_{fc} OL_{\ell ct} = \sum\limits_{\ell_2, c_2} CF_{f\ell\ell_2 c_2 t} + \sum\limits_{m} MF_{mf\ell t}\end{array}\right\} \ell \in L, c \in C, t = 1, \ldots, T$$

input-output balances (4.18)

$$\left\{\begin{array}{l}\sum\limits_{f,\ell} MF_{mf\ell t} = \sum\limits_{p} M_{mpt},\ m \in M,\ t = 1, \ldots, T \\ MA_{mpt}^{-} \leq M_{mpt} \leq MA_{mpt}^{+}\, m \in M,\ p \in P(m),\ t = 1, \ldots, T\end{array}\right\}$$

market demand (4.19)

Equations (4.10–4.13) are accounting equations for capital cost, market revenues, operating costs and transportation costs with the ρ, o, τ in Eqs. 4.3, 4.4 and 4.5 being the appropriate revenue/cost coefficients. There are several different ways of modelling capacity costs ranging from distinct choices, fixed costs plus linear or piecewise linear. We have left these unspecified in Eq. (4.10). Equation (4.11), together with Eq. (4.19), models market m as having piecewise constant demand curves (Lebow et al. 2003; Gilless and Buongiorno 1985). Equation (4.14) keeps track of capacity. The initial capacities $Cap_{\ell c0}$ are given data and subsequent capacities computed from additions. Equation (4.15) limits the wood flows of every flow type from each district to an assumed amount given in each period. This model has a simplified notion of processing. Each facility is assumed to have a recipe that dictates the inputs and outputs of the facility. Equation (4.16) limits the amount of processing to the available capacity. Equation (4.17) limit the relative proportions of each type of forest wood flow (*WF*) and intermediate product wood flow (CF) to be within certain proportions of recipe inputs. The two types of equations given by (4.18) indicate that the inputs, which may be in different units (tonnes, m^3, cords), adjusted using conversion factors β_{fc}, have to add up to the total recipe amount in operating level units. The outputs of the recipe are in fixed proportions α_{fc}. The outputs of the recipe can be intermediate products that go to other facilities (*CF*) and/or products that go to final markets (*MF*). Equation (4.19) specify that flows to a market m are sold at one of a variety of prices with the amount that can be sold at a given price range p is limited by lower and upper bounds. This enables the use of price endogenous demand models (Lebow et al. 2003).

For simplicity, capacity in the model described appears to last forever. Friedenfelds (1981) discusses a variety of ways of modelling capacity which is either time limited, diminishes over time, or is retired. However, this raises the issue of time horizon? Typically, time scales for industrial investment are 20 years or less, but the value of end of horizon capacity is probably not zero. How to value that capacity is an important question. These problems are closer to classic location problems, particularly in the realm of supply chain design (Melo et al. 2009) in that they involve the direct location of facilities, such as various types of sawmills, pulpmills and energy facilities.

In these problems there is little direct interaction between facilities of the same type but there are strong capacity influences in terms of a total capacity to process forest products and the number, size and location of mills to provide this capacity with substantial economies of scale in these capacity decisions. There are strong interactions with the forest regions and the flows of wood that emanate from these regions. There are also strong interactions with mills of other types. For example, sawmills depend on pulpmills, and energy facilities to make use of the substantial (more than 50 %) byproducts of the sawing process in terms of chips, sawdust, shavings and bark. The relative location, and hence the transportation cost to these other facilities matters.

In much of the literature on capacity planning for the forest products industry, there has actually been quite limited discussion on the location of the main processing facilities (see the review in D'Amours et al. 2008). Much of the design has focused on the flow of materials through the supply chain with some emphasis on the location of material handling/transformation facilities to facilitate this flow and on the choice of processing technologies at the various facilities. (e.g., Gunnarsson et al. (2004, (2006; Carlsson and Ronnqvist 2005; Carlsson et al. 2009 and Chauhan et al. 2009). To some extent this has been caused by the depressed state of the industry which has led to greatly reduced production and the closure of many mills (see Canada 2009). Few new mills have been built in the past 30 years. However, going forward, as mills continue to age the need to replace obsolete capacity will create demand for new capacity even if inherent demand does not increase. Importantly, as we move to the new biorefineries, the question will arise where to put them and how big they should be.

4.8 Dealing with the Dicotomy: Strategic Harvest; Strategic Capacity

As discussed above, if we want to know how to deal with the locational issues of harvest with some reasonable approach to the economics of the forest, we need to know something about the capacity to use the wood produced and where this capacity is located. This can be about more than economics in some jurisdictions. For example, if a certain species, present in a large number of stands, which can only be used by a few (or zero) mills of limited capacity for that species, then what we see is a total harvest of other species that is much less than the total capacity of the mills. This implies we need a possibility of disposal at low, (possibly negative) value so as to make possible the harvest of the more desireable species. This is the extreme case but in general we can see that the value of the delivered wood, the capacity to use the wood and the cost of transport from the harvest region to the location of the capacity will determine the amount of harvest. Moreover the amounts actually harvested need to correspond in some reasonable way to the amounts planned to be harvested. This is because of the regeneration and growth issues of the forest. If a stand is not harvested, it cannot be regenerated and hence cannot grow as rapidly.

This is the fundamental problem with the more conventional practice of using models without the economics in Eqs. (4.6–4.8). It has been observed in several cases that the aggregate harvests carried out in early periods may be approximately those in the plan for several important species but well below plan for some less desireable species. Some will equate this under usage of certain species as overall sustainability but this is not the case. The failure to harvest consistent with the plan will lead to less wood being available over the long term. (see Paradis et al. 2013). It is a case of use it or lose it! In summary, as shown in the recent work in Martin (2013) and Martin et al. (2014b) the harvest plan depends on the capacity to actually use the wood produced.

The mirror image is in the capacity model (4.9–4.19). The key feature is in the Eq. (4.15) where the amount of wood of a given type consumed from a given region in a particular time period must be less than or equal than the amount available. If these availabilities come from a plan developed using the strategic harvesting model and the capacity model solution satisfies (4.15) as equality, then the capacity plan will be consistent with the forest capacity as computed. There are two main issues here. If the amounts of a wood type used from given locations and time periods in the capacity plan is inconsistent (i.e., $<$) with the amounts provided in the strategic harvest plan, then the strategic harvest plan will not come true not just for the given wood type but possibly for all wood types. The second problem is that if the original harvest plan was developed using estimates of profitability that discouraged the production of a certain type of wood, then this wood will not be available to the capacity planning model and hence the capacity that might profitably make use of it cannot be installed. Martin (2013) showed that by installing a mill of the right type in the right place, it is possible to raise the economical harvest to be close to that achieved under volume maximization.

One answer is to combine (4.1–4.8) and (4.9–4.19) into a single model with the availability $A^{Res}_{f,d,t}$ in Eq. (4.15) being the yields $Y_{\ell rt}$ of Eq. (4.3) where the yield type l corresponds to the wood flow type f and the region r corresponds to harvesting district d of Eq. (4.5). The obviously difficulty with this approach is the complexity of solving the mathematical model. Although the model (4.1–4.8) is just a linear program, with a substantial number of *SFM* indicators and a large number of stands, these models become very large. Combing these with the integer programming model (4.9–4.19) is likely to be unsolvable. A current Ph.D. project (Gunn and Shereshti 2012) will look at iterative strategies for joint analysis.

4.9 Understanding Harvest and Access Costs: Relative Location of Stands and Roads

The Strategic Harvesting model (1)–(8) ignores the question of can we actually access the stands to be harvested at the time we want to harvest them within a given region. This region could be an ecodistrict, a watershed or a timbershed. These are

often referred to in forestry as tactical problems. There are at least two main issues here. One is adjacency of stands, the other is roads to access the stands.

If we harvest a set of adjacent stands simultaneously, this creates an opening in the forest canopy. Many jurisdictions have constraints on the area of the opening that can be created either within one period or in terms several periods required for harvested stands to "green up" to a forested condition. The question is can we meet a certain target trajectory of harvests of certain wood types from a region over time without violating these adjacency constraints. Initial approaches to this problem emphasized heuristics such as simulated annealing (e.g., Lockwood and Moore 1993). Snyder and Revelle (1996) was an early mathematical programming approach. More recent approaches, Weintraub and Murray (2006) have emphasized exact integer programming approaches. These can be simple models to formulate but quite complex in their data creation process requiring an enumeration of either all clusters (possible sets of allowable openings, see Goycoolea et al. 2005), all paths (possible sets of forbidden openings, see Tóth et al. 2013). Alternatively, one can use somewhat more complex models (Constantino et al. 2008). All three approaches become challenging as one gets to problems of realistic scale. For example a timbershed can easily contain 10,000 stands. Gunn and Richards (2005) pointed out that if a simple "stand centred" adjacency constraint were used, large scale problems could be solved relatively easily. The constraint states that if a particular stand k is harvested then the harvest from all stands immediately adjacent to it must be less than the maximum allowable opening minus the area of the stand under consideration. It takes the form

$$\sum_{i \in A(k)} a_i x_i \leq M_k(1 - x_k) + (A^{\max} - a_k)$$

where k indexes the stand under consideration, $A(k)$ is the set of stands adjacent to k, a_i is the area of stand i and A^{max} is the maximum allowable opening size. The x_i are binary variables with $x_i = 1$ if stand i is harvested. The value M_k is set to a "sufficiently large" number. Gunn and Richards show how these constraints can be improved through solving strengthening and lifting subproblems.

This constraint only applies to those stands whose area plus that of its adjacent stands exceeds the allowable opening. Thus the number of adjacency constraints is less than or equal to the total number of stands in the timbershed. The method is not rigorous in the sense that it does not prevent "long stringy openings," where stands $i_1, i_2, i_3, \ldots, i_{k\text{-}1}, i_k$ are harvested and stand i_1 is adjacent to $i_2, \ldots, i_{k\text{-}1}$ is adjacent to i_k but stand i_1, and i_k are not adjacent. Gunn and Richards point out that in the optimal solution to practical situations the number and total area of such long stringy openings is very small relative to the number of stands and the total area of the region.

The road building problem is quite closely related to the adjacency problem. If we plan to harvest certain stands over time, then when a particular stand is harvested, there must be a road available to allow equipment to access the stand and to remove the wood produced. Modeling the connection between the dynamics of stand harvesting and road building are demonstrated in Andalaft et al. (2003) although this paper does not deal with adjacency issues. Nelson and Finn (1991) formulated problems of road building with adjacency and solve these with random search heuristics.

However they dealt with cut blocks, not stands. Stands had been pre-assigned to cut blocks of such a large size that cutting any two adjacent cut blocks would violate adjacency. Richards and Gunn's (2000) tabu search approach included solving this problem with individual stand adjacency. In this case the problem of building roads to access the stands scheduled for cutting is a dynamic Steiner tree problem (Imase and Waxman 1991). Richards (1997) developed an efficient heuristic for this Steiner problem, which served as an evaluator for any particular choice of harvest decisions during the tabu search. Richards and Gunn (2006) demonstrated that the same large scale problems solved in Richards and Gunn (2000) can be solved as integer programming problems by using the stand-centred adjacency constraints together with tight linkages to the road network. Sample problems with 1039 stands, 135 road segments and 5 planning periods (5 year) produce optimality gaps of less than 1 % within 300 s in 24 of 30 problems. Gaps less than 5 % are produced within 500 s in all cases. The road building problem in Richards and Gunn deals with a higher level with fairly long potential road segments. Real road building problems are more complex. Meignan et al. (2012) discuss more detailed road location problems that take the stand harvests as given.

In terms of the longer term strategic forest harvesting and capacity models, these tactical models serve two roles. One is to estimate how much wood can actually be produced once adjacency is accounted for (Eq. 4.15). The second is to estimate the cost of accessing this wood. Per unit road costs are a key part of the cost coefficients in (4.1) and in (4.13) and the particular road segments constructed also influence the hauling distance to access the main highway network. Again we have a dichotomy. We need solutions to the strategic harvesting model to estimate the amount of wood we want to harvest from a particular region and we need the solution to these adjacency/road building problems to estimate access costs and assess realistic harvest levels.

4.10 Understanding Materials Handling and Logistics

The costs of moving material from a harvest to the particular facilities depend on the logistics of the movement. In both the forest harvesting strategy models (4.1–4.8) and the capacity models (4.9–4.18), one needs the cost of harvesting wood and delivering it to some facility for processing. However the supply chain is more complicated than this. We have choices of how to harvest, with many different types of machines and systems. These harvesting systems may produce either full trees, or cut-to-length wood assortments or wood chips or combinations of all three. Full trees may be the cheapest harvest in the woods but the trucks that can access the woods operations are inefficient for long hauls. Also the full tree can be broken down into different products (sawlogs, studwood, pulpwood, biomass chips) that might need to go to different locations. The contribution by Epstein et al. (2007a) reviews some of the issues in designing the forest transportation system. D'Amours et al. (2008) treat this as the tactical problem of designing the procurement for the

forest products supply chain. From a location theory point of view, perhaps the main interest here involves the decisions to create processing terminals where wood from the forest, perhaps tree length, perhaps otherwise, can be sorted and processed (cut to length, chipped), and transferred to more efficient trucks or possibly rail for movement to other facilities. How many such terminals, the types of processing available, the transport modes and their location become the design problem. One could look at such terminals as just additional processing facilities with products that go to other facilities in the context of the capacity model (4.9–4.19). However, as can be seen in Gunnarsson et al. (2004), the modelling of these logistics facilities rapidly produces very large models, even in an environment where energy facilities are given.

Transportation can be between 30 and 50 % of the costs of delivering wood to mills from the forest. Thus the logistics system can have significant effects on where we harvest and what capacity we build. Again we see the situation where in order to properly to develop forest harvest strategy and capital investment strategy, we need good estimates of transportation costs and we need supply chain logistics design models to estimate these costs. However the supply chain logistics models need to know where the wood is coming from and the facilities to which it can go. Thus we need solutions to the forest harvest strategy models and capacity investment models to develop good logistics systems.

4.11 Operational Location of Camps

In some settings in Canada (the boreal forest), harvesting occurs at considerable distance from communities with housing and other services. Companies who wish to harvest have a choice of dealing with the cost and inefficiencies of transporting workers long distances daily, or establishing camps with accommodation, recreation and other facilities where crews can be accommodated near the workplace. In the forest harvesting strategic model we have a single period consisting of 5 years of harvests in a region. However in order to understand the harvesting costs associated with this plan we need to know how the workers will be accommodated and transported. The problem is dynamic in two senses, first the harvest moves around the region over the 5 years. However, there are also strong seasonal influences. Many of the stands in the region can only be harvested in the winter when the ground is frozen sufficiently to permit both harvesting and transport. Jena et al. (2012) have shown that the problem of deciding the location, number and capacity of these camps is a very interesting dynamic location problem. The modular aspect where capacity is provided in terms of fairly standard trailers which can be moved from one location to another if necessary, and the dynamic aspect, where camps can be open and closed at various times during the year, make the solution of this dynamic location problem challenging. They have also been able to show quite substantial savings in total costs (transportation costs plus camp location and relocation costs) through the application of their approach.

Again we have the situation that in order to understand the costs of harvesting for the forest harvesting model, we need to know how much it will cost to accommodate and transport the workers. However, solving this problem of where and how to accommodate the workers depends on the original 5 year plan of total harvest in the region.

4.12 Operational Problems; Machine and Crew Scheduling; Product Transport

As we work down the hierarchal planning system, it will not be surprising to find that there are a host of problems in which location plays a major role Epstein et al. (2007b) characterizes these into seven main areas. We mention three of these here. These include the machinery location problem, the related crew scheduling problem and the detailed transportation planning problems that examine truck routing and load assignment.

Epstein et al. (2006) looks at the complexity of machinery location and road design. These problems are applied to somewhat small landbases (1000 ha or less) and are often known in the industry as the landing location problem. This was one of the early applications of location analysis in forestry (Dykstra and Riggs 1977). They can be very challenging (Legues et al. 2007). A landing is a point on the road network to which logs are delivered by the harvesting machinery so that they can then be sorted and trucked onwards to the logistics system. These problems typically deal with time periods of days and weeks and quite fine land divisions ($10\,m^2$) for sources of timber. The problem is how many landings to have, where they should be located and what connecting roads to build to remove the harvested wood. These problems are particularly important in mountainous areas where road building is expensive and the landing may correspond to the location of logging towers with aerial cables to deliver the wood from where it is cut to the landing. On more level ground, the landing is the site from which mobile skidders/tractors travel to the cut trees, pick them up and deliver them back to the landing. The more landings that you have, the easier it is to remove the wood from the forest to the landing. However, all landings require a fixed cost to establish. Epstein et al. (2007a) provide a history of the problem, a mixed integer programming formulation and a heuristic algorithm for solving these complex problems.

The crew scheduling problem is the dynamic problem of assigning a set of cutting crews to harvest various cut locations. There are three main issues governing the costs. First, different crews may have different types of equipment and, depending on topography, each type of equipment may be more or less productive on a given cut location. This affects both the time and the cost to harvest the location. Second, each crew may have a home base and there is a cost of daily travel from the home base to the cutting location. Third, once the crew has finished harvesting a given location, their equipment must be moved to the next location. This cost of relocation can be substantial and depends on the distance between the cutting locations. Complicating this is an ongoing demand for various types of logs from a set of

customer mills. Each of the cutting locations may differ in terms of the species and quality of logs produced by harvesting. Murphy (1998) deals with a static problem of assigning crews to stands to maximize equipment productivity. Mitchell (2004) deals with the complete dynamic problem. Mitchell's work utilizes David Ryan's (1992) set covering and constraint branching ideas, originally applied in air crew scheduling, to solve these complex problems quite efficiently. In some cases it is hard to quantify the ability of certain crews to work in a given location's topography and soil conditions. Gallant and O'Brien (2004) developed an approach where the company scheduler can decide a sequence of harvest locations for each crew. A *VBA* application within *MS* Access computes the timing and costs for each crew and a series of simple network flow problems solves the allocation to customers. The user can assess whether or not harvest location assignments and sequence should be changed.

Product transport is an enormous area; one that we can't deal with in any detail in this setting. The short message is that location implies the need for transportation. Thus transportation systems are challenged by the location of harvest activities and the location of the facilities that make up the supply/value chain. Conversely, efficient transportation systems can be the means of overcoming the location challenges in forestry. Two of the most noteworthy groups in this area have been the Chileans and the Swedes although some considerable progress is being made by the *FORAC* group in Quebec as well as the more broadly based *VCO* Network. The paper by Epstein et al. (1999) is illustrative of the long standing work in Chile. The *ASICAM* system, described further in Weintraub et al. (1996), deals with daily harvesting and truck movements with substantial savings in costs. One of the interesting location issues discussed in Epstein et al. (1999) is that the way trees are optimally bucked (cut up into logs of various sizes) depends on the location of the stand relative to the various mills that can use these logs. The Swedish work is illustrated in Carlgren et al. (2006), Forsberg et al. (2005), Karlsson et al. (2003) and Carlsson and Ronnqvist (2005). The latter of these is notable for the inclusions of the possibility of backhauls into the transportation calculations. Frisk et al. (2010) point out that in multi-company situations where companies have their own harvesting locations and facility locations, that an overall global solution that ignores company issues can produce substantial transportation savings over and above the sum of the individual company optimal transport costs. This raises questions of how these companies can collaborate, sharing benefits appropriately, to capture these potential savings while making all companies better off.

4.13 Uncertainty

Uncertainty is fundamental to forestry (see Martell et al. 1998). This includes uncertainty in products prices/demands, fuel costs, labour costs, and productivity that are typical of all production planning problems. However, there are fundamental forestry uncertainties. These include uncertainty in the amount and quality of trees

growing on a piece of land. Although new technologies are now enabling quite precise inventories (Pitt and Pineau 2009; Næsset 2007), some uncertainty in the amount and product breakdown of harvest will always remain. Similarly uncertainty in harvesting machine productivity, moisture levels and process yields will always affect the overall cost structure. These are natural products and come with all the variability of nature. This suggests that a well structured hierarchical planning process is a necessity (Gunn 1996). Within this hierarchical process, location issues play a large role in decision making.

Natural disturbance is a fundamental characteristic of forests. Fires, insect and disease can quite rapidly fundamentally alter the availability of forests for harvesting and the location of this availability. It can also change the various spatial indicators of biodiversity, ecosystem condition, and watershed condition. (James et al. 2007) This means that following a major disturbance that substantial replanning of forest harvest strategy and possibly industrial location strategy may be necessary.

There is a location issue that comes up in dealing with natural disturbance; namely how to allocate the prevention effort. Martell (1994) has indicated ways of valuing the forest fighting effort and indicating the return for this effort to prevent fires. This is part of a large body of work stemming from Martell's group at the University of Toronto (see Martell 2015). At a level of more spatial detail, Minas et al. (2014) have pointed out that appropriate cutting patterns can reduce the spread of fire and limit adverse impacts. Insects, such as the spruce budworm have also resulted in considerable spatial analysis in developing decision support tools (see MacLean et al. 2001, for example). Daugherty and Fried (2007) link fuel treatments to location of biomass facilities.

At an operational level the detection and the fighting of disturbance also poses locational challenges. MacLellan and Martell (1996), Islam and Martell (1998) and Islam et al. (2009) study how the location of bases for forest fire airtankers, the allocation of the airtankers to these bases and the strategy of deployment of the airtankers to the fire locations affects both cost and effectiveness of the fire suppression effort.

4.14 Summary

We have initially focused on the models for strategic forest management and industrial capacity location. The forest management models start with the requirement to deal with the spatial entities required to characterize biodiversity, ecosystem condition and watershed condition, key criteria of sustainable forest management. Strategic decisions on the indicator levels for these spatial entities will drive or constrain harvest strategies. Other criteria of sustainable forest management, namely multiple economic benefits require us to deal with location of the forest resource relative to the processing plants and markets and relative to the resource dependent communities, including First Nations communities.

We have seen that for harvests to be economical we need to plan these harvests with regard to the location of the sawmills, pulpmills and energy facilities that will be the destination for the wood produced. Harvest costs at the forest stands, transportation costs to the facilities and between facilities and the prices at the facilities will determine what harvest plans make sense. Put simply in order to plan harvests we need to know where the wood will go. However, we have also seen that determining what facilities make sense in terms of number, type, size, and location that we need to know how the amount of wood harvested varies over time by species type, quality/size, location and harvest cost. Again put simply, in order to decide on facility capacity and location, we need to know where the wood will come from. This dichotomy probably cannot be resolved completely algorithmically. There are issues both of model size and solvability but there are also issues of how system level benefits can be shared among the parties.

However it is also important to be aware that both of these strategic problems depend on more detailed models. The assumed availability and cost of wood in a region depends on the ability to access that wood, both in terms of ecological acceptability as portrayed through the adjacency models and in terms of road building. Similarly the ability to transport wood efficiently from the forests may well depend on investments in the logistics system in terms of investments in terminals to facilitate intermediate processing and transfer to more efficient long distance transport, including multimodal. These are often seen as tactical issues, but in reality they are also quite strategic. Investments here change costs, which permit revisiting the harvest and capacity strategies.

This story is repeated as we become more operational. Better logistics processes again change costs and again we should revisit out strategies. Even without uncertainty, understanding the complexity of forestry location related decisions requires a hierarchical family of models used not to get ultimate answers but to continually explore for opportunities of integration, collaboration and continuous improvement. Once uncertainty raises its head, a hierarchical approach is inevitable.

References

Andalaft N, Andalaft P, Guignard M, Magendzo A, Wainer A, Weintraub A (2003) A problem of forest harvesting and road building solved through model strengthening and Lagrangean relaxation. Oper Res 51(4):613–628

Anthony RN (1965) Planning and control systems: a framework for analysis. Harvard University, Graduate School of Business Administration, Boston

Bare BB, Briggs DG, Roise JP, Schreuder GF (1984) A survey of systems analysis models in forestry and the forest products industries. Eur J Oper Res 18(1):1–18

Baskent EZ, Keles S (2005) Spatial forest planning: a review. Ecol Model 188(2):145–173

Canada Parliament Senate. Standing Senate Committee On Agriculture And Forestry (2009) The Canadian forest sector: past, present, future, interim report of the Standing Senate Committee on Agriculture And Forestry, Dec 2009. http://www.parl.gc.ca/Content/SEN/Committee/402/agri/rep/repintdec09-e.pdf. Accessed 13 Apr 2015

Canadian Council of Forest Ministers. (2003). Defining Sustainable Forest Management in Canada 2003. http://ccfm.org/pdf/CI_Booklet_e.pdf. Accessed 13 Apr 2015
Carlgren CG, Carlsson D, Rönnqvist M (2006) Log sorting in forest harvest areas integrated with transportation planning using backhauling. Scand J For Res 21(3):260–271
Carlsson D, Rönnqvist M (2005) Supply chain management in forestry—-case studies at Södra Cell AB. Eur J Oper Res 163(3):589–616
Carlsson D, D'Amours S, Martel A, Rönnqvist M (2009) Supply chain planning models in the pulp and paper industry. INFOR: Inf Syst Oper Res 47(3):167–183
Chauhan SS, Frayret JM, LeBel L (2009) Multi-commodity supply network planning in the forest supply chain. Eur J Oper Res 196(2):688–696
Church RL (2007) Tactical-level forest management models. In: Weintraub A, Romero C, Bjørndal T, Epstein R (eds) Handbook of operations research in natural resources, Springer, New York, pp 317–341
Church RL, Murray AT, Weintraub A (1998) Locational issues in forest management. Locat Sci 6(1):137–153
Church R, Daugherty PJ (1999) Considering intergenerational equity in linear programming-based forest planning models with MAXMIN objective functions. Forest Sci 45(3):366–373
Constantino M, Martins I, Borges JG (2008) A new mixed-integer programming model for harvest scheduling subject to maximum area restrictions. Oper Res 56(3):542–551
D'Amours S, Rönnqvist M, Weintraub A (2008) Using operational research for supply chain planning in the forest products industry. INFOR: Inf Syst Oper Res 46(4):265–281
Daugherty PJ, Fried JS (2007) Jointly optimizing selection of fuel treatments and siting of forest biomass-based energy production facilities for landscape-scale fire hazard reduction. INFOR: Inf Syst Oper Res 45(1):17–30
Davis LS, Johnson KN, Bettinger PS, Howard TE (2001) Forest management, 4th edn. McGraw-Hill, New York
Dykstra DP, Riggs JL (1977) An application of facilities location theory to the design of forest harvesting areas. AIIE Trans 9(3):270–277
Epstein R, Morales R, Seron J, Weintraub A (1999) Use of OR systems in the Chilean forest industries. Interfaces 29(1):7–29
Epstein R, Weintraub A, Sapunar P, Nieto E, Sessions JB, Sessions J, Bustamente F, Musante H (2006) A combinatorial heuristic approach for solving real-size machinery location and road design problems in forestry planning. Oper Res 54(6):1017–1027
Epstein R, Rönnqvist M, Weintraub A (2007a) Forest transportation. In: Weintraub A, Romero C, Bjørndal T, Epstein R (eds) Handbook of operations research in natural resources, Springer, New York, pp 317–341
Epstein R, Karlsson J, Rönnqvist M, Weintraub A (2007b) Harvest operational models in forestry. In: Handbook of operations research in natural resources. Springer, New York, pp 365–377
Forsberg M, Frisk M, Rönnqvist M (2005) FlowOpt–a decision support tool for strategic and tactical transportation planning in forestry. Int J For Eng 16(2):101–114
Freidenfelds J (1981) Capacity expansion: analysis of simple models with applications. North-Holland
Gallant R, O'Brien S (2004) A decision support system for Bowater's Mersey woodland operations, senior year project. Department of Industrial Engineering. Dalhousie University, Halifax
Gibson DF, Rodenberg J (1975) Location models for the forest products industry and other applications. AIIE Trans 7(2):143–152
Gilless JK, Buongiorno J (1985) PELPS: Price-endogenous linear programming system for economic modeling. Research division of the college of agricultural and life sciences. University of Wisconsin, Madison
Goycoolea M, Murray AT, Barahona F, Epstein R, Weintraub A (2005) Harvest scheduling subject to maximum area restrictions: exploring exact approaches. Oper Res 53(3):490–500
Gunn EA (1996) Hierarchical planning in forestry: a stochastic programming, decision analytic perspective. In: Martell DL, Davis LS, Weintraub A (eds) Proceedings of the workshop on hierarchical approaches to forest management in public and private organizations. University of

Toronto, Petawawa National Forestry Institute, Information Report PI-X-124, Toronto, Ontario, Canada, 25–29 May, Canadian Forest Service (1996), pp 85–97
Gunn EA (2005) Sustainable forest management: control, adaptive management, hierarchical planning. In: Bevers M, Barrett M, tech. comps. 2005. System analysis in forest resources: proceedings of the 2003 symposium. Gen. Tech. Rep. PNW-GTR-656. U.S. Department of Agriculture, Portland, OR, Forest Service, Pacific Northwest Research Station, pp 7–14
Gunn EA (2007) Models for strategic forest management. In:Weintraub A, Romero C, Bjørndal T, Epstein R (eds) Handbook of operations research in natural resources, Springer, New York, pp 317–341
Gunn, Martell (2015) Decision support needs for strategic planning of Canadian forest value chains, chapter 4. In D'Amours S., Ouhimmou M., Audy J.-F., Feng Y (eds) Forest value chain optimization and sustainability. CRC Press/Taylor & Francis.
Gunn E (2009) Some perspectives on strategic forest management models and the forest products supply chain. INFOR: Inf Syst Oper Res 47(3):261–272
Gunn EA, Richards EW (2005) Solving the adjacency problem with stand-centred constraints. Can J For Res 35(4):832–842
Gunn E, Shereshti N (2012) What are the benefits and risks associated with tightly integrated industry structure. What can one do in the design of the industry to mitigate the risk of uncorrelated final product markets? Project 1.17, NSERC/FPInnovations Strategic Research Network, Value Chain Optimization in the Forest Industry (2012–2015)
Gunnarsson H, Rönnqvist M, Lundgren JT (2004) Supply chain modelling of forest fuel. Eur J Oper Res 158(1):103–123
Gunnarsson H, Rönnqvist M, Carlsson D (2006) A combined terminal location and ship routing problem. J Oper Res Soc 57(8):928–938
Imase M, Waxman BM (1991) Dynamic Steiner tree problem. SIAM J Discret Math 4(3):369–384
Islam KM, Martell DL (1998) Performance of initial attack airtanker systems with interacting bases and variable initial attack ranges. Can J For Res 28(10):1448–1455
Islam KS, Martell DL, Posner MJ (2009) A time-dependent spatial queueing model for the daily deployment of airtankers for forest fire control. INFOR: Inf Syst Oper Res 47(4):319–333
James PM, Fortin MJ, Fall A, Kneeshaw D, Messier C (2007) The effects of spatial legacies following shifting management practices and fire on boreal forest age structure. Ecosystems 10(8):1261–1277
Jena SD, Cordeau JF, Gendron B (2012) Modeling and solving a logging camp location problem. Ann Oper Res 1–27 doi:10.1007/s10479-012-1278-z
Johnson KN, Scheurman HL (1977) Techniques for prescribing optimal timber harvest and investment under different objectives—discussion and synthesis. For Sci 23 (Suppl. 18):a0001–z0001
Karlsson J, Rönnqvist M, Bergström J (2003) Short-term harvest planning including scheduling of harvest crews. Int Trans Oper Res 10(5) 413–431
Lebow PK, Spelter H, Ince PJ (2003) FPL-PELPS, a price endogenous linear programing system for economic modelling; supplement to PELPS III, V 1.1, USDA forest service, forest products laboratory research paper FPL-RP-614, Madiaon, WI. 32 p
Legües AD, Ferland JA, Ribeiro CC, Vera JR, Weintraub A (2007) A tabu search approach for solving a difficult forest harvesting machine location problem. Eur J Oper Res 179(3):788–805
Lockwood C, Moore T (1993) Harvest scheduling with spatial constraints: a simulated annealing approach. Can J For Res 23(3):468–478
MacLean DA, Erdle TA, MacKinnon WE, Porter KB, Beaton KP, Cormier G, Morehouse S, Budd M (2001) The spruce budworm decision support system: forest protection planning to sustain long-term wood supply. Can J For Res 31(10):1742–1757
MacLellan JI, Martell DL (1996) Basing airtankers for forest fire control in Ontario. Oper Res 44(5):677–686
Marques AF, Borges JG, Sousa P, Pinho AM (2011) An enterprise architecture approach to forest management support systems design: an application to pulpwood supply management in Portugal. Eur J For Res 130(6):935–948
Martell DL (1994) The impact of fire on timber supply in Ontario. For Chron 70(2):164–173

Martell DL (2015) Fire management systems laboratory. http://www.firelab.utoronto.ca/publications/dlm_pubs.html. Accessed 9 Apr 2015
Martell DL, Gunn EA, Weintraub A (1998) Forest management challenges for operational researchers. Eur J Oper Res 104(1):1–17
Martin AB (2013) A linear programming framework for models of forest management strategy. MASc Thesis, Department of Industrial Engineering, Dalhousie University. http://dalspace.library.dal.ca/handle/10222/37840. Accessed 9 Apr 2015
Martin AB, Richards EW, Gunn EA (2014a) A comparison of model one and model two for modeling of forest management strategy. Working paper (unpublished). Department of Industrial Engineering, Dalhousie University
Martin AB, Gunn EA, Richards EW (2014b) Modelling forest management strategy with industry representation. Working paper (unpublished), Department of Industrial Engineering, Dalhousie University
Meignan D, Frayret JM, Pesant G, Blouin M (2012) A heuristic approach to automated forest road location. Can J For Res 42(12):2130–2141
Melo MT, Nickel S, Saldanha-da-Gama F (2009) Facility location and supply chain management—a review. Eur J Oper Res 196(2):401–412
Minas JP, Hearne JW, Martell DL (2014) A spatial optimisation model for multi-period landscape level fuel management to mitigate wildfire impacts. Eur J Oper Res 232(2):412–422
Mitchell SA (2004)Operational forest harvest scheduling optimisation: A mathematical model and solution strategy. (Doctoral dissertation, ResearchSpace@ Auckland)
Montreal Process (2014) Montréal process criteria and indicators for the conservation and sustainable management of temperate and boreal forests. Technical Notes on Implementation of the Montréal Process Criteria And Indicators Criteria 1–7, 3rd edn. June 2009 (Rev. July 2014). http://www.montrealprocess.org/documents/publications/techreports/MontrealProcessTechnicalNotes3rdEditionRevisedJuly2014.pdf. Accessed 9 Apr 2015
Murphy G (1998) Allocation of stands and cutting patterns to logging crews using a tabu search heuristic. J For Eng 9(1):31–37
Næsset E (2007) Airborne laser scanning as a method in operational forest inventory: status of accuracy assessments accomplished in Scandinavia. Scand J For Res 22(5):433–442
Nelson JD, Finn ST (1991) The influence of cut-block size and adjacency rules on harvest levels and road networks. Can J For Res 21(5):595–600
Paradis G, LeBel L, D'Amours S, Bouchard M (2013) On the risk of systematic drift under incoherent hierarchical forest management planning. Can J For Res 43(5):480–492
Pitt D, Pineau J (2009) Forest inventory research at the Canadian Wood Fibre Centre: notes from a research coordination workshop, June 3–4, 2009, Pointe Claire, QC. For Chron 85(6):859–869
ReVelle CS, Eiselt HA (2005) Location analysis: a synthesis and survey. Eur J Oper Res 165(1):1–19
Richards EW (1997) A tabu search method for a tactical forest planning problem, Ph. D. Thesis, Dept. of Industrial Engineering. Dalhousie University, Halifax, N.S. Canada, p 247
Richards EW, Gunn A (2000) A model and tabu search method to optimize stand harvest and road construction schedules. For Sci 46(2):188–203
Richards EW, Gunn EA (2006) Integer programming models for tactical harvest and access planning. INFORMS Conference, Pittsburg, PA. (unpublished)
Ryan D (1992) The solution of massive generalised set partitioning problems in aircrew rostering. J Oper Res Soc 43(5):459–467
Snyder S, ReVelle C (1996) The grid packing problem: selecting a harvesting pattern in an area with forbidden regions. For Sci 42(1):27–34
Tóth SF, McDill ME, Könnyü N, George S (2013) Testing the use of lazy constraints in solving area-based adjacency formulations of harvest scheduling models. For Sci 59(2):157–176
Ware GO, Clutter JL (1971) A mathematical programming system for the management of industrial forests. For Sci 17(4):428–445
Weintraub A, Murray AT (2006) Review of combinatorial problems induced by spatial forest harvesting planning. Discret Appl Math 154(5):867–879
Weintraub A, Epstein R, Morales R, Seron J, Traverso P (1996) A truck scheduling system improves efficiency in the forest industries. Interfaces 26(4):1–12

Chapter 5
Layout Planning Problems in Health Care

Ines Arnolds and Stefan Nickel

5.1 Introduction

In general, layout planning problems can be classified as in-house location problems where the aim is to minimize traveling or material handling costs based on distances by deciding on the relative positions of any kind of organizational units inside a building. This class of operations research problems originates from industrial applications, for example, planning the location of different machines of an assembly line needed to manufacture a product or the arrangement of racks and shelves within a warehouse.

The special case of layout planning problems in health care has been first introduced by Elshafei in 1977 (Elshafei 1977). He modeled a hospital layout problem as a quadratic assignment problem (*QAP*) and developed heuristics to solve it. In the framework for hospital planning and control the hospital layout planning problem is classified as a resource capacity planning problem on a strategic level (Hans et al. 2012). Although it is a long-term decision, the spatial organization within hospitals also directly influences the quality and efficiency of health care and secondary services of the daily routine (Choudhary et al. 2010; Hignett and Lu 2010) as well as patient satisfaction (Chaudhury et al. 2005). In practice, hospital buildings are commonly planned by architects based on experience, design aspects and legal regulations. Instead of that, it is important to develop and follow a holistic approach in order to combine the architectural and legal aspects with logistics, i.e., patient, personnel and material flows inside the future hospital building. In this context, the established operations research methodologies, especially optimization and simulation techniques, can be applied in order to support finding an optimal or robust hospital layout. On the one hand, optimal can mean to minimize traveling costs for

I. Arnolds (✉) · S. Nickel
Discrete Optimization and Logistics at the IOR, Karlsruhe Institute of Technology,
Englerstr. 11, Gebäude 11.40, 2. OG, 76131 Karlsruhe, Germany
e-mail: ines.arnolds@kit.edu

S. Nickel
e-mail: stefan.nickel@kit.edu

H. A. Eiselt, V. Marianov (eds.), *Applications of Location Analysis,* International Series in Operations Research & Management Science 232, DOI 10.1007/978-3-319-20282-2_5

personnel or traveling distances and/or times for patients and/or material. Although these objectives might be conflicting, they not only result in more efficient workflows and, thus, in patient and personnel satisfaction but also in economic efficiency. On the other hand, robustness implies that a layout plan has a good performance for different scenarios with uncertain input data, for example, uncertain clinical pathways depending on the patients' recovery.

A hospital layout planning problem where all functional departments, wards, surgery rooms and other necessary and supporting areas have to be assigned to locations inside the hospital building is referred to as a layout planning problem on the macro level. In contrast, when only planning the layout of a single functional department, ward, etc. in the building it is called a hospital layout planning problem on the micro level. In the next section an overview of the literature on hospital layout planning problems on both levels is given. Nevertheless, the focus of the applications detailed in Sect. 5.3 as well as the framework presented in Sect. 5.4 lies on the macro level. In Sect. 5.5 a summary is given and some practical challenges are discussed.

5.2 Literature Review

A survey on layout planning problems was conducted by Drira et al. (2007). Furthermore, Tompkins et al. (2010), Heragu (2008), and Francis et al. (1992) published textbooks presenting different modeling and solution techniques for layout planning problems in general. In most applications the layout is considered as a long-term decision (Drira et al. 2007). In such static layout problems all relevant parameters are assumed to remain constant during the entire planning horizon. Nevertheless, there may also come up issues that make it necessary to rearrange a given layout, for example, the development of new products with different production processes or new treatment procedures that change the clinical pathways of patients with a specific disease. Thus, dynamic layout problems were developed in order to consider varying input data during the planning horizon. Two approaches exist to reflect this variability (Drira et al. 2007; Moslemipour et al. 2012): developing a robust layout that is best in sum over all periods during the planning horizon (see, for example, Kouvelis 1992; Benjaafar and Sheikhzadeh 2000; Azadivar and Wang 2000; Aiello 2001; Kulturel-Konak et al. 2004; Enea et al. 2005; Braglia et al. 2005; Norman and Smith 2006; Pillai et al. 2011; Arnolds and Nickel 2013b), or developing a layout plan for multiple periods where layout adaptations are allowed for while incurring rearrangement costs (see, for example, Lacksonen 1994; Urban 1998; Yang and Peters 1998; Kochhar and Heragu 1999; Balakrishnan and Cheng 2000; Chang et al. 2002; Krishnan et al. 2006, 2008; Kulturel-Konak 2007a; Ulutas and Islier 2009; Bashiri and Dehghan 2010). Reviews on dynamic layout problems and solution approaches are given in Balakrishnan and Cheng (1998), and Kulturel-Konak et al. (2007).

Regarding hospital layout planning problems, an exhaustive search for relevant literature follows. The objective of this literature review is threefold: First, it is designed to give an overview on recent advances in layout planning problems in health care with a focus on hospitals both on the macro and micro level. Second, a taxonomy, i.e., a structured way of classifying the reviewed papers, is developed to support discovering linkages between various publications as well as comparing them. Third, the extent to which diverse issues on layout planning in health care have already been covered in the literature is identified. Thus, existing research gaps can be revealed.

The following search string was used in the search engine Scopus:

TITLE-ABS-KEY ((“layout” OR “facility planning” OR “facilities planning” OR “facility design” OR “facilities design”) AND (“hospital” OR “clinic”) AND (“heuristic*” OR “optimization” OR “mixed integer program*” OR “mathematical program*” OR “integer program*” OR “linear program*” OR “binary program*” OR “quadratic program*” OR “dynamic program*” OR “goal program*” OR “discrete event simulation” OR “discrete-event simulation” OR “discrete-event-simulation”)).

The asterisk (*) may be replaced by different character combinations. For example, searching for “mathematical program*” can result in, for example, “mathematical program” or “mathematical programming”.

Using the presented search string, 59 papers were retrieved. The abstracts of these papers were screened in order to identify irrelevant articles and sort them out. After that, 22 relevant papers remained. Based on these, both a forward and a backward search were conducted. As an indicator for relevant papers the titles, abstracts and keywords of the forward and backward search results were scanned for layout and health care relevant applications. Furthermore, the papers cited by the literature reviews of Jun et al. (1999), Günal and Pidd (2010), Forsberg et al. (2011) were examined such that an additional 33 papers were retrieved.

In order to categorize the total of 55 papers, the following taxonomy was developed where most of the papers can be categorized into at least one topic of each category:

- Scope: This category differentiates between hospital layout planning problems on the macro and micro level. The micro level is further broken down to different organizational units such as operating theater, ward, radiology, emergency department or other patient service centers.
- Modeling technique: This category refers to the modeling approach such as quadratic assignment problems (*QAP*), mixed integer programs (*MIP*) or discrete-event simulation models (*DES*).
- Solution technique: In this category it is differentiated, for example, between optimal solution approaches, heuristics or process analysis.
- Objectives: This category distinguishes between general facility design aspects, patient or resource centered objectives, amongst others.

For each of the categories of the developed taxonomy a table is built which gives an overview on the retrieved research papers (see Tables 5.1–5.4).

Table 5.1 Scope of the reviewed papers

	Review	Micro level							Macro level			
		General	Corridor allocation problem	Emergency department	Gynecology/ Obstetric clinics	Laboratory	Wards/ Inpatient units	Operating theater	Patient service center/ Ambulatory care/Outpatient clinics	General	Hospital/ health care setting	Other settings
Acar et al. (2009)												x
Amaral (2012)			x									
Amladi (1984)									x			
Arnolds and Nickel (2013)							x					
Arnolds and Nickel (2013a)											x	
Arnolds and Nickel (2013b)											x	
Ashby et al. (2008)							x					
Assem et al. (2012)								x				
Baumgart et al. (2009)								x				
Becker and Parsons (2007)											x	
Borzo (1992)											x	
Boucherie et al. (2012)											x	
Bromley (2012)											x	
Burn (1982)									x			

Table 5.1 (continued)

	Review	Micro level							Macro level			
		General	Corridor allocation problem	Emergency department	Gynecology/ Obstetric clinics	Laboratory	Wards/ Inpatient units	Operating theater	Patient service center/ Ambulatory care/Outpatient clinics	General	Hospital/ health care setting	Other settings
Butler et al. (1992)											x	
Butler et al. (1992a)											x	
Ceglowski et al. (2005)				x								
Choudhary et al. (2010)							x					
Elshafei (1977)											x	
Fetter and Thompson (1965)											x	
Forsberg et al. (2011)	x									x	x	
Freeman et al. (1974)	x	x								x		
Gibson (2007)											x	
Grigg et al. (2009)							x	x				
Günal and Pidd (2010)	x										x	
Hahn and Krarup (2001)											x	
Halpern et al. (2012)				x								

Table 5.1 (continued)

	Review	Micro level							Macro level			
		General	Corridor allocation problem	Emergency department	Gynecology/ Obstetric clinics	Laboratory	Wards/ Inpatient units	Operating theater	Patient service center/ Ambulatory care/Outpatient clinics	General	Hospital/ health care setting	Other settings
Hamacher et al. (2002)											x	
Hancock et al. (1978)											x	
Hassan and Tucker (2010)											x	
Hassan and Tucker (2010a)											x	
Ierardo et al. (2008)							x					
Iskander and Carter (1991)									x			
Jiang and Hu (2012)											x	
Jun et al. (1999)	x										x	
Kiechle et al. (2005)						x						
Lee et al. (2012)											x	
Leung (1992)												x
Levy et al. (1989)												
Lin et al. (2015)								x				
Mahachek and Knabe (1984)					x						x	

Table 5.1 (continued)

	Review	Micro level							Macro level			
		General	Corridor allocation problem	Emergency department	Gynecology/ Obstetric clinics	Laboratory	Wards/ Inpatient units	Operating theater	Patient service center/ Ambulatory care/Outpatient clinics	General	Hospital/ health care setting	Other settings
Molyneux (2010)	x			x								
Murali (1988)								x				
Nassar (2010)											x	
Nickel et al. (2000)											x	
noAuthor (1975)	x											
Pagell and Melnyk (2004)												x
Seelye (1982)	x						x					
Sepulveda et al. (1999)									x			
Swisher et al. (2000)									x		x	
Thompson and Goldin (1975)	x						x					
Thorwarth and Arisha (2012)				x								
Vos et al. (2007)											x	
Walus et al. (1997)	x										x	
Weiss et al. (2002)								x				
Yeh (2006)											x	

Table. 5.2 Modeling techniques of the reviewed papers

	Simulation	Discrete event simulation	Quadratic assign-ment program	Multicriteria QAP	Stochastic QAP	Subset QAP	Pedestrian flow modeling	Process modeling	Mixed integer program (MIP)	Binary linear program (BLP)	Integer linear program (ILP)	Linear model	Graph-theoretical model	Hopfield neural network	Quadratic integer goal program
Acar et al. (2009)		x							x						
Amaral (2012)									x						
Amladi (1984)		x													
Arnolds and Nickel (2013)										x	x				
Arnolds and Nickel (2013a)		x	x												
Arnolds and Nickel (2013b)		x	x												
Ashby et al. (2008)		x													
Assem et al. (2012)													x		
Baumgart et al. (2009)								x							
Boucherie et al. (2012)		x													
Butler et al. (1992)	x														x
Butler et al. (1992a)	x		x												
Choudhary et al. (2010)												x			
Elshafei (1977)			x												
Fetter and Thompson (1965)		x													
Forsberg (2011)		x													
Gibson (2007)		x													
Günal and Pidd (2010)		x													
Hahn and Krarup (2001)			x												

Table. 5.2 (continued)

	Simulation	Discrete event simulation	Quadratic assign-ment program	Multicriteria QAP	Stochastic QAP	Subset QAP	Pedestrian flow modeling	Process modeling	Mixed integer program (MIP)	Binary linear program (BLP)	Integer linear program (ILP)	Linear model	Graph-theoretical model	Hopfield neural network	Quadratic integer goal program
Hamacher et al. (2002)				x	x										
Hancock et al. (1978)	x														
Hassan and Tucker (2010)	x						x								
Hassan and Tucker (2010a)							x								
Iskander and Carter (1991)	x														
Jiang and Hu (2012)						x									
Jun et al. (1999)		x													
Lee et al. (2012)							x								
Leung (1992)													x		
Levy et al. (1989)	x														
Mahachek and Knabe (1984)	x														
Nassar (2010)							x								
Nickel et al. (2000)			x												
Pagell and Melnyk (2004)	x														
Sepulveda et al. (1999)	x														
Swisher et al. (2000)		x													
Thorwarth and Arisha (2012)		x						x							
Vos et al. (2007)		x													
Yeh (2006)														x	

Table. 5.3 Solution techniques of the reviewed papers

	Simulation	Optimization/ Exact algorithm	Simulation-optimization	General	Simulated annealing	Hill climbing	Genetic algorithm	Evidence based design/ Experience	Process analysis	Questionnaire/ Survey/Interview	Graph-theoretical approach	Systematic layout planning (SLP)	Fuzzy constraint theory	Spiral technique	Social network analysis	Fractional factorial design
Acar et al. (2009)			x													
Amaral (2012)		x		x												
Amladi (1984)	x															
Arnolds and Nickel (2013)		x														
Arnolds and Nickel (2013a)			x													
Arnolds and Nickel (2013b)			x													
Ashby et al. (2008)	x															
Assem et al. (2012)											x			x		
Baumgart et al. (2009)									x						x	
Becker and Parsons (2007)								x								
Borzo (1992)								x								
Boucherie et al. (2012)	x		x													

Table. 5.3 (continued)

	Simulation	Optimization/ Exact algorithm	Simulation-optimization	General	Simulated annealing	Hill climbing	Genetic algorithm	Evidence based design/ Experience	Process analysis	Questionnaire/ Survey/Interview	Graph-theoretical approach	Systematic layout planning (SLP)	Fuzzy constraint theory	Spiral technique	Social network analysis	Fractional factorial design
Bromley (2012)				x						x						
Burn (1982)								x								
Butler et al. (1992)			x													
Butler et al. (1992a)			x													
Ceglowski et al. (2005)									x							
Elshafei (1977)				x												
Fetter and Thompson (1965)	x															
Forsberg (2011)	x															
Francis et al. (1992)											x					
Gibson (2007)	x															
Grigg et al. (2009)									x							
Hahn and Krarup (2001)				x												
Halpern et al. (2012)								x								

Table. 5.3 (continued)

	Simulation	Optimization/ Exact algorithm	Simulation-optimization	General	Simulated annealing	Hill climbing	Genetic algorithm	Evidence based design/ Experience	Process analysis	Questionnaire/ Survey/Interview	Graph-theoretical approach	Systematic layout planning (SLP)	Fuzzy constraint theory	Spiral technique	Social network analysis	Fractional factorial design
Hamacher et al. (2002)		x														
Hancock et al. (1978)	x															
Hassan and Tucker (2010)	x			x	x	x										
Hassan and Tucker (2010a)				x	x	x	x									
Ierardo et al. (2008)										x						
Iskander and Carter (1991)	x															
Jiang and Hu (2012)				x												
Jun et al. (1999)	x		x													
Kiechle et al. (2005)																
Lee et al. (2012)	x			x												
Leung (1992)				x							x					

Table. 5.3 (continued)

	Simulation	Optimization/ Exact algorithm	Simulation-optimization	General	Simulated annealing	Hill climbing	Genetic algorithm	Evidence based design/ Experience	Process analysis	Questionnaire/ Survey/Interview	Graph-theoretical approach	Systematic layout planning (SLP)	Fuzzy constraint theory	Spiral technique	Social network analysis	Fractional factorial design
Levy et al. (1989)	x															
Lin et al. (2015)												x	x			
Mahachek and Knabe (1984)	x															
Nassar (2010)		x														
Nickel et al. (2000)		x														
Pagell and Melnyk (2004)	x								x							
Sepulveda et al. (1999)	x															
Swisher et al. (2000)	x															x
Thompson and Goldin (1975)								x								
Thorwarth and Arisha (2012)	x															
Vos et al. (2007)	x															
Weiss et al. (2002)								x								
Yeh (2006)				x	x											

Table. 5.4 Objectives of the reviewed papers

	Patient centered	Personnel centered	Resource utilization	Occupancy rate	Facility design	Distance minimization	Workflow optimization/Efficiency	Pedestrian/ Occupant flows	Cost minimization	Closeness rating maximization
Acar et al. (2009)						x				
Amaral (2012)									x	
Amladi (1984)	x				x					
Arnolds and Nickel (2013)	x								x	
Arnolds and Nickel (2013a)	x	x			x	x				
Arnolds and Nickel (2013b)	x	x			x					
Ashby et al. (2008)	x				x					
Assem et al. (2012)										x
Becker and Parsons (2007)	x	x								
Borzo (1992)	x	x								
Boucherie et al. (2012)	x	x			x					
Bromley (2012)	x									
Burn (1982)	x									
Butler et al. (1992)					x	x				
Butler et al. (1992a)				x		x				
Ceglowski et al. (2005)			x		x					
Choudhary et al. (2010)		x								
Elshafei (1977)	x					x				

Table. 5.4 (continued)

	Patient centered	Personnel centered	Resource utilization	Occupancy rate	Facility design	Distance minimization	Workflow optimization/Efficiency	Pedestrian/ Occupant flows	Cost minimization	Closeness rating maximization
Fetter and Thompson (1965)					x		x			
Gibson (2007)	x	x			x					
Grigg et al. (2009)		x			x					
Hahn and Krarup (2001)						x				
Halpern et al. (2012)	x	x			x					
Hamacher et al. (2002)	x	x				x				
Hancock et al. (1978)					x					
Hassan and Tucker (2010)								x		
Hassan and Tucker (2010a)								x		
Ierardo et al. (2008)	x									
Iskander and Carter (1991)					x					
Jiang and Hu (2012)						x				
Jun et al. (1999)	x								x	
Kiechle et al. (2005)							x			
Lee et al. (2012)								x		
Leung (1992)										x
Levy et al. (1989)	x		x		x					
Lin et al. (2015)										x
Mahachek and Knabe (1984)	x	x			x					
Molyneux (2010)	x									

Table. 5.4 (continued)

	Patient centered	Personnel centered	Resource utilization	Occupancy rate	Facility design	Distance minimization	Workflow optimization/Efficiency	Pedestrian/ Occupant flows	Cost minimization	Closeness rating maximization
Murali (1988)		x	x						x	
Nassar (2010)								x		
Nickel et al. (2000)	x	x				x				
Pagell and Melnyk (2004)							x			
Sepulveda et al. (1999)	x		x							
Swisher et al. (2000)	x									
Thompson and Goldin (1975)	x						x			
Thorwarth and Arisha (2012)	x	x								
Vos et al. (2007)	x	x					x	x		
Weiss et al. (2002)							x			
Yeh (2006)					x					

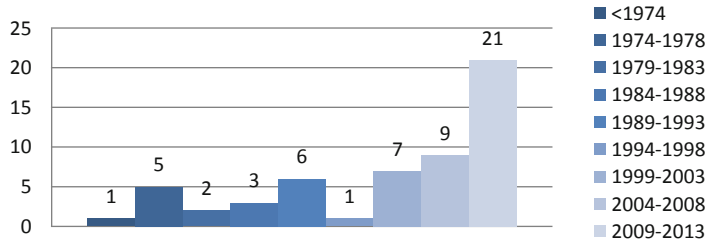


Fig. 5.1 Number of published papers per 5-year interval

Figure 5.1 shows the statistics with respect to the number of published papers per 5-year interval from 1965 to 2013. Obviously, the topic gets more and more attention within the scientific community. This could particularly be observed during the last 5 years where the number of published papers more than doubled.

5.3 A Graph-Theoretical Layout Planning Approach Applied to a Major German Hospital

As already mentioned, hospital buildings still are mainly planned by architects. They are experts in the field of design, usually have experience from other hospital planning projects and know the relevant legal regulations with respect to hospital buildings. Furthermore, there exist first example projects where architects take into account the processes, i.e., patient, personnel and material flows, when planning a hospital. But, lack of knowledge can still be detected in some cases and, consequently, no application of decision supporting operations research methodologies such as optimization or simulation techniques. Nevertheless, in the last years the authors have had the experience that responsible hospital planners are becoming more and more interested and open minded towards such kind of decision supporting operations research methodologies. Last but not least, this could be a result of the increasing financial pressure on hospitals.

In what follows, a project is detailed where the authors applied a graph-theoretical heuristic to support developing a layout plan for a new building of a major German hospital.

5.3.1 The Cooperating Hospital

The project was initiated by the Institute of Operations Research of the Karlsruhe Institute of Technology (http://dol.ior.kit.edu) and a major German community hospital with more than 1500 beds. Over the years, the hospital has grown on an area of 155,758 m^2 and currently comprises 23 buildings. The historical development of

the different buildings results in long travel distances and times for both patients and personnel. This impedes process efficiency and, thus, treatment quality as well as economic efficiency. At the same time, the technical infrastructure of the existing buildings is not expandable such that, for example, the requirements for the installation of medical technology cannot be covered anymore. Summarizing, there is a need for a new building which is also manifested in the target planning program of the hospital for the upcoming years. According to the hospital's plan, the new building shall comprise the following organizational units:

- Wards of different intensities of care.
- Interdisciplinary inpatient and outpatient operating theaters.
- Walk-in clinics for different disciplines.

5.3.2 Goal and Methodology

For the hospital, the objective of the layout planning project was to reduce the long travel distances and times for patients and personnel which can be observed in the current setting. This can be achieved by an efficient planning of the location of the organizational units inside the new building according to the patients' clinical pathways and logistic processes that will take place during daily routine.

Three main characteristics of the given layout planning problem have to be considered when deciding for an appropriate modeling and solution technique: Firstly, the high number of wards and functional departments to be located; secondly, the different sizes of the organizational units, and, thirdly, the necessity to plan a multi-floor building. These characteristics make the problem too complex to find an optimal solution by formulating and solving a mathematical model, as, for example, a quadratic assignment or a mixed integer program. Moreover, these models are rather appropriate if organizational units have to be (re-)located within a given structure of an existing building with defined dimensions (length and width of each level of the building). Contrarily, in this project there still were some degrees of freedom with respect to the architectural dimensions since a completely new building had to be developed where only the dimensions of the ground level were fixed beforehand. To solve this kind of layout problems, a graph theoretical approach was developed (see, for example, Francis et al. 1992). Particularly, an advantage of this procedure is that it is illustrative and visually presentable which makes it easier to convince the responsible hospital managers to apply the approach.

5.3.3 A Graph-Theoretical Approach for Layout Planning

The graph-theoretical approach for layout planning and its theoretical background is thoroughly described by Francis et al. (1992). In this section a short overview of

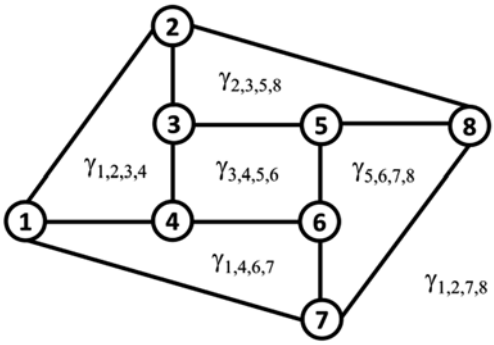


Fig. 5.2 Planar graph with five inner and one outer facet

the procedure and the underlying theory is given. For more detailed information the interested reader is referred to Francis et al. (1992). As a basis for the application an interaction or flow matrix is needed, where each entry represents the number of interactions between two organizational units, for example, during 1 year. This number can also be interpreted as the importance of locating two organizational units close to each other. The higher the interaction between two organizational units, the higher is the importance of a direct adjacency of the corresponding units. Organizational units can be, for example, machines in a manufacturing environment, offices in any service setting or wards and functional departments in a hospital. The idea of the approach is to construct a graph in the first step and then to derive a block layout from its dual graph in the second step. The objective is to maximize the sum of the interactions of organizational units that are adjacent in the resulting layout plan.

In the graph to be constructed each node represents one organizational unit. An edge between two nodes means that the represented units are adjacent in the resulting block layout. In order to derive a layout from the graph the latter has to be planar. A graph is called planar if it can be drawn on a plane without any crossover of edges (see Fig. 5.2). Furthermore, a planar graph is called a maximally planar graph if and only if the characteristic of planarity gets lost when adding a further edge (see Fig. 5.3).

By depicting a graph, the plane is subdivided in facets γ, i.e., a set of one or more inner facets and one outer facet. An inner facet is an area in the plane bounded by nodes and edges where, first, the bounding nodes and edges form an elementary circle (consisting only of different nodes and edges), second, a pairwise intersection of two facets is empty, and, third, no subset of a facet features the former two characteristics. In Fig. 5.2 the following areas are inner facets: $\gamma_{1,2,3,4}$, $\gamma_{1,4,6,7}$, $\gamma_{2,3,5,8}$, $\gamma_{3,4,5,6}$ and $\gamma_{5,6,7,8}$. An outer facet is the area in the plane which is not covered by inner facets. In Fig. 5.2 the outer facet is $\gamma_{1,2,7,8}$.

A facet which is bounded by the three nodes *i, j* and *k* and the three edges [*i, j*], [*j, k*], and [*k, i*] is called a triangle $\delta_{i,j,k}$ (see Fig. 5.3). If all facets of the graph are

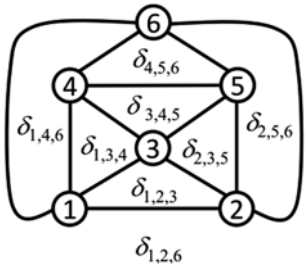


Fig. 5.3 Maximally planar graph (deltahedron) consisting of eight triangles (seven *inner* and one *outer* facet)

triangles it is called a deltahedron. A planar graph is maximal if and only if it is a deltahedron.

A graph is connected if there is a walk between every pair of vertices. A walk in a graph is an alternating sequence of vertices and edges $W = v_0, e_1, v_1, ..., e_n, v_n$ such that for $j = 1, \ldots, n$ the vertices v_{j-1} and v_j are the endpoints of edge e_j. A simple graph is a graph that has no self-loops or multi-edges. A simple graph is a complete graph if every pair of vertices is joined by an edge.

If a planar graph representing the adjacency requirements of several organizational units can be constructed, then it is possible to derive a compatible block layout using its dual graph. Given a connected, undirected and planar graph, then the dual graph is built as follows:

- The dual graph includes exactly one node for each facet of the primal graph.
- The dual graph includes exactly one edge for each edge in the primal graph that separates two facets. In the dual graph, this edge connects the two nodes that represent these two facets in the primal graph.

To solve the layout problem by applying the graph-theoretical approach, first, a planar graph has to be determined in which the sum of the edge weights, which represent the corresponding entries of the interaction matrix, is maximized. Thus, the search for such a graph can be limited to maximally planar graphs and, consequently, to deltahedrons.

If a graph is a deltahedron with n nodes and m edges, then the following relation holds: $m = 3n - 6$ (see Francis et al. 1992). Using this relation, an upper bound for the sum of the edge weights of a layout problem can be derived by simply summing up the $3n - 6$ highest values of the interaction matrix.

Given an interaction matrix $\mathbf{U} = (u_{ij})$ and a simple, complete and undirected graph $G = [V, E, \mathbf{U}]$ with nodes $v \in V$ and edges $e \in E$ which are weighted with the values in U. The problem to identify a maximally planar subgraph $G' = \left[V, E', U\right]$ of G with $E' \subset E$, which has the highest sum of edge weights, can be formulated as follows:

Decision variables:

$$x_{ij} = \begin{cases} 1 & \text{if edge } \left[i, j\right] \in E' \\ & \qquad \forall i, j \text{ with } i < j \\ 0 & \text{else} \end{cases}$$

Model:

$$\begin{aligned} \text{Max} \quad & \sum_{i=1}^{n} \sum_{j=1}^{n} u_{ij} x_{ij} \\ \text{s.t.} \quad & G' = [V, E'] \text{ is a maximal planar graph} \\ & x_{ij} \in \{0, 1\} \forall i, j \text{ with } i < j. \end{aligned}$$

Since the problem is **NP**-hard, large instances cannot be solved to optimality. Thus, the following heuristic procedure was developed to construct a maximally planar graph (see Leung 1992).

- Prerequisite: Calculate the row sums of the interaction matrix and sort them according to monotonic decreasing values.
- Initialization: Build an initial deltahedron using the four nodes with the highest row sums.
- Iterations: Integrate the remaining nodes into the deltahedron in the order of decreasing row sums. Always include them into the triangle where the objective function value is increased the most. Here, including means connecting the new node with the three existing nodes of the triangle. Thus, the new node is connected with the nodes of that triangle with which it has the most interactions.

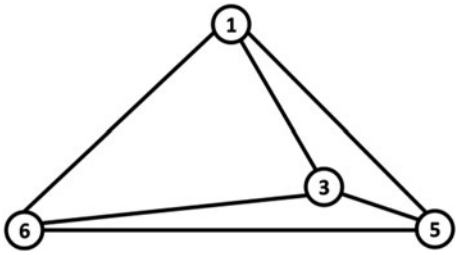
Regarding the quality of this heuristic, Leung (1992) showed that the sum of the edge weights of the obtained graphs using this heuristic lie between 92.4 and 99.8 % of the upper bound on the optimal value generated by summing up the weight of the $(3n - 6)$ edges of maximum weight.

Example 5.1. Given the following interaction matrix $\mathbf{U} = (u_{ij})$ of a complete graph with six nodes $v \in V$ and row sums u_i:

$$\mathbf{U} = \begin{pmatrix} - & 4 & 7 & 1 & 7 & 10 \\ 4 & - & 2 & 5 & 2 & 4 \\ 7 & 2 & - & 8 & 8 & 1 \\ 1 & 5 & 8 & - & 1 & 3 \\ 7 & 2 & 8 & 1 & - & 5 \\ 10 & 4 & 1 & 3 & 5 & - \end{pmatrix}, \; u_i = \begin{pmatrix} 29 \\ 17 \\ 26 \\ 18 \\ 23 \\ 23 \end{pmatrix}$$

Initialization

- Nodes with highest row sums: $V' = \{1,3, 5,6\}$
- Corresponding edges: $E' = \{[1,3], [1,5], [1,6], [3,5], [3,6], [5,6]\}$
- Corresponding triangles: $\Delta = \{\delta_{135}, \delta_{136}, \delta_{156}, \delta_{356}\}$
- Objective function value: $Z = 7 + 7 + 10 + 8 + 1 + 5 = 38$
- Resulting graph (Note: In order to construct the initial deltahedron, one node has to be chosen to be put in the center, here node 3 is chosen.):



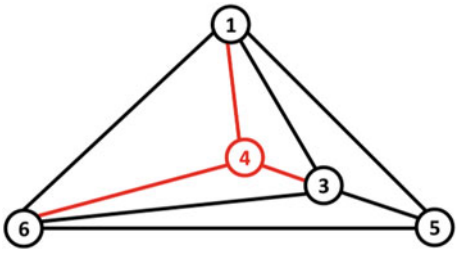
Iteration $it = 1$

- Choose $v = 4$, because $u_4 = \max\{u_2 = 17, u_4 = 18\}$

it	1
v	4
δ_{135}	$1 + 8 + 1 = 10$
δ_{136}	$1 + 8 + 3 = 12^*$
δ_{156}	$1 + 1 + 3 = 5$
δ_{356}	$8 + 1 + 3 = 12$

Note: The table shows the values representing the increase of the objective function value when integrating node 4 in the existing triangles. The highest increase is marked with an asterisk, ties are broken arbitrarily.

- Resulting graph:

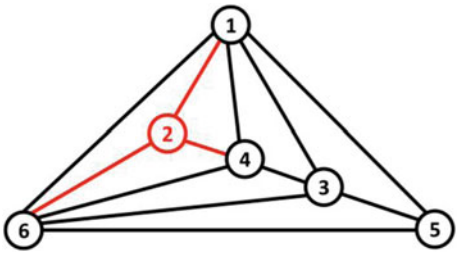


Iteration $it = 2$

- Choose the last remaining node $v = 2$

it	1	2
v	4	2
δ_{135}	10	8
δ_{136}	12*	–
δ_{156}	5	10
δ_{356}	12	8
δ_{134}		11
δ_{146}		13*
δ_{346}		11

- Resulting graph:



If all nodes are inserted the objective function value can be calculated by summing up the edge weights (see Fig. 5.4): $Z = 63$

Before constructing the dual graph and deriving a block layout, an additional node has to be integrated into the maximally planar graph. This node represents the environment of the new building and, thus, the origin of all flows before entering the building. Constructing the block layout from the dual graph without this additional node would result in one organizational unit located outside the building. In Fig. 5.5, node (7) represents the environment. This new node has to be connected to all the nodes that formerly established the outer facet of the maximally planar graph.

When constructing the dual graph it has to be ensured that the node representing the environment lies outside the dual graph (see Fig. 5.6).

One possible block layout to be derived from the dual graph is shown in Fig. 5.7. Note that no area restrictions are given for the organizational units.

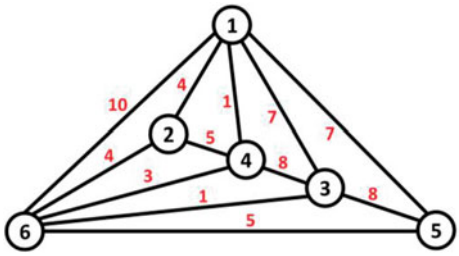


Fig. 5.4 Constructed graph with edge weights

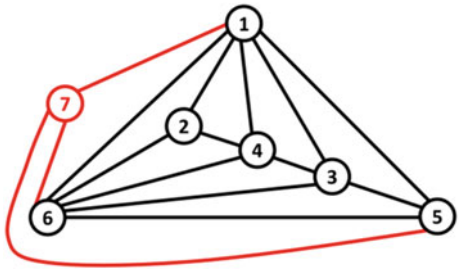


Fig. 5.5 Additional node (7) for the environment

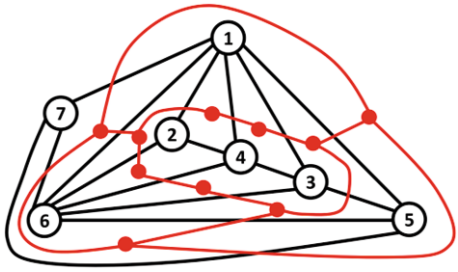


Fig. 5.6 Construction of the dual graph (*red*)

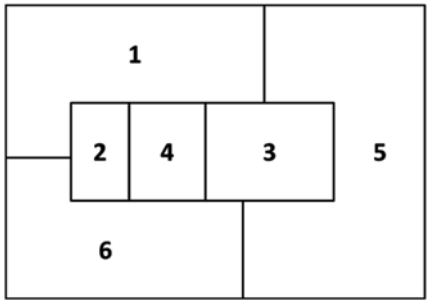


Fig. 5.7 Block layout derived from the dual graph

5.3.4 Application of the Graph-Theoretical Heuristic to Construct a Hospital Layout

The hospital layout planning problem instance addressed in this chapter consists of 25 organizational units to be located (see Table 5.5). Each organizational unit comprises a set of rooms. In accordance with the hospital management, the rooms were grouped to organizational units since for the total of more than 700 rooms it would not have been possible to reliably derive an interaction matrix from the available data sources. By applying the graph-theoretical heuristic a two-dimensional block layout in the plane, i.e., representing one level, can be constructed. Thus, in the last step the organizational units have to be assigned manually to the different floors of the new building.

5.3.4.1 Data

In order to derive a layout by applying the graph-theoretical heuristic, the data described in the following sections has to be obtained.

5.3.4.1.1 Area of each Organizational Unit

For planning purposes the hospital developed a so-called space allocation plan for the new building. This plan details the kind and area of each organizational unit. Table 5.5 gives an overview of the space allocation plan.

5.3.4.1.2 Further Restrictions Regarding the Layout

Before applying the described approach, the hospital management already had defined the following basics for the new building: The building shall have six levels where with each level the floor area decreases. The ground level shall have a dimension of 214 m of length and 54 m of width, including aisles, waiting areas, technical equipment areas, elevators, stairs etc. For the remaining levels of the building the area had not been fixed yet.

5.3.4.1.3 Interaction Matrices

The data that was delivered from the hospital to infer the interaction matrix for the organizational units was the hospital's specific case mix:

- Number of cases per discipline, for example, eye clinic, dermatology or women's clinic.
- Number of surgery cases per discipline, for example, trauma surgery or children's surgery.

Table 5.5 Space allocation plan

	Organizational unit	Area [m^2]
1	Emergency department (ED)	312
2	Medical services (MS)	336
3	Outpatient clinic for anesthesiology (A)	304
4	Outpatient clinic for eyes (OC1)	304
5	Outpatient clinic for abdominal surgery (OC2)	224
6	Outpatient clinic for vascular surgery (OC3)	280
7	Outpatient clinic for otolaryngology (OC4)	680
8	Outpatient clinic for oral and maxillofacial surgery (OC5)	352
9	Outpatient clinic for urology (OC6)	272
10	Outpatient cancer center (OC7)	152
11	Inpatient surgery unit (IS)	2720
12	Outpatient surgery unit (OS)	272
13	Standby rooms (SR)	192
14	General care unit 1 (GCU1)	1064
15	General care unit 2 (GCU2)	1064
16	General care unit 3 (GCU3)	1064
17	General care unit 4 (GCU4)	1064
18	Intermediate care unit 1 (IMC1)	1064
19	Intermediate care unit 2 (IMC2)	1064
20	Intensive care unit 1 (ICU1)	794
21	Intensive care unit 2 (ICU2)	794
22	Intensive care unit 3 (ICU3)	794
23	Intensive care unit 4 (ICU4)	794
24	Sterile goods supply (SG)	1000
25	Medical device supply (MD)	96

The assumptions which had to be made in order to derive the quantitative interactions between all pairs of organizational units from the case mix data are specified in the following section. Since no detailed information on quantitative flows was available directly, it had to be derived by making these assumptions. Furthermore, patient, personnel and material flows were not weighted differently such that the entries in the interaction matrix equal the sums of the different kinds of flows.

5.3.4.2 Assumptions

As it is usual in practical applications, not all data which is needed to apply the approach was available directly. This section deals with the assumptions that had to be made in order to derive quantitative information regarding the general, surgical and emergency patient flows as well as the sterile goods and medical device supply and the flows to and from the standby rooms for physicians. Based on these assumptions and (derived) data the interaction matrix was then set up. Since this is sensitive data for the hospital management, the absolute values, i.e., the entries of the interactions matrix, cannot be shown here.

The section on general patient flow refers to assumptions relevant for all in- and outpatients. Also, the surgical and emergency patient flows are not necessarily excluding each other, that means, an emergency patient may, for example, have to undergo a surgery. In most of the cases only patient flows were considered as these could be derived more reliably from the available data than the usually highly variable personnel and material flows. Nevertheless, for some organizational units, where patients do not have access to, for example, the central sterilization unit or standby rooms for physicians, relevant personnel and material flows were inferred from expert interviews.

Furthermore, there is assumed a 15 % increase on the number of patients per year based on the current numbers. This rise was stipulated by the hospital management.

5.3.4.2.1 General Patient Flow

Regarding the general patient flow, the first contact point inside the hospital for all outpatients is the organizational unit called “medical services”. In contrast, all inpatients are directly guided to their assigned ward. Outpatients leave the hospital from the corresponding outpatient clinics and inpatients from the discharging ward.

From the case mix data an average length of stay of 6 days from admission to discharge was calculated and taken as a basis for all patients and intensities of care. During their treatment process, patients recover and are transferred to less intense care units. It is assumed that due to reconvalescence 50 % of the intensive care patients are directly transferred to a general care unit, whereas the remaining 50 % have to be transferred to an intermediate care unit first before being moved to a general care unit later. The transfers take place in equal shares to each of the four general care and two intermediate care units, respectively. Furthermore, patients are only discharged from general care units.

5.3.4.2.2 Surgical Patient Flow

Regarding the surgical patient flow, it is divided between inpatient (59 %) and outpatient (41 %) surgeries, with a current total of about 21,600 surgeries per year. It is assumed that each outpatient is operated once. Contrarily, inpatients are at least

operated once during their hospital stay, and 10 % have to undergo a second surgery. Patients may come from each of the care units or from the emergency department. After surgery, the patients are transferred back to the care unit where they came from or, in case of the emergency patients, are assigned to one of the different care units in equal shares.

Before surgery, each outpatient has to sign a consent form and is given a surgical clearance in the clinic for anesthesiology. Since the clearance has to be completed at least 1 day before surgery the patients leave the hospital afterwards and return on the day of their surgery.

5.3.4.2.3 Emergency Patient Flow

About 40 % of all arriving patients are emergency patients. Less than 1 % of those emergency patients have to be transferred to one of the intensive care units immediately. From the remaining emergency patients 20 % are assigned to one of the outpatient clinics and 80 % have to be operated immediately.

5.3.4.2.4 Sterile Goods Supply

A sterile good unit is a defined unit of volume used for sterilization of materials in an autoclave. Sterile good units are needed for each surgery. For the transportation of sterile goods between the operating theaters and the unit for sterile goods supply the personnel uses trolleys with a capacity of 12 sterile good units. From the case mix data it is known that 59 % of the trolleys have to be transported to/from the inpatient operating theater and 41 % to/from the outpatient operating theater.

5.3.4.2.5 Medical Device Supply

The unit for medical device supply prepares all devices which are needed in the operating theaters. These are, for example, respiration machines or medical appliances needed for treatment or examination during a surgery. Again, 59 % are transported to/from the inpatient operating theater and 41 % to/from the outpatient operating theater.

Furthermore, in the intensive care units artificial respiration devices are used that have to be cleaned and sterilized for new patients. These devices have also to be transported between the intensive care units and the unit for medical device supply. It is assumed, that each intensive care unit needs the same amount of respiration devices.

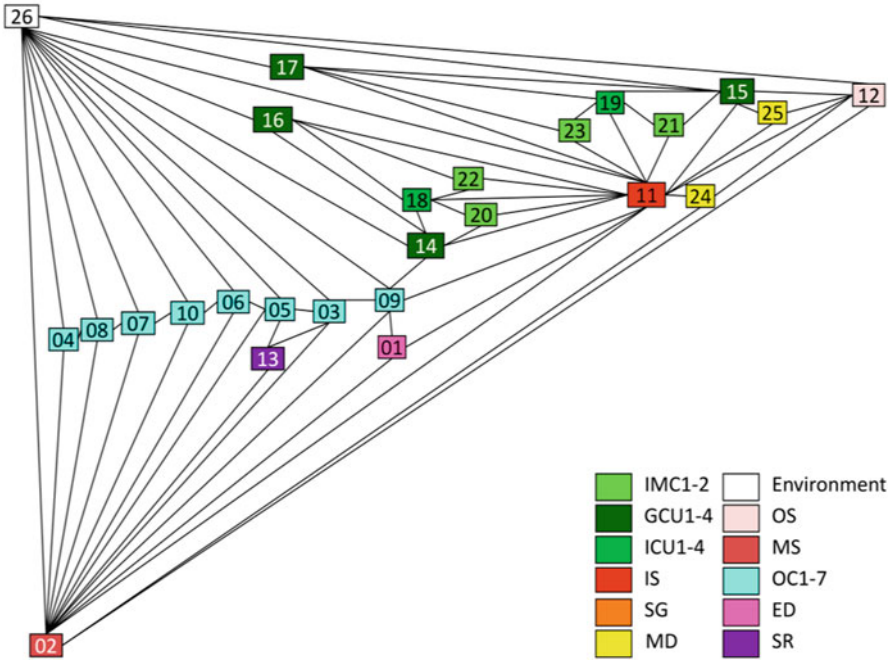


Fig. 5.8 Primal graph

5.3.4.2.6 Standby Rooms for Physicians

According to the medical personnel's information, it can be assumed that each night each of the 13 physicians on standby is called twice on average. The physician then either has to go to the emergency department or to one of the care units. After the patient's treatment the physician usually returns to the standby room.

5.3.4.3 Results

The solution of the implemented graph-theoretical heuristic procedure shows how to construct the resulting maximally planar graph. This primal graph is shown in Fig. 5.8 with a color code that highlights the different groups of organizational units. It can be observed that the inpatient surgical unit (11) is connected to the other organizational units with the highest number of edges or, mathematically spoken, it is the node with the highest degree.

The environment of the building is represented by node (26). The heuristic has to be slightly adapted in order to guarantee that this node is part of the outer facet in the resulting graph. Since all patients enter the hospital building from the environment this node is one of those chosen in the initialization step of the heuristic. Regarding

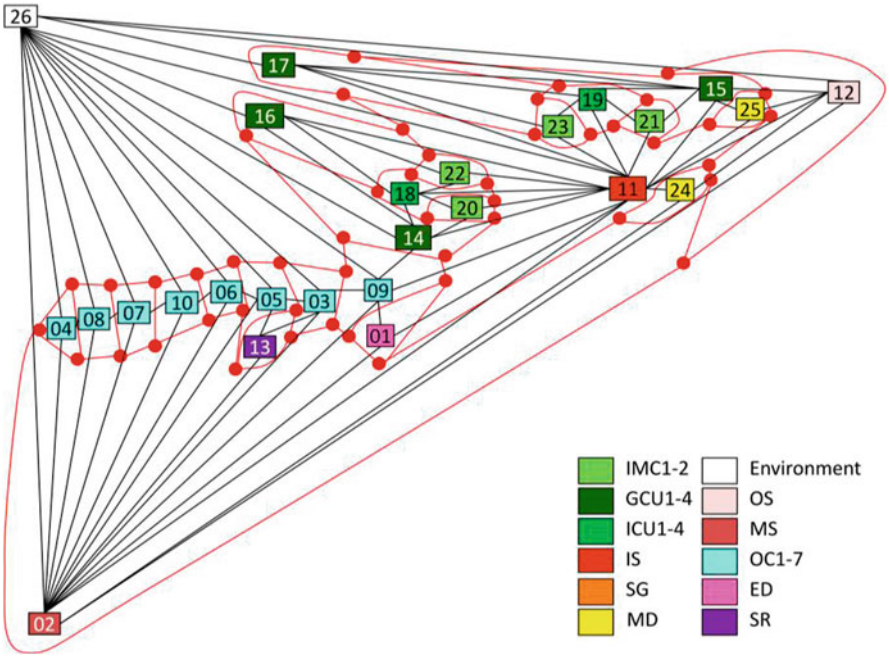


Fig. 5.9 Construction of the dual graph

the further steps, it then has to be ensured that new nodes are only inserted into one of the existing inner facets.

Figure 5.9 shows how to construct the dual graph that is needed to derive the single-floor block layout. Within each facet of the primal graph a new node is located which is colored red. The red nodes are connected such that each new (red) edge cuts one edge of the primal graph. Node (26) which represents the environment is not included in any of the inner facets because, obviously, it has to be outside the building.

After removing the primal graph from Fig. 5.9 a cluster of outpatient clinics can be recognized as well as two care clusters, each consisting of one intensive care unit, two intermediate care units and two general care units (see Fig. 5.10). The inpatient surgical unit (11) has a quite central position and the medical service unit (02) which is the first contact point for all outpatients is adjacent to the environment where all the patients come from.

In the next step, the edges of the dual graph are rectified. The result is a single-floor block layout as depicted in Fig. 5.11. Again, both the outpatient cluster as well as the two care clusters are clearly recognizable. Up to now, the real areas of the organizational units have not been taken into account yet. The sterile goods supply (24) and the medical device supply (25) units are close to the inpatient (11) and outpatient (12) surgical units. The emergency department (1) is located at a central

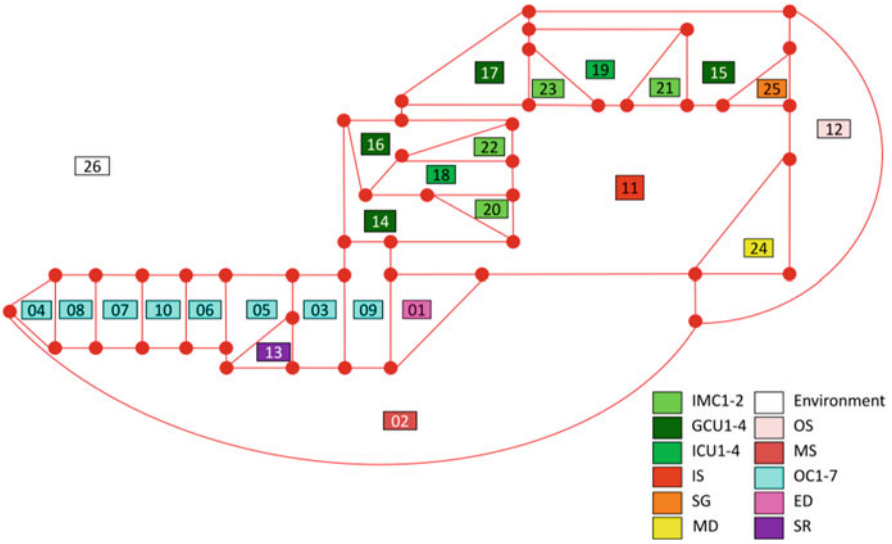


Fig. 5.10 Dual graph

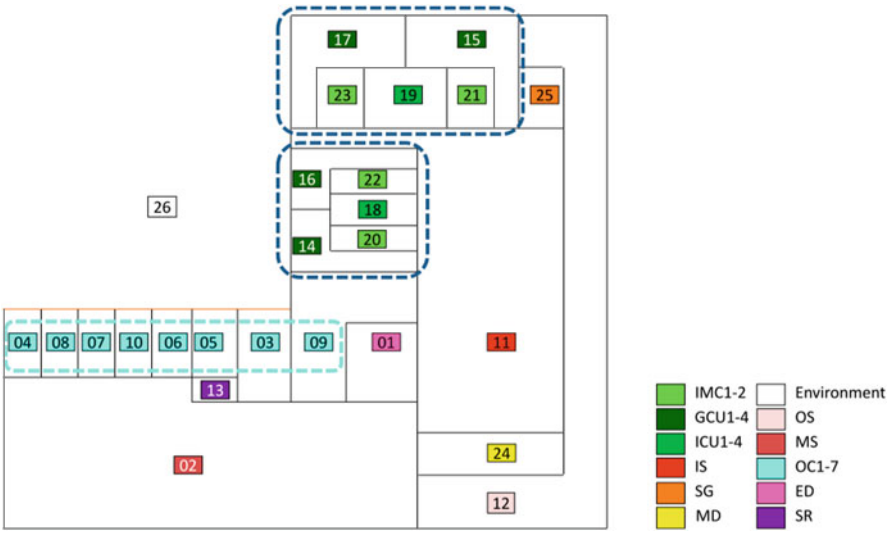


Fig. 5.11 Single-floor block layout without area restrictions

position next to the medical service for admission (2), the cluster of outpatient clinics (3)-(10) and the outpatient surgery unit (12). The standby rooms for physicians (13) are placed between the outpatient clinics for anesthesiology (3) and general and abdominal surgery (5) as well as the medical service (2).

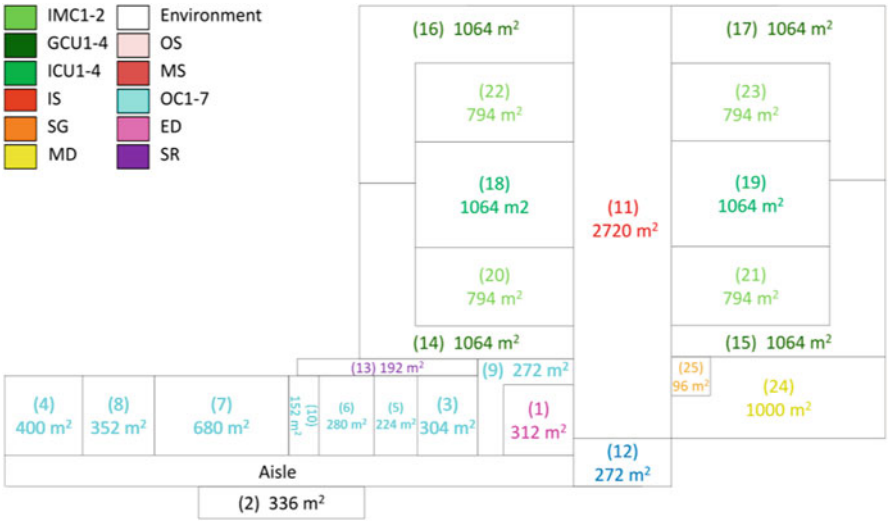


Fig. 5.12 Single-floor block layout with area restrictions

Considering the given areas for each organizational unit, the layout plan shown in Fig. 5.12 can be constructed. Still, this is a single-floor block layout, but now it considers the dimensions of each organizational unit. In the center, the inpatient surgical unit (11) is located. To its left and right the two care clusters are placed. The sterile goods (24) and medical device (25) supply units are adjacent to both the inpatient (11) and outpatient (12) surgical units. The emergency department (1) and the outpatient clinics (3)–(10) are located in the south-western part of the floor plan. The standby rooms are placed between the outpatient clinics for anesthesiology (3), abdominal surgery (5), vascular surgery (6), urology (9), the cancer center (10) and one of the intensive care units (14). As all outpatients first have to contact the medical service unit (2), from there they can easily reach the outpatient clinics via the aisle that has been incorporated additionally.

It can be observed that due to the given area restrictions of each organizational unit, some of the required adjacencies cannot be preserved anymore. But in accordance with the hospital management, the identified outpatient and care clusters are kept. Furthermore, the hospital management had already decided for a building with six levels such that the single-floor block layout had to be adapted accordingly. During this manual adaptation step, the desired width-length dimensions of the ground floor as well as the requirement that with an increasing level the floor areas shall decrease (see Sect. 5.3.4.1) have to be kept.

Figure 5.13 depicts the manually generated multi-floor layout plan where the outpatient and care clusters can still be identified. Level-01 contains the inpatient surgery unit (11), the sterile goods (24) and medical device (25) supply units as well as the emergency department (1).

Fig. 5.13 Multi-floor layout plan

All outpatient clinics (3)–(10) as well as the outpatient surgical unit (12), the doctor's standby rooms (13) and the first contact point for all elective patients, i.e., the medical services unit (2) are located on level 00.

Through the assignment of the emergency department (1) and the medical services unit (2) to different levels, the emergency and elective patient arrivals can be separated from each other so that overcrowding is prevented. Although the two surgery units for inpatients (11) and outpatients (12) are located on different levels, they could easily be connected via "sterile stairs." These are connecting stairs which only the personnel may use in order to move from one unit to the other without the necessity to, for example, change clothes due to hygienic requirements. Nevertheless, the assignment of personnel to either the inpatient or outpatient surgery

unit during one shift is usually fixed. Consequently, the usage of the "sterile stairs" would in most cases only be necessary at the beginning or end of a shift, during the rest period or due to personnel shortage.

Levels 01 and 02 each contain one care cluster consisting of one general, one intermediate, and one intensive care unit. Also, levels 03 and 04 each comprise one general and one intensive care unit. The arrangement in care clusters is advantageous regarding, on the one hand, the possibility of a transfer from one care level to another within the same specialty located on the same floor due to an improvement or deterioration of the patient's medical condition. On the other hand, transfers between the same care levels of different specialties which are located on different floors are very unusual such that vertical transports can be avoided. But, by locating the intensive care units on levels 03 and 04 directly above the intermediate care units on levels 01 and 02, the transfer of patients via an elevator between the intensive and the intermediate care unit is as easy and convenient as possible since it includes as less as possible horizontal movements. Furthermore, each care group can be directly connected to the inpatient surgical unit via elevators what makes the transport to and from the operating theater very comfortable.

5.3.4.4 Discussion

In this section the results of applying the graph-theoretical heuristic to the real-world hospital layout problem will be discussed thoroughly with respect to the assumptions, the multi-floor aspect, possible layout distortions, and the development of the layout plan. Although all the potential drawbacks of the presented approach have been addressed during the application process with common sense solutions there is still some potential for future research on a more methodological level.

5.3.4.4.1 Assumptions

As already detailed in Sect. 5.3.4.2, some assumptions had to be made to derive the necessary data for the graph-theoretical heuristic. As far as possible, the available data was used to set up assumptions, especially regarding realistic patient, personnel and material flows through the hospital. Here, the focus was laid on the patient flow because the available target planning program of the hospital, which served as data basis to plan the layout, contained no information on the personnel and material flows but only on the patients' case mix related data such as the number of cases or surgery numbers per discipline. Furthermore, on the one hand, the personnel usually belongs to and, consequently, moves within an organizational unit and, on the other hand, the material flow is highly variable and, therefore, hardly predictable. Obviously, this assumption regarding the input data influences the result of the graph-theoretical heuristic. Nevertheless, according to the hospital management the constructed layout shows a reasonable arrangement of organizational units which justifies the data-based assumptions. The results have not been compared

to another hospital due to the difficulty of finding a similar hospital with respect to patient flows which are the essential and critical input data for the proposed approach.

Furthermore, only data relevant for interactions between organizational units within the new building but not to and from other buildings on the hospital's campus were taken into account. Nevertheless, some organizational units which will not be moved into the new building as, for example, the radiology or endoscopy departments will have interactions with organizational units which will be located in the new building. These interactions are not taken into account since the patient and personnel flows within the building are assumed to be more important than the flows between buildings. The only effect which the flows between buildings could have on the solution would be a shift of organizational units inside the new building nearer to the entrance if they have high interaction rates with organizational units in the other buildings. But the distance within the new building has a much smaller share on the total distance than the distance to the corresponding organizational unit in another building anywhere on the campus. Obviously, the latter part of the distance cannot be influenced by solving the layout problem because the locations of all organizational units in any of the other buildings are fixed. This means that the total distances between organizational units in the existing and new buildings would only be influenced marginally by different layouts. Furthermore, the hospital management's focus was laid on the operational processes inside the new building and not between buildings.

5.3.4.4.2 Multi-Floor Layout

The layout generated by the graph-theoretical heuristic only comprises one level but due to the limited space available, the new building has to be a multi-floor building. Consequently, a manual adaptation is necessary which was justified by identifying different clusters in the single-floor layout. Nevertheless, some of the adjacencies had to be neglected when deriving the multi-floor layout plan from the single-floor layout plan. Obviously, in this step, other layout plans could also be identified but since expert opinions were included when deriving the clusters and the final layout the solution could be proven to be adequate. Nevertheless, future research aims at measuring the potential error of the graph-theoretical approach and developing a procedure to avoid high errors.

5.3.4.4.3 Layout Distortion

The graph-theoretical heuristic does not consider any area restrictions of the organizational units. In the presented application the areas of the organizational units differ with a factor of up to 1:10. Consequently, it has to be ensured manually that small organizational units are not stretched too much in order to be adjacent to other organizational units. This problem can be solved quite easily by fixing the shape of

small departments and then adjusting the shape and position of the larger departments accordingly. For example, neighboring organizational units do not have to be adjacent along the whole width or length but only along a part of it.

5.3.4.4.4 Development of the Layout Plan

When developing the layout plan using the graph-theoretical heuristic, each organizational unit is regarded as a whole. In our application this means, for example, that all standby rooms for medical doctors are handled as a group. The possibility to include single standby rooms in other organizational units is not considered by the approach. On the one hand, this shortcoming can be overcome by an appropriate grouping of rooms to organizational units. On the other hand, care needs to be taken to ensure that for each organizational unit the entries in the interaction matrix are needed and that the problem becomes more complex with more organizational units to be located.

5.4 An Iterative Simulation-Optimization Approach for Hospital Layout Planning

The presented graph-theoretical method is only one of many possible approaches for layout planning. As most of the procedures that can be found in the literature, it assumes deterministic data regarding patient, personnel and material flows. In this section an innovative framework for hospital layout planning is presented that takes into account the impact of strategic layout decisions on the operational performance with uncertain process flows (Arnolds and Nickel 2013b). Taking into account this stochastic influence distinguishes the simulation-optimization approach from the formerly presented graph-theoretical approach for layout planning where the data is assumed to be deterministic.

In a preparatory step, the term operational performance has to be defined depending on the application. For example, if the aim is to improve patient and personnel flows through the hospital building, the total travel times for patients and personnel as well as the patients' waiting times for elevators or personnel can be evaluated.

In order to incorporate process uncertainties in the layout planning phase, optimization is combined with discrete event simulation (*DES*). While solving a mathematical model results in an optimal layout under deterministic data, simulation scenarios help to find a robust layout which will show a good performance even when patient, personnel and material flows are uncertain. At a later point in time, the DES model can further be used to test, for example, new schedules for working hours or the influence of building modifications on workflows.

The idea of an iterative simulation-optimization approach for hospital layout planning is adapted from Acar et al. (2009) who presented a generic approach to combine mixed integer programming and simulation in order to solve combinatorial

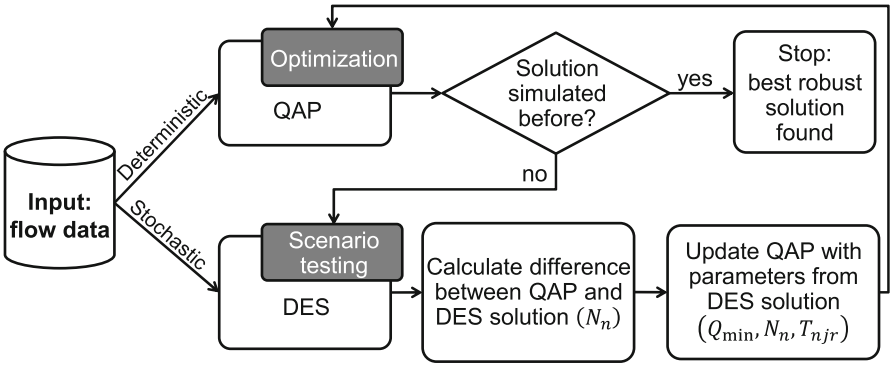


Fig. 5.14 Iterative simulation-optimization approach

problems under uncertainty. Regarding the application of this approach for a hospital layout planning problem, an optimization model for layout planning has to be set up in a first step. In general, any kind of problem formulation can be used. For example, for small instances one can apply the well-known quadratic assignment problem (*QAP*) which has been introduced by Koopmans and Beckmann (1957). By solving the *QAP*, each organizational unit is assigned to a predefined location inside the building. That means, the layout is not constructed from scratch but the structure of the building with defined dimensions (length and width of each level of the building) and locations is given. The following flow chart depicts the iterative simulation-optimization approach adapted to the interaction of a *QAP* and a *DES* model.

As Fig. 5.14 shows, the models interact multiple times, i.e., iteratively until the stopping criterion is met. The first iteration consists of four basic steps: First, the *QAP* is solved assuming deterministic flow data (i.e., generation of a candidate layout *n*). Second, the *DES* model is applied to test the generated *QAP* layout with stochastic flow data. In this step, it is important to ensure that the simulated objective value is worse than the optimal objective value. For example, this can be guaranteed by assuming an increasing amount of arriving patients. Third, the difference between the *QAP* and *DES* objective values (i.e., impact of uncertainty N_n of candidate layout *n*) is calculated. Fourth, the relevant parameters of the QAP model are updated:

- N_n: The impact of uncertainty of candidate layout n as calculated in step three.
- Q_{min}: Minimum simulation result obtained so far from any simulation run.
- T_{njr}: Indicator that shows if the binary variable y_{jr} in candidate layout *n* is 1, i.e., if organizational unit *j* is assigned to location *r* in candidate layout *n* of the *QAP* result.
- *i*: Current iteration − 1.

In the next iteration, solving the updated *QAP* either results in a new candidate layout and the described steps are repeated or it results in a solution that has already

been found in an earlier iteration such that the procedure stops. In the latter case, the last found *QAP* solution is saved as the best robust layout. In order to realize the feedback between the optimization and the simulation models, a generic *DES* model was developed, which can be easily adapted to different hospital layout plans according to the *QAP* solution.

The *QAP* model formulation according to Koopmans and Beckmann (1957) is as follows:

Parameters:

f_{jk}: Flow between organizational units j and k
$d_{r\ell}$: Distance between locations r and ℓ ·
m: Number of organization units and locations.

The decision variables are:

$$y_{jr} = \begin{cases} 1 & \text{if organizational unit } j \text{ is assigned to location } r \\ 0 & \text{else} \end{cases}$$

The model can then be written as:

$$\text{Min} \sum_{j=1}^{m} \sum_{k=1}^{m} \sum_{l=1}^{m} \sum_{r=1}^{m} f_{jk} d_{rl} y_{jr} y_{kl} \tag{5.1}$$

$$\text{s.t.} \quad \sum_{r=1}^{m} y_{jr} = 1 \; \forall \; j \in \{1, ..., m\} \tag{5.2}$$

$$\sum_{j=1}^{m} y_{jr} = 1 \;\; \forall \, r \in \{1, \ldots, m\} \tag{5.3}$$

$$y_{jr} \in \{0,1\} \quad \forall \; j, \; r \in \{1, \ldots, m\} \tag{5.4}$$

The objective function (1) minimizes the total travel distance between the organizational units. Constraints (2) ensure that each organizational unit is assigned to exactly one room whereas constraints (3) guarantee that each room is only occupied by one organizational unit. Constraints (4) define the domain of the decision variables.

In order to facilitate a feedback loop between the optimization and *DES* solutions according to Acar et al. (2009), additional parameters, decision variables and constraints have to be introduced in the *QAP*. Furthermore, the objective function has to be adapted.

The additional parameters are:

i: Current iteration − 1
M: large number
N_n: Impact of uncertainty of candidate solution n
Q_{min}: Minimum simulation result obtained so far from any simulation run

$$T_{njr} = \begin{cases} 1 & \text{if binary variable } y_{jr} \text{ in candidate solution } n \text{ is 1} \\ 0 & \text{else} \end{cases}$$

Additional decision variables are:

$$Z_n = \begin{cases} 1 & \text{if candidate layout } n \text{ has already been suggested in a previous iteration} \\ 0 & \text{else} \end{cases}$$

The resulting model (a modified *QAP*) is then:

$$\text{Min } Z = \sum_{j=1}^{m}\sum_{k=1}^{m}\sum_{l=1}^{m}\sum_{r=1}^{m} f_{jk} d_{rl} y_{jr} y_{kl} + \sum_{n=1}^{i} N_n Z_n \tag{5.5}$$

$$\text{s.t.} \quad \sum_{r=1}^{m} y_{jr} = 1 \; \forall \; j \in \{1, \ldots, m\} \tag{5.6}$$

$$\sum_{j=1}^{m} y_{jr} = 1 \; \forall \; r \in \{1, \ldots, m\} \tag{5.7}$$

$$y_{jr} \in \{0,1\} \; \forall \; j, \; r \in \{1, \ldots, m\} \tag{5.8}$$

$$\sum_{j=1}^{m}\sum_{r=1}^{m} (2T_{njr} \; y_{jr} - y_{jr} - T_{njr}) \leq Z_n - 1 \; \forall \; n \in \{1, \ldots, i\} \tag{5.9}$$

$$\sum_{j=1}^{m}\sum_{r=1}^{m} (2T_{njr} \; y_{jr} - y_{jr} - T_{njr}) \geq M \; (Z_n - 1) \; \forall \; n \in \{1, \ldots, i\} \tag{5.10}$$

$$Z \leq Q_{\min} \tag{5.11}$$

A penalty factor is added in the objective function (5) which reflects the impact of uncertainty N_n of a candidate layout n that has already been found and simulated in an earlier iteration (see Fig. 5.14). Constraints (6)–(8) are the same as constraints (2)–(4) in the original *QAP*. Constraints (9) and (10) ensure that the impact of uncertainty of a previously simulated solution is only incorporated if this solution is currently considered. Constraints (11) define an upper bound for the objective function value equal to the smallest found objective value of any previously simulated solution.

The advantages of the presented simulation-optimization approach are manifold. Firstly, the impact of the strategic layout decision on the operational performance with uncertain process flows and increasing demand in the future is considered. Secondly, the performance and robustness of hospital layouts can be compared and

improved for various scenarios. Scenarios can be defined by changing both input data (extrinsic configuration) and factors, which are revealed during the simulation run (stochastic influences). The former comprises items like the control, capacity and number of elevators, the existence of dedicated personnel elevators or personnel shifts and schedules for breaks. The latter incorporates issues like the uncertainty of clinical pathways depending on the state of health of patients, the used means and speed of transportation, patient arrival rates or service durations. Thirdly, by applying a *DES* model factors like the aforementioned which are hard to integrate in mathematical models or heuristics for hospital layout planning can be studied. Fourthly, the performance of the hospital layout can be evaluated separately for different patient types. Thus, a fairness factor can be established. Patient types can be defined, for example, by severity of illness or level of mobility.

5.5 Conclusion

Regarding the application of operations research methodologies to health care problems in general and to hospital layout problems in particular, it still remains a challenge to convince the people in charge as, for example, hospital managers, medical doctors or nurses of the positive and supporting effects of these methods. Until recently, hospitals were built the way they had been built through the centuries, i.e., hospital designers and planners used to rely mostly on experience and existing campus outlays for inspiration (Arnolds and Nickel 2013a). Today, hospital design is shifting towards patient logistics, opening up totally new perspectives. Hospital budgets, quality of care and patient satisfaction will profit from this transformation.

Mostly, long-term perspectives regarding resource and capacity planning are in the focus of hospital design. However, the emerging building will also significantly influence short-term aspects, i.e., operational workflows. Furthermore, most architectural designers assume that the information they take into account is fixed and deterministic. However, uncertainty can impact data, for example, on future patient figures for certain diseases, as can processes, i.e., the flow of patients, personnel and materials, depending on outcomes and reconvalescence. This uncertainty should be reflected during the design process. Processes should determine how buildings are designed, and not vice versa. Consequently, planning should integrate methods for logistical analysis. Prior to entering the design phase for a new construction project, an analysis of processes needs to be carried out. In particular, clinical pathways for the patients to be cared for in the building should be investigated, providing information on the paths of movement of patients, personnel and material.

The distances travelled can be reduced by an efficient location of organizational units according to the processes which take place in the building, including the flows of patients, personnel and material. Reducing distances means savings in time and, consequently, in resources. Increased efficiency leaves more time to spend on care, which in turn leads to improved patient and personnel satisfaction. To support these

improvements in efficiency it is planned to build a layout planning data base for benchmarking.

The challenge for researchers is now to work together with architects and hospital managers and to convince them that operations research methodologies can be used as additional decision support tools besides expert domain knowledge for hospital layout planning problems. Here, particularly discrete event simulation models can act as a door opener for the practitioners' acceptance of operations research methodologies not only in hospital layout planning but also in other health care topics.

References

Acar Y, Kadipasaoglu SN, Day JM (2009) Incorporating uncertainty in optimal decision making: integrating mixed integer programming and simulation to solve combinatorial problems. Comput Ind Eng 56(1):106–112

Amaral A (2012) The corridor allocation problem. Comput Oper Res 39(12):3325–3330

Amladi P (1984) Outpatient health care facility planning and sizing via computer simulation. Winter Simulation Conference Proceedings. pp 705–711

Arnolds IV, Nickel S (2013) Multi-period layout planning for hospital wards. Socio-Econ Plan Sci 47(3):220–237

Arnolds IV, Nickel S (2013a) An Iterative Simulation-Optimization Approach for Hospital Layout Planning. Working paper, Institute of Operations Research, Discrete Optimization and Logistics, Karlsruhe Institute of Technology

Arnolds IV, Nickel S (2013b) Patient Logistics Drive Hospital Construction. http://www.european-hospital.com/en/article/10973-Patient_Logistics_Drive_Hospital_Construction.html. Accessed 13 April 2015

Ashby M, Ferrin D, Miller M, Shahi N (2008) Discrete event simulation: optimizing patient flow and redesign in a replacement facility. Winter Simulation Conference Proceedings. pp 1632–1636

Assem M, Ouda B, Wahed M (2012) Improving operating theatre design using facilities layout planning. 2012 Cairo International Biomedical Engineering Conference, CIBEC 2012. pp 109–113

Azadivar J, Wang F (2000) Facility layout optimization using simulation and genetic algorithms. Int J Prod Res 38:4369–4383

Balakrishnan J, Cheng CH (1998) Dynamic layout algorithms: a state-of-the-art survey. Omega 26:507–521

Balakrishnan J, Cheng CH (2000) Genetic search and the dynamic layout problem. Comput Oper Res 27(6):587–593

Bashiri M, Dehghan E (2010) Optimizing a multiple criteria dynamic layout problem using a simultaneous data envelopment analysis modeling. Int J Comput Sci Eng 2(1):28–35

Baumgart A, Denz C, Bender HJ, Schleppers A (2009) How work context affects operating room processes: using data mining and computer simulation to analyze facility and process design. Qual Manage Health Care 18(4):305–314

Becker F, Parsons KS (2007) Hospital facilities and the role of evidence-based design. J Facil Manage 5(4):253–274

Benjaafar S, Sheikhzadeh M (2000) Design of flexible plant layouts. IIE Trans 32:309–322

Borzo G (1992) New patient tower mixes aesthetics with practicality. Health Care Strateg Manage 10(9):18–20

Boucherie R, Hans E, Hartmann T (2012) Health care logistics and space: accounting for the physical build environment. Winter Simulation Conference Proceedings

Braglia M, Zanoni S, Zavanella L (2005) Layout design in dynamic environments: analytical issues. Int Trans Op Res 12(1):1–19
Bromley EB (2012) Building patient-centeredness: hospital design as an interpretive act. Soc Sci Med 75(6):1057–1066
Burn J (1982) Facility design for outpatient surgery and anesthesia. Int Anesthesiol Clin 20(1):135–151
Butler TW, Karwan KR, Sweigart JR, Reeves GR (1992) An integrative model-based approach to hospital layout. IIE Trans 24(2):144–152
Butler T, Karwan K, Sweigart J (1992a) Multi-level strategic evaluation of hospital plans and decisions. J Oper Res Soc 43(7):665–675
Ceglowski R, Churilov L, Wassertheil J (2005) Facilitating decision support in hospital emergency departments: a process-oriented perspective. Proceedings of the 13th European Conference on Information Systems, Information Systems in a Rapidly Changing Economy, ECIS 2005
Chang M, Ohkura K, Ueda K, Sugiyama M (2002) A symbiotic evolutionary algorithm for dynamic facility layout problem. Proceedings of the 2002 Congress on Evolutionary Computation, vol 2. pp 1745–1750
Chaudhury H, Mahmood A, Valente M (2005) Advantages and disadvantages of single- versus multiple-occupancy rooms in acute care environments: a review and analysis of the literature. Environ Behav 37(6):760–768
Choudhary R, Bafna S, Heo Y, Hendrich A, Chow M (2010) A predictive model for computing the influence of space layouts on nurses' movement in hospital units. J Build Perform Simul 3(3):171–184
Drira A, Pierreval H, Hajri-Gabouj S (2007) Facility layout problems: a survey. Annu Rev Control 31(2):255–267
Elshafei AN (1977) Hospital layout as a quadratic assignment problem. Oper Res Q 28(1):167–179
Enea M, Galante G, Panascia E (2005) The facility layout problem approached using a fuzzy model and a genetic search. J Intell Manuf 16:303–316
Fetter RB, Thompson JD (1965) The simulation of hospital systems. Oper Res 13(5):689–711
Forsberg H, Aronsson H, Keller CD, Lindblad SE (2011) Managing health care decisions and improvement through simulation modeling. Qual Manage Health Care 20(1):15–29
Francis RL, McGinnis LF, White JA (1992) Facility layout and location: an analytical approach, 2nd edn. Prentice-Hall, Englewood Cliffs
Freeman J, Grimes R, Greene L (1974) Proceedings of the fifth Annual Conference of the Hospital and Health Services Division A.I.I.E., Houston, Texas, Feb. 1974. Abstracts of Hospital Management Studies 11/2, 12627MN, 268 p
Gibson I (2007) An approach to hospital planning and design using discrete event simulation. Simulation Conference, 2007 Winter. pp 1501–1509
Grigg SJ, Garrett SK, Miller MK (2009) Helping a hospital shine. Ind Eng: IE 41(10):24–29
Günal M, Pidd M (2010) Discrete event simulation for performance modelling in health care: a review of the literature. J Simul 4(1):42–51
Hahn P, Krarup J (2001) A hospital facility layout problem finally solved. J Intell Manuf 12(5–6):487–496
Halpern P, Goldberg S, Keng J, Koenig K (2012) Principles of Emergency Department facility design for optimal management of mass-casualty incidents. Prehosp Disaster Med 27(2):204–212
Hamacher H, Nickel S, Tenfelde-Podehl D, Chamoni P, Leisten R, Martin A, Minnemann J, Stadtler H (2002) Facilities Layout for Social Institutions. Operations Research Proceedings 2001. Springer-Verlag, Berlin, pp 229–236
Hancock W, Magerlein D, Storer R, Martin J (1978) Parameters affecting hospital occupancy and implications for facility sizing. Health Serv Res 13(3):276–289
Hans EW, van Houdenhoven M, Hulshof PJH (2012) A framework for healthcare planning and control. In: Hall R (ed) Handbook of healthcare system scheduling. Springer, US 168:303–320
Hassan FB, Tucker A (2010) Using cellular automata pedestrian flow statistics with heuristic search to automatically design spatial layout. Proceedings—International Conference on Tools with Artificial Intelligence ICTAI 2. pp 32–37

Hassan FB, Tucker A (2010a) Using uniform crossover to refine simulated annealing solutions for automatic design of spatial layouts. ICEC 2010—Proceedings of the International Conference on Evolutionary Computation. pp 373–379

Heragu SS (2008) Facilities design, 3rd edn. CRC Press, Boca Raton

Hignett S, Lu J (2010) Space to care and treat safely in acute hospitals: recommendations from 1866 to 2008. Appl Ergon 41(5):666–673

Ierardo G, Luzzi V, Vestri A, Sfasciotti G, Polimeni A (2008) Evaluation of customer satisfaction at the Department of Paediatric Dentistry of "Sapienza" University of Rome. Eur J Paediatr Dent 9(1):30–36

Iskander WH, Carter D (1991) A simulation model for a same day care facility at a university hospital. Winter Simulation Conference Proceedings. pp 846–853

Jiang H, Hu Y (2012) Subset quadratic assignment problem. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 7345 LNAI. pp 226–230

Jun J, Jacobson S, Swisher J (1999) Application of discrete-event simulation in health care clinics: a survey. J Oper Res Soc 50(2):109–123

Kiechle F, Holland C, Karcher RB (2005) Laboratory design. J Clin Ligand Assay 25(4):186–197

Kochhar JS, Heragu SS (1999) Facility layout design in a changing environment. Int J Prod Res 37(11):2429–2446

Koopmans T, Beckmann M (1957) Assignment problems and the location of economic activities econometrica. Econom Soc 25(1):53–76

Kouvelis P, Kurawarwala AA, Gutiérrez GJ (1992) Algorithms for robust single and multiple period layout planning for manufacturing systems. Eur J Oper Res 63(2):287–303

Krishnan KK, Cheraghi SH, Nayak CN (2006) Solving dynamic facility layout problems using dynamic from between charts. 2006 IIE Annual Conference and Exhibition

Krishnan KK, Cheraghi SH, Nayak CN (2008) Facility layout design for multiple production scenarios in a dynamic environment. Int J Ind Syst Eng 3(2):105–133

Kulturel-Konak S (2007a) Approaches to uncertainties in facility layout problems: perspectives at the beginning of the 21st Century. J Intell Manuf 18:273–284

Kulturel-Konak S, Smith AE, Norman BA (2004) Layout optimization considering production uncertainty and routing flexibility. Int J Prod Res 42(21):4475–4493

Kulturel-Konak S, Smith AE, Norman BA (2007) Bi-objective facility expansion and relayout considering monuments. IIE Trans 39(7):747–761

Lacksonen TA (1994) Static and dynamic layout problems with varying areas. J Oper Res Soc 45(1):59–69

Lee HY, Yang IT, Lin YC (2012) Laying out the occupant flows in public buildings for operating efficiency. Build Environ 51:231–242

Leung J (1992) A new graph-theoretic heuristic for facility layout. Manage Sci 38(4):594–605

Levy J, Watford B, Owen V (1989) Simulation analysis of an outpatient services facility. J Soc Health Syst 1(2):35–49

Lin QL, Liu HC, Wang DJ, Liu L (2015) Integrating systematic layout planning with fuzzy constraint theory to design and optimize the facility layout for operating theatre in hospitals. J Intell Manuf (26):87–95

Mahachek AR, Knabe TL (1984) Computer simulation of patient flow in obstetrical/gynecology clinics. Simulation 43(2):95–101

Molyneux E (2010) Paediatric emergency care in resource-constrained health services is usually neglected: time for change. Ann Trop Paediatr 30(1):165–176

Moslemipour G, Lee T, Rilling D (2012) A review of intelligent approaches for designing dynamic and robust layouts in flexible manufacturing systems. Int J Adv Manuf Technol 60:11–27

Murali NM (1988) Model design layout of circular operation theatres complex and ICCU for hospitals. IEEE/Engineering in Medicine and Biology Society Annual Conference, Publ. by IEEE, Piscataway, NJ, United States 10/4. pp 1838–1839

Nassar K (2010) A model for assessing occupant flow in building spaces. Autom Constr 19(8):1027–1036

Nickel S, Tenfelde D, Inderfurth K, Schwödiauer G, Domschke W, Juhnke F, Kleinschmidt P, Wäscher G (eds) (2000) Planning and organisation in the hospital. Oper Res Proc 1999:548–553 (Springer Berlin Heidelberg)

no Author (1975) Examination of case studies in health facilities planning. Abstr Hosp Manage Stud 11(3):13047 AR:397 p

Norman BA, Smith AE (2006) A continuous approach to considering uncertainty in facility design. Comput Oper Res 33(6):1760–1775

Pagell M, Melnyk S (2004) Assessing the impact of alternative manufacturing layouts in a service setting. J Oper Manage 22(4):413–429

Pillai VM, Hunagund IB, Krishnan KK (2011) Design of robust layout for dynamic plant layout problems. Comput Ind Eng 61(3):813–823

Seelye A (1982) Hospital ward layout and nurse staffing. J Adv Nurs 7(3):195–201

Sepulveda JA, Thompson WJ, Baesler FF, Alvarez MI, Cahoon LE III (1999) Use of simulation for process improvement in a cancer treatment center. Winter Simulation Conference Proceedings, IEEE, Piscataway, NJ, United States vol 2. pp 1541–1548

Swisher J, Jacobson S, Jun J, Balci O (2000) Modeling and analyzing a physician clinic environment using discrete-event (visual) simulation. Comput Oper Res 28(2):105–125

Thompson J, Goldin G (1975) The hospital: a social and architectural history. Yale University Press, New Have

Thorwarth M, Arisha A (2012) A simulation-based decision support system to model complex demand driven healthcare facilities. Proceedings—Winter Simulation Conference

Tompkins JA, White JA, Bozer YA, Tanchoco JMA (2010) Facilities planning, 4th edn. Wiley, Hoboken

Ulutas BH, Islier AA (2009) A clonal selection algorithm for dynamic facility layout problems. J Manuf Syst 28(4):123–131

Urban TL (1998) Solution procedures for the dynamic facility layout problem. Ann Oper Res 76:323–342

Vos L, Groothuis S, Van Merode G (2007) Evaluating hospital design from an operations management perspective. Health Care Manage Sci 10(4):357–364

Walus Y, Ittmann HC, Hanmer L (1997) Decision support systems in health care. Methods Inf Med 36(2):82–91

Weiss G, von Baer R, Riedl S (2002) Einfluss des Raumkonzepts einer Operationsabteilung auf die Nutzungseffizienz. Der Chirurg 73(2):174–179

Yang T, Peters BA (1998) Flexible machine layout design for dynamic and uncertain production environments. Eur J Oper Res 108(1):49–64

Yeh IC (2006) Architectural layout optimization using annealed neural network. Autom Constr 15(4):531–539

Part II
Public Services

Chapter 6
Modeling the Potential for Critical Habitat

Richard L. Church, Matthew R. Niblett and Ross A. Gerrard

6.1 Introduction

Planners frequently have some objective or outcome that they wish to address or understand. Location modeling is an approach that is regularly used as an aid in understanding the effect(s) of a particular policy or plan, or to provide possible alternatives for expanding, reorganizing, or contracting a system of facilities. For example, the Location Set Covering Problem (*LSCP*, Toregas et al. 1971) has often been employed to determine the minimum number of fire stations required to "cover" an entire population or set of demands. A demand is said to be "covered" if it is within the limits of a specified maximum travel-time or distance standard from a located station. The *LSCP* model can be used to determine the most efficient configuration of stations such that all neighborhoods of a city can be reached within the standard. By comparing an ideal *LSCP* solution to an existing set of stations, one can estimate the efficiency with which a city has deployed their fire protection services. For example, if a city uses 30 fire stations when it can be served by an optimal placement of 25 stations, then it might be possible to reduce the needed number of stations by relocation (ReVelle 1991). In fact, it is a relatively simple task to modify a model like that of the LSCP to generate and test possible alternatives of system expansion, relocation, and consolidation. The location set covering problem is just one of many location models that have been applied in a variety of modeling contexts. Such modeling contexts range from civil applications, such as fire station placement (Toregas et al. 1971), to environmental applications such

R. L. Church (✉) · M. R. Niblett
University of California, 1832 Ellison Hall Santa Barbara, Santa Barbara, CA 93106, USA
e-mail: church@geog.ucsb.edu

M. R. Niblett
e-mail: mniblett@geog.ucsb.edu

R. A. Gerrard
USDA Forest Service, PSW Research Station, 1731 Research Park Dr., Davis, CA 95618, USA
e-mail: rgerrard@fs.fed.us

H. A. Eiselt, V. Marianov (eds.), *Applications of Location Analysis,* International Series in Operations Research & Management Science 232, DOI 10.1007/978-3-319-20282-2_6

as the natural reserve site selection and design problems (Underhill 1994; Church et al. 1996; Williams et al. 2004).

Location models have been used for analysis and design in a number of environmental problems. Select examples include: designing a groundwater monitoring network (Hudak et al. 1993; Meyer and Brill Jr. 1988); natural reserve site design and selection (Church et al. 1996; Cova and Church 2000; Fischer and Church 2003; Malcolm and ReVelle 2005; Matisziw and Murray 2006; Williams et al. 2004); wildlife management (Downs et al. 2008); and conservation planning (Church 2013). Interest in environmental applications for which location modeling may be used to provide greater insight and planning capabilities when used in conjunction with Geographic Information Systems (*GIS*) have been on the increase.

This chapter focuses on the application of a location model in habitat analysis associated with the California Spotted Owl (*Strix occidentalis occidentalis*). The spotted owl is a species of concern in forest planning in California, and forest managers and biologists have spent a great deal of effort in estimating their habitat needs as well as planning activities (e.g., fuels removal) in patterns that minimize the impact to the spotted owl. The application reported here involves the use of the anti-covering location problem, which was first proposed by Moon and Chaudhry (1984). This anti-covering location problem has also been referred to as a packing problem (see, for example, Stephenson 2005). One of the main attributes for the anti-covering location problem is that any two located facilities may not be closer than a minimum separation distance standard. Although the application doesn't actually involve locating a configuration of facilities, it can be viewed as locating a set of circular territories and their centers, such that all centers are located at least a minimum separation distance from all other centers, just as in the anti-covering location problem. One can think of each circular territory as an idealized spotted owl territory centered about a feasible nest site or center location. Spotted owls are quite territorial and tend to keep other owls from encroaching in their area. In essence, maximizing the number of circular territories that can be placed within a region is equivalent to estimating the largest number of mated pairs of spotted owls that an area can support. This represents an estimate of the carrying capacity of a given habitat area, which is a critical metric in habitat planning. When planned activities (e.g. thinning of stands to reduce potential fire severity) and natural disturbances (such as an invasion of the bark beetle causing high tree mortality or a wildfire) alter both the shape and size of a habitat, the model can be used to estimate whether a change in carrying capacity has occurred as well. The anti-covering location model can be used in a similar manner for other territorial species as well.

6.2 Background

Conservationists tend to focus on protecting and sustaining what they believe to be core habitat, hot spots of high species richness, and locations containing rare endemic species. "Core habitat" is a concept that is often difficult to quantify and

is frequently estimated subjectively by expert opinion (Church 2013). Often, core habitats represent the interior of a continuous swath of good habitat that is free of edge effects. It also must satisfy some notion of minimum size. For example, medium sized core habitat for the black bear and moose in Vermont is an unfragmented landscape of suitable habitat comprised of an area ranging between 1500 and 10,000 acres in size (Austin et al. 2006). Also, core may be a smaller area, but be reasonably close to another area of suitable habitat. For example core habitat for the prairie chicken is defined as a mixed grass or tall-grass prairie of at least 5000 acres or an area between 1250 and 5000 acres that is no more than 6.5 miles from another habitat of at least 1250 acres (Hagen et al. 2004). So, size, quality, and distance relationships may be important in defining core habitat. Modeling suitable habitat, let alone core, can be a daunting task. The classic approach is to develop a wildlife habitat relationship model (*WHR*) that can be used to predict the presence of a specific animal based upon the geographic distribution of needed habitat elements, which may vary with season and vary when they are raising offspring. For example, the State of California maintains a geographic database that depicts the possible presence and suitability of 694 terrestrial vertebrates based upon the application of *WHR* models to geographic data.

Our focus here is on the estimation of the carrying capacity of an area of suitable habitat. Conservation planners often estimate the carrying capacity of a habitat for a territorial species by first estimating the total area of suitable habitat and then dividing that estimate by the average size of a given territory. Such estimates are approximate at best given that suitable habitat is often distributed unevenly across a landscape.

6.3 Determining the Carrying Capacity Using Core Habitat Maps

To describe the basic problem of determining the carrying capacity of an area associated with a territorial species, consider the hypothetical landscape depicted in Fig. 1.1. In this figure each raster cell is classified in one of three ways: core habitat, habitat, and matrix lands for a mated pair of fox. This type of classification follows that used in the well-known *FRAGSTATS* program, which is designed to analyze habitat patterns from the perspective of a number of metrics (McGarigal and Marks 1995). Core habitat in this case involves high quality habitat units that are surrounded by other core units or by units that are classified as suitable, but not core. Those areas that are classified as habitat are either of lower quality than what is needed to be classified as core, or they are of high quality, but are not completely surrounded by high quality habitat. Matrix lands are classified as everything else and may even be hostile areas for the fox. Suppose that the habitat requirements for this territorial species has been defined as needing approximately 37 raster cells of habitat, of which at least 30 cells should be considered core habitat. The area in Fig. 1.1 contains 34 units of core habitat, so one might readily conclude that this

habitat could support one mated pair of this species. But, what if we consider a territory of approximately 37 units in size? Figure 1.1 depicts two circles of such a size. Notice that it is not possible to find a compact area containing all core habitat or nearly so, which would meet the standards of being a feasible territory. Thus, the territory needs to be larger than the circular territory and even somewhat oblong. The larger a territory has to be in order to circumscribe enough core to support the fox and the more elongated that territory is, the greater the energy requirements are for the kit fox to hunt prey and return to their den to feed their young. The habitat might support two mated pairs of fox if the habitat within each of the circles is of high enough quality that each pair can survive. But, it is more likely that the core habitat is spread out across the region to the extent that it cannot support any mated pairs of fox, as the energetics required to find enough prey is too great a struggle. In general, territories that are compact and almost circular are preferred to those that are not.

Note here that the needed habitat for a mated pair is approximately 37 units of which 30 units are core. Here there are 34 such units classified as core, so a simple mathematical estimate of (34 units of core)/(30 units of core needed $= 1.13$ would suggest that the habitat could support one mated pair of fox, even though it probably couldn't support any mated pairs of fox. Thus, the proximity and pattern of core habitat cannot be captured by a simple mathematical equation when calculating whether a feasible species territory exists.

There are several ways in which one might compute a potential territorial area that can support a demographic unit of a species (e.g., a mated pair of fox). The simplest way is to assume that the best possible territory is compact and mostly circular. Another way is to develop a model which can identify an area that is a contiguous cluster of spatial cells that is comprised of enough good habitat to support a species, but at the same time contains little if any hostile areas (Church et al. 2003). In this chapter we will assume that estimated territories are compact and circular.

If we assume that territories are compact and circular, a more accurate estimate of the carrying capacity of an area for a territorial species can be generated through the use of a location model that involves the placement of non-overlapping circular territories, as the limit on carrying capacity is bounded by the number of such territories that can be placed within core habitat. Before we describe this application and our solution approach, the next section will briefly discuss the literature on the anti-covering location problem (Fig. 6.1)

6.4 The Anti-Covering Location Problem

The *Anti-Covering Location Problem* (*ACLP*) involves identifying a configuration of facilities that must be separated from each other by at least a minimum specified distance. Such a problem can be defined on a network or a continuous plane. For networks it is often assumed that each node is a potential facility site and that on the plane all points may be considered feasible within some bounded region.

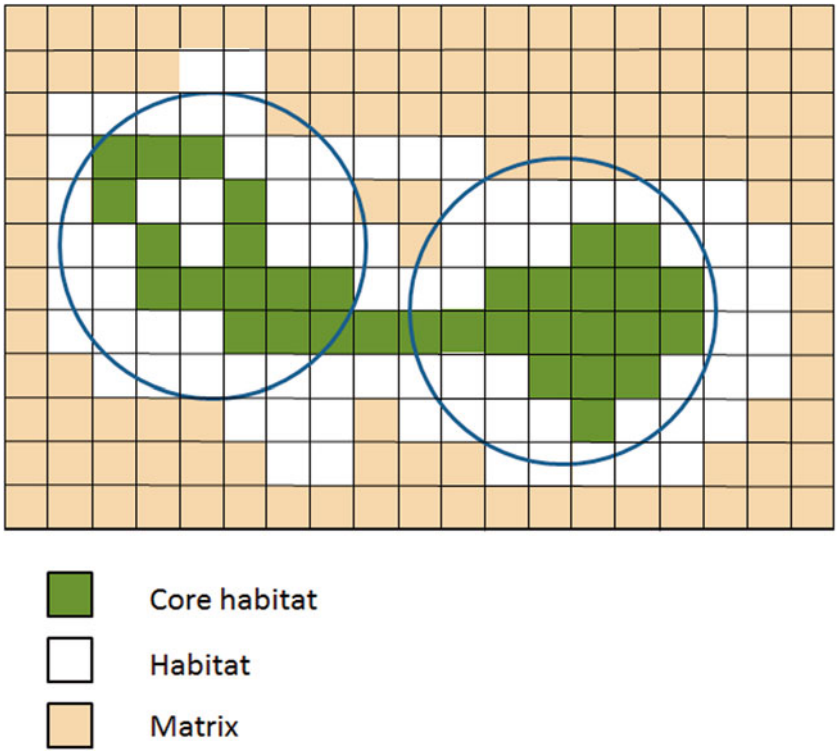


Fig. 6.1 A hypothetical region containing habitat for the kit fox. The circles represent a compact region of a size that if filled principally with core habitat would meet the requirements in supporting mated pair of kit fox

The anti-covering location problem on a bounded area can be thought of as a circle packing problem (Stephenson 2005) or as a special type of packing problem that can include spheres, boxes and other shapes in two or more dimensions (Dowsland and Dowsland 1992). From this point we will assume that the set of feasible locations is represented as a discrete set of points on a Euclidean plane. From among this set, the objective is to locate as many facilities as possible while ensuring that each located facility is at least a minimum distance, r, from its nearest neighboring facility. For the purposes of application, we will assume that the problem is defined on a Euclidean plane, where all feasible points are defined in advance and are represented by their x,y coordinates. The distances between any pair of facilities will be calculated as Euclidean distance. The Anti-Covering Location Problem was first described and formulated by Moon and Chaudhry (1984). It has also been called the r-separation location problem (Erkut et al. 1996). Consider the following notation:

i, j indexes representing potential location sites (the entire set is denoted by I)
d_{ij} Euclidean distance from potential site i to potential site j where $i, j \in I$

r Euclidean distance separation standard (no two facilities may be closer to each other than this distance)

$N_i = \{j : d_{ij} < r \ \ \& \ \ i \neq j\}$;

n_i A sufficiently large number to impose locational restrictions (see discussion)

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

Using the above notation we can formulate a model for the anti-covering location problem as follows:

$$\text{Max } Z = \sum_{i \in I} x_i \tag{6.1}$$

s.t.

$$n_i x_i + \sum_{j \in N_i} x_j \leq n_i \ \forall \ i \in I \tag{6.2}$$

$$x_i \in \{0,1\} \ \forall \ i \in I \tag{6.3}$$

The objective function (1) maximizes the number of facilities that are located among the feasible set of sites. Constraints of type (2) enforce the separation requirements among the located facilities. This constraint is written for each candidate site location. If a facility is located at site i, the sum of the selections at all other facilities j within r-distance of i, the neighborhood N_i, must sum to zero and are thus precluded from being selected. In essence, any site j that is strictly within r-distance of site i can be selected for a facility only if site i has not been selected. This type of constraint is often termed a neighborhood constraint. The final constraint (3) lists the integer restrictions on the decision variables. This is a rather simple binary integer programming problem.

The value of n_i was originally defined by Moon and Chaudhry to be M, a large number. Yoshimoto and Brodie (1994) suggested that the value of n_i did not need to be any larger than the number of members in the set N_i. This fact alone helps to produce a tighter formulation when solving a problem with integer programming software. Murray and Church (1996) have shown that the approach of Yoshimoto and Brodie can be further reduced to a theoretical limit defined by a vertex packing problem. More recently, Niblett (2014) has shown that without any loss of generality the value of N_i for a Euclidean based *ACLP* problem can be set at the value of 5.

There are alternate forms for the *ACLP* that have been developed which involve some form of pairwise or higher ordered restrictions. Suppose that two sites i and j are too close to each other to be both chosen for facilities. In this case, one could write a constraint of the form:

$$x_i + x_j \leq 1 \tag{6.4}$$

Constraint (4) is called a pairwise constraint and prevents the simultaneous selection of both sites i and j. If three sites$i$, j, k, are within r distance of each other a triplet constraint can be written as follows.

$$x_i + x_j + x_k \leq 1 \tag{6.5}$$

This constraint will allow the selection of at most one of these three sites to be selected, and represents pairwise restrictions for (*i*, j), (*j, k*), (*i, k*). Such conditions are called clique constraints and can help reduce the computational effort when solving *ACLPs* by general purpose software. When solving an *ACLP* to optimality, the best modeling approach reported in the literature has been found to be a combination of neighborhood constraints and clique constraints (Erkut et al. 1996); one such form is the maximal clique model of Murray and Church (1996).

Of interest here is the potential for using the *ACLP* model in estimating the carrying capacity of a given region of habitat. Clearly, the most desirable territories are compact and circular, which requires an individual animal to "protect" or defend the smallest amount of boundary as well as represents a shape that would require the least amount of energetics in foraging for food. If a territory of suitable habitat needs to be of a certain size, then this needed size can be used to estimate the radius, *s*, of a circle representing the territory. The separation distance between territory centers, *r*, would be twice the needed radius of a circular territory, *s*. Thus, $r = 2 \times s$. Given this needed radius and a habitat map of suitability, then one can solve an *ACLP* problem to estimate the maximum carrying capacity of the landscape for such a territorial species. What is important is that this methodology would be sensitive to the pattern and undulations in the habitat within a region. Several researchers have taken this approach including Gerrard (2006), Downs et al. (2008), and Church (2013) for habitat analysis. Downs et al. (2008) have employed the *ACLP* to estimate the carrying capacity for a population of greater sand-hill cranes (*Grus canadensis tabida*), in critical habitat in the State of Ohio. The greater sand-hill crane is a territorial species that rigorously defends and maintains a territory of roughly 3 km about their nesting sites. Estimates of the maximum carrying capacity of an area were then compared to that of the current population, so that restoration efforts could be directed towards those areas that had not reached their maximum potential. In this chapter we report on the work of Gerrard (2006) and Church (2013) that involved using the *ACLP* model in analyzing the potential carrying capacity for the California Spotted Owl in forests in California.

Before we describe the work on using the *ACLP* in modeling spotted owl habitat, it is important to discuss one final element associated with habitat modeling and planning. Although an ideal territory may be circular and compact, and many territories may be somewhat circular is shape, they may also be somewhat oblong, non-circular in shape. Gerrard et al. (2001) have described the features of constructing a map of suitable habitat values for the San Joaquin Kit Fox based upon expert opinions. The suitability values were based upon a ratio-scale, which means that they could be added across a habitat patch to define a composite score for such a patch. The suitability scores for each cell were defined in such a way that "poor" cells detracted from the value of a patch and high quality cells contributed to the value of the patch. Experts defined a minimum threshold for which the composite

sum of a patch needed to exceed this amount in order to be considered a viable supporting habitat patch for some defined demographic unit (in the case of the kit fox, it was a mated pair). Church et al. (2003) have developed a patch generation heuristic that can generate a large sample of viable patches. The important point here is that the *ACLP* model can also be used to estimate the carrying capacity of a region by maximizing the selection of these non-circular patches, while ensuring that no two selected patches overlap. Thus, the *ACLP* model can be tailored to handle non-circular shapes by generalizing the forms in which neighborhood and clique memberships are defined.

6.5 The California Spotted Owl and its Habitat

The California Spotted Owl (*Strix occidentalis occidentalis*) is a bird of prey that resides primarily in the coniferous, hardwood-coniferous, and hardwood forests of California's Coast Range and Sierra Nevada mountain ranges. Because the spotted owl is a protected species, forest managers must prepare management plans focused on preserving a large portion of existing habitat with an eye towards enhancing and increasing core habitat over time. This is also true for forest product companies that are required to maintain spotted owl populations. Many of the areas for which the spotted owl has been observed are in old growth conifer and mixed conifer forests (Verner et al. 1992). Nesting sites have been associated with dense canopy cover and a variety of tree sizes from medium to large structures (Gutiérrez et al. 1992; LaHaye et al. 1997; Gerrard 2006). In addition to preference of larger trees and dense canopy, the spotted owl is an intrasexual territorial species, which preferentially select nest sites (Verner et al. 1992).

In fact, "An individual or a pair may be dominant in the core area of its home range but not in the periphery. This tends to produce a *regular dispersion* [emphasis added] by effectively excluding other individuals from breeding in the core without necessarily excluding their presence there as subordinates engaged in other activities (Brown and Orians 1970, p. 244)." Such nest sites are typically associated with having high canopy closure and larger trees in the area surrounding the nest site as well. For spotted owls, the average of the home range core surrounding a nesting area is approximately 2.4 km^2 or a circle of radius 879 m, which is equivalent to an area of 600 acres (0.937 square miles).

California spotted owl core habitat can be mapped following the approach of Gerrard et al. (2001). In this case, core habitat can be mapped using a vector dataset containing the vegetation class polygons available from the US Forest Service. This data contains estimates of the mix of trees (when present), an age/seral class, as well as the level of canopy closure within each polygon. Often such polygons are determined through the use of air-photo interpretation or with remotely sensed satellite data, which is usually verified through a random set of on-the-ground survey plots. These polygons can be converted to a raster form. Figure 6.2 shows a map with the rasterized polygons which have been classified based upon the amount of canopy

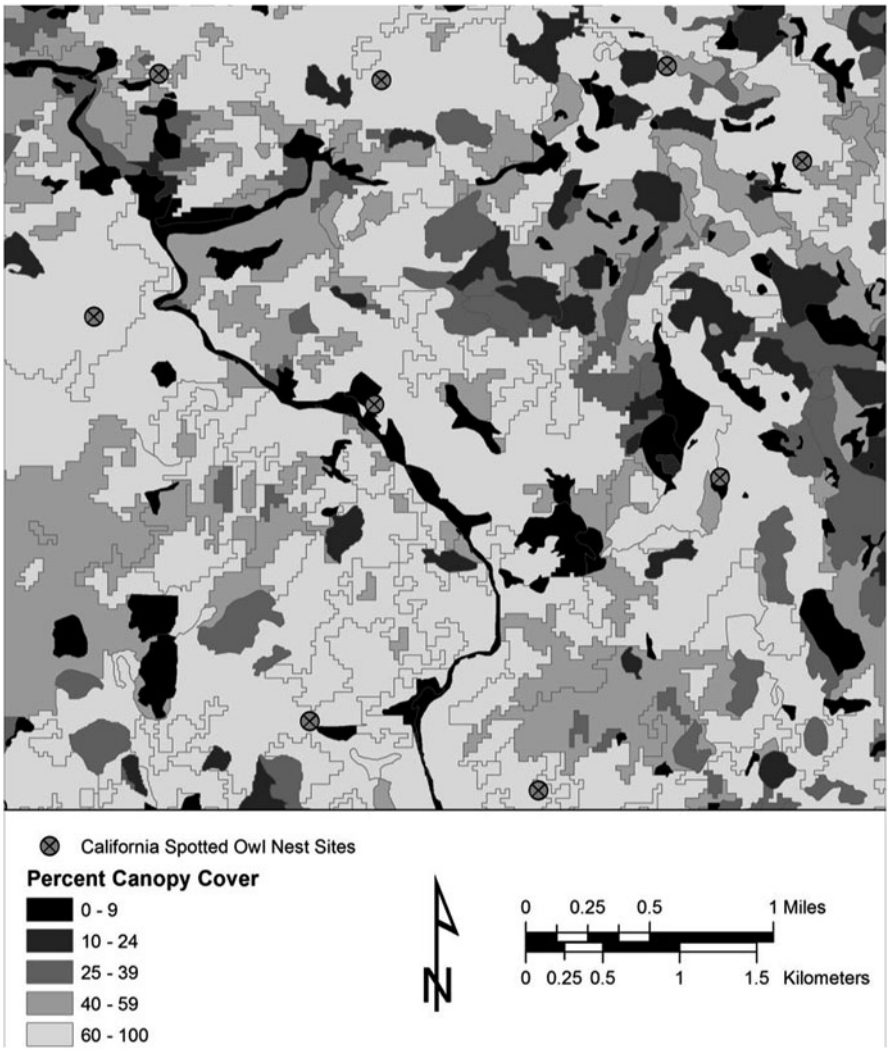


Fig. 6.2 Rasterized map of canopy cover polygons with California spotted owl nest sites

closure. Forest biologists defined core habitat to be those areas of mature to late seral stage forests in which a given location and the surrounding 879 m circle about it has an average canopy closure of 60 % or greater, and habitat to be those areas in which the canopy closure within a 879 m circle surrounding a point is greater than 50 %. It is relatively easy to generate a map depicting habitat and core habitat using *GIS* (Geographic Information System) functionality. For example, the Focal Mean Score function of *Arc/GIS* of *ESRI* Corporation can be used to calculate the average canopy surrounding a given pixel for a circular region of 600 acres and map this

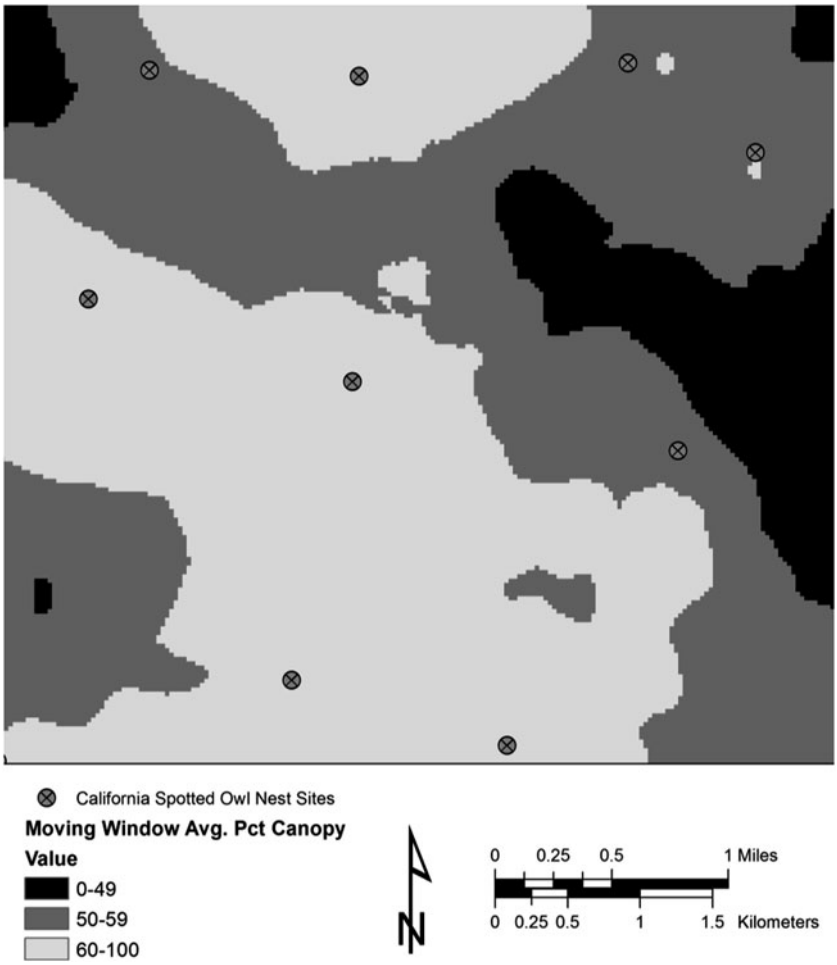


Fig. 6.3 Core habitat map derived using *GIS* functionality. Each raster cell is scored based upon the average canopy values within a circle of 879 m

average value for that given pixel. Figure 6.3 shows the same area as Fig. 6.2 and shows the results of using the focal mean function where each pixel is classified in one of the 3 canopy cover ranges, based upon the average canopy surrounding each pixel in an 879 m circle (600 acres). One can think of the areas that have an average canopy $> 60\,\%$ as being core habitat and potential nest site locations. This figure also shows known nest site locations. All but three of the nest sites are in the densest canopy class. None of the sites are in locations with less than 50 % canopy cover. Those sites that are located in areas with an average of 50–59 % canopy cover are generally very close to those locations that have dense canopy ($\geq 60\,\%$). In addition,

one can observe that each of the spotted owl nest sites are roughly evenly distributed from one another.

6.6 Utilizing the Anti-Covering Location Model for Estimating Carrying Capacity for the Spotted Owl

One of the main problems in solving the anti-covering location problem in habitat analysis is the fact that problems can be of considerable size. For the raster discussed above for the Kings River Project Area (*KRPA*), nearly 75,000 cells are

possible candidates for selection as potential nest centers out of approximately 1,000,000 cells that comprise this forested area. This easily translates to a model size that exceeds the capabilities of most commercial solvers. In fact, given that there is one binary integer variable for each potential nest site center, a simple application of the *ACLP* to the *KRPA* would involve nearly 75,000 binary variables and 150,000 constraints (assuming 1 neighborhood constraint for each potential nest center and one maximal clique constraint for each potential nest center), which is clearly out of range of a commercial solver. There are two possible tacks that one can use in solving a problem of this magnitude. The first is to develop a heuristic and the second is to solve a set of smaller integer programming problems. The second approach can be accomplished by dividing the region into a number of smaller areas and applying the *ACLP* model to each of these subareas, or by reducing the total number of nest sites by selecting a sample of nest sites distributed across the region and solving the *ACLP* on such a sample. This second approach is also a heuristic in nature, as the true optimum may not be identified when a problem has been divided into a set of smaller regions or when only a sample of sites are actually used. We chose to develop both approaches (a heuristic and an IP based heuristic), although our sponsor wanted a process that could quickly solve the problem to within a reasonable estimate of the carrying capacity. Overall, we wanted to use the *ACLP* model to give us a level of confidence that our heuristic could generate close to if not optimal solutions to the *ACLP*.

The *KRPA* does exhibit several distinct areas of feasible nest centers, and it is possible that we could solve the *ACLP* model on each of these distinct areas. Because such a distribution is not always so disconnected, we decided to develop a form of the IP in which a selected sampling of sites would be used to generate a model problem instance. We developed a visualization routine in Visual Basic to map out a given solution to the *ACLP*. We added a module to this routine to produce an MPS file that represents a given *ACLP* model instance and used this as input to the *XPRESS* solver (ver. 25.01.05) of *FICO* Corporation. Our module selects an X% sample of cells of all feasible nest center cells as potential sites. This module generated for each potential nest center, one core clique constraint (see Erkut et al. 1996) and needed pairwise constraints to represent all separation restrictions. We found

Table 6.1 Selected sample problems involving the application of the ***XPRESS*** solver to an integer linear optimization model of an ACLP

% of sites used	Number of site selection variables	Number of constraints	Objective value	Solution times (secs.)
4	3406	242,434	62	230.1
5	4066	334,723	62	989.6
6	4989	492,793	63	3,899.1
7	5551	649,360	63	52,568.5

this to be the best formulation representing this problem. Table 6.1 presents data associated with the solution of a range of problem instances based upon the percentages of sites that were selected for generating the problem instance. Note that for a 5 % sample of sites, a problem instance contained 4066 binary variables and 334,723 constraints. Some of these constraints were redundant and were eliminated by the presolve function of *XPRESS*. The solution time for this problem was 989.6 s (16.5 min) and involved the selection of 62 nesting site centers and associated circular territories. Over all sample problems, the largest value for the *ACLP* objective for the *KRPA* solved to optimality was 63. It should be noted here that each X% sample of sites represents the entire region of feasible sites. By the very fact that it contains fewer constraints, one might conclude that it may be possible to site more than what would be otherwise possible, with a fully representative model. But, given that a solution based upon an X% sample of sites is a feasible anti-covering solution for the entire problem means that it cannot involve the positioning of any more than the maximum possible number of circular territories. Therefore, a solution to a sample based problem instance could be optimal to the original, fully specified problem, but also may be less than optimal.

Figure 6.4 presents a map of the solution generated for the problem solved with a 6 % sample of sites. This solution involved the location of 63 circular territories. Since this is a sample based solution, it is possible that more could be located, although it is probably unlikely.

We have also developed a heuristic solution process for the *ACLP* problem when defined on a raster, where all feasible cells are considered as possible sites rather than restricting site selection to a small sample of sites as required when solving an integer linear program. Our heuristic design was based upon two simple observations associated with a solution of nest centers. The first observation is that for a solution to be close to optimal, it is likely to have the circular territories packed as closely together as possible, so that as many of these circular territories are deployed as possible. The second is if such circular territories are densely packed and often touch boundaries of neighbors, then local adjustments of a given solution are likely to violate the separation requirements. We can view this property as one of brittleness, that is readjusting a site center of a densely packed solution may trigger the need to readjust many other nest centers in order to meet all of the separation

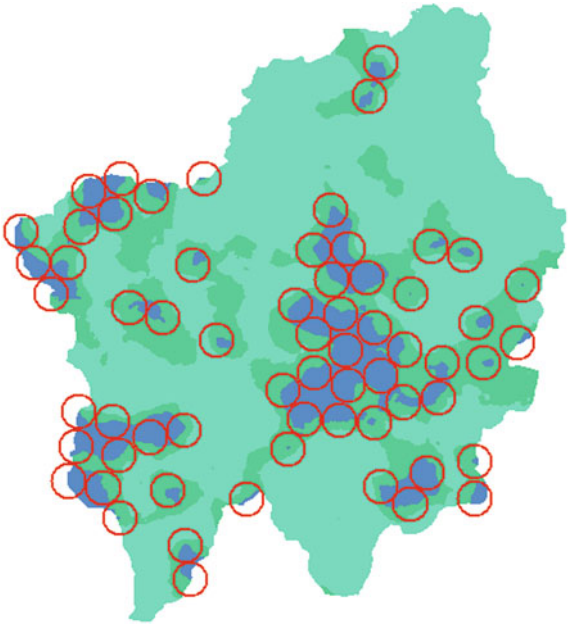


Fig. 6.4 A map of the 63 selected nest centers and their circular territories associated with the optimal *ACLP* model solution generated when the sample was set at 6 %. The darkest regions represent those pixels that can serve as territory centers

conditions among all circular territories. That is, adjusting the position of one territory will potentially generate a domino effect in needed readjustment of a string of nest centers. Thus, readjusting a solution to test for local improvement may take a considerable amount of computational effort with no guarantee that it will be fruitful. Because of these two observations, we elected to concentrate on developing a heuristic that could generate a densely packed feasible solution. We could then run the heuristic a large number of times and keep the best solution found among all of the restarts.

The basic premise is to start the heuristic by selecting one of the feasible nest sites as the first center for a circular territory of radius r. The second circular territory is chosen as that feasible center which is as close as possible to the first chosen center. If there are several possible candidates that are equally close, this second site is chosen at random from among the closest candidates. The third site that is chosen, must be a site which minimizes a combined distance to the first two chosen centers while maintaining the desired separation among the chosen sites. The fourth site is the site that is chosen at random from among those sites that minimize a combined distance from the other three sites, subject to the condition that no separation conditions are violated. The product of this first step, often results in a kernel of four clustered sites and circular territories. Subsequent sites are chosen which minimize the combined distance to this four site kernel, while maintaining that no site separation conditions

are violated among the chosen sites/circular territories. The process continues until it is not possible to add any further sites without violating the separation standard. The process is repeated for a set number of times, among which the best solution found is reported as the final result.

There are two elements of randomness, which make this process akin to a semi-greedy process. The first, involves the selection of the initial "seed" center, about which other selected sites will be clustered. The second is associated with the fact that at any stage of selection, if there are ties as to which site is closest to the growing kernel or kernel (of the first four chosen sites), the selection is made at random with respect to the candidates tied for best. As this is a network of cells defined along a raster (or grid), it is common that such ties exist. Consequently, this process tends to expand in a semi-greedy fashion about the first chosen site. As a beginning seed locations are chosen at random from among the set of feasible nest centers, a large number of restarts tends to sample many of the areas of a problem landscape, a form of diversification. Finally, this heuristic by its very nature produces a densely packed arrangement of nest centers and circular territories, a property often visualized in optimal *ACLP* arrangements. We call this heuristic, *Packer*, as it tries to pack as many circular territories into a given region.

The application of the Packer heuristic to the *KRPA* for 1000 restarts is presented in Table 1.2. There are four columns in this table. The first column presents a given deployment of circular territories. For example on the first line, it can be seen that 54 territories were located. The second column indicates how many times out of 1000 trials, this result was obtained. In this case it shows that 8 times out of a thousand trials, 54 territories were located. The third column indicates the frequency with which this particular number of territories was located, and the fourth indicates how much such solutions depart from the best solution that was found among the 1000 restarts. For example, the highest value found was 62, which was found 21 times out of a 1000 restarts, which was approximately 2.1 % of the time. The average over all restarts was 57.6, and 27.5 % of the time the heuristic found a solution which was within 5 % of the best found. Overall, the heuristic performance is close to expectations, as it found a number of close to optimal solutions within a reasonable amount of computation time (about 4000 s) given the enormous number of sites and the very nature of a packed solution. It is important to note that the optimal process was able to locate only one additional territory than the maximum of the heuristic Table 6.2.

6.7 Final Comments and Conclusions

In this chapter we have described an application of the anti-covering location problem as a tool to estimate the possible carrying capacity of a territorial species. The basic idea is one of locating as many compact and circular territories as possible, while ensuring that each circular territory does not overlap with any other located territories. Each circular territory must be centered at a location which is considered

Table 6.2 The results of the Packer heuristic applied 1000 times to the Kings River Ranger District

Number of located circular territories	Number of times out of 1000	Frequency (%) in which a given number are located	% departure from best solution found
54	8	0.8	12.9
55	54	5.4	11.3
56	181	18.1	9.7
57	282	28.2	8.1
58	200	20.0	6.5
59	163	16.3	4.8
60	32	3.2	3.2
61	59	5.9	1.6
62	21	2.1	0.0

to be ideal habitat as well as contain within the circle enough core habitat to be considered feasible in the sense that is could support such a species. The number of located territories, then, is an estimate of the maximum carrying capacity of the habitat for that species. The anti-covering location problem is an exact representation of this problem when applied to a Euclidean plane, as the objective is to locate as many facilities (territory centers) as possible such that they are all separated by at least $2r$-distance from each other (so circular territories of radius r do not overlap).

We show how the *ACLP* was applied to estimate the number of spotted owls that can be supported in a given forest region. To do this requires a definition of suitable habitat as well as what would constitute a potential nest/territorial center with respect to the site and its surrounding circular shaped territory. We present an application of this model to the Kings River Project Area of the Sierra National Forest in California, which was comprised of nearly 75,000 feasible sites, among which it was possible to locate 63 non-overlapping circular territories.

We described the solution of this problem from the perspective of two heuristic approaches. The first involved solving an integer linear programming problem where a sample of all sites is used as it was not possible to solve this problem to optimality when using all sites as potential points for selection. The second approach was a heuristic algorithm that was designed to generate solutions which mimic the conditions that can be observed among optimal solutions, that is dense tightly packed arrangements. This process was designed to be repeated a large number of times as a form of diversification where the best found is saved for display and review. Results of both approaches applied to the Kings River Ranger District are presented. Overall, the heuristic is able to identify close to if not optimal solutions for such large problems.

The *ACLP* heuristic was developed along with a visualization routine, so that forest planners and biologists could apply this program to any raster data set that was stored in an *ASCII*-grid format, a raster data format that can be exported from several well-known geographical information systems (*GIS*) such as *Arc/GIS* (a product

of *ESRI* corporation, Redlands, CA). This stand-alone program can be used to estimate the carrying capacity of other territorial species as long as one has an estimate of the required size of circular territories, geographic information system database on habitat, and accepted definitions for estimating whether a given territory location can support the species under consideration. This work is now being tested on habitats involving the fisher (*Pekania pennanti*) as well as expanded for varying radii of separation, where larger radii are used for habitats of good quality and smaller radii are used for habitats of high quality. Finally, feasible noncircular territories are being modeled for the San Joaquin Kit Fox (*Vulpes macrotis*) and habitat plans are being generated by optimizing the selection of these non-circular areas while each territory is allowed to overlap by a pre-specified amount with other territories.

Acknowledgements We would like to acknowledge the research funding of the US Forest Service which supported the development of the program Packer and the visualization routine. We also would like to acknowledge the help and assistance from Pat Manley, Pete Stine, and John Keane of the US Forest Service.

References

Austin J, Viani K, Hammond F (2006) Vermont wildlife linkage habitat analysis: a GIS-based, Landscape-level identification of potentially significant wildlife habitats associated with the State of Vermont Roadways. Vermont Agency of Transportation

Brown JL, Orians GH (1970) Spacing patterns in mobile animals. Annu Rev Ecol Syst 1:239–262

Church RL (2013) Identification and mapping of habitat cores. In: Craighead FL, Convis CL Jr (eds) Conservation planning: shaping the future. Redlands. ESRI Press, California, (219–244)

Church RL, Stoms DM, Davis FW (1996) Reserve selection as a maximal covering location problem. Biol Conserv 76:105–112

Church RL, Gerrard RA, Gilpin M, Stine P (2003) Constructing cell-based habitat patches useful in conservation planning. Ann Assoc Am Geogr 93/4:814–827

Cova TJ, Church RL (2000) Contiguity constraints for single-region site search problems. Geogr Ann 32/4:306–329

Downs JA, Gates RJ, Murray AT (2008) Estimating carrying capacity for sandhill cranes using habitat suitability and spatial optimization models. Ecol Model 214:284–292

Dowsland KA, Dowsland WB (1992) Packing problems. Eur J Oper Res 56:2–14

Erkut E, ReVelle C, Ulkusal Y (1996) Integer-friendly formulations for the *r*-separation problem. Eur J Oper Res 93:342–351

Fischer DT, Church RL (2003) Clustering and compactness in reserve site selection: an extension of the biodiversity management area selection model. For Sci 49/4:555–565

Gerrard RA (2006) Measuring habitat quality and carrying capacity for the California spotted owl: case study using data from Sierra national forest. Interim report compiled for Sierra Nevada research center and Plumas-Lassen-Tahoe national forests

Gerrard R, Stine, P, Church R, Gilpin M (2001) Habitat evaluation using GIS: a case study applied to the San Joaquin Kit Fox. Landsc Urban Plan 52:239–255

Gutiérrez R, Verner J, McKelvey KS, Noon BR, Steger GN, Call DR, Senser JS (1992) Habitat relations of the spotted owl. In: The California spotted owl: a technical assessment of its current status. Pacific Southwest Research Station, US Dept. of Agriculture. US Forest Service, Gen. Tech. Rep. PSW-GTR-133. Albany, CA, p. 79–98

Hagen CA, Jamison BE, Giesen KM, Riley TZ (2004) Guidelines for managing lessor prairie-chicken populations and their habitats. Wildl Soc Bull 32/1:69–82

Hudak PF, Loaiciga HA, Schoolmaster FA (1993) Application of geographic information systems to groundwater monitoring network design. J Am Water Resour Assoc 29/3:383–390

LaHaye W, Gutiérrez R, Call D (1997) Nest-site selection and reprodutive success of California spotted owls. Wilson Bulletin 109/1:42–51

Malcolm SA, ReVelle C (2005) Representational success: a new paradigm for achieving species protection by reserve site selection. Environ Model Assess 10:341–348

Matisziw TC, Murray AT (2006) Promoting species persistence through spatial association optimization in nature reserve design. J Geograph Syst 8:289–305

McGarigal K, Marks B (1995) FRAGSTATS spatial analysis program for quantifying landscape structure. General Technical Report PNW-GTR-351, USDA Forest Service, Pacific Northwest Research Station

Meyer, Philip D, Brill Jr, E Downey (1988) "A method for locating wells in a groundwater monitoring network under conditions of uncertainty". Water Resour Res 24(8):1277–1282

Moon ID, Chaudhry SS (1984) An analysis of network location problems with distance constraints. Manage Sci 30/3:290–307

Murray AT, Church RL (1996) Analyzing clique constraints for imposing adjacency restrictions in forest models. Forest Science 42:166–175

Niblett MR (2014) The anti-covering location problem: new modeling perspectives and solution approaches. PhD Dissertation. University of California, Santa Barbara

ReVelle, Charles (1991) "Siting ambulances and fire companies: new tools for planners". J Am Plan Assoc 57(4):pp 471–484

Stephenson K (2005) Introduction to circle packing: the theory of discrete analytic functions. Cambridge University Press, New York

Toregas C, Swain R, ReVelle C, Bergman L (1971) The location of emergency service facilities. Oper Res 19/6:1363–1373

Underhill LG (1994) Optimal and Suboptimal Reserve Selection Algorithms. Biol Conserv 70:85–87

Verner J, Gutiérrez R, Gould GI Jr (1992) The California spotted owl: general biology and ecological relations. In: The California spotted owl: a technical assessment of its current status. Pacific Southwest Research Station, US Dept. of Agriculture. US Forest Service, Gen. Tech. Rep. PSW-GTR-133. Albany, CA, p. 55–77

Williams JC, ReVelle CS, Levin SA (2004) Using mathematical optimization models to design nature reserves. Front Ecol Environ 2/2:98–105

Yoshimoto A, Brodie JD (1994) Comparative analysis of algorithms to generate adjacency constraints. Can J For Res 24:1277–1288

Chapter 7
Saving the Forest by Reducing Fire Severity: Selective Fuels Treatment Location and Scheduling

Richard L. Church, Matthew R. Niblett, Jesse O'Hanley, Richard Middleton and Klaus Barber

7.1 Introduction

Wildfire is a natural process which can lead to a variety of conditions in a forested landscape, some beneficial and some not. For example, in coastal chaparral communities some seeds are activated and germinate only by the heat of a fire. After a fire, young plants can grow somewhat free of competition. As another example, the giant sequoia (*Sequoiadendron giganteum*) is shade intolerant and fire is necessary to clear an understory so that light can reach seedlings. Unfortunately, wildfire may also be so destructive that sensitive habitats may be reduced in size or significantly damaged, even to the point of reducing its viability in supporting threatened or endangered species. Some believe that such destructive fires are the result of long term fire suppression (Gruell 2001; Miller et al. 2009; North et al. 2007), however, this is a hotly debated issue. Fortunately, no one questions that destructive fires often occur when litter (woody debris from trees) and ladder fuels are abundant leading to more destructive crown fires.

Finney (2004) has shown with a fire simulation model that fire severity and extent may be reduced by selectively reducing forest fuels across a forest. Fuels reduction

R. L. Church (✉) · M. R. Niblett
1832 Ellison Hall Santa Barbara, University of California, Santa Barbara, CA 93106, USA
e-mail: church@geog.ucsb.edu

M. R. Niblett
e-mail: mniblett@geog.ucsb.edu

J. O'Hanley
Kent Business School, Kent University, Canterbury, UK
e-mail: J.Ohanley@kent.ac.uk

R. Middleton
Los Alamos National Laboratory, Los Alamos, NM, USA
e-mail: rsm@lanl.gov

K. Barber
US Forest Service Region 5, Vallejo, CA, USA
e-mail: 4khbarber@gmail.com

H. A. Eiselt, V. Marianov (eds.), *Applications of Location Analysis,* International Series in Operations Research & Management Science 232, DOI 10.1007/978-3-319-20282-2_7

may involve mechanical removal of litter and ladder debris, prescriptive burns of understory, and thinning of stands. Thinning of stands helps separate the crowns of trees reducing the possibility of a spreading crown fire. In order to make a forest system more resilient to catastrophic fire losses, Finney demonstrated that treatments do not need to encompass an entire forest, but be strategically placed, scattered and encompass approximately 25 % of the forest. In fact a set of strategically placed fuels treatment areas can be almost as effective as a cost-prohibitive approach of treating an entire forest. This means that there is considerable flexibility in where treatments can be placed. Further, since fuels treatments are costly, such treatments must be scheduled over a planning horizon.

In order to conceptualize, develop, and apply a location model for strategic and tactical decision making, one must first understand the problem as well as the audience of users and decision makers. Within an industrial organization, a location model formulation and application may well be done by an in-house systems modeling team or by a consultant. Often these location decisions are one-of a kind, or are accomplished periodically over a period of years in support of expansion, consolidation, merging, and other strategic decisions (Geoffrion and Powers 1980; Camm et al. 1997). In comparison, large governmental agencies may divide their operations geographically into districts, in which similar problems and decisions are made, but involve local subject matter experts. These subject matter experts in local issues and concerns, may find it necessary to solve a variant of a given problem. Without flexibility in how data is acquired and stored and how such data might be used to populate a model as well as having flexibility in defining the model itself, a location model may find use in one district and not another. Thus, there are hurdles to overcome when a model is to be applied locally among a number of district organizations, rather than globally by one organization. In fact it is this dimension of modeling that was the major hurdle in developing an approach that would be applied across ranger districts of National Forests in California.

The US Forest Service operates 17 National Forests in California, of which 15 are located in the Sierras or in the northern third of California. Altogether these forests cover 31,264 square miles (approximately the area of the Czech Republic). All but two of these forests have been the subject of fuels removal planning. Until the 1980's the US forest service managed harvesting operations in many of the western National Forests. These activities were planned at the strategic level by the use of large scale linear programming models (Kent et al. 1991) and were assigned to watershed units by a tactical level model, which also involved linear programming (Church et al. 1998). Concern for environmental impacts to watersheds and habitat changed the major focus of the *USFS* to habitat conservation and protection. Because of this, large scale harvesting virtually disappeared in public forests in California in the 1990s, with the exception of salvage operations after major fires. Today, stands may be thinned by harvesting selected tree sizes; such operations are performed on a very selective basis for fire protection and stand health and must retain any sizable trees (called green tree retention). The planning process for selective thinning and woody debris removal is described in the next section. Removal

of dead woody debris and tree thinning can reduce fire severity and aid in stand survival, which is the major objective of fuels removal programs. Following that, we describe a spatial location-scheduling model that has been developed to assist in fire resilience planning. After the description of the model and a short description as to how it is solved, we describe its implementation, the key element to application, as well as give an example of the results of an application. This model system has now been used in the majority of forests in California.

7.2 Background

The organization of the US Forest Service (*USFS*) is a well-defined hierarchy. The headquarters is in Washington, DC. They direct 9 regions; each of them operate independently of the others and are each directed by a Regional Forester. Each region is comprised of a number of National Forests (*NF*). Most of these are divided into three to five ranger districts (*RD*). Day-to-day operations are directed by the staff of each ranger district. Policies and directives are developed by the Washington office and regional headquarters. Planning, environmental review, monitoring, assessment, etc. is accomplished by *NF* staff and *RD* staff, with input and assistance from the regional staff. Historically, modeling assistance has been provided by the top analyst in a region (the regional analyst). The tools, models, and focus of a given *NF* can vary considerably from other *NFs* in a region, based upon the priorities, budget, and targeted conservation species.

In the last decade, operations have been directed towards the protection of critical habitat from catastrophic, ground clearing fires as well as protecting the wildland-urban interface (*WUI*). The interface between wildland and urban areas contain cabins, homes, recreational facilities, and scattered commercial services. Protecting *WUI* to the greatest extent possible is a political necessity and protecting habitat is a prime objective. These two goals frequently are in conflict, as critical habitat and *WUI* lands are often some distance apart and budgets for fuels removal are limited.

In order to understand how limited resources might be used to protect habitat and *WUI* areas from severe fires, it is first necessary to understand the state of the art in modeling fire. *USFS* and other agencies have supported fire modeling research for quite some time. There are several fire simulation programs that have been developed over the last several decades, and the ones that have been used the most are part of a suite of programs. These include *FARSITE* (Finney 2004), Behave (Andrews et al. 2003), and *FLAMMAP* (Finney 2006).

Fire simulation models use data on terrain, vegetation coverage, moisture conditions, etc. and simulate the spread of a fire over time from a starting location. One can determine the potential impact of fuels removal on fire spread, by simulating a fire before and after fuels treatment. Essentially, a fuels treatment plan will effect vegetation and woody debris conditions and changes fire spread, severity, etc. between pre and post fuels treatment can be estimated by comparing the two simulations (Collins et al. 2010). Finney (2001) demonstrated with the use of

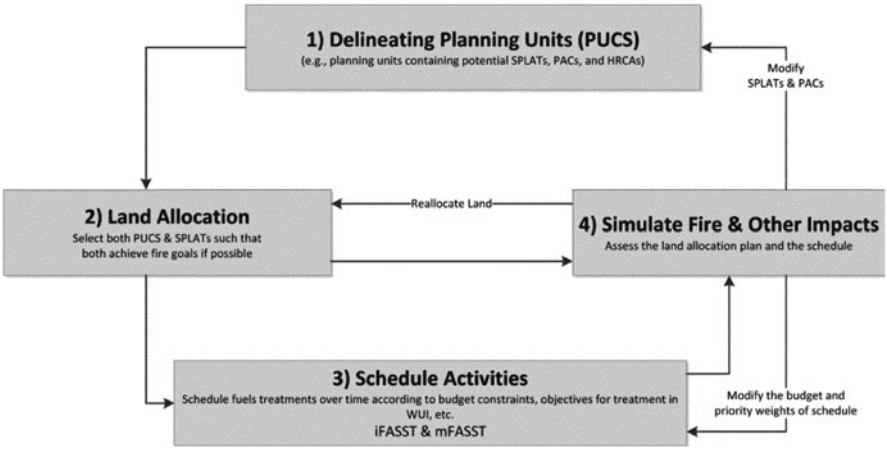


Fig. 7.1 Planning work flow for the Fire Cadre

simulation that a set of distributed fuels treatment areas which together comprise approximately 25 % of the landscape can reduce fire severity as if the entire landscape had been treated. Imagery of recent fires, e.g., Rodeo-Chediski fire in 2002 in Arizona, showed that fire intensity was considerably lower in those areas that had been thinned. Others have documented that fuels removal of litter and ladder fuels also tends to lessen the intensity and speed of a fire (Collins et al. 2010). In 2001, the Regional Forester of Region 5 (Pacific Southwest) directed that the forests of the Sierras be managed to protect the existing habitat of the spotted owl and address the fire hazard risk to those areas by strategically locating fuels treatments (Powell 2001). With the outbreak of several large fires (including the Rodeo-Chediski) in 2002, the US Congress acted to improve forest health with the passage of the "Healthy Forests Restoration Act of 2003." In response, the Regional Forester of Region 5 directed that a plan "aggressive enough to reduce the risk of wildfire to communities in the wildland urban interface (*WUI*) while modifying fire behavior over the broader landscape" be implemented (Blackwell 2004). This directive also ordered that habitat of other species of concern be protected as well. This included the Pacific fisher, the Willow flycatcher, the Great Gray Owl, as well as others.

To meet the directive, each *NF* was called upon to develop fuels treatment plans. At the heart of such plans is the designation of those areas that will be treated for fuels reduction (mechanical removal, thinning, and controlled burns), called *SPLATs* (*S*trategically *P*laced *L*andscape *A*reas for fuels *T*reatments). To aid in this process, a team of experts, called the Fire Cadre (*FC*), was assembled. The *FC* had experts in fire modeling, treatment cost estimation, vegetation modeling, forest biology and *GIS*. Together this team aided each *NF* in modeling fire with and without treatments. The *FC* developed a planning process which is outlined in Fig. 7.1.

Figure 7.1 depicts the 4 major steps of the planning process (see Bahro et al. 2007 for a detailed description). The first step (Box 1) is associated with delineating those

areas that are part of protected activity centers (*PACs*) or home range core areas (*HRCAs*) for the spotted owl. Other protected habitats must also be identified. Further, potential *SPLATS* are identified and a proposed fuels treatment method based upon: location, vegetation condition, terrain, etc, is determined. The second step (Box 2) involves selecting a set of *SPLATs* that are spatially distributed. Given this set, the vegetation cover map is modified to reflect this set of treatments. Then, a fire simulation model is run for a number of different fire scenarios (wind direction, starting location, etc.) in order to assess the potential of the system in protecting critical habitat, *WUI*, and promoting tree survival (Box 4). This is done in order to generate a preliminary assessment as to the effectiveness in *SPLAT* treatments. It may be necessary to readjust some *SPLATs*, generate more *SPLATs*, or be possible to reduce the number of *SPLATs* to generate a cost effective treatment plan for fire resilience. Once a feasible preliminary plan has been developed then a schedule of treatments is generated by a spatial-scheduling model (*FIRS*) using the Initial Forest Activities Spatial Scheduling Tool (*iFASST*) Decision Support System (*DSS*), the main subject of this chapter. After a schedule is generated, then fire Simulations (Box 4) are run on the impact of the first 5 years of treatments (or some other timeframes) to estimate the impact of the schedule on overall area protection, etc. This may mean that some readjustments in the schedule need to be made. The remainder of this chapter describes the spatial scheduling model and its implementation.

Before scheduling, nearby *SPLATs* are aggregated into small project areas, called *PUCS*, a short term for *p*lanning *u*nits *c*ontaining *SPLATs*. *SPLATs*, themselves, are too small to contract out for a private company. For private companies to be interested in bidding, it is necessary to have enough SPLATS aggregated together so that it is worth their while and will occupy a work team for several months. The *PUCS* are scheduled by the iFASST model.

Our description so far has purposefully neglected to describe the spatial elements of the scheduling problem in order to keep our description from being needlessly complicated. But now that we have described the overall planning process, it is now necessary to specify in greater detail the spatial elements in scheduling. As previously stated, *SPLATS* and therefore *PUCS* need to comprise approximately 25 % of the land area. Finney (2004) demonstrated that fuels removal at this percentage of the landscape can be as effective as reducing fuels across the entire landscape. This has been called the Finney effect. To achieve this condition, SPLATS need to be distributed and strategically placed in terms of potential fire spread, etc. When an area represented by a *PUCS* has been treated, enough fuels have been removed so that the Finney effect or condition has been reached. Essentially, the objective is to achieve the Finney condition across the entire forest. But, budgets limit the extent to which this can be achieved and to treat all *PUCS* may take 20 years. Besides the budget preventing a quick resolution to fuels treatment needs, there is an important spatial issue. If isolated *PUCS* are treated, then fire severity will presumably be lessened in those *PUCS* areas. But, if localized critical habitat is surrounded by several *PUCS*, some treated and some not, then that critical habitat is not fully protected until all remaining neighboring *PUCS* have been treated. In addition to this, scheduling neighboring *PUCS* together or in subsequent years helps to extend the Finney

effect to larger blocks of land, an added benefit in slowing the rate of fire spread. If all landscape elements were considered equal, then the scheduling problem would involve starting treatment in one *PUC* and then clustering subsequent treatments among its neighbors, and continuing that process until all *PUCS* have been treated. But, in fact, the record of decision discussed above directs the planners to protect as a first priority, areas of critical habitat and *WUI*. Thus, the scheduling problem is one in which *PUCS* are scheduled and clustered among *WUI* areas and surrounding areas of critical habitat.

There are other constraints of a practical matter that must also be considered. If at all possible, at least one *PUC* project should be scheduled each year in each ranger district of a *NF*. This helps to spread the workload of those managing and monitoring fuels treatment projects across the ranger districts. Also, many *NFs* wish to treat approximately the same acreage in one year as another, as this tends to spread out the work evenly across the planning horizon. There are other nuances of the spatial-scheduling system that are better described in the overall model presentation, which is the subject of the next section.

7.3 The Spatial-Scheduling Model *FIRS*

In order to formulate the Fire Intensity Reduction Scheduling (*FIRS*) model, we begin by defining needed notation:

i, j	indices used to represent a specific project areas, where $j = 1, 2, 3, \ldots, n$
t	an index used to represent planning periods, where $t = 1, 2, 3 \ldots, m$
k	an index used to represent a specific planning area, where $k = 1, 2, \ldots, p$
c_{jt}	the cost of treating project j in time period t
a_j	the size of treatment area in project j in acres
f_j	the total acreage of project (treated and untreated)
w_j	the amount of wildland—urban interface acres present in project j
h_j	the number of acres of sensitive habitat present in project j
θ_j	the earliest time period in which project j can be scheduled
H_t	the total number of acres of habitat that can be disturbed in time period t
d_t	the discount factor for time t, where value is zero in time one and increases with time
B_t	the available budget for fuel removal projects in time period t
Γ_t	$\{j$: project j can be assigned for treatment in period $t\}$
Ω_j	$\{t$: project j can be assigned for treatment in period $t\}$
P_j	$\{j$: project j is part of planning area $k\}$
E	$\{(i, j)$: project areas i and j are adjacent where $i < j\}$
s_t	deviation above the average yearly level of treatment area for time t
v_t	deviation below the average yearly level of treatment area for time t

In addition to the above notation, we will need the following decision variables:

$$x_{jt} = \begin{cases} 1, \text{ if project } j \text{ is scheduled for treatment in time period } t \\ 0, \text{ otherwise} \end{cases}$$

$$u_{kt} = \begin{cases} 1, \text{ if planning area } k \text{ is not assigned a project in period } t \\ 0, \text{ otherwise} \end{cases}$$

$$z^0_{ijt} = \begin{cases} 1, \text{ if project } i \text{ and project } j \text{ are both scheduled in time period } t \\ 0, \text{ otherwise} \end{cases}$$

$$z^1_{ijt} = \begin{cases} 1, \text{ if project } i \text{ is scheduled in time } t \text{ and project } j \text{ is scheduled in time } t+1 \\ 0, \text{ otherwise} \end{cases}$$

$$z^1_{jit} = \begin{cases} 1, \text{ if project } j \text{ is scheduled in time } t \text{ and project } i \text{ is scheduled in time } t+1 \\ 0, \text{ otherwise} \end{cases}$$

Given the above notation and decision variables, the *FIRS* model can be formulated as follows:

$$\text{Max } Z_1 = \sum_{t=1}^{m} \sum_{j \in \Gamma_t} f_j x_{jt} \tag{7.1}$$

$$\text{Min } Z_2 = \sum_{k=1}^{p} \sum_{t=1}^{m} u_{kt} \tag{7.2}$$

$$\text{Min } Z_3 = \sum_{t=1}^{m} \sum_{j \in \Gamma_t} d_t w_j x_{jt} \tag{7.3}$$

$$\text{Min } Z_4 = \sum_{t=1}^{m} (s_t + v_t) \tag{7.4}$$

$$\text{Max } Z_5 = \sum_{(i,j) \in E} \left(\sum_{t=\max\{\theta_i, \theta_j\}}^{T} z^0_{ijt} + \sum_{t=\max\{\theta_i, \theta_j\}}^{T-1} z^1_{ijt} + z^1_{jit} \right) \tag{7.5}$$

subject to the following constraints:

If at all possible, assign at least one project to each planning unit in each period:

$$\sum_{j \in P_k \cap \Gamma_t} x_{jt} + u_{kt} \geq 1 \quad \text{for each } k = 1, 2, 3, ..., p \ \ \& \ \ t = 1, 2, 3, ..., m \tag{7.6}$$

Do not exceed the allowable budget in each year:

$$\sum_{j \in \Gamma_t} c_{jt} x_{jt} \leq B_t \quad \text{for each } t = 1, 2, 3, ..., m \tag{7.7}$$

Treatment within sensitive habitat in each year is limited:

$$\sum_{j \in \Gamma_t} h_{jt} x_{jt} \leq H_t \quad \text{for each } t = 1, 2, 3, ..., m \tag{7.8}$$

Each project can be scheduled at most once:

$$\sum_{t=\theta_j}^{m} x_{jt} \leq 1 \quad \text{for each } j = 1, 2, 3, ..., n \tag{7.9}$$

Track when adjacent treatments are scheduled:

a. adjacent treatments in the same year

$$z^0_{ijt} \leq x_{it} \quad \text{for each } (i, j) \in E \ \ \& \ \ \max\{\theta_i, \theta_j\} \leq t \leq m \ \ \& \ \ j > i \tag{7.10}$$

$$z^0_{ijt} \leq x_{jt} \quad \text{for each } (i, j) \in E \ \ \& \ \ \max\{\theta_i, \theta_j\} \leq t \leq m \ \ \& \ \ j > i \tag{7.11}$$

a. adjacent treatments in previous or subsequent year

$$z^1_{ijt} \leq x_{it} \quad \text{for each } (i, j) \in E \ \ \& \ \ \max\{\theta_i, \theta_j\} \leq t \leq m-1 \ \ \& \ \ j > i \tag{7.12}$$

$$z^1_{ijt} \leq x_{jt+1} \quad \text{for each } (i, j) \in E \ \ \& \ \ \max\{\theta_i, \theta_j\} \leq t \leq m-1 \ \ \& \ \ j > i \tag{7.13}$$

$$z^1_{jit} \leq x_{jt} \quad \text{for each } (i, j) \in E \ \ \& \ \ \max\{\theta_i, \theta_j\} \leq t \leq m-1 \ \ \& \ \ j > i \tag{7.14}$$

$$z^1_{jit} \leq x_{it+1} \quad \text{for each } (i, j) \in E \ \ \& \ \ \max\{\theta_i, \theta_j\} \leq t \leq m-1 \ \ \& \ \ j > i \tag{7.15}$$

Track treatment level in each year and compute deviation from average:

$$\tfrac{1}{m} \sum_{t=1}^{m} \sum_{j \in \Gamma_t} a_j x_{jt} - \sum_{j \in \Gamma_t} a_j x_{jt} = v_t - s_t \quad \text{for each } t = 1, 2, 3, ..., m \tag{7.16}$$

Restrictions on variables:

$$x_{jt} = \{0, 1\} \quad \text{for each } j = 1,2, 3, ..., n \ \ \& \ \ t = \theta_j, \theta_j + 1, ..., m \tag{7.17}$$

$$u_{kt} = \{0, 1\} \quad \text{for each } k = 1,2, 3, ..., p \ \ \& \ \ t = 1, 2, 3, ..., m \tag{7.18}$$

$$v_t \geq 0 \quad \text{for each } t = 1, 2, 3, ..., m \tag{7.19}$$

$$s_t \geq 0 \quad \text{for each } t = 1, 2, 3, ..., m \tag{7.20}$$

$$z^0_{ijt} = \{0,1\} \quad \text{for each } (i, j) \in E \ \ \& \ \ t = \max\{\theta_i, \theta_j\}, ... \tag{7.21}$$

$$z^1_{ijt} = \{0,1\} \quad \text{for each } (i, j) \in E \ \ \& \ \ t = \max\{\theta_i, \theta_j\}, ..., m-1 \tag{7.22}$$

$$z^1_{jit} = \{0,1\} \quad \text{for each } (i, j) \in E \ \ \& \ \ t = \max\{\theta_i, \theta_j\}, ..., m-1 \tag{7.23}$$

The *FIRS* model is a multi-objective integer-linear programming model. It contains five types of objectives. Each project, a Planning Unit Containing Splats (*PUCS*), represents a set of fuel removal activities as a cluster of *SPLATS* within the project

area. If all the activities are accomplished within the project area, then the entire area is defined as having met the "Finney" threshold. The first objective, Z_1, involves maximizing the total number of acres over the entire planning period that is classified as meeting the Finney threshold. The second objective, Z_2, involves minimizing the number of time periods any planning area has not been assigned at least one project. Planning areas are relatively large subdivisions of the forest and usually represent ranger districts. Since each ranger district is a somewhat independent operating unit within a National Forest, it is important to keep planning and operations staff involved in each district and time period if at all possible. The third objective, Z_3, is designed to minimize or maximize a discounted function of treated wildland urban interface (*WUI*) lands over time. This effectively prioritizes treatments in *WUI* lands in the minimization sense to be treated earlier in the planning horizon. Prioritizing such treatment areas earlier in the planning horizon helps to protect cabins, residential, and commercial areas on the edge of the urban-forest interface from significant fire as early and quickly as possible. In addition, scheduling projects containing *WUI* acreage as early as possible reduces the cost of fire suppression in these areas, as fires in treated areas do not burn as intensely. The fourth objective, Z_4, minimizes year-to-year variation in treated acreage. Including this objective is crucial, as many fuels treatment activities are handled by outside contractors and forest service personnel who would like to keep the contracting and supervisory workload consistent from year to year. The fifth objective, Z_5, maximizes project adjacency in the same (Z^0_{ijt}) and subsequent (Z^1_{ijt}) time period. This objective enables a planner to cluster projects in a given year, or in a subsequent year, or to spread projects out if the objective is minimized.

The constraints also restrict the model solution in several key ways. The first constraint (6) assigns *PUCS,* x_{jt}, to each planning unit k in each time period t, or it forces $u_{kt} = 1$. This constraint thus tracks whether or not a project has been assigned to a planning area (ranger district) for time period t. This constraint, in conjunction with the second objective, attempts to keep a ranger district and its personnel occupied with at least one *PUCS* project each year.

Constraint (7) forces the total cost of projects x_{jt} scheduled in a time period t to be less than or equal to the allotted budget for time period t. Constraint (8) similarly forces every scheduled project x_{jt} in time period t to impact no more than H_t acres of sensitive habitat for that time period. Constraint (9) stipulates that each project j can be scheduled at most once between the earliest possible time project j can be scheduled and the last year of the planning horizon.

Constraints (10)-(15) track whether or not a scheduled project is adjacent to another scheduled project. Constraints (10)-(11) track whether a project has been scheduled adjacent to another project in the same time period t, considering the earliest planning period the project may be scheduled through the end of the planning horizon. Constraint set (12)-(15) tracks whether or not a project in time period t has been scheduled adjacent to another project in time period $t + 1$, considering the earliest planning period the project may be scheduled through the end of the planning horizon minus one. The reason for considering the earliest time a project may

be scheduled to the end of the planning horizon minus one ($m - 1$) is because a project cannot be scheduled in the year past the end of the planning horizon.

Constraint (16) tracks the treatment level in each year and computes the deviation from the average. This constraint is needed to define the level of deviation that is used to support objective 4. Constraints (17)-(23) define the restrictions on the variables; in this case to be binary or non-negative in value.

Although the above formulation captures the fuels treatment scheduling and location problem, it contains several properties which make it difficult to solve optimally. First, National Forests typically contain three or more Ranger Districts. Each of the Ranger Districts may in themselves contain 50–100 project areas (*PUCS* containing *SPLATS*). Consequently, the number of *PUCS* in a forest is usually larger than 160. Since the planning period has been defined to be 20 years, and with each *PUC* typically being schedulable in each of those time periods, yields a problem with the number of integer scheduling variables exceeding 5000. Couple this large number of integer scheduling variables with two "budget" constraints (one on costs and one on habitat disturbance) for each time period, and it becomes a large multi-dimensional knapsack problem. Finally, add to this a large number of "spatial-temporal clustering constraints and variables" and the result is a very large integer linear programming problem.

The *FIRS* model is embedded in a modeling and maping decision support system called the *i*ntial *F*orest *A*ctivities *S*patial *S*cheduling *T*ool (*iFASST*). This modeling system is described in detail in the next section. This system is capable of producing a model output file that can be read by solvers like *GUROBI* and *CPLEX*. A routine was developed to read a solver output file and produce a scheduling file that could be mapped and analyzed by the *iFASST* program. A specialized heuristic was also developed to solve the problem as well. The optimal approach was used to help fine tune the heuristic. The specific details of the heuristic are given in Niblett et al. (2014); suffice it to say that it employes elements of GRASP (Feo und Resende 1995; Resende und Ribeiro 2005), and path relinking (Glover et al. 2000). As most *NFs* do not have access to solvers like *IBM CPLEX, GUROBI*, or *FICO Xpress* nor the time to spend seeking true optimal solutions, the heuristic was relied on by forest analysts.

7.4 Decision Support System Implementation

As noted above, US Forest Service managers have been tasked with reducing fire severity within their forests by using varying fuels reduction strategies with species of concern and ecosystem restoration objectives in mind. The *DSS* must therefore be flexible in its ability to handle a variety of objectives and data inputs from several sources and users. It should also be able to: set up planning scenarios by an analyst; display solution(s) to a scenario in mapped and data-graphic form; and to allow the analyst to modify a scenario based upon a previously obtained solution. In addition, the *DSS* should be easily deployed for use by forest managers and analysts. The

Initial Forest Activities Spatial Scheduling Tool (*iFASST*) *DSS* was designed with these capabilities in mind.

The *iFASST DSS* was designed to run on the Microsoft Windows operating system, the main operating system of US Forest Service computers, with the ability to open a variety of data types. The data types used by *iFASST* are either derived from Geographic Information Systems (*GIS*) or from processed data exported from software such as Microsoft Excel or Microsoft Access. In addition to its built-in heuristic and integer linear programming model building, *iFASST* also has the capability to provide data-graphs and maps of solutions derived from heuristic or solver solutions and to export solution data and statistics. The next section focuses on how a scenario file may be generated and the data requirements.

7.4.1 Building a Scheduling Scenario

Building a scheduling scenario requires spatial and non-spatial data inputs. Building several planning scenarios for a variety of forests and problem variations involving several users is even harder. The *iFASST DSS* was designed to enable such planning scenarios to be built easily using a Graphical User Interface (*GUI*). In addition, the *DSS* was designed specifically to handle spatial data to manipulate and view fuels treatment schedules and data graphs, but also to utilize data manipulated externally by other software. It was also necessary that spatial data generated in a *GIS* be in shapefile form. Shapefiles are used because they may be read in both proprietary and open-source *GIS*.

A simple *GUI* was designed to enable analysts to quickly set up a scenario in the *iFASST DSS*. Figure 7.2 shows an example of the *GUI* used to build the scenario file, called a Basefile in *iFASST*. The Basefile stores all of the data-file linkages and settings required to solve a fuels treatment scheduling problem, and how to display results. In the left side of Fig. 7.2 there are several drop-boxes within the section titled “Required Files for Basefile”. Each of these drop-down boxes contain a list of files or options representing a data-type. For example, the first drop-down box references the shapefile representing the spatial data of the forest. When a file is chosen from the drop-box list, a map showing the Planning Units Containing Splats (*PUCS*) of the forest is shown, and the shapefile is checked for required attributes used for scheduling and solution output purposes. In addition, the “Shapefile Field Selector” portion of the Basefile Creator window becomes active (not greyed out), and the drop-box for the next required data setting becomes active if the minimum required spatial attribute fields are present.

The shapefile field selector lets the analyst select spatial attributes that they would like to use in modeling, as well as determining how they want an attribute displayed. An attribute could be displayed in chart form, or as a sum total on the “Left” portion of the *iFASST DSS* display.

When specifying the external data-file to use, the “Data File Field Name Settings” area becomes active when an item is selected from the “BaseData File Name”

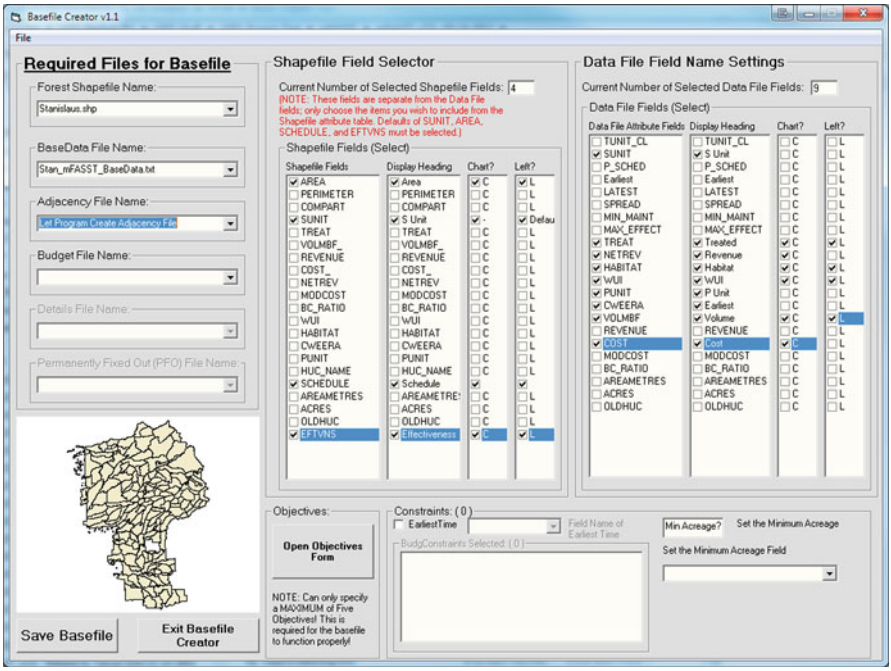


Fig. 7.2 Setting up and saving a scenario as a basefile using the basefile creator

drop-box. The selected file is checked to ensure that required attribute field headers are present. The Data File contains data that has been processed externally from a *GIS* that an analyst wishes to use in the *iFASST DSS*. The "Data File Field Name Settings" area specifies how attributes contained in the Data File should be displayed in *iFASST*.

The other "Required Files" drop-boxes contain options or links to required data. For example, the "Adjacency File Name" drop-box contains a setting that allows the user to specify if the *iFASST DSS* should create the adjacency file, or if a previously computed one should be used. The adjacency option identifies those *PUCS* that are adjacent to one another. A budget data file must also be selected which contains a fiscal budget for the forest by year for the planning horizon, as well as the maximum level of sensitive habitat that may be disturbed by year for the planning horizon. A details file contains information specifying the planning horizon to be considered, as well as the number of ranger districts within the forest to consider scheduling projects in. The permanently fixed out (*PFO*) file contains a list of *PUCS* that should not be considered by the *iFASST DSS* because they are private inholdings that are managed outside the purview of the forest service, if present.

In addition to the "Required Files" the Basefile Creator also contains a "Constraints" section where constraints of the scenario may be specified. For example, the "Earliest Time" constraint may be included by selecting the check-box. If the constraint is selected for inclusion, an attribute field header representing the earliest

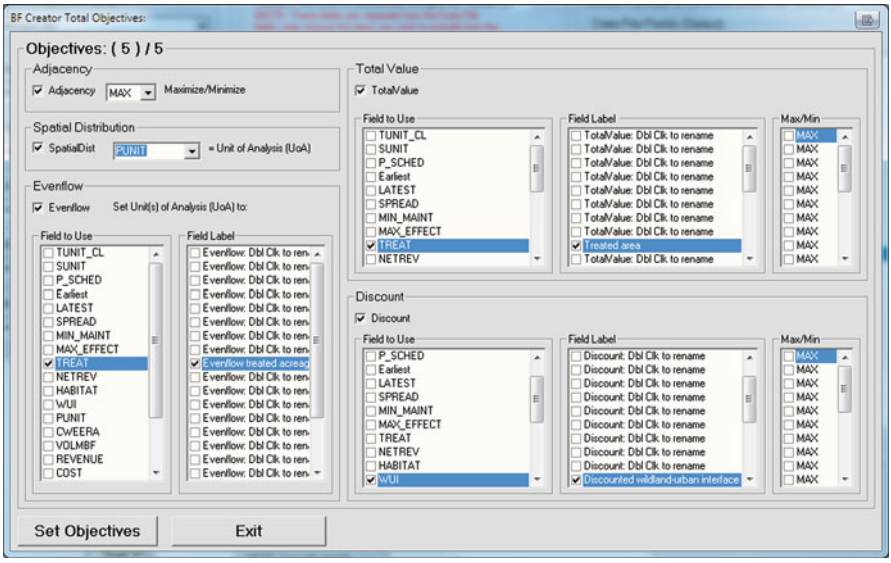


Fig. 7.3 Basefile creator objectives specification

time a project could theoretically be scheduled must be chosen from the now-active drop-box. Similarly, budget constraints may be selected. Budget constraints represent attributes used for fiscal budgeting or for environmental impact such as that related to habitat. In addition, a constraint specifying the minimum level of treatable acreage to be considered as a viable project may also be specified. Once all of the data-files and constraints that an analyst wishes to use to represent a planning scenario are selected, the analyst may specify the objective or objectives that they want to have considered in the *iFASST DSS* planning scenario.

Figure 7.3 represents the window where the objectives to be used in the planning scenario are specified. A maximum of five objectives may be specified for consideration in the *iFASST DSS*. In Fig. 7.3, five example objectives have been selected. One of them is the adjacency objective. If adjacency is maximized, *PUCS* will tend to be scheduled adjacent to one another (clustered) in the same time period or in the next subsequent time period if possible. If adjacency is minimized, projects will be spread out at the same or over time (dispersive tendency) with an attempt to not schedule two adjacent *PUCS* in the same time period t or in time $t+1$. The workflow (called "Spatial Distribution" in Fig. 7.3) objective involves projects scheduled in ranger districts. When this objective is specified, the iFASST DSS will attempt to schedule at least one project in each of the ranger districts of the forest in each time period so that ranger district personnel are not idle.

The "Evenflow", "Total Value", and "Discount" sections of the scheduler represent an objective function that may be desirable for more than one attribute. The numbers of attributes selected represent unique objectives. For example if two "Evenflow" fields are selected, this would represent 2 objectives. The "Evenflow"

objective attempts to minimize the year-over-year variance of projects scheduled over the planning horizon. For example, if amount of treated acres (*TREAT*) is selected, the year-over-year variation in the number of treated acres is minimized. Similarly, the "Total Value" objective maximizes the sum total of an attribute. For example, if treated acres is selected, the total number of treated acreage would be maximized. The "Discount" objective provides a "discount" to projects scheduled toward the beginning or end of the planning horizon. For example, if an analyst wants to preferentially treat Wildland Urban Interface (*WUI*) lands earlier in the planning horizon, the analyst would maximize the *WUI* field as a discounted objective value. If the analyst wished to schedule *WUI* treatments later in the planning horizon, the analyst would minimize the *WUI* field by checking the box in the "Max/Min" list.

Once all of the objectives an analyst wishes to include in a scenario have been set, the analyst clicks the "Set Objectives" button and the specified objectives are saved. The analyst then saves the scenario data as a file. If the analyst wishes to modify a scenario, a previously saved scenario may be opened, updated, and then saved as a new scenario. The next section highlights how a scenario is solved and how solutions may be mapped and displayed as data-graphics.

7.4.2 Using the iFASST DSS to Solve a Scenario and Display Results

The *iFASST DSS* solves a planning scenario by loading the Basefile representing it. The analyst may then weight the previously specified objectives to give added emphasis to a particular, or set of, objectives.

Figure 7.4 shows the *iFASST DSS* Graphical User Interface (*GUI*) with the Heuristic solution window, the Chart window, and a mapped solution. The map, heuristic, and chart are all given in the right region of the *GUI*. On the left side of the *GUI*, the "Left", totals are given in this panel. Those elements appearing in this panel were specified in the Basefile as "Left" for "Left Side Display". The map legend is given in the upper left side of the main display. Here each year is given a color, so the thte schedule can be viewed as a transition of colors on the map itself. The Heuristic window allows the analyst to set weights for objectives specified in the planning scenario. In addition, it enables the analyst to manually schedule a *PUCS* during a fixed time period, or to place it out of consideration as part of a solution. The analyst can save the *PUCS* settings which may be re-opened for use later. The Chart window allows an analyst to chart a particular attribute as a function of the planning horizon. Figure 7.4 shows the treated acreage for the mapped schedule for example. Clicking on a charted attribute during a particular year will highlight the *PUCS* scheduled for that year in the map as well.

Mapping a solution interactively is also very helpful for an analyst. To achieve this, the displayed map will update and color *PUCS* according to the year they were scheduled in, if they were not able to be scheduled, or if there was not enough

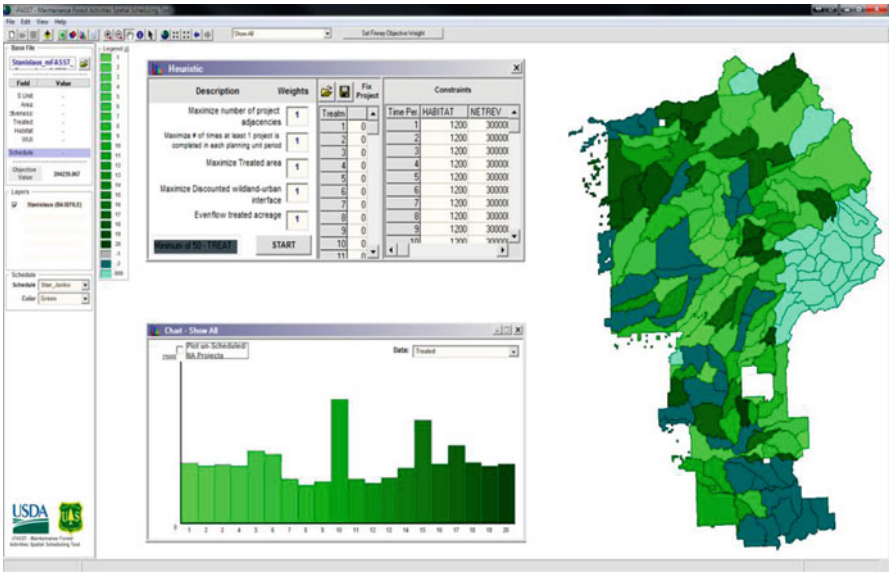


Fig. 7.4 Example of *iFASST DSS* Graphical User Interface (*GUI*) with Heuristic parameters, data-graph, and mapped

acreage present for treatment based upon the minimum treatment acreage constraint. There are also several map-query functions that have been built into the *iFASST* display map. For example, identifying projects scheduled for a particular year may also be done by left clicking on the map schedule legend entry for that year. Similarly, projects scheduled up to that particular year may also be shown by right clicking on a year. In addition, the map may be re-colored by clicking the color next to a year and colored using a color that an analyst chooses. Being able to change color schemes is very important for those who have some level of color blindness. Additionally, pre-set color ramps are also included for quick re-coloring using a variety of color schemes. Mapping functions are also included to allow the user to query individual *PUCS* for attribute and schedule information to be displayed on the left hand side information area. Also included are map re-zooming functions and the ability to display a local mini-map showing the forest as a whole and the larger mapped portion.

All of these *DSS* features enable an analyst to quickly set up, run, and tweak a planning scenario. In addition, they enable an analyst to interact with a solution using the map functions and chart features. This ability has enabled this *DSS* to be used on a variety of forests seeking to reduce forest fuel loads while also meeting several other varied objectives. Without this ability, forest fuels treatment scheduling in California would not have been as effective.

7.5 Application and Implementation Issues

As we stated in the introduction, the US Forest service is divided into operational regions. Each region contains a number of national forests, and each is directed by a supervisor. This means that there are 17 *USFS* supervisors, each directing an almost self-contained organization within region 5. The Regional Forester is the administrative head of the region. Each regional headquarters contains a staff, which helps support different functions across the *NFs* as well give main directives to each supervisor, however, there is often considerable flexibility as to how such directives are to be met within each *NF* and their subunits, ranger districts. Consequently, operations are loosely coordinated as well as differ due to local circumstances. For example, chaparral scrub and oak woodlands, sierra mixed conifer forests and redwood forests all differ considerably. Because of the flexibility in how the *iFASST* model could be used, applications across these different domains were possible.

In those forests where the *iFASST* program was used extensively, results were used to help create integrated vegetation and fuels management plans. Its use was valuable for three principal reasons. The first reason was that the heuristic was designed to solve a problem to near optimality in less than 60s. This meant that analysts could test a number of different constraints and levels, and quickly generate an understanding of what could be realistically achieved. That is, the model could be used to learn the interplay among the constraints in a timely manner, something that could not be done if an optimal solver took days to solve a given problem. The second main reason for its use is that one could easily generate an estimate of the cost of "promises" that had been made to special interests (or to cabin owners for that matter) or desires of forest managers, by fixing in specific treatments in the early years of the schedule. Then the differences in outcomes (of treated land, protected habitat, etc.) between an unfettered solution and one with such desires fixed in would represent the cost of meeting these "deals/negotiations." The third main reason was that the program helped to "cap" expectations. Most planners assumed that far more could be done in terms of treatments and meeting "Finney conditions" than what was possible in a given time frame. Even though the program would be used to schedule treatments out 20 years, the interest was fixed on what needed to happen in the first 5 years, especially ensuring what was scheduled in the first 5 years did not preclude long term targets from being met. Foresters realize that conditions change over time, and that virtually every *NF* will need to consider possible changes in their plan 5 or 10 years from now, based upon natural disturbances, like pests, drought, and fire. Overall, the *iFASST* program helped in understanding the implications of making a near term (5-yr) schedule on a 20 year plan. One forester characterized this as an attempt at making near term actions while ensuring long term sustainability.

7.6 Final Comments and Conclusions

Wildfire can be very destructive and, when severe, produce ground-clearing events. Past harvesting activities have reduced the size and number of oldgrowth stands in California, and any further losses threaten the viability of a number of threatened species, including the California Spotted Owl. To address the concerns of biologists, the public, and special interest groups, Congress passed the Healthy Forests Restoration Act in 2003. This along with the directive of the Regional Forester in California outlined a policy for reducing fire severity by fuels treatments in order to protect threatened old-growth stands as well as protect cabins, homes, recreational facilities within the interface between urban and wildland areas. These activities include prescribed burns during appropriate weather and plant moisture conditions, mechanical removal of litter and downed woody debris, and selective thinning. Past research by Finney (2004) demonstrated through the use of fire simulation models that fire severity could be lessened by selective fuels treatments, rather than wholesale treatments across the entire forests. A planning process was developed by the Region 5 analysts and researchers at *UCSB* to meet this challenge. At the center of this process is a specialized fire modeling team, the Fire Cadre, who assisted staff at National Forests in generating a plan of treatments as well as a 20-year schedule, that address spatial, budget, and management objectives. At the heart of this planning paradigm is a fire simulation model and a decision support system, *iFASST*.

This chapter has presented the details of the spatial optimization model embedded in *iFASST*. The model addresses the major issues in producing the "Finney condition" across the forest as a whole with a special emphasis in clustering activities spatially around the *WUI* as early as possible. As each National Forest has similar but different problems, *iFASST* was designed to be flexible in both constraints and objectives. It is because of this flexibility in model formulation, data manipulation, and scenario generation, that the *iFASST* has been used in the largest of these National Forests. Overall, the system allows planners to test different budget levels, objective weights, and even fix the time in which some *PUCS* are treated, aiding the planners to test a variety of strategies. At this time *iFASST* has now been used in 11 of the 17 forests in California. Due to its success, work was recently completed to extend the *iFASST* system to maintenance treatments so that the Finney condition is maintained over time.

Acknowledgements We would like to acknowledge the research funding of the US Forest Service which supported the development of the *FIRS* model and the *iFASST*/*mFASST* program. We also would like to acknowledge the help and assistance from Tanya Kohler and other staff members at the Region 5 headquarters of the US Forest Service.

References

Andrews PL, Bevins CD, Robert C (2003) BehavePlus fire modeling system, version 2.0: Users Guide. Gen. Tech. Rep. RMRS-GTR. U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station, Ogden

Bahro B, Barber KH, Sherlock JW, Yasuda DA (2007) Stewardship and fireshed assessment: a process for designing a landscape fuel treatment strategy. USDA Forest Service Gen Tech. Rep. PSW-GTR-2003, 41–54, Vallejo, CA, USA

Blackwell J (2004) Sierra Nevada Forest Plan Amendment–Final Supplemental Environmental Impact Statement. Record of Decision, Department of Agriculture, Forest Service, US Government

Camm JD, Chorman TE, Dill FA, Sweeney DJ, Wegryn GW (1997) Blending OR/MS, judgment, and GIS: restructuring P & G's supply chain. Interfaces 27(1):128–142

Church RL, Murray AT, Weintraub A (1998) Locational issues in forest management. Locat Sci 6:137–153

Collins BM, Stephens SL, Moghaddas JJ, Battles J (2010) Challenges and approaches in planning fuel treatments across fire-excluded forested landscapes. J For 108(1):24–31

Feo T, Resende MGC (1995) Greedy randomized adaptive search procedures. Glob Optim 6:109–133

Finney MA (2001) Design of regular landscape fuel treatment patterns for modifying fire growth and behavior. For Sci 47(2): 219–228

Finney MA (2004) FARSITE: fire area simulator—model development and evaluation. U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station, Ogden

Finney MA (2006) An overview of FlamMap Fire modeling capabilities (in Fuels Management-How to Measure Success: Conference Proceedings), USDA Forest Service Proceedings RMRS-P-41; Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station. 213–220

Geoffrion AM, Powers RF (1980) Facility location analysis is just the beginning (if you do it right). Interfaces 10(2): 22–30

Glover F, Laguna M, Martí R (2000) Fundamentals of scatter search and path relinking. Control Cybern 29(3):653–684

Gruell GE (2001) Fire in Sierra Nevada forests: a photographic interpretation of ecological change since 1849. Mountain Press, Missoula, MT

Kent B, Bare BB, Field RC, Bradley GA (1991) Natural resource land management planning using large-scale linear programs: the USDA Forest Service experience with FORPLAN. Oper Res 39(1):13–27

Miller JD, Safford HD, Crimmins M, Thode AE (2009) Quantitative evidence for increasing forest fire severity in the Sierra Nevada and Southern Cascade Mountains, California and Nevada, USA. Ecosystems 12(1):16–32

Niblett MR, O'Hanley JR, Church RL (2014) Scheduling fuels reduction projects using spatial optimization. Unpublished manuscript

North M, Innes J, Zald H (2007) Comparison of thinning and prescribed fire restoration treatments to Sierran mixed-conifer historic conditions. Can J For Res 37(2):331–342

Powell, Bradley E. (2001). Sierra Nevada Forest Plan Amendment Environmental Impact Statement. Forest Service, Department of Agriculture. US Government.

Resende MGC, Ribeiro CC (2005) GRASP with path-relinking: recent advances and applications. In: Ibaraki T, Nonobe K, Yagiura M (eds) Metaheuristics: progress as real problem solvers, vol. 32. Springer, Boston, pp 29–63

Chapter 8
Locating Intelligent Sensors on a Transportation Network to Facilitate Emergency Response to Traffic Incidents

Tejswaroop Geetla, Rajan Batta, Alan Blatt, Marie Flanigan and Kevin Majka

8.1 Introduction and Motivation

According to the Federal Highways Administration (*FHWA*) study presented in Office of Highway Policy Information (2009) there are 254 million registered motor vehicles in the U.S. Each year this number continues to grow, increasing utilization of the road transportation network. In 2009 alone there were an estimated 2.2 million injuries related to traffic incidents and 33,808 fatalities from these injuries. In addition, traffic incident related fatalities ranked sixth in the list of preventable fatalities in the U.S.

Apart from being one of the major causes of fatalities, traffic incidents also represent one of the major causes of roadway congestion. According to a survey presented in (Cambridge Systematics and Texas Transportation Institute 2005), traffic incidents account for 25 % of all non-recurring congestion reported on roads in the US. Observing the ill effects of traffic incidents, the *FHWA* turned to Intelligent Transportation Systems (*ITS*) to help increase safety and reduce congestion on current roadways.

Nagare (2012) define Intelligent Transportation Systems (*ITS*) as an application of advanced technology to solve transportation problems. *ITS* uses advanced communication, sensor and computer technology to address transportation problems and enhance the efficiency of the system for movement of people and goods. There are many uses of *ITS* systems. These include systems that help reduce the likelihood of a crash, usually based on sensors that issue warnings regarding approaching vehicles. They also include systems that provide congestion data to users in a form that aims to impact driver behavior on route choices. Another area of application is in controlling entry of traffic onto highways to control congestion. There are

R. Batta (✉) · T. Geetla
Department of Industrial and Systems Engineering, University at Buffalo (SUNY), Buffalo, NY, USA
e-mail: batta@buffalo.edu

A. Blatt · M. Flanigan · K. Majka
Center for Transportation Injury Research, CUBRC, Buffalo, NY, USA

H. A. Eiselt, V. Marianov (eds.), *Applications of Location Analysis,* International Series in Operations Research & Management Science 232, DOI 10.1007/978-3-319-20282-2_8

emerging application areas of *ITS*, one of which is traffic-incident management. This real-world application is the major focus of this chapter.

An effective traffic-incident management system needs the following functions:

- Rapidly detect a crash
- Characterize a crash and crash scene for incident management
- Deliver fast emergency medical services to the injured
- Provide crash related information to drivers and highway managers to reduce congestion

At the core of any incident-detection technology is the use of sensors and wireless communication technology. Sensors are responsible for incident detection and the role of wireless communication channel is to transfer sensor readings to a data collection node/location. A sensor is an instrument that measures a physical quantity and reports to the user the measured quantity. In the road network systems, sensors quantify the real road environment (e.g., congestion, weather conditions) and serve as tools to alert and help drivers and traffic managers. For example, many cars are now equipped with an on-board *GPS* with functionalities like navigation. The modern navigation design on a car can collect traffic and congestion data through radio communication to alert a driver of potential congestions and re-route a trip to improve their travel time see, e.g., BMW USA (2008). This capability arises from the use of *GPS* along with data communication capabilities.

There are many sensors used by *ITS* systems for incident detection and management. We classify the sensors broadly into stationary and mobile. A stationary sensor has a fixed location on the road network and monitors incidents on the road network close to its location. Acoustic sensors and traffic cameras are some examples of stationary sensors. Mobile sensors move continuously in and out of traffic. These mobile sensors are mounted on vehicles, which could be regular vehicles or public utility vehicles like buses and taxis. They can serve both as sensors which operate only in a crash situation or as sensors that relay continuous information. They can also serve as computational engines that fuse or aggregate data from neighboring sensors.

These mobile sensors may be operational only during some part of the day. For example, Srinivasan et al. (1997) proposes and uses motor vehicle probes that move along with traffic collecting vital road data like traffic flow, travel time, etc. The mobile sensors that act as incident detection sensors observe and monitor only the surroundings of their real-time position with certain functionality activated only if they are involved in an incident. For example, a product offered by General Motors (*GM*) called OnStar Automatic Crash Notification (*ACN*) will detect a crash (we note that an earlier version of the OnStar product was a factory-installed *ACN* system). Other examples of mobile sensor systems used in crash detection include the upgraded OnStar (factory installed) system called Advanced *ACN* (*AACN*) see, e.g., (OnStar 2012), BMW (2010) and White et al. (2011).

Both mobile and stationary sensors with communication capabilities have an added advantage in that they can support data fusion. Data fusion is the application of mathematical techniques to combine data from multiple sensors. The data

in question can be of many different types, e.g. sound, change of speed, weather conditions. The key feature of data fusion comes in when data of different types in collected related to a common event, e.g. a crash. In this case, association of this data towards this event needs to be made and furthermore appropriate fusion of this data leads to a more comprehensive understanding of the event, since different sensors collect information on different features of the event. Application of this technique to incident detection will help combine data from multiple sensors classes to better detect traffic incidents. In this chapter, we assume that all sensors on the road network (stationary or mobile) have computational capability and are able to use data fusion techniques.

With all the sensors available for incident detection, present day-road sensor networks are essentially multi-sensor networks, implying that they have a variety of different sensors. In a multi-sensor environment, each independent sensor observation becomes a data point and all data points are collected and analyzed together to create a situational awareness for the observer, meaning that various aspects related to the event (situation) are developed. All sensor data needs processing to convert the data into useful information. For example, two cars are involved in a frontal impact motor vehicle crash. Assume that one of the cars is equipped with an *AACN* and there is an acoustic sensor located very close to the crash. Both sensors generate unique crash data. These two independent data sets have to be pooled together first and then combined to extract crash information. Multi-sensor data fusion techniques propose a systematic data combination to extract accurate information.

Researchers developed multisensor data fusion to combine data generated by a large set of sensors for Department of Defense (*DoD*) applications in the 1980s. This development combined ideas from various research disciplines like signal processing, statistics and artificial intelligence. In recent years, data fusion has developed into a research field of its own. Data fusion is a collection of efficient methods for combining data from disparate sensor resources and databases to generate a unified and in some ways better information compared to individual sensor data, see Hall et al. (1997). This book also contains many lucid examples of how data fusion can provide better information. In fact it provides cases where data fusion can be used to predict future events. Our use of data fusion is at an elementary level, in that we are trying to better characterize an event that has happened for appropriate emergency response, not predict future occurrences of events.

One of the advantages of data fusion is it helps in authenticating a physical quantity measurement observed by multiple sensors (for example, if there are multiple sensors in a chosen area making temperature measurements, which become critical in winter driving situations as they serve as a surrogate for slippery driving conditions). Each sensor records a temperature measurement. Data fusion techniques, using statistical estimation, help to improve confidence, reliability and reduce ambiguity in temperature readings. The other advantage of data fusion is that once this estimation of a physical quantity is complete, other observations use the estimate in other data fusion processes. For example, using the temperature estimate generated by multiple sensors, the acoustic sensors estimate the speed of sound (which varies with air temperature and use it to estimate position of the source of sound.

Data fusion provides the availability of estimated data across all sensors to generate a unified view of the real event.

Working together using data fusion all sensors in the *ITS* systems used in incident detection should actively aim to provide complete information, devoid of any ambiguity, of every crash on US roads. However, considering all the budgetary constraints in deploying such a massive infrastructure, operations research techniques can help in maximizing the potential of partially deployed *ITS* systems. Similarly, the use of simulation can aid in evaluation with relatively small investments.

Purchase of sensors, sensor maintenance and the support data systems may be the three important cost drivers of any *ITS* project. Keeping the budgetary constraints and to maximize the incident detection capability, the placement of the stationary sensors becomes critical. The good news is that placement of stationary sensors can be optimized using operations research techniques.

Many factors can lead to motor vehicle crashes. Some of the commonly known factors are speeding, carelessness, inexperience, distraction, loss of control over the vehicle, etc. All the conditions are mistakes originating from the driver side and the last condition may occur due to unforeseen road weather conditions. Driver mistakes are the major cause of accidents, which happen more often at intersections and entry-exit ramps, where a large number of vehicles move through and interact. An assessment of historical crash data can provide insights into accident-prone zones or areas that we can target for sensor installation. In this chapter, we use previous accident data to guide sensor placement with a focus on acoustic sensors.

We proceed to state the problem and its application, which is to understand sensor technology used in incident detection and use this knowledge to improve the placement of stationary sensors. This chapter does not consider issues related to data communication or data transmission. The assumption is that every sensor can communicate with all the sensors and are capable of transmitting their data to a central base location or a local data collection sensor. This assumption will be true in the future as the *DOT* is working on improving communications as part of the Vehicle-to-Vehicle (*V2V*) and Vehicle-to-Infrastructure (*V2I*) initiatives.

In Sect. 8.2, we present background on sensor coverage models, present the solution methodology for the sensor placement problem, and develop a simulation model used in the sensor placement model as both an evaluation tool and as an optimization technique. In Sect. 8.3, to demonstrate the proposed solution methodology, we use omnidirectional sensors as a case study and present the results. In Sect. 8.4, we summarize the findings from our study and propose future research directions.

8.2 Sensor Location Problem

The benefits of data fusion are only realized when multiple sensors observe or detect the same incident. This requires efficient placement of sensors to provide at least double coverage (i.e., a minimum of two sensors covering the same part of the road segment). Single coverage implies that only one sensor covers a part of the

road segment. This single coverage is a less favorable outcome for the placement of sensors than double coverage.

The placement of sensors will greatly influence the effectiveness and utility of data fusion technology. With a limitation on the number of sensors, a tradeoff between multiple sensors covering a smaller area versus single sensor coverage of a larger area becomes apparent. The output from the data fusion techniques is dependent on the distance between the sensors and the crash location. Data fusion of the information from multiple sensors will take into consideration the real world uncertainties. Since the output of data fusion is very dependent on the location of the crash, we need a mechanism to capture the uncertainty in the location of future crashes when assessing the value of using data fusion. One way to assess this uncertainty in crash location is through simulation. Randomized crashes generated using simulation help evaluate the detection capability of acoustic sensors and data fusion opportunities. Statistical evaluation of the effectiveness of a particular sensor placement design is necessary. Experimental design serves as a tool to achieve this goal.

Operations research techniques have helped in innumerable sensor placement problems and facility location problems to place sensors. The research question here is to use the same techniques to help place incident detection sensors in a traffic environment to detect incidents and provide a data fusion-capable environment where multiple sensors interact. Section 8.2.1 presents the literature background for the sensor placement problem. The section also highlights the research gap this chapter aims to fill.

8.2.1 Literature Review of Sensor Placement Problem

Sensor placement problem is a common mathematical problem arising in a large distributed sensor network. One easily relatable sensor location problem is the security guard location problem (also called an art gallery problem) in a museum, whose objective is to cover every display piece inside the museum by using the smallest number of guards. The problem has to take into consideration the geometric shape of the display area inside the room. Other research interests in the sensor placement problem involve development of heuristics or algorithms for the sensor placement problem. Dhillon und Chakrabarty (2003) presents two algorithms for placement of sensors in a sensor field modeled as a grid. Nie et al. (2012) develops the problem of locating bomb-detecting sensors in airports or shopping malls, to counter terrorist attack by increasing the detection capability. Gentili und Mirchandani (2005) presents a location of traffic flow sensors on a road environment for applications like flow measurements and time to travel a path. One of the important measures in traffic management is time of travel between an origin and destination pairs. Gentili und Mirchandani (2005) proposes the use of traffic flow sensors to estimate the time to travel between an O-D pair. A set-covering formulation is developed. Since this is an NP-hard problem, they use a heuristic solution algorithm.

8.2.1.1 Coverage Problems

Building on the location set-covering concept, the maximal coverage location problem proposed in Church und ReVelle (1974) aims to maximize coverage using limited resources. For example, in a cellphone base-station location model, the demand can be customers who receive and transmit radio waves; or in a warehouse-location-problem, the demand is from potential shops or malls. Coverage problem formulations model demand in three ways. In the first method, the demand occurs at specific nodes. The warehouse location problem is a good example that uses the node demand, where malls represent the demand nodes. In the second method, demand occurs on paths. For a bus-route selection problem where each trip has a specific demand, the path based demand model is used. In the third method, demand occurs at both nodes and paths. A node and path based demand model is used to locate cellphone base stations and emergency helicopter base stations, see Erdemir et al. (2008a, 2008b). In Erdemir et al. (2008a) cellphone base-station location problem demand arises from customers using cellphone from nodes (housing areas, shopping malls, etc.) and paths (roads and highways). For the sensor location problem considered in this chapter, roadway crashes constitute demand since the objective is to maximize the incident detection. However, roadway crashes can happen anywhere on roadways; there are no predictive mechanisms to forecast incident occurrence. To plan the sensor location problem we assume previous crash data serve as a predictor for future accidents. However, a key observation from previous crash data is that the crash demand split into crash paths and crash nodes. Crash nodes represent high accident-prone locations (road intersections, entry and exit ramps of a highway) and crash paths representing the rest of the road network. Crash paths represent the other road segments that are not part of the crash nodes. The crash paths can be parts of highways where accidents are equally likely to happen at any part of the road segment.

Using the above observation from past crash data, where crashes happen on both crash nodes and crash paths, the major goals of the present research are to improve both incident detection and data fusion capabilities in this multi sensor network. However, there is not much research on how to improve data fusion in a coverage problem. A regular max cover problem has a single objective function that maximizes the coverage of every sensor (we note that Berman et al. 2013, have recently considered continuous coverage). There can be extensions made to this formulation to increase secondary coverage that might increase the data fusion capabilities.

There are many benefits provided by improving the data fusion capabilities, which translates into requiring secondary coverage. The main drawback of data fusion lies in the fact that it is made possible through secondary coverage, which implies a reduction in primary coverage. While primary coverage emphasizes detection capability, secondary coverage emphasizes data fusion capabilities. In this chapter, we plan to explore the interesting research question of the trade-off between these two objectives.

8.2.2 Sensor Classes and Solution Framework

There are many sensor classes used for incident detection. For this chapter we chose three sensor classes:

i. Omnidirectional sensors
ii. Traffic cameras
iii. Traffic flow sensors/traffic count sensors

Acoustic sensors used in traffic incident management belong to the class of omnidirectional sensors (unlike microphones, which are directional acoustic sensors). There are many *ITS* applications that use acoustic sensors. Traffic cameras are visible at major intersections and these sensors actively monitor road segments. A representative traffic-count sensor (whose role, for example, is to reveal sudden changes in traffic flow which is a typical sign that a crash has occurred) is an inductive loop detector embedded inside the road segment. Although this chapter focuses on placing these three sensor classes, extensions or modifications are possible to the general approach presented here for placement of sensors to suit other sensor classes as well.

Considering the three sensor classes, we propose a solution methodology that involves three steps, which can be independently applied to all three sensor classes. The first step formulates the problem as a coverage problem. We call this the explicit model. The explicit model resembles a maximal coverage problem. The second step involves a pure geometric approach to solve the sensor location problem. We call this the implicit model. Unlike the explicit model that only considers a path covered if and only if the entire path is covered, the implicit model allows partial coverage. The implicit model to solve the problem lets the decision maker/s decide the percentage of the path that should be overlapped by the coverage of two or more sensors. This overlap area provides a good opportunity for collecting more data and improving the quality of information through data fusion. The implicit model can be used as a separate method to solve the sensor location problem, independent of the explicit model. The solution of the explicit model provides an excellent starting point for the implicit model.

As illustrated in Fig. 8.1, a simulation based optimization procedure follows after the implicit model's application. This is the third step in our solution methodology. Simulation-based optimization first evaluates all the generated feasible solutions and then proceeds to use a local-search heuristic to generate other feasible solutions. Solutions that result in an improved objective function will iteratively use local search to reach better solutions. The addition of simulation-based optimization to a coverage problem is unique and to the best of our knowledge never tried before. In the next section, we present the details of the explicit and implicit models.

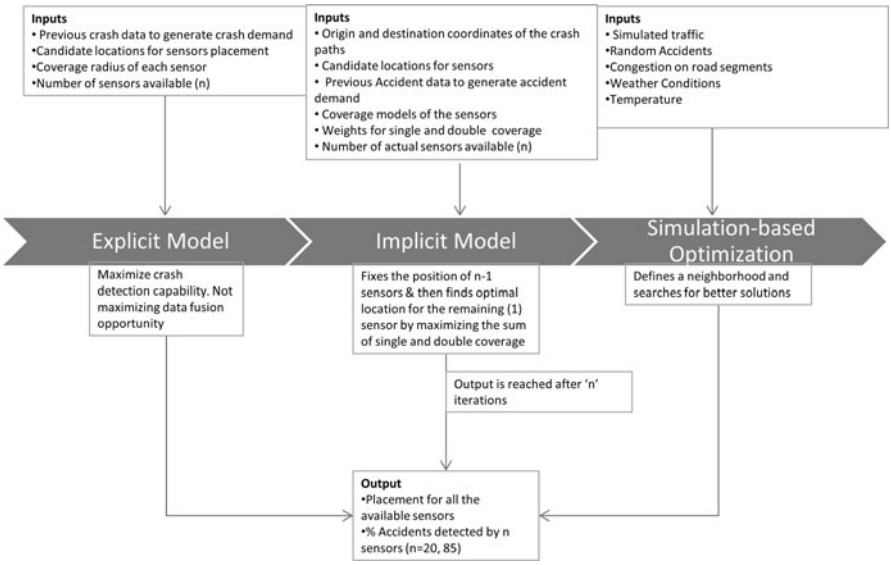


Fig. 8.1 Solution methodology for the sensor placement problem

8.2.3 Explicit Model

The inputs to the explicit model are:

i. Node and path demands (in our case, calculated using previous crash data)
ii. Candidate locations for sensors (in our case, a set of feasible sensor locations on infrastructure with access to power)
iii. Coverage criteria for every sensor class
iv. Number of sensors available from each sensor class

The output of this model is the location of the available sensors.

A node is 'covered' if it is in the coverage zone of at least one sensor. A path is 'covered' if each point on the path is in the coverage zone of at least one sensor. The model enforces these coverage rules by the constraints. We now proceed by introducing relevant notation, followed by our model formulation. Our model is identical to the one found in Erdemir et al. (2008b) except that a single period case is considered.

The *input* data for this model is as follows:

M set of potential sensor locations
N set of nodes
P set of paths
p_1 total number of sensors to be located from omnidirectional sensors
p_2 total number of traffic cameras to be located
p_3 total number of traffic flow sensors to be located

H_j weight of node j
H_k weight of path k
A (Ai,j,m), where Ai,j,m = 1, if sensor at location i of type m covers node j, and 0, otherwise
B (Bi1,m1,i2,m2,k), where Bi1,m1,i2,m2,k = 1, if sensors at i1 of type m1 and i2 of type m2 cover path k, and 0, otherwise
B' (B'i1,m1,k), where B'i1,m1,k = 1, if sensor at i1 of type m1 covers path k, and 0, otherwise

The *outputs* of the model are as follows:

$x_{i,m}$ 1, if a sensor of type m, is located at i, and 0, otherwise
z_j 1, if node j is covered, and 0, otherwise
$\ell_{k,1}$ 1, if path k is covered by a single sensors, and 0 otherwise, and
$\ell_{k,2}$ 1, if path k is covered by a pair of sensors, and 0, otherwise

The formulation of the model is as follows:

$$\text{Max} \sum_{j \in N} H_j z_j + \sum_{k \in P} H_k(\ell_{k,1} + \ell_{k,2}) \forall m \in \{1,2,3\} \tag{8.1}$$

$$\text{s.t.} \sum_{i \in M} x_{i,m} \leq p_m \forall m \in \{1,2,3\} \tag{8.2}$$

$$\sum_{j} \sum_{m} A_{i,m,j} x_{i,m} \geq z_j \forall j \in S \text{ and } m \in \{1,2,3\} \tag{8.3}$$

$$\sum_{i1,m1,i2,m2} B_{i1,m1,i2,m2,k} x_{i1,m1} x_{i2,m2} \geq \ell_{k,1} \forall k \in P \tag{8.4}$$

$$\sum_{i1,m1,i2,m2} B'_{i1,m1,k} x_{i1,m1} \geq \ell_{k,2} \forall k \in P \tag{8.5}$$

$$\ell_{k,1} + \ell_{k,2} \leq 1 \tag{8.6}$$

$$x_{i,m} \in \{0,1\} \tag{8.7}$$

$$z_j \in \{0,1\} \tag{8.8}$$

$$\ell_{k,t} \in \{0,1\} \forall t \in \{1,2\} \tag{8.9}$$

In the model, the objective function (1) maximizes the total coverage, first term maximizing the coverage from the nodes and the second term maximizing the coverage from the path. Constraint (2) limits the number of acoustic sensors to p_1, limits the number of traffic cameras to p_2, and limits the number of traffic flow sensors to p_3. Constraint (3) defines node coverage and declares that a node j is "covered" if and only if at least one sensor covers node j. Similarly, constraint (4) defines path coverage and declares a path "covered" if at least one pair of acoustic sensors cover it or a pair of traffic cameras cover it, or a pair of traffic flow sensors cover it. Constraint 5 defines path coverage and declares a path 'covered' if a single traffic flow sensor,

a single acoustic sensor, or a single traffic camera covers the entire path. Constraint (6) ensures that a single sensor or a pair of sensors can "cover" a path, and only one term can enter the objective function. Constraint (7), (8) and (9) ensure that the decision variables are all binary.

The paper Erdemir et al. (2008a) uses a Greedy Paired Adding Heuristic (*GPAH*) to solve the above quadratic maximum coverage location problem (*qMCLP*). *GPAH* is an efficient and computationally faster algorithms compared to other algorithms. From computational complexity theory, we know that *MCLP* is a hard problem to solve, we apply the same *GPAH* heuristic with minor variations to solve *qMCLP*. The *GPAH* solution is within 6.6 % of the optimal in the computational tests performed Erdemir et al. (2008a). Since the *GPAH* algorithm performs well on large-scale problems, we use it for our situation to obtain a starting solution for the implicit model. The *GPAH* is a greedy heuristic that places the sensors by first allocating sensors to the highest weighted node or a path. At every step, the greedy heuristic searches for the highest weighted uncovered node or path. If a node is the highest weighted uncovered, then the greedy heuristic locates one of the sensors near the node. If a path is the highest weighted uncovered, then the greedy heuristic places a pair of available sensors at the closest potential sensor locations on the same path.

Note that objective function or constraints in the explicit model do not include terms dealing with the data fusion capabilities. The implicit model detailed in the next section captures this aspect.

8.2.4 Implicit Model

The implicit model in Erdemir et al. (2008a) provides a geometric (not based on mathematical programming) method to solve the coverage problem with demand originating from both nodes and paths. This model allows the users the flexibility to adjust single, double and triple coverage. We adopt this model for the sensor location problem but only utilize single and double coverage. In the sensor location problem, single coverage implies that a single sensor covers the entirety of a path and double coverage implies coverage of a path by two sensors. If there is a crash in the area covered by two sensors, each sensor makes an independent observation of the crash. Every sensor deployed in the field has a detection range/zone. Inside this detection zone, each sensor observes an incident with a positive probability.

For example, consider a crash scenario inside the detection range of a sensor with true positive probability of 0.7. This means that if there is a crash in its detection area, the sensor detects it 70 times out of hundred. However, if the crash occurs in a region covered by two sensors (assuming the sensors have the same detection rate and are independent), the combination of data from two sensors observes 91 crashes out of 100, which is an improvement of 30 %. The advantage double

coverage provides is with data fusion techniques on data obtained through multiple sensors. The advantages of *ITS* systems with data fusion capabilities are both improved true positive probability and an improved crash characterization.

In this geometric approach, we measure the coverage achieved by a particular placement of sensors on every path and every node. For example, consider a path A with a crash demand of 0.04. Suppose that in a specific sensor placement strategy $\frac{3}{4}$ of the path is covered by at least one sensor, and . . . of the path is covered by two sensors and we assign a weight (the value of the weight reflects the relative importance given to double coverage) of $0.7\alpha_1$ to single coverage and $0.3\alpha_2$ to double coverage. Then this path contributes $0.04[(3/4)0.7 + (1/4)0.3]$ to the objective function.

$$\begin{aligned} \text{Max } & f[(x_1, y_1), (x_2, y_2), ..., (x_{n-1}, y_{n-1}), (x_n, y_n)] \\ & = \alpha_1 s_1[(x_1, y_1), (x_2, y_2), ..., (x_{n-1}, y_{n-1}), (x_n, y_n)] \\ & \quad + \alpha_2 s_2[(x_1, y_1), (x_2, y_2), ..., (x_{n-1}, y_{n-1}), (x_n, y_n)] \end{aligned} \tag{8.10}$$

where s_1 is a function of single coverage and α_1 is the weight of single coverage, and, similarly, s_2 is a function of double coverage and α_2 is the weight of double coverage.

The objective function defined in (10) maximizes the single and double coverage with associated weights.The decision variables are the location of sensors $[(x_1, y_1), (x_2, y_2), \ldots, x_n, y_n)]$. The heuristic methodology fixes the position of n-1 sensors and then finds the optimal location for the remaining sensor using the objective function (8). Each step randomly selects a sensor location from the current solution. Removing a sensor from the chosen location and keeping the rest of the (n-1) sensor locations intact, a new location for the displaced sensor is needed. The new sensor location chosen from the previously un-occupied sensor locations that maximizes objective function in (8). For this heuristic to work effectively, the authors start the heuristic with the solution from the *GPAH* algorithm. We adapt the same solution methodology. There are other approaches to generate improved solutions. For example, a variation to this approach is to find the best location for two sensors in the heuristic. This heuristic can fix the location of n-2 sensors and find the best location for the two sensors. Using the first or the second approach to solve the implicit model does not guarantee an optimal solution.

The inputs for the implicit model are: (*i*) the coordinates of the two end-points of the paths, (*ii*) coordinates of potential sensor locations, (*iii*) mean demand per unit length of a crash path, (*iv*) coverage radius of the sensors, (*v*) weights for single and double coverage, and (*vi*) the number of actual sensors available. Output for the implicit model is the coordinate locations of sensors.

8.2.5 Simulation as an Evaluation Tool

Computer simulation is an imitation of a real-world system using abstract mathematical models representing the real system using computers. Computer simulation

or simulation is used in a number of manufacturing facility systems, road network and service industries to study and improve stochastic processes. See Kelton et al. (1997) and Law and Kelton (1999), for general introduction on use of simulation and general applications.

Simulation or field-testing captures this complex interaction of traffic movement, sensor detection and sensor data accumulation. Simulation as used here is a cost-effective evaluation tool to quantify the effectiveness of the placement strategy. There is thus a need to construct a simulation of traffic on a road network with *ITS* sensors, where crashes randomly occur and where the placement of *ITS* sensors creates data fusion opportunities.

One such simulation tool currently in development is the Automated Situational Awareness Platform "*ASAP*" using Arena (essentially it is a simulation capable of modeling vehicular movement throughout a study area and a validation of the emergency vehicle travel times through historical crash response data in an existing traffic network, see Henchey et al. (2013). In designing *ASAP*, a major goal was to address issues associated with the response to highway-related emergencies. In particular, *ASAP* is used to:

1. Understand how data is accumulated and used in an *ITS* systems to create situational awareness of an emergency event
2. Use the information generated from data fusion to generate emergency response decisions and then evaluate these decisions.

The *ASAP* platform provides the necessary tools to embed sensor modules that can interact with the road network and the traffic. These sensor modules need to reflect the abstract detection model of each sensor. For example, the simulation models a traffic-monitoring sensor like an inductive loop detector. Based on the type of inductive loop detector these sensors can count the motor vehicles moving on the road segments. In the simulation, the inductive loop sensor can be a checkpoint, which counts the motor vehicle entities passing on the road segment. In the real world, a sensor always has error estimating the traffic counts/traffic flow. This real world error will also be included in the simulation. The traffic sensor handbook, see Federal Highways Administration (2009), contains error rates.

The simulation developed in Henchey et al. (2013) consists of three modules. The first module creates a traffic network based on the road network extracted from the case study. More specifically, the traffic network it creates has road segments, intersections, intersection signal lights, motor vehicles and traffic flow at intersections. The module generates traffic using real world data based on expected traffic flow as a function of the time of day, type of vehicle and weather. This module also simulates traffic in different weather conditions. For example during a heavy rain condition, the simulation automatically decreases the traffic speed on the roads to model traffic movement in severe weather conditions. We call the first module "traffic generation module." What we intend to do is to validate the approach on a realistic simulation model, as a surrogate to an actual field test.

The second module simulates a random crash on the road network. This module uses the previous crash data collected for the case study to choose a crash location.

For example, the module chooses a path to create a crash. A random point on the path serves as the actual location of the crash for this run of the simulation. Then the crash creation module creates all the associated parameters, e.g., the number of vehicles involved, number of people in each vehicle, delta velocity (difference in velocities of the objects involved in crash at impact) of the impact, number of impacts, principal direction of force (*PDOF*), and sensor relevant data for the crash. For example, a traffic camera needs a crash scene image for the video image processor software and an acoustic sensor needs a crash acoustic signature for signal processing. The crash creation module uses past accident image and acoustic signature information to create this crash related sensor data. In order to facilitate a random accident we need a large collection of crash sensor information.

The third module is the sensor module. This part of the simulation is for the sensor placement problem. This module imitates a real sensor behavior in a data fusion environment. Significant functions of the sensor module are to use the crash relevant sensor ground-truth data (here ground truth refers to the actual events and conditions, as opposed to that observed by sensors which can have errors) generated by the crash generation module and to create crash specific data sets for every sensor. After the sensor module receives the crash data, the sensor module evaluates the crash. For example if the crash generation module generates a random crash at (43.024592, − 78.798994) with an acoustic signature, the acoustic sensor located at (43.026178, − 78.800012) receives the acoustic signature from the crash. The sensor module calculates the distance between the crash and the sensor location. Through the traffic module, the sensor collects the congestion data (e.g., on the particular road on is 600 vehicles/h) and the local temperature from weather station data. Using this information the sensor module estimates the sound decay and generates a new acoustic signature with decay. This new decayed acoustic signature will be the ground truth for this particular acoustic sensor. Now the sensor module performs incident detection using the decayed amplitude versus time data. If the crash is detected it will calculate the error rate in its estimation. Each acoustic sensor module makes the necessary calculations of sound decay from the acoustic signature of the crash and then uses a threshold detection value to sense the crash.

Each simulation starts first by creating traffic on a road network on a regular day from 7:00 a.m. to 9:00 p.m., with a predefined location of sensors. The crash creation module creates a random crash from the crash-weights given to each road segment from past data. Once the crash generation module creates a crash and the necessary calculation in the sensor modules is complete, the simulation collects the data from nearby sensors and stops the simulation. The sensor data obtained from the nearby sensors is analyzed using signal-processing techniques. The output from the signal processing is the input for data fusion techniques. The output from the data fusion creates an estimate of the crash location and other crash characteristics. To gather statistically significant performance measures like percentage detection, error in estimation of the crash location and the quality of crash estimates, the simulation is run multiple times generating random crashes in every simulation run and keeping the position of the sensors unchanged. We note that the issue of false positives is

not specifically considered here, in that we assume that the necessary emergency response units/resources are available to respond to any detected incident.

8.2.5.1 Simulation-Based Optimization

The next step in the solution methodology is to use a simulation-based optimization procedure. Simulation-based optimization uses simulation as an evaluation tool and integrates an optimization method into the simulation to reach better performing solutions. The simulation evaluates the stochastic system and the optimization tool uses a heuristic, approximation or response surface methodology. Further reading material on simulation-based optimization see Law und Kelton (1999) and Dong (2007).

8.3 Location of Acoustic Sensors to Detect and Characterize a Crash

So far, we have discussed the use of sensors in incident management systems and the sensor placement model for the stationary sensors. A hybrid solution methodology using explicit-implicit and a simulation-based optimization procedure is proposed. In this section, we apply the proposed solution methodology only to the omnidirectional sensors, for which acoustic sensors act as representative sensors.

8.3.1 Case Study for the Acoustic Sensors Placement Problem

We chose a case study to test the explicit-implicit-simulation-based optimization methodology for the omnidirectional sensors. We still follow the methodology proposed but use specific solution methodologies only applicable to acoustic sensors and in general omnidirectional sensors.

We chose a road network that covers a 9-mile square area near the University of Buffalo's North Campus in Buffalo for the case study. This road network consists of an urban-arterial road network with part of an interstate-highway running through it. The area chosen has a university, shopping mall, three schools and other commercial areas. This area represents a road network with moderate traffic flow with peak traffic typically in the early mornings and evenings. The average annual traffic data is a reference to recreate the traffic flow patterns in the simulation, see Greater Buffalo-Niagara Regional Transportational Council (2010). The area chosen is without a hilly terrain and large industries, which is ideal for understanding and testing the placement of acoustic sensors (large industries and hilly terrain would create an added level of complexity into the simulation model). We assume that road noise

from the traffic is the most influential background source in the signal processing for the acoustic sensors, which act as the only incident detection sensor.

The simulation case study collects two important performance measures from every simulation run. The first performance measure is, "percentage of crashes detected" by the acoustic sensors with a given location strategy. The second performance measure is "% crashes detected by two acoustic sensors." This performance metric measures the data fusion opportunity arising from having multiple sources of crash sensor data.

There are many factors, which influence these performance measures. One such factor is the placement of the sensors, which influences both the performance measures. So far, in this study the solution methodology produces three placements of the acoustic sensors, the first from the explicit model, the second from the implicit model and the third, from the solutions obtained through simulation-based optimization.

8.3.1.1 Explicit Model Solution

The explicit model for the acoustic sensor placement problem is an approximate mathematical model. We utilize the mathematical formulation in 2.3 to suit the acoustic sensor placement problem. The objective function remains the same, maximizing the coverage achieved by a fixed number of sensors.

In the area chosen for the case study, there were 3600 reported crashes over a 5-year period. From observing the crash data, we identified 11 locations that have 20 or more crashes occurring in very close proximity. These 11 locations have 784 crashes in total and serve as nodes in the path and node model. The weight (demand) of a node was set equal to the number of crashes that occurred at the node divided by the total number of crashes. The remaining 2816 crashes occurred on the remaining road network. The road network divides into road segments that serve as paths in the explicit model. The weight (demand) of a path was set equal to the number of crashes on the path divided by the total number of crashes. We chose to have 760 paths in the model and of these 760 paths, 208 had zero demand.

To construct candidate sensor locations, we first chose all the ends of a path as potential sensor locations. This choice yielded 400 potential sensor locations. We also generated 469 additional locations by first randomly selecting 469 paths and then for each selected path we choose a random location in the interior of the path. The total number of candidate sensor locations is therefore 869.

The *GPAH* algorithm presented in Sect. 8.2.3 can generate feasible solutions for the acoustic sensor placement problem. In Fig. 8.2, we present the performance of the *GPAH* algorithm with different number of sensors. The results presented in Fig. 8.2 show that the percentage of covered crashes increases as we increase the number of sensors. In Fig. 8.3, we show the sample solution of the explicit model on the case study chosen near the University of Buffalo's North Campus with an 'n' value of 55. The *GPAH* algorithm (developed in *JAVA*®) runs is under 15 min using

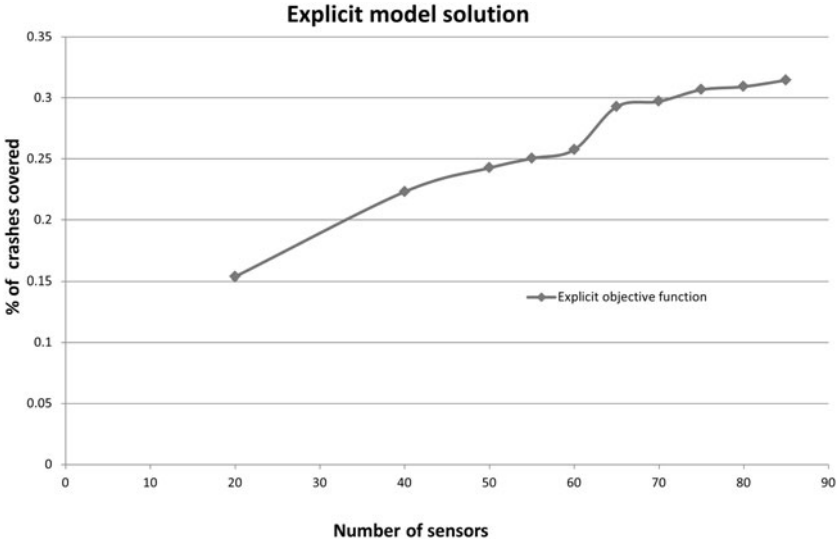


Fig. 8.2 Explicit model solution

Intel® Core™ 2 Duo with clock speed of 3.17 *GHz* and 4 GB *RAM* for the case study chosen.

8.3.1.2 Implicit Model Solution

The Implicit model proposes an alternate solution methodology to the sensor placement problem. The implicit model is inherently a local search heuristic to improve a solution using a different objective function from the explicit model.

Observing the objective functions in Fig. 8.4 gives an impression of how much the implicit model improves over the explicit model. This behavior is expected as the implicit solution uses the explicit solution as a starting solution and improves it using the local search heuristic.

Figure 8.5 presents the implicit model solution on the same case study. Comparing Figs. 8.5 and 8.3, observe that sensor placements in the implicit model solution are clustered (i.e. several sensors located nearby) whereas the explicit model solution has scattered placement of sensors. This clustering of sensors in the implicit model is the result of having a secondary coverage term in the objective function. Such secondary coverage consideration is not part of the explicit model.

From a computational perspective, this implicit model (developed in JAVA®) uses the explicit model solution as a starting solution and takes a computation time of 3 h using *Intel*® *Core*™ 2 Duo with clock speed of 3.17 Hz and 4 GB *RAM*, compared to the computation time of less than 15 min for the explicit model. These computation times reported here are for the case study presented in Sect. 8.2.5.

Fig. 8.3 Map of acoustic sensor location using explicit model strategy

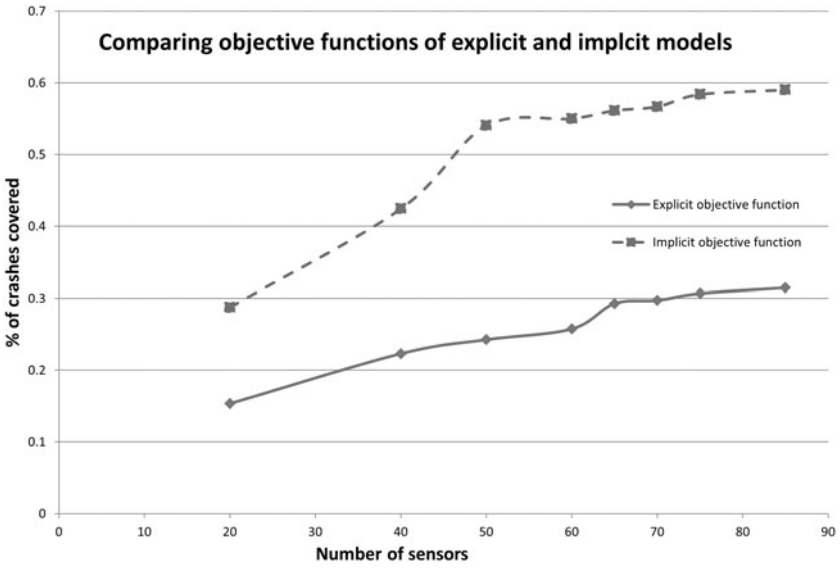


Fig. 8.4 Comparing objective functions of explicit and implicit model

Fig. 8.5 Map of acoustic sensor location using implicit model strategy

8.3.1.3 Simulation-Based Optimization in Acoustic Sensors

As noted in the case study, there are 869 candidate sensor locations and we plan to place 20–85 sensors in some of them. The number of possible ways to locate the sensors is large, e.g., C_{20}^{869} solutions exist for 20 sensors, far too many to evaluate using simulation. Note that each solution needs multiple simulation runs to reach statistically significant results to make a comparison with other placement strategies. Given these constraints, the goal of the simulation-based optimization is to use neighborhood search to identify a few solutions with potentially better performance than the initial solution (obtained by placing sensors from implicit model as shown in Fig. 8.5 for the case study).

A neighborhood search procedure for the acoustic sensor location problem should generate alternate placement strategies from the present solution and move in the direction of improved performance measures. There is no guarantee that this method will generate a better solution because the method is an exploratory search method. To generate a feasible set of neighbor solutions we propose a cluster-based exchange for the acoustic sensors.

A cluster of sensors refers to a geographically close set of sensors. In the cluster-based exchange, we move a sensor from one cluster to an unoccupied potential sensor location in another cluster as shown in Fig. 8.6.

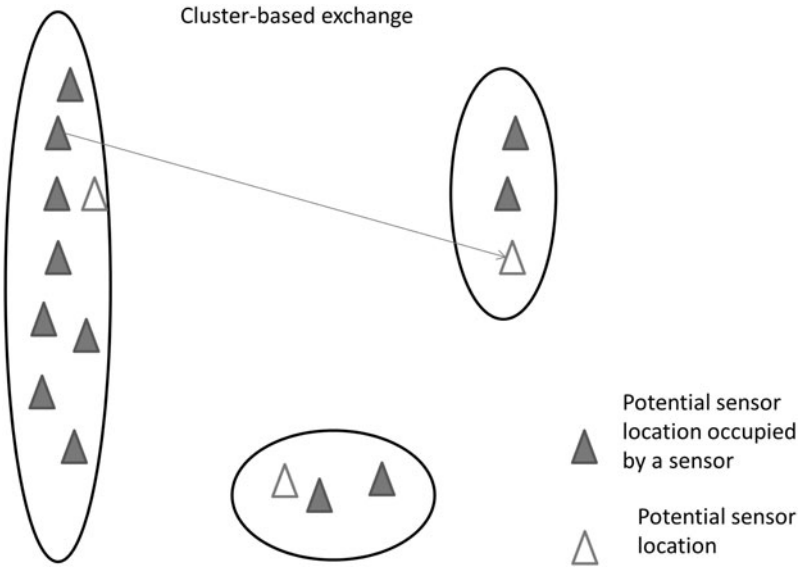


Fig. 8.6 A demonstration of cluster-based exchange

To implement the cluster-based exchange we first divide the set of potential sensor locations occupied by sensors into clusters. For example, in the case of 75 sensors we chose six clusters. The reason for choosing six clusters is that we expect roughly 10–15 sensors per cluster. We use the Lloyd's algorithm to form the six clusters, see Lloyd (1982). Figure 8.7 illustrates the six clusters obtained by applying Lloyd's algorithm on the 75-sensor problem for our case study. We note that the number of sensors in each cluster varies significantly. To proceed with generating a neighboring solution, we select a geographically distant sensor (from the

Fig. 8.7 Cluster-based exchange generating neighbor solutions

center of one cluster) and remove the sensor allocated to this location. A new potential sensor location is chosen (preferably closer to the center of the cluster) from another random cluster and a sensor is allocated to this new location.

There are other methodologies to create feasible neighbor solutions; however, the cluster-based proposed here preserves the data fusion opportunities provided from using implicit model.

8.3.1.4 Preliminaries Related to Computational Experiments using Simulation

In order to compare statistically significant estimates for the performance measures we use 20 experiments for each acoustic sensor placement considered. Each experiment consists of 50 individual simulation runs. At the end of each experiment, the performance measures are collected. For example, if the first experiment results in the detection of 30 crashes by the acoustic sensors, the performance measure "% of crashes detected" is 30/50 or 60 %. We repeat this experiment 20 times generating 19 more observations for the "% of crashes detected" performance measure. In total, we run 1000 simulation runs for an acoustic sensor placement strategy. Running this *Arena*® simulation using *Intel*® *Core*^{2} Duo with clock speed of 3.17 Hz and 4 Gb *RAM*, takes on average 45 s to complete a simulation run and takes on average 12.5 h to complete evaluation of a single placement of sensors.

Three points of clarification are necessary regarding our experiments. First, since the computation time for evaluation of a single placement strategy is 12.5 h only a few more evaluations of additional sensor placement alternatives are possible during the neighborhood search procedure. Second, each simulation run should always start with a different random number seed in Arena; this ensures that the results collected from the simulation runs are statistically independent. Third, the fact that we run 20 experiments with each run generating a statistically independent value for both the performance measures studied, allows us to generate 95 % confidence intervals for the performance measures.

8.3.1.5 Explicit Model Versus Implicit Model Using Simulation Evaluation

In this section, we evaluate the results of the explicit and the implicit model using simulation. We use the simulation model as an evaluation tool for both the performance measures in the explicit and implicit model. For comparison, we place 75 sensors using the explicit and the implicit model. Note that the solution method for the implicit model starts by using the explicit model solution as the starting point.

For the "% of crashes detected" performance measure, explicit model placement has a mean of 23.1 % and the implicit model has a mean of 43.5 %. From the 95 % confidence intervals in Fig. 8.8, the explicit model has higher variance in "% of crashes detected" performance measure. Similarly, for "% of crashes detected by

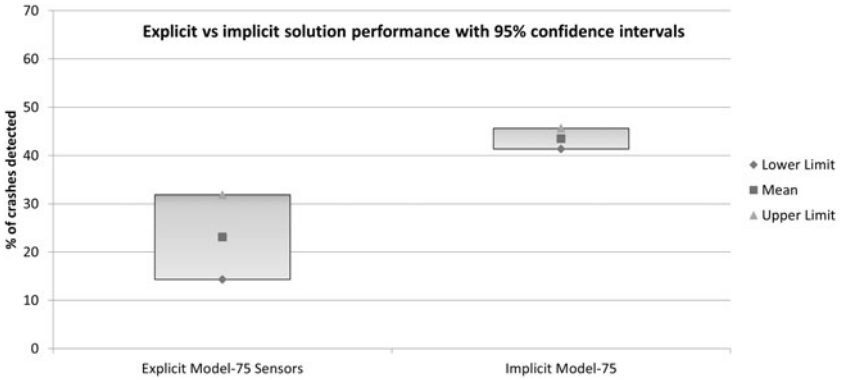


Fig. 8.8 Explicit versus implicit solution evaluation in simulation with 95 % confidence intervals

two sensors," the explicit model placement has a mean of 0.3 % and implicit model placement has a mean of 11.2 %.

8.3.1.6 Implicit Model Performance with Different Number of Sensors

The second factor that will affect the performance measures in the implicit model is "number of sensors." Ideally, the number of sensors available depends on budgetary constraints. In this section, we present the results of changing the number of sensors on both the performance measures.

In Fig. 8.9, we present the 95 % confidence intervals for the "% of crashes detected" performance measure with different number of sensors (50, 55, 60, 65, 70 and 75). The mean of "% of crashes detected" performance measure is found to increase with the number of sensors.

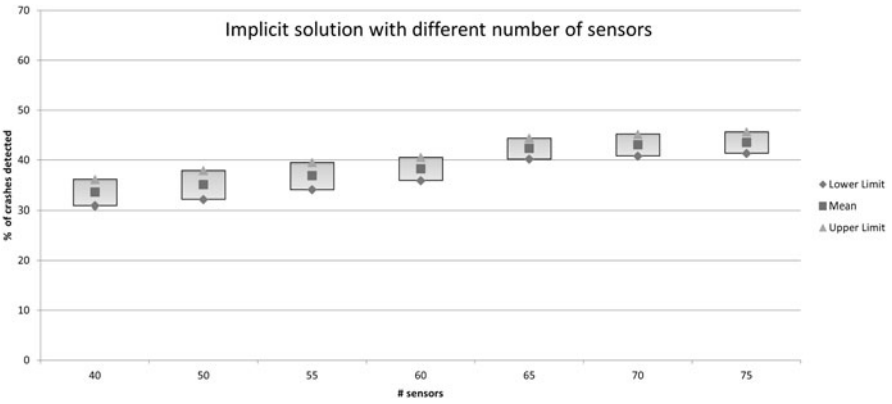


Fig. 8.9 Implicit solution with different number of sensors evaluated using simulation

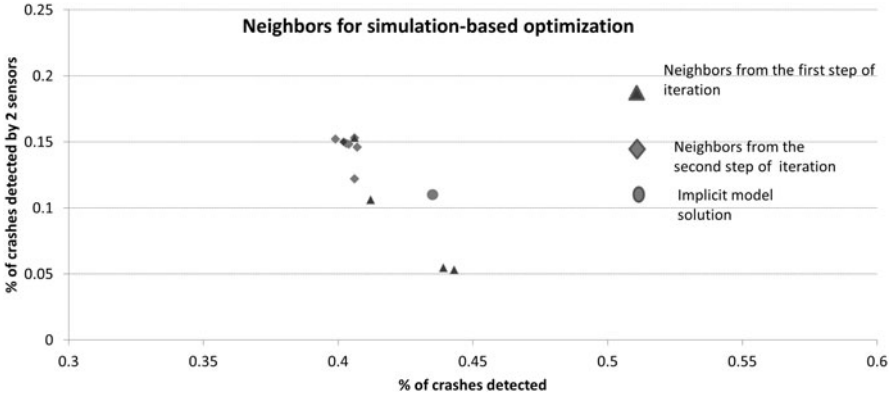


Fig. 8.10 Neighbor solutions generated using two iterations of simulation-based optimization

Another observation from Fig. 8.9 is the mean of "% crashes detected" performance measure is not diminishing as the number of sensors increase; in particular going from 55 to 60 sensors yields less improvement than going from 60 to 65 sensors. The solution the implicit model reaches is a function of the explicit solution; we attribute this discrepancy to having a comparatively better starting solution going from 60 to 65 sensors, i.e. the explicit solution for 65 sensor is a better solution than 60 sensors and the heuristic approach of the implicit model found better solutions starting from 65 sensors, compared to starting at 60 sensors. We note that other approaches like starting the implicit model from a random solution were not considered.

8.3.1.7 Results of Simulation-Based Optimization Using a Neighborhood Search Heuristic

Simulation-based optimization adds a local search method to the simulation to search for better solutions in the neighborhood of a particular solution. For the acoustic sensor placement problem, we proposed a cluster-based exchange to provide neighbor solutions as described in Sect. 8.3.1.3.

We initiate the cluster-based exchange procedure with the implicit model solution and generate a set of five neighbor solutions that are different from each other. Each generated solution has a different sensor placement strategy. In the next step, simulation evaluates each placement strategy. In Fig. 8.10, the neighbor solutions are around the implicit model solution.

Since the implicit model solution is still one of the efficient/non-dominated solutions after the first step, we use this as the starting point to generate a new set of five neighbor solutions for the second step. We always choose a non-dominated solution to start the search, since the local search process plans to move only in the improved direction.

Figure 8.10 shows two steps of neighborhood search. In every Step 5 new solutions are generated and using simulation the two performance measures are evaluated. In Fig. 8.10 the solutions from first step are marked using the red triangles whereas the solutions from the second step are marked in blue. The goal of any neighborhood search is to generate new solutions, which are better in both performance measures. The newly generated neighbors are evaluated using simulation to observe both the performance measures. From the results in Fig. 8.10, the implicit model solution is still a non-dominated solution in both the performance measures after two iterations. Other solutions from the neighborhood search are not efficient. It proves our intuition that the implicit model is a good solution for the acoustic sensor placement problem.

8.3.2 Exploration of Three Special Cases: Higher Detection Radius, Solution Robustness and Single and Double Coverage Weights

In this section, we explore three cases of special interest in the acoustic sensor placement problem. These cases demonstrate the use of simulation as an experimental tool to evaluate future scenarios or specific questions useful for a practitioner. In Sect. 8.3.2.1, we present the change in performance measures if there is an increase in the detection radius of the acoustic sensors. Section 8.3.2.2 presents a robustness test to evaluate the implicit model solution. Section 8.3.2.3 presents the effects of changing single and double coverage values chosen in the implicit model on the two performance measures.

8.3.2.1 Evaluation of a Sensor with Higher Detection Radius

Detection radius is a key input for both the explicit and implicit model. A favorable change for a sensor location problem is the improvement in the detection radius. An improvement in detection radius for a sensor is an important factor as it usually has a direct impact on system performance measures. Section 8.3.1 assumes a detection radius of 50 m (which is current specification for commonly available acoustic sensors). This low detection radius for commonly available acoustic sensor is attributed to the decay of sound in atmosphere. In this section we study the impact of a higher detection radius (60, 70, 80 and 90 m) on the "% of crashes detected" and "% of crashes detected by two sensors" performance measures. A higher detection radius is possible if the acoustic sensors start using noise isolation and noise filtering.

From Fig. 8.11, we notice the increase in "% of crashes detected" performance measure, when the detection radius increases from 50 to 90 m is about 4 %. The increase in "% crashes detected by two sensors" is 26 % when the detection radius changes from 50 to 90 m. Using a higher detection radius, the sensors increase the

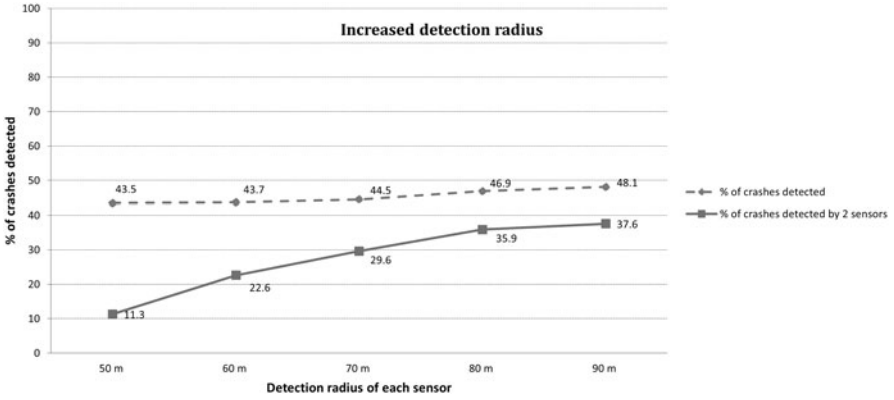


Fig. 8.11 Simulation results for increased detection radius of the two performance measures

overlap region but considering the 9-mile square area chosen, the sensors are unable to cover a large part of the road segments previously uncovered by sensors with 50 m detection radius. This explains increased gains in the "% crashes detected by two sensors" performance measure. This also raises the need to explore potential savings (in terms of higher costs for sensors versus better detection of incidents or false positives).

8.3.2.2 Solution Robustness

A key input for both the explicit and the implicit model is the past crash/demand data. We use past crash data to construct demand for each road segment in the explicit and the implicit model. In this chapter we assume that the past crash data serves as a good indicator for the future crash data that guides the placement of the incident detecting sensors. There are equally good arguments made to support and reject the above assumption. For example, accidents are more common on or near entry and exit roads of a highway where large volumes of traffic merge. However, not all entry and exit highways are equally likely to have an accident. The other reasons like road construction and traffic flow rates play a major role in causes of accident. A favorable argument to use the past accident data is statistically some highway entry and exit paths have a higher probability of accidents. Law of large numbers suggests that if there is a large volume of data, the mean number of accidents remains the same. Most accidents are attributed to driver distraction, this argument assumes that driver distraction is the major cause and no amount of past data can predict future crashes.

In order to alleviate some of the concerns from the assumption, we design a robustness test, to evaluate the performance of the implicit model through an experiment. In this robustness test, the implicit solution strategy is experimented via changes in the crash generation module. In general, the crash generation module

Table 8.1 Shows the construction of new demand for nodes and paths for the robustness test

Path/node	Probability of crash	Robustness test ($p = 0.9$)	Robustness test ($p = 0.7$)
Path a	0.2	0.2(0.9)	0.2(0.7)
Path b	0.3	0.3(0.9)	0.3(0.7)
Path c	0.1	0.1(0.9)	0.1(0.7)
Node d	0.4	0.4(0.9)	0.4(0.7)
Path e	0	$(1 - 0.9)/2$	$(1 - 0.7)/2$
Node f	0	$(1 - 0.9)/2$	$(1 - 0.7)/2$

uses past crash data to simulate crashes in the simulation. We modify the approach that uses past crash data in the simulation and use the implicit model placement strategy for the sensors.

In our study area, we found that close to 30 % of the road segments have no crashes in the period 2004–2009, but may clearly have future crashes. We use a parameter p to signify the likelihood that past crash data accurately predicts future crash data, with $p = 1$ implying that perfect predictability. In this section we test the performance of the placement suggested in Sect. 8.3.1 with different values of p ($p = 0.9$ and $p = 0.7$). If the value of p is chosen as 0.9, in that experiment 90 % of the crashes happen on road segments where the crashes have occurred previously and 10 % of the crashes occur on road segments that did not have an crash previously. Table 8.1 shows the construction of the demand for the two cases of $p = 0.9$ and $p = 0.7$. We construct a new path and node demand values and evaluate the placement suggested by the implicit model using 75 sensors.

From the results shown in Figs. 8.12 and 8.13, comparing the original ($p = 1$) results with the robustness test ($p = 0.9$ and $p = 0.7$), both the performance measures decrease. The percentage of crashes detected by a single sensor decreases by 10 % as the value of p decreases by 0.1. However, double coverage decreases drastically and as p-value approaches to 0.7 there are just 1 % of the crashes covered by two sensors. These results also indicate to a user/practitioner that if 30 % of future crashes happen on road segments, which previously have no crashes, the data fusion opportunities shrink to zero.

8.3.2.3 Effects of Changing Single/Primary (α_1) and Double/Secondary Coverage (α_2) on Performance Measures

The implicit model allows a practitioner/user to choose between primary/single coverage and secondary/double coverage and assign values for (α_1) and (α_2) in Eq. 8.10. This choice of single and double coverage weights affects sensor placement and affects both performance measures. In the results shown in Sect. 8.3.1, we selected $\alpha_1 = 0.7$ and $\alpha_2 = 0.3$. In this section, we experiment with various values of α_1 and evaluate the changes in both the performance measures.

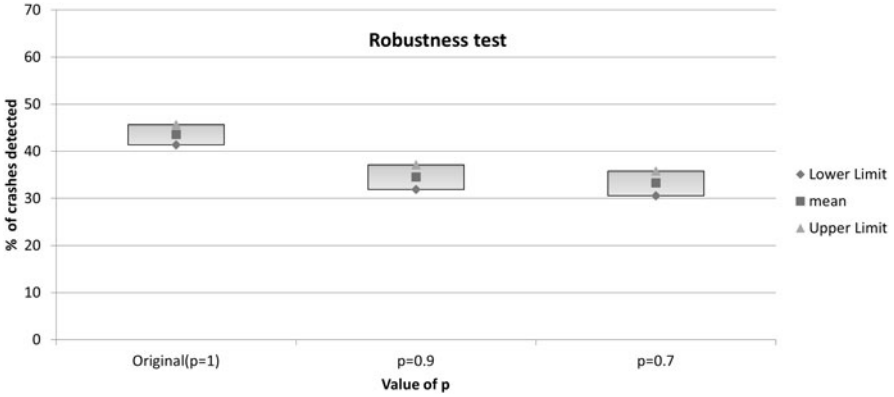


Fig. 8.12 Results from the robustness test using simulation

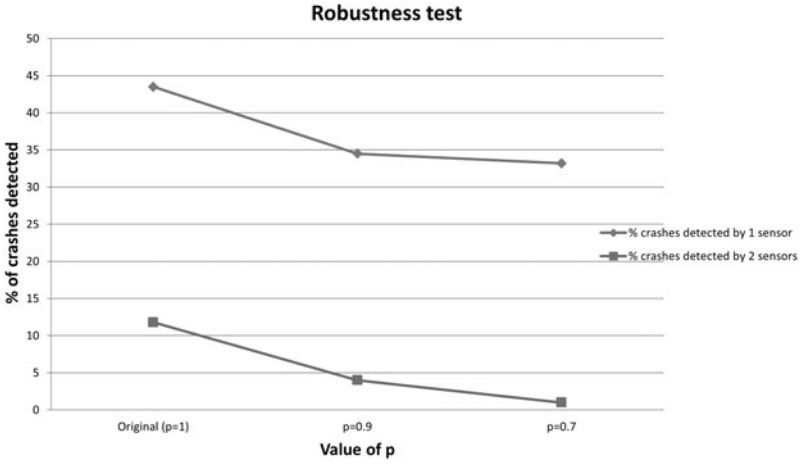


Fig. 8.13 Robustness test results showing both the performance measures

Table 8.2 A table showing different weights chosen for the primary and secondary coverage split

	Single coverage α_1	Double coverage α_2
Experiment 1	0.4	0.6
Experiment 2	0.5	0.5
Experiment 3	0.6	0.4
Experiment 4 (original)	0.7	0.3

In order to understand the effect of single and double coverage values on the performance measures, we chose three experiments with α_1 and α_2 values as shown in Tab. 8.2. For example, Experiment 1 assigns a weight of 0.4 to α_1 and 0.6 to α_2. Experiment 4 references the results in Sect. 8.3.1.

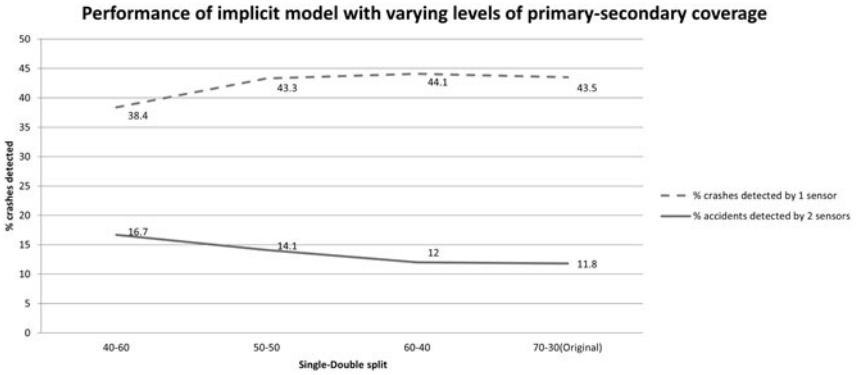


Fig. 8.14 Performance of implicit model with varying levels of primary and secondary coverage evaluated using simulation

From Fig. 8.14 we observe a clear positive correlation (except for Experiment 3 and Experiment 4) between the α_1 value and "% of crashes detected" performance measure and similarly there is a positive correlation between "% of crashes detected by two sensors" performance measure and α_2. Upon further inspection of data from sensor placement strategy 3 and 4, there is very little variation in performance measures and a large overlap in the 95 % confidence intervals of both performance measures, attributing this exception to random nature of the simulation. Using this knowledge of the relation between the weights for the implicit model and both the performance measures should help the user/practitioner to select the appropriate weights before using the implicit model.

8.3.3 The Impact of Mobile Sensors

In Sect. 8.3.1 and 8.3.2, we discussed the placement of acoustic sensors and the evaluation of the placement in various scenarios. This section presents the uses of stationary sensors working in tandem with mobile sensors to generate a situational awareness of an incident. We use *AACN* sensors as a representative sensor for mobile sensors. Both *AACN* and acoustic sensors use a signature crash characteristic to detect a crash. Acoustic sensors use the unique crash sound signature to detect a crash, whereas *AACN* sensors use accelerometers and contact sensors to detect a crash. Both the sensors are capable of detecting the crash characteristics:

- Crash detection
- Crash location
- Crash time
- Rollover
- Number of impacts

For example, a motor vehicle involved in a crash very close to an acoustic sensor and the automobile involved is equipped with an *AACN* sensor; there are two sets of dissimilar crash data for the above crash characteristics. The acoustic sensor attempts to extract a sound feature that consists of a high amplitude noise for a short duration of time. If the acoustic signal processing succeeds in extracting the crash sound feature, the signal processor accomplishes detection and estimates other crash characteristics. Similarly, the *AACN* reports a crash when the accelerometers report unusual acceleration/deceleration observed in motor vehicle crash. Since a single acoustic sensor observes this crash, we assume that the acoustic sensor reports a point estimate for the crash and sends its (own) *GPS* location as an estimate for crash location. Every *AACN* includes a *GPS* sensor that estimates the position of the crash. The acoustic sensor and the *AACN* both report different sets of information to determine crash time, e.g., the observation times of a crash-like high-amplitude sound and high accelerometer reading. An *AACN* sensor also may include a 3-d acceleration sensor system that can record unusual readings to detect a rollover. Once the two crash data sets are available, the next step in the data fusion schema is to send this data from individual sensors to local data fusion motes (a mote is a data processing unit that can receive data from sensors and use data fusion algorithms).

We assume that a backbone fusion network, capable of combining information from both acoustic sensors and sensors in the *AACN* system exists. Suppose there is a roadway incident involving two cars in a data fusion capable environment, populated with acoustic sensors and cars enabled with *AACN*. The tree diagram shown in Fig. 8.15 identifies the data fusion opportunities.

The tree diagram in Fig. 8.15 identifies the levels of data fusion needed when acoustic sensors and *AACN* sensors are considered. Consider the case of a two-car crash in which one car is equipped with *AACN*. There is just one acoustic sensor near the crash, two independent crash data sets are available through the acoustic sensor and *AACN* and there are two levels of data fusion possible, L0 and L2. L0 data fusion associates the data from acoustic sensor and *AACN*. After Level 0 data fusion, the two data sets are combined using L2 level data fusion to generate crash data useful for situational awareness.

Expanding upon Fig. 8.15, in Fig. 8.16 we present the data fusion opportunities with 75 acoustic sensors and assuming 10 % of the vehicles are equipped with *AACN*. The probability values reported for various scenarios of detection are results from simulation.

Table 8.3 summarizes the simulation results and shows that the combined sensor system detects 53.5 % of the crashes. Seventy-five acoustic sensors (without the aid of *AACN* sensors) could detect 43.5 % of the crashes in the same system. With *AACN* sensors alone (in 10 % of vehicles), 16.9 % of crashes were detected.

By simulating the *AACN* sensors alone, the error in position in 169 detected crashes was 4.23 m. Using acoustic sensors alone in a data fusion environment, the error in position was 5.42 m in 435 crashes. Combining the *AACN* with the 75 acoustic sensors, the error in position in 535 detected crashes was 5.04 m. Finally, using the combined *AACN* and the acoustic sensor system, the error in detecting multiple impacts was 5 %.

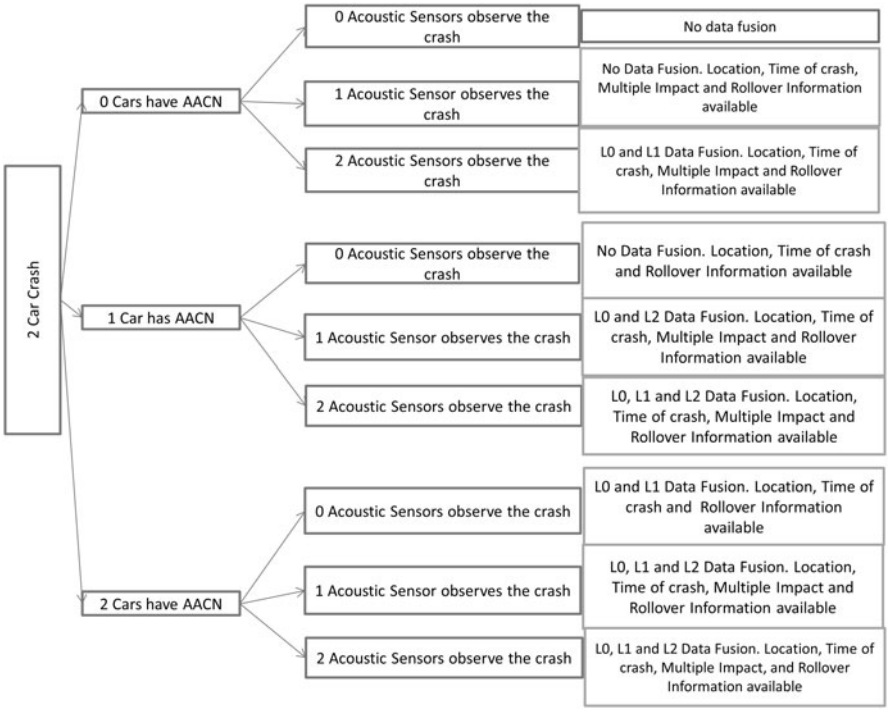


Fig. 8.15 A tree diagram identifying the data fusion opportunities when acoustic sensors and *AACN* act as incident detection sensors

Using both *AACN* and acoustic sensors together in an incident detection system capable of data fusion reduces the error in position (relative to acoustic sensors alone) to 5.04 m. Although this reduction may seem small, it can be significant in areas containing multiple road segments over bridges and through underpasses. Five-meter accuracy in the position of a crash is close to lane level accuracy and very helpful for emergency response personnel. The simulation evaluation assumes that in the case study, *GPS* sensors work with an error of 10 m (Hubrich and Curran 2009) and the road network in the system has no tunnels and other forms of interruptions for *GPS* systems. Even in cases with *GPS* interruption, we believe that a system with acoustic sensors and *AACN* is effective and will help detect and characterize a crash.

Table 8.3 demonstrates the positives of using a data fusion capable environment with different sensor types observing a crash. The simulation evaluation results point towards the improvements in crash detection rate for the combined system. It also shows a decrease in error for the combined system relative to the acoustic sensor alone, in detecting both the crash position and number of impacts, which can be critical in assessing the crash.

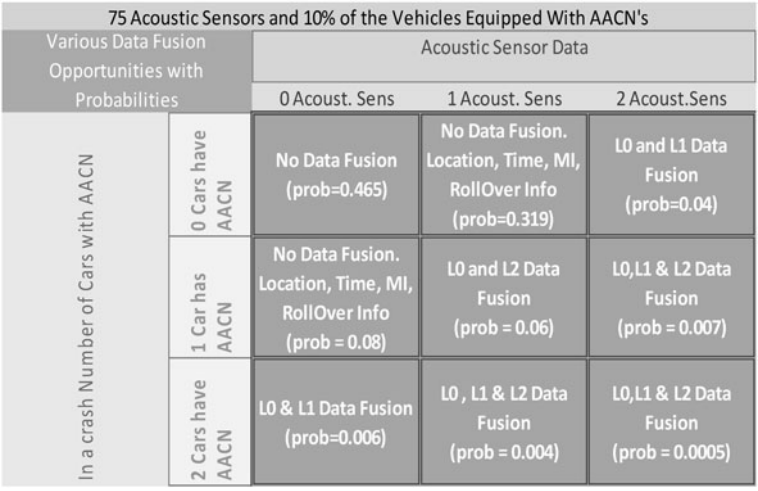


Fig. 8.16 Data fusion opportunity in an incident detection system with 75 acoustic sensors and 10 % vehicles equipped with *AACN*

Table 8.3 Summary of the advantages of using AACN and acoustic sensors in incident detection

	Acoustic sensors	*AACN* sensor	Acoustic and *AACN* sensor
% of crashes detected	43.5	16.9	53.5
Average error in estimation of accident location (m)	5.42	4.23	5.04

8.4 Concluding Remarks

Multi sensor data fusion is essential in future road-transportation systems. For *ITS* systems used in incident-detection, data fusion is paramount and increases the usability and reliability of data from individual sensors. If stationary sensors are a part of the incident detection system, their detection capability and quality of data fusion are dependent on the placement of the sensors. There is abundant literature on placement of sensors for maximizing incident detection. However, little to no literature exists on placement models that consider both quality of data fusion and detection capability in placement of stationary sensors. This chapter describes a sensor placement problem for multiple sensor classes, where both single coverage and double coverage are important in placement of sensors. Single coverage is a surrogate for indirectly measuring the detection capability and double coverage is a surrogate for indirectly assessing data fusion opportunities.

This chapter proposes an explicit-implicit model, followed by simulation-based optimization to incorporate single and double coverage in the sensor placement problem. This is a unique approach developed for this chapter. The explicit-implicit model is a mathematical model that approximates the real situation, and proposes

an objective that aims to cover crash demand from both road segments (paths) and intersections (nodes) where crashes are more likely to happen. The explicit model is a quadratic maximal coverage problem and uses a greedy paired adding heuristic to generate a feasible solution. The implicit model is an alternate geometric approach to solve the sensor placement model. The explicit model placement strategy acts as the starting solution for the implicit model. The simulation-based optimization generates additional feasible solutions using a local search and improves the best sensor placement strategy. This combined approach has yielded good solutions to the omnidirectional sensor placement problem.

The simulation model borrowed from Henchey et al. (2013) and developed in this chapter imitates traffic movement in a real road network and the functioning of omnidirectional (acoustic) sensors and mobile (*AACN*) sensors. The simulation model has unique features that create a crash, measure the performance of the sensors and has built-in weather-effects, congestion and road noise. The simulation model developed here evaluates a placement strategy for the acoustic sensors. The main performance metrics considered are "% of crashes detected" and "% of crashes detected by two sensors".

Stochastic systems use simulation as an evaluation tool built on abstract models to understand and quantify the real-system. This evaluation process goes well with an improvement procedure to an existing system by using a local search procedure to evaluate local solutions and move in the direction of the improved performance. This approach used here to evaluate the acoustic sensor placement strategy and provide more local search results for a decision maker to optimize their criteria.

This chapter makes significant contribution in using simulation as a beneficial and cost-effective evaluation tool to answer questions pertaining to sensor deployment in a road network environment. We demonstrate that by choosing the location of just 75 acoustic sensors with a detection radius of 50 m, 43.5 % of the crashes are observed by a single sensor and two sensors make independent detections in 11 % of the crashes. The combined incident detection system of *AACN* and acoustic sensors detects 53.5 % of the accidents and the average error in estimation of the crash location is 5.04 m.

The chapter evaluates the data fusion capable system with both acoustic and *AACN* sensors. Many papers discuss the capabilities of an advanced system involving multiple sensor types, but none has been able to evaluate the real world applications. This work demonstrates the effectiveness of such systems and quantifies the improvements.

Several factors affect the overall performance of the omnidirectional sensor placement. Three factors explored in this chapter are radius of detection, single and double coverage weights for the implicit model and the change in crash demand data. In Sect. 8.3.2, we evaluate all these factors individually to understand the relation between the factors and performance measures. A future research direction would be to explore the combined effects of changing multiple factors on the performance measures.

References

Administration FH (2009) Traffic detector handbook, 3rd edn., Vol I. US Department of Transportation. http://www.fhwa.dot.gov/publications/research/operations/its/06108/. Accessed 4 Dec 2015

Berman O, Drezner Z, Krass D (2013) Continuous covering and cooperative covering problems with a general decay function on networks. J Operat Res Soc 64:1644–1653

BMW USA (2008) BMW innovations, Real time traffic information. www.bmwusa.com/Standard/Content/Innovations/Engineering/RTTI/Default.aspx. Accessed 4 Dec 2015

BMW (2010) BMW Insights safety. http://www.bmw.com/com/en/insights/technology/connecteddrive/2010/safety/emergency_call/emergency_call_information.html. Accessed 4 Dec 2015

Cambridge Systematics & Texas Transportation Institute (2005) Traffic congestion and reliability: trends and advanced strategies for congestion mitigation. U.S. Department of Transportation, Federal Highway Administration

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

Dhillon SS, Chakrabarty K (2003) Sensor placement for effective coverage and surveillance in distributed sensor networks. Proceedings of IEEE Wireless Communications and Networking Conference, pp 1609–1614

Dong G (2007) Simulation-based optimization. Ph. D. Dissertation, University of Wisconsin-Madison

Erdemir ET, Batta R, Rogerson P, Speilman S, Blatt A, Flanigan M (2008a) Optimization of aeromedical base locations in new mexico using a model that considers crash nodes and paths. Accid Anal Prev 40:1105–1114

Erdemir ET, Batta R, Rogerson P, Speilman S, Blatt A, Flanigan M (2008b) Location coverage models with demand originating from nodes and paths: application to cellular network design. Eur J Oper Res 190(3):610–633

Gentili M, Mirchandani P (2005) Locating active sensors on traffic networks. Ann Oper Res 136:229–257

Greater Buffalo-Niagara Regional Transportational Council (2010) Geographic information system. http://www.gbnrtc.org/index.php/planning/gis/. Accessed 4 Dec 2015

Hall D, Llinas J (1997) Introduction to multi-sensor data fusion. Proc IEEE 85:6–23

Henchey M, Batta R, Blatt A, Flanigan M, Majka K (2013) A simulation approach to studying emergency response in an advanced transportation system. J Simula 8:115–128

Hubrich S, Curran K (2009) Optimizing mobile phone self-location estimates by introducing beacon characteristics to the algorithm. J Loc Based Serv 3:55–73

Kelton D, Sadowski R, Sadowski D (1997) Simulation with arena. McGraw-Hill, New York

Law A, Kelton D (1999) Simulation modeling and analysis. McGraw-Hill, New York

Lloyd S (1982) Least square quantization in PCM. IEEE Trans Info Theory 28(2):129–137

Nagare A (2012) Brief introduction to intelligent transportation systems. www.freeway.gov.tw/UserFiles/File/Traffic/A1%20Brief%20introduction%20to%20Intelligent%20Transportation%20System,%20ITS.pdf. Accessed 4 Dec 2015

Nie X, Parab G, Batta R, Lin L (2012) Simulation-based selectee lane queuing design for passenger checkpoint screening. Eur J Oper Res 219:146–155

Office of Highway Policy Information (2009) Highway statistics. (Office of Highway Policy Information). http://www.fhwa.dot.gov/policyinformation/statistics/2009/tc202c.cfm. Accessed 4 Dec 2015

OnStar (2012) Automatic crash response. https://www.onstar.com/web/portal/emergencyexplore?tab=1. Accessed 4 Dec 2015

Srinivasan K, Jovanis P (1997) Determination of number of probe vehicles required for reliable travel time measurement in urban network. Transp Res Rec 1537:15–22

White J, Thompson C, Turner H, Dougherty B, Schmidt D (2011) WreckWatch: automatic traffic accident detection and notification with smartphones. J Mob Netw Appl 16(3):285–303

Chapter 9
Location Models for Preventive Care

Vedat Verter and Yue Zhang

9.1 An Introduction to Preventive Care

Health conditions are much easier to prevent than to treat, and recovery is often more likely if an illness is diagnosed at an early stage. The substantial savings in the costs of diagnosis and therapy as well as the relatively lower capital investment associated with preventive care programs have been recognized for a long time (Walker 1977). Preventive care programs can save lives and contribute to a better quality of life by reducing the needs for radical treatments. For example, there is evidence that mammograms taken on a regular basis have the potential to reduce deaths from breast cancer for women between the ages of 50 and 69 by up to 40 % (Health Canada 2007).

According to World Health Organization (2002), although many diseases can be prevented, the current healthcare systems do not make the best use of their available resources to support preventive programs. Most of these systems are based on responding to acute problems, urgent needs of patients, and pressing concerns. Preventive care is inherently different from healthcare for acute problems, and in this regard, current healthcare systems worldwide fall remarkably short. For instance, only 5 % of the $ 1.4 trillion spent on direct health care in the United States goes to preventive care measures and the promotion of general health (Falkenheimer 2004).

Preventive care programs can be categorized into three groups with regards to their objective: (1) *primary prevention* aims at reducing the likelihood of diseases in people with no symptoms, e.g., immunizations of healthy children; (2) *secondary prevention* aims at identifying and treating people who have risk factors or are at

V. Verter (✉)
Desautels Faculty of Management, McGill University, 1001 Sherbrooke Street West, Montreal, H3A 1G5, QC, Canada
e-mail: vedat.verter@mcgill.ca

Y. Zhang
College of Business and Innovation, University of Toledo, 2108 West Bancroft Street, Toledo, 43615, OH, USA
e-mail: yue.zhang@utoledo.edu

H. A. Eiselt, V. Marianov (eds.), *Applications of Location Analysis,* International Series in Operations Research & Management Science 232, DOI 10.1007/978-3-319-20282-2_9

very early stage of diseases, e.g., pap smears to detect early forms of cervical cancer; (3) *tertiary prevention* aims at treating symptomatic patients in an effort to decrease complications or severity of disease, e.g., sugar control in a diabetic in order to mitigate vision and nerve problems. Flu shots, blood tests, mammograms, and anti-smoking advice are among the most well-known preventive care services.

An effective way to improve the efficiency of a regional healthcare system with limited recourses is to increase the number of people receiving preventive care services, which has been an integral part of many healthcare reform programs within the past two decades (Goldsmith 1989). However, the achievement of desired participation level continues to be a challenge to many preventive care programs. For example, every province in Canada had an organized program offering biennial mammography screening by 2003. The average participation rate, however, reached only 44 % in 2006, and none of the programs achieved the nationally established target of 70 % participation (Public Health Agency of Canada 2006). Thus, maximizing the level of participation to preventive care programs has been a common objective for policy makers.

In contrast to sick people who need urgent medical attention, people who seek preventive services have more flexibility as to when and where to receive preventive care services. The accessibility of the facilities is an important factor for the success of a preventive care program. Zimmerman (1997) found through a survey that the convenience of access to the facility was a very important factor in a client's decision to have prostate cancer screening. Facione (1999) revealed that the perceptions of lack of access to services were related to the decrease of mammography participation. Therefore, most studies in the literature focus on accessibility related to the number, type, concentration, and location of preventive care facilities, as well as transportation to services and availability of providers.

Preventive care has a number of differentiating features in comparison with other types of care such as emergency, acute and long-term care. First, the number of people who seek the services at the facility is not controlled by the policy maker, i.e., preventive care is a *user-choice* environment in terms of the allocation of clients to facilities. Second, when people have to wait for a long time to receive the services due to limited capacity, their willingness to participate in preventive programs could decrease significantly. As a result, *congestion* at the facilities is a crucial factor considered in most studies in the literature. Third, it is essential to consider the relationship between demand volume and quality of preventive care services. For example, U.S. Food and Drug Administration (1999) requires a radiologist to interpret at least 960 mammograms and a radiology technician to perform at least 200 mammograms in 24 months to retain their accreditations. As a result, in order to ensure service quality, preventive care facilities cannot be operated unless the size of their clientele fulfills a *minimum workload* requirement.

This chapter focuses on the problem of designing a network of preventive care facilities so as to maximize accessibility and hence, the level of participation. We provide an overview of the prevailing literature and report on a real-life application. The remainder of the chapter is organized as follows. The next section presents the common characteristics of the fundamental models. Sects. 9.3 and 9.4 discuss

two categories of models in detail, respectively. Sect. 9.5 presents an illustrative example in order to provide the reader with a comparative understanding of the alternative modelling frameworks. Sect. 9.6 discusses alternative solution methods. Sect. 9.7 summarizes an application of location models for designing the network of preventive breast cancer screening centres in Montreal. Several limitations of the reviewed studies and extensions to the current literature are discussed in Sect. 9.8. The final section concludes the chapter.

9.2 Overview of the Fundamental Models

As mentioned above, the objective is to maximize the level of participation by determining the most appropriate configuration of the preventive healthcare facility network. This comprises determining the number of facilities to be established, the location of each facility as well as its capacity. In order to ensure service quality, preventive care facilities cannot be operated unless the size of their clientele fulfills a minimum workload requirement. Therefore, the estimation of the user's choices as to which facilities they will patronize constitutes an essential component of the model formulation.

In many cases, the problem involves the accreditation of a subset of the existing facilities, rather than building a set of brand new facilities, so as to provide the service. As a result, most of the existing studies do not incorporate a fixed setup cost, and the minimum workload requirement actually works as the accreditation criterion. In order to focus on the resource allocation strategy, the total number of servers to be allocated to open facilities is often fixed. Consequently, the number of facilities that could be opened is limited, mainly due to the minimum workload requirement as well as the total number of available servers.

As mentioned earlier, the accessibility of the facilities is a crucial factor to influence people's decision to participate. Several papers used travel distance or travel time alone as a proxy for the accessibility of a facility, including Verter and Lapierre (2002) and Zhang et al. (2012). In contrast, another stream of research also incorporates congestion into the model, such as Zhang et al. (2009 and 2010). They used the total (travel, waiting, and service) time as a proxy for accessibility.

The user-choice concerning where to receive the service is also a critical part of the model structure. There are typically two modeling alternatives in the literature to address the user choice: (1) the *deterministic-choice model* or optimal-choice model; and (2) the *probabilistic-choice model* or stochastic-choice model. The former requires a fully-informed and rational set of clients, who patronize the facility with the highest attractiveness. (In most location models, for example, it is common to assume that each customer will seek services from the closest open facility.) The latter is based on random utility theory and assumes that the user-choice depends on a deterministic component comprising identified variables and a random component that captures the impact of all unobserved factors. This amounts to assuming that each client can visits any facility with a certain probability, which is increasing

with the identified attractiveness (utility) of the facility. Note that both modeling alternatives could be realistic depending on the specific circumstances.

9.3 Deterministic-Choice Models

9.3.1 Distance-Based Models

To the best of our knowledge, Verter and Lapierre (2002) is the first paper on location analysis in preventive care. With no capacity decisions considered, their problem is therefore to locate an undetermined number of facilities to maximize the level of participation subject to the minimum workload requirement. They used travel distance or travel time alone as a proxy for the accessibility of a facility. With the determination-choice modeling, this leads to the assumption that clients go to the closest facility.

Let $N(|N| = n)$ be a set of nodes, representing the neighborhoods of a city or the population zones. There is a finite set of potential facility sites $M(|M| = m)$. Let $S \subset M$ be a set of open facilities. The shortest travel time between nodes i and j is denoted by t_{ij}. The fraction of clients residing at node i is denoted by h_i. The number of clients who require the service over the entire network denoted by λ, and thus the number of clients from each node i is λh_i. A facility cannot be established at node j unless a_j exceeds a minimum workload requirement denoted by $R_{\min}$.

Assume that the faction of clients at node i who would seek at site j (i.e., the participation rate), denoted by a_{ij}, is a decreasing function of the travel time t_{ij}. Further assume that this participation function is linear with an intercept A_i and a slope γ, i.e.,

$$a_{ij} = \begin{cases} A_i - \gamma t_{ij} \text{ if } t_{ij} \le \dfrac{A_i}{\gamma} \\ 0 \qquad\quad \text{otherwise} \end{cases}, i \in N, j \in M. \tag{9.1}$$

Let L_{ij} denote the set of alternative facility sites that are closer to population zone i than site j, i.e., $L_{ij} = \{\ell : t_{il} < t_{ij}, \ell \in M\}$.

Define two sets of binary decision variables:

$$y_j = \begin{cases} 1 \text{ if } a \text{ facility is located at node } j \\ 0 \text{ otherwise;} \end{cases}$$

$$x_{ij} = \begin{cases} 1 \text{ if population zone } i \text{ is served by a facility at site } j \\ 0 \text{ otherwise.} \end{cases}$$

The problem is formulated as a mixed integer program (MIP), which is referred to as **Model 1**:

$$\underset{y,x}{\text{Max}}\ \lambda \sum_{i=1}^{n} h_i \sum_{j=1}^{m} a_{ij} x_{ij} \tag{9.2}$$

$$\text{s.t.} \sum_{j=1}^{m} x_{ij} \leq 1, i \in N \tag{9.3}$$

$$\lambda \sum_{i=1}^{n} h_i a_{ij} x_{ij} \geq R_{\min} y_j, j \in M \tag{9.4}$$

$$x_{ij} \leq y_j, i \in N, j \in M \tag{9.5}$$

$$x_{ij} \leq 1 - y_\ell, \ell \in L_{ij}, i \in N, j \in M \tag{9.6}$$

$$y_j \in \{0,1\}, x_{ij} \in \{0,1\}, i \in N, j \in M \tag{9.7}$$

The objective (9.2) is to maximize the level of participation. Constraints (9.3) ensure that at most one facility can serve a population zone. Constraints (9.4) impose the minimum workload requirements on open facilities. Constraints (9.5) stipulate that a population center can only be served by an open facility. Constraints (9.6) ensure that each population center is assigned to the closest open facility. The integrality requirements on the decision variables are expressed by constraints (9.7).

Zhang et al. (2012) presented a deterministic-choice model following Verter and Lapierre (2002). They also assumed that clients would obtain the service from the closest facility. In contrast, they extended the earlier model by considering: (1) congestion at the facilities in a constraint of the model, i.e., a service level constraint; and (2) capacity allocation among the open facilities subject to a fixed total number of available servers.

Specifically, to incorporate congestion into the model, they assumed that the number of clients who require the service over the entire network is Poisson distributed. They also assumed homogeneous servers and exponentially distributed service times. Thus, each facility is modeled as an $M/M/c$ queue, where c denotes the number of servers. To guarantee a timely service, they assumed that the mean waiting time at any facility cannot exceed a maximum acceptable level. Other types of service level constraints can be applied in the model as well. Similar service level constraints have been used in the location literature, such as Marianov and Serra (2002), Wang et al. (2002), and Berman and Drezner (2006), as well as in practice. Marianov et al. (2004) adopt similar constraints on the wait times in a model where public service centres compete with private centres.

In summary, the problem is to maximize the total participation rate by finding the optimal set of locations and the number of servers at each open facility, subject to the service level constraint, the minimum workload requirement, and the given number of total available servers. They formulated the problem as a mixed integer programming problem, referred to as **Model 2**.

9.3.2 Time-Based Models

Zhang et al. (2009) extended Verter and Lapierre (2002) by incorporating congestion at the facilities into the participation modeling. Specifically, using the expected total (travel, waiting, and service) time as a proxy for accessibility of facilities, they assumed that clients would patronize the facility with the minimum expected total time. This is a major difference from Model 2 described above, where congestion at the facilities is considered in a service level constraint. Moreover, they assumed that clients at the same population zone would patronize the same facility. However, this assumption may not be realistic in the context of preventive care, and it also in fact prevents them from finding an equilibrium allocation of clients to facilities in general. Zhang et al. (2010) further extended the previous model by (1) relaxing the assumption that clients at the same population zone would patronize the same facility, and (2) considering capacity allocation among the facilities given a fixed total number of available servers.

In general, it is important to note the relationship between the arrival rate and the congestion at a facility for this type of time-based models. As the congestion or expected time at a facility is considered an attractiveness determinant, it influences the participation rate, i.e., the arrival rate to a facility depends on its congestion. On the other hand, the congestion, in turn, clearly depends on the arrival rate. Thus, one must determine the the arrival rate and the congestion in the steady state. Moreover, since we consider a network of multiple facilities, it implies that one must solve an equilibrium problem to determine the arrival rate and the congestion for each facility. This is called a *user-equilibrium* problem in the operations research literature.

Due to this nature, Zhang et al. (2010) formulated the problem as a bilevel model, where the lower level model is to find a user-equilibrium satisfying that clients choose the facility with the minimum expected total time, while the number, locations, and capacities of the facilities are determined at the upper level model. For similarity, this section primarily introduces the lower level model, given the set of open facilities S.

As each facility is modeled as an $M/M/c$ queue, the general formula for the mean waiting time is (Kleinrock 1975):

$$\bar{W}(a,c) = \frac{C(c,u)}{s}\frac{1}{\mu(1-\rho)} + \frac{1}{\mu} \tag{9.8}$$

where a denotes the arrival rate and

$$u = \frac{a}{\mu}, \rho = \frac{a}{c\mu}, C(c,u) = \frac{1-K(u)}{1-\rho K(u)}, K(u) = \frac{\sum_{l=0}^{c-1}\frac{u^l}{l!}}{\sum_{l=0}^{c}\frac{u^l}{l!}}. \tag{9.9}$$

Define two sets of decision variables for the lower level model:

s_j = number of servers at facility j;
x'_{ij} = fraction of clients from population node i who request service from facility j.

Denote the arrival rate of clients at facility j by a'_j, i.e.,

$$a'_j = \lambda \sum_{i=1}^{n} h_i x'_{ij},\ j \in S. \tag{9.10}$$

Denote by $\bar{T}_{ij}$ the average total time that clients from node i spend in order to receive service at facility j. The average total time $\bar{T}_{ij}$ comprises of two components: (1) the travel time from node i to facility j denoted by t_{ij}; and (2) the average time clients spend at the facility possibly waiting and receiving service which we denote by $\bar{W}(a'_j, s_j)$, i.e.,

$$\bar{T}_{ij} = t_{ij} + \bar{W}(a'_j, s_j),\ i \in N,\ j \in S. \tag{9.11}$$

Denote the total participation rate (fraction) at node i by p_i, i.e.,

$$p_i = \sum_{j \in S} x'_{ij},\ i \in N. \tag{9.12}$$

As clients choose the facility with the minimum expected total time, denote by T_i this shortest time incurred by clients at node i. Assume that the total participation rate p_i at node i is a linear decreasing function of the shortest time T_i with an intercept A'_i and a slope γ' (the participation function), i.e.,

$$p_i(T_i) = A'_i - \gamma' T_i,\ i \in N. \tag{9.13}$$

Denote by $T_i(p_i)$ the inverse participation function, i.e.,

$$T_i(p_i) = \frac{A'_i - p_i}{\gamma'},\ i \in N. \tag{9.14}$$

In fact, $T_i(p_i)$ represents a threshold time, i.e., the total time clients at node iare willing to incur for participating in the service, while the actual time incurred by clients at node i to facility j is $\bar{T}_{ij}$.

As mentioned earlier, given the set of open facilities S and the number of servers at each open facility s_j, $j \in S$, the lower level model involves the clients' facility choices so as to minimize their expected total time. This is a user-equilibrium problem, and at equilibrium, no client wants to change her choice. This equilibrium condition can be stated as: given S and s_j, $j \in S$, for all pairs of (i, j), $i \in N$, $j \in S$,

$$\bar{T}_{ij} = t_{ij} + \bar{W}(a'^*_j, s_j) \begin{cases} = T_i(p_i^*) & \text{if } x'^*_{ij} > 0 \\ \geq T_i(p_i{}^*) & \text{if } x'^*_{ij} = 0, \end{cases} \tag{9.15}$$

This condition (9.15) states that if there is a flow of clients from node i to facility j, then the actual time incurred by clients at node i to facility j must be equal to the

threshold time (the longest time that clients would accept to go to the facility); and if the actual time exceeds the threshold time, there is no flow.

To find $x_{ij}^{\prime *}$ in (9.15), we have to solve the following nonlinear complementarity problem, where $a_j^{\prime}$ and p_i are expressed by $x_{ij}^{\prime}$,

$$t_{ij} + \bar{W}\left(\lambda \sum_{i=1}^{n} h_i x_{ij}^{\prime}, s_j\right) - \frac{A_i^{\prime} - \sum_{j \in S} x^{\prime}{}_{ij}}{\gamma^{\prime}} \geq 0, i \in N, j \in S$$

$$x_{ij}^{\prime}\left[t_{ij} + \bar{W}\left(\lambda \sum_{i=1}^{n} h_i x_{ij}^{\prime}, s_j\right) - \frac{A^{\prime}{}_i - \sum_{j \in S} x^{\prime}{}_{ij}}{\gamma^{\prime}}\right] = 0, i \in N, j \in S$$

$$x^{\prime}{}_{ij} \geq 0, i \in N, j \in S \tag{9.16}$$

Alternatively, using vector-matrix notation, the complementarity problem (9.16) can also be rewritten as a variational inequality (Zhang et al. 2010). The upper level model with constraints (9.16) is called as a mathematical program with equilibrium constraints (*MPEC*), which is referred to as **Model 3**. Please refer to Zhang et al. (2010) for more details of the bilevel model.

9.4 Probabilistic-Choice Models

Probabilistic-choice models are represented by a variety of spatial interaction models in the literature. The first spatial interaction model that represents the probabilistic-choice behavior is by Huff (1962). Following Huff's model in which client utility is expressed by a simple gravity formula, a variety of studies in econometrics and marketing focused on developing better representations of the utility function, such as the multiplicative competitive interaction (*MCI*) model (Nakanishi and Cooper 1974). Both Huff's model and the *MCI* model were developed based on aggregate flows between population zones and facilities. Another well-known model, the multinomial logit (*MNL*) model (McFadden 1974), was originally proposed at the disaggregate (individual) level for discrete choice modeling, but it has been applied at the aggregate level as well (Gupta et al. 1996).

In general, the *MNL* model posits that the total utility of individual ichoosing alternative *j, j* ∈ *J*, denoted by U_{ij}, can be decomposed into two parts: (1) a deterministic component denoted by V_{ij}, which is usually explained by a linear function of attributes or characteristics; and (2) a random component denoted by ε_{ij}, which is assumed to be independent, identically extreme value distributed,

$$U_{ij} = V_{ij} + \varepsilon_{ij}. \tag{9.17}$$

Denote the probability of individual *i* choosing alternative *j* by P_{ij}. With utility-maximizing, P_{ij} is given by,

$$P_{ij} = \frac{e^{V_{ij}}}{\sum_{p \in J} e^{V_{ip}}}. \tag{9.18}$$

Note that the *MNL* model can only identify utility differences. Thus, one arbitrary alternative has to be defined as the reference alternative, and the utility associated with this alternative has to be set to an arbitrary value, usually normalized to 1.

In the context of preventive care, Gu et al. (2010) first applied a Huff-based location model for allocating clients to the facilities. Zhang et al. (2012) presented a probabilistic-choice model analogous to Model 2. They applied the *MNL* model for the client allocation with the assumption that travel time (distance), t_{ij}, to be the primary attractiveness determinant. In other words, $V_{ij} = -\beta t_{ij}$, where $\beta > 0$ is a parameter that denotes the sensitivity to travel time and can be estimated empirically. For simplicity, β is assumed to be identical for any population zone i or facility j. Again, this section focuses only on the allocation part of the entire problem using the *MNL* model.

To formulate the problem as a mathematical program, define two sets of decision variables:

$$s'_{jk} = \begin{cases} 1 \text{ if node } j \text{ has } k \text{ or more servers} \\ 0 \text{ otherwise,} \end{cases}$$

q_{ij} = probability (or fraction) of clients at zone i who request the service from facility j.

Note that q_{ij} is actually an auxiliary decision variable, and that $s'_{j\ell}$ in fact denotes the location decision at node j. According to the *MNL* model, define the choice of not visiting any facility as the reference alternative, and normalize its utility to 1. Now, q_{ij} can be expressed as:

$$q_{ij} = \frac{e^{-\beta t_{ij}} s'_{j\ell}}{1 + \sum_{\ell \in M} e^{-\beta t_{i\ell}} s'_{\ell 1}}, i \in N, j \in M. \tag{9.19}$$

This formulation ensures that clients must require service only from open facilities. Expression (9.19) can be rewritten as:

$$q_{ij} + \sum_{\ell \in M} e^{-\beta t_{i\ell}} q_{ij} s'_{\ell 1} = e^{-\beta t_{ij}} s'_{j\ell}, i \in N, j \in M. \tag{9.20}$$

Since $s'_{\ell\ell}$ is a binary variable and q_{ij} is a continuous variable, expression (9.20) can be linearized as follows by defining $z_{ij\ell}$ as an artificial continuous variable,

$$q_{ij} + \sum_{\ell \in M} e^{-\beta t_{i\ell}} z_{ij\ell} = e^{-\beta t_{ij}} s'_{j\ell}, i \in N, j \in M \tag{9.21}$$

$$z_{ij\ell} \leq q_{ij}, i \in N, j, \ell \in M \tag{9.22}$$

$$z_{ij\ell} \leq B_1 s'_{\ell 1}, i \in N, j, \ell \in M \tag{9.23}$$

$$z_{ij\ell} \geq q_{ij} - B_2(1 - s'_{\ell 1}), i \in N, j, \ell \in M \tag{9.24}$$

$$z_{ij\ell} \geq 0, i \in N, j, \ell \in M, \tag{9.25}$$

where B_1 and B_2 denote two big numbers. For this problem, they can be set to 1, the upper limit of q_{ij}.

The mathematical model with constraints (9.21) to (9.25) and others can then be formulated as a mixed integer programming problem, which is referred to as **Model 4**. Please refer to Zhang et al. (2012) for the detailed formulation.

9.5 An Illustrative Example

This section presents a small example to illustrate the differences between the above models. Suppose there are 5 open facilities and 10 population zones. Table 9.1 gives the travel times t_{ij} in hours and the fractions of clients h_i, and set $\lambda = 10$ clients/hour. The parameters used for Model 1 include $A_i = 1.0$ and $\gamma = 1.4$, those for Model 3 include $A'_i = 1$, $\gamma' = 0.2$, and $\mu_j = 2.5$, and that for Model 4 is $\beta = 0.9$. Note that we are only interested in the allocation results now, which are independent of the other parameters. These parameter values are chosen so that the total participation rates derived from these models are all approximately equal to 75 %.

Tables 9.2, 9.3, and 9.4 show the allocation results of Model 1, Model 3, and Model 4, respectively. Although the total participation rates are close, the allocations of clients to the facilities are very different. In particular, the main difference between Tables 9.2 and 9.3 is that clients at the same population zone, e.g., zones 4 and 9, may go to different facilities at equilibrium in Table 9.3. Note that although they go to multiple facilities, the utilities visiting these facilities are identical at equilibrium.

We also note that, in this example, the user-choice modeling does not have much impact on the number of clients from each zone, whereas its impact on the number of clients visiting each facility is significant. For instance, in Table 9.2, facilities 2 and 3 serve the smallest and largest number of clients, respectively; in contrast, the two facilities interchange their roles in Table 9.4.

9.6 Solution Methodologies

As some of the models, e.g., Models 1, 2, and 4, are formulated as MIPs, they can be solved directly using a professional solver, such as *CPLEX*. However, large-sized problems may not be solved within a reasonable time. Therefore, the existing studies

Table 9.1 Travel times and fractions

	Facility					
Zone	1	2	3	4	5	Fraction
	0.41	0.68	0.29	1.01	0.88	0.06
	0.84	0.43	0.78	0.18	0.41	0.15
	1.04	0.38	0.63	0.60	0.52	0.07
	0.84	0.53	0.42	0.44	1.07	0.07
	1.07	0.43	0.70	0.60	1.04	0.10
	0.31	0.62	0.94	0.53	0.64	0.02
	0.69	0.47	0.65	0.48	0.82	0.02
	0.19	0.28	1.05	0.25	0.63	0.18
	0.20	0.61	0.15	0.78	0.27	0.16
	0.95	0.44	0.99	0.70	0.22	0.17

Table 9.2 Allocation result of model 1

	Facility					
Zone	1	2	3	4	5	Total
	0	0	0.44	0	0	0.44
	0	0	0	1.26	0	1.26
	0	0.45	0	0	0	0.45
	0	0	0.39	0	0	0.39
	0	0.55	0	0	0	0.55
	0.15	0	0	0	0	0.15
	0	0.12	0	0	0	0.12
	1.44	0	0	0	0	1.44
	0	0	1.36	0	0	1.36
	0	0	0	0	1.33	1.33
Total	1.59	1.12	2.19	1.26	1.33	7.48

often proposed improved exact solution methods or heuristic methods to reduce the running time.

Verter and Lapierre (2002), for instance, presented two exact solution procedures for Model 1. Both procedures are intended to take advantage of the problem structure during the solution process. Specifically, recall that constraints (9.6) impose the closest facility allocations. However, majority of these constraints may not be binding at the optimal solution, they therefore suggested to add only the necessary constraints in each iteration of the procedure.

In contrast, other models with nonlinear formulations, e.g., Model 3, are difficult to be solved exactly. Hence, developing efficient heuristic methods are the

Table 9.3 Allocation result of model 3

	Facility					
Zone	1	2	3	4	5	Total
	0	0	0.45	0	0	0.45
	0	0	0	1.15	0	1.15
	0	0.53	0	0	0	0.53
	0	0.05	0.02	0.38	0	0.46
	0	0.69	0	0	0	0.69
	0.16	0	0	0	0	0.16
	0	0.16	0	0	0	0.16
	1.35	0	0	0	0	1.35
	0	0	1.08	0	0.13	1.22
	0	0	0	0	1.31	1.31
Total	1.51	1.43	1.56	1.53	1.43	7.48

Table 9.4 Allocation result of model 4

	Facility					
Zone	1	2	3	4	5	Total
	0.11	0.08	0.12	0.06	0.07	0.45
	0.17	0.25	0.18	0.31	0.25	1.15
	0.07	0.13	0.10	0.11	0.11	0.52
	0.08	0.11	0.12	0.12	0.06	0.49
	0.10	0.18	0.14	0.16	0.10	0.68
	0.04	0.03	0.02	0.03	0.03	0.16
	0.03	0.04	0.03	0.04	0.03	0.17
	0.34	0.32	0.15	0.32	0.23	1.36
	0.29	0.20	0.31	0.17	0.27	1.25
	0.19	0.30	0.18	0.23	0.36	1.25
Total	1.42	1.63	1.36	1.55	1.53	7.49

primary focus for these models. The heuristic methods are often based on the location-allocation framework:

Given a set of facility locations and the associated capacities, identify client flows from population zones to the facilities; determine the best set of locations and the associated capacities.

In this framework, **Alloc P** serves as a sub-routine for **Loc P**. For any set of locations and the associated capacities, with **Alloc P**, the number of clients from each population zone to each facility and the objective function value can be determined.

9.6.1 Allocation Algorithms

A variety of allocation algorithms have been proposed according to the nature of the different problems. This section primarily introduces an allocation algorithm developed by Zhang et al. (2010) for Model 3, which is to determine equilibrium flows given a set of facilities S. For the ease of exposition, consider all the facilities with one server, i.e., $s_j = 1$, $j \in S$.

When $s_j = 1$, each open facility becomes an $M/M/1$ queuing system. Thus, the mean waiting time formula (9.8) reduces to

$$\bar{W}(a'_j) = \frac{1}{\mu - a'_j}, \; j \in S. \tag{9.26}$$

Zhang et al. (2009) initially studied this allocation problem and proposed an allocation heuristic for **Alloc P**. Furthermore, Zhang et al. (2010) developed an exact allocation method. Specifically, they showed that the nonlinear complementarity problem (9.16) can be reformulated as the following optimization problem,

$$\begin{aligned} \text{Min} - \sum_{j \in S} \ln\left(\mu - \lambda \sum_{i=1}^{n} h_i x'_{ij}\right) + \lambda \sum_{i=1}^{n} h_i \sum_{j \in S} \left(t_{ij} - \frac{A'_i}{\gamma'}\right) x'_{ij} \\ + \sum_{i=1}^{n} \frac{1}{2\gamma'} \left(\sum_{j \in S} x'_{ij}\right)^2 \\ \text{s.t. } x'_{ij} \geq 0, i \in N, j \in S. \end{aligned} \tag{9.27}$$

They proved that the above optimization problem is convex. Since this is almost an unconstrained problem, it can be solved by a variety of optimization methods, such as the gradient projection method (Kelley 1999). They further proved the existence of the user-equilibrium, whereas the equilibrium in general may not be unique.

They showed that the above allocation method can directly be applied to find exact equilibrium flows when $s_j \leq 2$, as the mean waiting time formula for an $M/M/2$ queue has a similar structure to expression (9.26). However, the mean waiting time formula becomes more complex when $s_j > 2$. To make the allocation method applicable, they developed an approximation for the mean waiting time formula when $s_j > 2$, which also has a similar structure to expression (9.26). The above allocation method can then be used to find approximate equilibrium flows. Refer to Zhang et al. (2010) for details of the approximation.

9.6.2 Location Algorithms

Many heuristic algorithms have been developed for **Loc P**. Zhang et al. (2009) described four location heuristics for location decisions only, i.e., no capacity decisions considered, including a greedy-type heuristic and three meta-heuristics based

on a probabilistic adaptive search. Each of these heuristics applies three basic neighborhood move procedures as follows.

Remove Procedure The procedure starts with all potential facilities open. In each iteration, it removes one facility at a time until feasibility is obtained. The facility to be removed is chosen so that the total participation served by the remaining facilities is maximized.

Add Procedure The procedure is the opposite of **Remove Procedure**, starting with a feasible facility set. In each iteration, it adds a new facility. The new facility is chosen so that the total participation served by the located facilities is maximized, while maintaining feasibility.

Exchange Procedure This procedure attempts to improve a given feasible solution by swapping a facility in the current solution with a potential facility that is not currently open and then executing **Add Procedure**.

The greedy-type heuristic applies **Remove, Add**, and **Exchange** procedures in succession. Although it is fast to produce a feasible solution, it often results in a local optimum. In an effort to overcome this disadvantage, they also introduced three probabilistic adaptive search heuristics which are meta-heuristics. These meta-heuristics run the **Remove, Add**, and **Exchange** procedures in succession repeatedly. In each procedure, the heuristic is based on the randomized choice of available alternatives (alternatives here can represent facilities to be removed, added, or facility pairs to be swapped). The probability of selecting an alternative is proportional to the change in total participation once this alternative is selected.

Zhang et al. (2012) implemented a genetic algorithm to determine the best set of locations. They further compared the performances of one probabilistic adaptive search heuristic above and the genetic algorithm. In their computational experiments, they showed that the former exhibits slightly better accuracy but less reliability than the genetic algorithm.

Zhang et al. (2010) developed a tabu search procedure to make both location and capacity decisions for (**Loc P**), subject to a given number of available servers. Assuming that a facility can be allocated to at most K servers, each potential facility is divided into K pseudo-facilities. Similar to the genetic algorithm above, a solution is then composed of a series of binary numbers that correspond to each pseudo-facility. The tabu search heuristic starts from an initial feasible solution. Then, in each iteration, it applies **Remove** and **Add** procedures in succession. The procedure repeats until a stopping criterion is reached. When neighborhood moves are selected, tabu restrictions are used to prevent moving back to previously investigated solutions. Define a tabu list in which each value is associated with a pseudo-facility to represent its tabu status. Once removed or added to the set of open pseudo-facilities, a node is classified as tabu with a certain length, which represents the number of iterations in which the node typically will not be selected for removing or adding. However, even for a tabu node, it can still be selected for adding, if an aspiration criterion is satisfied. A typical criterion states that if a move produces a feasible

solution which is better than the best known feasible solution, then the tabu status is disregarded and the move is executed.

9.7 A Real-Life Application

Verter and Lapierre (2002) adopted their distance-based model (Model 1) in working with the Quebec Ministry of Health for determining the breast cancer screening centers in Montreal that should be included in the Ministry's program to subsidize mammograms for women between the ages of 50 and 69. At the time, there were 194,475 women in Montreal in this age group and 36 existing facilities with mammographic equipment. The problem was to determine which facilities to be accredited so as to maximize participation to the program. The Ministry made a policy decision to require a minimum of 4000 mammographies per year for facilities to be accredited. The Ministry's preference was to make their decisions based on the assumption that women travel a maximum of 7 miles to take mammograms. There were 497 population centers in representing spatial distribution of the potential clients, and a maximum participation rate was assumed to be 0.95.

Verter and Lapierre (2002) formulated the problem using Model 1. They have observed that the number of closest facility assignment constraints (9.6) proliferates with the problem size. Since the majority of these constraints may not be binding at the optimal solution, they experimented with alternative solution procedures that add only the necessary constraints at each iteration. Indeed, further computational experiments showed that the way the closest facility assignment constraints are formulated has a significant impact on the computational effort required to solve large scale problems. In tackling the Montreal application, they were able to reduce the computational time significantly by replacing (9.6) by constraints of the form suggested by Rojeski and ReVelle (1970). The basic premise of this formulation is as follows: each population center should be assigned to its closest alternative site, if there is a facility at that site. The assignment should be to the second closest alternative site, if the closest site is not open and there is a facility at the second closest site. The argument can be extended to third, fourth etc. closest alternative sites.

The optimal solution of Verter and Lapierre (2002) suggested accreditation of 17 of the 36 existing mammography facilities. The maximum expected participation would be around 50 %. They presented the solution to the Ministry, with the recommendation that a follow up study is performed to better understand the behavioural aspects of the clients' participation to the preventive breast cancer screening programs. Under time pressure to implement a solution, however, the Ministry went ahead and approached the proposed 17 mammography centres to recruit them to the program. A few of the centres have refused to be part of the subsidy program, and hence the Ministry accredited some additional centres, which were willing to work with the subsidy system albeit they were not part of the optimal solution.

Unfortunately, significant wait times were observed at some of the facilities after the system was running. One of the reasons for this is that Model 1, based on

distance only, does not take congestion into consideration. This motivated several follow-up studies (Zhang et al. 2009 and 2010) concerning the incorporation of congestion into the models. For example, Zhang et al. (2009) demonstrated that ignoring congestion would lead to significant overestimation of the participation rate, and that increasing the capacity of the accredited facilities would be a better policy than accrediting a larger number of small facilities. They also suggested, to some extent, to allocate resources on health promotion projects targeted to potential clients rather than capacity expansion. These latter recommendations have not been implemented due to the budgetary pressures in the health sector.

Turning to the big picture, note that breast cancer is the most common cancer diagnosed in Canadian women and is second only to lung cancer as the most common cause of cancer deaths among women. Annually, around 25,000 women are diagnosed with this condition and unfortunately around 5000 lose the battle to breast cancer. Just over 50 % of all breast cancer cases occur in women between the ages of 50 and 69. Accordingly, screening recommendations in Canada include a biennial mammogram of the breast and clinical breast examination in asymptomatic women in this age group. It is important that the breast cancer screening programs offer timely service and capture a significant proportion of the women in the target age group. The national target in Canada is 70 %.

9.8 Extensions

From the modeling perspective, all of the studies introduced above considered facility accessibility, such as distance and congestion, as the main determinant for participation. None of these studies provides empirical support for their assumptions. However, past empirical studies have suggested that factors other than accessibility may play a more important role in the decisions (Rimer et al. 1989; Kee et al. 1992; Munn 1993). Moreover, the models taking congestion into consideration represent walk-in facilities, where appointments are not needed. This enables the effective use of queuing models. Many preventive care services, including flu shots, anti-smoking advice, drug and alcohol abuse, nutrition advice, are offered in walk-in facilities. Many other preventive care programs, such as preventive cancer screening programs, however, are offered by the use of appointment systems. One of the main challenge concerning appointment systems is the fact that wait times for appointments are often not commensurate with the travel and service times at the facilities.

Recently, aiming at optimizing the configuration of a preventive cancer screening facility network, Kucukyazici et al. (2014) proposed a novel methodology that includes a variety of empirical and analytical methods. They first identified potential facility attributes that affect the client choice of preventive cancer screening facilities based on a focus group meeting and interviews. They then used stated preference discrete choice modeling (*SPDCM*) to model clients' preference over the identified attributes. In other worlds, they designed a survey and investigated the client choice

from survey data with discrete choice modeling. Results from the *SPDCM* analysis show that nursing staff's manner and knowledge regarding the target disease, waiting time for an appointment, and parking availability are the most influential attributes, among other significant factors such as travel time, communication during the screening process, and waiting time for result. Kucukyazici et al. (2014) also conducted latent class analysis to explore the existence of client heterogeneity. Further, they made use of the results from the econometric analysis to develop an empirically based simulation model, and incorporated meta-heuristic algorithms into the simulation. This simulation optimization approach was used to optimize the configuration of a facility network, by relocating a fixed number of servers within the network. To the best of our knowledge, this is the first study that integrates empirical analysis and simulation optimization to provide valuable insights into how to improve preventive cancer screening system performance through network design and configuration.

Aboolian et al. (2014) developed an ε-optimal algorithm for a variant of Model 3. In their model, the service rate at each facility is a decision variable rather than the number of servers. They reformulated the nonlinear problem as an MIP. The idea of their reformulation is to replace the decision variables, the service rates, by the waiting times and then to approximate the waiting time functions using the *TLA* method developed in Aboolian *et al*. (2007). Aboolian et al. (2014) present a realistic example based on the hospital network of Toronto in order to highlight the potential use of their model in gathering policy insights.

Although Model 4 is formulated as a mixed integer programming problem, medium- or large-sized instances cannot be solved to optimality within a reasonable time. Haase and Muller (2013) presented an alternative formulation and showed that medium-sized instances can be easily solved to optimality, or at least close to optimality, by a professional solver in a reasonable time. They further proposed an approach to derive a tight lower bound to the problem.

9.9 Conclusion

This chapter reviews several studies that apply location analysis for designing preventive care programs. The generic problem is to design a preventive care facility network so as to maximize the level of participation. The number of facilities to be established, the location of each facility, and the capacity at each facility are the main determinants of the configuration of a facility network. The existing models in the literature typically address some of the differentiating features of preventive care, such as elastic participation, congestion, and the user-choice environment. Accessibility of facilities is usually considered to influence the participation rate. Some papers used travel distance or travel time alone as a proxy for the accessibility, while others also take congestion into consideration. Moreover, the deterministic-choice and probabilistic-choice models are two alternatives in the literature to address the client choice on where to receive the service. In terms of solution methodology, both

exact and heuristic solution methods have been proposed to solve the problems. Evidently, this is a fledgling research stream and as discussed in the previous section there are a number of ways that the state of the art can be improved.

References

Aboolian R, Berman O, Verter V (2014) Maximal accessibility network design in the public sector. Forthcoming in Transportation Science

Berman O, Drezner Z (2006) Location of congested capacitated facilities with distance sensitive demand. IIE Trans 38:213–231

Facione NC (1999) Breast cancer screening in relation to access to health services. Oncol Nurs Forum 26:689–696

Falkenheimer SA (2004) The adequacy of preventive health care: does the health care provider matter? http://www.cbhd.org/content/adequacy-preventive-health-care-does-health-care-provider-matter. Accessed 13 April 2015

Goldsmith J (1989) A radical prescription for hospitals. Harvard Bus Rev 67:104–111

Gu W, Wang X, McGregor SE (2010) Optimization of preventive health care facility locations. Int J Health Geogr 9:17–32

Gupta S, Chintagunta PK, Kaul A, Wittink DR (1996) Do household scanner data provide representative inferences from brand choices: a comparison with store data. J Mark Res 33:383–398

Haase K, Muller S (2013) Insights into clients' choice in preventive health care facility location planning. Working Paper. University of Hamburg

Health Canada (2007) Mammography. http://www.hc-sc.gc.ca/hl-vs/iyh-vsv/med/mammog-eng.php. Accessed 13 April 2015

Huff DL (1962) Determination of intra-urban retail trade areas. Real Estate Research Program, UCLA

Kee F, Telford AM, Donaghy P, O'doherty A (1992) Attitude or access: reasons for not attending mammography in northern Ireland. Eur J Cancer Prev 1/4:311–316

Kelley CT (1999) Iterative methods for optimization. Frontiers in applied mathematics. SIAM, Raleigh

Kleinrock L (1975) Queueing system I: theory. Wiley, New York

Kucukyazici B, Song L, Verter V, Zhang Y (2014) Incorporating client choice into facility network design of preventive cancer screening programs. Working Paper. McGill University

Marianov V, Serra D (2002) Location-allocation of multiple-server service centers with constrained queues or waiting times. Ann Oper Res 111:35–50

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

McFadden D (1974) Conditional logit analysis of quantitative choice behavior. In: Zarembkar P (ed) Frontiers in economics. Academic Press, New York

Munn EM (1993) Nonparticipation in mammography screening: apathy, anxiety or cost? N Z Med J 106/959:284–286

Nakanishi M, Cooper LG (1974) Parameter estimation for a multiplicative competitive interaction model: least squares approach. J Market Res 11(3):303–311

Public Health Agency of Canada (2006) Organized breast cancer screening programs in Canada—Report on program performance in 2005 and 2006. http://www.phac-aspc.gc.ca/cd-mc/publications/cancer/obcsp-podcs05/index-eng.php Accessed 14 July 2015.

Rimer BK, Keintz MK, Kessler HB, Engstrom PF, Rosan JR (1989) Why women resist screening mammography: patient-related barriers. Radiology 172/1:243–246

Rojeski P, ReVelle CS (1970) Central facilities location under an investment constraint. Geogr Anal 2:343–360

Verter V, Lapierre SD (2002) Location of preventive health care facilities. Ann Oper Res 110: 123–132

Walker K (1977) Current issues in the provision of health care services. J Consum Aff 11:52–62

Wang Q, Batta R, Rump CM (2002) Algorithms for a facility location problem with stochastic customer demand and immobile servers. Ann Oper Res 111:17–34

World Health Organisation (2002) Integrating prevention into health care. http://apps.who.int/bookorders/anglais/detart1.jsp?codlan=1&codcol=15&codcch=739. Accessed 13 April 2015

Zhang Y, Berman O, Verter V (2009) Incorporating congestion in preventive healthcare facility network design. Eur J Oper Res 198:922–935

Zhang Y, Berman O, Marcotte P, Verter V (2010) A bilevel model for preventive healthcare network design with congestion. IIE Trans 42/12:865–880

Zhang Y, Berman O, Verter V (2012) The impact of client choice on preventive healthcare facility network design. OR Spectr 34:349–370

Zimmerman S (1997) Factors influencing hispanic participation in prostate cancer screening. Oncol Nurs Forum 24:499–504

Chapter 10
Modeling the Location of Retail Facilities: An Application to the Postal Service

Anthony M. Yezer and James W. Gillula

10.1 Introduction

There are many well-developed theoretical models of facility location for various types of retail activity. Given a spatial distribution of consumers with known demand curves and a function that transforms distance into transportation cost, it is possible to specify the spatial demand curve facing a firm located at any point in economic space. A plant level cost function, usually a combination of fixed and variable cost, must be added to the model along with assumptions regarding the nature of spatial competition from producers of close substitutes. Once this is done, the model can be solved, and the revenue and cost consequences of alternative facility location and marketing strategies can be computed in a fairly straightforward fashion.

The model developed here is based on a well established theoretical literature that began with Launhardt's (1885) model of the spatial demand curve. Consumers spread over space face two "prices" when they make a consumption decision. The first is the price charged for the good or service at the facility where it is sold (this is the free on board (*FOB*) or mill price). The second is transportation cost to and from the facility. Therefore, the cost of acquisition rises with distance to the facility and accordingly the quantity demanded falls. Subsequently, Reilly (1931) called this the "law of retail gravitation" because demand declines with distance, although the effect of distance on transportation cost is generally found to be either linear or concave. Neidercorn and Becholod (1972) showed that the spatial demand curve could be derived from consumers maximizing a utility function. These insights regarding spatial demand formed the basis for modern location theory whose historical development is reviewed nicely by Beckmann and Thisse (1986). The specific theoretical issues involved in modeling facility location are considered in Ye and Yezer (1992).

A. M. Yezer (✉)
George Washington University, Washinton, D.C., USA
e-mail: yezer@gwu.edu

J. W. Gillula
IHS Economics, Washinton, D.C., USA
e-mail: James.Gillula@ihs.com

H. A. Eiselt, V. Marianov (eds.), *Applications of Location Analysis,* International Series in Operations Research & Management Science 232, DOI 10.1007/978-3-319-20282-2_10

For a recent statement of the theoretical problem of spatial competition see Braid (2011, 2012).

Going from the available theory to an operational empirical model should not be difficult given sufficient data, particularly micro data on individual household demand and location. However, privacy concerns generally make public disclosure of such data problematic. Accordingly, while there is little doubt that many fine models of spatial competition for retail activities exist, they have been developed for the private use of individual firms. Individual retailers can collect information on the purchasing patterns and location of customers by providing "buyer loyalty" discount cards. Public access to these models is limited. Furthermore the extreme requirements for micro data needed to estimate and implement these models limit the ability to use them as the basis for model calibration. Data relating consumption behavior of households, including both what is purchased and precisely where it is purchased, to their individual characteristics is seldom available to researchers or practitioners.

Instead, researchers must generally be content with information on facility sales and characteristics of the market area in which the facility is located as a basis for estimating spatial demand functions. The study described in this Chapter uses this readily available data to formulate a rigorous model of a spatial retail market that is capable of serving as the basis for a business model of the costs and benefits of alternative patterns of facility size, type, and location. The model can simulate a full range of economic outcomes of alternative patterns of production that can be used for both positive and normative economic analysis. The method proposed here accommodates the possibility that demand is based on both the residential location and the workplace of potential consumers. Perhaps the closest model, at least in terms of spatial demand modeling, to the problem considered here is Davis' (2006) effort to characterize spatial competition among movie theaters. The specific application here is to retail postal services, including facilities providing, stamps, package services, post office boxes, and special services. The fact that the United States Postal Service (*USPS*) is a spatial monopolist is an aid in constructing the model, but the techniques developed here could be applied to other forms of spatial competition. The model is complicated because the *USPS* has two different modes of delivering retail postal services, retail postal stores (*PSs*) and limited-service contract postal units (*CPUs*). The scale of the model is massive, applying to the entire contiguous 48 U.S. states.[1]

10.2 General Statement of the Postal Location Problem

The *USPS* retail facility location problem is a special version of standard problems where facilities subject to increasing returns to scale serve spatially distributed consumers. Based on the location of their residence or place of work as well as their

[1] The retail postal store is commonly known as a main post office, branch, or station. It provides a full range of postal services.

income and personal characteristics, individuals demand services provided by a facility. The facility has production costs that increase at a decreasing rate with output, i.e., increasing returns to scale at the facility level. This suggests producing in a small number of large facilities to minimize production cost. But consumers must pay both the cost of production and the cost of accessing the facility, which suggests producing in a large number of small facilities to minimize transportation cost. The optimal facility location problem is to select the size and spacing of facilities which maximizes the net willingness of consumers to pay for postal services, i.e., which maximizes the net benefit produced by the location of retail facilities.

The general statement of the problem provided below could be applied to a variety of facility location problems. The price of services is fixed at all plants so that it is possible to write total revenue per square mile, R, in the following form:

$$R = PQ = f(D,\ r), \tag{10.1}$$

where P is the fixed price per unit service, Q is the quantity of services provided, D is the density of postal customers (households and firms) per square mile, r is the market radius of the store, and the function $f(D,\ r)$ is an empirical relation to be determined using USPS data on revenue and market characteristics. It is anticipated that $\partial f/\partial D = f_D > 0$ because greater density of customers yields greater demand density and that $f_r < 0$ because a longer market radius makes access to the store more expensive.

Retail postal services at a store are produced through "windows." There is a fundamental production relation between the number of required windows and revenue that is given by[2]

$$W = g(R). \tag{10.2}$$

Note that the equations in this section measure relations per unit area so that R is revenue per square mile and W is windows per square mile. Based on *a priori* expectations, window requirements should increase at a decreasing rate in revenue. Estimates of $g(R)$ using *USPS* data on revenue and windows of facilities serving different market areas confirmed the expectation, i.e., that the first derivative of $g(R)$, $g_R > 0$ and its second derivative $g_{RR} < 0$. There is a capacity constraint on the ability to deliver retail services and generate revenue from a given number of windows. However, there is an added complication because customers will not wait in long lines for most postal services. This customer impatience is one reason that empirical measurement of the relation between windows and revenue is necessary.

There is a relation between the number of windows in a postal store and the required labor, so that labor per square mile of market area L is given by

$$L = h(W). \tag{10.3}$$

[2] For facilities that are connected to the Point Of Sale (*POS*) system, a window is a register that is generating significant amounts of revenue. The ability to observe revenue generation at the individual register level is a major advantage in the calibration of cost functions. This will be discussed in some detail below.

Windows generate staffing requirements and, based on *USPS* data, there is a clear relation between the number of windows operating and the number of *USPS* personnel in the facility so that $h_W > 0$. Once again the fact that the analysis is conducted in terms of labor and windows per unit market area has no effect on the sign of the relation between the two variables.[3]

Similarly, there is a relation between interior space per square mile of market area S, at a store and the number of windows given by

$$S = i(W), \tag{10.4}$$

where $i(W)$ must also be determined empirically using *USPS* data and design criteria. But it is clear that $i_W > 0$ *and* $i_{WW} < 0$ because more space is required to accommodate the additional windows.

Translating clerical labor and facility space into *USPS* costs is fairly straightforward. Labor cost per square mile, C_L, is the product of the number of workers per square mile and the earnings per worker, E, or $C_L = EL$. The cost of space to serve demand generated by each square mile of market area will depend on space per square mile and on the size of the market area, i.e., on r, *viz.*,

$$C_S = j(S, r) \tag{10.5}$$

where: C_S is the cost per square mile of market area, and $j(S, r)$ is a facility rent equation that gives the relation between the space per square mile of market and the facility rent per square mile. Unlike labor cost, space rental costs vary with location. Consistent with real estate literature, estimates from *USPS* data indicate that $j_S > 0$, $j_{SS} < 0$, and $j_r < 0$ because additional space costs more but larger facilities that serve a larger market radius cost less per square foot.

One function of facility location models is normative, i.e., to solve for the optimal spacing between facilities. In most cases, this maximization problem is straightforward once a criterion for judging optimal facility provision is selected.[4] One standard choice is maximization of the surplus of revenue over cost. Note that this is not a monopoly solution because price is fixed and the *USPS* is a non-profit producer. Instead this corresponds to a welfare criterion for consumers who ultimately must pay a higher price for the cost of additional facilities.[5] Of course, the model also functions in a positive, i.e., non-normative, mode because it can simulate the implications of changes in location and other operating decisions for revenue, cost, etc.

As in most facility location models, there is a revenue recognition problem. Because *USPS* is producing more than retail services all of the revenue should not be

[3] Employees who staff windows have other duties. Some of these are directly related to revenue generation, handling incoming mail. Others may involve sorting mail for delivery. These issues will be discussed later in this chapter.

[4] For an excellent discussion of the literature on location theory see Beckmann and Thisse (1986) or Beckmann (1999).

[5] For a specific discussion of the public facility location problem in continuous space see Ye and Yezer (1992, 1993).

attributed to retail operations. Revenues collected at the retail level must also cover costs of mail handling, processing, transportation, and delivery, hereafter referred to as "mailing cost." These operations impose costs on the *USPS* that are a substantial fraction of total revenue. Mailing costs have been estimated as 75 % of *USPS* operating cost. It is not at all clear that this average cost figure should be deducted from revenue for purposes of the analysis conducted here, particularly given that marginal cost is likely well below average cost. At this point, a fraction, $0 < (1 - \Psi) < 1$ of revenue is attributed to covering mailing cost, which leaves a fraction Ψ of revenue to be counted as net revenue associated with retail services.

Based on Eqs. 10.1–10.5 above, the welfare criterion may be stated as

$$\left.\begin{aligned} \omega &= \Psi R - C_L - C_S = \Psi f(D, r) - Eh(W) - j(i(w), r) \\ \omega &= \Psi f(D, r) - Eh(g(f(D, r))) - j(i(g(f(D, r))), r) \end{aligned}\right\} \tag{10.6}$$

where ω is surplus per unit area and maximizing ω maximizes surplus for the entire system. The important insight from Eq. (10.6) is that, given the density of customers that determines demand per unit area, welfare maximization is based on the selection of postal store spacing at intervals of $2r$. This also determines facility size because D and r determine postal store revenue from Eq. (10.1), $R = f(D, r)$. Once revenue is determined, all the other operating parameters of the model such as number of windows, facility size, and workers per facility follow recursively. Thus, the facility location problem can be solved by choosing the appropriate plant spacing, r, and tracing the implications of this for all other postal store characteristics using the model.

For given demand density, surplus is maximized by setting the derivative of (10.6) with respect to r equal to zero, i.e.,

$$\frac{d\omega}{dr} = \psi f_r - Eh_W g_R f_r - j_S i_W g_R f_r - j_r \tag{10.7}$$

Equation (10.7) can be solved to determine the optimal market radius between stores in areas where demand density is given. Recall from the above discussion that $f_r < 0$, $g_R > 0$, $h_W > 0$, $i_W > 0$, $j_r < 0$, and $j_S > 0$, which implies that the first term on the right side of (10.7) is negative, while the other terms are positive. As expected, increasing r tends to have a negative effect on surplus per unit area due to the fall in revenue as average demand per square mile falls, but there is a compensating rise in welfare as larger sales per postal store lowers space cost per square foot. It is apparent that labor cost per square mile of market area falls as revenue falls. Clearly the necessary condition to maximize welfare in Eq. (10.7) is a function of demand density D. In these models, the effects of demand density are complex because a rise in demand density tends to increase costs of additional r because the fall in revenue with added market area is larger but the fall in cost is larger also.

10.3 Specific Statement of the Postal Location Problem

This section shows how the standard setup for a facility location problem can be adapted to the specific retail location issues confronting the *USPS*. Much of this adaptation is necessary because of the limited data availability, and other aspects are necessitated because it is not appropriate to assume that current operations of the *USPS* are technically efficient, i.e., that current facility location, size, and operation are designed to minimize operating cost. Put another way, current operating costs of the *USPS* are not based on facility sizes and locations that minimize cost or maximize net revenue. This means that the cost function for providing postal services must be constructed and calibrated based on first principles rather than on the actual costs observed at different facilities.

Going from the general statement of the postal store location problem in the previous section to a specific model that can be calibrated and applied requires that the general functional forms be replaced by specific equations. There are two steps in this process. First, functions must be replaced with equations that can be justified in terms of economic theory. Second, the parameters of the equations must be estimated using *USPS* data on revenue and facility characteristics.

This section discusses the first of these tasks. As suggested in the general solution, there are several fundamental relations to be understood: the determinants of revenue per square mile ($f(D, r)$), and then the relations between revenue and windows ($g(R)$), space and windows ($i(W)$), employees and windows ($h(W)$), and space and rents ($j(S, r)$). As noted above, each of these relations should be based on a facility operating at or near capacity. In this section, each of these issues is discussed in turn.

10.3.1 Determinants of Revenue Per Square Mile of Market Area

Demand for retail postal facilities is based on local employment and residential population. Following standard practice in the literature discussed in Beckmann (1999), assume that demand for retail postal services is given by

$$q = \alpha - \beta(P + tx) \tag{10.8}$$

where q is the quantity of services used by an individual household or firm, P is the price of postal services, t is the cost of transportation to and from the facility, x is the distance to the nearest facility and $\alpha > 0$ and $\beta < 0$ are parameters. This is a simple linear demand curve where the total cost of the services is the sum of the price and the transportation cost, tx, to access the facility. The problem is simplified by the fact that P is known to be constant across facilities.

Assuming that facilities serve a circular market area of radius r so that the market area is $\pi r^2 = A$, the distance between facilities is $2r$, and that demand for postal services comes from households or firms (the illustrative equation could apply to

either group although α, β, and t would be different for households and firms), total demand at any facility is given by

$$QA = \frac{f(D,r)}{P} = \int_0^r 2\pi x D(\alpha - \beta(P + tx))dx, \tag{10.9}$$

where Q is revenue per square mile, QA is total sales per facility, and D equals the density of households or firms measured as households or employees per square mile. There is a direct connection between Eqs. (10.9) and (10.1) above, and multiplying by P gives an expression for total revenue that is equivalent to (10.1):

$$RA = PQA = Pf(D,r) = P\int_0^r 2\pi x D(\alpha - \beta(P + tx))dx \tag{10.10}$$

Evaluating (10.10) we find that total revenue per facility can be written as:

$$RA = Pf(D,r) = P2\pi r^2 D\left[\left(\frac{a}{2}\right) - \left(\frac{bP}{2}\right) - \left(\frac{brt}{3}\right)\right] \tag{10.11}$$

Recognizing that P is a constant, this equation states that total revenue of a facility is the product of the total number of customers in the market area ($\pi r^2 D$) and a constant, $2P[(a/2) - bP/2)]$, and a constant multiplied by the radius of the facility service area, $2P(-bt/3)r$. It is convenient to write Eq. (10.11) as

$$RA = Pf(D,r) = (\pi r^2 D)\left[P\,(a - bP) - \left(\frac{2bPt}{3}\right)r\right] \tag{10.12}$$

or

$$RA = Pf(D,r) = (\pi r^2 D)[\theta - \tau r], \tag{10.13}$$

where $\pi r^2 D = N$ is the total number of households or employees (i.e. potential customers) in the service area of the facility, θ is the constant ($aP - bP^2$), and τ is another constant, ($2Pbt/3$) which includes the effects of transportation cost on demand. Note that the effect of increasing market radius on revenue depends on two effects. First is the negative effect due to the τr term as higher transportation cost cuts demand by effectively raising the cost of accessing retail services. As suggested above, $f_r < 0$ as average revenue per square mile falls when r rises. However, raising r increases market area and hence raises the number of customers ($dN/dr = 2\pi rD > 0$). The second effect on $f(D,r)A$ dominates, and total revenue per facility rises when r increases.[6] In the next section, Eq. (10.13) will be shown to provide the basis for the empirical estimation of the facility revenue equation that is vital to the exercise conducted here.

[6] There are technical reasons for this result that need not concern us here.

10.3.2 Relation Between Revenue and Retail Input Needs

Postal stores produce retail postal services principally using inputs of labor and space.[7] The challenge is to relate revenue generated, principally walk-in revenue, at stores to required inputs of labor and space that can then be priced to form a cost function. Fortunately, the *USPS* collects revenue data through the point of sale (*POS*) system of wired registers in a substantial number of retail stores. It is possible to observe the relation between the number of registers in a store and the amount of revenue. There are obvious technical reasons to expect that a given quality of service can only be maintained in a postal store experiencing additional walk-in demand by increasing the number of operating registers. In the *POS* data some registers are clearly dedicated to retail services and are termed "front" and others are "back," although they sometimes take in significant amounts of revenue. Some front registers are located at formal front windows and others at counters. In the *POS* data it is possible to identify all registers that perform a significant revenue-generating function and, as noted above, these are termed windows. If W denotes windows per square mile, then total windows for a facility with market area A is WA. It is expected that there should be a relation between total postal store revenue and required windows that takes the following specific functional form

$$WA = \kappa(RA)^{\lambda}, \tag{10.14}$$

where κ is a constant approximately equal to the reciprocal of the annual revenue expected from operation of a single window and λ is the relation between additional revenue and the need for added windows.[8] It is expected that λ will be slightly less than unity because adding more registers permits specialization in services and should raise revenue per window slightly. Note that (10.14) expresses the relation between annual revenue per store and number of windows per store assuming that the windows are kept busy. Clearly, the causality runs from revenue to windows. Doubling windows should have no effect on revenue unless the previous number of windows was inadequate to deal with demand and significant lines and wait times were experienced by patrons. Sometimes window services are used by customers who are not generating revenue. For purposes of this model, it is assumed that these activities are proportional to revenue generation and simply reduce the potential of windows to generate revenue. It is possible in the *USPS* data to find cases in which facilities with several windows generate very little revenue. It is important to evaluate (10.14) based on facilities where the registers are operating near capacity and, given technical efficiency cannot be assumed, this creates an empirical challenge for the model.

---

[7] The emphasis here is on operating inputs and operating costs, and general and administrative costs are ignored.

[8] Equation (10.14) relates to the function in (10.2) as $W = g(R) = \kappa(RA)^{\lambda}/A$

10.3.3 Relation Between Windows and Labor Requirements

Registers must be staffed. While employees have duties beyond direct revenue generation, the data reveal a reliable relation between windows and the number of clerical employees operating the registers at each facility.[9] The general form of this relation was determined to be

$$LA = \varphi(WA)^{\chi} = \varphi(\kappa(RA)^{\lambda})^{\chi} \tag{10.15}$$

where φ indicates the clerical employees required to staff a postal store with a single window and $0 < \chi < 1$ indicates the rate at which employment expands with additional windows.[10] Returns to scale in staffing arising from the ability to relieve workers systematically suggests that doubling the number of windows does not require doubling the number of employees, and the data confirm this expectation.

Translating labor inputs into labor cost requires an estimate of unit labor cost for employees. Because workers can be reassigned across facility size categories in a flexible manner, a constant labor cost was used for all units of clerical labor.

10.3.4 Relation Between Windows and Postal Store Space Needs

There are two ways to develop the relation between windows and required postal store space needs. The first is to use the traditional economic approach of taking data from actual *USPS* operations to estimate a production relation. Alternatively, design criteria for facilities and operating standards developed for services can be used to infer a relation between revenue and space needs. Current postal store designs relate number of windows to overall area. These two approaches will be discussed in turn.

First, the traditional economic approach to relate inputs to outputs is through a production function. In this case, the labor input is ignored, leaving only space input, and the general form of the postal services production function may be written as

$$SA = \mu(WA)^{\nu} = \mu(\kappa(RA)^{\lambda})^{\nu}, \tag{10.16}$$

where SA is the total size of the facility in square feet expressed as the product of size per square mile, S and market area in square miles, A, W is windows per unit, and μ and ν are parameters to be determined empirically.[11] It is anticipated that μ reflects

[9] The analysis relies on the assumption that the non-register activities of employees observed in the data do not vary systematically with the number of windows. Estimation of equations designed to explain the number of retail employees consistently revealed that the number of windows is the dominant factor determining employment. To the extent that this is not true, some adjustment of the relation between size and labor cost attributed to revenue generation is needed.

[10] Equation (10.15) relates to the function in (10.3) as $L = h(w) = \varphi(WA)^{\chi}/A$

[11] Equation (10.16) relates to the function in (10.4) as $S = i(W) = \mu(WA)^{\nu}/A$.

the space required for a facility with a single window and $0 < \nu < 1$ because space needs increase at a decreasing rate with windows due to certain indivisible space requirements for doorways, bathrooms, etc. The discussion here is in terms of total space per facility because that is what is observed in the data and this section is designed to set up the empirical analysis to follow.

The second method that could be used to relate postal store size and revenue would be to use facility design criteria that have been established by the *USPS* along with current criteria that relate window operation time to retail revenue. This might be termed an engineering approach to the production function relation. Currently, the *USPS* has postal store design criteria that relate the number of windows and number of carriers assigned to a facility to the physical size of the facility. This is done through a form of a table, in which different combinations of windows and carriers are related to different designs.[12]

Both methods were employed and found to be generally consistent. However, the final model calibration was based primarily on the second, engineering, approach because it presumably reflects current USPS thinking about the relation between space and windows. The economic approach gave larger space requirements, which may reflect past policies of providing greater space per window.

10.3.5 Relation Between Postal Store Size and Cost

The *USPS* rents over 70 % of its facilities. The lease information allows estimation of hedonic rent equations that give the expected rent per square foot in different geographic locations as a function of facility size and other characteristics. Based on this work, as well as the general literature on the relation between the size and rent of office space, the costs of space should increase at a decreasing rate with facility size. This leads to the following basic specification of a postal store rental cost equation

$$C_s A = \Upsilon(SA)^{\phi} = \Upsilon(\mu(WA)^{\nu})^{\phi} = \Upsilon(\mu(\kappa(RA)^{\lambda})^{\nu})^{\phi} \tag{10.17}$$

where $C_S A$ is the total rental cost of the facility written as the product of rent per square mile of market area (C_S) and market area (A), and Υ and ϕ are parameters to be determined. Because rent increases at a decreasing rate with size, it is expected that $0 < \phi < 1$. Rents vary by location and this effect is captured by Υ, which varies based on estimated rental hedonic equations for different locations.

[12] At the same time, USPS has standards for labor productivity that relate window operating time to revenue. Taken together, these standards would allow one to work backwards from retail revenue expectations for each facility to the size of facility needed to meet those expectations.

10.3.6 Solving the Model

Recall from Eq. (10.6) above that surplus per unit area may be written as the difference of revenue net of mailing cost and the sum of labor and space cost per square mile or that

$$\omega = \Psi f(D, r) - Eh(g(f(D, r))) - j(i(g(f(D, r))), r) \tag{10.6}$$

or

$$\omega = \Psi R - C_S - C_L. \tag{10.18}$$

The discussion presented above has demonstrated that

$$RA = (\pi r^2 D)[\theta - \tau r], \quad C_s A = \Upsilon(SA)^{\phi} = \Upsilon(\mu(WA)^{\nu})^{\phi} = \Upsilon(\mu(\kappa(RA)^{\lambda})^{\nu})^{\phi}$$
$$= \Upsilon(\mu(\kappa[(\pi r^2 D)(\theta - \tau r)]^{\lambda})^{\nu})^{\phi}$$

and that

$$C_L A = LA = E\varphi(WA)^{\chi} = E\varphi(\kappa(RA)^{\lambda})^{\chi} = E\varphi(\kappa[(\pi r^2 D)(\theta - \tau r)]^{\lambda}))^{\chi}.$$

The market area of these circular markets is given by $A = \pi r^2$. Dividing the expressions for *RA, $C_S A$,* and $C_L A$ by A and collecting terms gives an expression for ω that is not too messy:

$$\begin{aligned}\omega =& \psi D(\theta - \tau r) - E\varphi\kappa^{\lambda\chi}\pi^{\lambda\chi-1}r^{2(\lambda\chi-1)}(D(\theta - \tau r))^{\lambda\chi} \\ &- \Upsilon(\mu^{\phi}\kappa^{\lambda\nu\phi}\pi^{\lambda\nu\phi-1}r^{2(\lambda\nu\phi-1)}(D(\theta - \tau r))^{\lambda\nu\phi}\end{aligned} \tag{10.19}$$

The problem of choosing facility location and size to maximize surplus is thus reduced to choosing r to maximize ω. This is done by setting the derivative of (10.18) equal to zero, i.e., $d\omega/dr = 0$. To get some insight into this result note that λ, ν, φ, χ are all on the (0, 1) interval. When they are estimated empirically using *USPS* facility data for purposes of model calibration, they all lie on the [0.8, 0.95] interval. This means that the exponents of r such as $2(\lambda\chi - 1)$, and $2(\lambda\nu\phi - 1)$, are both negative. Consider the three terms in (10.18). The derivative of the first with respect to r is negative and the derivatives of the second two terms with respect to r are both positive. The second derivative of the first term with respect to r is zero and the second derivatives of the other terms with respect to r are negative. Taken together, this means that, provided the first derivative is initially positive for lower values of r, Eq. (10.18) produces an internal maximum of welfare for some positive level of r. There is, however, no guarantee that surplus is maximized at a positive value of

ω. In these cases, given the business model underlying the postal store, there is no positive *r* for which surplus is positive and the model minimizes the expected loss.[13]

In particular, when demand density is low, as it is in much of the U.S., there is no value of *r* that produces a positive surplus. However, service must be universal as an obligation of the *USPS*. In addition, there are significant transaction costs in modifying the number and size of facilities in an area. Finally, there is a choice between operation as a postal store or substituting some lower cost contract postal units into the retail mix. For these reasons, the business model constructed is designed to quickly simulate a variety of alternative configurations of *PS* and contract postal units (*CPUs*) in a given market area so that the user can trace the pattern of incremental change in surplus associated with changing the level and mix of retail services. Given the nature of the model, the surface that maximizes surplus is quickly traced and the mix associated with a maximum is readily apparent. It was only after the model was constructed and operated that this feature of its performance was apparent. Furthermore, operation revealed that there are often a number of alternative configurations of retail facilities that generate similar levels of surplus. In the presence of significant transaction costs, the configuration that involves the least disruption to current service is likely preferred to another that promises a small gain in surplus but causes significant disruption.

10.4 Calibration of the Location Model

The central challenge to constructing a business model of spatial location is calibration, particularly specification and estimation of the spatial demand relations. Based on the above discussion, it is evident that estimates of a number of critical parameters, specifically ψ, θ, τ, κ, λ, μ, ν Υ, and ϕ will be needed to evaluate the model. This estimation problem is solved by relying on the theory discussed in the previous sections, and specifically on Eqs. (10.13), (10.14), (10.15), (10.16) and (10.17).

However, this discussion was based on fairly straightforward application of standard theory. In the case of postal services, as with many other retail and wholesale activities, the actual market situation is not really accurately represented in the standard theory. This is particularly true of the demand for retail postal services, which is discussed in some detail in the next section. Issues with the calibration of the supply side of the model are then considered, followed by a presentation of the results of the entire calibrated model.

[13] The possibility of a corner solution at $r = 0$ is not important here. As a practical matter, solutions for sufficiently small radius are not feasible because they would imply operation with fractional windows and employees. There is a minimum size of feasible facility due to the need to have at least one window and retail worker.

10.4.1 Estimating and Calibrating a Model of Retail Demand

Consider the standard approach to spatial demand where consumers are located in space and assumed to face a delivered price based on the *FOB* price plus transportation cost from the plant. In the case of many retail purchases, the origin of the trip is not the residence, but it may be the place of employment instead. Furthermore, shopping trips may have multiple purposes and the postal service may be relatively unimportant compared to other reasons for travel. Therefore, the "origin" of a retail postal service shopping trip may be a residence in the market area, an employment location in the market area, or another location in the market area that motivated the consumer to be in the vicinity of the postal store. However, as long as the origin of the trip is a random location within the market area and density of trip origins is constant across the area, the total demand for retail postal services can be expressed in terms of Eq. (10.13) as

$$RA = Pf(D, r) = (\pi r^2 D)[\theta - \tau r], \tag{10.20}$$

except that this equation can be applied to each individual type of customer. That is, the expression for total revenue based on trips from home would involve D_R or residential density and could be written as the sum of two terms, $\theta\pi r^2 D_R$ and $(-tr)\theta\pi r^2 D_R$. The first term is the total demand that would be experienced in the absence of transportation cost, i.e., if all customers in the market were located adjacent to the postal facility. The second term is the fall in demand that arises with increasing market radius. Thus, demand is determined by the number of consumers in the market and the radius over which they are spread. In addition to residential demand, employment based demand will depend on the density of employment in the market area, or $\pi r^2 D_E$ $(\theta - tr)$. Further experimentation indicated, as might be expected, that the income elasticity of demand for postal services is positive and hence household demand is increasing in income of residential households in the market area.

These special characteristics of retail demand, rather than serving as an impediment to the estimation of a postal services demand relation, actually simplified the estimation and calibration of the model. While it is very difficult to get individual household, firm, or employee data on use of postal services, observing the number of households and employees in the market area of postal stores is straightforward. Indeed, census data allow disaggregation of population by demographic characteristics and income and employment by industrial sector by market area. This allowed specification and estimation of a model of total revenue, and even revenue by product type, using revenue observed by postal store.

One other consideration influencing total retail revenue at *USPS* facilities is the presence of competition from other mail delivery services. Two different measures of local competition are available. One is the total employment in private mail services located in the *ZIP* codes served by the postal store, and the other is the number of these facilities within the market area of the facility. Given that either measure of competition could reduce effective demand, both are included in the estimated

revenue equation. Finally, transportation cost, t, differs geographically. The primary determinant of variation in t should be density of population and employment, or urban location. Accordingly, the estimation allows the effect of r on revenue, and thus by implication the effect of t, to vary for facilities located in urban areas.[14]

Based on these considerations, the final form of the total revenue equation is

$$RA = \alpha_0+\alpha_1 N_E + \alpha_2 N_E U + \alpha_4 N_H + \alpha_5 N_H U + \alpha_6 N_H M + \alpha_7 N_E I_P + \alpha_8 N_E I_C + \alpha_9 N_E r + \alpha_{10} N_E U r + \alpha_{11} N_H r + \alpha_{12} N_H r U + \varepsilon \tag{10.21}$$

where N_E is total employment in the market area of the facility, N_H is total households in the market area, U is a 0–1 dummy variable equal to unity in large urban areas, M is median household income in the area, I_P is an index of private mail employment in the area, I_C denotes an index of competing private mail establishments in the market area, r is the market radius of the facility (half the distance to nearby postal facilities), ϵ is an iid random error term, and the α*'s* are parameters to be estimated.[15] There is particular interest in α_9, α_{10}, α_{11}, and α_{12}, because these reflect the effect of increased radius on demand, that is, they reflect transportation cost to the facility, represented by τ in Eq. (10.13). The terms multiplied by the urban dummy U, allow the effect of households and firms on demand to be different in large cities and the effects of market radius to differ because of higher transportation costs. The other estimated coefficients reflect the components of θ in Eq. (10.13).

A number of alternative statistical techniques were used to estimate Eq. (10.19) using data for postal stores located in the 48 contiguous U.S. states. Alaska, Hawaii and U.S. territories were excluded based on differences in topography and/or climate. Facilities reporting revenue under $ 100 and those with a market area less than one-tenth of a square mile were eliminated. The preferred estimation was accomplished using robust regression and is displayed below as Eq. (10.20):[16]

$$\begin{aligned} RA =& 13{,}244.7 + 971.9\, N_E + 13.6\, N_E U + 48.8\, N_H - 15.2\, N_H U \\ &+ 0.0000344\, N_H M - 188.2\, N_E I_P - 117.3\, N_E I_C - 14.0\, N_E r \\ &- 39.9\, N_E r U - 4.7\, N_H r + 3.5\, N_H r U \end{aligned} \tag{10.22}$$

The number of households and employees in the market area of the store has a major effect on revenue. Additional employment is more important than additional households. This may reflect demand by workers based on their workplace or demand by

[14] A variety of efforts to customize the estimation to differences in local characteristics that might influence transportation costs have demonstrated some promise but are not discussed here in order to focus on the main elements of the calibration effort.

[15] The index I_P is 0 when there is 0 private mail employment in the service area of the facility, 1 if employment is between > 0 and < 10, 2 if employment is at least 10 and < 20, 3 if employment is at least 20 and < 30, etc. Similarly, the I_C index is 0 if there are 0 competitors in the service area, 1 if there is 1 competitor, 2 if there are 2 competitors, 3 if there are 3–5 competitors, and 4 if there are 5 or more competitors.

[16] This equation was estimated using 21,898 observations with $F(11, 21{,}886) = 38{,}000$. The t-ratios of the estimated coefficients were all larger than 4.0.

the firms themselves. Higher household income has a modest positive effect on demand. As expected, the presence of competitors reduces demand, and the effect of a larger market area also reduces demand. Location in a large urban area increases the effect of market radius on demand for employment as expected, but it also reduces the effect of households on demand. It is not surprising that, relative to household demand, employment demand is more important in larger urban areas. In general, the effect of households on demand for *USPS* services is lower in large urban areas. All estimated coefficients are highly statistically significant. This will be used as the basic demand equation in the simulation experiments performed in the next section.

The results in Eq. (10.20) provide some insight into the determinants of demand at individual postal stores. It is encouraging that the signs of the estimated coefficients and their general magnitudes agree with prior expectations. This effort to understand demand is directed toward the purpose of a spatial model in which the effect of proximity is of major importance. The results presented here should not be confused with a more fundamental examination of the determinants of the demand for retail services which should involve, at a minimum, greater disaggregation of both household and employment by type as well as greater consideration of the relation between distance and actual transportation cost.[17]

Note that demand is estimated at the facility level rather than aggregated over market areas that involve several facilities. Actual operation of the model involves prediction of outcomes for larger market areas, generally counties, because a change in facility size and spacing generally involves transition of several facilities. However, aggregation must be done with some caution because comparisons will be based on averages for a given area but averages are a linear aggregation of magnitudes and spatial aggregation is often nonlinear. Indeed, the model itself is very nonlinear.

Consider the follow illustration of a potential confusion that can arise in the spatial aggregation process. Assume that there are two facilities and that each has a market radius of 2 miles. The average radius will be 2 miles and the average market area will be $3.14(2^2) = 12.56$ square miles because each market area is identical. Now assume that there are two facilities whose market radii are 1 and 3 miles. The average market radius is still 2 but the market areas are 3.14 (for the 1 mile radius) and $3.14(3^2) = 28.26$ (for the 3 mile radius). Thus the average market area is $(3.14 + 28.26) = 15.7$. The average area in the "real world" where the radius varies, will be larger than average area in the model solution where radius is constant. In general, the greater the variance in the radius, the larger the excess market area above that calculated based on the average radius.

When comparing results from the model simulations where radius is constant with actual operating results, the difference between the market area of an average radius when radius is uniform and the real world, where the variance in the radius is significant, should be considered. The difficulty can usually be overcome, in this case, by working in average area per postal store. Alternatively, analysis can only be

[17] It would also be useful to consider the precise shape of the market area.

done for facilities which all have a similar market radius. In the calibrations reported here, one or the other of these approaches is taken to avoid aggregation problems.

10.4.2 Estimating and Calibrating a Postal Services Cost Function

Ordinarily a cost function relates physical output to costs of producing that output. However, measures of the quantity of postal services produced at individual locations are not available. Therefore, costs must be expressed as a function of revenue. The fact that prices are uniform nationally, means that there is potentially a linear relation between prices but variation in output mix across facilities can be significant and therefore there is a real possibility that revenue differences do not reflect services.

The cost function calibration begins by estimating the relation between revenue and number of windows in the postal store. Measures of the number of "front" registers and "back" registers along with their time of operation and revenue generation are included in the available data along with the number of windows in the store.

The general form of Eq. (10.14) relating windows to revenue, $WA = \kappa(RA)^{\lambda}$ appears straightforward. There are difficulties in estimating the values of the parameters κ and λ. First, the total number of windows, WA, is an integer, whose value at $RA = 0$ should, of course, equal 0. The analysis should focus on values for $WA = 1$, 2, 3, or 4 as these constitute the relevant range for most retail stores. Second, this is a frontier relation in that it reflects the revenue per window achievable under x-efficiency, i.e., given the way postal stores operate, they are producing given services at lowest cost.[18] Observations from the many facilities that have more windows than necessary, as evidenced by the fact that they are never, or hardly ever, shown to be open should be ignored.

Because of the need to observe a frontier, where the relation between revenue and windows is based on high rates of utilization, is so important, three approaches were taken to eliminate observations from inside the frontier. First, the mean revenue per window for different numbers of windows was computed using only observations in which revenue per window was more than one standard deviation above the mean. Second, only facilities in which all the windows were operating for most of the time during selected high-demand weeks (the second week in December and April) were used. Third, the demand equation in (10.21) was used to estimate the demand at each facility, and those facilities, in which the demand per window was predicted to be one standard deviation above the mean were used. In all three cases, a clear

[18] The cost function used in the model is based on x-efficiency in that it takes the current rules and procedures for operating postal stores as given. The term "technical efficiency" used here is based on x-efficiency. It may be that different work rules or ways of deploying labor and capital, perhaps on-line retail aids, could raise output per facility. These innovations are not considered here, although they could be added to the model easily by recalibrating the cost functions.

pattern was observed in which, for facilities judged to be on the frontier, revenue per window for facilities with one window was approximately $ 400,000 per year and revenue per window increased slightly as the number of windows increased.[19] As a result of this, the value of κ was set at 0.000005 and the value of λ was set at 0.95. Using these parameters, the relation between windows and revenue reflects a postal store that is technically efficient.[20]

Next it is necessary to calibrate the relation between windows in operation and labor requirements. The labor per window function, given in Eq. (10.15) above as $LA = \varphi(WA)^{\chi} = \varphi(\kappa(RA)^{\lambda})^{\chi}$, is also a frontier concept. It reflects use of labor in a technically efficient fashion, i.e., at facilities where revenue per employee and window is high. There are some facilities for which the ratio of clerical workers per window is quite high, i.e., greater than two. In such cases, it is likely that these workers are performing a variety of duties unrelated to revenue generation. Fortunately, the *USPS* point of sale data allow observation of the hours spent logged in to a register by individual employee identification code by day for each facility. It is possible to identify the number of different employees who spent a significant amount of time at terminals during each of the 3 weeks that were observed (the second weeks in December, April, and August). Employees who logged on for less than 10 h/week were not counted as retail employees. Estimates of the parameters in Eq. (10.15) were based on facilities where the same numbers of workers were spending significant time logged on to a register during each of these 3 weeks.[21] The estimates were all based on facilities, in which the retail operation appears to be operating near capacity so that they were judged to be technically efficient. In this case, facilities with revenue per window above $ 250,000/year were selected.

As anticipated, the number of employees per window falls with the number of windows, and the fitted parameters of Eq. (10.15) were $\varphi = 1.6$ and $\chi = 0.9$. Thus, within the data on facilities with revenue per window greater than $ 250,000, those with one window had an average of 1.6 employees per window and this ratio fell slowly as the number of windows increased. Calibration of the parameters ϕ and χ is accomplished based on a table of the mean values of retail workers for facilities with different numbers of windows subject to the requirement that revenue

[19] The pattern of revenue per window was also tested for the two high demand weeks, and the small increase in revenue per window with number of windows was observed.

[20] A further check on the calibration of revenue per window was done by using data from the *USPS* Window Operations Survey (*WOS*) for FY 2008. The *USPS* has a standard for time per transaction that determines the hours "earned" at a facility window. At facilities reporting actual hours equal to earned hours, the average annual revenue per hour of actual window operation was computed and, when this was multiplied by annual hours per window, the average annual revenue per full time window in operation at these facilities was found to be $ 642,720, well above the $ 400,000 set as efficient revenue per window in this model calibration. Evidently the standard for technically efficient window operation adopted for calibrating the cost function is not particularly rigorous compared to *USPS* standards for earned hours.

[21] This does not mean that the employees in 1 month were the same as those in another month. The employee counts were based on a single week in each month. It does mean that the number of retail workers counted in each of the 3 months was constant.

per window exceeded $ 250,000. The value of ϕ is the mean for facilities with one window, and χ followed based on the rate of decline in retail workers per window as windows expanded beyond one.

The functional form selected for the relation between interior space and number of windows was given in Eq. (10.16) by $SA = \mu(WA)^{\nu} = \mu(\kappa(RA)^{\lambda})^{\nu}$. This is also a frontier relation which should be measured for facilities operating under *x*-efficiency. Empirical estimation of (10.16) produced estimates of space required that were well in excess of the standards embodied in current USPS facility design criteria for new facilities. Given the forward-looking nature of this analysis, an engineering approach was taken to the evaluation of the parameters of (10.16) using the standards for new retail facilities. These standards allow for P.O. Box sections and carrier workroom space that is an increasing function of the number of windows. While there are several alternative designs, it appears that 1800 square feet is an average standard for a postal store with a single window and that space increases at a slightly decreasing rate with the number of counters. Accordingly, Eq. (10.16) was parameterized by setting $\mu = 1800$ and $\nu = 0.9$. Thus, a facility with one window occupies 1800 square feet of interior space, including the service lobby, P.O. Box lobby, workroom for mail processing, bathrooms, etc. This increases to 3359 square feet for two windows, 4839 square feet for three windows, etc.

Estimation of cost per square foot has been accomplished using the *USPS* Facility Database (*FDB*) in support of P.O. Box fee-setting for over a decade. This provides substantial evidence on the parameters of Eq. (10.17), *viz.*,

$$C_s A = \Upsilon(SA)^{\phi} = \Upsilon(\mu(WA)^{\nu})^{\phi} = \Upsilon(\mu^{\phi}(\kappa^{\nu}(RA)^{\lambda})^{\nu})^{\phi}. \tag{10.23}$$

Specifically, it is well established that the cost per square foot falls with facility size and that the elasticity of cost with respect to size is about 0.8, i.e. $\phi = 0.8.\Upsilon$ is the cost per square foot appropriate for the real estate market under consideration and ranges from under $ 5/square foot in rural areas to over $ 45/square foot in the higher density portions of larger urban areas.

10.5 Demonstrating the Location Model

The calibrated model can be solved for operating revenue and cost as a function of market radius and the levels of key variables such as population and employment density, household income, competitors, urban character and rental price of space in the area being served. Because of the recursive nature of the model, determination of the radius or facility spacing implies unique values for all the operating characteristics of the postal stores serving the market area.

Investigation of the demand for *USPS* services indicated that the structure of demand was different in large urban areas than in the rest of the U.S. This is evident in the estimation results for Eq. (10.20) reported above where the effect of households and employment on demand is different in large urban areas than in the rest of the data. Compared to residential households, employment is a relatively more important determinant of demand in large urban areas. Also, the effect of market radius on demand for services is greater in large urban areas.

Table 10.1 Estimated revenue and cost at different radii for large urban areas

Radius (miles)	Revenue ($)	Windows	Empl.	Interior space (sf)	Cost ($)	NetRev ($)/SqMi
0.5	146,186	0.4	0.7	795	60,711	15,773
0.75	302,700	0.8	1.3	1481	112,208	22,161
1	510,656	1.3	2.1	2316	174,552	25,725
1.5	1,055,126	2.6	3.8	4308	322,503	29,025
2	*1,728,068*	*4.2*	*5.8*	*6568*	*489,733*	*29,801*
2.5	2,477,950	5.9	7.9	8939	664,648	29,265
3	3,253,245	7.7	10.0	11,282	837,166	27,935
3.5	4,002,422	9.4	12.0	13,469	998,012	26,081
4	4,673,953	10.8	13.7	15,379	1,138,339	23,858
4.5	5,216,309	12.0	15.0	16,892	1,249,452	21,368
5	5,577,959	12.8	15.9	17,889	1,322,572	18,680
5.5	5,707,375	13.1	16.2	18,243	1,348,563	15,846
6	5,553,028	12.8	15.8	17,820	1,317,554	12,907
6.5	5,063,387	11.7	14.6	16,468	1,218,306	9900
7	4,186,925	9.8	12.4	13,998	1,036,897	6867

Employment density 355 workers/square mile, *residential density* 2259 households/square mile, *income* $ 62,382

In addition to differences in the parameters of the demand equation, the costs of space are higher in large cities. There are also differences in the density of households and employment and in the density of competition. These differences lead to rather different implications for the allocatively efficient size and spacing of facilities between large urban areas and the rest of the U.S., and the nature of postal store operations is generally very different. All this is illustrated in Tables 10.1 and 10.2 and the accompanying figures below.

Table 10.1 and the associated Fig. 10.1 show the solution of the model for postal stores located in large urban areas based on both the demand function in those areas and the higher density of households, firms and high space cost. A total of 3423 postal stores fall into this category.[22] The average household density in these areas is 2259/square mile, employment density is 355/square mile, and median household income is $ 62,382/year. The model shows that allocative efficiency is achieved when net revenue per square mile is maximized at about $ 29,801 with a market radius of 2 miles. The postal store associated with this maximum has approximately 4 windows, 6 retail workers, 6600 square feet of interior space and $ 1.73

[22] The 3423 also reflects the number of facilities with sufficiently complete data to be used in computation of characteristics of the actual condition of facilities in this group.

Table 10.2 Estimated revenue and cost at different radii for non-urban areas

Radius (miles)	Revenue ($)	Windows	Empl.	Interior space (sf)	Cost ($)	NetRev ($)/SqMi
0.5	32,799	0.1	0.2	222	16,505	− 134
1	89,932	0.3	0.5	525	38,814	1959
1.5	182,347	0.5	0.9	960	70,721	2895
2	307,748	0.8	1.3	1502	110,308	3469
2.5	463,839	1.2	1.9	2133	156,320	3852
3	648,324	1.7	2.5	2841	207,797	4118
3.5	858,907	2.2	3.2	3613	263,945	4303
4	1,093,291	2.7	3.9	4441	324,072	4430
4.5	1,349,181	3.3	4.7	5316	387,561	4514
5	1,624,280	4.0	5.5	6230	453,848	4564
5.5	1,916,292	4.6	6.4	7175	522,408	4587
6	*2,222,922*	*5.4*	*7.2*	*8146*	*592,751*	*4589*
6.5	2,541,872	6.1	8.1	9136	664,408	4572
7	2,870,847	6.8	9.0	10,138	736,934	4540
7.5	3,207,551	7.6	9.9	11,146	809,899	4495
8	3,549,687	8.3	10.8	12,155	882,889	4438
8.5	3,894,960	9.1	11.7	13,159	955,498	4373
9	4,241,073	9.9	12.6	14,153	1,027,335	4298
9.5	4,585,731	10.6	13.4	15,130	1,098,012	4216
10	4,926,636	11.4	14.3	16,087	1,167,151	4128
11	5,588,007	12.8	15.9	17,916	1,299,323	3934
12	6,206,816	14.2	17.4	19,600	1,420,905	3721
13	6,764,695	15.4	18.8	21,096	1,528,985	3493
14	7,243,276	16.4	19.9	22,366	1,620,654	3251

Employment density 26 workers/square mile, *residential density* 130 households/square mile, income $ 45,269

million/year in revenue. The table also shows the performance characteristics of choosing alternative patterns of postal store spacing.

The actual radius of all the postal stores in these urban areas was 2.1 miles and revenue per facility was $ 1.25 million. The model estimate of an allocatively efficient radius of 2.0 miles matches the average for this group of 3423 postal stores. The fact that spacing of postal stores in this sample of urban areas is consistent with the spacing suggested for allocative efficiency does not imply that the size or spacing of these facilities is efficient for two reasons. First, the analysis is based on "average" spacing and, in a sample that aggregates over all large urban areas, the average may be appropriate but its detailed distribution on a city by city basis may be flawed. However, the evidence suggests that, the spacing of these facilities is not

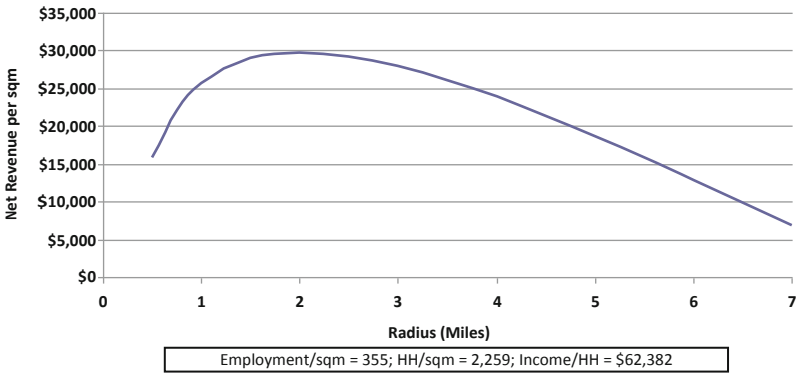


Fig. 10.1 Radius vs. net revenue per square mile: large urban areas

grossly inefficient. There may still be problems of lack of technical efficiency as the analysis is disaggregated to the individual *MSA* level. Nevertheless, the fact that the average size for all postal stores in the sample is about 7000 square feet, compared to the optimal 6600 square feet associated with the radius of 2 miles indicates that facility size is not, on average, inconsistent with allocative efficiency either.[23]

Second, the discussion thus far has not considered technical efficiency of the actual postal stores found in these large urban areas. The results for allocative efficiency above were based on the assumption that the actual size, in terms of interior space, number of windows, and clerical staffing was technically efficient. Technical efficiency requires that the actual sizes be efficient given the actual demand for services at each postal store. Measurement of technical efficiency for many postal stores is limited because the number of windows in operation and the level of clerical staffing are only observed for facilities with *POS* terminals. Nevertheless, a rough estimate of technical efficiency can be obtained by using data from the *POS* facilities to impute windows and retail workers for the non-*POS* postal stores. In the case of the following analysis, this imputation was done by relating facility size (interior space) to numbers of windows and workers for those facilities with *POS* terminals and then using the resulting relation to estimate windows and workers for non-*POS* postal stores.

The result of the imputation is that the average numbers of windows, retail workers, and interior space can be substituted into the cost equations and the imputed characteristics of the actual postal stores in large urban areas can be compared to the technically efficient costs. In Table 10.1 below, a technically-efficient facility serving the allocatively efficient average market radius of 2 miles has 4.2 windows, 5.8 retail workers, 6568 square feet of interior space. The total cost for the 12.56 square mile market area is $ 489,733 or $ 38,991/square mile. The imputed characteristics of the current average postal store in large urban areas, where actual market

[23] Specifically, average facility size for those facilities with $< 20{,}000$ square feet is 7000 square feet. Larger facilities likely include significant mail processing and/or vehicle storage and maintenance activities.

radius is about 2 miles, are 2.95 windows, 4.63 workers, and 16,484 square feet of interior space, yielding an imputed cost per square mile of $ 33,275. This is a remarkable $ 5716 less than the technically efficient cost per square mile.

Compared to the technically efficient postal store, actual facilities have fewer windows and retail workers but more interior space on average. Is it possible that the actual operation of *USPS* retail postal stores in large urban areas has achieved some kind of super efficiency? Are the standards for technical efficiency in the model too low? Alternatively, it may be that the numbers of windows and employees provided in these facilities in large urban areas are too small resulting in long lines and poor service. Insight into these questions is gained from comparing the actual revenue per facility with model estimates based on the demand equation. Actual revenue per facility is $ 1.25 million compared to the demand model estimate of $ 1.7 million, a difference of almost $ 40,000/mile of market area. The reason that the model requires more windows and retail workers is that the model generates estimates of significantly more revenue than the actual postal stores generate. Although these are rough estimates based on averages over thousands of postal stores, the suggestion is that, while the average spacing of postal stores (and hence the average number of facilities) is consistent with allocative efficiency, the number of windows and clerical staffing per facility appears to be too low given the demand for postal services.

It may be that significant waiting time at postal stores in large urban areas deters potential customers from using the *USPS*. A rough estimate of the net revenue loss per square mile due to this under provision of window space would be $ 40,000 of extra revenue per square mile less approximately $ 6000 in extra cost per square mile or $ 34,000/square mile. This is slightly larger than the net benefit per square mile of the average postal store in a large urban area under allocative efficiency, i.e., it is substantial. Some caution is warranted here because these estimates are based on averages of very large and diverse facilities located across all large urban areas. Still, the analysis of technical and allocative efficiency is consistent in the conclusion that the number and spacing of postal stores in large urban areas is appropriate on average and that the number of windows and workers or alternative mechanisms for providing retail services should, if anything, be larger.

Table 10.2 and Fig. 10.2 illustrate the model solution in the average area outside large cities. This includes smaller cities, towns, and rural areas. The average density of households is dramatically lower at 130/square mile as opposed to 2259 for large cities. Employment density is 26/square mile compared to 355 and there are fewer competitors. The model is solved for the revenue, employment, and size implications of alternative market radii separating postal stores under conditions of technical efficiency. Allocative efficiency is achieved when net revenue per square mile is maximized at $ 4589 when market radius is 6 miles. This is a substantial contrast to the large city solution where net revenue per square mile is larger by a factor of 6.5 and radius was one-third as great. The postal store that maximizes net revenue per square mile has approximately 5 windows, 7 employees and 8000 square feet of interior space.

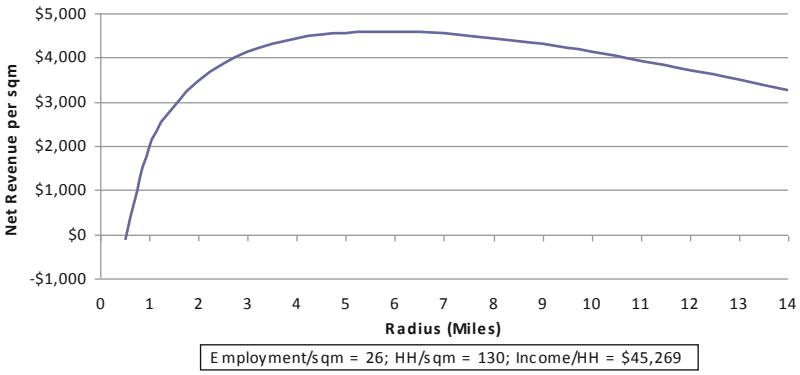


Fig. 10.2 Radius vs. net revenue per square mile: non-urban areas

Data on actual postal stores in this outside large city group indicates that the average radius is 4.7 miles, revenue is only about $ 410,000/year, and the average facility size is 3100 square feet.[24] There are 22,811 postal stores in this group.

Comparing allocative efficiency with a radius of 6 miles with the actual radius of 4.7 miles from Table 10.2 it appears that net revenue per square mile is $ 75 ($ 4589 − $ 4514) per mile higher with a radius of 6 miles. This may seem like a small margin but the 22,811 postal stores in question have an average market area of 106 square miles and this means that the total difference in annual net revenue per square mile implies a total difference of $ 180 million/year. This is a rough estimate of gains that could be achieved because it is an average over areas with very different levels of service. In general, the greater the geographic disaggregation, the larger the welfare gains from right-sizing and right-spacing the postal stores.

It is also possible to conduct an analysis of the technical efficiency with which retail services are currently produced. As discussed above for the case of postal stores in large urban areas, the challenge for measuring costs of current operations is that numbers of windows and retail workers are only observed for POS postal stores. Once again, the answer to the missing observations is to impute the numbers of windows and retail workers. The imputation produced estimates, for the average facility, of 2.13 windows, 3.46 retail workers, and 4764 square feet of interior space for the 22,811 postal stores in the non-urban sample. This produced a cost per square mile of market area of $ 4486. These magnitudes are all substantially above the technically efficient averages of 1.1 windows, 1.75 workers, 1964 square feet of interior space, and $ 2264 cost per square mile for technically efficient facilities located with a market radius averaging 4.5 miles. The difference in cost per mile, $ 2222 = $ 4486 − $ 2264, when multiplied by the 22,811 postal stores and the 106 square miles of average market area produces an increase in total cost due to technical inefficiency of $ 5.4 billion/year. The standard cautions regarding estimates

[24] Note that predicted revenue at a radius of 4.5 miles is much larger than actual revenue. This illustrates the difficulty of applying the demand model to large diverse areas and also may point to a further problem in revenue generation in low demand density area.

based on imputed values of windows and employees as well as averages computed over all facilities located outside large urban areas are in order here. It may be that the relation between the size and number of active retail windows in non-*POS* facilities is much lower than that for *POS* facilities. If true, the technical inefficiency estimate would be substantially lower. Furthermore, the ratio of employees per window for non-*POS* facilities may be significantly lower than that for *POS* facilities. This would also lower the $ 5.4 billion/year estimate of technical inefficiency. Even if this cost figure were reduced by half due to systematic differences between *POS* and non-*POS* facilities, the annual cost of technical inefficiency is impressive and the question of differences between facility types should be investigated.

Clearly, location in large cities generates far larger net revenue per unit area than in the rest of the country. Total surplus per postal store is the multiple of area and surplus per unit area. Because market area increases with the square of the radius, the total surplus of the large city postal store is $ 374,300 while the total surplus of the optimal facility outside a large city is $ 518,740. Differences in surplus per postal store are far smaller than differences in surplus per square mile, and total surplus is actually larger for the facility outside large cities because the market area is 113 square miles instead of 12.5 square miles.

Maximization of total surplus across the nation is achieved by maximizing net revenue per square mile and then multiplying by the 4.1 million square miles of U.S. territory. The dataset used in this study has 28,110 postal stores whose market area totals 2,681,694 square miles. The mean overall market radius is 4.33 miles. The mean market radius in large urban areas is 2.1 miles, virtually identical to the optimal radius produced in the results in Table 10.1. The mean market radius in the remainder of the sample is 4.7 miles, considerably smaller than the 6.0 miles characterizing the surplus-maximizing solution in Table 10.2. These results suggest that the actual market radius of postal stores, and by extension the number of postal stores, in larger urban areas is close to the optimum and that there are too many postal stores placed too close together in the rest of the country. Note that the implication for the number of facilities of having a radius of 6 miles is dramatically different than for 4.7 miles. The ratio of market areas is approximately 2 to 1 (actually 113/69 or 1.64 to 1), implying that the optimal number of facilities in the portion of the U.S. outside the largest cities is 61 % of the current number.

The analysis of technical efficiency in large urban areas compared to the rest of the U.S. (the “non-urban areas”), although requiring imputation of some very important data, produces results that are quite complementary with those from the analysis of allocative efficiency. For large urban areas, the number and spacing of current *USPS* facilities appears consistent with allocative efficiency and the numbers of windows and workers available appears to be slightly below that needed for technical efficiency. It may be that, in large urban areas, failure to provide adequate retail inputs is costing the *USPS* significant amounts of revenue. Further study of waiting times and revenue would be needed to make such a determination. For the rest of the U.S., just as allocative efficiency requires fewer postal stores located farther apart, technical efficiency implies a substantial reduction in the number of windows and workers. The estimated gain from moving from the current level of operations in

these non-urban areas to a technically efficient condition is $ 5.4 billion/year, and the further gain if these changes were associated with a move to an allocatively efficient number and spacing of postal stores is $ 180 million/year. Although these are estimates and better data on non-*POS* facilities might change them significantly, their magnitudes suggest that using modern modeling of facility location could achieve substantial gains for the *USPS*.

10.6 Modeling the Effects of Alternative Outlets

The *USPS* has recently been expanding the availability of retail postal services through the use of Contract Postal Units (*CPUs*) as an alternative to postal stores modeled above. The *CPU* provides a range of postal services including virtually all mailing services but generally does not provide post office boxes.[25] There are approximately 3500 *CPUs* in operation, which means that there are many markets without significant *CPU* penetration. Adding *CPUs* to the model provides an opportunity to demonstrate the manner in which the approach can be generally adapted to a range of alternative retail environments.

The process of adding *CPUs* involves substantial negotiation and locations of the facilities are generally predetermined based on other types of retail activity. In some cases, agreements are signed with a single retailer for *CPUs* to be located at a number of existing facilities. However, the intent of the *USPS* is to provide additional retail postal service capacity without incurring the full cost of a postal store. Accordingly, in contrast to postal stores where size and location are often the result of a historical process that has little to do with current economic conditions, *CPU* activity should be motivated by potential demand in the area. Furthermore, given that compensation of *CPU* owners is based on a percentage of revenue, *CPU*s are unlikely to remain in operation unless they generate significant revenue.[26]

Estimation of a revenue model such as that in Eq. (10.20) for *CPUs*, is therefore problematic because the presence of a *CPU* in the area is likely to be a result as well as a cause of the level of revenue generated in the market. *CPUs* are not likely to open and continue in operation unless the market generates significant amounts of postal revenue. In view of this, a modified instrumental variables approach was used to estimate *CPU* revenue generation. The first stage involved estimating a market revenue equation like Eq. (10.20) for all market areas that did not have a *CPU*. The predicted value of market area revenue was then used in a second stage equation to predict the revenue generated by *CPUs* in markets that actually had one or

[25] The *CPU* should not be confused with stamps sold on consignment by a range of ordinary retailers who do not provide mailing services.

[26] There is a substantial annual birth and death rate for *CPUs* both due to economic forces and changes in the business model of the retailer hosting the operation.

more *CPUs*.[27] Formally, the two stage estimation procedure involved the following approach shown as the system (10.22) and (10.23):

$$RA = \alpha + \beta_E N_E + \beta_{EU} N_E U + \beta_H N_H - \beta_{HU} N_H U + \beta_{HM} N_H M - \beta_{EP} N_E I_P - \beta_{EC} N_E I_C - \beta_{Er} N_E r - \beta_{EUr} N_E U r - \beta_{Hr} N_H r + \beta_{HUr} N_H U r + \varepsilon_{RA} \tag{10.24}$$

$$R_{CPU} = \rho_{CPU} N_{CPU} RA^* + \rho_{CPUr} N_{CPU} RA^* r + \varepsilon_{CPU} \tag{10.25}$$

Here the Eq. (10.22) is simply the facility revenue equation estimated for all markets without a *CPU*, and the lower *CPU* revenue equation relates total *CPU* revenue in the market area of the postal store to the product of the number of *CPUs*, N_{CPU}, and the estimated total revenue for the market from (10.22) noted RA^*. The next term is this same product multiplied by the radius of the market area, $N_{CPU} RA * r$, ρ_{CPU}, and ρ_{CPUr}, are parameters to be estimated and ε_{CPU} is an iid error term. Note that (10.23) is specified without a constant term because logically $R_{CPU} = 0$ when $RA^* = 0$.

Estimates of (10.23) reflecting the result of estimating the system of Eqs. (10.22) and (10.23) are displayed in Table 10.3. In keeping with the general finding that density of population has a significant effect, likely through transportation cost differentials, on the spatial demand curve, different versions of the equation are shown for areas with 4 different levels of population density, ranging from very low (less than 50 households per square mile) to high (more than 500 households per square mile). Estimates of the share of revenue going to a *CPU*, as indicated by estimated values of the ρ_{CPU}, and ρ_{CPUr} parameters, tend to increase with density. As transportation costs rise with density, *CPUs* become more attractive substitutes for postal stores within a market area of given size.[28]

Of course, part of the revenue at *CPUs* is displaced from postal stores. Modeling the effects of *CPUs* requires estimating both revenues at the CPU and declining revenue at postal stores. For the reasons noted above, total market revenue, including both postal stores and *CPUs* located in a given market area, is likely a cause of *CPU* revenue. Therefore, the effect of adding *CPUs* on total market revenue was estimated using the same type of two equation system described by Eqs. (10.22) and (10.23) except that a total revenue equation was substituted for (10.23):

$$R_T = \xi_S RA^* + \xi_{CPUr} N_{CPU} RA^* r + \varepsilon_T \tag{10.26}$$

[27] One complication was the case in which a *CPU* appeared to fall within the geographic market area of more than one postal store. In this case, the *CPU* was shared out to the market area proportionally based on distance from each of the postal stores.

[28] Of course market size falls significantly with density ranging from 5.2 miles at the lowest density to 1.4 miles for the highest density areas. The share of revenue going to the *CPU* also increases with market radius based on the second parameter in the *CPU* revenue equation.

Table 10.3 Estimates of equations (10.23) & (10.24): Determinants of *CPU* revenue and market revenue with *CPUs*

Dependent variable	$\rho_{CPU}(\rho_{CPUr})$ in Eq. (10.23)	*t*-ratio	Dependent variable	$\xi_S(\xi_{CPUr})$ in Eq. (10.24)	*t*-ratio
Lowest density coefficients					
N_{CPU} *RA**	0.122***	16.5	*RA**	0.992***	246
N_{CPU} *RA** *r*	0.0019	1.13	N_{CPU} *RA**	0.027***	25
Modest density coefficients					
N_{CPU} *RA**	0.115***	10.1	*RA**	0.967***	129.7
N_{CPU} *RA** *r*	0.0083**	2.38	N_{CPU} *RA** *r*	0.045***	18
Medium density coefficients					
N_{CPU} *RA**	0.093***	6.23	*RA**	0.996***	88.6
N_{CPU} *RA** *r*	0.023***	3.81	N_{CPU} *RA** *r*	0.039***	7.9
High density coefficients					
N_{CPU} *RA**	0.127***	25.6	*RA**	1.00***	90.6
N_{CPU} *RA** *r*	0.0089***	5.04	N_{CPU} *RA** *r*	0.055***	6.7
All density coefficients					
N_{CPU} *RA**	1.27***	25.7	*RA**	1.064***	271
N_{CPU} *RA** *r*	0.0088***	5.04	N_{CPU} *RA** *r*	0.045***	27

* significant at the 10 % level;
** significant at the 5 % level;
*** statistically significant at the 1 % level

where R_T is total market area revenue, *RA** is predicted revenue from areas with no *CPUs*, N_{CPU} is the number of *CPUs* in the market area, *r*, is the market radius and $_t$ is an error term. The estimate of ξ_{CPUr} is the fractional change in total market revenue of adding a single *CPU* in a market area whose radius is 1 mile.

Estimation results for Eq. (10.24) are shown in Table 10.3, so that they can be compared to the results for Eq. (10.23). As was the case with the ρ_{CPUr} parameter estimates in Eq. (10.23) the estimates of ξ_{CPUr} rise with density of the market area. These results reflect the same increased substitution effect as transportation costs rise. Note that the predicted revenue of *CPUs* expressed as a fraction of *RA** based on estimates of Eq. (10.23) is generally about twice as large as the expected effect of adding a *CPU* on total area revenue using Eq. (10.24) estimates. This difference reflects the revenue displacement effect, which appears to be close to 50 %. Put another way, adding a *CPU* to an average sized market area for a given density (where market area falls significantly with density) results in *CPU* revenue that is about 12 % of postal store revenue, of which about 7 % is revenue that is displaced from the postal store and 5 % is additional revenue. The ratio of the total revenue effect to the displacement effect rises with the radius of the market area as expected.

These results are easily translated into a calibrated model of the effects of adding or subtracting *CPUs* on net revenue because the cost function of *CPUs* is simply a

constant fraction of revenue. The model can be run consecutively with a changing mix of postal stores and *CPUs* to determine the net revenue implications of changes. The model does not solve for an "optimum" presence of *CPUs* because they do not provide all postal services and the USPS has an obligation to provide a full range of retail services to the public. Nevertheless, model simulations of alternative mixes of stores and *CPUs* make apparent the tradeoff between net revenue and the mix of postal stores and *CPUs* in a market area. Presumably, the additional revenue generated by the *CPUs* reflects a gain in value of service to the public that should be considered in evaluating the net position of the public when *CPUs* are substituted for stores.

10.7 Conclusions for Modeling Retail Location

Applying standard location theory to retail activity is complicated because consumer shopping trips do not necessarily originate from the primary residence. Unfortunately, at most, information on consumers is available by place of residence. There is information on places of employment and the location of other types of retail activity that could influence shopping trips, but characteristics of these potential customers or customers of rivals, is not available. This paper demonstrates that the standard theory can be modified to provide implications for total facility revenue as a function of the spatial distribution of residential population, employment, and rival outlets in a market area. Empirical estimates of the revenue equation implied by this theory produce results that agree well with prior expectations based on theory and can be used to calibrate models that produce plausible results.

The model developed can be used to maximize net revenue per unit area, and hence net revenue for any area being analyzed. It provides guidance regarding the spacing, size, revenue, cost, and input requirements for a spatial retail activity. The specific case of *USPS* activity refers to a case in which there are limits on entry because it is a public enterprise. The model could also be used to estimate the optimal characteristics of a spatially competitive industry. In this case, free entry would generally be expected to produce a pattern of that departs from the optimum by having too many facilities and smaller than optimal market areas. The model could be used to measure the welfare effects of such departures from optimal size and spacing that have been noted by Capozza and VanOrder (1980) in the theoretical literature on spatial competition.

If the model is used to analyze a retail network that is already in place, it is helpful to generate the type of net revenue surface, as a function of plant radius shown in the figures. If, as is the case for the postal store network considered explicitly here, the net revenue function is relatively flat, a range of plant size and spacing solutions may produce roughly equivalent results. If there are significant transaction costs associated with altering the network, a departure from net revenue maximization may be warranted.

In the specific case of postal services, there are alternatives to postal stores for serving consumers. The example of *CPUs* is considered here and the manner in which this alternative technology for providing retail access modifies the opportunities for net revenue is demonstrated.

In this case, all estimation and model calibration was accomplished using cross section data from a single year. Clearly, panel data would have allowed adjustment for unobserved heterogeneity that is always present in spatial models. In addition, panel data would allow the research to identify either explicit or implicit quasi experiments in which the pattern of retail service delivery was modified and the consequences could be observed. These empirical techniques are well understood in economics and their ability to inform the model calibration should be clear.

Nevertheless, the final results show that the model is capable of identifying cases in which the spacing but not the size of facilities is optimal. It can distinguish areas where both the size and spacing of facilities is suboptimal. Finally, it can allow users to explore the consequences of implementing an alternative method of retail service delivery and to relate these consequences for alternative market areas with different demand and cost characteristics.

References

Beckmann MJ (1999) Lectures on location theory. Springer-Verlag, Berlin

Beckmann MJ, Thisse JF (1986) The location of production activities. In: Mills E, Nijkamp P (eds) Handbook of regional and urban economics, vol 1. Elsevier Science, New York, pp 21–95

Braid R (2011) Bertrand-Nash mill pricing and the location of firms with different product selections or product varieties. Pap Reg Sci 90(1):197–211

Braid R (2012) The location of firms on intersecting roadways. Ann Reg Sci 50:791–808

Capozza DR, VanOrder R (1980) Unique equilibria, pure profits, and efficiency in location models. Amer Econ Rev 70(5):1046–1053

Davis P (2006) Spatial competition in retail markets: movie theaters. Rand J Econ 37(4):964–982

Launhardt W (1885) Mathematische Begründung der Volkwirtschaftslehre. Engelmann-Verlag, Leipzig. (Translated by Schmidt H and edited and introduced by Creedy J as *Launhardt's Mathematical Principles of Economics*. Edward Elgar, Aldershot, 1993)

Neidercorn JH, Becholod VB (1972) An economic derivation of the law of retail gravitation: a further reply and a reformulation. J Reg Sci 12(1):127–136

Reilly W (1931) The law of retail gravitation. Putnam, New York

Ye M-H, Yezer AM (1992) Location and spatial pricing for public facilities. J Reg Sci 32(2): 143–154

Ye M-H, Yezer AM (1993) Local government and supervised spatial multiplant monopoly. Southern Econ J 59(4):733–748

Chapter 11
Rural School Location and Student Allocation

Ricardo Giesen, Paulo Rocha E Oliveira and Vladimir Marianov

11.1 Introduction

The School Location Problem (*SLP*) is a network design problem aiming at defining where to locate schools. Its solution requires also knowing how many and which students should attend each one of the located schools in order to provide school services to the population of a region, so it involves sizing of the schools, either as a preset parameter or as part of the solution. The determination of what students will attend which school can be seen as a districting problem, which adjusts the boundaries of the zone served by each school within a given school system or network (Caro et al. 2004; Mandujano et al. 2012). In synthesis, the *SLP* involves location, allocation, districting and sizing.

The *SLP* is rarely a straightforward p-median cost minimization problem as each school system has a set of individual peculiarities and constraints that must be taken into account, which vary significantly from case to case due to the different priorities and policies of various school systems around the world. It is further complicated by the fact that different students belong to different grades, which must be in general treated separately, but in the same buildings.

Also, the school location problem in practice is rarely a "pure" cost minimization location problem. Decision makers can belong to various government levels, and the time horizon of the solution can vary significantly from project to project

R. Giesen (✉)
Department of Transport Engineering and Logistics,
Pontificia Universidad Católica de Chile, Santiago, Chile
e-mail: giesen@ing.puc.cl

P. R. E Oliveira
IESE Business School, University of Navarra, Pamplona, Spain
e-mail: paulo@iese.edu

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.), *Applications of Location Analysis,* International Series in Operations Research & Management Science 232, DOI 10.1007/978-3-319-20282-2_11

(from a horizon that ends with the next elections, to long-term horizons considering the benefit of the community). The objectives are also very different (racial or social mix, cost minimization, more/less transportation involved, minimization of the maximum distance to the nearest school, optimization of the balance between school sizes, etc.). Therefore, there are many different formulations.

The problem of allocating children to schools first appeared as an optimization problem triggered by the U.S. policy of desegregating schools applied in 1954 (Franklin and Koenigsberg 1973). Since then, many authors have studied this problem considering different optimization techniques and constraints. Initial approaches used linear programming models (Franklin and Koenigsberg 1973; Sutcliffe et al. 1984). Other authors addressed this problem using integer and mixed integer programming models (Barcelos et al. 2003; Liggett 1973; Diamond and Wright 1987; Church and Murray 1993; Pizzolato and Fraga da Silva 1997; Caro et al. 2004; Pizzolato et al. 2004; Teixeira and Antunes 2008; Araya et al. 2012; Mandujano et al. 2012) or heuristics (Pizzolato 1994). Researchers concerned with the changes of students location and needs over time, addressed it using dynamic optimization models (Antunes and Peeters 2000; Lemberg and Church 2000). Finally, there is literature that determines school location within a logit equilibrium model, considering different types of providers, parental choices based on prices, locations and other issues (Martínez et al. 2011)

Among the distinctive characteristics often found in the formulation of the school location problem are the following:

1. *Parental choice*. In some school systems parents can choose in which school they want to register their children. The rules governing how this choice takes place vary significantly from one location to another and are often the subject of heated political debates. For example, in New Orleans, Louisiana (U.S.), the school system had to be completely planned from scratch after hurricane Katrina, and one of the most heated political discussions in that context was whether the city should have community schools or charter schools being families able to choose sending their children to their preferred school.
2. *School size constraints*. School size constraints can enter the model for a number of reasons. Firstly, because of the lot sizes available for school construction and expansion. Secondly, because of managerial and quality concerns: very small schools are inefficient from a cost perspective, and very large schools can be more efficient from an administrative point of view, but more impersonal. Quality also may depend on the size of the school. Finally, because of pedagogical concerns, the number of students per classroom could be constrained. There are conflicting theories as to whether it is preferable to have standardized schools (all same size) or schools of different sizes to cater each specific community's needs. Depending on which side of the debate the decision makers at hand fall, the corresponding mathematical constraint can be strict or more relaxed.

3. *Number of schools*. Restrictions as to the overall number of schools can be a consequence of school size constraints but can also emerge for other reasons. First, there can be a limit on the locations at which new schools can be built, exogenously setting a maximum number of schools. But also, there can be restrictions about closing schools, thus establishing a minimum number of schools. These decisions are often motivated by political (rather than cost-related) reasons.
4. *Total system capacity*. School location studies are sometimes initiated due to expected changes in demand (increase or decrease) due to demographic changes or migrations. In occasions, school location studies are triggered by required overhauls of scholar systems due to new policies regarding quality, the goal of attracting more students from the private school system to the public schools, and so on.
5. *Shift allocation*. Some school systems operate in single shifts, in that case there is no need to allocate students to shifts, but in other cases school systems operate in two or even three shifts. In the latter case, school location problems must also decide which grade levels to operate in what shift and which students will attend school in each shift. In some cases, a policy is established that assigns older or younger students together to same shift, but in other cases this could be a decision variable. In this case cost savings can be achieved by a better utilization of the infrastructure and the transportation system.
6. *Transportation*. In many places it is common for local governments to be in charge of meeting public school transportation needs. In these cases, the cost of optimal location of schools must also take into account the transportation costs of students to and from schools. This can bring significant complexity to the problem, and often requires two separate problems being solved: a long-term problem with approximate transportation costs, and a short-term problem that defines the bus routes. Additional complexities that come up in these situations include defining rules of access to transportation, which can be as simple as a maximum distance students can walk; or as complex as specifying times of day when students can walk, dangerous routes that cannot be covered by walking (in our experience this can be due to a variety of reasons, from high traffic to wild animals, including areas with high crime rates), and times of the year when the route can be walked (typically due to weather conditions). Given these constraints the set of bus stops and assignment of students to each one should be defined.
7. *Staffing levels*. In most cases, it is beneficial to predefine student/teacher ratios, administrative staff per school, and so on.
8. *School types*. Some systems have pre-defined types, e.g., Kindergarten, Basic, Middle School, High School. Some countries have systems where the grade levels of each school can differ within the same system.
9. *School districting*. Some school systems have very strict rules about school districting. For instance, all students must attend the nearest school to their home. Or the rules can include compactness and contiguity constraints, which are desirable, in particular, when parents decide what schools are their children

to attend, based on distance to the school. Rules can also include maximum distances from the residence of each student to the closest school.

10. *Student profile balancing*. The issue of racial or social balancing and mixed student bodies was a particularly important issue in the United States after school desegregation. In most of the world it is not relevant as such, but other types of student profile balancing can be important when it comes to allocating resources to students with special needs. These special needs can be learning disabilities or language difficulties, which can be particularly important in school districts with large immigrant or aborigine populations.
11. *Rurality*. In the case of the school location problem for rural areas, sparse population further complicates the solutions, as it favors the location of many schools, so avoiding long travel distances, but it also results in small size, less efficient schools. Drop-out ratios also tend to be higher in rural schools, because of quality and travel factors.

Five properties have been declared as desirable in the literature for school location:

- Students' grades should be modeled separately. As Caro et al. (2004) pointed out, ignoring students' grades could only be reasonable if all schools have the same grade structure and if the grade blend were homogeneous among bus stops. In rural areas this is hardly the case.
- Schools should consider Continuous Serving Zones (*CSZ*). As pointed out by Caro et al. (2004) and then restated by Teixeira and Antunes (2008) in a public facility planning context, users from neighboring centers should be prevented to be assigned to different facilities, because it makes the solution harder to understand and accept by users. This solution property was defined as *CSZ*.
- The model must allow schools' closure, opening, and expansion. As the main objective is to propose more efficient schools' networks, closing inefficient schools and expanding others to gain efficiency has to be allowed. This assumes the presence of economies of scale in schools costs.
- Schools' instruction costs should be included explicitly. The reason for including them explicitly is that administrative personnel and teachers' wages (instruction costs) are the major component of educational costs (White and Tweeten 1973).
- Multi-graded classrooms schools should be a possibility in rural areas. As previously pointed out, multi-graded classrooms schools are sometimes the most efficient solution in rural areas.

Table 11.1 shows a summary of how previous researchers have addressed these five desirable properties.

To the best of our knowledge, the only model that simultaneously considers all of these five desirable characteristics is the one proposed by Mandujano et al. (2012). In the next section, a case study is presented in which this model was used on instances with data from Brazilian rural-municipalities.

In this chapter, we address two different experiences of school location problems. The first case corresponds to Barao de Grajaú, in the northeast region of Brazil, and includes a model that prescribes opening, closing and size of schools, allocates students to them, and designs a transportation system for students, as in this case, the authority in charge also deals with transportation. The second application, covering

Table 11.1 Desirable properties of the school location and sizing problem

Reference	Grades modeled	Contiguity constraint	Closing and expanding schools	Instruction costs	Multi-grade schools
Franklin and Koenigsberg (1973)	Yes	–	–	–	–
Liggett (1973)	–	–	–	–	–
Church and Murray (1993)	–	–	–	–	–
Antunes and Peeters (2000)	–	–	Yes	Yes	–
Lemberg and Church (2000)	–	–	Yes	–	–
Pizzolato et al. (2004)	–	–	Yes	–	–
Caro et al. (2004)	Yes	Yes	–	–	–
Teixeira and Antunes (2008)	Yes	Yes	–	–	–
Araya et al. (2012)	Yes	–	Yes	Yes	Yes
Mandujano et al. (2012)	Yes	Yes	Yes	Yes	Yes

(–) No

the whole rural school system in the country of Chile, uses a very similar model, which minimizes total cost and students' travel distance. In this case, the transportation is not included, and the model was intended first as a study not necessarily to be implemented, but as an evaluation tool for the country's rural school system. However, after a very large earthquake in 2010, the model, with a few changes, was applied to municipalities as a tool for finding the best solutions for school reconstruction.

11.2 The Brazilian Application

In this section we present a case study using data from different rural municipalities in Brazil, in which we compared the current situation versus an optimized proposal. These proposals were obtained using the method developed by Mandujano et al. (2012), which is presented in the next subsection. After explaining the optimization model, we show some results of this methodology applied to some rural municipalities in Brazil. Then we discuss the main implementation difficulties, and some thought on how to overcome these complications.

11.2.1 Optimization Approach for the School Location and Size Problem

The main objectives of the School Location (and Sizing) Problem (*SLSP*) are deciding which schools among the current ones should be closed, expanded or maintained, and which students based on the bus stops wherein they board should be assigned to each school. The objective is to minimize the total cost of the school

Table 11.2 Set of parameters of the school location and sizing problem

Parameter	Description
$stud_{b,\,g}$	Number of students in educational grade g at bus stop b
$dist_{b\ell}$	Road distance between bus stop b and the possible school location ℓ
$mccap_e$	Multi-graded classroom capacity for educational level e
$sccap_g$	Single-graded classroom capacity for grade g
$mnclass_s$	Minimum number of classrooms for a school type s
$mxclass_s$	Maximum number of classrooms for a school type s
$shifts_s$	Number of operating shifts for school type s
$class_\ell$	Current number of classrooms at location ℓ

network, which includes: School Administrative Costs (*SAC*), Classrooms' Operating Costs (*COC*) that correspond to teachers' wages, and Classrooms' Construction Costs (*CCC*), which account for investments needed for expansions. In addition, a Classroom Exceeded Capacity Cost (*CEC*) is considered in order to include flexibility in terms of classroom capacity. Using these *CEC*, opening a new class in cases in which only a few students exceed the limit on the number of student in a class. Furthermore, a Student Distance Cost (*SDC*) is considered as a proxy for the student transportation cost.

The *SLSP* model decides for each current school location, among the previously defined set of school types, if any of those types should be located there. If the type of school to locate is larger than the current school, then new classrooms will be needed. The definition of school type includes the school size, which is the maximum number of classrooms, also whether multi-graded (*MS*) or single-graded (*SS*) classrooms are operated, and the number of operating shifts. By modifying the school assignment of the students in a particular bus stop, economies of scale can be obtained in terms of school administrative and class operating costs, *SAC* and *COC* respectively, since schools and classroom can be operated near to full capacity.

The mathematical programming model developed in Mandujano et al. (2012) for solving the *SLSP* is presented below. Indices b and ℓ represent the current location of bus stops and schools respectively. The set *BL* represent all the possible assignable pairs within bus stops and schools. Indices e and g represent school levels and grades correspondingly. School levels are group of grades that can be taught together in multi-graded classrooms, this relationship is represented by set $GE[e]$. Index s represents the school type. The decision variable $X_{s\ell}$ equals 1 if a school type s is located at ℓ, and 0 otherwise. The decision variable $C_{s\ell}$ is the number of new classrooms to be built for school type s at location ℓ. The decision variables $ZM_{e\ell}$ and $ZS_{g\ell}$ are the number of multi-graded and single-graded classrooms used daily for level e and grade g at location ℓ. The decision variables $WM_{e\ell}$ and $WS_{g\ell}$ are the proportion of exceeded capacity on multi-graded and single-graded classrooms respectively. The decision variables $YM_{b\ell}$ and $YS_{b\ell}$ equals 1 if bus stop b is allocated to a school located at ℓ, and 0 otherwise. Table 11.2 presents the set of parameters used in the model.

The optimization problem can then be written as follows:

$$\begin{aligned}\min & \sum_{\ell\in L}\sum_{s\in S}(SAC_s X_{s\ell} + CCC\ C_{s\ell}) + \\ &+ \sum_{e\in E}\sum_{\ell\in L}(MCOC_e ZM_{e\ell} + MCEC_e WM_{e\ell}) + \\ &+ \sum_{g\in G}\sum_{\ell\in L}(SCOC_g ZS_{g\ell} + SCEC_g WS_{g\ell}) + \\ &+ \sum_{(b,\ell)\in BL}\sum_{g\in G}[SDC\ stud_{bg}\ dist_{b\ell}(YS_{b\ell} + YM_{b\ell})] \end{aligned} \tag{11.1}$$

$$\text{s.t.} \sum_{s\in S} X_{s\ell} \leq 1\ \forall\ \ell \in L \tag{11.2}$$

$$\sum_{\ell:(b,\ell)\in BL}(YM_{b\ell} + YS_{b\ell}) = 1\ \forall b \in B \tag{11.3}$$

$$YM_{b\ell} \leq \sum_{s\in MS} X_{s\ell}\ \forall\ (b,\ell) \in BL \tag{11.4}$$

$$YS_{b\ell} \leq \sum_{s\in SS} X_{s\ell}\ \forall\ (b,\ell) \in BL \tag{11.5}$$

$$\sum_{s\in MS} shifts_s\ mnclass_s\ X_{s\ell} \leq \sum_{e\in E} ZM_{e\ell}\ \forall\ \ell \in L \tag{11.6}$$

$$\sum_{s\in SS} shifts_s\ mnclass_s\ X_{s\ell} \leq \sum_{g\in G} ZS_{g\ell}\ \forall\ \ell \in L \tag{11.7}$$

$$\sum_{e\in E} ZM_{e\ell} \leq \sum_{s\in MS} shifts_s\ mxclass_s\ \ X_{s\ell}\ \forall\ \ell \in L \tag{11.8}$$

$$\sum_{g\in G} ZS_{g\ell} \leq \sum_{s\in SS} shifts_s\ mxclass_s\ X_{s\ell}\ \forall\ \ell \in L \tag{11.9}$$

$$\sum_{b:(b,\ell)\in BL}\ \sum_{g\in GE[e]} stud_{bg}\ YM_{b\ell} \leq (ZM_{e\ell} + WM_{e\ell})mccap_e\ \forall\ E \in e,\ \ell \in L \tag{11.10}$$

$$\sum_{b:(b,\ell)\in BL} stud_{bg} YS_{b\ell} \leq (ZS_{g\ell} + WS_{g\ell})sccap_g\ \forall\ G \in g,\ \ell \in L \tag{11.11}$$

$$WM_{e\ell} \leq ZM_{e\ell}\ e \in E,\ \ell \in L \tag{11.12}$$

$$WS_{g\ell} \leq ZS_{g\ell}\ \forall g \in G,\ \ell \in L \tag{11.13}$$

$$C_{s\ell} \leq mxclass_s\ X_{s\ell}\ \forall\ s \in S,\ \ell \in L \tag{11.14}$$

$$\sum_{e \in E} ZM_{e\ell} \leq \sum_{s \in MS} shifts_s(C_{s\ell} + class_\ell X_{s\ell}) \; \forall \; \ell \in L \tag{11.15}$$

$$\sum_{g \in S} ZS_{g\ell} \leq \sum_{s \in SS} shifts_s(C_{s\ell} + class_\ell X_{s\ell}) \; \forall \; \ell \in L \tag{11.16}$$

$$YM_{b\ell} \leq YM_{c\ell} \; \forall (b,\ell) \in BL, \quad c \in PBL[b,\ell] \tag{11.17}$$

$$YS_{b\ell} \leq YS_{c\ell} \; \forall (b,\ell) \in BL, \quad c \in PBL[b,\ell] \tag{11.18}$$

$$X_{s\ell} \leq YM_{b\ell} \; \forall \; \ell \in L, \quad s \in MS, \quad b \in LZ[\ell] \tag{11.19}$$

$$X_{s\ell} \leq YS_{b\ell} \; \forall \; \ell \in L, \quad s \in SS, \quad b \in LZ[\ell] \tag{11.20}$$

$$X_{s\ell} \in \{0,1\} \; \forall \; s \in S, \quad \ell \in L \tag{11.21}$$

$$C_{s\ell} \in \mathbb{Z}^+ \cup \{0\} \; \forall \; s \in S, \; \ell l \in L \tag{11.22}$$

$$YM_{b\ell}, YS_{b\ell} \in \{0,1\} \; \forall \; (b,\ell) \in BL \tag{11.23}$$

$$ZM_{e\ell} \in \mathbb{Z}^+ \cup \{0\} \; \forall \; e \in E, \quad \ell \in L \tag{11.24}$$

$$ZS_{g\ell} \in \mathbb{Z}^+ \cup \{0\} \; \forall \; g \in G, \quad \ell \in L \tag{11.25}$$

$$WM_{e\ell} \in [0,1] \; \forall \; e \in E, \; \ell \in L \tag{11.26}$$

$$WS_{g\ell} \in [0,1] \; \forall \; g \in G, \quad \ell \in L \tag{11.27}$$

Constraints (2) to (5) are variations of the typical location constraints for the *p*-median problem. Constraints (2) and (21) ensure that no more than one type of school is located at each current location. For the allocation variables, constraints (3) and (23) ensure that all bus stops are allocated to a school (demand covering). As the allocation can be to multi-graded or single-graded schools, constraints (4) and (5) ensure that the allocation (multi-graded or single-graded) corresponds to the type of school located.

Constraints (6) to (16) define the number of operative classrooms (i.e., the number of teachers needed) at every located school. Because there is more than one shift for school daily operation, operative classrooms differ from physical ones. Constraints (6) and (7) define the minimum number of operative classrooms depending on the school type. If the school has multi-graded classrooms (*MS*) the comparison has to be in terms of levels and if it has single-graded classrooms (*SS*) in terms of grades. Constraints (8) and (9) define the maximum number of operative classrooms. Constraints (10), (11), (12) and (13) balance the number of operative classrooms for each level/grade with the allocated students at each level/grade. Constraints (14), (15) and (16) determine the number of new classrooms to build if necessary.

Constraints (17) to (20) help the solution to be better understood by users and decision makers. The set $PBL[b, \ell]$ define all bus stops that may be in the path between bus stop b and location ℓ. As was pointed out by Teixeira and Antunes (2008), if all bus stops in the path between bus stop b and location ℓ are also allocated to school at ℓ, the solution makes more sense for users (students and parents) and decision

Table 11.3 Comparison of total costs of current versus proposed solutions

Municipality	Year	Current total costs USD	Proposal total costs USD	Variation (%)
1	2012	3,521,350	2,828,450	– 19.7
2	2014	6,606,273	5,547,899	– 16.0
3	2014	28,854,974	34,012,037	17.9
4	2013	37,123,346	39,088,171	5.3
5	2014	17,682,912	19,391,104	9.7
6	2013	12,402,769	11,607,210	– 6.4
7	2013	8,600,210	8,467,699	– 1.5
8	2013	10,428,987	8,487,340	– 18.6
9	2013	3,411,440	2,996,249	– 12.2
10	2014	6,771,341	5,888,637	– 13.0

makers (educational board). The set $LZ[\ell]$ define a minimum serving zone for every current location, and then all bus stops within this zone have to be allocated to location ℓ if the school at ℓ is not closed. Finally, constraints (21) to (27) define the nature of the decision variables involved.

This model for the *SLSP* had been applied, together with a heuristic to solve the School Bus Routing and Shifts' Programming Problem (*SBRSPP*) (Mandujano et al. 2012), to instances from different municipalities in Brazil showing promising results. In the next subsection the result of these case studies is discussed.

11.2.2 Analysis of Results

Table 11.3 summarizes results of proposals in ten municipalities in different regions of Brazil, based on the results of the *SLSP* model presented in the previous subsection combined with the *SBRSPP* model. As shown in most municipalities significant gains are obtained. Gains reach up to nearly 20 %. However these savings vary widely among municipalities with some of them without savings. Moreover, in some instances the total costs of the proposal are higher than the current total costs. This can be explained since it is common that municipalities operate under level of service standards that are very different to those proposed by the model, e.g. more students per classroom, bus routes exceeding its capacity, students traveling more time than the maximum allowed. In those cases the process of collecting data to run the model highlighted those inefficiencies. Thus the proposed solution might have not reduced total costs, but for sure it improved the level of service.

Table 11.4 presents a more detailed comparison of instances 10 and 3 that permits showing the trade off between the level of service and efficiency in design of a school system in a municipality. For example in the instance 10, an increase in

Table 11.4 Comparison of current versus proposed solutions

Municipality	10		3	
Number of schools	20		183	
Number of students	3476		14,939	
	Current	Proposed	Current	Proposed
Avg number of students per class	20.3	20.3	20.9	18.4
Avg distance to school [Km]	0.6	0.7	1.4	1.2
Avg distance traveled [Km]	5.7	5.5	10.9	9.7
Educational costs [USD]	6,303,632	5,218,096	25,609,616	27,424,236
Transportation costs [USD]	467,709	670,541	3,245,358	6,587,801
Total costs [USD]	6,771,341	5,888,637	28,854,974	34,012,037

transportation costs allows an important reduction in educational costs while keeping the same number of students per class and slightly reducing the average distance traveled by students. On the other hand, in the instance 3 all the performance indicators associated with level of service to students increased, but transportation and educational costs increased, since in many cases the maximum distances were not respected, and in other cases the maximum number of students per class was not enforced.

Even though from a technical standpoint the results are encouraging, there are many implementation difficulties to translate a sound proposal into practice. These difficulties and some ideas to overcome them are discussed in the next subsections.

11.2.3 Implementation Difficulties

Rarely is a school system planned from scratch. Therefore, the idea of planning an optimal school network based on a mathematical model is a theoretical abstraction. In typical applications, models are used to analyze the current system and propose changes. In this sense, the optimal location problem can be used as benchmark to identify opportunities for improving the status quo.

One example of the successful use of school planning model is the work done in developing countries by *UNESCO-IIPE* (International Institute for Educational Planning). They offer education on "school microplanning," which includes school mapping and tools for deciding school location. While it is not very sophisticated in terms of optimization techniques, it is an example of a simple tool that is accessible to decision-makers that can improve the status quo.

In our experience working with more developed school systems, where mapping was not an issue, we have seen these models not being implemented for a number of reasons. Indeed, the authors have never seen the implementation of the optimal solution of an operational research method without modifications. Typical outcomes include a meeting with decision makers in which the report is met with varying

levels of acceptance and one or two ideas are taken up for further discussion. The study is typically requested by the department of education, and it implementation often involves other departments (e.g., city planning, transportation, construction). Furthermore, there are a number of additional stakeholders (politicians, teachers, parents, etc.) each of which has their own agenda.

In mathematical terms, this means that there are too many variables to take into account. So the way to proceed is to simplify the problem and make the required adjustments later. These adjustments, intended to overcome the barriers can sometimes hinder implementation overall. It is essential for the operations research modelers who want to see their solutions implemented to talk closely to the decision makers to understand how these barriers may play out.

11.2.4 Overcoming Difficulties

In this subsection we discuss some key difficulties faced when implementing optimization based proposals and some ideas to overcome them.

1. *Transition plan*. The outcome of the optimization model can be desirable, but it is often not clear how to go from the current state of things to the optimized proposal. Two aspects of a transition plan can help to make implementation more likely. First, it is important to consider a financial model. Construction is often required; these costs are typically not covered by the operational budget. Therefore, it is fundamental to take into account how should municipalities pay for these costs. Secondly, a logistics plan: if all the necessary construction cannot be done during school holidays, which school will the students attend and how will they get there during the transition phase? Decision-makers often do not know how to answer these questions, and in absence of an answer implementation is stopped.
2. *Talking to communities*. Implementation often involves closing schools. This can require working with the communities, justifying the new schools. Politicians may not want to close or open schools in certain locations for reasons that are completely unrelated to cost optimization.
3. *Dealing with unions*. Teachers unions, when present, could object to anything that changes the *status quo* in schools. In one situation we encountered, a very strong union did not want teachers commuting long distances. As a consequence, the students from the suburbs were being bused to schools in a central location in order to keep the teachers' commute within the acceptable level.
4. *Political Cycles*. Decision makers think in "election (e.g. 4-year) cycles." Depending on the political structure, the decision is made by a major who does not think beyond the next election. The benefits of the optimized solution are often not seen in such a short term.
5. *Educators in the way of the decision*. Because of the biases in their training, educators typically do not understand operations research models and methods

or their value. Rather than trying to engage with, and understand the models, the reaction we most often encounter is a defense of the *status quo* based on "pedagogical reasons."

6. *Renegotiations with suppliers*. Changing schools to different locations can require renegotiating contracts with a number of external service providers, such as caterers, security, and transportation. These negotiations can be politically difficult, particularly when there are long-term contracts in place.
7. *Too many decision makers*. Modelers need to understand the various decision makers involved and understand their role in the decision making process. School location decisions often go beyond the education authorities. Thus, to guarantee implementation, it is essential to understand the political environment.

11.3 Selected Experiences in Chile: Rural School Location and Sizing

Answering to a requirement of the Ministry of Education, an analysis was done of the effects of closing some rural schools, opening new schools and resizing existing schools, over the total cost of the system, as well as over the distances traveled by students to go to their closest school, in the whole country of Chile (15 regions). Since most of rural schools are public, the project considered only this type of schools. The results are reported in detail in Araya et al. (2012). Later, after the 8.8 magnitude earthquake that affected several regions of the country in 2010, the model was partially used, together with a heuristic, to provide the Ministry of Education with a tool for deciding what schools to reconstruct.

The Chilean school system is composed by public and private schools. The private sector, in turn, includes schools that are financed strictly by student fees, and subsidized schools. The last ones receive mostly middle and low-income students and, in some cases, the subsidy is complemented by a small monthly fee paid by students[1]. Around a 9.5 % of school students in Chile live in rural areas. School education in rural areas is provided by some 4000 schools, both public (managed by municipalities) and subsidized, many of them small and run by one teacher who teaches in a multigrade mode, i.e., to students belonging to several (up to four) grades simultaneously. In multigrade schools, when the number of students is very small, the quality of education can be good (as in home schooling). However, as the student number increases, the quality deteriorates quickly, which is the case in most of them in Chile. In regular schools, there are in all 13 grades: preschool, basic primary school to the eighth grade, and four high school grades. In general, the students must travel long distances to attend the school, which increases the drop-out rate.

[1] By the time this chapter is being written, the existence of these schools is being discussed.

Addressing these problems is complex and includes different actions, including relocating and merging schools, as well as integrating transportation networks into the analysis and the solution, as it was shown in the Brazilian example. Relocating and merging schools was possible in Chile to an extent. However, unfortunately, as opposed to the Brazilian case, the authority in charge of education in Chile does not have the freedom to establish its own transportation system, or hiring third party transportation, so the problem has to be addressed considering that students use public transportation or their own resources, e.g., cars or hired minivans. This is the main difference between the Brazilian and Chilean cases we discuss in this chapter.

The demand for schools comes from "rural entities", which are small villages or hamlets with populations that do not exceed 2000 people. There are around 26,000 of these entities. All the students from an entity must be assigned to the same school. Entities close enough to towns with urban schools are not considered, as students prefer attending urban schools when these are close enough. In our model, the different types of schools include existing schools as they are, new schools with different compositions of grades (preschool, basic, high, preschool-basic; basic-high, all-grade schools) and, existing schools that are to be closed as a consequence of the optimization. In existing schools, the number of classrooms can change according to the allocated students. Closures are limited in terms of percentage, as it is not politically viable to have too many closures in a region. Budget is also limited. The costs and parameters are very similar to the Brazilian case, as well as the model (except for the transportation component).

Almost 300,000 students were considered, 26,000 rural entities, more than 4000 existing schools and each entity could be a candidate location for a new school. With these numbers, the problem became almost intractable from the point of view of obtaining optimal solutions. Furthermore, many rural entities were too close to each other, which would result in many solutions with the same objective value that unnecessarily delay obtaining an optimal solution and multiply the number of alternate optima.

In order to reduce the number of potential locations of new schools, we first solved a Location Set Covering Problem (*LSCP*, Toregas et al. 1971) formulated as follows:

$$\text{Min} \sum_j C_j$$

$$\text{s.t.} \sum_{j \in N_i} E_j + \sum_{j \in N_i} C_j \geq 1 \quad \forall\, i,$$

where C_j is 1 if j becomes a candidate location; E_j is 1 if there is an existing school with at least 100 students at j, and N_i is the set of all sites within 10 km of location i. Note that the solution to this problem minimizes the number of candidate locations, so that all entities i have either a school or a candidate location within 10 km. The number of candidates is reduced in almost 90 %. Furthermore, the administrative division of the country enforces students from one region not attending schools

in a different region. Thus, the divisions between regions are a natural boundary for dividing and conquering this very large problem. Still, the largest regions have associated problems with millions of variables, out of which, hundreds of thousands are binary. We divided these regions in two or three parts, using natural boundaries such as rivers.

For the largest regions, we allowed closure of up to 40 schools. For the remaining regions, closures are bounded above by 20 or a 20 % of the existing schools. We also bounded the distance to be traveled by students to 30 km, except in the least populated, extreme regions in the far south (fjords) and far north (desert) to 50 km.

A model was used that is very similar to that of the Brazilian case. Finding a good solution required, in cases, a heuristic, as the optimal solution was not found within 3 h, even using an integrality gap of 5 %. For more details, see Araya et al. (2012).

A baseline solution was found, assuming all students assisting to their nearest schools. The results of the new model, compared to the baseline are shown in Table 11.5. We solve the problem in two versions: using the travel distance limit of 50 km in extreme regions of the country, and relaxing this constraint for these regions only, keeping, in both versions, the distance limit of 30 km in the remaining regions. Note that even though the number of schools is reduced in the new situations, the average travel distance is significantly reduced, as well as the maximum distance, when it is constrained in the extreme regions. Also, the cost decreases in 17 % when the total distance is constrained in the extreme regions, and in 18 % in the unconstrained case.

It is interesting to remark what were the practical issues that had to be addressed during the project. In the first place, the data required for carrying the project was scattered among several different divisions at the Ministry of Education. There are divisions in charge of financial aspects, quality aspects, political issues, and so on, and at the time of the project there was no sharing of the data among them, so the research group acted as a consolidation center for the information. Secondly, the existence of multi-grade schools had to be accepted, even though when such schools increase in size, the quality of the educational results decreases significantly. On the other hand, small multi-grade schools seem to provide a very good education, similar to home schooling. Thirdly, the investment money and the operational money come from different divisions, and each one seeks its own optimum. The operational expenses decrease importantly if the solutions are implemented; however, there is investment to be done, which makes the corresponding divisions oppose to any change.

Finally, closing a school is extremely unpopular, and a strong opposition appears whenever a school is in danger of being closed. This makes the implementation of the results of the model very difficult.

However, with a few changes, the model was used in practice after the February 27, 2010, 8.8 magnitude earthquake that affected a large area in Chile. After this event, approximately 2600 urban and rural schools were lightly damaged; 2200 schools suffered moderate damage; 210 severe damage and 90 were destroyed.

Table 11.5 Comparison of the baseline and the proposed solutions

		Baseline	Solution (50 km)	Solution (unconstrained distance)
Number of schools	Existing	4165	3851	3851
	New		235	173
	Closed		314	314
Reduced classrooms			471	522
Distance traveled by student [km]	Average	21.9	8.5	13.7
	Maximum	329.0	50.0	371.2
	Minimum	0	0	0
Students per school (Existing)	Average	76	75	80
	Maximum	2059	1658	1658
	Minimum	0	0	0
Students per school (New)	Average		64	78
	Maximum		565	565
	Minimum		1	1
Average unused capacity		28 %	30 %	15 %
Students per teacher		15.6	15.6	18.1
Total cost MMUS$		415.8	352.6	348.8

This time, the model was applied to both urban and rural schools, not as a prescriptive model, but as a helping tool for making decisions about which damaged schools to reconstruct; what schools to move to a different place, or whether pairs of schools were to be closed, and a new school opened in a new place instead. The objective of the model (cost, average distance from school to students' home) was complemented with a quality objective. In this case, a proxy for quality was obtained from standardized tests that are being applied to all schools in the country. The quality depends mainly on the personnel of each school (teachers, director). With these changes, the model was used by the Ministry of Education to negotiate with the Municipalities in charge of the schools the amount of investment to be assigned to each damaged school (or a replacement school in a different place), depending on cost, travel distance and quality. In some cases, the use of the model allowed the Ministry of Education to convince both Municipalities and parents of replacing the damaged schools by new schools at a different location.

The application of the model in this case, reduced the operational cost per student by US $ 27, and the travel distances in 6 %, while increasing the quality, measured by an average increase in 3.4 points in the standardized national tests, for the considered schools. Additionally, there were savings of US $ 22 million in damage schools repairs.

11.4 Conclusions

In this chapter, we presented a particular application of location models in public policy, namely the problem of school sizing and location and student assignment to schools in rural areas, discussing the main complexities and differences with common location problems. We discuss in detail a methodology for optimizing the number and location of public schools in rural areas which had been applied to different rural-cities in Brazil, and briefly describe a similar problem, but covering a whole country, in Chile. This School Location and Sizing Problem (*SLSP*) includes properties from school location models, school districting models and also peculiarities of rural schools allowing multi-graded classrooms in schools. The results in different regions in Brazil show gains in efficiency up to nearly 20 %, even though the benefits had a lot of variability, which can be explained since the level of service provided by current solutions vary widely among different cases and the level of service of the proposed solution respected standards such as maximum travel time per student. In the Chilean case, the average distance travelled by students is decreased in a 37 % or a 61 %, depending on the distance constraints in extreme regions. The total cost of the system is also significantly reduced.

In addition, a discussion on the main implementation difficulties and some ideas on how to overcome them were included. Among the key recommendation for a successful implementation of optimization-based proposal are: to provide a good transition plan; to implement a communication plan to persuade different communities about the benefits; to establish an agreement with unions and educators so that they become proactive and supporters of the plan; and understanding the window of opportunity in the political cycle, so that the plan could be implemented during the period of the champion of the project or with the support of different political groups, so that it can be continued after new authorities take control.

In other cases, as after the Chilean earthquake, circumstances help the authorities in the decision of using operations research tools.

We think that a critical challenge for operations research practitioners to have an impact on school locations problems is to carefully understand the priorities of all decision makers involved. In these politically motivated settings, the criteria for making the decision often evades what operations research modelers consider to be rational cost minimization. Yet, often these alternative criteria can be brought into the model in one way or another.

Therefore, we think that there are essentially two ways for operations researchers to have more impact in practice as far as the school location model goes. In the short term, implementation is more likely to the extent that models are customized to each decision setting. In the long term, it is in the interest of all involved parties that operations research practitioners engage with professionals and politicians involved in school planning in discussions about the relationship between school efficiency and school quality, particularly in what has relation with size of the schools and their cost. Thus, we see a big opportunity in exploring the relationship between the operational efficiency of school systems and education quality, bringing efficient

operations to the political agenda, and finding effective ways to convey the important of efficiency to the general public.

Acknowledgements Giesen and Oliveira gratefully acknowledge partial support by the Instituto Alfa e Beto, Brazil. Marianov gratefully acknowledges partial support by the Institute Complex Engineering Systems through Grants ICM MIDEPLAN P-05-004-F and CONICYT FBO16.

References

Antunes A, Peeters D (2000) A dynamic optimization model for school network planning. Socio-Econ Plan Sci 34(2):101–120

Araya F, Dell R, Donoso P, Marianov V, Martínez F, Weintraub A (2012) Optimizing location and size of rural schools in Chile. Int Transa Oper Res 19(5):695–710

Barcelos FB, Pizzolato ND, Lorena AN (2003) Localização de escolas do ensino fundamental com modelos capacitado e não-capacitado: caso de Vitória/es. Pesqui Operacion 24(1):133–149

Caro F, Shirabe T, Guignard M, Weintraub A (2004) School redistricting: embedding GIS tools with integer programming. J Oper Res Soc 55(8):836–849

Church RL, Murray AT (1993) Modeling school utilization and consolidation. J Urban Plann Dev 119:23–38

Diamond JT, Wright JR (1987) Multiobjective analysis of public school consolidation. J Urban Plann Dev 113(1):1–18

Franklin AD, Koenigsberg E (1973) Computed school assignments in a large district. Oper Res 21(2):413–426

Lemberg DS, Church RL (2000) The school boundary stability problem over time. Socio-Econ Plan Sci 34(3):159–176

Liggett RS (1973) The application of an implicit enumeration algorithm to the school desegregation problem. Manage Sci 20(2):159–168

Mandujano P, Giesen R, Ferrer J-C (2012) An optimization model for location of schools and student transportation in rural areas. Transp Res Rec 2283:74–80

Martínez F, Tamblay L, Weintraub A (2011) School locations and vacancies: a constrained logit equilibrium model. Environ Plan A 43:1853–1874

Pizzolato ND (1994) A heuristic for large-size *p*-median location problems with application to school location. Ann Oper Res 50(1):473–485

Pizzolato ND, Fraga da Silva HB (1997) The location of public schools: evaluation of practical experiences. Int Trans Oper Res 4(1):13–22

Pizzolato ND, Barcelos FB, Lorena NAL (2004) School location methodology in urban areas of developing countries. Int Trans Oper Res 11(6):667–681

Sutcliffe C, Boardman J, Cheshire P (1984) Goal programming and allocating children to secondary schools in Reading. J Oper Res Soc 35(8):719–730

Teixeira JC, Antunes AP (2008) A hierarchical location model for public facility planning. Eur J Oper Res 185(1):92–104

Toregas C, Swain R, ReVelle C, Bergman L (1971) The location of emergency service facilities. Oper Res 19(6):1363–1373

White F, Tweeten L (1973) Optimal school district size emphasizing rural areas. Amer J Agr Econ 55(1):45–53

Part III
Enforcement and First Responders

Chapter 12
Fire Station Siting

Alan T. Murray

12.1 Introduction

Location analysis and modeling has long been relied upon to support a wide array of public facility siting efforts (ReVelle 1987; Marianov and Serra 2002; Church and Murray 2009). As noted in Daskin and Murray (2012), public sector policy and planning often involves multiple constituents, various competing/conflicting objectives, politically driven decision making and constraints on available resources, making policy and planning problems "wicked" (see Liebman 1976). Addressing issues and making good decisions therefore requires a delicate balance of skill and finesse that is supported by data and quantitative methodologies (Brill 1979). In this chapter the focus is on a particular aspect of public sector policy and planning, namely the siting of facilities. Public facilities are necessary to provide services to people in communities, towns and cities, and often this must be done with access, accessibility, equity, fairness, fiscal prudence, etc., in mind. Examples of public facilities along these lines include post offices, parks, schools, police stations, firehouses, etc. Public financing and support of these services continues to be the norm because of their widely held necessity and value.

The emphasis in many modeling and siting approaches is the design of a new system of facilities. Often ignored or overlooked is the necessity of systematic re-evaluation and assessment of existing facilities in relation to anticipated future conditions. Public facilities in particular reflect evolving needs and incremental changes, so an existing system of facilities is a byproduct of additions and closings over time. When new facilities are needed to complement existing facilities, it is essential that systematic and comprehensive evaluation take place (Schilling et al. 1980; Francis et al. 1992; Badri et al. 1998; Murray 2010, 2013). Most urban areas have an existing system, referred to as a brownfield system (in contrast to the Greenfield case where no existing facilities exist). This is noteworthy because an

A. T. Murray (✉)
Center for Spatial Analytics and Geocomputation, College of Computing and Informatics, School of Public Health, Drexel University, Philadelphia, PA, 19104, USA
e-mail: amurray@drexel.edu

H. A. Eiselt, V. Marianov (eds.), *Applications of Location Analysis,* International Series in Operations Research & Management Science 232, DOI 10.1007/978-3-319-20282-2_12

existing system may suggest a prevailing idea of expansion, yet broader integrated planning may in fact be warranted that involves closing or relocating one or more existing facilities. Doing so may well turn out to be more cost effective and efficient over the long term, but cannot be attained if it is not sought or explored through planning and analysis.

The purpose of this chapter is to address the need for fire and emergency response in a region experiencing changing social, economic, residential, employment, etc. conditions. In particular, the goal is to design a fire station response system in an urban area. This need arises due to urban growth, development and/or decline, where an area already has a number of existing fire stations and must plan for anticipated future conditions. Thus, new facilities combined with a re-evaluation of existing services not only makes sense, it is a prerequisite for sound public management of scare resources targeted to provide essential services.

A number of things make the problem of locating fire stations interesting, special and different from the broader context of facility location modeling. First, fire stations are very much response oriented. It is extremely important that emergency personnel arrive to a call for service within a stipulated response standard, typically something in the range of 5–9 min. The reason for this is that quick response means that lives and property can be saved (Murray and Tong 2009; Murray 2013). Second, it is recognized that rapid response may not always be possible, but should occur most of the time. Third, there are real fiscal implications for building, equipping and manning a fire station. The costs per year to man a fire station are significant, essentially that of the fixed costs of establishing a station (Plane and Hendrick 1977; Badri et al. 1998; Murray et al. 2012). Care should therefore be taken to provide sufficient services, now and over time, without bankrupting or financially crippling a community. Finally, many important advances have been made in location modeling to accommodate the richness of spatial data, and application oriented studies dealing with public facilities must recognize and exercise care associated with both substantive context as well as technical considerations in carrying out analysis.

12.2 Review

As suggested throughout this book, there are a range of location models that have been applied in both descriptive and prescriptive ways to address a variety of public and private sector planning issues. There has, in fact, been an extensive amount of research devoted to the use of location models in fire and emergency response (ReVelle 1991; Marianov and ReVelle 1995; Badri et al. 1998; Sorensen and Church 2010; Murray 2013).

Hogg (1968) was among the early efforts to advocate systems analysis in fire protection planning, suggesting the problem of minimizing loss through the selection of a minimum number of stations (and where to locate them). This marked the beginning of theoretical research on location models to support such a planning situation, particularly oriented toward fire station siting. Toregas et al. (1971) is noteworthy

for establishing a connection to the notion of coverage, where response from a fire station is critical within a distance/time standard. They demonstrated that coverage of a demand point could be assessed for a potential facility location given a response standard (e.g., 6 min, 2 km, etc.). With this in mind, Toregas et al. (1971) detailed the location set covering problem (*LSCP*) as an approach for identifying the minimum number facilities, like fire stations, and where they should be sited in order to serve a region in a timely manner reflected in time/distance response standards. Church and ReVelle (1974) recognized that it may not be economically viable to expect all demand in a region to be served as imposed in the *LSCP* for emergency response facilities such as fire stations. That is, while most of the time response should be within the desired standards, on occasion slight delays in response may be unavoidable and acceptable. To address this, Church and ReVelle (1974) proposed the maximal coverage location problem (*MCLP*) seeking to maximize service demand coverage provided through the location of a pre-specified number of facilities. They demonstrated that considerable cost savings could be achieved compared to the number of facilities identified using the *LSCP*, but a high percentage of demand could still be ensured coverage within desired service standards.

A number of fire station siting studies have followed based on coverage models. Plane and Hendrick (1977) applied the *LSCP* to site fire stations in Denver, Colorado. Schreuder (1981) also utilized the *LSCP* for fire station placement in Rotterdam, Netherlands. Schilling et al. (1980) proposed an *MCLP* based approach for fire response in Baltimore, Maryland. Reilly and Mirchandani (1985) developed a combined median and coverage approach for fire service placement in Albany, New York. Badri et al. (1998) structured a model to minimize the maximum distance demand was from a sited fire station, applying this in Dubai, United Arab Emirates. They ensured that a maximum distance standard was not exceeded, which is essentially what is imposed in the *LSCP*. In Singapore, separate studies by Liu et al. (2006) and Huang and Fan (2011) relied on an *MCLP* based approach for fire station siting. Coverage models were also used in Catay (2011) as well as Aktas et al. (2013) for fire station planning in Istanbul, Turkey. Murray et al. (2012) and Murray (2013) report the use of an *MCLP* oriented model in fire station system design for the community of Elk Grove, California. Chevalier et al. (2012) applied an extension of the *MCLP* to site fire brigades in Belgium.

What is evident then is that coverage models like the *LSCP* and *MCLP* are important and widely applied in fire station siting because they explicitly account for response time standards. Not only this, but such response standards are often articulated in community, state and/or federal guidelines regarding fire service provision (Murray 2013). While clearly important and reflective of fire response needs and goals, the *LSCP* and *MCLP*, as well as extensions and variants, have been shown to be sensitive to scale and demand unit representation issues. Murray and O'Kelly (2002) demonstrated that point based representations intended to reflect continuous potential service demand across a region actually under-estimates spatial coverage provided when the *LSCP* is relied upon. Alternatively, Murray et al. (2008) showed that polygon based representations of a region will result in an excessive number of facilities identified as needed using the *LSCP*. Similar findings are reported for

the *MCLP* in Tong and Murray (2009). The implication is that spatial representation using *LSCP* or *MCLP* based modeling approaches to support facility siting could be a complicating factor capable of biasing analysis results in unintended (and undesirable) ways. More importantly, this could lead to poor strategic planning and decision making. Extensions to address bias/error issues in coverage models have been developed in Murray (2005), Murray et al. (2010), Tong (2012), among others, and coverage location modeling remains an active area of research.

Coverage models are important approaches that encapsulate a range of concerns central to fire response, as well as other public facilities, and therefore are an essential component of strategic planning involving the placement stations. While accounting for existing facilities is recognized as important, few studies have attempted to incorporate potential system changes in utilized approaches, yet such considerations are fundamental for the long term fiscal management of a public service system. Contemporary approaches that take advantage of geographic information system (*GIS*) capabilities and spatial data richness have seen limited use in fire station siting studies. The contributions of this chapter revolve around addressing the above issues in support of planning an integrated system of fire stations.

12.3 Location Planning Context

As reported in Murray et al. (2012), the author was commissioned by the Community Services District (*CSD*) Fire Department in Elk Grove, California to provide analysis supporting insight into how many future fire stations would be needed and where they should be sited. When the study began, Elk Grove had a population of approximately 144,000 people over 10,900 ha served by nine existing stations (three stations were not yet operational). The city and fire stations are shown in Fig. 12.1.

Their established standard was response within 5 min of a call for service. Analysis of call history, performance and response conducted by the fire department suggested that this length of time corresponded to an effective distance of 1.90 km from a station. Overall, the Elk Grove *CSD* Fire Department strived to maintain a 5 min response standard for at least 80 % of service calls.[1] This suggests that an *MCLP* based location model to support analysis and planning is particularly appropriate as this is an approach that accounts for coverage based standards intent on maximizing service to as much demand as possible.

The above conditions were to be assumed in the analysis. Further, the initial study was only to focus on the portion of the city where growth and development

[1] This standard is not unlike those found in many communities, though the time and performance percentage may vary. The National Fire Protection Association (NFPA) (2010), as an example, advocates that fire stations in urban areas be sited so that response times of nine minutes are achieved at least 90 % of the time.

Fig. 12.1 Elk Grove region with 9 existing fire stations

was anticipated. The population of Elk Grove is expected to more than double, with most of this anticipated in the southern half of the region. Expansion of the fire response system is seen as essential. New potential station sites were deemed possible anywhere in the area of expected growth. The fire department was not interested, at the time, in altering existing stations. Based on stipulated conditions, the analysis reported in Murray et al. (2012) indicated that 9 additional fire stations would be necessary, bringing the total number of fire stations to 18.

An important question remains about whether a better long term plan is possible for the community. As noted previously, existing systems are a byproduct of evolving needs over time. The resulting configuration may therefore become inefficient in a number of ways, potentially stressing or overburdening services in some areas while leaving some stations underutilized. This of course creates inequities, but also suggests system inefficiencies. It is precisely these issues that are important to address, especially when there are long term financial implications for inefficient system design. Most studies view (or are forced to view) existing services to be fixed and given, with the analysis focused on where new stations should be sited. This was essentially the case in Kanoun et al. (2010) and Murray et al. (2012), as an example. Plane and Hendrick (1977), Badri et al. (1998), Schilling et al. (1980), Catay (2011), Murray (2013) and Aktas et al. (2013), among others, discuss and/or provide evidence that simply expanding an existing system is problematic in the sense that system inefficiency may be introduced or perpetuated. Certainly there is a fiduciary responsibility to ensure that expenditures on fire response services are kept

to a minimum, and doing so may mean that reconfiguration of the existing system is paramount in order to realize increased efficiency.

As suggested above, some studies have attempted to address these issues within the context of an existing system in various ways. Plane and Hendrick (1977) accounted for a tradeoff in using a combination of new and existing facilities. Badri et al. (1998) incorporated fixed and annual costs in their goal programming approach. Schilling et al. (1980) detailed an approach to add a constraint restricting the number of existing facilities to utilize. This too was incorporated into the approach of Murray (2013). A different sort of approach is to assume that no facilities exist in the first place, then carry out the analysis to provide comparison of what could be possible if a region were able to begin anew. This was done most recently in the study reported in Aktas et al. (2013). Unfortunately, each of these approaches is lacking in certain ways, detached from specifically addressing issues of long term expenditures. Murray (2013), in fact, had to examine each of the various combinations of keeping existing facilities, then calculate the financial implications for a specific period of time. This was done following the analysis (and model application), requiring many different scenarios to be considered. What really is necessary is an explicit approach to address long term costs for a fire station response system, combining existing and new station decision making.

The following notation will be utilized in the derivation and discussion of the applied location model:

i index of demand areas to be served;
a_i demand for service in area i;
N_i set of potential fire stations that serve any portion of area i within the standard;
j index of potential future fire station locations;
c_j annual staffing and maintenance costs for fire station j;
f_j fixed cost to build a fire station at site j;
b_{ij} portion of area icovered by fire station j within stipulated service standard;
Φ set of sites where a new fire station could be built;
Ψ acceptable level of regional demand suitably served;
λ temporal horizon for which cost considerations apply;

$$X_j = \begin{cases} 1 & \text{if fire station at site } j \text{ is selected to be in system} \\ 0 & \text{otherwise} \end{cases}$$

Z_i amount of area i demand served by any fire station

With the above notation, it is possible to formalize the measures and criteria relied upon in the location model. To begin, a critical issue is system costs. In siting new stations, the prevailing cost would be building the fire station and acquiring the land. Given the set of potential locations for a new station, Φ, the fixed cost for site j would be f_j. The total costs for new facilities would therefore be:

$$\sum_{j \in \Phi} f_j X_j \tag{12.1}$$

where X_j denotes decisions about which facilities are selected. It is not uncommon that such fixed costs would essentially be the same. Thus, in the *LSCP* the objective is simply to minimize the total number of facilities, which is equivalent to minimizing total costs when the fixed cost at each potential fire station site is the same. Of course, fixed costs are not the only concern. Plane and Hendrick (1977), Badri et al. (1998) and Murray (2013) report that the annual costs of operating a fire station is equivalent to the fixed cost. The total annualized operational cost of the fire response system is the following:

$$\sum_j c_j X_j \tag{12.2}$$

This is across all stations, existing or new. If even one less fire station is possible through more efficient system design, taking into account existing and new stations, there are tremendous savings possible over the long term. This is precisely why it is important to explicitly consider existing and new stations when there are changing demands for service.

The objective of the *LSCP* is minimizing total fixed costs while ensuring that each demand is suitably covered. This is accomplished by imposing a service response constraint for each demand i:

$$\sum_{j \in N_i} X_j \geq 1 \tag{12.3}$$

This guarantees that the minimum number of facilities found using the *LSCP*, objective (12.1), are configured so that at least one fire station is located within the response standard for each demand. Noted above was that the Elk Grove response standard is service within 5 min (1.90 km). In contrast to the *LSCP*, the *MCLP* recognizes that serving the entire region under strict response standards may not be cost effective. The implication is that constraints (12.3) must effectively be relaxed for select demand. To address this, the *MCLP* has as an objective to maximize demand suitably served within the standard. Specifically, Elk Grove has stipulated that response within 5 min should occur for at least 80 % of service calls. Based on our notation here, the *MCLP* objective is:

$$\sum_i Z_i \tag{12.4}$$

as Z_i is the demand covered by the fire station siting configuration. The *MCLP* also imposes a particular level of system investment:

$$\sum_j X_j = p \tag{12.5}$$

where p is the number of fire stations to site, specified a priori. Murray and Tong (2009) highlight that analysis must therefore be conducted to identify the minimum

investment level, p, that provides the stipulated level of regional coverage, Ψ. Typically the tradeoff curve showing regional demand coverage for each value of p, 1 facility, 2 facilities, 3 facilities, etc., is consulted, as illustrated in Church and ReVelle (1974). It is possible, however, to relate coverage in percentage terms as follows:

$$100\left(\sum_i a_i\right)^{-1} \sum_i Z_i \tag{12.6}$$

The terms before the summation simply standardize Eq. (12.4) based on total demand in the region, giving a percentage of demand covered by the fire station configuration. One can then impose a bound on this percentage as a model constraint. Murray and Tong (2009) demonstrate that it is possible to use this as a replacement for objective (12.4) and constraints (12.5), representing an extension of the classic *MCLP* that they refer to as the *Threshold Coverage Model*.

What remains to be discussed is tracking demand coverage using decision variables Z_i. The *LSCP* and *MCLP* originally were conceived assuming that demand was represented as a point. Coverage in this case is fairly straightforward. A point is either covered or it is not. Some ambiguity arises, however, when demand is actually a line, ployline, polygon or other geometric object. Assessment of coverage may not be clear when the demand is only partially covered/served within the response standard. As noted previously, it is precisely this representational issue that research has found to introduce bias, error and uncertainty in *LSCP* and *MCLP* application (see Murray and O'Kelly 2002; Murray et al. 2008; Tong and Murray 2009). To address this issue, Murray (2005) introduced the idea of complementary coverage for the *LSCP* in the case where demand is a line, polyline, polygon, etc. The idea is that there is an implicit expectation of demand being served based on spatial complementarity, but this must be structured in the model. An alternative is to explicitly account for portions of demand being covered by a facility siting configuration. This was detailed for the *MCLP* in Tong and Murray (2009) and requires enumeration of specific coverage configuration. A summary review of implicit and explicit coverage approaches can be found in Murray et al. (2010). For the *MCLP*, Tong (2012) proposed an implicit approach tracking the portion of demand area i served by selected facilities:

$$\sum_{j \in N_i} b_{ij} X_j \tag{12.7}$$

where N_i is the set of facilities within the coverage standard of demand i and b_{ij} represents the amount of area i served by a facility sited at location j. While not exact in terms of the actual portion of demand i served, empirical evidence suggests that it can be used in an effective manner to track coverage. This basic approach was relied upon in Murray (2013).

With the above details and discussion, it is possible to now formalize an *MCLP* based model extension to support fire station siting, taking into account existing and new services as well as spatial representation issues. The model is as follows:

$$\text{Min} \sum_{j \in \Phi} f_j X_j + \lambda \sum_j c_j X_j \tag{12.8}$$

$$\text{s.t. } 100 \left(\sum_i a_i \right)^{-1} \sum_i Z_i \geq \Psi \tag{12.9}$$

$$\sum_{j \in N_i} b_{ij} X_j \geq Z_i \quad \forall\, i \tag{12.10}$$

$$Z_i \leq a_i \quad \forall\, i \tag{12.11}$$

$$X_j = \{0,\ 1\} \forall\, j \tag{12.12}$$

$$Z_i \geq 0 \forall\, i \tag{12.13}$$

The objective, (12.8), minimizes total costs over time horizon λ, including fixed and annual recurring expenses. Constraint (12.9) establishes the percentage of regional demand that must be covered within the stipulated service standard. Constraints (12.10) implicitly track the portion of a demand served by sited stations. Constraints (12.11) bound possible coverage of a demand area. Binary integer and non-negativity requirements are imposed in constraints (12.12) and (12.13), respectively.

This model is an extension of the *MCLP*, related to the Threshold Coverage Model detailed in Murray and Tong (2009). This model further extends the so called Complementary Threshold Coverage Model detailed in Murray (2013), enabling long term financial expenditures over time, λ, to be explicitly account for. Unlike the *MCLP*, the number of facilities required is endogenously determined based on achieving regional coverage requirements stipulated in constraints (12.9). The model is also structured to accommodate alternative spatial representations of demand, including points, lines, polygons, etc., addressing issues that could arise in the application of the *LSCP* and *MCLP* for non-point demand.

12.4 System Insights

The analysis is carried out using a Windows-based computer (Intel Xeon processor at 2.53 GHz) with 6 GB of *RAM*. The location model was structured using Python, relying on *GIS* based libraries (Shapely), and solved using commercial optimization software (Gurobi).

There are 638 neighborhoods representing potential demand for fire response (all polygons in Fig. 12.1). As fire stations can technically be sited anywhere in continuous space, processing to identify a finite dominating set of discrete potential locations was carried out. Murray and Tong (2007) derived a finite dominating set for coverage based problems like the *LSCP* and *MCLP* involving non-point demand

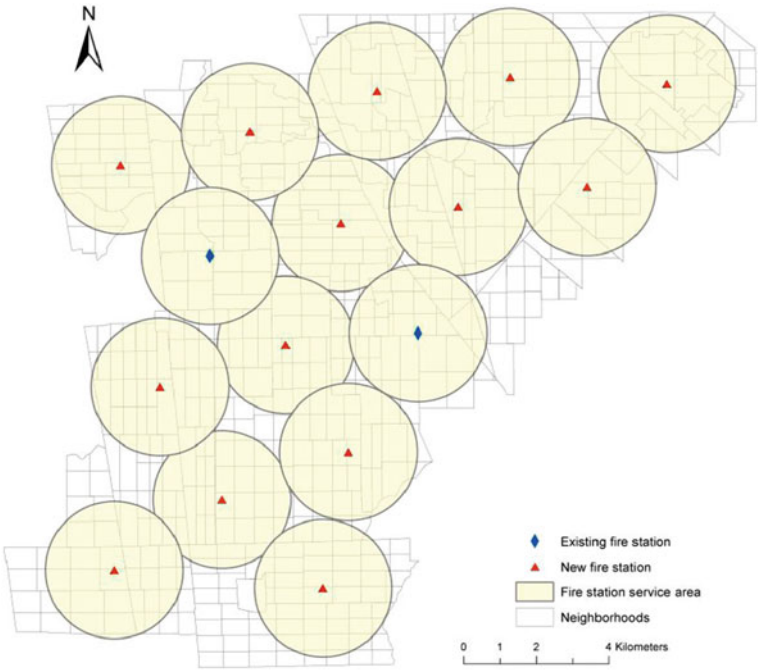


Fig. 12.2 Minimum long term cost siting configuration

objects. They proved that this set would contain an optimal solution for the continuous space context when facilities may be located anywhere in the region. The resulting set in this case consisted of 1638 points. This set combined with the nine existing station sites gives 1647 potential fire station locations for consideration. A time horizon of 10 years was used ($\lambda = 10$). The assumed fixed cost for a new fire station is \$2 million (f_j for $j \in \Phi$), with the annual cost per station of \$2 million (c_j). The location model contained 2285 decision variables (1647 integer and 638 continuous) and 1277 constraints. The data processing time to structure the model was 325.4 s and solution time was 3389 s.

As noted above, if the existing system is fixed and given with new stations to be sited in order to cover currently unserved portions of region, then 18 fire stations (9 existing and 9 new) would be necessary to achieve at least 80 % system response stipulated in constraint (12.9). That is, response within 5 min to at least 80 % of calls for service would be ensured for a configuration of 18 stations. Considering a time horizon of $\lambda = 10$, the total costs for Elk Grove to provide fire service over 10 years would be \$378 million (18 stations costing \$2 million a year per station over 10 years plus nine new stations at a cost of \$2 million per station). The question of interest is whether this total cost can be decreased.

Applying the location model in order to consider siting decisions involving existing as well as new fire stations finds that 16 fire stations are sufficient. This configuration of fire stations is shown in Fig. 12.2. Evident is that 2 existing stations

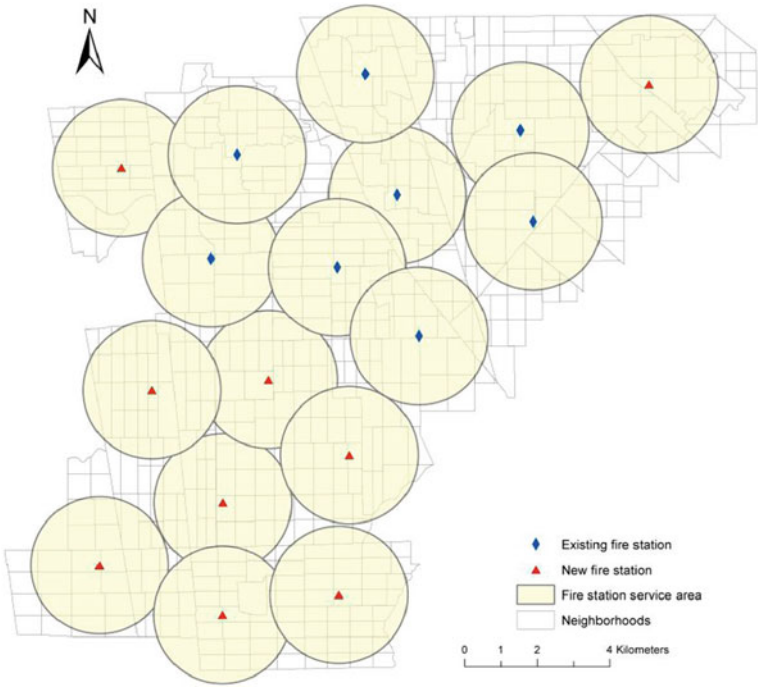


Fig. 12.3 Minimum short term cost (1–4 year horizon) siting configuration

are maintained supplemented by 14 new stations. It is therefore possible to provide suitable response to over 80 % of demand in Elk Grove using only 16 fire stations. The cost to provide this service will be $348 million over 10 years. Needless to say, this is a substantial reduction in costs over a rather arbitrary time horizon. Still, $30 million over 10 years is a significant savings for any community. Of course, a longer period of time would mean greater realized savings. Worth noting is that if λ is varied and the model reapplied, alternative recommendations could be reached based upon a shortened return on investment window. For example, 1–4 years would suggest a 17 fire station system consisting of 8 existing stations and 9 new stations. This configuration is shown in Fig. 12.3. While appealing because is makes use of most of the existing system, after 5 years it becomes financially more prudent to select the minimum number of stations (16) depicted in Fig. 12.2.

12.5 Future Steps

There are many issues that arise in any application oriented study. These issues span articulation of the problem but also practical considerations regarding data, projected demand for service, response, personnel availability, etc., not to mention

decision making context, political and social preferences. Bridging theoretical and application concerns is no simple task, making it a challenging area to work in. There is no doubt more to be done in this area given better supporting data and greater insights. Highlighted below are some of the primary considerations driving planned future research.

Somewhat unexpected was that the Elk Grove *CSD* Fire Department merged with a neighboring community, the Galt Fire Protection District, shortly after the completion of the project. This combined with the economic downturn continues to thwart plan implementation, with no changes to date to the fire response system. The future remains uncertain at this point, yet the population has increased by over 10 %.

Irrespective of system implementation or change issues, the research did not address timing considerations of bringing a new system online. Of course this is not trivial as communities like Elk Grove have real budgetary limitations, but also growth and development is a process that is controlled by economic and individual decision making. A second issue involved expected demand for service. The currently served portions of Elk Grove are largely developed. However, the unserved areas are mostly agricultural with residential development expected over the coming years. Addressing expected demand required an assumption of uniformly distributed demand for service. The *CSD* Fire Department made this assumption based on call history, but there does remain a source of uncertainty regarding projected service demand. This may or may not impact fire station location decisions. Extensions to address projected growth over time combined with annual investment possibilities over time are interesting areas for future research. Of course, this requires good future demand predictions and cost estimates to better refine model parameters, like a_i, c_j and f_j, where temporal specificity is possible. If t represents an index of time periods, it is conceivable that demand, annual cost and fixed cost could be indicated temporally as a_{it}, c_{jt}, and f_{jt}, respectively. No doubt there would need to be subsequent modification of the model to account for this enhanced detail and specificity.

Further location model attention would be spatial representation and detail. The reviewed literature has discussed recognized issues in spatial representation known to impact results identified using coverage models. While the complementary coverage approach reported here is certainly an improvement, it is not free of all potential error or uncertainty (Murray et al. 2010; Tong 2012). This is precisely why research continues on coverage location model approaches for reducing and/or eliminating representation error and uncertainty, such as that by Murray and Wei (2013) and Wei and Murray (2014, 2015). This too remains a promising area for future research.

Acknowledgements This work benefitted from assistance and comments by Dr. Ran Wei (University of Utah).

References

Aktas E, Ozaydın O, Bozkaya B, Ulengin F, Onsel S (2013) Optimizing fire station locations for the Istanbul metropolitan municipality. Interfaces 43:240–255

Badri MA, Mortagy AK, Alsayed A (1998) A multiobjective model for locating fire stations. Eur J Oper Res 110:243–260

Brill ED (1979) The use of optimization models in public-sector planning. Manage Sci 25(5): 413–422

Catay B (2011) Siting new fire stations in Istanbul: a risk based optimization approach. OR Insight 24:77–89

Chevalier P, Thomas I, Geraets D, Goetghebeur E, Janssens O, Peeters D, Plastria F (2012) Locating fire-stations: an integrated approach for Belgium. Socio-Econ Plan Sci 46:173–182

Church RL, Murray AT (2009) Business site selection, location analysis and GIS. Wiley, New York

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

Daskin MS, Murray AT (2012) Modeling public sector facility location problems. Socio-Econ Plan Sci 46:111

Francis RL, McGinnis LF, White JA (1992) Facility layout and location: an analytical approach, 2nd edn. Prentice Hall, New Jersey

Hogg JM (1968) The siting of fire stations. Oper Res Q 19:275–287

Huang Y, Fan Y (2011) Modeling uncertainties in emergency service resource allocation. J Infrastruct Syst 17:35–41

Kanoun I, Chabchoub H, Aouni B (2010) Goal programming model for fire and emergency service facilities site selection. INFOR: Inf Syst Oper Res 48:143–153

Liebman JC (1976) Some simple-minded observations on the role of optimization in public systems decision-making. Interfaces 6:102–108

Liu N, Huang B, Chandramouli M (2006) Optimal siting of fire stations using GIS and ant algorithm. J Comput Civil Eng 20:361–369

Marianov V, ReVelle C (1995) Siting emergency services. In: Drezner Z (ed) Facility location: a survey of applications and methods. Springer, New York, pp 199–223

Marianov V, Serra D (2002) Location problems in the public sector. In: Drezner Z, Hamacher HW (eds) Facility location: applications and theory. Springer, Berlin, pp 119–150

Murray AT (2005) Geography in coverage modeling: exploiting spatial structure to address complementary partial service of areas. Ann Assoc Am Geogr 95:761–772

Murray AT (2010) Advances in location modeling: GIS linkages and contributions. J Geograph Syst 12:335–354

Murray AT (2013) Optimising the spatial location of urban fire stations. Fire Saf J 62:64–71

Murray AT, O'Kelly ME (2002) Assessing representation error in point-based coverage modeling. J Geograph Syst 4:171–191

Murray AT, Tong D (2007) Coverage optimization in continuous space facility siting. Int J Geogr Inf Sci 21:757–776

Murray AT, Tong D (2009) GIS and spatial analysis in the media. Appl Geogr 29:250–259

Murray AT, Wei R (2013) A computational approach for eliminating error in the solution of the location set covering problem. Eur J Oper Res 224:52–64

Murray AT, O'Kelly ME, Church RL (2008) Regional service coverage modeling. Comput Oper Res 35:339–355

Murray AT, Tong D, Kim K (2010) Enhancing classic coverage location models. Int Reg Sci Rev 33:115–133

Murray AT, Tong D, Grubesic TH (2012) Spatial optimization: expanding emergency services to address regional growth and development. In: Stimson R, Haynes K (eds) Studies in applied geography and spatial analysis. Edward Elgar, UK, pp 109–122

National Fire Protection Association (NFPA) (2010). NFPA 1710: Standard for the organization and deployment of fire suppression operations, emergency medical operations, and special operations to the public by career fire departments. http://www.nfpa.org/. Accessed 13 April 2015

Plane DR, Hendrick TE (1977) Mathematical programming and the location of fire companies for the Denver Fire Department. Oper Res 25:563–578
Reilly JM, Mirchandani PB (1985) Development and application of a fire station placement model. Fire Technol 21:181–198
ReVelle C (1987) Urban public facility location. In: Mills E (ed) Handbook of regional and urban economics, volume II. Elsevier, New York, pp. 1053–1096
ReVelle C (1991) Siting ambulances and fire companies—new tools for planners. J Am Plan Assoc 57:471–485
Schilling DA, ReVelle C, Cohon J, Elzinga DJ (1980) Some models for fire protection locational decisions. Eur J Oper Res 5:1–7
Schreuder JAM (1981) Application of a location model to fire stations in Rotterdam. Eur J Oper Res 6:212–219
Sorensen P, Church R (2010) Integrating expected coverage and local reliability for emergency medical services location problems. Socio-Econ Plan Sci 44:8–18
Tong D (2012) Regional coverage maximization: a new model to account implicitly for complementary coverage. Geogr Anal 44:1–14
Tong D, Murray A (2009) Maximizing coverage of spatial demand for service. Pap Reg Sci 88: 85–97
Toregas C, Swain R, ReVelle C, Bergman L (1971) The location of emergency service facilities. Oper Res 19:1363–1373
Wei R, Murray AT (2014) Evaluating polygon overlay to support spatial optimization coverage modeling. Geogr Anal 46(3):209–229
Wei R, Murray AT (2015) Continuous space maximal coverage: insights, advances and challenges. Computers and Oper Res (66: 325–336)

Chapter 13
Locating Vehicle Identification Sensors for Travel Time Information

Monica Gentili and Pitu B. Mirchandani

13.1 Introduction

Currently, transportation networks are significantly instrumented with various types of sensors, ranging from simple inductive loop detectors embedded in the pavement that produce a signal when a vehicle is over it, to image detectors which use cameras machine vision to detect activities over the pavement, to mechanical gauges to measure the strain on the pavement and associated infrastructure like bridges, to *GPS*-equipment in the vehicles that is be able to track vehicles in the network (commonly referred to as traffic probes when they are used to obtain a congestion measure of traffic such as travel times). Machine vision and laser-based readers are used to automatically collect tolls on a highway. Traffic sensors may also be used to detect incidents on a road by observing a sudden change in traffic flow patterns. All these uses are ultimately related to (a) managing traffic on a network by monitoring traffic and making traffic control decisions to decrease congestion, energy use, air pollution, etc. and (b) planning networks so that the demand using the road network is appropriately served. Thus, locating sensors optimally improves our traffic management and network planning decisions.

The problem of optimally locating sensors on a traffic network has been the object of growing interest in the past few years. Sensor location decision models differ from each other according to the type of sensors that are to be located and the objective function that one would like to optimize. Many different models have been proposed in the literature as well as corresponding solution approaches. The proposed existing models could be classified according to two main criteria: (i) the

M. Gentili (✉)
Department of Mathematics, University of Salerno, Fisciano, Salerno, Italy
e-mail: mgentili@unisa.it

P. B. Mirchandani
School of Computing, Informatics and Decision Systems Engineering,
Arizona State University, Tempe, AZ, USA
e-mail: pitu@asu.edu

H. A. Eiselt, V. Marianov (eds.), *Applications of Location Analysis,* International Series in Operations Research & Management Science 232, DOI 10.1007/978-3-319-20282-2_13

typology of sensors to be located on the network (e.g., counting sensors, image sensors, Automatic Vehicle Identification (*AVI*) readers), and, (ii) the objective function to be optimized, such as minimizing errors in estimating Origin- Destination (OD) trips, route flows, link flows, or travel times. Among the first models, Lam and Lo (1990) addressed the problems of locating sensors to estimate OD flows and studied how different deployments of sensors on the network could affect OD flow estimation. Since then, studies on locating sensors have grown rapidly due to the development of new methods and technologies. Indeed, traditionally, counting sensors have collected speed and flow volume data, but sensors based on technologies such as vehicle recognition and *GPS* can give much more information about the operation of the network. Hence new objective functions have been studied: route flow estimation and observability, link flow estimation and observability, travel time measurement and estimation, etc.

The authors have recently reviewed the literature related to sensor location for flow observability and estimation (Gentili and Michandani 2011, 2012). The focus of this chapter is on the optimal location of sensors to monitor travel time related information.

13.2 Travel Time Measurement and Estimation

Real-time measurement and estimation of travel times is a useful tool for traffic management and for providing information to travelers. For example, this knowledge is used for dynamically updating travel time predictions. Traditionally, collection of information for travel time estimation has been carried out by transportation agencies through an extensive use of counting sensors (i.e., loop inductance detectors) uniformly located on freeways and highways usually at distances of 0.5–1.0 miles. The origin of such a practice is based on past analysis that has shown uniform spacing among detectors is useful for Automatic Incident Detection algorithms (*ADI*) (the role of *ADI* algorithms based on loop detector data is addressed in Chap. 7 of this volume). However, it should be noted that given the massive use of *GPS* enabled smart phones, the data from the resulting probe vehicles' travel times are increasingly being used for incident detection as well as travel time estimation. Nevertheless, for situations where this kind of information is not readily available, either because of low market share of smart phones or when data communication networks are not sufficiently advanced or when communication of vehicle data is expensive, the location of fixed permanent sensors is still an important and relevant approach to gather real time data for estimating travel times. The effect of counting sensor deployment on the quality of the measurements and estimation of travel times has been widely studied (see for example, Ban et al. 2009; Bartin et al. 2007; Chan and Lam 2002; Chaudhuri et al. 2011; Danczyk and Liu 2011; Edara et al. 2010; Fujito et al. 2007; Liu et al. 2006). Travel time measurement can also be obtained using technology such as Automatic Vehicle Identification (*AVI*) readers. Examples include sensors that use machine vision

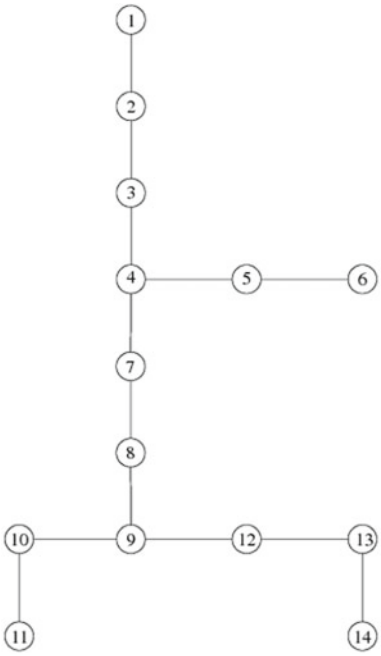


Fig. 13.1 A simple network example

to read license plates, sensors that read unique blue tooth signals from each vehicle that passes in its vicinity, and sensors that read a radio electronic tag much like the *RFID* tags used to track parts in manufacturing facilities. The remainder of the chapter will review existing approaches to optimally locating *AVI* readers on networks to estimate travel times and related measures.

13.3 Travel Time and Related Information on Networks

The use of *AVI* readers for estimating roadway travel times is a recent development. Their primary application has been for electronic toll collection. *AVI* reader systems rely on a combination of passive tags attached to vehicles and electronic readers installed on the roads (above lanes, or on roadsides). A vehicle is detected and identified every time it passes a reader. Upon detection of a vehicle, the system transmits the information to a data processing center where a measurement of the travel time between two consecutive readers can be computed. For example consider the network in Fig. 13.1. By locating two *AVI* readers, one on link (4, 7) and one on link (8, 9), all the vehicles that pass both locations are intercepted, and, by comparing the interception times, information on the travel time on the portion of the intercepted network can be determined. Obviously, one could be interested in determining travel

Table 13.1 OD pairs, OD trips and OD routes related to the example network in Fig. 13.1

OD pair	Route ID	Trips	Links
1–3	1	100	(1, 2) (2, 3)
1–9	2	10	(1, 2) (2, 3) (3, 4) (4, 7) (7, 8) (8, 9)
7–11	3	10	(7, 8) (8, 9) (9, 10) (10, 11)
5–13	4	10	(5, 4) (4, 7) (7, 8) (8, 9) (9, 12) (12, 13)

times on links of the network, on particular routes in the network, or travel times between OD pairs in the network. Optimal location decisions depend on what travel times one wants to estimate.

Assume, for example, the network of Fig. 13.1 has the following OD pairs $W = \{(1, 3), (1, 9), (7, 11), (5, 13)\}$, each of them connected, in this simple example, by a single route (these are listed in Table 13.1). To get information about travel times between OD pair $w = (1, 3)$ one needs to locate *AVI* reader at the upstream and downstream end-points of the route connecting w, that is one on link (1, 2) and the other on link (2, 3). Hence, to have information about travel times among all the OD pairs in the network, *AVI* readers should be located on links (1, 2), (2, 3), (7, 8), (8, 9), (10, 11), (5, 4) and (12, 13). This would provide knowledge of the travel times on the entire routes connecting all the OD pairs in the network. However, in case of multiple routes between OD pairs and also when there are budget limits to acquire sensors, such detailed knowledge is unlikely to be collected. In this case, one cannot get information about travel times on all entire routes but only on portions of some of them. For example, by locating readers on links (4, 7) and (8, 9), travel times on parts of routes two and four could be collected. Consider now the OD trips listed in Table 13.1. If the purpose of determining travel times is to disseminate such information among commuters, then the majority of users would not benefit from the reader located on links (4, 7) and (8, 9). It would be preferable to locate one of the readers on link (1, 2) and the other on (2, 3) to intercept as many commuters as possible for as long as possible.

It is clear, then, that different factors should be considered when locating *AVI* readers. Some of them are:

- *Number of AVI readers*: this is a measure of the installation cost of the system.
- *Number of readings*: a *reading* is obtained when the same vehicle is intercepted at two different locations. This measure relates to the amount of travel time data that is obtained from the *AVI* system.
- *Length between readings*: a reading obtained by two readers that are far apart has more travel time information compared with readings obtained by two readers that are close to each other.
- *Number of OD pairs covered*: For an O-D pair to be covered, at least one reading should be obtained from one of the routes between the OD pair. This is useful

for estimating OD flows, and estimating OD travel times when the network is in equilibrium (Sheffi 1984).

- *Travel time reliability*: If travel time along a link has little variance, then a real-time travel time reading on that link has little value; if variance is large, then real-time travel times are very useful. Hence, sensors can collect real-time information to estimate travel time reliability.

The models reviewed in the next section take into account some of these factors in developing the locational decision criteria and associated models. In particular, six different mathematical formulations are presented to optimally locate readers on the links of the network, which we refer to as *AVI* Location Problems or *AVIL* models.

13.4 Mathematical Models

We can represent a traffic network by a graph $G=(V, A)$ where the set of nodes V represents intersections in the network and the set of links A, joining node pairs, represents roads. We denote by W the set of OD pairs of the network and by R_w the set of routes connecting the OD pair $w \in W$. The entire set of routes in the network is denoted by R and we have $R = \bigcup_{w \in W} R_w$. We denote by h^w the average number of trips connecting the OD pair w for the period of study, and by f_r the average flow volume on route r. Additional notation will be introduced as needed.

The first model (model $AVIL_1$), presented by Sherali et al. (2006), works under the assumption that there is a single route connecting each OD pair. This model associates different weights (benefit) to OD pairs and seeks to maximize the total benefit accrued by locating at most k readers, respecting a given budget constraint. An OD pair is considered to be covered if the readers are located on the route connecting the OD pair at the upstream and the downstream end-points of the route (particularly, at the beginning and ending links). In particular, let us consider the OD pair w $\in W$ and consider the shortest route r_{ij} connecting it, that starts with link i and ends with link j. Due to the simplifying assumption of having one route for each OD pair, we can use route r_{ij} to represent its corresponding OD pair. Route r_{ij} is defined to be covered if readers are located both on link i and on link j. When the route r_{ij} is covered, a benefit u_{ij} is obtained. Let c_j be the cost of installing a reader on link j. Define z_j to be a binary variable associated with link $j \in A$ that assumes a value of 1 if a reader is located on the link and 0 otherwise. The $AVIL_1$ model is the following:

$$AVIL_1\text{: Max} \sum_{r_{ij} \in R} u_{ij} z_i z_j \tag{13.1}$$

$$\text{s.t.} \sum_{j \in A} z_j \leq k \tag{13.2}$$

$$\sum_{j \in A} c_j z_j \leq C_{max} \tag{13.3}$$

$$z_j \in \{0,1\} \ \forall \ \ j \in A \tag{13.4}$$

The objective function (13.1) maximizes the total benefit accrued by locating the readers. Constraint (13.2) limits the sensors to at most k readers. Constraint (13.3) imposes the total budget to be at most C_{max}. A possible definition of the benefit factor is also proposed by Sherali et al. (2006). They defined the benefit factor u_{ij} to describe the relevance of route r_{ij} (and hence, of its corresponding OD pair) in capturing information about the variability of travel times in the network. Such factors are computed according to the coefficient of variation associated with each link j, defined as follows:

$$COV_j = \frac{\sigma_j}{\mu_j}, \tag{13.5}$$

where μ_j and σ_j are estimates of the mean and standard deviation of travel times on link j, respectively. For each route r_{ij}, the corresponding benefit factor u_{ij} is computed considering the coefficients of variation (13.5) associated with the links that belong to the route. In particular, the suggested benefit u_{ij} is computed as:

$$u_{ij} = COV_{ij} = \sqrt{\frac{\sum_{a \in r_{ij}} \sigma_a^2}{\sum_{a \in r_{ij}} \mu_a^2}}. \tag{13.6}$$

Sherali et al. (2006) linearized model $AVIL_1$ and applied the Reformulation Linearization Technique (Adams and Sherali 1986, 1990) to solve it. They applied the model to data from Interstate-35 Freeway, San Antonio, Texas, USA.

The more realistic assumption of having multiple routes between each OD pair is considered in the other models $AVIL_i$, $i = 2, \ldots, 6$. These models assume an OD pair to be covered if there are at least two readers located on each of the routes connecting the OD pair, without any restriction on where on the routes they are located. Model $AVIL_2$ to follow is basically a set covering formulation aimed at minimizing the total number of readers to be located on the network to ensure the coverage of all the OD pairs. Let z_j be the number of sensors located on link j. Model $AVIL_2$ is as follows (Chen et al. 2004).

$$AVIL_2\text{: Min} \sum_{j \in A} z_j \tag{13.7}$$

$$\text{s.t.} \sum_{j \in r} z_j \geq 2 \ \ \forall\, r \in R_w \ \ \forall \ \ w \in W \tag{13.8}$$

$$z_j \in \{0,1,2,\ldots\} \ \ \forall \ \ j \in A \tag{13.9}$$

The objective function (13.7) minimizes the total number of readers to be located on the network. Constraints (13.8) ensure that each OD pair is covered. Since the location variable z_j associated with link j denotes the total number of readers that can be installed on the link, the total number of readers located on a link can be

greater than 1. The maximum number of readers that can be installed on each link of the network can be determined by the spacing requirement between pairs of readers. Moreover, note the slightly different definition of coverage of an OD pair compared to the one used in model $AVIL_1$: an OD pair w is considered to be covered when there are at least two readers installed on each route $r \in R_w$ connecting the OD pair. Unlike model $AVIL_1$ where for each route the only locations for readers that are considered are on the first and last link of the route, readers for model $AVIL_2$ can be located anywhere on the route.

When the total number of readers is limited, a different objective function has been proposed. The following model, $AVIL_3$ (Chen 2004), is a variant of the well known maximal covering location problem (Church and ReVelle 1974) where, for a fixed number of available readers to be installed, the objective is to maximize the total number of covered OD pairs.

$$AVIL_3\text{: Max} \sum_{w \in W} y_w \tag{13.10}$$

$$\text{s.t. } y_w \leq max \left\{ 0, \sum_{j \in r} z_j - 1 \right\} \ \forall \ r \in R_w \ \forall \ w \in W \tag{13.11}$$

$$\sum_{j \in A} z_j \leq k \tag{13.12}$$

$$z_j \in \{0,1,2,\ldots\} \ \forall \ j \in A \tag{13.13}$$

$$y_w \in \{0,1\} \ \forall \ w \in W \tag{13.14}$$

Variables y_w, defined for each OD pair, assume value equal to 1 if the OD pair w is covered and 0 otherwise; and, therefore, objective (13.10) maximizes the total number of ODs covered. The coverage is ensured by the set of constraints (13.11) which force y_w to be equal to 0 unless there are at least two readers located on each route connecting w. The total number of installed readers is limited to k through constraint (13.12).

The main differences between model $AVIL_2$ and $AVIL_3$ with respect to the classical set covering problem and the maximal covering problem, respectively, are: (*i*) multiple location of readers on the same link is allowed, and (*ii*) the definition of OD pair coverage requires the installation of at least two devices on each covered route connecting the OD pair.

The following model $AVIL_4$ locates k readers on the links of a network to maximize the weighted sum of the total number of readings and of the total number of covered OD pairs (Teodorovic 2002):

$$AVIL_4\text{: Max} \frac{h_1}{H} \sum_{r \in R} f_r \left(\sum_{j \in r} z_j - 1 \right) + \frac{h_2}{|W|} \sum_{w \in W} y_w \tag{13.15}$$

$$\text{s.t.} \quad y_w \leq \max \left\{ 0, \sum_{j \in r} z_j - 1 \right\} \quad \forall \ r \in R_w, \ \forall \ w \in W \tag{13.16}$$

$$\sum_{j \in A} z_j \leq k \tag{13.17}$$

$$z_j \in \{0, 1, 2, \ldots\} \ \forall \ j \in A \tag{13.18}$$

$$y_w \in \{0, 1\} \ \forall \ w \in W \tag{13.19}$$

where (*i*) h_1, h_2, are the weights assigned to model the importance of the two components in the objective and (*ii*) H is the total number of trips in the network (i.e., $H = \sum_{w \in W} h^w$). The criterion (13.15) maximizes the weighted sum of two components in the objective: the first component is the ratio between the total number of readings obtained from the locations of the readers and the total trips in the network; and the second component of the objective function is the ratio between the total number of covered OD pairs and the total number of OD pairs in the network. When a weight u_w is assigned to each OD pair w as a measure of importance, then the second component of the objective function could also be: $\frac{h_2}{\sum_{w \in W} u_w} \sum_{w \in W} u_w y_w$. The coverage of the OD pairs is considered by the set of constraints (13.16) which force y_w to 0 unless there are at least two readers located on each route connecting w. Constraint (13.17) assumes that the total number of installed readers cannot be greater than k.

Distances among two located readers (that is, the length of a reading) is not considered in any of the above models. In $AVIL_5$ this measure is defined as the total vehicle-miles monitored. In $AVIL_6$ it is a general benefit factor. These two models are explained next.

Model $AVIL_5$ solves the problem of optimally locating k readers on the links of the network to maximize the number of vehicle-miles (or vehicle-travel times) monitored. For each link $j \in A$ and each route $r \in R$ the parameter d_j^r denotes the distance from the origin of route r to a potential reader located on link j. The following sets of variables are defined for $AVIL_5$: the binary variable z_j equals to 1 if a reader is located on link j and 0 otherwise; the binary variable y_{rj} equals to 1 if a reader on link j is the most downstream reader located on route r and 0 otherwise; the binary variable x_{rj} equals to 1 if a reader on link j is the most upstream reader located on route r and 0 otherwise. The mathematical formulation is the following (Mirchandani et al. 2009):

$$AVIL_5\text{: Max} \sum_{r \in R} \sum_{j \in r} \left(y_{rj} - x_{rj} \right) d_j^r f_r \tag{13.20}$$

$$\text{s.t.} \sum_{j \in A} z_j \leq k \tag{13.21}$$

$$y_{rj} \leq z_j \ \ \forall \, r \in R \, \forall \ j \in A \tag{13.22}$$

$$x_{rj} \leq z_j \quad \forall \; r \in R \quad \forall \; j \in A \tag{13.23}$$

$$\sum_{j \in r} y_{rj} \leq 1 \quad \forall \; r \in R \tag{13.24}$$

$$\sum_{j \in r} x_{rj} \leq 1 \quad \forall \; r \in R \tag{13.25}$$

$$\sum_{j \in r} y_{rj} - \sum_{j \in r} x_{rj} = 0 \quad \forall \; r \in R \tag{13.26}$$

$$z_j \in \{0,1\} \quad \forall \; j \in A \tag{13.27}$$

$$y_{rj} \in \{0,1\} \quad \forall \; r \in R \quad \forall \; j \in A \tag{13.28}$$

$$x_{rj} \in \{0,1\} \; \forall \; r \in R \quad \forall \; j \in A \tag{13.29}$$

Objective function (13.20) maximizes the total vehicle-miles monitored. Note that this function takes into account both the length of each reading and the total number of readings. Constraint (13.21) requires locating no more than k readers on the network. Constraints (13.22) and (13.23) are logical constraints linking the variables and ensuring that there cannot be a most upstream (or a most downstream) reader on a route if no reader is located on it. Constraints (13.24) and (13.25) ensure that there is no more that one most upstream reader and no more than one most downstream reader on each route. Finally, constraint (13.26) states that if there is a most upstream reader on route r, there must also be a most downstream reader on the same route and vice versa. Similarly, if there is no most upstream reader on route r, then there cannot be a most downstream reader on the same route and vice versa. It is also true that if on a route there is only one reader installed, then it is both the most upstream and downstream reader on that route, and the vehicle miles monitored on that route is zero.

The objective function of $AVIL_5$ can be modified so that the model tries to locate AVI readers to obtain travel time statistics and improve predictability of estimated travel times. This model, referred to as $AVIL_{5'}$, is illustrated in Sect. 5 where its application is discussed.

A generalized surveillance benefit (Li and Ouyang 2012) is maximized in model $AVIL_6$. Consider a route r, and, without loss of generality, let us assume every link of the route is a candidate location for a reader. If s_r readers are located on the route, then we can divide the route into $s_r - 1$ sections where each section has a reader on its upstream boundary link and its downstream boundary link. Without loss of generality, we also assume there are two virtual locations b and t for each route, located at the very beginning and at the very end of the route with two virtual readers located at positions 0 and $s_r + 1$, respectively. Denote the portion of the route between link j and link h by r_{jh}. We say that the portion r_{jh} is monitored if (i) a reader is located on link j and on link h, and (ii) there is no reader located in between (that is, the two readers on j and h are consecutive). If r_{jh} is a monitored portion, then the two links j and h are said to be paired on route r. More specifically, given

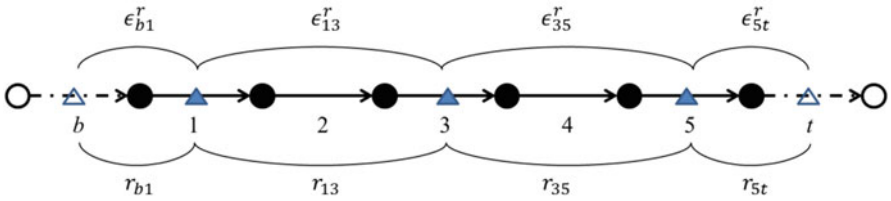


Fig. 13.2 Monitored portions of a route and corresponding benefit factors

a route r and s_r readers located on it, the link of the route where the i-th reader is located is paired with the link where the $(i+1)$-st reader is located, and with the link where the (i–1)-st reader of the route is located. The virtual initial link b at position 0 is paired with link j_1 where the first real reader is located; the virtual last link t at position $s_r + 1$ is paired with link j_{s_r} where the last real reader, with position s_r, is located. Figure 13.2 depicts a route r with five links, that is $j = 1, 2, 3, 4, 5$. Two virtual links b and t are added to the route, and two virtual readers are assumed to be located on them. Suppose $s_r = 3$ readers are located on the route, on links one, three, and five. Then, the monitored portions of the route are: r_{b1}, r_{13}, r_{35}, and r_{5t}. The virtual link b is paired with link one which is also paired with link three, which in turn is paired with link five, and so on.

Finally, if the i-th reader of the route is located on link j and the $(i+1)$-st reader on the route is located on link h, then we denote by $\varepsilon^r_{jh} = \varepsilon^r_{j_i j_{i+1}}$ the benefit associated with the monitored section r_{jh} (see example in Fig. 13.2). The total benefit accrued with the given locations of the readers is then:

$$\varepsilon = \sum_{r \in R} \sum_{i=0}^{s_r} \varepsilon^r_{j_i j_{i+1}} \tag{13.30}$$

To introduce model $AVIL_6$, which seeks to locate a limited number of readers to maximize (13.30), some additional notation is needed. Given a route r and a link $j \in r$, the set F^r_{j+} is the set of links of the route *after* link j where a reader can be located, while the set F^r_{j-} is the set of links of the route *before* link j where a reader can be located. The set F^r_{jh} is the set of links of route r between link j and link h where a reader can be located. Hence, for example, if we consider the route in Fig. 13.2, the set $F^r_{1+} = \{2,3,4,5,t\}$, the set $F^r_{1-} = \{b\}$ and the set $F^r_{25} = \{3,4\}$. Finally, let us define the binary variable y^r_{jh} to be equal to 1 if the portion of route r between two readers located on link j and link h is monitored and 0 otherwise. The mathematical formulation for $AVIL_6$ is as follows (Li and Ouyang 2012):

$$AVIL_6\text{: Max} \sum_{r \in R} \sum_{j \in F^r_{t-}} \sum_{h \in F^r_{j+}} y^r_{jh} \varepsilon^r_{jh} \tag{13.31}$$

$$\text{s.t.} \sum_{j \in A} z_j \leq k \tag{13.32}$$

$$\sum_{h \in F^r_{j+}} y^r_{jh} = z_j \quad \forall\, j \neq t \in r \quad \forall\, r \in R \tag{13.33}$$

$$\sum_{j \in F^r_{h-}} y^r_{jh} = z_h \quad \forall\, h \neq b \in r \quad \forall\, r \in R \tag{13.34}$$

$$z_b = z_t = 1 \tag{13.35}$$

$$z_j \in \{0,1\} \;\; \forall \;\; j \in A \tag{13.36}$$

$$y^r_{jh} \in \{0,1\} \;\; \forall \;\; h, \;\; j \in r \;\; \forall\, r \in R \tag{13.37}$$

Constraint (13.32) imposes that the total number of readers to be installed be less than or equal to k. Constraints (13.33) pair link j of route r where a reader is located (that is, when $z_j = 1$), with only one link h in the route that follows j. Constraints (13.34) pair link h of route r where a reader is located (that is, when $z_h = 1$), with only one link j in the route that precedes h. Finally, the objective function (13.31) maximizes the total benefit accrued by the location as defined by (13.30). For possible examples of how benefit factors ε^r_{jh} can be defined, the reader can refer to Li and Ouyang (2012).

Models $AVIL_i$, $i = 1, 2, \ldots, 6$ were introduced and developed in the last few pages to indicate how these models have been evolving, introducing more and more constraints and objectives. The first model (model $AVIL_1$) assumes that there is a single route connecting each OD pair, which happens only in special sparse networks, for example in rural areas. Model $AVIL_2$ (which minimizes the total number of readers to be located on the network to ensure that all OD routes are observed) and model $AVIL_3$ (which maximizes the total number of OD routes observed with a limited numbers of available readers), should be applicable in the cases when the observation effort does not increase as the volume of traffic on the network changes, and hence applicability is limited. On the other hand, when the traffic manager or the transportation planner wants to monitor all the traffic on the network or to maximize the amount of traffic monitored, then model $AVIL_4$ is more appropriate.

Distances among two located readers is not considered in any of the Models $AVIL_i$, $i = 1, \ldots, 4$. In $AVIL_5$ this measure is defined as the total vehicle-miles monitored. This is generalized in $AVIL_6$ where a general benefit factor is considered for an OD route monitored, which could be some function of travel distance, travel time, or some other measure of benefit to the planner or manager. The application described next focuses on the travel time reliability of the OD routes monitored, where here "reliability" essentially means how well the travel time can be predicted. If there is no variance in travel time, then it is highly predictable, whereas when the travel time has large variance its predictability is low. Installing sensors in the latter case will help the traffic manager predict travel times better as real-time measurements are made.

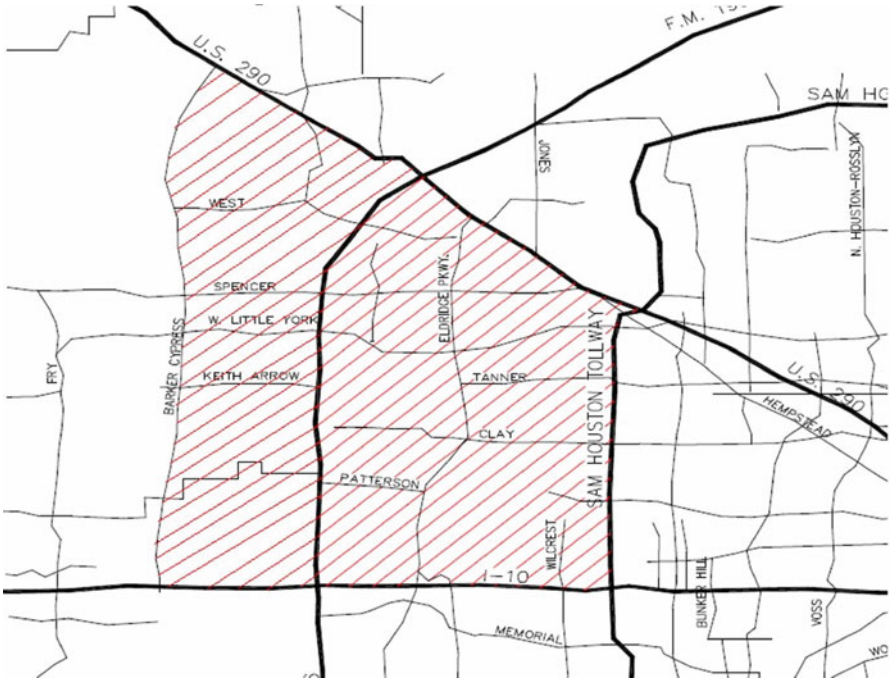


Fig. 13.3 Study area of Harris County, TX

13.5 Applications

In this section we discuss two applications of model $AVIL_5$ to a portion of the Harris County (Texas) network (see Fig. 13.3). One of the authors was working with the County on a traffic management research project, and as a by-product they were interested in if they could use toll tag readers to monitor traffic in a certain area. Consequently, we developed two different models for (*i*) the problem of optimally locating the toll tag readers (these are effectively *AVI* readers) on the arcs of the network to maximize the number of vehicle-miles (or vehicle-travel times) monitored (model $AVIL_5$ presented in the previous section) and (*ii*) the problem of optimally locating *AVI* readers on the arcs of the network to maximize the ability to predict travel times on the network. The motivation underlying the first problem is for traffic managers to have an idea of the traffic volume in the region at any given time. The motivation underlying the second problem is to monitor travel times and communicate to commuters some predictions of the mean travel times of their routes. In this case, the underlying model is assumed to be stochastic. That is, each arc (or route segment) will have a mean travel time that changes slowly over time during the day, while the actual travel time has a further fluctuation which is due to differences in driving characteristics of the commuters; this latter fluctuation may be treated as some sort of additive noise. In particular, let τ_j denote the travel time on link j. τ_j

is therefore a random variable with mean μ_j and variance σ_j^2. This variance σ_j^2 is assumed to be constant over time and known for all links in the network. It is also assumed that for any time period, the a priori distribution of μ_j is known, which is Normal with mean η_j and variance γ_j. Therefore, the variance of travel time on link j has two components, that is, $var(\tau_j) = \sigma_j^2 + \gamma_j$. For the short period of time when travel time data is collected, the expected travel time on link j is η_j. Knowing the values of η_j and var (τ_j), a confidence interval for τ_j can be obtained. If n observations of, the average value of the observed travel times (denoted $\bar{\tau}_j$) are collected, then according to the Central Limit Theorem $\bar{\tau}_j$ can be approximated by a normal distribution with mean μ_j and variance $\frac{1}{n}\sigma_j^2$. With an observation of $\bar{\tau}_j$, the a priori distribution of μ_j can be updated by using Bayesian statistical theory. The updating can be performed using the following formulas:

$$\hat{\eta}_j = \frac{\frac{1}{n}\eta_j\sigma_j^2 + \bar{\tau}_j\gamma_j}{\frac{1}{n}\sigma_j^2 + \gamma_j} \tag{13.38}$$

$$\hat{\gamma}_j = \frac{\frac{1}{n}\gamma_j\sigma_j^2}{\frac{1}{n}\sigma_j^2 + \gamma_j} \tag{13.39}$$

Therefore, the expected value of τ_j is $\hat{\eta}_j$ and $var(\tau_j) = \sigma_j^2 + \hat{\gamma}_j$. Note that the variance of μ_j is reduced after the updating:

$$\Delta\gamma_j = \gamma_j - \hat{\gamma}_j = \frac{\gamma_j^2}{\frac{1}{n}\sigma_j^2 + \gamma_j} \tag{13.40}$$

Equation (13.40) shows that the larger the sample size (n) of τ_j, the more the variance is reduced and, hence, the more reliable the travel time estimate. The above analysis can be extended to an entire route r (or a portion of it) under the assumption of independence among the travel times on different links of the same route. Consider the portion r_{jh} of route r between link j and link h and assume *AVI* readers are located both on link j and link h. The route travel time $\tau_{r_{jh}}$ on such a portion is then a random variable with mean $\mu_{r_{jh}} = \sum_{i \in r_{jh}} \mu_i$ and variance $var(\tau_{r_{jh}}) = \sum_{i \in r_{jh}} \sigma_i^2 + \sum_{i \in r_{jh}} \gamma_i$. The updating on the distribution of $\mu_{r_{jh}}$ would lead to the computation of the reduction of variance:

$$\Delta\gamma_{r_{jh}} = \gamma_{r_{jh}} - \hat{\gamma}_{r_{jh}} \tag{13.41}$$

where $\gamma_{r_{jh}}$ and $\hat{\gamma}_{r_{jh}}$ are the *a priori* and the *a posteriori* variances of the $\mu_{r_{jh}}$.

Model $AVIL_5$ can be applied to solve the problem of locating a given number k of *AVI* readers on the links of a network to maximize the reliability of travel time estimates. The resulting model formulation, $AVIL_{5'}$, is:

$$AVIL_{5'}\text{: Max } \frac{1}{2}\sum_{r \in R}\sum_{i \in r}\sum_{j \in r} y_{ri} + z_{rj}\Delta\gamma_{r_{ij}} \tag{13.42}$$

s.t. (13.21) – (13.29).

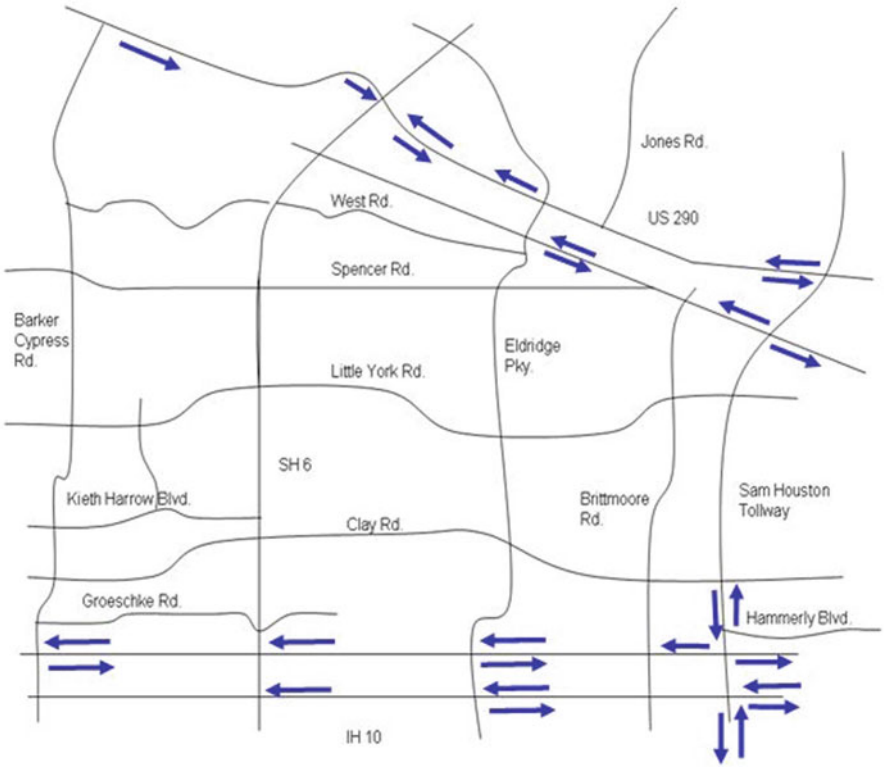


Fig. 13.4 Existing AVI Readers in the study area of Harris County, TX

13.5.1 The Study Area

The study area is bounded by Barker Cypress Road, Interstate 10, US 290, and Sam Houston Tollway (represented by the shaded area in Fig. 13.3) in Texas, USA, with 470 total arcs and 230 total nodes. The total number of OD pairs is 930, which amounts to a total of 1700 routes, 12,075 miles, 90,020 vehicles and 459,234 vehicle-miles. There are already a number of AVI readers installed on the traffic network. Figure 13.4 shows the approximate location and direction of those devices. It can be seen that most existing vehicle identification devices are located on freeways and highways such as Interstate 10, US 290, and Sam Houston Tollway.

13.5.2 Results

Figures 13.6, 13.7, and 13.8 give the optimal locations for 5, 10 and 15 additional *AVI* readers, respectively, obtained when solving $AVIL_5$ to maximize the total

Table 13.2 Performance of *AVI* reader location strategies in the study area of Harris County, when maximizing total vehicle-miles

Additional *AVI* readers installed	Total vehicle-miles		Total monitored flow		Total route length	
	Monitored	% of total	Monitored	% of total	Monitored	% of total
0	181,776	39.6	50,296	55.9	2236	18.5
5	291,040	63.4	61,764	68.6	3040	25.2
10	348,973	76.0	64,940	72.1	4183	34.6
15	380,530	82.9	68,818	76.4	4986	41.3

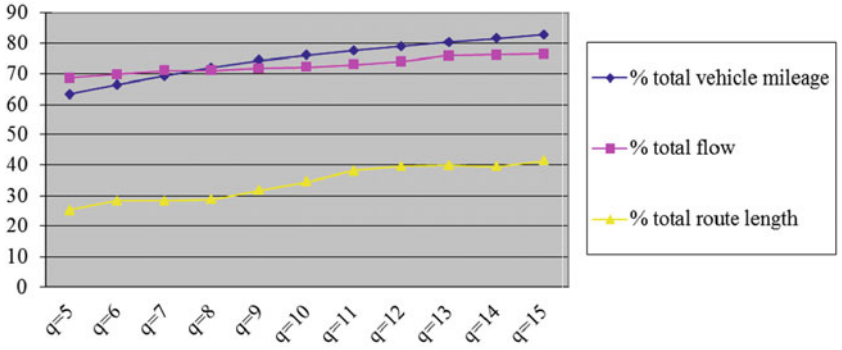


Fig. 13.5 Total benefit of the new sensors installed, from 5 to 15 sensors

vehicle-miles monitored (that is, the objective function (13.20)) by using *CPLEX* 7.0. Performance results for installing different numbers of additional *AVI* readers are given in Table 13.2 where: (a) the first column reports the total number of additional *AVI* readers installed; (b) the second and third columns report the vehicle-miles monitored and the percentage with respect to the total, respectively; (c) the fourth and fifth columns report the total monitored flow and the percentage with respect to the total, respectively; (d) and the two last columns show the route length monitored and the percentage with respect to the total, respectively.

We can observe (refer to Table 13.2) that the performance of the existing readers is such that only 39.6 % of the total vehicle mileage, 55.9 % of the total flow and 18.5 % of the total route length is monitored. By locating additional *AVI* readers, these percentages, obviously, increase. However, we can notice that the marginal increase in the total vehicle-miles monitored decreases with additional readers installed. Indeed, the first five additional readers produce an increase of 23.8 % points; by adding five additional readers the increase is 12.6 % points and, with 5 more additional readers, the increase is equal to 6.9 % points.

Figure 13.5 shows the total benefit of the new sensors installed, from 5 to 15 sensors. Note the decreasing slope of the %total vehicle mileage function.

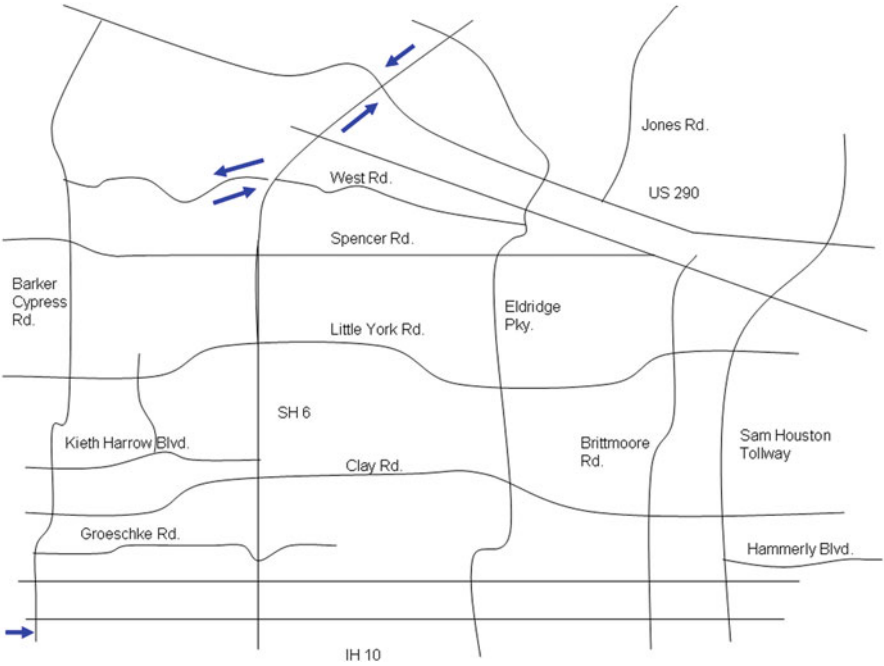


Fig. 13.6 Optimal location of 5 additional *AVI* readers in the study area of Harris County, when maximizing total vehicle-miles

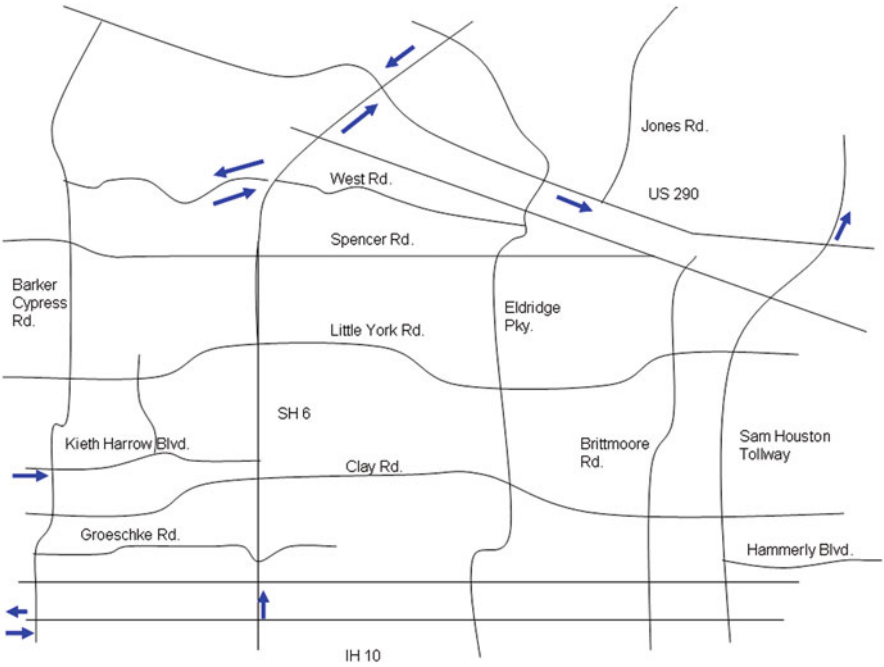


Fig. 13.7 Optimal location of 10 additional *AVI* readers in the study area of Harris County, when maximizing total vehicle-miles

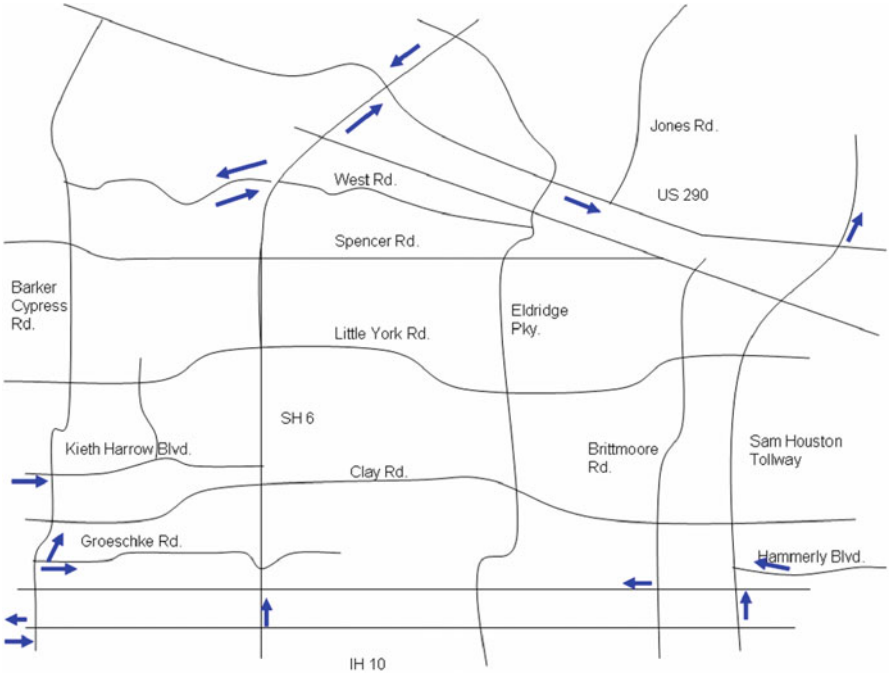


Fig. 13.8 Optimal location of 15 additional *AVI* readers in the study area of Harris County, when maximizing total vehicle-miles

Figures 13.9, 13.10, and 13.11 give the optimal locations for 5, 10 and 15 additional *AVI* readers, respectively, obtained when solving $AVIL_{5'}$ to maximize the reliability of travel time estimates (that is, the objective function (13.42)), again using *CPLEX* 7.0. We assumed that the travel time variance is proportional to travel time on that link. That is, longer travel times have larger variances. The total variance of the mean route travel times in the study area is 12,579. The traffic monitoring results of installing different numbers of *AVI* readers are summarized in Table 13.3 where the first column reports the total number of additional *AVI* readers installed, the second column reports the total reduction in variance, and the last column reports the percentage of the reduction with respect to the total variance. With the existing *AVI* readers located, the total variance reduction is 3361, which is 26.72 % of the total variance. By installing additional *AVI* readers, the reduction of the variance increases to 7578.79 with 5 additional readers, to 8997.69 with ten additional readers and to 9908.40 with 15 additional readers. Note again that the marginal decrease in the total variance decreases with additional readers installed. Indeed, the first five additional readers produce a decrement of 33.53 % points; by adding 5 additional readers the decrement is 11.28 % points and, with 5 more additional readers, the decrement is equal to 7.24 % points.

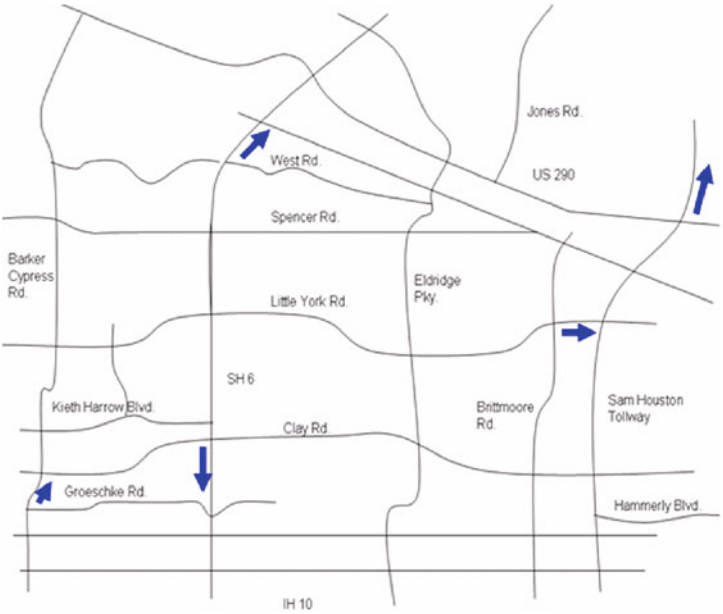


Fig. 13.9 Optimal location of 5 additional *AVI* readers in the study area of Harris County, when maximizing travel time reliability

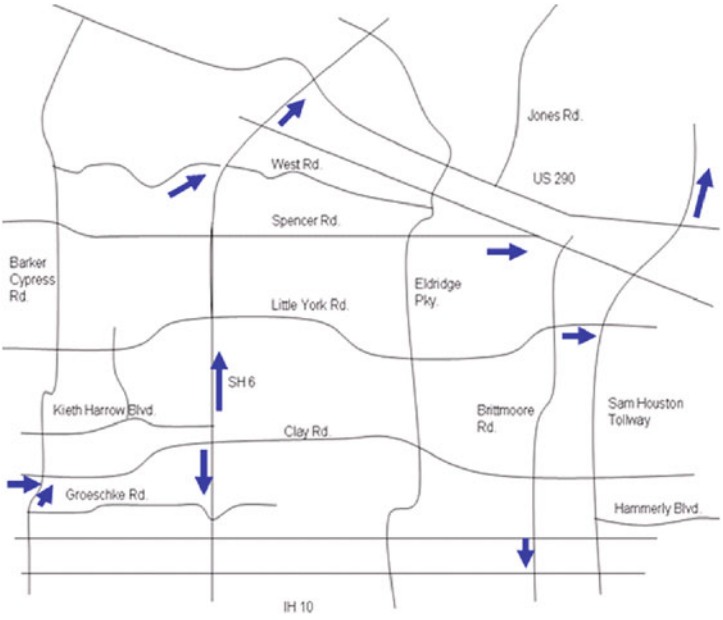


Fig. 13.10 Optimal location of 10 additional *AVI* readers in the study area of Harris County, when maximizing travel time reliability

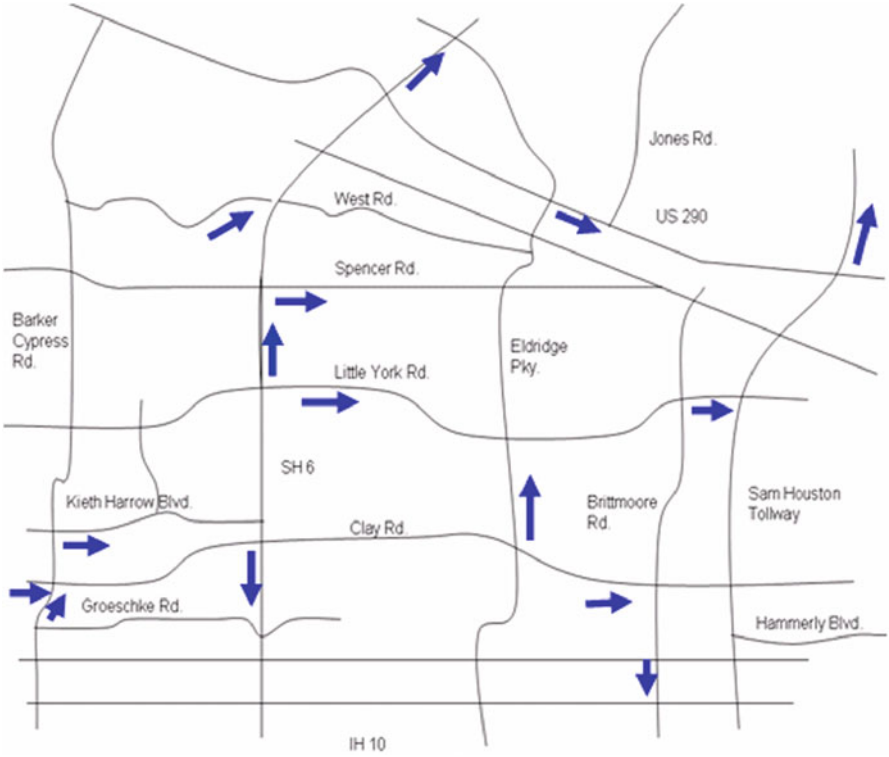


Fig. 13.11 Optimal location of 15 additional *AVI* readers in the study area of Harris County, when maximizing travel time reliability

Table 13.3 Travel time reliability results of *AVI* reader location strategies in the study area of Harris County

Additional *AVI* readers installed	Variance reduction	Percentage of total variance (%)
0	3361.08	26.72
5	7578.79	60.25
10	8997.69	71.53
15	9908.40	78.77

13.6 Conclusions

Although much has been written about sensor location problems, this chapter is the first comprehensive summary of the use of vehicle identification sensors to monitor OD route coverage, travel times and related information on traffic networks. In particular, application of such models was demonstrated for (a) maximizing the

vehicles-miles monitored and (b) maximizing the predictability of travel time estimates where the resulting variance reduction is especially useful for routes that are well traveled and *a priori* travel time variances are high.

Further uses of such sensors and attendant models by planners and traffic managers for the applications discussed, as well as future applications that one has yet to discover, attest to the value of this research and these models.

Acknowledgements The authors thank Dr. Yang He, former Ph.D. student of the second author, for the computational results. The authors acknowledge the partial support from TranStar, Houston, Texas and *USDOT* support to the *ATLAS* Center, University of Arizona. The authors also acknowledge support from National Science Foundation Grant n.1234584. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the above mentioned agencies. Finally, the authors thank the co-editors of this book for inviting them to provide this contribution for their timely book.

References

Adams WP, Sherali HD (1986) A tight linearization and an algorithm for zero-one quadratic programming problems. Manage Sci 32(10):1274–1290

Adams WP, Sherali HD (1990) A hierarchy of relaxations between the continuous and convex hull representations for zero–one programming problems. SIAM J Discret Math 3(3):411–430

Ban X, Herring R, Margulici JD, Bayen AM (2009) Optimal sensor placement for freeway travel time estimation. In: William H, Lam K, Wong SC, Lo HK (eds) Transportation and traffic theory 2009: Golden Jubilee. Springer US, Golden, pp 697–721

Bartin B, Ozbay K, Iyigun C (2007) Clustering based methodology for determining optimal roadway configuration of detectors for travel time estimation. Transp Res Rec 2000:98–105

Chan KS, Lam WHK (2002) Optimal speed detector density for the network with travel time information. Transp Res Part A Policy Pract 36(3):203–223

Chaudhuri P, Martin PT, Stevanovic AZ, Zhu C (2011) The effects of detector spacing on travel time prediction on freeways. Int J Eng Appl Sci 7(1):1–10

Chen A, Chootinan P, Pravinvongvuth S (2004) Multiobjective model for locating automatic vehicle identification readers. Transp Res Rec 1886:49–58

Church R, ReVelle C (1974) The maximal covering location problem. Pap Reg Sci Assoc 32(1):101–118

Danczyk A, Liu HX (2011) A mixed-integer linear program for optimizing sensor locations along freeway corridors. Transp Res Part B 45:208–217

Edara P, Smith B, Guo J, Babiceanu S, McGee C (2010) Optimizing freeway traffic sensor locations by clustering global-positioning-system-derived speed patterns. J Transp Eng 137(3):155–173

Fujito I, Margiotta R, Huang W, Perez WA (2007) Effect of sensor spacing on performance measure calculations. Transp Res Rec 1945:1–11

Gentili M, Mirchandani PB (2011) Survey of models to locate sensors to estimate traffic flows. Transp Res Rec 2243:108–116

Gentili M, Mirchandani PB (2012) Locating sensors on traffic networks: models, challenges and research opportunities. Transp Res C Emer Technol 24:227–255

Lam WHK, Lo HP (1990) Accuracy of o-d estimates from traffic counts. Traffic Eng Control 31:435–447

Li X, Ouyang Y (2012) Reliable traffic sensor deployment under probabilistic disruptions and generalized surveillance effectiveness measures. Oper Res 60(5):1183–1198

Liu Y, Lai X, Chang G (2006) Detector placement strategies for freeway travel time estimation. In Intelligent Transportation Systems Conference, ITSC '06. IEEE, pp 499–504

Mirchandani P, Gentili M, He Y (2009) Location of vehicle identification sensors to monitor travel-time performance. IET Intell Transp Syst 3(3):289–303

Sheffi Y (1984) Urban transportation networks: equilibrium analysis with mathematical programming techniques. Prentice Hall, Englewood Cliffs

Sherali HD, Desai J, Rakha H (2006) A discrete optimization approach for locating automatic vehicle identification readers for the provision of roadway travel times. Transp Res Part B 40:857–871

Teodorovic D, Van Aerde M, Zhu F, Dion F (2002) Genetic algorithms approach to the problem of the automated vehicle identification equipment locations. J Adv Transp 36:1–21

Chapter 14
Shape and Balance in Police Districting

Victor Bucarey, Fernando Ordóñez and Enrique Bassaletti

14.1 Introduction

Districting is a classic design problem when attempting to provide an efficient service to a geographically dispersed demand. There exists a natural trade-off between aggregation, which allows the pooling of resources, and individualization, where each individual demand has its own resources and response is as efficient as possible. Police and security providers are no strangers to this phenomenon. Having a single set of resources to satisfy the demand over an entire service area avoids the duplication of resources, coordination problems, and uneven workloads that can occur in districting. However, when the service area is large, service times at certain locations can exceed acceptable levels, making it more attractive to service this demand from distributed resources. Furthermore, districting allows for specialization of the resources to efficiently service a diverse demand in different areas. Such specialization causes additional complexity if managed from a centralized pool of resources. This creates the basic problem of separating a demand area into subregions to organize the service process.

The problem of districting in general, and police districting in particular, consists of dividing a geographical region in subregions (also referred to as districts or quadrants) in a way that improves some objective measure. Perhaps the most well known version of this problem is the political districting problem, where due to population changes political districts are frequently reshaped, sparking intense debate (Hess et al. 1965; Fleischmann and Paraschis 1988; Hojati 1996). There are however a number of other applications where districting is regularly used, including territory

V. Bucarey (✉) · F. Ordóñez
Department of Industrial Engineering, Universidad de Chile, Av. República 701, Santiago, Chile
e-mail: vbucarey@ing.uchile.cl

F. Ordóñez
e-mail: fordon@dii.uchile.cl

E. Bassaletti
Department of Criminal Analysis, Carabineros de Chile, Av. Libertador Bernardo O'Higgins 1196, Santiago, Chile

H. A. Eiselt, V. Marianov (eds.), *Applications of Location Analysis,* International Series in Operations Research & Management Science 232, DOI 10.1007/978-3-319-20282-2_14

design for sales and services, health care districting, police and emergency service districting, and districting in logistics operations (Kalcsics et al. 2005).

Important considerations when designing districts include balancing the demand for resources and creating districts that have geographical contiguity and compactness (Wright et al. 1983). Existing literature has addressed the need for demand balance constraints in districting models. In particular, not considering demand balance in the police districting problem has been shown to create difficulties such as lack of indicators to compare the performance of the different districts by police management, staff imbalance, and morale problems, among others Kistler (2009). Compactness of a district is important because response time, a key measure of quality of service, is related to the distance that has to be traversed. This makes it desirable that all points in a district be close together, as measured by travel time, to be serviced within a small response time. Therefore, travel time can be used to measure the compactness of a given district. Finally, contiguity is desirable because in addition to the need for fast response times simple districts are easier to administrate.

A natural mathematical model of the districting problem can be built on a graph representation of the geographical area and, as we describe in the literature review section, previous work has introduced such models before. However, although the requirements of balanced, contiguous and compact districts have been dealt with before, these are rarely combined in a single model. In this work we introduce such a mathematical model for a specific police districting problem of the Chilean national police force in urban areas, including the requirements of balanced, contiguous, and compact districts.

To build this mathematical model we consider a graph where nodes represent a city block (or groups of city blocks) and arcs connect adjacent nodes (blocks), or nodes that can be accessed thought the road network, see for example Fig. 14.1. In addition, this network should consider information regarding the demand of police resources at every node and information regarding distances (or travel times) on arcs between nodes. On this network structure, the districting problem becomes a graph partitioning problem. The demand balance constraints can be represented as capacity constraints on the total demand on each district, where the lower and upper bounds are calculated as a percentage deviation respect to the average demand of resources or workload. The information on distances between nodes can be used to characterize compactness and contiguity. Hence, the proposed graph partitioning model at the heart of a districting problem can take into account the features of balance, contiguity and compactness of each district.

Regrettably, introducing the requirements of balance, contiguity and compactness makes the districting problem more difficult to solve. For instance, consider the example in Fig. 14.2. This example shows two districting solutions, one for a p-median problem without balance, contiguity or compactness constraints and one for a p-median problem with balance constraints. For this problem which defines 9 districts out of 407 nodes, the solution to the p-median problem on the left is obtained in a few minutes, while the problem with balance constraints can take hours to solve exactly.

We note that the p-median districting problem results in districts that seem reasonably compact and contiguous. The reason for this is because the p-median problem forms clusters around centers so that the total distance of nodes to these

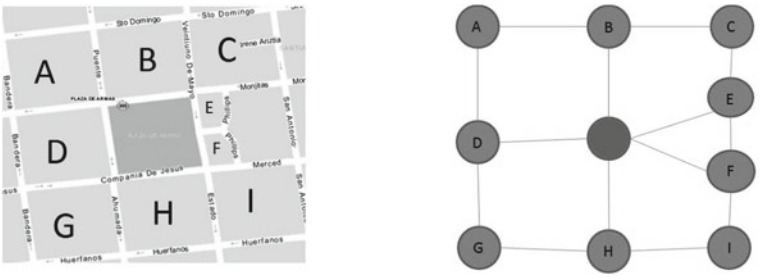


Fig. 14.1 Adjacency graph

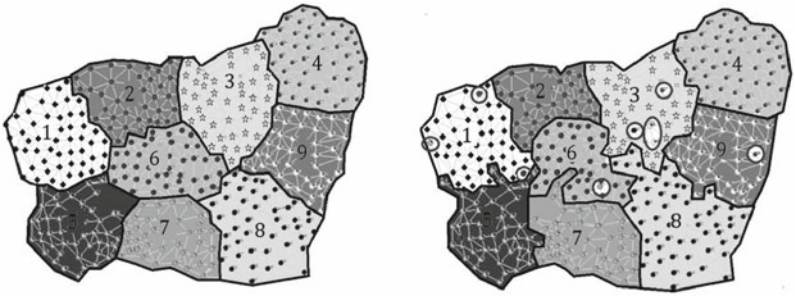


Fig. 14.2 Comparison between the results of a *p*-median problem (*left*) and a *p*-median with balance constraint of 1 % of deviation with respect to the average workload (*right*)

centers is minimized, see Daskin (2013).This total distance objective is minimized when clusters are compact and contiguous around selected centers. If the *p*-median problem has balance constraints, however, then it creates districts that are not contiguous. For example districts 4, 8 and 9 include nodes that are in the interior of district 3. It also creates districts with branches that protrude into neighboring districts, see for instance the branches out of districts 5, 7 and 8. Districts that are not contiguous or have long boundaries tend to be less compact according to existing definitions (Taylor 1973; Horn 1995).

It therefore becomes important to explicitly include contiguity and compactness constraints in optimization models where balance requirements are enforced, to obtain acceptable districts. This is particularly relevant in the police districting problem considered, because the current districts are built with a detailed manual procedure that immediately discards solutions that significantly violate contiguity and compactness requirements.

The main objective of this chapter is to present a mathematical model of a realistic police districting problem. We are inspired by the urban districting problem faced by Chile's national police force (Carabineros de Chile) in their Preventive Security Quadrant Plan (*PCSP* for its acronym in Spanish). This districting problem has been gradually implemented in Chile since 2001. By the year 2013 it had

been implemented in the 120 most populous municipalities. In this work we propose a mathematical optimization formulation of the districting problem that Chile's national police force solves as part of the *PCSP*. This optimization problem also takes into consideration balance, contiguity and compactness constraints.

In the next section we present a brief literature review on previous related work on districting problems. In Sect. 14.3 we describe the general characteristics of Chile's national police force *PCSP* program and in Sect. 14.4 we present a mathematical formulation of this particular districting problem. Section 14.5 presents our computational results evaluating the tradeoffs between the different model objectives and shows how the districting solution of the proposed model can improve current practice for an example corresponding to a real municipality in Santiago.

14.2 Literature Review

The police districting problem can be considered as a particular case of the territory design problem, which consists in grouping small geographic areas (such as city blocks) into larger geographic clusters or districts (in our case called quadrants) in such a way the clusters are acceptable according to relevant planning criteria (Kalcsics et al. 2005). Some important applications of the territory design problem are political districting (Hess et al. 1965; Fleischmann and Paraschis 1988; Hojati 1996), territory design for sales and services (Hess and Samuels 1971), and the assignment of students to public schools (Ferland and Guénette 1990; Caro et al. 2004), among others. This previous literature on territory design usually does consider at least one of the following three concepts: balance of demand, geographical contiguity, and compactness. We separate this literature review in describing how these three concepts have been addressed by previous work on territory design and in previous work on police districting. We begin by pointing out that the concepts of balance, contiguity, and compactness are present in political districting problems and touched on by the practice of "gerrymandering", which involves redesigning districts to make electoral gains (Erikson 1972; Vickrey 1961).

The concept of balance in territory planning refers to building quadrants where geographic measures or characteristics are similar across the quadrants formed. These measures or characteristics could be population (important, for instance, in political and school districting), workload or demand (important in services such as police districting), number of buildings or linear kilometers of roads, to mention some possibilities. Related to police districting, the work in Kistler (2009) considers a combined measure that takes into account the amount of historic calls, the average response time, the linear kilometers of roads, the total district area, and the population.

Geographical contiguity means that every quadrant is geographically connected. Optimal solutions to standard p-median problems with a metric or distance function between locations have contiguous quadrants. Indeed, since the objective of a p-median problem minimizes the sum of distances to each center, each node is

assigned to the center that it is closest to. Therefore, the triangular inequality implies that the line between the point and the center it is assigned to must also belong to the same quadrant, making it contiguous (Daskin 2013). This is not necessarily true for every objective and in particular, optimal solutions to the p-center problem, where the objective is to minimize the maximal distance to the corresponding center, need not be contiguous. This because nodes whose distance to more than one center is smaller than the maximal distance could be assigned to either without changing the objective. Partitions of an area formed by having each node/block assigned to the center it is closest to, for some distance metric, are known as Voronoi tessellations. As discussed above, such partitions form naturally contiguous quadrants. However, even without using a distance metric, the geographical contiguity can be enforced by explicit linear constraints. These constraints ensure that for every node assigned to a district, an entire path of nodes to that district's center are assigned to the same district. This is done for example in Williams (2002) for a land acquisition problem and in Drexl and Haase (1999) for a sales force deployment.

There are a number of different notions of compactness that are used in territory planning, which do not refer to the standard topological definition. Some of the compactness measures proposed include the ratio of the perimeter to the area, comparisons to a related compact shape such as a rectangle or circle, or moment of inertia. See Li et al. (2013) for a recent review of different compactness measures. To illustrate the variety in measures of compactness we review a few definitions that arise in some political districting examples.

For instance, Niemi et al. (1990) investigate how different measures of compactness are related to racial and partisan political discrimination in a political districting problem. These different compactness measures are treated as a multidimensional property of a district capturing: the *dispersion* and the *perimeter* of the district, and the geographic dispersion of the *population*. The *dispersion* of a district was quantified by one of four measures: the value $|L_i - W_i|$ or L_i/W_i, where L_i and W_i are the maximum length and width of district i, respectively, a comparison between the area of the district and the area of a compact shape circumscribing the district (such as square, circle or hexagon) and a measure of the distances within a district, given by the sum of the distances of every block to the center of gravity of its district. Two *perimeter* measures of a district are considered: the length of the boundary of the district (Horn 1995), and the difference of the perimeter of the district with a circumference with equal area, or conversely, the difference between the area of a district with the area of a circumference of equal perimeter. Finally, compactness measures that considered the *population* are: the moment of inertia of the population (Papayanopoulos 1973), and the ratio between the population of the district and the population existing in the maximum convex figure or circle circumscribed in the district.

A different compactness measure of a district, proposed in Theobald (1970), is the absolute deviation of the district area with respect to the average area. Young (1988) proposes eight measures of compactness of legislative districts. Many of these measures are similar to the ones listed above, except for two: a visual test, in which a human evaluator gauges the compactness of a district, and the Taylor's test

(Taylor 1973). The Taylor's test quantifies the indentations of the districts by the difference between reflexive angles and non-reflexive angles on the border divided by the total number of angles. For example a convex figure has ratio 1 (compact), and a five pointed star has ratio 0 (non compact).

Summarizing, there are many definitions of compactness measures which can be used, depending on the application considered. To the best of our knowledge, to date there is no previous work that builds compactness measures for territory planning from axioms that the measure should satisfy, even for a specific application. Identifying desirable first principles for compactness measures is an interesting area of future research that could help simplify these definitions. Compactness measures in use focus on different characteristics of districts that should be taken into account. Which characteristic is most important depends on the application being considered. For example, in political districting measures of how population is distributed in the district are important, while in territory planning for emergency service applications a compactness measure that bounds the longest distance, such as a deviations from a circle of the same perimeter, can be a desirable objective. Another consideration is the difficulty in computing the compactness measure from geographical information. Recent work has shown that a moment of inertia measure of compactness can be more efficiently computed than isoperimetric ratios (Li et al. 2013).

A different application where the shape and characteristic of an area is important is the problem of designing habitat reserves, see Marianov et al. (2008) or Church (2015) in this volume. Habitats that are suitable to hosting a given species of animals must satisfy certain shape constraints that depend on the species considered. These shape constraints can impose restrictions on compactness, size, and connectivity. Similar issues arise in the problem of harvest scheduling for environmental reasons in the forestry industry (Barahona et al. 1992).

Previous work on police districting has already considered the issues of demand balance and districts shape constraints. For instance, in Mitchell (1972) an optimization model is proposed that aims to minimize the moment of inertia of the expected workload of each area subject to balance constraints. An applied police districting model for the Buffalo Police Department is presented in Sarac et al. (1999). The paper discusses a set partitioning problem formulation and a practical approach based on census tacks. The optimization approach was abandoned because the multiple objectives (with regards to demand balance, contiguity and compactness of the districts) made the problem computationally challenging for problems of real size. This difficulty is surmounted by limiting the flexibility of the districts to existing census tracks. In D'Amico et al. (2002) the problem of police districting is considered as a graph partitioning problem subject to compactness, contiguity, convexity, and size constraints. The problem also considered a constraint which forced an upper bound on the average response time in each quadrant. This consideration leads to a non-linear approach, which is solved by local search techniques such as simulated annealing. The Constraint-Based Polygonal Spatial Clustering (*CPSC*) method is another heuristic solution method used in police districting (Joshi et al. 2009). The *CPSC* consists of designing quadrants by adding blocks to a seed block until an objective, such as a compactness score, is met. This districting method has been

used for the Charlottesville Police Department and evaluated with an agent-based model (Zhang and Brown 2013).

Another optimization model is used in Curtin et al. (2010) to determine optimal police patrol areas. This work considers a maximum covering model (Church and ReVelle 1974) to formulate a first phase problem. Then, a second phase problem determines how to distribute resources to meet a set of options with backup coverage, i.e. maximizing the blocks that are covered two or more times without considering any geographic issue or balance. In Verma et al. (2010) Voronoi Tessellations are used to design police districts testing the incremental changes in the improvement of measures such as balance of workload, 911 response time, geographical area fit, and citizen satisfactions among others.

The police districting problem considered in this chapter is represented as a modified *p*-median problem with multiple objectives and balance constraints. One of the objectives is related to the police's definition of preventive policing, which is non-linear in the quadrants formed. The combination of balance constraints and this non-linear objective cause the *p*-median problem to provide solutions that are not contiguous or compact. We therefore include a compactness measure that aims to minimize the quadrant boundaries (or nodes adjacent to a different quadrant) similar to minimizing the length of the boundary in (Horn 1995). As noted in Sarac et al. (1999), this multiple objective problem is challenging to solve for real size instances, we therefore adapt a location-allocation heuristic (Teitz and Bart 1968) to solve this problem.

14.3 The Chilean Police Quadrant Plan

The Chilean national police force, Carabineros de Chile, is the principal internal police force in Chile. Its mission includes maintaining and re-establishing order and security throughout the country and patrolling the borders. Starting in 2001 Carabineros de Chile developed its own police districting methodology for urban areas, known as the Preventive Security Quadrant Plan (*PCSP*). The objective of this plan is to define smaller districts within police precincts in order to (1) quantify the need for policing resources in each district, (2) establish a closer connection with the population by providing a point of contact and information for each district to residents, (3) organize and facilitate deployment of police resources to the community.

The process of implementing the *PCSP* in a municipality begins with the division of the precincts being analyzed into quadrants. This division is currently performed by the expert judgment of the police officers conducting the analysis. For each of the quadrants identified, different requirements for police resources are tallied and converted to a uniform policing unit, referred to as an Equivalent Unit of Vigilance (or *UVE* in Spanish). This exercise helps organizing the expected policing activities that have to be conducted in each quadrant, and dimensioning the need for policing resources in the area under consideration. The *PCSP* also considers a number of novel policing practices aimed at establishing a closer connection between the police

and area residents, reduce response time, as well as institutionalizing improvement practices.

Our work seeks to automate the design process of the *PCSP* in a given area. In particular we aim to develop mathematical models that will form optimized quadrants that can improve the utilization of the police resources. Having an aid in the planning of *PCSP* can help streamline the territory design process allowing a more frequent revision of the information in different municipalities where this districting methodology is deployed. We begin by detailing different aspects that are considered by Carabineros de Chile in the *PCSP* methodology as described in the *PCSP* manual Dirección General de Carabineros de Chile (2010). In particular, the process of dividing a precinct into quadrants takes into account the following four criteria:

- **Patrolling constraints**: There should be enough quadrants so that each quadrant can be patrolled in one shift. The police has estimated that a police vehicle patrolling a neighborhood can traverse 82 km during an 8 h shift. Therefore the number of quadrants has to be at least the total linear kilometers in the region divided by 82.
- **Geographical considerations**: Quadrants should not be divided by avenues, roads, train lines, rivers, mountains or other elements that can make it difficult to patrol.
- **Activities related considerations**: It is desirable that the design of quadrants does not partition areas of activities, such as commercial or civic districts, residential neighborhoods, etc.
- **Concentric radial design**: It is also desired that quadrants be formed following a concentric distribution around a civic quadrant or central plaza.

Of these, the first two considerations are strictly enforced by experts during the design of quadrants, while the last two are desirable features that may not necessarily be implemented.

Over each of these quadrants the *PCSP* methodology determines the amount of demand for police resources in a standardized unit of *UVEs*. Each *UVEs* corresponds to the amount of patrolling that a police car with three police officers is able to conduct in one 8 h shift. The demand for police resources in each quadrant is divided in two components: a reaction component and a prevention component. The reaction component measures the procedures and other operational functions that are fulfilled by Carabineros on average during an 8 h shift on the quadrant. This quantity is made up of four factors:

- **Deployments**: Quantifies an expected number of deployments to service events such as theft, injury, damage to property, alcohol law, disorder on public streets, drugs, traffic accidents.
- **Court orders**: The expected number of activities due to court orders, such as: arrests, detentions, notices, citations, closures, protective measures, injunctions and evictions.
- **Monitoring Establishments**: Includes all resources needed to monitor establishments such as liquor stores, restaurants, banks, gas stations, nightclubs, among others.

Table 14.1 Parameters for preventive patrolling in *PCSP*

Factor	Constant	Units
Km	82	km/*UVE*
RC	23,500	Yearly reported crime/*UVE*
Pop	50,000	People/*UVE*

- **Extraordinary services**: This factor is caused by special events that generate a non-repeated large request for police resources such as: sporting events, concerts, protests, and visits of foreign dignitaries or other authorities.

The prevention component of the demand for police resources is defined as the maximum of three factors over the quadrant: a measure of the criminal activity (amount of reported crime *RC*), the population (*Pop*), and the total kilometers of roads in the quadrant (*Km*). Each of these factors influences the need for police resources used for preventive patrols. For instance: reported crime can be used as proxy for possible future criminal activity, the population can be proportional to the level of criminal activity, and finally the total linear kilometers in a quadrant is related to the number of resources needed to patrol this distance in a fixed period. The methodology, described in the Dirección General de Carabineros de Chile (2010), considers normalizing constants, given in Table 14.1, to translate these factors in terms of *UVEs*.

Therefore, the demand for police resources for prevention according to the PCSP is given by

$$D = \max\left\{\frac{1}{82}Km, \frac{1}{23500}RC, \frac{1}{50000}Pop\right\}. \tag{14.1}$$

This demand is defined as the maximum because any one of these three factors can be the cause for additional police resources for prevention. If a quadrant has a very large population, or reported crime, or total kilometers of roads, it can require additional resources to conduct adequate preventive patrolling. For example, a large, sparsely populated district with little reported crime can have a demand for preventive police resources determined by the *Km*, while a small, densly populated district with high reported crime will have the number of police resources for prevention determined by either *RC* or *Pop*.

An automatic method of constructing quadrants would naturally assign the information regarding demand for police resources to each block, then depending on the resulting quadrants, compute the reactive and preventive components of demand over them. We note that the reactive component of demand is defined as the sum of the demand of each of the blocks. Therefore over the whole area the reactive demand remains constant, independent of how the quadrants are constructed. The same is not true for the preventive component of demand for police resources. Since this quantity is defined by a maximization over three factors, the total number of resources needed to cover a district depends on the shape of the quadrants the district is divided on.

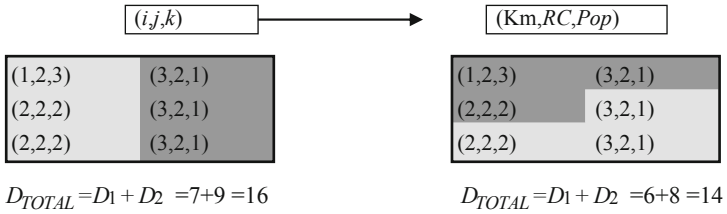


Fig. 14.3 Two ways of splitting an area with 6 nodes showing differences in the prevention demand component

To illustrate this consider Fig. 14.3. This example considers a region formed of six building blocks that can be organized in two quadrants. Each cell represents a building block, and each number represents the normalized amount of Kilometers, Reported Crime and Population. Two ways of splitting the region are represented, each district is identified by the different shades of gray and the demand for each district is calculated as the maximum between the sums of each of the three components of demand prevention. The split of the area on the left achieves a total prevention demand component of 16, while the one on the right equals 14. The example shows that the total number of resources required over both quadrants depends on how the quadrants are formed.

The demand of the district p is

$$D_p = \max \left\{ \sum_{\ell \in P} Km_\ell, \sum_{\ell \in P} RC_\ell, \sum_{\ell \in P} Pop_\ell \right\}.$$

This example shows that, in addition to automating the police districting problem for the *PCSP* methodology, it is possible to optimize the amount of resources used to provide the preventive component of demand by adjusting the shape of the quadrants formed. This is a key component of the mathematical model of the *PCSP* methodology presented in the next section and a unique feature of this districting problem.

14.4 A Police Districting Model

The model is based on a p-median model with a balance constraint in the demand of police resources. Given a set of blocks I and a set of candidates to be centers of districts $J \subseteq I$. The p-median model locates p facilities and assigns blocks to each facility, so as to minimize the sum of the distances of each block to its center. We refer to this measure as the total distance to centers. The p-median model with balance constraints considers the previous model with lower and upper bound constraints on the demand allocated to each quadrant.

In order to avoid non contiguities and "bad shapes," we consider the compactness as a two dimensional factor which minimizes the total distance to centers and the

length of the boundary of each district. To minimize the boundary of each district we define an adjacency graph in which each node represents a block, and an edge (i, j) exists if the block i share a piece of boundary with the block j, in this case we say that $i \in N(j)$ and vice-versa. We call the set $N(i)$ as the neighborhood of i. Also, in this model, constraints could be easily added that avoid dividing conflictive areas, e.g., hot spots, or quadrants being crossed by rivers or main roads. Finally we include the preventive demand component that minimizes the sum of the maximum between the reported crime level, the population, and the total kilometers of streets in each quadrant.

We use the variables of the classical p-median problem: x_j is a location binary variable which takes the value 1 if the center $j \in J$ is assigned and 0 otherwise, and y_{ij} are variables which take the value 1 if the block $i \in I$ belongs to the center is $j \in J$ and 0 otherwise. Also we use variables z_{ik} which take value 1 if the block $i \in I$ is allocated to a different center than $k \in N(i)$. This $N(i)$. This variable allows to count the length of the boundary of each district. Finally we use a continous variable D_j which takes the value of the prevention component of the demand. With these definitions the districting model is as follows:

$$\text{Min } \theta_1 \sum_{i \in I, j \in J} \ell_{ij} y_{ij} + \theta_2 \sum_{i \in I} \sum_{m=1}^{|N(i)|} \kappa_m u_{im} + \theta_3 \sum_{j \in J} D_j \tag{14.2}$$

$$\text{s.t.} \sum_{j \in J} x_j = p \tag{14.3}$$

$$\sum_{j \in J} y_{ij} = 1, \ i \in I \tag{14.4}$$

$$x_j \geq y_{ij}, \ i \in I, \ j \in J \tag{14.5}$$

$$x_j L \leq \sum_{i \in I} y_{ij} dem_i \leq x_j U, \ j \in J \tag{14.6}$$

$$z_{ik} \geq y_{ij} - y_{kj}, \ i \in I, \ k \in N(i), \ j \in J \tag{14.7}$$

$$z_{ik} \geq -y_{ij} + y_{kj}, \ i \in I, \ k \in N(i), \ j \in J \tag{14.8}$$

$$u_{im} \leq 1, \ i \in I, \ m \in \{1, \cdots, |N(i)|\} \tag{14.9}$$

$$\sum_{m=1}^{|N(i)|} u_{im} \geq \sum_{k \in N(i)} z_{ik}, \ i \in I \tag{14.10}$$

$$D_j \geq f_{pop} \sum_{i \in I} y_{ij} pop_i, \ i \in J \tag{14.11}$$

$$D_j \geq f_{km} \sum_{i \in I} y_{ij} km_i, \ j \in J \tag{14.12}$$

$$D_j \geq f_{rc} \sum_{i \in I} y_{ij} rc_i, \ j \in J \tag{14.13}$$

$$x_j, \ y_{ij} \in \{0,1\}, i \in I, j \in J \tag{14.14}$$

$$D_j, z_{ik}, u_{im} \geq 0, \ i \in I, \ j \in J, \ k \in N(i), \ m \in \{1, \ldots, |N(i)|\} \tag{14.15}$$

The objective (14.2) minimizes the multi-objective weighted function, where θ_1 is the weight of the total distance to centers (where ℓ_{ij} represents the distance of block $i \in I$ to its allocated center $j \in J$), θ_2 is the weight of the penalty function of size of the boundary and θ_3 weights the prevention demand component.

The Eq. (14.3) states the number of centers or quadrants. The set of constraints (4) states that each block must be assigned to a quadrant. The set of inequalities (5) establishes that a block could be assigned to a center only if the center is activated. The set of constraints (6) balances the demand between the quadrants. This means that, denoting dem_i the demand from block i, the total demand assigned to each center cannot be more than an upper bound $U = (1+\alpha)\sum_{i \in I} dem_i/p$ and cannot be less than the lower bound $L = (1-\alpha)\sum_{i \in I} dem_i/p$. Here, α is the maximum acceptable percentage of deviation from the average of the demand of the reaction component (e.g., 5 %). Expressions (7) and (8) define z_{ik} as the absolute value of the difference between y_{ij} and y_{kj} for every pair of adjacent blocks. By defining auxiliary variables u_{im}, the sets of constraints (9) and (10) allow to express $\sum_{i \in I} \left(\sum_{k \in N(i)} z_{ik}\right)^{\beta}$ as a piecewise linear approximation $\sum_{i \in I} \sum_{m=1}^{|N(i)|} \kappa_m u_{im}$, where $\kappa_m = \kappa_m(\beta)$ depends of the convexity of the function. These variables basically count the number of adjacent blocks not assigned to the same median, so disconnected blocks can be penalized. The set of constraints (11–13) defines D_j as the maximum of three factors: the linear kilometers km_i, the amount of reported crime for the block i, rc_i, and the population of the block, pop_i. All of them are normalized by a factor f_w. Finally, expressions (14) and (15) define the domain of the variables.

Realistic instances of this optimization problem can be difficult to solve exactly. If we consider a problem on n nodes, where each node has at most ρneighbors, then this optimization problem would have $O(n^2)$ binary variables, $O(n\rho)$ non-negative continuous variables, and $O(n^2\rho)$ inequality constraints. The realistic case study presented in this chapter considers $n = 1266$ blocks and up to $\rho = 7$ neighbors, which gives a problem with millions of constraints and binary variables. Even reducing the problem size by aggregating variables or limiting the possible centers and possible block to center assignments we obtain problems whose size are a challenge for existing exact solution methods. We therefore consider a heuristic solution algorithm based on a Location-Allocation heuristic (Teitz and Bart 1968) to find good solutions to this problem efficiently. This heuristic implemented is as follows:

1. *Step 1:* Generate an initial set of p centers. Any rule can be used to generate this set, for instance random selection or solving the p-median problem.
2. *Step 2:* (*Allocation phase*): Solve the mixed integer optimization problem that allocates every block to a given set of centers, so that the balance constraints hold, and the objective is minimized.
3. *Step 3*: (*Location phase*): For every quadrant generated, find the best geometrical center, and go back to the step 2 until there are no changes.

We repeat this heuristic with multiple random seeds or initial solutions to further search the solution space. One of the best features of this heuristic is that in every step a feasible solution is obtained. This gives several districting plans within a reasonable computational time that could be evaluated by the decision maker with other criteria, as desired. We note that this heuristic can be accelerated by solving the linear relaxation in the allocation phase, assigning the fractional blocks using a heuristic rule. This acceleration makes it more difficult to satisfy the balance constraints.

In the next section we show how this model with the heuristic and the adequate parameters of the objective function (θ, κ) gives compact and balanced districts for a real instance.

14.5 A Case Study

In this section we show the most important results of the model applied to a real instance in Ñuñoa, Santiago de Chile. The current districts have a configuration as shown in Fig. 14.4. The data of each block was extracted from demographic census data and historical data of Carabineros de Chile. This instance has a size of 1266 blocks. The current districting plan has a maximum deviation of the reactive component of demand for police resources that deviates 77 % from the average value in this region (Bustamante 2011).

This map, after a clustering of some blocks was represented in an adjacency graph as shown in Fig. 14.5. The clustering was performed with a grid of 200 $\times$ 200-m cells. Each node represents a cell of the grid that contains a block of the original map. After this clustering the graph has a size of 407 nodes.

If a p-median is used to find quadrants, the results would be nice and compact shapes. However, the maximum deviation from the workload balance could reach large figures, similar to the 77 % shown by the current quadrants. On the other hand, the p-median with balance constraints will find an optimal solution, but the shape of the districts will be irregular forming non compact quadrants. Even non contiguities could be observed in this graph.

We applied the model and the heuristic for different values of the parameters of the objective function $(\theta = (\theta_1, \theta_2, \theta_3))$ and found that the quality of solutions and the performance of the heuristic are very sensitive in this parameter. For instance, Fig. 14.6 shows the output for different parameters θ.

Fig. 14.4 Current quadrants

Figure 14.6a shows the result when the relative importance of the three factors in the objective function is the same, i.e., all three are weighted uniformly. In this case non-contiguities are observed and the shape of the districts is irregular. This situation is improved when the total distance to centers is weighted with greater relative importance, which gives a solution closer to the *p*-median problem as in the case of Fig. 14.6b. In that graph each district generated has contiguous blocks. The case of Fig. 14.6c shows the result where the size of the boundary has greater relative importance in the objective function. Due to the symmetries in the optimization model, using the proposed objective function, the allocation integer optimization problem in the heuristic becomes quite difficult to solve. We therefore fixed a 300 s limit for each iteration of the heuristic. Figure 14.6d shows the graph generated by the heuristic where the most important factor is the prevention demand. As expected, non contiguities and non compactness are present in the result. Finally we weighted with greater relative importance both the total distance to centers and size of the boundary. Figure 14.6e shows better shapes, while the balance between the quadrants is maintained. In particular the improvement of this solution over the one depicted in Fig. 14.6b is that because of the greater weight on the boundary, nodes which have many neighbors tend not to be on the boundary.

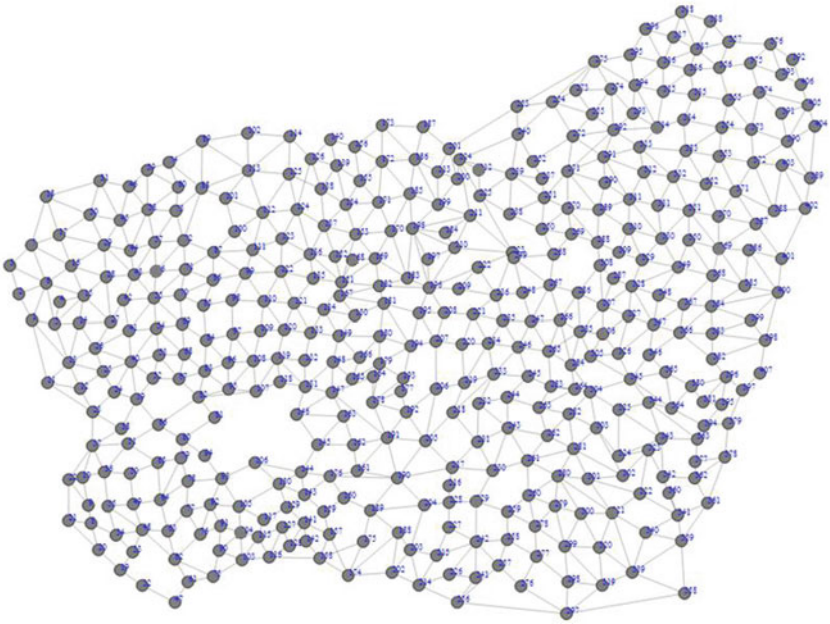


Fig. 14.5 Adjacency graph that represents Ñuñoa

We analyze these solutions further with the results presented in Table 14.2. This table shows the average solution time per iteration of the heuristic and the contributions to each part of the objective (distance to centers, boundary, prevention demand) of the five solutions presented in Fig. 14.6. We note that the total distance to centers component of the objective is not very sensitive to changes in the weights and varies at most 6 % of the mean value, while the objective components of boundary and prevention demand vary greatly (with a maximal variation of 60 % and 44 % with respect to the mean value respectively). Another important aspect is the difficulty of solving each iteration in the heuristic. When the objective is primarily driven by the size of the boundary we observe that the heuristic reaches the run time limit. This is due to the difficulty of solving the allocation step when the contribution of this nonlinear objective is significant. It is interesting to note that, for the best shapes found (experiment e), the time required was just two seconds. This solution method also provides many feasible solutions (one during each iteration) which can be evaluated afterwards by decision makers.

Finally, in Fig. 14.7 we show a new districting plan which is the output of the heuristic where the parameters of the objective function θ weight most importantly the geometrical aspects as in the experiment (e). Notice that this districting solution has a demand balance within 5 % of the mean demand value and forms contiguous and compact districts.

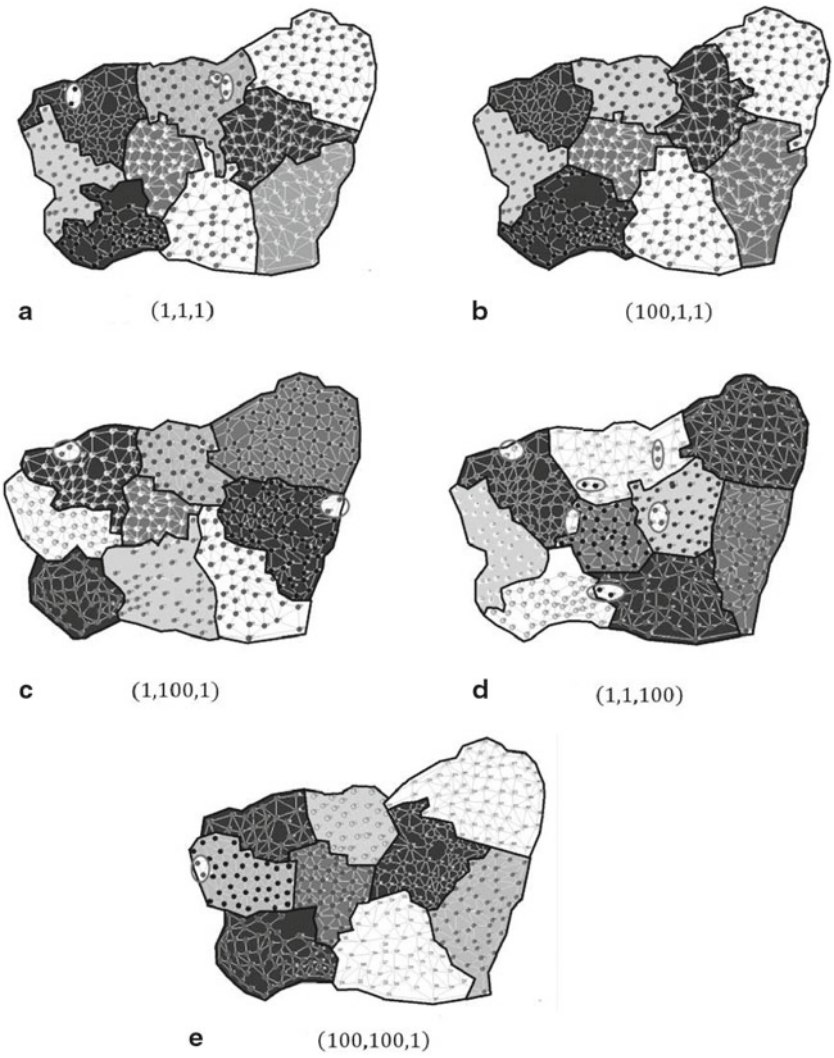


Fig. 14.6 Comparison for the output of the model for differents values of $(\theta_1, \theta_2, \theta_3)$

Table 14.2 Average resolution time and contributions of each objective component for each experiment

Experiment label	Parameters			Time [sec]	Distance to Centers $\sum d_{ij} y_{ij}$	Boundary Component $\sum \kappa_m u_{im}$	Prevention Demand $\sum \gamma_j$
	θ_1	θ_2	θ_3				
a	1	1	1	5	229,528	4308	185,895
b	100	1	1	3	222,551	4173	288,158
c	1	100	1	300	235,451	2131	221,071
d	1	1	100	18	229,635	4186	185,895
e	100	100	1	2	222,639	3350	288,326

Fig. 14.7 Districting plan given for the heuristic using $\theta = (100, 100, 1)$ and $\beta = 3$

14.6 Conclusions

Districting is a difficult problem arising in many contexts where resources must be allocated and divided to provide an efficient service to a distributed demand. In particular, police institutions are no stranger to this challenge, where the demand for police resources must accomplish both preventive and reactive actions to provide security to an area.

In this work, done in collaboration with Chile's national police force, we formulated an optimization problem to represent Carabineros' *PCSP* police districting problem. This leads to a modified *p*-median model with a preventive demand non-linear objective, balance constraints on the reactive demand component, and an objective of minimizing the boundary. The balance requirements and non-linear preventive demand objective not only make the districting problem much harder, but give a problem that has optimal solutions that do not have contiguous or compact quadrants. Adding the compactness measure that minimizes the boundary to this multi-objective problem helps provide solutions that achieve efficient objectives with acceptable quadrant shapes.

The proposed model and location-allocation heuristic provide several efficient solutions that a planner can compare to trade-off additional less tangible criteria in selecting the optimal districting plan. We applied the heuristic to a real case in Santiago, Chile, and obtain districts with a reactive demand imbalance of at most 5 % from the average quadrant demand, as opposed to an imbalance of 77 % obtained by the current manual solution. This benefit comes at the expense of districts which have more complicated boundaries, which become operationally more challenging to manage. Improvements of this model should include additional constraints on the boundaries to achieve quadrant shapes that are simple to patrol. Ongoing work includes evaluating the use of this methodology in other precincts of Santiago with real police data and adding geographical and operational considerations of what nodes can be used as boundaries of quadrants. Lieutenant Colonel Bassaletti, who participated in this work, has stated that this model and solution method are part of the techniques that the Chilean national police force is taking into account in their ongoing revision of the *PCSP* districting plan.

Acknowledgments This work was partially supported by Conicyt through grant ACT87.

References

Barahona F, Weintraub A, Epstein R (1992) Habitat dispersion in forest planning and the stable set problem. Operations Res 40(Suppl 1): S14–S21

Bustamante S (2011) Metodología para el rediseo de los cuadrantes utilizados por carabineros de chile en el plan cuadrante de seguridad preventiva. Masters thesis. Universidad de Chile, Facultad de Ciencias Físicas y Matemáticas, Santiago, Chile

Caro F, Shirabe T, Guignard M, Weintraub A (2004) School redistricting: embedding GIS tools with integer programming. J Oper Res Soc 55(8):836–849

Church R, ReVelle C (1974) The maximal covering location problem. Pap Reg Sci 32(1):101–118

Curtin KM, Hayslett-McCall K, Qiu F (2010) Determining optimal police patrol areas with maximal covering and backup covering location models. Netw Spat Econ 10(1): 125–145

Church RL, Niblett MR, Gerrard RA (2015) Modeling the Potential for Critical Habitat, pp 155–172. Applications of Location Analysis, Springer NY

D'Amico SJ, Wang S-J, Batta R, Rump CM (2002) A simulated annealing approach to police district design. Comput Oper Res 29(6):667–684

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

Dirección General de Carabineros de Chile (2010) Nuevo Manual Operativo del Plan Cuadrante de Seguridad Preventiva. Santiago, Chile
Drexl A, Haase K (1999) Fast approximation methods for sales force deployment. Manage Sci 45(10):1307–1323
Erikson RS (1972) Malapportionment, gerrymandering, and party fortunes in congressional elections. Am Polit Sci Rev 66(4):1234–1245
Ferland JA, Guénette G (1990) Decision support system for the school districting problem. Oper Res 38(1):15–21
Fleischmann B, Paraschis JN (1988) Solving a large scale districting problem: a case report. Comput Oper Res 15(6):521–533
Hess SW, Samuels SA (1971) Experiences with a sales districting model: criteria and implementation. Manag Sci 18(4):41–54
Hess SW, Weaver JB, Siegfeldt HJ, Whelan JN, Zitlau PA (1965) Nonpartisan political redistricting by computer. Oper Res 13(6):998–1006
Hojati M (1996) Optimal political districting. Comput Oper Res 23(12):1147–1161
Horn MET (1995) Solution techniques for large regional partitioning problems. Geogr Anal 27(3):230–248
Joshi D, Soh L-K, Samal A (2009) Redistricting using heuristic-based polygonal clustering. In Proceedings ICDM'09. Ninth IEEE International Conference on Data Mining, pp. 830–835, IEEE
Kalcsics J, Nickel S, Schröder M (2005) Towards a unified territorial design approach: applications, algorithms and GIS integration. Top 13(1):1–56
Kistler A (2009) Tucson police officers redraw division boundaries to balance their workload. Geogr Public Saf 1(4):3–5
Li W, Goodchild MF, Church R (2013) An efficient measure of compactness for two-dimensional shapes and its application in regionalization problems. Int J Geogr Inf Sci 27(6):1227–1250
Marianov V, ReVelle C, Snyder S (2008) Selecting compact habitat reserves for species with differential habitat size needs. Comput Oper Res 35(2):475–487
Mitchell PS (1972) Optimal selection of police patrol beats. J Crim Law Crim Police Sci 63(4):577–584
Niemi RG, Grofman B, Carlucci C, Hofeller T (1990) Measuring compactness and the role of a compactness standard in a test for partisan and racial gerrymandering. J Polit 52(4):1155–1181
Papayanopoulos L (1973) Quantitative principles underlying apportionment methods. Ann NY Acad Sci 219(1):181–191
Sarac A, Batta R, Bhadbury J, Rump C (1999) Reconfiguring police reporting districts in the city of buffalo. OR Insight 12:16–24
Taylor PJ (1973) A new shape measure for evaluating electoral district patterns. Am Polit Sci Rev 67(3):947–950
Teitz M, Bart P (1968) Heuristic methods for estimating the generalized vertex median of a weighted graph. Oper Res 16:955–961
Theobald HR (1970) Equal representation: a study of legislative and congressional appointment in Wisconsin, pp 71–260. The State of Wisconsin Blue Book, Wisconsin
Verma A, Ramyaa R, Marru S, Fan Y, Singh R (2010) Rationalizing police patrol beats using Voronoi tessellations. In Proceedings of 2010 IEEE International Conference on Intelligence and Security Informatics (ISI), pp 165–167
Vickrey W (1961) On the prevention of gerrymandering. Polit Sci Quart 76(1):105–110
Williams JC (2002) A zero-one programming model for contiguous land acquisition. Geogr Anal 34(4):330–349
Wright J, ReVelle C, Cohon J (1983) A multiobjective integer programming model for the land acquisition problem. Reg Sci Urban Econ 13:31–53
Young HP (1988) Measuring the compactness of legislative districts. Legislative Stud Quarterly 13(1):105–115
Zhang Y, Brown DE (2013) Police patrol districting method and simulation evaluation using agent-based model & GIS. Secur Inf 2:7

Chapter 15
Location and Sizing of Prisons and Inmate Allocation

Vladimir Marianov

"People shouldn't live better in jail than they do on the outside. Here in my jails, they don't."
—Joe Arpaio, Sheriff, Maricopa County, AZ.

15.1 Introduction

According to the International Centre for Prison Studies (Walmsley 2013), 10.2 million people were held in jails and prisons worldwide in 2013. The total world population at the beginning of 2013 was of 7.1 billion, which means that 144 people out of every 100,000 were prison inmates. The US not only has the largest inmate population in the world, according to those figures (2.24 million), but it also has the highest prison population rate (716 people per 100,000 inhabitants.) The Russian Federation has a healthy figure of 475 inmates per 100,000, and the lowest rates (with the exception of San Marino, with 6,) are found in some European Northern countries (58–88), Japan (51), and India (30). Some Western African countries have remarkably low rates (28–75). Chile has a rate of 266 per 100,000.

As these figures show, unless incarceration ceases to exist—a not very likely event—societies cannot live without penal facilities. Although their basic function is debated (it is not clear whether it is to punish, keep inmates from committing more crimes, maintaining them apart from society, reeducate them, or several of the above), there is agreement in considering jails and prisons indispensable. Some studies even conclude that they boost the economy of the areas in which they are located (Cherry and Kunce 2001). Such positive externalities are disputed by other authors, though (Hooks et al. 2010; King et al. 2003.)

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.), *Applications of Location Analysis,* International Series in Operations Research & Management Science 232, DOI 10.1007/978-3-319-20282-2_15

Some definitions are required at this stage. The term "jail" refers to places where people are awaiting or under trial or serving sentences that last for less than 1 year. A prison, or correctional, or penitentiary[1], is a place where convicted felons serve their sentences. According to the National Institute of Corrections (Hall 2006), due to their nature, a jail turns over its beds far more frequently than a prison: 36 times a year against 0.75 times a year in average. Our application refers to penal facilities that fulfill both requirements.

There are high, medium and minimum security closed prisons, from where inmates are not allowed to leave at any time. However, in some countries, industries may contract inmates as employees. For example, the program Federal Prison Industries ("Unicor") provides prison labor in the US. (NBC 2012). In New Zealand, there is a policy oriented to provide jobs to inmates with the involvement of the private sector (New Zealand Department of Corrections 2001) Some of these prisons could have forced labor, although this is a decreasing tendency[2]. There are also open prisons, allowing prisoners to be free to move with minimum supervision, and even to leave. These include training, psychiatric and drug abuse treatment centers. Finally, there are the halfway houses, intended for rehabilitation of inmates who have recently been released from prison and are in the process of rehabilitation. In this chapter we do not deal with open prisons or halfway houses, which, from the point of view of location, have requirements that are different from those of prisons, since they tend to require some contact by inmates with people in the surrounding community. Interested readers are referred to Johnson (2006).

Penal facilities are generally considered undesirable. *NIMBY* (not in my back yard), *NIMTO* (not in my term of office), *LULU* (locally unacceptable land use), *BANANA* (build absolutely nothing anywhere near anything) are among the terms that have been used to denote facilities that people do not want to be located in their neighborhood. Landfills, jails and prisons, polluting industries and power plants are examples of such facilities. Among locally undesirable facilities, penal institutions have one very distinctive feature: they house people who do not like being there, and most of the people in their surroundings do not like to have them around, either. This feeling is not surprising: being locked with other inmates with not-so-clean records and not being able to leave is not the best state of affairs for anybody. For neighbors, living close to a place full of delinquents generates some uneasiness due to the likelihood of escapes, decline in property values and the constant presence of visitors who are relatives, friends and colleagues of those inside.

However, not everybody wants prisons and jails located far away: families of the inmates, especially of those sentenced to long stays, naturally prefer closeness to the place where their relatives are, so they can visit often. Facility employees and service providing personnel want to be close to their workplace, which is relevant

[1] Or papa's house (Comfort 2006).

[2] In reference with forced labor and other therapies, Joe Arpaio declares: "I want to make this place so unpleasant that they won't even think about doing something that could bring them back." Inmates in his prisons do work.

considering that, in 2005, there was one employee per each 3.2 inmates in US federal correctional facilities (Stephan 2008).

The penal facility location problem could include different decision levels, which are usually addressed separately. For a federal prison, a first level considers the decision about the state, province and county in which the prison will be located (Marianov and Fresard 2005; Hernández et al. 2012). At this level, the location factors are related mainly to transportation costs and, in minor degrees, with public opposition. After the first-level decision has been made, a second level concerns the decision about the specific lot in which the prison will be built. At this level, the main issues are availability and cost of appropriate and accessible land, as well as existence and cost of utility services; public concerns and environmental impact. A good description of the issues can be found in Ricci (2006), specific criteria have been developed by a number of US counties, and are available on the web, and an interesting example can be found in Chan (1996). A good literature review up to that date, on impacts of a jail on a community, can be found in Fehr (1995).

A third level includes the architecture or layout design of the facility (Chan 1996). When locating a federal prison, the three levels must be addressed, but the problem of locating a county jail includes only the second and third levels. An example of the site selection process for county jails can be found in San Mateo County Sheriff's Office (2009). A very good description of these levels and the issues in prison location can be found in Engel (2007).

The academic literature on jail and prison location is very scarce, even though mathematical methods have been used in crime and justice problems (Maltz 1996). One of the few articles related to the subject is Korporaal et al. (2000), who propose a model for assessing the needs for prison capacity in the Netherlands. They formulate a queuing model of the prison population, which can be solved for different scenarios of capacity expansion of the system, computing in each the blocking probability (probability of inmates sent home for lack of capacity). However, there is a number of reports, manuals and leaflets developed by practitioners and users, about criteria and considerations to be taken into account when locating a prison. Rather than reviewing this literature, we describe the problem, and provide an example of penal facility location problem, based on a real case in Chile in the early 2000s. In our application we locate multiple facilities, and we include only the first decision level, i.e., the determination of what province or county should the facilities be located in. Additionally, we determine the optimal expansion schedule for existing facilities and allocate inmates to facilities. In the application, the facilities fulfill the needs for jails and prisons.

15.2 Issues in a Prison Location Problem

15.2.1 Location Criteria

As with other undesirable facilities, most of the public would want prisons and jails as far as possible from their own locations. This criterion has indeed been used in the past. Examples are Siberia for Russian inmates and Australia for English criminals. This criterion is still valid for large prisons, up to a certain point, and in some countries (Moran et al. 2011; Lawrence and Travis 2004). However, as mentioned above, prisons and jails need to be reachable by inmates' families (contact with family is one of the determinants of rehabilitation), and by people in charge of custody and services (lawyers, drug rehabilitation personnel, prison personnel, catering, health services, etc.) As jails house inmates under trial or expecting it, they need to be not too far from courts, a requirement not necessary for prisons. In consequence, there is a tradeoff between public concerns and operating (transportation) cost, and the solution in each case is highly dependent on the context and budget availability. For instance, in the case of a county jail, location in a rural area keeps the jail far from the bulk of population, which is concentrated in urban areas, but also far from families and courts and, in most cases, requires an extension of services to the location. An urban location is closer to courts but requires more expensive land, poses security risks, and faces more opposition.

15.2.2 Planning Horizon and Population Forecast

Locating a jail or prison is a strategic matter, i.e., the planning needs to consider a long term horizon, since jails cannot be easily moved[3]. Consequently, prison population must be forecast as accurately as possible, using the historical trends in prison population, past and present crime trends and other influencing factors, as well as all the uncertainties: future changes in incarceration policies, new crime trends, changes in police efficiency and effectiveness, and so on. Modern methods of population forecast are reviewed in Berk (2008), a recent method used in New South Wales, Australia is described in Wan et al. (2013), while other examples of forecast can be found in de Silva et al. (2006), Scalia (2004) and Ministry of Justice UK (2013). Practical considerations in forecasting can be found in *NAATAP* (2014).

[3] With the exception of the (in-) famous Tent City (for different points of view on it, access http://www.mcso.org/MultiMedia/PressRelease/Tents%20Birthday.pdf and http://www.motherjones.com/politics/2013/05/10-worst-prisons-america-joe-arpaio-tent-city), established by the Maricopa County sheriff Joe Arpaio, whose quotes decorate this chapter. (http://thinkexist.com/quotes/joe_arpaio/).

15.2.3 Overcrowding in Prisons

Many prisons in the world suffer from overcrowding or overpopulation. According to the International Centre for Prison Studies[4], in 2013, in Haiti there were more than four inmates locked in a space meant for one (an occupancy level of 416.3 %). In Venezuela, almost three people were sharing the space built for one (270.1 %). In Italy, the occupancy level was of 121.6 %. In Chile, 111 %. Although in the US as a whole, the occupancy level in 2013 was officially 99 %, in California there was an occupancy level of 146 %[5]. Overcrowding is due to insufficient prison space and it has several causes: fluctuating or unforeseen crime rates; changes in laws; and increased public pressure on law enforcement bodies and courts, that makes them act more harshly than expected by planners; new offenses and, last but not least, the delay between the time at which prison population starts exceeding the capacity and the time the decisions are made and actions required taken for the opening of new prisons.

The effects of overcrowding are increasing inmate misconduct; insufficient services such as education and drug treatment programs; and decreasing safety and security of inmates and personnel (US Government Accountability Office 2012).

It is out of the scope of this chapter to analyze the causes and effects of overpopulation and the required measures to decrease it. However, from the point of view of prison planning, it is reasonable to expect some overpopulation because of the uncertain nature of future population, and because, if large penal facilities are to be built, it may not be reasonable to build 1 with a capacity of, say, a 1000 inmates, as soon as the prison population exceeds the current capacity by 1 inmate.

15.2.4 Dealing with Public Opposition

Being undesirable facilities, prisons generate public opposition (Martin and Myers 2005.) As with other locally undesirable infrastructure, there are basically two policies when it comes to dealing with public concerns. The first is the authoritarian approach, consisting of maintaining secrecy about the location of the facility, and once the location is decided, deal with the public by defending the location. This approach has been termed *DAD* (Decide—Announce—Defend) by Kleindorfer and Kunreuther (1994), see Eiselt and Marianov (2015). Naturally, elected officials do not like this policy, as it may lead to losing a re-election.

A second policy considers including the public in the decision, and compensating those most affected by the facility. A nice example of this policy can be found in Armstrong (2012), and a discussion of the different policies in Farkas (1999).

---

[4] See http://www.prisonstudies.org/highest-to-lowest/occupancy-level?field_region_taxonomy_tid=All.

[5] See http://www.economist.com/blogs/graphicdetail/2013/08/daily-chart.

15.2.5 Location Models

Different levels of decision use different models. The location of a large federal prison (which we termed a first level decision) could be decided by an optimization model in which the cost is the main objective (land acquisition, building, opening, operating, transportation, etc.), and will consider also other lower-level objectives, e.g., minimization or penalization of overcrowding and penalization of travel distances covered by all the involved personnel and service providers, as well as by visitors of the inmates. Note that this is a cost that is not incurred by the government, but by providers, employees and relatives; a social cost. The constraints must include satisfaction of the forecast demand, opening and closing schedule over the horizon, number of facilities to be built, and others.

At the second level (determination of the specific community and the lot on which the prison will bebuilt), the decision is usually made by comparing the alternative sites one by one, using a multiobjective approach as the analytic hierarchy process (Saaty 1980) or *PROMETHEE* (Brans and Vincke 1985). On this level, costs are again important, but other objectives are also taken into account: neighborhood socioeconomic characteristics as level of poverty, crime rate and unemployment rate; site attributes including development cost, availability of services and transportation; distance to the courts; environmental and public opposition issues, etc.

15.2.6 Offender Flow and Inmate Segregation/Classification

Although the organization of the crime treatment system differs between countries, the offender flow follows similar stages everywhere: once a crime is committed and the offender is arrested, a short period of holding in a police premise is the usual first stage. The second stage corresponds to an initial hearing, after which there is either no case and the suspect is released, or there is a case, in which situation the suspect is detained in jail or released on bail. A trial follows and, again, there can be no case (the suspect is released), or the suspect may be either acquitted and released or sentenced. The last stage corresponds to the classification of the sentenced individual into one of several types of prisons or centers (prison, training center, open prison, psychiatric center, drug-addiction treatment center, or similar facilities). Furthermore, prisons can usually be of maximum, medium or minimum security, and there could be segregation by other inmate features, including gender and health status. Finally, defendants may go free or use a halfway house.

During the time they serve their sentences, prison inmates can go through several different stages depending on the correctional system. In some cases, the whole time includes induction, regular incarceration and pre-release stages, while in other cases a sentenced individual can access some benefits that can include a supervised (or unsupervised) time in open prisons or even an early release in the cases of a good behavior.

15.3 An Application

15.3.1 The Setting

In 2000, Gendarmería de Chile, the institution in charge of the penitentiary system in Chile, commissioned a study, whose main objective was to determine the capacities required for housing an increasing number of inmates over a 20-year time horizon in the whole country. The facilities in the country serve the needs of both jails and prisons, and they also house short-term detainees. For this reason, from now on, we indistinctly refer to the facilities as jails or prisons. These capacities could be fulfilled by increasing the size of existing prisons and by building new, large-size prisons. The locations (provinces) for these new prisons were to be determined, as well as a general allocation of inmates to facilities. Furthermore, the schedule of the new prison openings, the timing of each expansion of both new and old prisons, and the final size of all prisons at the end of the time horizon (in terms of number of inmates) was to be also determined. In regard with the locations, the main idea was to solve only the first level of the problem, i.e., the provinces in which the facilities should be located, as opposed to finding the exact lots for each facility or the architectural features of the establishments. The objective was to minimize the cost of maintaining the penitentiary system, including opening and expansion costs, overcrowding penalties, transportation costs and some social and environmental costs. Finding the required capacities required a forecast of the demand over the whole project horizon (20 years).

The prison population in Chile includes different types of inmates: individuals under pre-trial detentions (staying in a facility for up to 5 days), defendants under trial (ranging from a few weeks up to 5 years[6]) and sentenced offenders staying in prison from a few months to life. In the year 2000, there were some 33,000 inmates, and the flow of detainees in the jails was of approximately 200,000 people. Detainees and inmates under trial must be housed in facilities close to the respective courts, because they have to frequently travel to and from court. Sentenced criminals do not need to be close to courts, but the rehabilitation rate increases if they stay in contact with their families, so they must be kept not far from their relatives. All types of inmates are kept in the same prisons, although there is internal segregation.

The system comprises small local jails as well as regional or multi-jurisdictional large jails, each covering one of the jurisdictions of several courts. The capacities range from a few tens to a few thousands of inmates. At the time of the study, the country was divided in 13 regions and 51 provinces. Each province was composed by the jurisdiction of one or several courts, depending on its population. The jurisdiction of one particular court falls always entirely within one province. Since large prisons can cover more than one court jurisdiction and more than one province's demand, the problem is not separable by districts.

[6] Since then, criminal justice went through a deep reform, and the trials are much shorter.

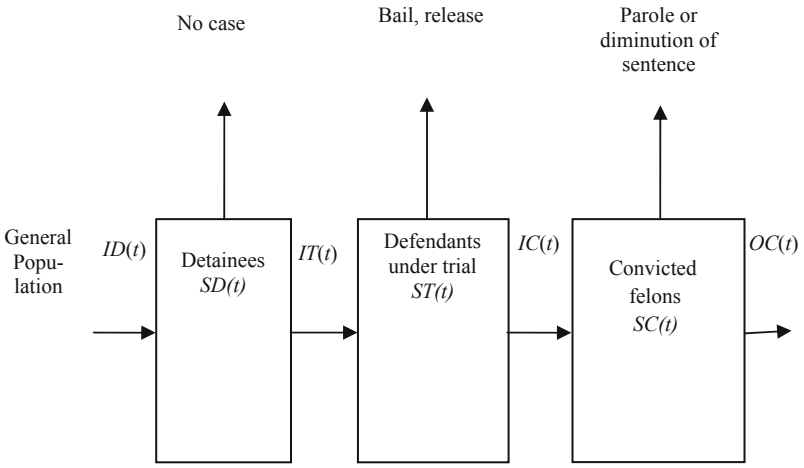


Fig. 15.1 Inmate flow

15.3.2 Demand Forecast

The demand for the 20-year horizon was estimated for each geographic unit (province) by using a flow model in which each type of inmate (detainees, defendants under trial, and convicted felons) is modeled as a process with an input rate, a stock and an output rate, all changing in time according to demographic and policies changes. Some authors use models in which recidivism is also considered (Baker and Lattimore 1994); however, information on recidivism was not available to us. Fig. 15.1 shows the diagram of the model we used.

Note that in Fig. 15.1, the number $ID(t)$ of individuals entering the "Detainees" category, is the number of people that are arrested, which is proportional to the population of the province. The proportionality factor depends on the police efficiency and practices. Detainees stay as such for a maximum of 5 days, so in a year, the output of the "Detainees" process $IT(t)$ is basically equal to the input $ID(t)$. However, there is a stock $SD(t)$, which is computed as the total number of detainees in a year (approximately 200,000) divided by 365 and multiplied by the average length of stay of the detainees in the jail. A detained person can be released or become a defendant under trial. As a result, the input to the process "Under Trial", $IT(t)$, is proportional to $ID(t)$. The proportion of defendants that are bailed is computed from historical data. So are the proportions of defendants that stay for different periods of time in the jail. Since at the time of the study, people could stay under trial up to 5 years, the stock of people under trial can be computed as

$$ST(t) = \sum_{i=0}^{4} \left(\sum_{k=i+1}^{5} b_{kt} \right) \lambda_1^{i+1} IT(t-i),$$

where b_{kt} are the proportions of defendants whose trials finish after k years. The subscript t reflects the fact that these proportions vary in time. The parameter λ_1 is the probability of a person under trial not being bailed out. A similar reasoning is used to compute the stock of convicted felons, except that $IC(t)$ consists of people who were convicted after being under trial for 1, 2, 3, 4 and 5 years, and the stock of people comprises people who have been in prison for times that go from less than a year up to life, although it is uncommon to find people that have been there for more than 19 years. More specifically, we obtain

$$IC(t) = c \sum_{i=0}^{5} b_{it} IT(t-i)$$

$$SC(t) = \sum_{i=0}^{18} \left(\sum_{k=i+1}^{19} c_{kt} \right) IC(t-i)$$

All of the coefficients and factors are computed by fitting the series to historical data.

Once a forecast was obtained by fitting the series to the historical data, three different inmate population growth scenarios were obtained using different projected values of the most significant parameters: police efficiency, duration of trials, sentencing and custodial policies of courts, changes in crime rates due to new types of crimes, and changes in the percentage of offenders sentenced to programs that are alternatives to incarceration. Also, a fourth scenario was proposed by the decision maker, which considered a strong decrease in the number of inmates. The uncertainty and hence, the need for different scenarios, was mainly due to a penal law amendment affecting sentencing policies and trial modalities, whose application was starting at the time of the project. As a consequence, inmate population could equally likely increase or decrease, because there are opposite effects: more alternatives to incarceration and shorter trials would decrease the prison population, while the public reaction to a more effective justice would increase the number of trials and inmates.

Additional scenarios were required because of the fact that prisons were heavily overpopulated at the time of the study. As a consequence, some of the inmates from provinces with overpopulated jails were being sent to prisons in nearby provinces. Thus, the base time series of some provinces did not necessarily reflect the real number of inmates originating in that province. Rather, it reflected a saturated condition of nearby provinces. In order to correct this distortion, the fact was used that, because of administrative reasons, inmates were never moved out of their region of origin (provinces are grouped in 13 regions), i.e., the inmates in a region belong to that region only. Thus, aggregating all the inmates in a region and redistributing them later among provinces in proportion to the provinces' population, would tend to correct this distortion, provided that the crime rate is the same across the region. Then, four new scenarios were devised, equivalent to the basic ones, but in which the inmate population of a whole region was first aggregated and then, distributed among provinces, proportionally to their populations.

Table 15.1 Eight population growth scenarios *S*1 to *S*8

	Scenario description	What is affected	Effect
S1	Time series	Total number of inmates	–
S2	Time series, plus aggregation of inmates in a region and subsequent redistribution among provinces, proportionally to their population	Total number of inmates	Redistribution between provinces
S3	Increase in police effectiveness	Number of detainees	Increases in 15 %
	Shorter trials	Number of inmates under trial	Decreases in 15 %
	Less alternatives to incarceration	Number of sentenced inmates	Increases in 20 %
S4	Same parameters as scenario 3, with province redistribution as in scenario 2	See scenario 3	See scenarios 2 and 3
S5	Increase in police effectiveness	Number of detainees	Increases in 4 %
	Shorter trials	Number of inmates under trial	Decreases in 4 %
	Less alternatives to incarceration	Number of sentenced inmates	Increases in 6 %
S6	Same parameters as scenario 5, with province redistribution as in scenario 2	See scenario 5	
S7	More alternatives to incarceration	Number of inmates under trial and sentenced	Decreases in 50 %
	Increase in police effectiveness	Number of detainees	Increases in 15 %
S8	Same as scenario 7, with province redistribution as in scenario 2	See scenario 7	See preceding scenario and scenario 2

The model was solved for the eight demand scenarios, described in detail in Table 15.1.

15.3.3 The Model (Marianov and Fresard 2005)

For modeling effects, demand was aggregated at the geographical centers of the provinces. The same points are the locations of courts and existing facilities, and candidates for locations of the new facilities. Each province has a known existing capacity for inmates. Each existing facility has a known maximum growth capacity (in number of inmates). New facilities have preset initial and maximum capacities.

Some of the provinces are candidates to location of new jails, and each candidate province has a maximum number of new facilities. Construction activity (opening of new facilities and expansions) is considered only in the 1st, 4th, 8th, 12th, 16th, and 20th years, the "active" years.

The system was modeled as a network, in which the nodes are the geographical centers of the provinces. The lengths of the arcs are the road distances between province centers. Detainees or inmates under trial must remain in the same province of their respective courts. However, if the capacity in their province of origin is not sufficient, sentenced offenders can be sent to other provinces. In such a case, inmates cannot be located farther than 150 or 400 km from their families, depending on which part of the country they are being held. The model penalizes transportation distances.

The *variables* of the model are the following:

$Sent_{ijt}$ Number of sentenced inmates which are transported from courts in province *i*, to facilities in province *j*, on year *t*. These variables are defined only for provinces *i* and *j* closer to each other than the maximum transportation distance (150 or 400 km).

Loc_{jt} Number of new jails that are opened (located) in province *j* in year *t*.

$Open_{jt}$ Number of new jails that are either opened or remain open in province *j* in year *t*. This variable does not need to be declared integer, since the model structure forces it to be integer.

Cap_{jtt} Capacity (number of inmates) of existing facilities in province *j* in year *t*.

$CapN_{jt}$ Capacity (number of inmates) of new facilities in province *j* in year *t*.

Exp_{jt} Expansions (number of inmates) of existing facilities in province *j* in year *t*.

$ExpN_{jt}$ Expansions (number of inmates) of new facilities in province *j* in year *t*. The initial capacity of the facility is included.

$Overp_{jt}$ Overpopulation in province *j* during year *t*.

The *parameters* and *sets* are the following:

FC_{jt} Fixed opening cost of a new facility at province *j* in year *t*.

EC_{jt} Expansion cost per inmate at province *j* in year *t*. This may be different for new and existing jails.

TP_{ij} Transportation penalty per inmate sent from province *i* to province *j*.

OP_{j} Overpopulation penalty per inmate in province *j*.

$DemSent_{it}$ Demand of space for sentenced inmates in province *i* in year *t*.

$DemTr_{it}$ Demand of space for detainees and defenders under trial in province *i* in year *t*.

$CurrCap_{j}$ Current capacity in province *j*.

$PotCap_{j}$ Maximum potential capacity in existing facilities in province *j*.

$IniCap_{j}$ Initial capacity of new facilities in province *j*.

$PotCapN_{j}$ Maximum capacity of new facilities in province *j*.

SC_{ij} Social cost per inmate of visitors' transportation, as a percentage of the transportation penalty of sentenced inmates.

p_j	Maximum number of new facilities in province j.
N_i	Set of provinces j not farther to i than the allowed maximum distance (150 or 400 km)

The model can then be written as follows:

$$\min \sum_{j \text{ candidate}} \sum_{t} \left(FC_{jt} Loc_{jt} + EC_{jt} ExpN_{jt}\right) + \sum_{j} \sum_{t} EC_{jt} Exp_{jt} + \\ + \sum_{i} \sum_{j} \sum_{t} (1 + SC_{ij}) TP_{ij} Sent_{ijt} + \sum_{j} \sum_{t} OP_j Over_{jt} \tag{15.1}$$

$$\text{s.t.} \sum_{j \in N_i} Sent_{ijt} = DemSent_{it} \; \forall i, t \tag{15.2}$$

$$DemTr_{jt} + \sum_{i} Sent_{ijt} \leq Cap_{jt} + CapN_{jt} + Over_{jt} \forall j, t \tag{15.3}$$

$$Cap_{jt} = Cap_{j(t-1)} + Exp_{jt} \forall j, t \tag{15.4}$$

$$CapN_{jt} = CapN_{j(t-1)} + ExpN_{jt} \forall j, t \tag{15.5}$$

$$Cap_{jt} \leq PotCap_j \forall \; j, t = \text{last year} \tag{15.6}$$

$$CapN_{jt} \leq PotCapN_j * Open_{jt} \forall \; j, t \tag{15.7}$$

$$CapN_{jt} \geq IniCap_j * Open_{jt} \forall \; j, t \tag{15.8}$$

$$Open_{jt} = Open_{j(t-1)} + Loc_{jt} \forall \; j, t \tag{15.9}$$

$$\sum_{t} Open_{jt} \leq p_j \forall \; j \tag{15.10}$$

The first term in the objective (1) minimizes the fixed cost of opening new facilities (cost of land, basic services; environmental and political costs, all depending on the province, its characteristics and population) and the expansion costs of new facilities, including common areas and services. The second term minimizes the expansion costs of existing facilities, which are lower than those of new facilities. The third term penalizes inmate transportation costs and visitors' distance costs. The last term penalizes overpopulation. This cost, per inmate, is higher than the expansion cost (otherwise, the model will always increase overpopulation, and there will be no capacity expansions). It was set so that a new facility opens when the overpopulation in its area has reached a level that makes its opening cost effective. In agreement with the planners, we set this level to one third of the opening capacity. In this application, the minimum capacities of new prisons were 1000 and 1650, depending on the province. New prisons were opened when the overpopulation in their "service areas" reached 333 and 550 inmates, respectively, meaning that the overpopulation

cost of 333 inmates is equal to the opening cost of a 1000 inmate prison, and the overpopulation cost of 550 inmates is equal to the opening cost of a 1650 inmate prison.

Constraint (2) requires that at any year t and province i, all the demand for space must be satisfied. Constraint (3) states that at any year t and province j, the capacity of existing and new facilities must be sufficient to satisfy the demand, and that some overpopulation is allowed. Constraints (4) and (5) indicate for existing and new facilities respectively, that for every year t, the capacity of in province j is their capacity in the previous year, t–1, plus the expansions on year t. For the first period, the previous capacity of existing facilities is $CurrCap_j$ and zero for new facilities. Constraints (6), (7) and (8) set the maximum and minimum capacities for all facilities. By virtue of constraints (9), in any candidate province, the new facilities remain open once they are located ($Open_{j0} = 0$). Finally, constraint (10) sets the maximum number p_j of facilities that can be located at any candidate province.

15.3.4 Scenario Analysis

The model was solved for each scenario independently, obtaining for each scenario the locations of the new facilities and the years in which they should be open; the expansion schedule for all facilities; the inmate allocation to facilities, and the overpopulation that could be expected in each facility for each active year.

In most scenarios, the existing jails are expanded to their maximal sizes during the time horizon, because this is the least expensive way of increasing the capacity of the system. However, as Table 15.2 (Marianov and Fresard 2005) shows, locations and numbers of new facilities strongly depend on the scenario. In the Table, the numbers correspond to the opening year (4, 8, 12, 16, 20.)

In different scenarios, different numbers of prisons are open, ranging from 3 to 12. As there is no information about what scenarios could be most likely, we assumed equally likely scenarios and used a minimum regret analysis to recommend the locations for the new facilities. For the subsequent runs, instead of solving the problem with a free number of new facilities, we agreed with the decision maker that it would be better to fix the number of new facilities at 10. This decision, although political, is justified because, no matter what is the result of the exercise, it must be repeated every few years to adapt the planning to the actual scenario as it develops in time. If too few facilities are planned at the beginning of the horizon, some of the locations could result highly suboptimal if a populated scenario actually happens. Furthermore, in no scenario more than 10 new facilities are open in the first 10 years, and there is always time to add new facilities. Because of political reasons, also, the new facilities should be planned to be open in the eighth year.

Table 15.2 Opening years of new jails, for all eight scenarios *S*1 to *S*8

	S1	S2	S3	S4	S5	S6	S7	S8	Total number
Province									
24	12		8	20	12				*4*
27			4	8	8				*3*
29			20						*1*
33			16	12					*2*
34	4	4	4	4	4	4	4	4	*8*
35					20	20			*2*
37		20	8	8	16	16			*5*
38			12		16	20			*3*
39	4	4	4	4	4	4	4	4	*8*
40				12					*1*
43	20			16					*2*
44			12	16					*2*
45	4	4	4	4	4	4	4	4	*8*
46			16		16				*2*
47		16, 20	12	8, 12		12, 16			*7*
Total number	*5*	*6*	*12*	*12*	*9*	*8*	*3*	*3*	*58*

15.3.5 Regret Analysis

At this stage, we required an approach to decide the location of the new prisons, that would allow reducing as much as possible the decision error, no matter what scenario happened in practice. This worst-case error minimization is achieved through regret analysis.

The regret analysis is performed by solving an aggregated model that integrates several scenarios simultaneously. In this model, all objectives, constraints and variables of the original model are repeated for each scenario, except for the location variables corresponding to new facilities, which are left as common to all scenarios and make the problem non-separable by scenarios. The objective to be minimized is the highest cost among all scenarios.

Five scenario combinations (cases) were run separately:

- Aggregation of scenarios 1, 3, 5, 7, i.e., those that are generated using the current prison population without redistribution among provinces.
- Aggregation of scenarios 2, 4, 6, 8, i.e., those in which the demand is redistributed among provinces according to their population.
- Three combinations of all scenarios (1 to 8), with different penalties on overpopulation: a “high overpopulation” case, with a low penalty on overpopulation, thus

Table 15.3 Locations that minimize the worst-case scenario

Province	Scenarios 1, 3, 5, 7	Scenarios 2, 4, 6, 8	High over-population	Medium over-population	Low over-population
24	*			*	*
27					*
29	*	*		*	
34	*	*	*	*	*
36	*	*		*	*
37	*	*	*	*	*
38	*	*	*	*	*
39	*	*	*	*	*
41			*		
43		*	*		
44			*		
45	*	*	*	*	*
46	*			*	*
47 – 1	*	*	*	*	*
47 – 2		*	*		
Total	*10*	*10*	*10*	*10*	*10*

The asterisks mark the provinces in which prisons are open for each scenario group.

allowing more overcrowding; a "low overpopulation" case, with high penalty on overcrowding and resulting in low overpopulation; and an intermediate value on overcrowding penalty (in this case, the value used for the penalty ensures opening of a new jail when overpopulation reaches one third of its capacity).

The results of the analysis are shown in Table 15.3 (Marianov and Fresard 2005).

In Table 15.3, the first column shows the numbers of the provinces in which jails are open in the different runs. The numbers 47 – 1 and 47 – 2 in this column represent a first and a second jail in province 47 (a very populated province). In the second column, the asterisks mark the provinces in which prisons are open in year 8, when regret analysis is applied to the aggregation of scenarios 1, 3, 5, 7 (the scenarios designed using the current prison population, without correcting facility saturation). The third column shows the provinces in which prisons are open when the regret analysis is performed aggregating scenarios 2, 4, 6 and 8 (the scenarios with the corrected population distribution). The remaining columns show the provinces with prison openings, for regret analysis performed by aggregating all scenarios, with different values of penalization on overpopulation.

In our final analysis, we maintain first those locations that seem to be robust, i.e., locations that appear in every possible result in Table 15.3. These are the new facilities in provinces 34, 37, 38, 39, 45 and 47. We also recommended a jail in province 36, since it appears in all cases, except in the not very desirable case in

which high overpopulation is allowed. For the three remaining prisons, we choose those candidates that appear more frequently. Finally, the proposed set of locations, minimizing the worst case cost, is in provinces 24 (or 29), 34, 36, 37, 38, 39, 43, 45, 47 and either 46 or a second jail in 47.

If these new prisons are open, the overpopulation is reasonable in the whole country, except for some small provinces in the central zone, which could be an artifact due to the current offenders-prison assignment. If the prison population in these provinces shows to be large in practice, small, inexpensive jails (150 inmates) could be easily open in the provinces 42, 19, 16, and 13.

Furthermore, the overpopulation is low in all cases until the middle of the time horizon. Thus, there is time enough to correct the situation, if needed.

15.3.6 Final Outcome of the Project

Given the uncertainty in the future demand for capacity, it was strongly recommended to update the study every active year or even more frequently. Furthermore, once the project was completed and its results handled to the authorities, political considerations not taken into account previously came into play, and the planning was partly changed. Finally, the office in charge decided to plan for the building of 10 prisons in two stages. In the first stage, 5 prisons were to be built in 5 of the proposed provinces: 34, 36, 39, 44, and 45. The construction for this stage started right away. A second stage would include provinces 33, 38, 43, 46 and 47. This second stage was not to be initialized until some years later, and only after checking the new change rates in prison population, which development could show a new trend, as a product of the recent law amendments.

A second recommendation addressed the availability of data. For this project, data had to be collected from many different sources, and there were inconsistencies that made the data analysis very complicated. A list of useful data to be recollected in the future was suggested, which could be used in subsequent updates of the study.

Some years after the project was finished, a different approach to addressing uncertainties in the problem was explored, in which a tree of scenarios was defined (Hernández et al. 2012). In this tree, there is branching at each "active" period of time. Branches correspond to different uncertain parameter realizations in that particular period. At each active period, decisions can be made regarding construction. Thus, the number of possible scenarios to analyze grows significantly. Once the tree is defined, a model is formulated which aggregates individual models for each scenario (similar to the model used during the project,) and constraints are added to ensure non anticipativity and consistency between scenarios; i.e., a given scenario does not use information available only in a later scenario, and scenarios originating in a common branch of the tree share those parameters and context that correspond to the common branches. In order to solve the large model, a technique is used called branch and fix coordination (Alonso-Ayuso et al. 2003).

This new approach requires far more computational power, but provides solutions that reduce overpopulation and the expected cost (Hernández et al. 2012).

15.4 Conclusions

We here present an application aimed at designing the correctional system for a whole country. As such, it considers the expansion of the current capacities, as well as location and opening of several large prisons. Readers interested in location or dimensioning of individual jails, can check some available reports, e.g., Ricci Green Associates (2013).

A previous requirement was to forecast the prison population in the coming 20 years. In our application, the forecast was complicated by the high degree of uncertainty introduced by a law amendment, affecting the behavior of police corps, judges and population, passed shortly before the study was started. For the forecast to be useful, a time series is not enough: it must be analyzed and changed according to new trends affecting crime, e.g., the increase in drug-related crime in Chile at the time of the study, as well as the demographic and sociological changes in the country. In synthesis, all the factors that have an influence on the number of inmates must be analyzed and embedded into the forecast. In our case, there was an additional difficulty: historical information about prison population was obtained from the yearly figures of existing prisons. However, some of the facilities were saturated and inmates in the province had to be sent to prisons in other provinces. This fact introduces a distortion in the forecast that must be corrected by reassigning inmates between provinces, as described here.

Another particular feature of this application is the fact that the prison population is composed by three different types of inmates: detainees staying as inmates for 5 days or less; defendants under trial, in the system for up to a few years; and sentenced offenders. Each type of inmate has requirements in terms of location and closeness to courts or families. In other countries, these requirements are fulfilled by opening at least two types of facilities: jails and prisons. Jails are close to court, while prisons need to be reachable by visitors and employees and service providers.

Finally, the location model considered cost as an objective, plus a penalization of long distance travelled by families and visitors of the sentenced inmates. Additionally, instead of considering hard capacity constraints, we allowed some overpopulation, also penalized in the objective. For the locations to be found, the model also assigned inmates to existent and new facilities. It also provided, as part of the solution, the scheduling of the opening of the new facilities and the expansion works for existing facilities.

Together with the results of the study, some recommendations were offered to the decision maker, regarding the size of the future facilities and the data requirements for future studies. Regarding facility size, although large facilities can take advantage of economies of scale in staffing, supplies and services, from the point of view of overcrowding it seems better to open facilities with small opening sizes (150–400 inmates), since large facilities require waiting until overcrowding in existing prisons justifies their opening; otherwise they would be operating with a few inmates for some time, which is not an efficient use of resources. These sizes also make rehabilitation more likely, as show some studies (Hall et al. 2013; Johnsen 2011).

This is probably due to the fact that the likelihood of rehabilitation increases with the quality of life in prison, which, in turn, tends to be better when the prison size is smaller. In any case, these issues require much more discussion, as it seems that there are no conclusive studies.

Acknowledgments Partial support by FONDECYT grant 130265, as well as by the Institute Complex Engineering Systems through Grants ICM MIDEPLAN P-05-004-F and CONICYT FBO16 is gratefully acknowledged.

References

Alonso-Ayuso A, Escudero LF, Ortuo MT (2003) BFC, a branch-and-fix coordination algorithmic framework for solving some types of stochastic pure and mixed 0–1 programs. Eur J Oper Res 151:503–519

Armstrong S (2012) Siting prisons, sighting communities: geographies of objection in a planning process. Working Paper, Scottish Centre for crime and justice research, University of Glasgow. http://ssrn.com/abstract=2117840. Accessed 4 Nov 2015

Baker J, Lattimore P (1994) Forecasting demand using survival modeling: an application to US prisons. Australas J Inf Syst 2(1). http://journal.acs.org.au/index.php/ajis/article/view/411/372. Accessed 4 Dec 2015

Berk R (2008) Forecasting methods in crime and justice. Annu Rev Law Soc Sci 4:219–238

Brans JP, Vincke PH (1985) A preference ranking organization method, the PROMETHEE method. Manage Sci 31:647–656

Chan H (1996) Medium security prison. Master of Architecture Thesis, The University of Hong Kong (Pokfulam, Hong Kong). http://hub.hku.hk/bitstream/10722/26060/1/FullText.pdf?accept=1. Accessed 4 Nov 2015

Cherry T, Kunce M (2001) Do policymakers locate prisons for economic development? Grow Change 32(4):533–547. http://www.appstate.edu/ cherrytl/papers/prison_location.pdf. Accessed 4 Nov 2015

Comfort ML (2002) 'Papa's House': the prison as domestic and social satellite. Ethnography 3(4):467–499

de Silva N, Cowell P, Chow T, Worthington P (2006) Prison population projections 2006–2013, England and Wales. Report of the Great Britain home office research development and statistics directorate. https://www.ncjrs.gov/App/abstractdb/AbstractDBDetails.aspx?id=236652. Accessed 4 Nov 2015

Eiselt HA, Marianov V (2015) Location modeling for municipal solid waste facilities. Computers and operations research 62:305–315

Engel MR (2007) When a prison comes to town: siting, location, and perceived impacts of correctional facilities in the Midwest. PhD Thesis, The graduate college at the University of Nebraska. Lincoln, Nebraska, August 2007

Farkas MA (1999) 'Not in my backyard': the issues and complexities surrounding prison siting. Justice Prof 12:95–109

Fehr L (1995) Literature review of impacts to communities in siting correctional facilities. Report, Washington council on crime and delinquency, Seattle, WA. http://www.tippecanoe.in.gov/egov/docs/1203448928_100158.pdf. Accessed 4 Nov 2015

Hall D (2006) Jails vs. prisons. American correctional association website. http://www.thefreelibrary.com/Jails+vs.+prisons.-a0184267124. Accessed 4 Dec 2015

Hall MJB, Liu WB, Simper R, Zhou Z (2013) The economic efficiency of rehabilitative management in young offender institutions in England and Wales. Soc Econ Plan Sci 47:38–49

Hernández P, Alonso-Ayuso A, Bravo F, Escudero L, Guignard M, Marianov V, Weintraub A (2012) A branch-and-cluster coordination scheme for selecting prison facility sites under uncertainty. Comput Oper Res 39(9):2232–2241

Hooks G, Mosher C, Genter S, Rotolo T, Lobao L (2010) Revisiting the impact of prison building on job growth: education, incarceration, and county-level employment, 1976–2004. Soc Sci Quart 91(1):228–244

Johnsen B, Granheim PK, Helgesen J (2011) Exceptional prison conditions and the quality of prison life: prison size and prison culture in Norwegian closed prisons. Eur J Criminol 8(6):515–529

Johnson MP (2006) Decision models for the location of community corrections centers. Environ Plan B 33:393–412

King RS, Mauer M, Huling T (2003) Big prisons, small towns: prison economics in rural America. The sentencing project, Washington, DC. http://prison.ppjr.org/files/tracy%20huling%20prisons%20economy%20study.pdf. Accessed 4 Nov 2015

Kleindorfer PR, Kunreuther HC (1994) Siting of hazardous facilities. In: Pollock SM, Rothkopf MH, Barnett A (eds) Handbooks in operations research and management science 6. Elsevier North-Holland, Amsterdam, The Netherlands, pp 403–440

Korporaal R, Ridder A, Kloprogge P, Dekker R (2000) An analytic model for capacity planning of prisons in the Netherlands. J Oper Res Soc 51:1228–1237

Lawrence S, Travis J (2004) The new landscape of imprisonment: mapping America's prison expansion. Research Report CPR04 0121, Urban Institute, April 2004, available at http://www.urban.org/UploadedPDF/410994_mapping_prisons.pdf. Accessed 4 Nov 2015

Maltz MD (1996) From Poisson to the present: applying operations research to problems of crime and justice. J Quant Criminol 12(1):3–61

Marianov V, Fresard F (2005) A procedure for the strategic planning of locations, capacities and districting of jails: application to Chile. J Oper Res Soc 56:244–251

Martin R, Myers D (2005) Public response to prison siting: perceptions of impact on crime and safety. Crim Justice Behav 32(2):143–171

Ministry of Justice UK (2013) Prison population projections 2013–2019 England and Wales. Ministry of justice statistics bulletin, 7th November 2013. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/255574/prison-population-projections-2013-2019.pdf. Accessed 4 Nov 2015

Moran D, Pallot J, Piacentini L (2011) The geography of crime and punishment in the Russian Federation. Eurasian Geography Econ 52(1):79–104

NAATAP (2014) Population profiles, population projections and bed needs projections. Native American and Alaskan technical assistance project. https://www.bja.gov/Publications/PopProfiles.pdf. Accessed 4 Nov 2015

NBC (2012) Inside the secret industry of inmate-staffed call centers. http://usnews.nbcnews.com/_news/2012/01/12/10140493-inside-the-secret-industry-of-inmate-staffed-call-centers. Accessed 4 Nov 2015

New Zealand Department of Corrections (2001) Inmate employment policy. http://www.corrections.govt.nz/__data/assets/pdf_file/0007/676087/inmateemployment.pdf. Accessed 4 Nov 2015

Ricci K (2006) Jail site evaluation and selection. Bulletin from the jails division of the National institute of corrections, April 2006. https://s3.amazonaws.com/static.nicic.gov/Library/021280.pdf. Accessed 4 Nov 2015

Ricci Green Associates (2013) Validation study of the dutchess county criminal justice system needs assessment May 28:2013. http://www.co.dutchess.ny.us/CountyGov/Departments/CriminalJusticeCouncil/CJValidationStudyofNeedsAssessment.pdf. Accessed 4 Nov 2015

Saaty TL (1980) The analytic hierarchy process: planning, priority setting, resource allocation. McGraw Hill, New York

San Mateo County Sheriff's Office (2009) Site selection. http://www.smcsheriff.com/jail-planning/site-selection. Accessed 4 Nov 2015

Scalia J (2004) Federal prisoner detention: a methodology for projecting federal detention populations. Report of the office of the federal detention trustee. http://www.justice.gov/archive/ofdt/rpd_methodology2004.pdf. Accessed 4 Nov 2015

Stephan J (2008) Census of state and federal correctional facilities 2005. bureau of justice statistics, U.S. department of justice, October 2008. http://www.bjs.gov/content/pub/pdf/csfcf05.pdf. Accessed 4 Nov 2015

US Government Accountability Office (2012) Growing inmate crowding negatively sffects inmates, dtaff, and infrastructure. Bureau of prisons, Report GAO-12-743. Published September 12:2012. http://www.gao.gov/products/GAO-12-743. Accessed 4 Nov 2015

Wamsley R (2013) World prison population list (10th ed.) International centre for prison studies. http://www.prisonstudies.org/sites/prisonstudies.org/files/resources/downloads/wppl_10.pdf. Accessed 4 Dec 2015

Wan W-Y, Moffatt S, Xie Z, Corben S, Weatherburn D (2013) Forecasting prison populations using sentencing and arrest data. Crime and justice bulletin, NSW bureau of crime statistics and research, Nr. 174, October 2013

Chapter 16
Vessel Location Modeling for Maritime Search and Rescue

Ronald Pelot, Amin Akbari and Li Li

16.1 Introduction

16.1.1 Locating Maritime Search and Rescue Vessels

The location-allocation problem serves to deploy assets effectively to respond to a known or estimated geographically distributed demand for service, including providing emergency services. At first glance, the rescue vessel location problem in the Search and Rescue (*SAR*) domain is similar to the emergency vehicle location problem such as for ambulance location. In both cases, mathematical location models can be formulated to maximize the number of incidents that can be serviced by a specified number of resources (vehicles) within a pre-specified amount of time, or, alternatively, we can minimize the time it would take a vehicle to arrive at the scene of the incident. However, several differences exist. First of all, in the case of emergency vehicle location, all response units are generally assumed to have the same capability and speed. Conversely, the Canadian Coast Guard (*CCG*) has many different *SAR* rescue vessel types that were designed or purchased with specific tasks in mind, and not all are equally effective at handling different incident types. Also, the ranges vary greatly among different types of rescue vessels, so rescue vessel capabilities need to be considered in our study. Furthermore, the method of computing distances to the incidents is different as rescue vessels are patrolling on the sea, thus requiring a land-avoidance algorithm to calculate the travel distance rather than Euclidean or Manhattan distance metrics. However, land-avoidance distance is calculated before performing the optimization in this study, so this distinction is moot in this instance.

R. Pelot (✉) · A. Akbari · L. Li
Department of Industrial Engineering, Dalhousie University, Halifax, B2H 4R2, NS, Canada,
e-mail: Ronald.Pelot@Dal.Ca

A. Akbari
e-mail: Amin.Akbari@Dal.Ca

H. A. Eiselt, V. Marianov (eds.), *Applications of Location Analysis,* International Series in Operations Research & Management Science 232, DOI 10.1007/978-3-319-20282-2_16

This research is based on earlier models in the literature, with some modifications for our Maritime Search and Rescue situation. Planning the future location of rescue vessels requires the following:

- Data on incident numbers for each grid cell, as well as the incident date and time
- Candidate sites for locating rescue vessels
- Stratification of incidents by severity and determination of associated response time category
- Examination of differences between types of rescue vessels and categorization of rescue vessels based on the capability to response to various incident classes
- Modification and application of Location-Allocation models for Maritime Search and Rescue scenarios

The Canadian Coast Guard determines each incident's categorization based on severity and they also establish the response time requirements for different categories of severity. Categorization of rescue vessel types is based on their respective capabilities (Cameron and Pelot 2005).

16.1.2 Research Scope

16.1.2.1 Study Area

The entire Atlantic Canada region serves as our research area, partitioned into 2263 grid cells (Fig. 16.1).

16.1.2.2 Study Period

The incident data from 2000 to 2003 were used for this study and checked and cleaned for quality control. Two sizes of grid squares were generated in the GIS software. The grid size cannot be too large or small. If it is too large, the calculated distances will not be very precise. On the other hand, if it is too small, the distance calculation will be computationally expensive. However, greater resolution is advantageous near the shore where the incident density is highest. Therefore, a specified distance of 150 nautical miles (nmi) from shore was chosen to delineate small and large grid square areas (see the dotted line in Fig. 16.1). For the area outside the line, a size of 1×1 degree square is used, while for the area nearer to shore a smaller size of 0.25×0.25 degrees is used. The historical incidents within these grids are used as reasonable representations of demand for SAR service.

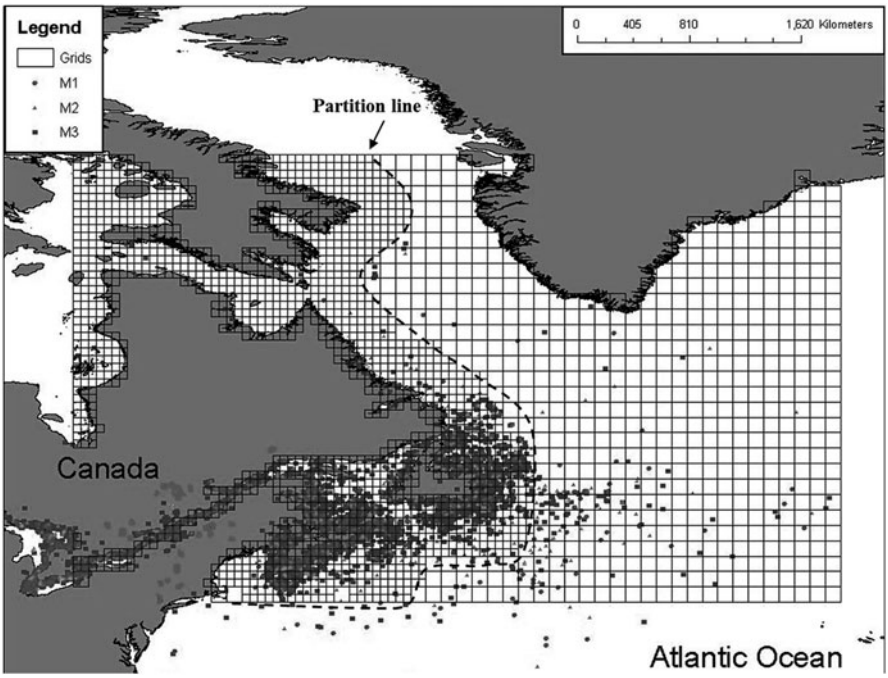


Fig. 16.1 Eastern Canada SAR incidents (2000–2003) by severity class: M1 (*circle*), M2 (*triangle*) and M3 (*square*)

16.1.2.3 Choice of SISAR Incidents by Classification

All incidents are categorized into severity classes, of which the following three are relevant for this study: Class M1 are distress incidents, class M2 are potential distress incidents, and class M3 are non-distress incidents. The study scope is shown on the map in Fig. 16.1.

16.1.3 Objectives

This study examines the optimal locations of rescue vessels in Atlantic Canada. The ultimate objective is to ensure the maximum likelihood of saving lives and secondarily of mitigating property loss using available resources. The access time is an important performance criterion for emergency service systems and the likelihood of saving lives and property is highly related to the access time. The access time is defined in this study as the elapsed time from the departure of each rescue vessel to the arrival at the scene of an incident. Since all of the vessels categorized as primary type vessels (i.e. regular lifeboats) have similar speed, access time can be approximately reflected by response distance. Moreover, primary rescue vessels have a

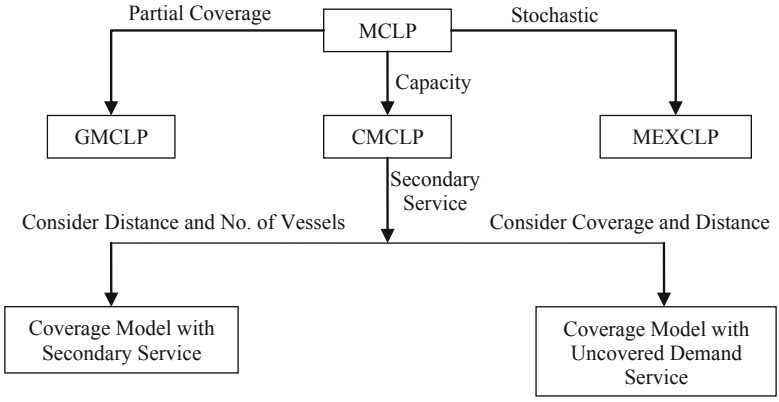


Fig. 16.2 *MCLP* model and extensions

limited range, so the extent of coverage should also be a key concern in this system. Therefore a surrogate objective is to maximize the coverage within a predetermined access time limit or to minimize the average response distance by optimizing the location of rescue vessels.

16.2 Literature Review

Classical mathematical programming approaches for discrete location problems fall into three main categories: covering problems, median problems (minsum models) and center problems (minmax models). The covering problem is concerned with covering demands within a specified response time standard, while the median problem aims to minimize the system-wide average response time. In the center problem, the objective is to minimize maximum distance from the facilities. Covering and median problems are widely used in the area of emergency location analysis which is the case in this study. The maximal covering location problem (*MCLP*) model provided by (Church and ReVelle 1974) has proven to be one of the most useful facility location models from both theoretical and practical points of view. It is a tool for siting decisions where worst case performance is a primary concern. This is often appropriate for emergency service systems such as police, fire, and ambulance service. Berman (1994) describes the maximal covering location problem on networks. Demands and potential facility sites are represented on nodes, so the problem space forms a network of nodes. In our case, potential rescue vessel stations are limited and incidents are represented on gridded nodes which is a discrete problem, not continuous.

The fundamental *MCLP* model and some of its possible extensions can be illustrated in Fig. 16.2 as follows. Based on the fundamental model, *MCLP*, three extensions are produced by including considerations of partial coverage (*GMCLP*),

capacity limits (*CMCLP*) and stochastic factors (*MEXCLP*) respectively. In order to consider the service to uncovered demand, two coverage models with secondary service are given. These two models will be modified to apply to our research, where response units (rescue vessels) have different capabilities.

Berman and Krass (2002) presented a generalized maximal covering location problem. They used the concept of partial coverage in their study and defined the degree of coverage as a decreasing function of the distance to the closest facility. They showed that this problem is equivalent to the general uncapacitated *MCLP*. It is assumed that the demand is fully covered within the minimum critical distance *S*, partially covered up to a maximum critical distance *T*, and not covered at all outside of the maximum critical distance.

The location-allocation problem is an essential model for many applications, each of which has distinguishing characteristics. The ambulance and fire station location problems are "server-to-customer" type systems. Conversely, the distribution center and warehouse location problem is "customer-to-server" type system. In the Maritime Search and Rescue situation, the rescue vessel location problem is a "server-to-customer" emergency service system, which is similar to the emergency vehicle location problem.

There is a very large number of studies in the literature dealing with analyzing the location of emergency services facilities such as health centers, ambulances, fire stations and Maritime Search and Rescue stations. Example applications are locating hospitals (Sinuany-Stern et al. 1995), emergency medical services (Pirkul and Schilling 1988), blood banks (Jacobs et al. 1996), and ambulances (Ball and Lin 1993). Marianov and ReVelle (1996) discussed the problems and applications in siting emergency services. They proposed a model for a queueing maximal availability location problem taking into account the randomness of servers' availability. Harewood (2002) formulated a bi-objective programming problem to locate ambulances on the island of Barbados. Two objectives in the model are minimizing the cost of serving customers and maximizing multiple coverage given a certain distance standard.

A review of location models specifically applied to healthcare is found in (Daskin and Dean 2004). Goldberg (2004) reviewed the literature of operations research applications in emergency services vehicles. Another review study was performed by (Brotcorne et al. 2003) on the evolution of models in the area of ambulance location and relocation. They divided the studies into deterministic and probabilistic models. Griffin et al. (2008) developed an optimization model to determine the best location and number of new Community Health Centers in a geographical network as well as what services each such center should offer at which capacity level. The objective is to maximize the weighted demand coverage of the target population subject to budget and capacity constraints.

Nguyen and Kevin (2000) incorporated maximal covering and *p*-median location problems into a goal programming model to assess the level of service of the current *SAR* system (in terms of location of *SAR* aircraft and helicopters) and compare it to the optimal solution. They also used simulation and queueing theory to examine the performance of *SAR* aircraft in terms of average time that incidents spend in

Table 16.1 Incident class breakdown of *SISAR* data for 2000–2003

Incident class	Description	Incident count	Percentage (%)
M1	Distress	658	4.83
M2	Potential distress	1094	8.02
M3	Non-distress	10,105	74.09
Subtotal		*11,857*	
M4	False alarm/hoax	1733	12.71
M1P	Serviced by others	48	0.35
Total		*13,638*	

queue (i.e. waiting for a response unit) for both the current situation and their proposed solution, with the latter showing a significant improvement over the current situation. In a similar study for maritime *SAR*, Li (2006) employed both location modelling and simulation to establish optimal vessel placement as well as produce several system statistics related to utilization rates and workload balance. This work forms the basis for some of the modelling in this paper.

Azofra et al. (2007) proposed a tool for assignment of sea rescue resources to incidents. Their methodology, based on gravitational modeling, provides a coefficient for each possible assignment based on the appropriateness of the rescue vessel to the incident. This study only evaluates different solutions but is not trying to propose an optimal solution.

As shown in this section, many studies have been performed in the area of emergency location analysis. However, there are still some gaps remaining, particularly in the area of Maritime Search and Rescue location modelling.

16.3 Overview of Data

16.3.1 Incident Data

All maritime incidents in which Canadian Coast Guard Search and Rescue branch is contacted are recorded in the *SISAR* (Search and Rescue Program Information Management System) database.

There are 21 types of incidents, including: grounded, capsized, stranded, etc. The incidents are also categorized into five classes based on severity (Table 16.1)

Different incident classes may have different service times. M1 is a distress incident, M2 is a potential distress incident, and M3 is a non-distress incident. Only M1, M2, and M3 are included in the model objective as these classes of incidents contribute to the busy time for response utilization computations. Conversely, M4 incidents correspond to false alarms for which no resource is dispatched or, if one is, the rescue vessel is usually called back before arriving on scene. Thus, false alarms

Table 16.2 Speed and range by rescue vessel type

Rescue vessel class	Max speed		Average speed		Range	
	(knots)	(km/h)	(knots)	(km/h)	(nm)	(km)
Lifeboat	25	46.3	22.5	41.7	100	185
Icebreaker	20	37.0	18.0	33.3	6518	12070
Offshore patrol vessel	16	29.6	15.5	28.7	6518	12070
Inshore rescue boat (IRB)	45	83.3	45.0	83.3	4.35	8

are not included in the objective but sometimes contribute to the busy time of rescue vessels. M1P comprises distress or potential distress incidents which are responded to by some other non-Canadian Coast Guard (*CCG*) vessel.

16.3.2 Rescue Vessels' Capabilities Study

The *CCG* has limited documentation available about their rescue vessels and their rescue vessels' capabilities. Cameron and Pelot (2005) conducted a preliminary study on categorizing and analyzing the capabilities of CCG rescue vessels. The results for four main types of rescue vessels are summarized in Table 16.2.

The icebreakers and offshore patrol vessels have relatively long range which exceeds the maximum distance in our study area. Therefore, we can assume that these two types of vessels have unlimited range for the purposes of this analysis.

For the *MCLP* model with consideration of incident severity presented in Sect. 16.4.2, we consider two types of rescue vessels with limited ranges:

- Regular lifeboat with a range of 185 km,
- IRB with a range of 8 km.

For the two modified models in Sect. 16.4.5, we consider the primary rescue vessels and special boats with unlimited range:

- Primary Rescue vessels: Regular lifeboats with a range of 185 km,
- Special boats: Icebreakers and Offshore Patrol vessels with an unlimited range.

16.4 Solving Location Models for Maritime SAR

In this section, we apply three models, notably *MCLP* (Church and ReVelle 1974), *CMCL* (Pirkul and Schilling 1991), and *MEXCLP* (Daskin 1983) to our Maritime Search and Rescue system. In order to consider different types of response vessels, we also make some modifications to the coverage model by incorporating workload

capacities and backup service (Pirkul and Schilling 1988) and also extend the basic coverage model with uncovered demand service (Pirkul and Schilling 1991).

16.4.1 Maximal Covering Location Problem Model

The maximal covering location problem (*MCLP*) model developed by (Church and ReVelle 1974) has proven to be one of the most useful facility location models from both theoretical and practical points of view. It is a tool for siting decisions where worst case performance is a primary concern. This is often appropriate for emergency service systems such as police, fire, and ambulance service.

The objective of the fundamental *MCLP* model is to maximize the total coverage of demand (incidents) within a desired service distance S by locating a fixed number of facilities (e.g., rescue vessels). The formulation is as follows:

$$\text{Max } z = \sum_{i \in I} a_i y_i \tag{16.1}$$

$$\text{s.t.} \sum_{j \in N_i} x_j \geq y_i \; \forall \; i \in I \tag{16.2}$$

$$\sum_{j \in J} x_j = p \tag{16.3}$$

$$x_j \in \{0, 1\} \; \forall \; j \in J \tag{16.4}$$

$$y_i \in \{0, 1\} \; \forall \; i \in I \tag{16.5}$$

where

I: denotes the set of demand nodes (incidents),
J: denotes the set of candidate location sites for facilities as supplied by the CCG,
a_i: demand in each demand zone (i.e., grid) i,
x_i: 1 if a facility is allocated to site j; 0 otherwise,
y_i: 1 if demand at node i is covered; 0 otherwise,
d_{ij}: distance between node i and facility at node j,
S: the specified maximum response distance,

$N_i = \{j \in J | d_{ij} \leq S\}$ is the cover set for each demand zone i within the desired response distance S of a facility site j,

p: the number of facilities to be located.

Considering the incidents as demand, and regular lifeboats as the facilities to be located, we apply the *MCLP* model in the Atlantic Canada area based on gridded incident data from *SISAR* 2000 to 2003. We obtain the solution for coverage according to the number of rescue vessels to be placed, as shown in Fig. 16.3. It is apparent

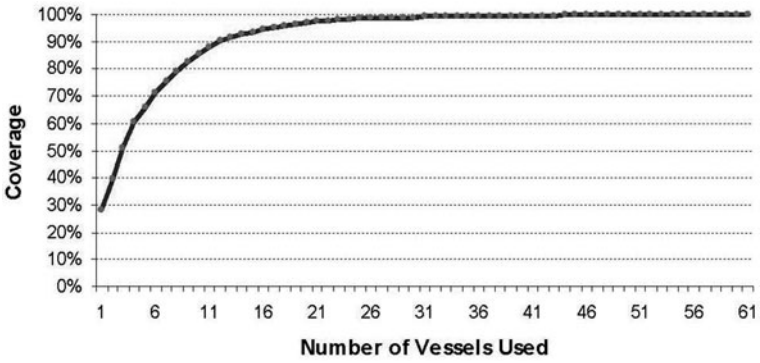


Fig. 16.3 Coverage curve for maximum response distance = 185 km

that the total coverage increases as the number of rescue vessels used increases. In addition, the figure reflects the fact that as additional rescue vessels are added to the study area, the marginal rate of additional coverage declines. As seen in the figure, all incidents can be covered by 57 rescue vessels with a range of 185 km, which is the range of a regular lifeboat (the primary rescue vessel in the *CCG*). Note however, that 95 % of the incidents could be covered by only 17 vessels optimally placed. This model yields the best coverage given a specified number of vessels to be located, but a limitation is that it does not consider the workload balance. Two extensions are applied in the following sections which smooth the workload among response units explicitly or implicitly.

16.4.2 MCLP Model with Weights on Incident Class

The above model weights each incident class equally and just considers the total number of all types of incidents for each grid. However, in reality, incidents with higher severity require more coverage and faster response time. Therefore, we want to cover as many high-severity incidents as possible. One straightforward strategy is to assign weights to different classes of incidents. Therefore the new objective is given as follows:

$$\text{Max} z = \sum_{i \in I} (w_1 a_{1i} + w_2 a_{2i} + w_3 a_{3i}) y_i,$$

where w_1, w_2 and w_3 are weights assigned to incident class M1, M2 and M3 respectively, and a_{1i}, a_{2i} and a_{3i} are the number of incidents of class M1, M2 and M3 occurring in grid i.

The resulting solutions for locating 20 rescue vessels according to five sets of weights are summarized in (Table 16.3)

Table 16.3 Solutions of *MCLP* with weights (locating 20 rescue vessels)

Weight			Coverage			
w_1	w_2	w_3	M1	M2	M3	Total
1	1	1	591	992	9927	11510
6	3	1	607	995	9864	11466
20	3	1	607	995	9860	11462
50	3	1	608	986	9786	11380
1	0	0	608	955	9709	11272

Table 16.4 Solutions of *MCLP* with weights (locating 5 rescue vessels)

Weight			Rescue vessels position	Coverage			
w_1	w_2	w_3	1 2 3 4 5	M1	M2	M3	Total
1	1	1	747, 759, 797, 934, 1011	319	552	6968	7839
6	3	1	747, 759, 797, 934, 1078	327	566	6069	6962
20	3	1	747, 759, 797, 934, 1078	327	566	6069	6962
50	3	1	747, 762, 934, 1011, 1078	328	558	6673	*7559*
1	0	0	747, 762, 934, 1011, 1079	328	558	6681	*7567*
1	1	0	747, 759, 797, 934, 1078	327	566	6059	6952

As the relative weight on M1 increases, the number of M1 incidents which can be covered is increasing slightly but the total coverage is reduced. In general, the coverage of M1 incidents is not very sensitive to the weights in this example. For comparison, the final solution with weight of (1, 0, 0) represents the maximum possible number of M1 incidents covered for 20 vessels with the specified 185 km range. The second-last solution maintains this optimal M1 coverage, but with better coverage of M2 and M3 incidents.

This exercise is repeated using only 5 rescue vessels, with results given in Table 4.4. Similarly, as more relative weight is placed on M1, the coverage of M1 incidents increases slightly but the total coverage reduced. The second-last solution represents the maximum coverage of M1 possible. The final solution with weights of (1, 1, 0) represents the maximum coverage of M1 and M2 combined.

The solutions for locating 5 rescue vessels are displayed on maps in Figs. 16.4 and 16.5. The circles represent M1, while the crosses denote both M2 and M3. In Fig. 16.4, we put equal weights of (1, 1, 1) on M1, M2 and M3. The shaded area represents the areas which can be covered. In this case, region I with many incidents of class M1 but relatively few total incidents is not covered whereas region II with more total incidents and proportionately fewer M1 is covered. In Fig. 16.5, we put all the weight on M1 and this time region I is covered and region II is not. (Table 16.4)

From the above two Tables 16.3 and 16.4, we can see that the advantage of putting weights on the incidents is that we can obtain comparatively higher coverage of the incidents with higher severity. However, the drawbacks are also obvious. First

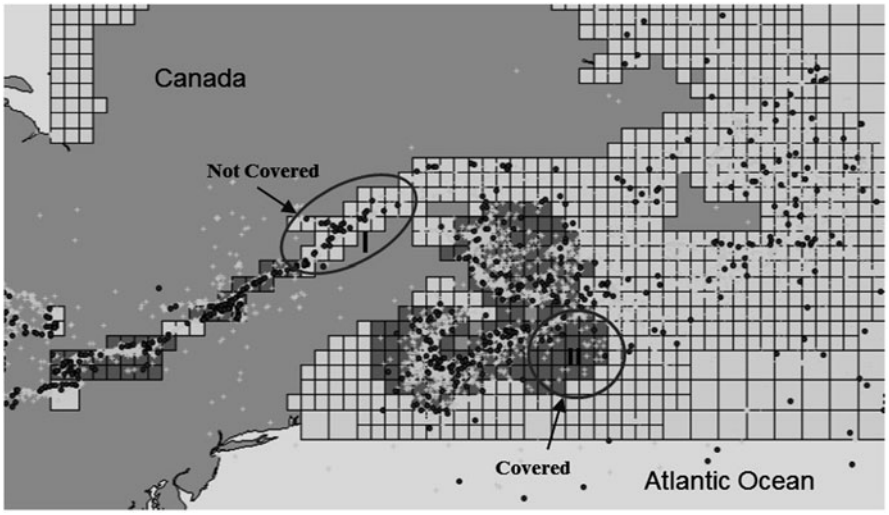


Fig. 16.4 Coverage with equal weights on M1, M2 and M3

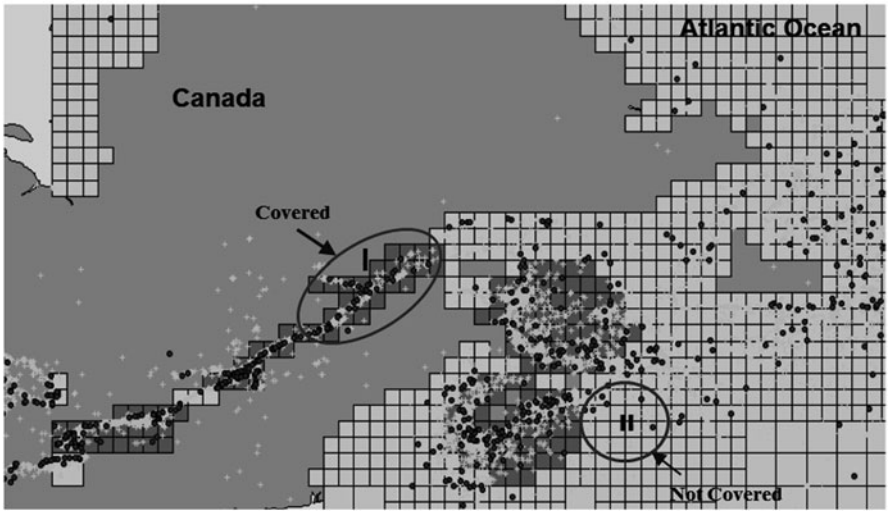


Fig. 16.5 Coverage with weight (1, 0, 0) on M1, M2 and M3

of all, we cannot quantitatively determine how much the coverage would improve by increasing the relative weight without running simulations on all possible combinations. Secondly, for a particular level of M1 and M2 coverage, we cannot guarantee that the total incident coverage is optimized. For instance, according to the fourth and fifth rows of Table 16.4, the total coverage given weights of (1, 0, 0) is higher than when the weights are set to (50, 3, 1), even though the coverage of M1 and M2 remains the same. The above drawbacks also cause the jump in coverage of M3

from the solution with weights of (20, 3, 1) to (50, 3, 1). Another possible reason for this irregularity could be that there may exist some other optimal solution with more M2 and fewer M3 which has the same total coverage.

A more general model given as follows can address the above drawbacks to some extent. We split the total incidents in the objective function into three terms indicating the incident numbers for each class M1, M2 and M3. Constraints of type (16.7), (16.8) and (16.9) allow us to specify the maximum response distance and create a specific covering set for each incident class. Moreover, we can give specific coverage requirements for M1, M2 and M3 as a percent of their respective totals in constraint sets (16.11), (16.12) and (16.13) respectively.

$$\text{Max} z = \sum_{i \in I_1} a_{1i} y_{1i} + \sum_{i \in I_2} a_{2i} y_{2i} + \sum_{i \in I_3} a_{3i} y_{3i} \tag{16.6}$$

$$\text{s.t.} \sum_{j \in N_{1i}} x_j \geq y_{1i} \quad \forall i \in I_1 \tag{16.7}$$

$$\sum_{j \in N_{2i}} x_j \geq y_{2i} \quad \forall i \in I_2 \tag{16.8}$$

$$\sum_{j \in N_{3i}} x_j \geq y_{3i} \quad \forall i \in I_3 \tag{16.9}$$

$$\sum_{j \in J} x_j = p \tag{16.10}$$

$$\sum_{i \in I_1} a_{1i} y_{1i} \geq r_1 T_1 \tag{16.11}$$

$$\sum_{i \in I_2} a_{2i} y_{2i} \geq r_2 T_2 \tag{16.12}$$

$$\sum_{i \in I_3} a_{3i} y_{3i} \geq r_3 T_3 \tag{16.13}$$

$$x_j \in \{0, 1\} \quad \forall j \in J \tag{16.14}$$

$$y_{1i}, y_{2i}, y_{3i} \in \{0,1\} \quad \forall i \in I \tag{16.15}$$

where

I_1: denotes the set of incident grids with incidents of class M1,
I_2: denotes the set of incident grids with incidents of class M2,
I_3: denotes the set of incident grids with incidents of class M3,
S_1: maximum response distance for an M1 incident,
S_2: maximum response distance for an M2 incident,
S_3: maximum response distance for an M3 incident,
d_{ij}: the land-avoided distance from grid i to grid j,

y_{1i}: 1 if demand class 1 at node i is covered; 0 otherwise,
y_{2i}: 1 if demand class 2 at node i is covered; 0 otherwise,
y_{3i}: 1 if demand class 3 at node i is covered; 0 otherwise,
x_i: 1 if a facility is allocated to site j; 0 otherwise,
a_{1i}: number of incidents of class M1 in grid i,
a_{2i}: number of incidents of class M2 in grid i,
a_{3i}: number of incidents of class M3 in grid i,
p: the number of rescue vessels to be located,
T_1: the total number of M1 incidents,
T_2: the total number of M2 incidents,
T_3: the total number of M3 incidents,
r_1: the minimum percentage of M1 incidents to be covered,
r_2: the minimum percentage of M2 incidents to be covered,
r_3: the minimum percentage of M3 incidents to be covered.
N_{1i}: $\{j \in J : d_{ij} \leq S_1\}$,
N_{2i}: $\{j \in J : d_{ij} \leq S_2\}$,
N_{3i}: $\{j \in J : d_{ij} \leq S_3\}$,

N_{1i} is the set of rescue vessel location sites eligible to cover the M1 class incidents in grid i, N_{2i} is the set of rescue vessel location sites eligible to cover the M2 class incidents in grid i, and N_{3i} is the set of rescue vessel location sites eligible to cover the M3 class incidents in grid i.

The objective is to maximize the total number of M1, M2 and M3 incidents which can be covered within the respective maximum response distances S_1, S_2 and S_3. Constraints of type (16.7) allow y_{1i} to equal 1 only when one or more rescue vessels are located within S_1 of incident grid i. Similarly for constraints of type (16.8) and (16.9), y_{2i} and y_{3i} can equal to 1 only when one or more rescue vessels are located within S_2 and S_3 respectively. Usually the more severe the incidents are, the shorter is the required maximum response distance, so $S_1 \leq S_2 \leq S_3$. The number of rescue vessels to be located must equal p in constraint (16.10). Constraints of type (16.11)–(16.13) specify the percentage of incidents M1, M2 and M3 which are required to be covered respectively.

In addition, the model can also be extended to rescue vessels with different ranges. As previously shown in Table 16.2, the lifeboats, special boats (Icebreaker and Offshore Patrol Vessel) and Inshore Rescue Boat (*IRB*) have very distinct ranges. However, since the model is based on the concept of coverage, the special boats which can reach anywhere in the study area should not be considered, otherwise all incidents could be covered by just one special boat. Therefore, only lifeboats and inshore rescue boats are considered in this model. Then constraints of type (16.7)–(16.9) in the previous model are modified as follows:

$$\sum_{j \in N_{1i}} x_j \geq y_{1i} \; \forall \, i \in I_1 \tag{16.16}$$

$$\sum_{j \in M_{1i}} z_j \geq y_{1i} \; \forall \, i \in I_1 \tag{16.17}$$

Table 16.5 Locating 5 rescue vessels (maximizing coverage of M1)

No. of lifeboats	r_1 (%)	Coverage of M1 (%)	Coverage of M2 (%)	Coverage of M3 (%)	Total coverage (%)
5	0	48.48	50.46	68.96	66.11
5	48.5	48.63	50.18	66.55	64.05
5	49.0	49.70	50.64	66.38	64.00
5	49.8	49.85	51.01	66.12	63.82

$$\sum_{j \in N_{2i}} x_j \geq y_{2i} \; \forall i \; \in I_2 \tag{16.18}$$

$$\sum_{j \in M_{2i}} z_j \geq y_{2i} \; \forall \, i \in I_2 \tag{16.19}$$

$$\sum_{j \in N_{3i}} x_j \geq y_{3i} \; \forall \, i \in I_3 \tag{16.20}$$

$$\sum_{j \in M_{3i}} z_j \geq y_{3i} \; \forall \, i \in I_3 \tag{16.21}$$

where x_j denotes regular lifeboats: $x_j = 1$ if a regular lifeboat is located in grid j; $x_j = 0$ otherwise. Similarly, z_j denotes inshore rescue boats: $z_j = 1$ if an inshore rescue boat is located in grid j; $z_j = 0$ otherwise.

$N_{1i} = \{j \in J : d_{ij} \leq \min\,(R_1, S_1)\}$ is the set of regular lifeboat location sites eligible to cover the M1 incident class in grid i. $R_1 = 185$ km is the range of a regular lifeboat. S_1 is the required maximum response distance of incident class M1.

$M_{1i} = \{j \in J : d_{ij} \leq \min\,(R_2, S_1)\}$ is the set of inshore rescue boat location sites eligible to cover the M1 incident class in grid i, and $R_1 = 8$ km is the range of an inshore rescue boat. Constraints of type (16.16) ensure that for any grid where a regular lifeboat is expected to respond to the M1 incident load, at least one lifeboat unit is located in a proximate cell within a distance that is the lesser of that boat type's range of 185 km and the associated required maximum response distance. Constraints of type (16.17) apply the same restriction for *IRB*s, and subsequent constraints of type (16.18)–(16.21) extend this logic to the other incident classes M2 and M3 respectively.

If we locate 5 lifeboats and want to cover as many incidents as possible of class M1 which is the most severe incident type, then we can set $r_2 = r_3 = 0$ to obviate constraints (16.12) and (16.13). Varying the value of r_1, the solution results are presented in Table 16.5. Initially letting $r_1 = 0$ yields the solution in the first row. Then increasing the M1 coverage a little bit at a time produces better and better coverage of M1, and meanwhile the total incident coverage decreases. When r_1 increases to 49.8 %, we obtain the best coverage of 328 M1 incidents for this example.

Table 16.6 Locating 5 rescue vessels (maximizing coverage of M1 and M2 combined)

No. of rescue vessels	r (%)	Coverage of M1 and M2 (%)	Coverage of M3 (%)	Total coverage (%)
5	0	49.71	68.96	66.11
5	49.8	49.89	68.20	65.50
5	49.9	50.29	66.38	64.00
5	50.3	50.57	66.12	63.82
5	50.6	50.74	65.36	63.20
5	50.8	50.97	60.04	58.70

Since the incidents of class M2 reflect potential distress, we may want to cover as many as possible M1 and M2 incidents combined. Setting $r_3 = 0$, the solutions are summarized in Table 16.6. Initially letting $r = 0$, where r is the minimum coverage required based on the percentage of M1 and M2 incidents combined, we get the solution in the first row. Then increasing the coverage of M1 and M2 a little bit each time, we can get better and better coverage of M1 and M2 combined, and meanwhile the total coverage decreases. When r increases to 50.8 %, we obtain the best coverage of 50.97 % for both M1 and M2 incidents combined. It is remarkable that the slight increase in M1 + M2 coverage in the last row results in a dramatic drop in the percentage of M3 incidents covered.

16.4.3 MCLP with Workload Capacities

The fundamental *MCLP* model does not consider balancing the workload among facilities, therefore some response units may cover a disproportionate number of the incidents while other units may not have much work to do. When designing a model, capacity or workload limits can serve as a means of smoothing the workload among the different response units. In order to impose the workload constraint on each facility, a binary decision variable x_{ij} is introduced into the *CMCLP* model objective function as follows (Pirkul and Schilling 1991):

$$\text{Max} \sum_{i \in I} \sum_{j \in J} C_{ij} a_i x_{ij} \tag{16.22}$$

$$\text{s.t.} \sum_{j \in J} y_j \leq p \tag{16.23}$$

$$\sum_{j \in J} x_{ij} = 1 \; \forall \; i \in I \tag{16.24}$$

$$x_{ij} \leq y_j \; \forall \; i \in I, \; j \in J \tag{16.25}$$

$$\sum_{i \in I} C_{ij} a_i x_{ij} \leq K_j \ \forall \ j \in J \tag{16.26}$$

$$y_j \in \{0, 1\} \ \forall \ j \in J \tag{16.27}$$

$$x_{ij} \in \{0, 1\} \ \forall \ i \in I, \ j \in J \tag{16.28}$$

where

I: the index set of all demand points,
J: the index set of all potential facility sites,
a_i: the demand at point i,
K_j: the workload capacity for a facility at site j,
p: the number of facilities to be sited,
S: the desired maximum response distance,
d_{ij}: the travel distance from facility j to demand i,
C_{ij}: 1, if $d_{ij} \leq S$; 0 otherwise,
x_{ij} : 1, if the demand at point i is assigned to a facility at j; 0, otherwise,
y_j : 1 if a facility is located at j; 0 otherwise.

x_{ij} equals 1 if the demand at point i is served by a facility in grid j, otherwise x_{ij} equals 0. Except for the capacity constraints, the *CMCLP* model is quite similar to the original *MCLP*. However, the introduction of binary decision variable x_{ij} makes the number of decision variables and constraints increase exponentially, which makes the model substantially more difficult to solve.

Our problem has 2263 demand zones which means x_{ij} is of dimension 2263×2263. This is a large-scale problem which cannot be solved easily using an exact approach. In actuality however, we do not need to consider all 2263 grids. First of all, not all of the grids have incidents. Secondly, each incident can just be covered by a limited number of grids given the lifeboat range of 185 km. Two methods to compress the data and reduce the problem size making it easier to solve are given as follows.

16.4.3.1 Pick Certain Candidate Sites

Since our objective is to maximize the total number of incidents that can be covered, we select 722 grids of the 2263 total in which one or more incidents occurred from 2000 to 2003 as the demand grid set. Usually rescue vessels would be better located at a position which has more incidents. So we pick 183 grids which have 10 or more incidents of the original 2263 grids as the potential location set. Then the problem size becomes, 722×183 which is still large but can be solved by the *CPLEX* solver (*IBM ILOG CPLEX*). This method is simple but not as precise because we assume the rescue vessels are located in grids with 10 or more incidents. However selecting candidate sites according to some geographic or other environmental factors is very practical in reality. Usually rescue vessels cannot be located just anywhere, and most of time vessels are located near shore. So the candidate sites can be narrowed down greatly.

Table 16.7 Solution results of two data compression methods

No. of rescue vessels	Range km	Capacity constraints	No. of candidate sites	Coverage by *CMCLP*		Coverage by *MCLP*	
20	185	≤ 790	183	11152	94.05 %	11510	97.07 %
20	185	≤ 790	2263	11454	96.60 %		
50	185	≤ 316	183	9235	77.89 %	11841	99.87 %
50	185	≤ 316	2263	9505	80.16 %		

16.4.3.2 Workload Capacity

We use the formula below to establish the maximum workload capacity:

$$K = \text{Average workload capacity} = \frac{\text{Total incidents number}}{\text{Number of vessels located}} \times \alpha$$

where α is the coefficient of capacity and $\alpha \geq 1$ as we assume that no rescue vessel will be assigned *a priori* a lower than average workload. $\alpha = 1$ means every response unit is assigned a strictly equal workload. We set $\alpha = 4/3$ in our problem.

Based on the data from 2000 to 2003 with a total number of 11,857 incidents, if locating 20 rescue vessels:

$$\text{Workload Capacity} = \frac{11857}{20} \times \frac{4}{3} = 790 \text{ incidents/vessel}$$

If locating 50 rescue vessels:

$$\text{Workload Capacity} = \frac{11857}{50} \times \frac{4}{3} = 316 \text{ incidents/vessel}$$

Table 16.7 shows that the more candidate sites we choose, the better the solution we obtain. However, selecting fewer candidate sites based on incident rate is fairly efficient. Choosing grids with 10 or more incidents as candidate sites yields 183 options for placing the rescue vessels. However, the total coverage is only 2.55 % less than choosing entire the 2263 grids as candidate sites for locating 20 rescue vessels, and for locating 50 rescue vessels there is only 2.27 % difference in the optimal *CMCLP* coverage.

In order to see how the total coverage is related to workload capacity balance when varying the number of rescue vessels to be located, we solve the above model varying the value of α from 1.1 to 3.5. The solutions are displayed in Fig. 16.6. While all solutions converge to very high coverage when workload balance is relaxed (except for the 10 vessel scenario), the more interesting result is when greater parity is maintained, such as $\alpha = 4/3$ as per the earlier assumption. In this case, the effectiveness varies significantly, with 20 vessels providing the best combination of coverage in this example.

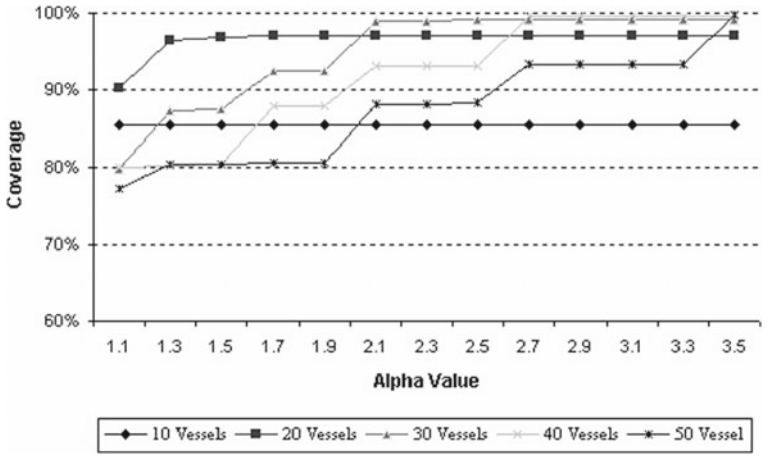


Fig. 16.6 *CMCLP* solution for varying number of rescue vessels and α

16.4.4 MCLP with Stochastic Considerations

All of the above models assume that each rescue vessel has response capability to handle any request for service at any time. However, this is not the case in reality for Maritime Search and Rescue. The maximal expected covering location model (*MEXCLP*) given by (Daskin 1983) extended the basic *MCLP* model by incorporating stochastic considerations of availability. The objective here is to maximize expected coverage given that rescue vessels are busy (unavailable) with a probability p. Given that a demand node can be covered by m facilities, then the probability that this node is covered by at least one available response unit is $1 - p^m$.

There are three assumptions in this model:

- Response units operate independently
- Response units have the same busy probabilities
- Response units' busy probabilities are invariant with respect to their locations

Under the above assumptions, the number of available facilities at any time follows a binomial distribution:

$$p(j \text{ facilities being available out of } M \text{ located}) = \binom{M}{j}(1-p)^j p^{M-j},$$

$j = 0,\ 1,\ \ldots, M$.

The formulation of (Daskin 1983) is as follows:

$$\text{Max} \sum_{k=1}^{N} \sum_{j=1}^{M} (1-p) p^{j-1} a_k y_{jk} \tag{16.29}$$

$$\text{s.t.} \sum_{i=1}^{N} C_{ki} x_i \geq \sum_{j=1}^{M} y_{jk} \; \forall \; k \tag{16.30}$$

$$\sum_{i=1}^{N} x_i \leq M \tag{16.31}$$

$$x_i = 0, 1, 2, \ldots, M \; \forall \; i \tag{16.32}$$

$$y_{jk} \in \{0, 1\} \; \forall \; j, k \tag{16.33}$$

where

M: number of facilities to be located,
N: number of nodes in the entire study area,
S: the desired maximum response distance,
d_{ki}: the travel distance from facility i to demand k,
C_{ki}: 1, if $d_{kl} \leq S$; 0, otherwise,
x_i: the number of facilities located at node i,
y_{jk} : 1 if node k is covered by at least j facilities; else 0, if less than j,
a_k: the demand amount at node k.

The objective function (16.29) can be rewritten as:

$$\sum_{j=1}^{M} \left((1-p) p^{j-1} \sum_{k=1}^{N} a_k y_{jk} \right).$$

The solutions of locating 10, 20, 30, 40 and 50 rescue vessels respectively with the busy (i.e., unavailable) probability ranging from 0.05 to 0.35 are graphically compared in Fig. 16.7. As the number of rescue vessels located increases, the total coverage is less sensitive to the probability of being busy.

Another useful tool for planning emergency response systems is the hypercube queuing model, first proposed by Larson (1974). In this model, each server in the system can be represented individually thus allowing for more complex assignments, while accounting for temporal and geographic characteristics of the area. An adjusted *MEXCLP*, model and a hypercube optimization procedure were given by (Batta et al. 1989), which relaxed the three assumptions in the original *MEXCLP* model. Based on their results, there is disparity among the original *MEXCLP*, adjusted *MEXCLP*, and hypercube optimization procedure for large busy probability p.

For our Maritime Search and Rescue system which does not experience heavy utilization, the hypercube model is not necessary. For locating 20 vessels, the average probability of rescue vessel being busy is $p \approx 0.15$; for locating 50 vessels, $p \approx 0.05$.

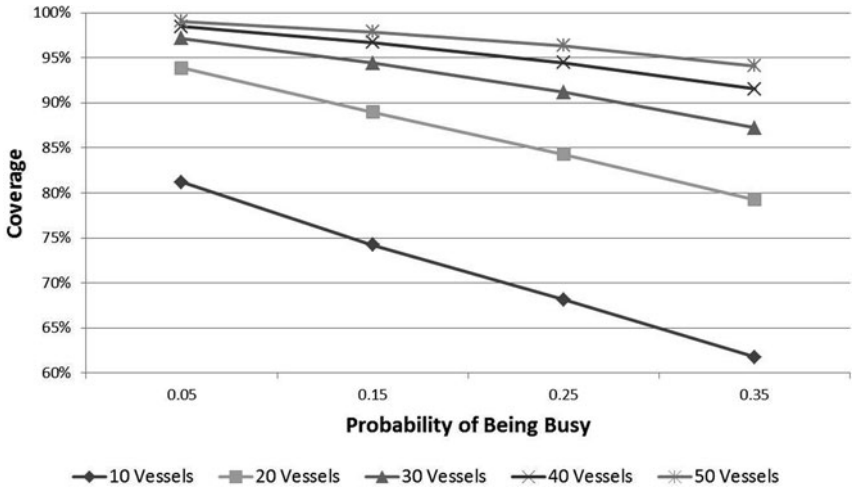


Fig. 16.7 Coverage percentage by varying the busy probability

16.4.5 Coverage Model with Response Units Having Unlimited Range

The three models applied above are for considering regular lifeboats, the principal type of rescue vessel. If we want to include the special boats with infinite range in the model, we have to modify the formulations.

16.4.5.1 Coverage Model with Secondary Service

The coverage model with workload capacities and backup service given by (Pirkul and Schilling 1988) ensures that each demand can be serviced by both a primary and secondary response units using the following two constraints:

$$\sum_{j \in J_i} x_{ij} = 1 \quad \forall\, i \in I,$$

$$\sum_{j \in J_i'} z_{ij} = 1 \;\forall\, i \in I,$$

The first constraint requires that each demand point i can be serviced by a response unit located at j within the maximum prescribed response distance. The second constraint requires that each demand point can also be responded to by a backup (secondary) response unit which can be beyond the maximum prescribed response distance.

Combining the two concepts of secondary service and workload capacities as a variation of the above model produces more realistic Maritime Search and Rescue

scenarios. The regular lifeboat with a range of 185 km can be used to give primary service, while special boats which can go to anywhere in our study area can provide secondary service. If an incident can be responded to by both a regular lifeboat and a special boat, the primary service is preferred. The above model requires both a primary and a secondary service for each demand point which is suitable for heavily-used systems. However, in our application, the average vessel utilization is less than 0.05 for locating 60 vessels, which is quite low. In addition, since the number of rescue vessels is limited, a term is added to the objective function to represent the respective costs of locating regular lifeboats and special boats. The modified model can be expressed as follows:

$$\text{Min} \sum_{i \in I} \left(\sum_{j \in J} c_{ij} a_i x_{ij} + \sum_{k \in K} c'_{ik} a_i y_{ik} \right) + \sum_{j \in J} v_j r_j + \sum_{k \in K} w_k s_k \tag{16.34}$$

$$\text{s.t. } x_{ij} \leq h_{ij} \ \forall \ i \in I, \ j \in J \tag{16.35}$$

$$\sum_{i \in I} x_{ij} \leq M r_j \ \forall \ j \in J \tag{16.36}$$

$$\sum_{i \in I} y_{ik} \leq M s_k \ \forall \ k \in K \tag{16.37}$$

$$p_1 - \sum_{j \in J} x_{ij} \leq M q_i \ \forall \ i \in I \tag{16.38}$$

$$\sum_{j \in J} x_{ij} - p_2 \leq M q_i \ \forall \ i \in I \tag{16.39}$$

$$\sum_{j \in J} h_{ij} \leq M(1 - q_i) \ \forall \ i \in I \tag{16.40}$$

$$1 - \sum_{k \in K} y_{ik} \leq M o_i \ \forall \ i \in I \tag{16.41}$$

$$1 - \sum_{j \in J} h_{ij} \leq M(1 - o_i) \ \forall \ i \in I \tag{16.42}$$

$$\sum_{j \in J} x_{ij} + \sum_{k \in K} y_{ik} = 1 \ \forall \ i \in I \tag{16.43}$$

$$\sum_{i \in I} x_{ij} a_i \leq K_j^R \ \forall \ j \in J \tag{16.44}$$

$$\sum_{i \in I} y_{ik} a_i \leq K_k^S \ \forall \ k \in \ \forall \ K \tag{16.45}$$

$$x_{ij} \geq 0 \ \forall \ i \in I, j \in J; y_{ik} \geq 0 \ \forall \ i \in I, k \in K; q_i \in \{0, 1\} \ \forall \ \ i \in I; i \in I;$$
$$o_i \in \{0,1\} \ \forall \ i \in I \tag{16.46}$$

where

I: the index set of all demand nodes,
J: the index set of all potential positions for regular lifeboats,
K: the index set of all potential positions for special boats,
M: a large number,
S: desired maximum response distance (range of regular lifeboat),
p_1, p_2: upper and lower bounds on the primary service percentage in a grid, which can be covered by grids other than itself,
v_j: the cost of locating a regular lifeboat in grid j,
w_k: the cost of locating a special boat in grid k,
a_i: the number of incidents in grid i,
K_j^R: the maximum demand that can be served by a regular lifeboat in grid j,
K_k^S: the maximum demand that can be served by a special boat in grid k,
x_{ij}: the fraction of incidents in grid i serviced by a regular lifeboat in grid j,
y_{ik}: the fraction of incidents in grid i serviced by a special boat in grid k,

$$h_{ij} = \begin{cases} 1, \text{ if } d_{ij} \leq S \\ 0 \text{ otherwise} \end{cases}$$

$$r_j = \begin{cases} 1, \text{ if a regular lifeboat is located in grid } j \\ 0, \text{ otherwise} \end{cases}$$

$$s_k = \begin{cases} 1, \text{ if a special boat is located in grid } k \\ 0, \text{ otherwise} \end{cases}$$

The objective function (16.34) is to minimize both fixed and variable 'costs' such that sufficient regular and special boats are located to provide each incident grid with primary or secondary service.

The parameter c_{ij} is the cost of providing primary service by a regular lifeboat from grid i to grid j. c'_{ik} is the cost of providing secondary service by a special boat from grid i to grid k. The cost can be interpreted in several ways. One simple and practical way is to replace it by distance. To scale costs between 0 to 1, we define:

$$c_{ij} = d_{ij} / \max d_{ij}$$

$$c'_{ik} = d_{ik} / \max d_{ik}$$

where

d_{ij}: land-avoided distance from incident grid i to grid j,
d_{ik}: land-avoided distance from incident grid i to grid k,
$\max d_{ij}$: maximum distance from all incident grids to any potential position for regular boat,

max d_{ik}*:* maximum distance from all incident grids to any potential position for special boat.

The constraints of type (16.35) state that incidents can only be serviced by a regular lifeboat (primary service) which is within the specified range. K is the index set of all potential rescue vessel positions for special boats. However, since the special boat has unlimited range and can go anywhere in the study area, we do not have a comparable set of constraints on y_{ik}, as any incident can be assigned to a special boat.

Constraints of type (16.36) specify that a regular lifeboat must be placed at j if it is to provide any primary service to incidents in grid i. Similarly, constraints of type (16.37) specify that a special boat must be placed at k if it is to provide any secondary service to incidents in grid i.

Constraints of type (16.38)–(16.42) specify the primary service percentage, which can be logically represented as follows:

For all ($i \in I$)

{If

$$\sum_{j \in J} C_{ij} > 0$$

then

$$p_1 \leq \sum_{j \in J} x_{ij} \leq p_2;$$

Else

$$\sum_{k \in K} y_{ik} = 1;$$

End if;

};

For each incident grid, if there exists one or more candidate rescue vessel positions for a regular lifeboat, we require that the primary service percentage be between p_1 and p_2. Otherwise the incident will be fully serviced by special boats.

Constraints of type (16.43) state that the total service amount of each incident grid by both regular lifeboat and special boat should be equal to 1, which means every incident has to be assigned to exactly one response vessel.

Constraints of type (16.44)–(16.45) place bounds on rescue vessel workload capacity. For a regular lifeboat we have a capacity of K_j^R and special boats have a capacity of K_k^S.

Applying this model to the Canadian Search and Rescue Problem, we can identify the locations for rescue vessels such that all incidents around the Atlantic area will receive either a primary or a secondary service or both. The demand network consists of 2263 grids covering the whole Atlantic study area, 722 of which have

Table 16.8 Solutions for varying primary service percentage

Primary service percentage (%)	No. of regular lifeboats	No. of special boats	Demand serviced of primary (%)	Avg. dist of primary km	Avg. dist of secondary km	Total avg. dist. km
0–10	1	46	263 (2.2)	48.1	74.6	74.0
10 – 20	33	40	1983 (16.7)	55.0	71.7	68.9
20 – 30	34	35	3116 (26.3)	55.5	78.6	72.6
30 – 40	36	30	4414 (37.2)	51.5	83.1	71.3
40 – 50	38	26	5439 (45.9)	49.0	88.6	70.4
50 – 60	40	24	6385 (53.9)	50.1	95.1	70.8
60 – 70	42	18	7441 (62.8)	47.5	128.3	77.6
70 – 80	46	13	8690 (73.3)	46.9	157.4	76.4
80 – 90	50	10	9847 (83.1)	45.8	193.7	70.9
90 – 100	52	3	11264 (95.0)	48.7	407.1	66.6

one incident or more. We select 183 of these nodes with more than 10 incidents as potential candidate sites for regular lifeboats and special boats. Based on the 11,857 total number of incidents and assuming a fleet of about 60 rescue vessels, choosing $\alpha = \frac{4}{3}$ to ensure equitable workload balance, the rescue vessel capacity can be defined as:

$$K_j = \frac{\text{Total incidents number}}{\text{Number of vessels located}} \times \alpha = \frac{11857}{60} \times \frac{4}{3} \approx 263$$

Therefore, we set

$$K_j^R = K_k^S = 263 \ \forall \ j \in J, \ k \in K.$$

We set fixed cost equals to 5 and vary the primary service percentage from [0, 0.1] to [0.9, 1], yielding the location configurations shown in Table 16.8. For example, when the primary service percentage is over 70 %, the total number of rescue vessels is 59 of which the number of special boats equals 13.

To describe the solution more clearly, we display it graphically in Fig. 16.8. We can see that as the required percentage of primary service increases, the number of regular lifeboats is increasing while the number of special boats drops. The average distance of service for regular lifeboats does not show much variation as their range is restricted. However, the average distance of service for special boats increases sharply as the service percentage increases beyond 60 %. The reason for this phenomenon is that although the number of special boats and the fraction of secondary service decrease as the primary service percentage increases, the secondary service extent is not reduced as fast as the number of special boats needed. Therefore the average response distance by secondary service increases dramatically, especially when the primary service percentage is relatively high. (Table 16.8)

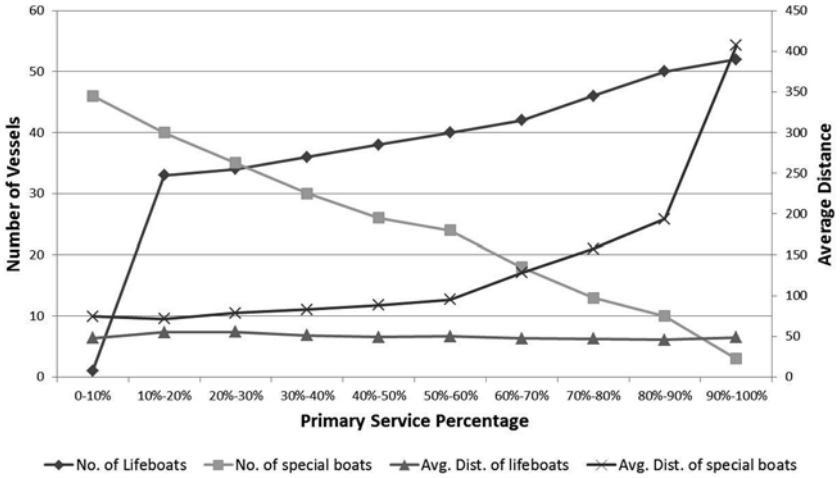


Fig. 16.8 Number of regular lifeboats and special boats needed, and average distance of primary and secondary service when varying the primary service percentage

The average distance of regular lifeboats (primary service) and special boats (secondary service) can be calculated as follows:

$$\text{Average distance per incident for regular lifeboats} = \sum_{i\in I}\sum_{j\in J} d_{ij}a_i x_{ij} \Big/ \sum_{i\in I}\sum_{j\in J} a_i x_{ij}$$

$$\text{Average distance per incident for special boats} = \sum_{i\in I}\sum_{k\in K} d_{ik}a_i y_{ik} \Big/ \sum_{i\in I}\sum_{k\in K} a_i y_{ik}$$

Generally the more rescue vessels to be located, the lower the total average travel distance obtained. When the primary service percentage is between 60 and 90 %, according to the results, the total number of regular lifeboats and special boats are close to the actual situation in Canadian Search and Rescue in that time period, giving a fairly short average response distance of primary service, average response distance of secondary service, and total average distance. Therefore, in the next part, when we examine the tradeoff between the travel distance and the fixed costs of locating rescue vessels, we will fix the primary service percentage to the range 70–80 %.

The objective function (16.34) in the preceding formulation serves to minimize both the number of rescue vessels needed and the total response distance. In order to trade off between these objectives, we include a weight W into the objective function:

$$\text{Min} \sum_{i\in I}\left(\sum_{j\in J} c_{ij}a_i x_{ij} + \sum_{k\in K} c'_{ik}a_i y_{ik}\right) + W\left(\sum_{j\in J} v_j r_j + \sum_{k\in K} w_k s_k\right)$$

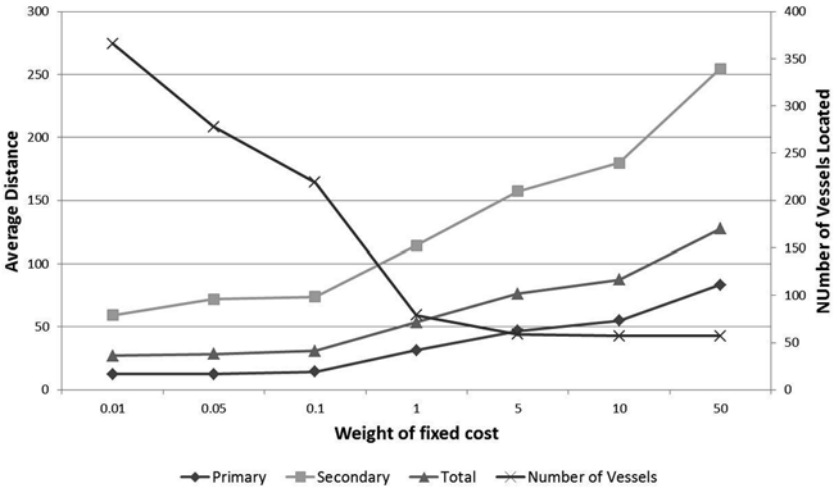


Fig. 16.9 Objective function tradeoffs between the number of rescue vessels located and the average response distance

Table 16.9 Solutions for objective function tradeoffs

W	No. of regular vessels	No. of special vessels	Total no. of vessels	Avg. dist of primary service km	Avg. dist of secondary service km	Total avg. dist km
0.01	183	183	366	12.6	59.2	27.1
0.05	177	101	278	12.5	72.0	28.5
0.1	143	76	219	14.3	74.0	30.8
1	56	23	79	31.4	114.3	53.5
5	46	13	59	46.9	157.4	76.4
10	45	12	57	54.9	179.8	87.5
50	45	12	57	83.4	254.4	127.4

Requiring the primary service percentage to lie between 70 and 80 %, and setting the fixed costs v_j and w_k to 1, seven location configurations are generated by varying W (Table 16.9). The average travel distance is used here rather than the total travel distance since the former is typically more meaningful when varying the number of rescue vessels located. As the weight on fixed cost W increases from 0.01 to 50, the number of rescue vessels located decreases and the average response distance increases. The first solution in Table 16.9, has the minimum travel distance and assigns a vessel to all the potential positions as there is actually little penalty for doing this (i.e. $W = 0.01$). The last solution represents the minimum fixed cost investment, but at the expense of increased response distance. The objective tradeoffs are graphically illustrated in Fig. 16.9.

16.4.5.2 Coverage Model with Uncovered Demand Service

The model described in Sect. 16.4.5.1 considers the total travel distance in the objective. Sometimes, service within range or maximum response distance can be treated having the same service level, which means if an incident can be responded to by primary service, we would not care much about the exact distance. Therefore, in the following model, we consider both the total coverage by primary service and the total travel distance to "uncovered" incidents which cannot be covered by regular lifeboats. For incidents inside the covering set which can be serviced by both the regular lifeboats and special boats, a preference is given to using regular lifeboats for response, while incidents which cannot be covered by any regular lifeboat must be responded to by special boats. We developed a modification to an existing coverage model with capacity limits and uncovered demand service (Pirkul and Schilling 1991) to make it suitable for Maritime Search and Rescue scenarios. We have two objectives: maximize the coverage by primary service and minimize the total distance to the incidents which cannot be covered by regular lifeboats. The multiple objective functions would be written as follows:

$$\text{Max} \sum_{i \in I} \sum_{j \in J} C_{ij} a_i x_{ij}, \ \text{Min} \sum_{i \in I} \sum_{k \in K} a_i d(i,k) y_{ik}$$

where:

a_i: is the number of incidents in grid i,
x_{ij}: 1 if an incident in grid i is assigned to a regular lifeboat in grid j; and 0 otherwise,
y_{ik}: 1, if an incident in grid i is assigned to a special boat in grid k; and 0 otherwise,
C_{ij}: 1 if grid i can be covered by a vessel in grid j; and 0 otherwise,
$d(i,k)$: is the distance from incident grid i to potential special boat position k.

The second objective can be expressed by a more practicable definition as (Pirkul and Schilling 1991):

$$Max \sum_{i \in I} \sum_{k \in K} v_{ik} a_i y_{ik},$$

where v_{ik} is the value of the service level for the incidents provided by secondary service:

$$v_{ik} = \begin{cases} 1, \text{ if } d_{ik} \le S \\ \alpha \left[1 - \frac{d_{ik} - S}{\max\{d_{ik} - S\}} \right], \text{ if } d_{ik} > S \end{cases}$$

S is the range of regular lifeboats and max d_{ik} is the maximum land-avoided distance from an incident grid to a potential position for special boats in our study area. If the

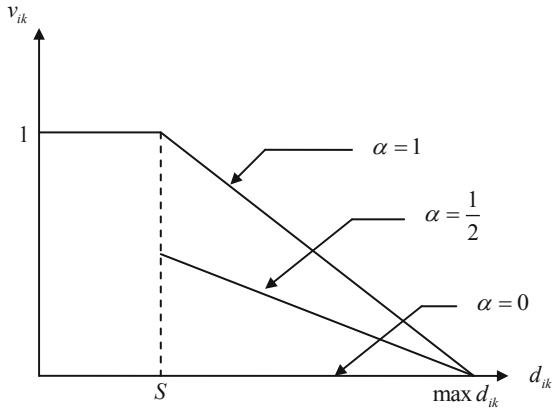


Fig. 16.10 Value of v_{ik} varying α from 0 to 1 (Pirkul and Schilling 1991)

distance is less than or equal to the range, we set $v_{ik} = 1$ representing full service. For the situation where distance is greater than the range, we let

$$v_{ik} = \alpha \left[1 - \frac{d_{ik} - S}{\max\{d_{ik} - S\}} \right], v_{ik} \in [0, \alpha), \alpha \in [0,1].$$

The longer the distance is, the lower the service level v_{ik} will be. The constant α acts as a relative weight tradeoff between the total incident number covered by regular lifeboats and the service level of uncovered incidents which are serviced by special boats. As α decreases from 1 to 0, less emphasis is placed on incidents beyond range S, and therefore more relative weight is given to incidents which are covered by regular lifeboats (Fig. 16.10).

To reflect that for an incident which can be served by both a regular lifeboat and a special boat, preference is given to the regular one, a penalty term on special boats is added to the objective function.

The model is as follows:

$$\text{Max} \sum_{i \in I} \sum_{j \in J} a_i x_{ij} + \sum_{i \in I} \sum_{k \in K} v_{ik} a_i y_{ik} - p \sum_{k \in K} s_k \tag{16.47}$$

$$\text{s.t. } x_{ij} + y_{ik} = 1 \; \forall \, i \in I, j \in J, k \in K \tag{16.48}$$

$$x_{ij} \leq r_j C_{ij} \; \forall \, i \in I, j \in J \tag{16.49}$$

$$y_{ik} \leq s_k \; \forall \, i \in I, k \in K \tag{16.50}$$

$$\sum_{i \in I} a_i x_{ij} \leq K_j^R \; \forall \, \text{j} \in \text{J} \tag{16.51}$$

$$\sum_{i \in I} a_i y_{ik} \leq K_k^S \; \forall \, k \in K \tag{16.52}$$

$$\sum_{j \in J} r_j + \sum_{k \in K} s_k = 60 \tag{16.53}$$

$$x_{ij} \in \{1,0\} \ \forall\, i \in I,\ j \in J;\ y_{ik} \in \{1,0\} \ \forall\, i \in I,\ k \in K;$$
$$r_j \in \{1,0\} \ \forall\, j \in J;\ s_k \in \{1,0\} \ \forall\, k \in K \tag{16.54}$$

where

I: the index set of all demand nodes,
J: the index set of all potential positions for regular lifeboats,
K: the index set of all potential positions for special boats,
a_i: the number of incidents in grid i,
K_j^R: the maximum demand that can be served by a regular lifeboat in grid j,
K_k^S: the maximum demand that can be served by a special boat in grid k,
x_{ij}: 1, if an incident in grid i is serviced by a regular lifeboat in grid j; and 0, otherwise,
y_{ij}: 1, if an incident in grid i is serviced by a special boat in grid k; 0, otherwise,

$$C_{ij} = \begin{cases} 1, & \text{if } d_{ij} \le S \\ 0, & \text{otherwise} \end{cases}$$

$$r_j = \begin{cases} 1, & \text{if a regular lifeboat is located in grid } j \\ 0, & \text{otherwise} \end{cases}$$

$$s_k = \begin{cases} 1, & \text{if a special boat is located in grid } k \\ 0, & \text{otherwise} \end{cases}$$

p: Penalty for locating special boats.

The objective function (16.47) is to maximize the total service level including both the incidents which can be covered by the regular lifeboats and the uncovered incidents which are serviced by special boats. Constraints of type (16.48) state that each incident grid must be responded to either by a regular lifeboat or by a special boat. Constraints of type (16.49) guarantee that incident i can be assigned to regular lifeboat j only if a regular lifeboat is located in grid j and the distance from i to j is within the range. Similarly, constraints of type (16.50) state that incident i can be serviced by a special boat at k only if there is a special boat located in grid k. The requisite workload limits for all potential positions are imposed by constraint sets (16.51) and (16.52). Finally, the total number of rescue vessels is restricted in constraint set (16.53).

Computational experiments show that the results obtained are sensitive to the value of p. Varying p from 0 to a very small value can make the number of special boats decrease greatly (Fig. 16.11). For example, setting p equal to 0.001 and $\alpha = 4$, the number of special boats drops dramatically from 29 to 10 compared with setting p equal to 0.

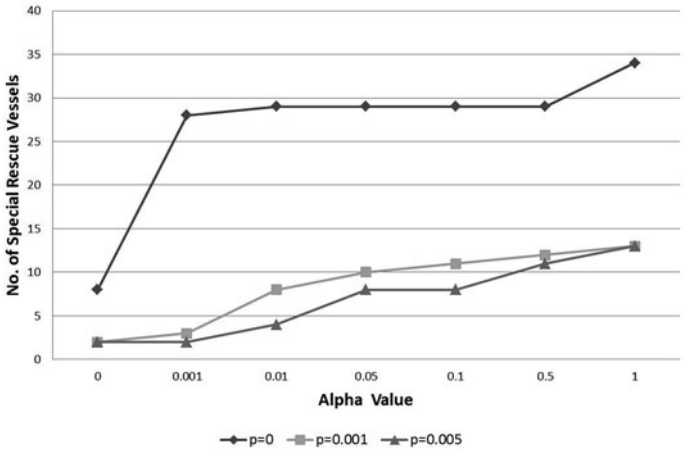


Fig. 16.11 Number of special boats by varying the value of p

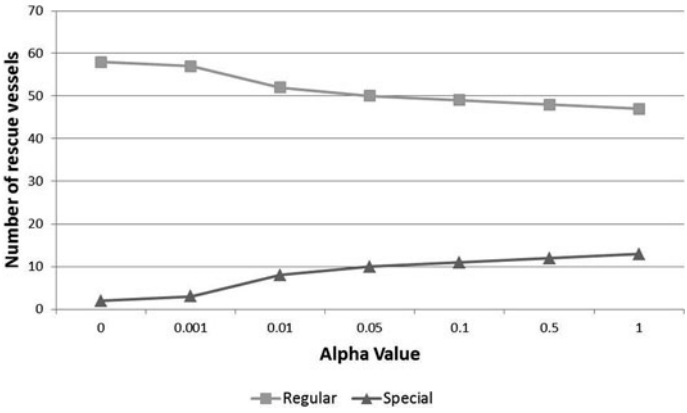


Fig. 16.12 Regular lifeboats and special boats located as a function of α

The solution configurations obtained when varying the value of α from 0 to 1 are summarized in Table 16.10. As α increases, more emphasis is placed on incidents beyond range S, hence less relative service value is given to incidents which are responded to by primary service (regular lifeboats). Therefore, the number of regular lifeboats located decreases and the number of special boats increases, while the coverage by regular lifeboats gets worse and the average distance of special boats improves. The solution results with $p = 0.001$ are graphically illustrated in Fig. 16.12 and Fig. 16.13.

Based on the above results from the coverage model with uncovered demand service, as the value of α increases from 0 to 1, the number of regular lifeboats decreases from 58 to 47, while the number of special boats increases from 2 to 13. The average distance of secondary service improves greatly without significantly

Table 16.10 Solutions by coverage model with uncovered demand service ($p = 0.001$)

Alpha	No. of regular lifeboats	No. of special boats	Coverage by primary service (%)	Avg. dist of special boats km
0	58	2	11467 (96.7)	1534.5
0.001	57	3	11449 (96.6)	528.7
0.01	52	8	11395 (96.1)	454.5
0.05	50	10	11366 (95.9)	437.5
0.1	49	11	11216 (94.6)	365.3
0.5	48	12	11035 (93.1)	304.4
1	47	13	11010 (92.9)	298.0

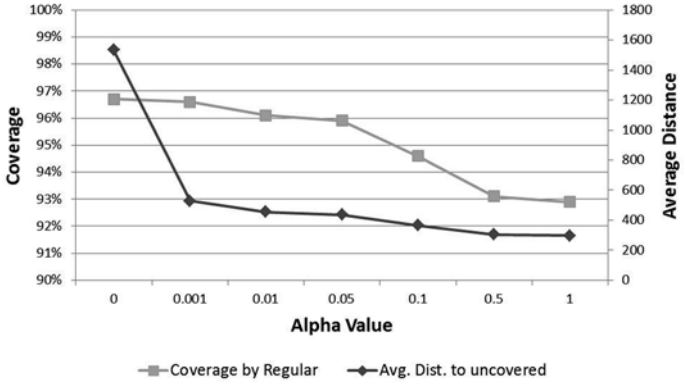


Fig. 16.13 Coverage by regular lifeboats and average distance of special boats

affecting the coverage of primary service. However, special boats are usually expensive to purchase and use. Therefore, we can obtain the solution for a prescribed number of special boats by adjusting the values of p and α. For example, if we have 12 special boats, we can set $p = 0.001$ and $\alpha \leq 0.5$.

16.5 Summary

Although the location–allocation problem has been thoroughly addressed by researchers in many applications, there is scope for further developments relating to resource allocation in the Maritime Search and Rescue field. Scenarios in Maritime Search and Rescue are similar to the emergency vehicle location problem except for the following two differences: a land-avoidance algorithm must be used to calculate the travel distance rather than Euclidean or Manhattan distance metrics; and response unit (vessel) capability needs to be considered. The review of previous research provided a framework of applicable methods for the maritime resource location problem.

The current study reviewed the methods and models in emergency vehicle location problem and provided the following results:

- Summarizes previous research on emergency vehicle location problems which can be useful for maritime resource location scenarios;
- Applies three optimization models of *MCLP*, *MEXCLP* and *CMCLP* to Maritime Search and Rescue for locating lifeboats, the principal response vessels of the Canadian Coast Guard;
- Modifies the coverage model with secondary service to consider special boats with (essentially) infinite range, considering tradeoffs between the two objectives of total response distance and the cost of locating and using regular lifeboats and special boats;
- Modifies the coverage model with uncovered demand service to consider special boats with unlimited range, tradeoffs between the two objectives of incident coverage by regular lifeboats (primary service) and average response distance to the "uncovered" demand (secondary service).

The models we applied and modified are either single objective or two objectives combined by some weight or coefficient. In the first modified model with secondary service, we put a weight W into the objective function to trade off between total response distance and fixed cost of locating rescue vessels. For the second modified model with consideration of uncovered incidents service, we use a coefficient α to determine the service level by special boats for responding to uncovered incidents. Also, in order to give the priority to primary service for the incidents which can be responded to by both primary and secondary service, we add a penalty term into the objective function associated with a penalty coefficient p. Although sensitivity analysis was conducted in this study, a drawback is that it is hard to determine reasonable values of weights or coefficients beforehand.

Typically, when some criteria get better, others will get worse. For instance, the *MEXCLP* and *CMCLP* models can give better response time and workload balance but worse total incident coverage than the *MCLP* model. Therefore, multi-criteria approaches could be applied to consider several criteria such as coverage, response time, workload balance, vessel utilization balance and budget simultaneously, with weights provided by subject matter experts.

Generally, the approach used in this study was to develop different mathematical location models for the case of Maritime Search and Rescue. Starting with simpler models that highlight the key elements of our *SAR* scenarios, the models progressively become more realistic by adding features and additional constraints. Of course adding constraints engenders a tradeoff of greater realism at the expense of a poorer objective value. Also, all the models do not have exactly the same objective. Although maximizing coverage is the main objective in most of the models, we have included other criteria such as cost and average distance from covered and uncovered demands as well. Hence, it is not useful to directly compare the solutions of different models to each other as they are not compatible. Overall, what we can say about the models' performance is that those incorporating multiple criteria and considering additional constraints such as capacity and server availability work well

in our case as they generate acceptable solutions with regards to the coverage compared to the basic models like *MCLP*. Ultimately, the best model to use among these alternatives could be selected based on the decision makers' needs and preferences.

In this research, we assume that all the grid squares can be used to locate rescue vessels, which can yield optimal solutions in our study area. However, optimal does not mean necessarily practical. In the actual system, rescue vessels are usually placed near the shore and around ports. The models in this research can be modified to be more practical by adding vessel position constraints based on more information about vessel location requirements from the Canadian Coast Guard.

While the model results described herein were not implemented directly by the Canadian Coast Guard, they do form part of a suite of investigations performed over many years by our MARIN (Maritime Activity & Risk Investigation Network) research group to address diverse aspects of maritime accident prevention and response. Our optimization models have been used to help guide decisions on the location of new lifeboat stations across the country, planning for opening of new waterways to remote areas, Search & Rescue Needs Analysis, and recommended adjustments to coverage when CCG vessels are out of service for planned or unplanned maintenance.

Acknowledgement The authors are grateful for the support provided by the Natural Sciences and Engineering Research Council of Canada for this work.

References

Azofra M, Perez-Labajos C, Blanco B, Achutegui J (2007) Optimum placement of sea rescue resources. Saf Sci 45:941–951

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

Ball M, Lin F (1993) A reliability model applied to emergency service vehicle location. Oper Res 41(1):18–36

Berman O (1994) The *p*-maximal cover—-*p*-partial center problem on networks. Eur J Oper Res 72:432–434

Berman O, Krass D (2002) The generalized maximal covering location problem. Comp Oper Res 29:563–581

Brotcorne L, Laporte G, Semet F (2003) Ambulance location and relocation models. Eur J Oper Res 147:451–463

Cameron A, Pelot R (2005) Search and rescue vessel capability study. MARIN group, Department of Industrial Engineering, Dalhousie University, Halifax

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

Daskin M (1983) A maximal expected covering location model: formulation, properties and heuristic solution. Transp Sci 17:48–70

Daskin MS, Dean LK (2004) Location of health care facilities. In: Brandeau ML, Sainfort F, Pierskalla W (eds) Operations research and health care: a handbook of methods and applications, Kluwer Academic Publishers, Drodrecht, pp 43–76

Goldberg JB (2004) Operations research models for the deployment of emergency services vehicles. EMS Manage J 1(1):20–39

Griffin PM, Scherrer CR, Swann JL (2008) Optimization of community health center locations and service offerings with statistical need estimation. IIE Trans 40(9):880–892

Harewood S (2002) Emergency ambulance deployment in Barbados: a multi-objective approach. J Oper Res Soc 53:185–192

IBM ILOG CPLEX http://www01.ibm.com/software/commerce/optimization/cplex-optimizer/

Jacobs DA, Silan MN, Clemson BA (1996) An analysis of alternative service areas of american facilities locations. Interfaces 26(3):40–50

Larson RC (1974) A hypercube queuing model for facility location and redistricting. Comput Oper Res 1(1):67–95

Li L (2006) Rescue vessel location modelling. Nova Scotia. M.A. Sc. in Industrial Engineering, Dalhousie University, Halifax

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

Nguyen BU, Kevin YKN (2000) Modeling Canadian search and rescue operations. Faculty of Administration, University of Ottawa, Ottawa

Pirkul H, Schilling D (1988) The siting of emergency service facilities with workload capacities and backup service. Manage Sci 37(7):896–908

Pirkul H, Schilling D (1991) The maximal covering location problem with capacities on total service. Manage Sci 37:233–248

Sinuany-Stern Z, Mehrez A, Tal A-G, Shmuel B (1995) The location of a hospital in a rural region: the case of the Negev. Locat Sci 3(4):255–266

Chapter 17
Military Applications of Location Analysis

John E. Bell and Stanley E. Griffis

17.1 Introduction

Over time, the decision-making needs of military commanders have had a strong influence upon the development of the field of operations research and analytic problem solving. The challenge of correctly positioning military units and resources within a geographical setting has vexed commanders and their staffs for thousands of years. However, it is only in the last 70 years that optimization methods have developed to the point where analysts can apply them to accomplish such goals. Along the way toward solving these narrowly defined military-focused problems, the advancement of the field has benefitted as generalizable techniques are extended beyond their origins to countless non-military applications. For example, military analysts first solved complex problems regarding routes for convoys of ships, code breaking, and materiel allocation mathematically during World War II ultimately leading to the development of linear and mathematical programming techniques by wartime scientists. Location analysis knowledge was similarly benefitted by the war, as commanders required the ability to spatially position munitions to destroy a target and covering a search area to find the enemy. The benefit is reciprocal however, as military strategy and planning since World War II have literally been redefined by the operations research field's ability to solve larger and more complex problems.

The dawn of the computer age and the explosion of the operations research field are understandably tied. Since the 1950s there has been a boom in problem solving ability, and a wide range of applications of operations research and location analysis methods have benefitted from the computing power, and increased analytic tools

J. E. Bell (✉)
Department of Marketing & Supply Chain Management, University of Tennessee, Knoxville, TN 37996, USA
e-mail: bell@utk.edu

S. E. Griffis
Department of Supply Chain Management, Michigan State University, East Lansing, MI 48824, USA
e-mail: griffis@broad.msu.edu

H. A. Eiselt, V. Marianov (eds.), *Applications of Location Analysis,* International Series in Operations Research & Management Science 232, DOI 10.1007/978-3-319-20282-2_17

to take advantage of this power. While military problems germinated many of the original techniques, over time we have seen these military problems fade into the background as industrial and urban planning priorities have fueled the research and applications of location modeling. Despite this general trend, military applications of location analysis remain an important and often unique contextual area for the application of solution methods. The military application area has seen a renewal in interest with increased number of publications over the past 15 years. In large part this is the result of geopolitical shifts brought on by the end of the cold war, and the post 9/11 realities for militaries. In particular, the current set of vexing problems possesses challenging decision variables, parameters and constraints unique to the current realities of military location modeling needs.

In this chapter, we will review the key military applications of location analysis published over the last several decades in the academic literature. In doing so, we will first outline the brief history and classic academic publications in the area. Then more recent military applications will be reviewed in order to highlight newer constraints and problems faced by 21st century military commanders. This review will lead to the development and presentation of a typology of common features that characterize military applications in location analysis. We will conclude with recommendations for further research and opportunities for location analysts.

17.2 A Brief History of Military Application of OR Techniques

The location of combat resources and the effect of geography on those military positions have played a critical part in military strategy and operations throughout history. Military theorists such as Sun Tzu, Clausewitz and Eccles have studied and emphasized the strategic importance of location decisions in the operations, tactics, and logistics of warfare (Griffith 1963; Howard and Paret 1989; Eccles 1959). In one famous historical example of the role of location in military strategy comes from the war between Rome and Carthage. The Carthaginians successfully maneuvered their forces across the desolate North African deserts to secure locations where vital food and water resources existed to sustain their armies (Roth 1998, p. 308). This was not chance on the part of the Carthaginian General Hannibal, he actively sought to induce the Romans into battle on a location of his choosing. The end result a rout of the Romans at Cannae (B.C. 216) in one of the most famous victories in military history. Location science and geography continued to create challenges for the Romans as the decline of the empire was brought on, in part, by the inability of the military to cover and secure its vast geographical areas. As a result, Rome failed to protect its holdings and hard-earned territories were abandoned (e.g., Britain). The basic covering problems that plagued Emperor Constantine and his military leaders have been revisited by modern location scientists including Stewart (1999), ReVelle and Rosing (2000), Henning (2003), and Cockayne et al. (2004), and have shown that solutions to the Roman maximal covering problem could have been achieved. Against this backdrop, we consider, “how have analytical models been used in recent history to make military location decisions?” Answering this question necessitates a

review of some of the prominent military applications of location analysis over the last several decades. This review reveals unique features of military application of location analysis, and highlights opportunities for continued research in this area.

Prior to the 20th century, strategic decision-making (including location analysis) was conducted primarily by the most senior military commanders. (Dantzig 1963, p. 12). In the 1830s, military science was advanced greatly by the Prussian theorist, Carl von Clausewitz. Despite being one of the most acclaimed military theorists in history, in his classic work, *On War*, Clausewitz specifically denied the usefulness of "geometric analysis" for the positioning of bases of operation and the locations needed to provide supplies to military forces, and indicated that such analysis could lead to "wrong tendencies"(Howard and Paret 1989, Ch. XV). However, by the time of World War I, planning for modern armies broadened to include a General Staff rather than a single leader to conduct planning processes and to help commanders make complex strategic decisions in shorter periods of time. This planning process, often called "programming", allowed broader views on planning issues than Clausewitz held, and grew significantly in the US military during World War II. These scientists and military analysts engaged in wartime programming that laid the groundwork for the development of the academic field of operations research. During World War II, many of the operational problems studied contained locational aspects including decisions for combat unit deployments to various geographic theaters around the globe (Dantzig 1963, p. 14). In addition, wartime analysts tackled problems to include developing methods to decide where to drop depth charges (location, depth, and pattern) to defeat enemy submarines, and the best methods to convoy merchant vessels across the Atlantic Ocean to avoid attacks (Morse and Kimball 1951). In addition, wartime scientists applied mathematical techniques and search theory for the "coverage "of a continuous space in order to find enemy targets within distinct geographic area (Morse and Kimball 1951, p. 86). However, this work was only a prelude to the advancement of operations research methods and the military applications in location analysis that were to come.

Following WWII, the further development of linear programming by the US Air Force, and the creation of the simplex method for generating solutions for linear programming by Dantzig were major steps forward for military problem solving. However, most of the early military applications of linear programming were primarily for dynamic program planning (Wood and Geisler 1951), routing of aircraft and ships, labor and equipment scheduling, and contract bidding (Dantzig and Fulkerson 1954; Jacobs 1955; Natrella 1955). To understand this apparent lack of attention on location analysis, an understanding of the focus of the Cold War military from the 1950s until the late 1980s is needed. As described by Bonder (2002), the focus of US Army operations analysis during this time was on tactical military operations and weapons system analysis. The locational aspects of future conflicts were considered relatively well known and stable for the US military and its scientists, the battles of the Cold War were expected in central Europe. In the US Navy, important work did continue through the 1960 and 1970s on problems such as screening for submarines and the proper escort positioning for ship convoys (Hughes 2002). But even then, the military's mission focused on countering the Soviet Union's larger forces with superior tactics and systems; while the location

of expected conflicts and needed facilities could be predicted with relative certainty (Bonder 2002). Given all the assumptions about where the fight would occur were 'resolved' advancement of location science within the US military stagnated as other, more pressing issues, were addressed. Reflecting this stagnation, relatively few military applications of location analysis appear in the literature in this time period.

Interestingly, while the military operations research scientists became less concerned with the location problem, academic location analysis picked up where they had left off, and the area developed rapidly. Access to the first computers in the early 1960s helped to spawn an explosion in location analysis techniques and many of the problem formulations, and algorithms were developed, and seminal articles published. Building on the economic works of von Thunen, Weber and Hoover and inspired by the seminal works of Hakimi (1964), Cooper (1963), Toregas et al. (1971) and others, location analysis methods exploded from the1960s through the 1980s. Through the end of the cold war period a wide range of location modeling applications in industrial, geographical science, urban planning, and communications emerge. As we will see in Sect. 17.3, commercial applications drive the location analysis agenda in the latter stages of the Cold War era, however, this would change with the close of the Cold War era.

Within a relatively short period of time between 1989 and 1992 the military focus of the world shifted (Bonder 2002). After the Cold War, geopolitical shifts resulted in renewed interest within the military to location analysis. Primary among these changes was uncertainty regarding the location of future conflicts. Throughout the Cold War, the location of the next potential conflict was fairly well assured, the evaporation of the USSR resulted in a dispersion of military might, and a resulting need to reconsider where countervailing forces needed to be positioned. Dozen of potential scenarios for future conflicts as well as the need to simultaneously support multiple conflicts (itself a significant shift in planning complexity) became priories for military planners (Bonder 2002; Bell 2003; Amouzegar et al. 2004a). The second major shift centered on the need to pre-position stores of military equipment and munitions as a way of scaling back the investments associated with standing, ready forces in foreign installations around the globe. This change in philosophy was most clearly seen in the European theater where dozens of facilities were permanently closed during the 1990s. At the same time, not only did the drawdown in forces affect the location problems confronting analysts, the objectives changed as well. In response to the "peace dividend", analysts were required to find more fiscally motivated solutions, unlike the "win at all costs" perspective of the Cold War. These factors combined to create the last shift, toward time-based contingencies as a priority for military planners. Now that fewer permanent locations were available around the globe, and the location of the next threat was unknown, and pools of supply were pre-positioned to address anticipated needs, analysts were confronted with developing models to facilitate the generation of rapid response forces to address short-notice 'needs' anywhere on the globe. As an example, the 1990s conflicts in Kosovo required US forces deploy to a new combat location, with relatively no notice, no logistical planning in place, and extremely short time windows. Further

adding to the complexity faced by analysts, the new realities of location analysis required greater consideration of the redeployment phase, that time period following the military engagement, when all the manpower and materiel needs to be returned and reset to their original locations (Bonder 2002). Taken together, and made even more acute after the terrorist attack of 9/11, all of these factors combined to create a need for new and different location analysis techniques, thereby reinvigorating the military's interest in the location sciences.

As we will show in Sect. 17.3, these new paradigms have resulted in a significant increase in the number of military location analysis applications in the literature since the late 1990s. As we review the primary military applications during the last several decades, this chapter will identify and describe those special attributes of military applications and the ever-changing conditions that create them.

17.3 Review of Military Applications

The common features and attributes of military operations research applications published in the recent literature appear in Table 17.1. Although, not all articles possess each attribute, trends in the research that distinguish military location applications from other areas become clear. For example, the multi-objective nature of military applications must consider not only the cost minimization and equitable service objectives seen in other location research (ie., the location of police and fire stations), but also the ability to maintain a secure and resilient operating location. This is because military locations themselves may also become the targets of opposing forces, and the capability and survival of the network also depends on the defense of locations within the network. Military models often treat factors with different severity than commercial models, stock-out costs being an example. While business models care about inability to cover or serve an area from the perspective of terms of lost sales or profit, for a military application an inability to serve a "demand" might translate into death for fielded troops or failed national objectives. Accordingly, a high premium is often put on achieving near perfect service levels in military applications, even at the expense of increased financial costs. Table 17.1 lists several other unique aspects of military applications, and classifies the articles reviewed in this section, based on the appearance of these aspects in the article.

17.3.1 The Cold War Years (1960 to 1990)

Military applications of location problems during the Cold War were fairly limited, compared to the preceding decades. This was due in part to military planners relative certainty as to the location of expected future conflicts with communist nations. Given the "base in place" finish to World War II, the location of forces in Western Europe and the Korean Peninsula were well known. Therefore, from the early

Table 17.1 Classification of military applications literature

Articles	Deterministic	Stochastic	Single Period (Static)	Multiple Period (Dynamic)	Multiple layers (Hierarchical)	Median (Min cost)	Coverage (Max service)	Dispersion (Avoidance)	Security (Protection)
Levin and Friedman (1982)	X			X		X		X	
Yamani (1990)	X		X			X			
Loerch et al. (1996)	X		X		X	X	X		
Murty and Djang (1999)	X		X		X	X	X		
Nguyen and Ng (2000)	X		X			X	X		
Segall (2000)		X	X			X	X	X	
Johnstone (2002,2004)	X		X		X		X		
Leinart et al. (2002)	X		X		X	X	X		
Bell (2003)	X		X			X	X		
Gue (2003)	X			X				X	
Flint et al. (2003)		X		X			X		
Kierstead and DelBalzo (2003)		X		X			X		
Amouzegar et al. (2004b, 2006)	X			X	X	X	X		X
Brown et al. (2005)		X	X				X		X
Farahani and Asgari (2007)	X				X	X	X		
Dawson et al. (2007)	X		X			X	X		
Ghanmi and Snow (2008)		X		X	X	X	X		
Overholts et al. (2009)	X			X		X	X		
Dell (2008)	X		X			X	X		
Killingsworth (2008)	X		X			X			

Table 17.1 (continued)

Articles	Deterministic	Stochastic	Single Period (Static)	Multiple Period (Dynamic)	Multiple layers (Hierarchical)	Median (Min cost)	Coverage (Max service)	Dispersion (Avoidance)	Security (Protection)
Chan et al. (2008)		X	X				X		
Kress and Royset (2007)		X		X			X	X	
Rappold and Van Roo (2009)		X	X			X			X
Salmeron et al. (2009)		X		X					
Ferrer (2010)	X		X			X	X		
Burks et al. (2010)	X			X		X	X		
Murphy et al. (2010)		X	X			X	X		X
Dong et al. (2010)							X		X
Ghanmi (2011)	X		X			X	X		
Toyoglu et al. (2011)	X			X		X		X	
Bell et al. (2011)	X		X			X	X		X
Brown et al. (2011)		X	X				X		X
Plastria (2012)	X		X			X	X		X
Tanerguclu et al. (2012)		X					X		
Qiu and Sharkey (2013)	X			X		X	X		
Dillenburger et al. (2013)		X	X				X	X	

1960s until 1990, the focus of much of the operational research modeling in the military was on other topics. In an overview of 40 years of military modeling, Hughes (1984) outlines the major application areas of military modeling during the Cold War period. Hughes outlines seven primary military application areas during this time period including (1) battle planning (2) wartime operations (3) weapon procurement (4) force sizing (5) human resource planning (6) logistics planning and (7) National policy analysis (Hughes 1984, p. 23). It is interesting to note that although location modeling and spatial decision-making are not considered major application areas within Hughes' framework, he does mention the strategic importance of facility location and layout decisions throughout this text. For instance, within the area of logistics planning, Hughes notes that the location and capacity of military maintenance and repair facilities is an important part of logistics modeling applications. In addition, he goes on to point out that much of the effort in facility location planning in these early years was geared towards the objective of peacetime efficiency (cost minimization) of military facilities and infrastructure, and that future military modeling would need to consider analytical models that focus on wartime realities including the achievement of the right level of "decentralization" of forces being threatened by an enemy, and the need for building-in excess capacity to facility networks in order to prepare for the uncertainties created by war (Hughes 1984, p. 32).

Given the reduced focus upon location analysis, resources were spent in other areas in this period. Inventory management research dug into issues related to improving repairable inventory decisions. Work by Zangwill (1966), Sherbrooke (1966), Muckstadt (1973), Hillestad and Carrillo (1980) and others focused on making the better decisions for inventory purchases to minimize the expected backorders in an inventory system, and to get the highest return for the inventory system given limited resources. Although inventory research was originally tactical in nature, over time, work by Muckstadt (1980) and Hillestad (1982) broadened to consider multiple levels or hierarchies of locations in the network, and then further extended to include fixed costs of facilities and variable capacities at different locations, resulting in more strategic models with integrated inventory and location decision-making (Drezner and Hillestad 1984, p. 211).

Transportation problems also moved to the foreground during this era. Following the work of Dantzig (1963) and Ford and Fulkerson (1954), military modeling projects addressed needs for algorithms to handle location-specific bottlenecks in transportation networks and to improve the network of locations (and routes) needed to transport military material and personnel (Garfinkel and Rao 1971; Berman et al. 1981). Such modeling highlighted the expense of building such transportation systems with the determination of the needed capacity to reach distant locations (Drezner and Hillestad 1984, p. 212). Foretelling the near-future, Drezner and Hillestad (1984, p. 215) observe that military planning in the cold war period "de-emphasized tactical conventional capabilities because of the nuclear umbrella. . ." and that the objectives, constraints and structure of future military models would have to take into account a much higher degree of uncertainty and that wars might

occur at any location on the globe. Additionally, the increased integration of various military models would be needed in the future, in order to understand important tradeoffs in military decisions, such as those that simultaneously account for location, inventory and transportation decisions by military decision makers (Drezner and Hilliard 1984, p. 218.)

In addition to direct military applications, several theoretical pieces were published on location modeling during this time period that were motivated by or directly related to military problems. First, it is interesting to note that in their review of private and public location models, ReVelle et al. (1970) stop short of classifying military applications in the public sector, along with civil defense, police stations and other emergency services. However, many of the aspects of the public location models discussed in their article do apply to military applications. For example, similar to other public location models the "the goals, objectives, and constraints" of military applications may not be "easily quantifiable . . . or easily defined." (ReVelle et al. 1970, p. 694). Additionally, policy, laws, political constraints and government budget restrictions often place unclear constraints on the feasible region of military problems. Such challenges cause military applications to often be multi-objective in nature, where according to ReVelle et al. (1970, p. 694), the hope is not "to be able to define solutions, but to at the least gain more information about the system under analysis." Therefore, understanding the tradeoffs in objectives and identifying the range of feasible solutions available to military commanders and political leaders is often an aspect of military location applications.

Another research area inspired somewhat by the military resource allocation challenges is the work of Kaplan (1974) to limit use scarce resources such as fuel, water, and ammunition to meet anticipated demands. Similarly, Litwhiler and Aly's (1979) theoretical work on the spherical Weber problem that consider non-Euclidean distances appears motivated by the need for improved mathematical methods for naval deployment and communications over large geographical areas, faced by the US military. Additionally, Woolsey's (1986) editorial contained mention of facility location modeling, and the Pentagon and the US military specifically. Although none of these articles utilize military-sourced data, they are motivated by military operations and continue to illustrate the need for improved location modeling techniques to improve real world military decision-making.

Levin and Friedman (1982) specifically apply dynamic warehouse location methods to deploy military units to the most effective locations and minimal cost locations. Their approach steps beyond theoretical mathematics to address a specific military problem and develop solutions for this problem. In doing so they specifically addressed how to determine the maximum number of support units needed to supply military forces, and where those units would need to be located. Interestingly, this application also considered military unit effectiveness decreases due to "exposure" to various hazards or threats when moved to a particular location. This aspect of the model also makes dispersion and survival of units a part of the objective.

Towards the end of this period, Yamani et al. (1990) extended work on the spherical Weber problem conducted by Drezner and Wesolowsky (1978), Litwhiler and Aly (1979), using US Air Force aircraft refueling data. This paper built a model of

single aircraft mid-air refueling with the objective of selecting refueling locations in order to minimize aircraft fuel consumption. Yamani et al. (1990) show that the problem faced by the military was actually a nonlinear convex location problem on a sphere, and presented an optimal procedure for generating solutions to the problem. This paper was a prelude to a large number of future military applications that would more deeply examine military oriented problems and apply actual military data in the process of generating solutions to those problems.

17.3.2 Transition in Focus: 1990–2001

The period from the beginning of Operations Desert Shield/Storm in 1990 until the terrorist strikes of 9–11 in 2001, represent a transition in military strategic planning that ushered in a renewed need for improved military location analysis. The sheer size of the forces and amount of equipment deployed to the Middle East during Desert Storm and the follow-on no-fly zones in Iraq during this period led to a re-examination of global military asset positioning. Given positioning of US military forces in Europe and Asia was done in anticipation of conflicts against anticipated Cold War opponents, most of the forces and equipment needed in the 1990s were poorly positioned to respond to needs in the Middle East. In addition, increasing pressure to decrease military budgets incentivized the nation and its military to focus more on smaller, more mobile forces, needed to respond quickly to a wide range of conflicts. During the 1990s, military researchers, including those at the *RAND* Corporation, focused on creating flexible and responsive approaches to location selection for the military. Global positioning and global reach became important aspects of military strategy, especially given constraints associated reducing the mobility "footprint" of forces to be moved during a future conflict. Accordingly, military location modeling applications during this period helped reflected an increased interest in location modeling in the US military.

One area the military began to reassess in the post Cold War environment was better ways to pre-position and transport critical wartime material. Loerch et al. (1996) studied the location strategy and positioning of US Army's facilities in Europe, making recommendations as to which bases to deactivate and close in the European theater of operations in the 1990s. Their integer programming model sought to minimize costs of the network subject to constraints related to minimum support and proximity (coverage). Their results provided insight into the cost trade-offs of opening and closing facilities in an pre-existing defense network, as well as resource utilization insights. Loerch et al. (1996) analysis was complicated by the fact that military units under consideration were required to located within a certain distance of their superior units, which are also being located in the model. Because of this requirement, the problem is more difficult to solve due to the multi-level hierarchy of locations at play simultaneously. Further complicating the analysis, support units in the model have different coverage requirements compared to combat units,

and capacities exist for both shared resources (equipment) and funding at the locations. Using multiple scenarios and data from the US Army the research was able to identify four alternate solutions to reach $ 49–58 million in savings annually in comparison to a manually generated solution created by Army staff experts.

Similar to efforts by the US Army, modeling efforts at the US Naval Postgraduate School by Anderson (1998) and Sentlinger (2000) were aimed at improving the locations for the storage of munitions by the US Navy. Anderson (1998) looked to improve weapons positioning across the Pacific region for the US Navy. Sentlinger (2000) considered heterogeneous weapons mixes with a mixed integer program to determine the best mix of weapons to preclude shortfalls in a number of potential combat scenarios. Both of these studies are limited-distribution to those working within US military circles.

Also during this time period, Murty and Djang (1999) developed a multiple layer strategy for the US's National Guard to locate mobile training units. In the first stage of their model, a *p*-median approach is used to select the best set of home bases for the mobile trainers. Then in the second stage, a set covering approach is used to select secondary training locations within a maximum distance of units needing training. Finally Murty and Djang (1999) determine routing to the secondary training locations. This study, the results of which were implemented by the US military, achieved a 70 % reduction in the expected annual mileages driven by the trainers and a significant cost savings of over $ 8 million yearly. The work stands as a classic example of the benefits achieved from applying location-modeling techniques to operational military problems.

Segall (2000) responds to a more mobile military and considers hospital location modeling, specifically the locations and capacity of a network of military hospitals needed to respond to chemical agent attacks. Segall's (2000) work describes a set of optimization models for determining the location of facilities when faced by six different cases of how medical facilities will be distributed over a planar region. Patient flows to military emergency facilities and the vulnerability of the hospitals and their staffs to the actual chemical weapons attacks that create the patients are also considered, as well as access for patients to care and cost minimization of facility damage; all while taking into account the spatial differences in the demand created by military combat scenarios.

Nguyen and Ng (2000) apply *p*-median and coverage techniques within a goal programming model to analyze search and rescue operations for the Canadian military. This approach was used to determine the service capability of the current military system and to make suggestions to the Canadian military for improving the non-optimal locations then being used by search and rescue aircraft, while prioritizing guaranteed service to as many distress incidents as possible within a critical coverage distance (similar to solving the maximal set covering problem). Additionally they sought to minimize the average distance to incidents that occur outside of the coverage areas of the selected bases, similar to a *p*-median approach. Using a simulated annealing metaheuristic Nguyen and Ng (2000) generate solutions to the problem and analyze the tradeoffs between maximizing coverage and minimizing

distances and costs. Then Nguyen and Ng (2000) use simulation and queuing theory to analyze the dynamic aspects of search and rescue operations by the Canadian military over large spatial areas. Military planners in Canada used the study's results to propose changes to their existing network and also applied findings from the follow-on simulation study to realize significant improvements in waiting time for rescues.

By the end of this era, work by Tripp et al. (1999) and others at the *RAND* Corporation began to consider issues related to the need for more agile combat logistics support framework for the US military. This research foreshadowed a continued shift in the US military away from permanent overseas facilities and toward a network of much smaller forward support locations and forward operating locations, positioned where potential conflicts might occur. This research specifically addressed the uncertainty of future conflict locations and the need for a flexible set of locations and operating capacities to handle multiple military combat scenarios at the same time. This type of flexible and agile thinking, and the changed-state of conflicts after 9–11 would help to trigger a significant increase in the number of military applications of location modeling.

17.3.3 Recent Military Applications of Location Modeling (Post 9–11 era)

The last few years have seen a veritable explosion of military application articles in the literature, as compared to the period prior to the new millennium. The relative uncertainty and instability of military operations in the 1990s renewed dormant interest in location analysis in contrast to the Cold War, when locations were considered more fixed (Bonder 2002). However, the last decade has seen the development of new strategic approaches such as "Flex-Basing" and "Forward Support Locations" proposed to the US military by researchers at the *RAND* corporation (Tripp et al. 1999; Killingsworth 2000; Geller et al. 2004). In addition, the US government's Base Realignment and Closure Commission (*BRAC*) sponsored and motivated location research to analyze which US military bases should remain open, and which should be closed. Changes to military basing strategies and the efforts to improve the global network of military locations led to an increase in military applications of location analysis by researchers. The remainder of this section has been broken into several sub-categories that detail the studies of this time period based on their scope and objectives. Similar to the previous sections, the common features and attributes of specific military applications are listed in Table 17.1.

17.3.3.1 Strategic Global Networks

Building a global network of locations to respond to uncertain conflicts became a priority for the US military at the beginning of the millennia. Conceptual pieces began to call for an improvement in the selection of forward operating locations around the globe for the military and for an improvement in the positioning of difficult to move military equipment (Tripp et al. 2000; Underwood and Bell 1999).

In order to consider building such an improved global network, modeling research by Johnstone (2002) (and Johnstone et al. 2004) examined the location and deployment decisions for Air Force munitions inventories used to respond to crises around the world. The objective of this model was primarily to minimize the time to move munitions, but it also considered and selected the best supply points for the pre-positioning of assets needed in potential conflicts. In addition, the work by Johnstone (2002) emphasized the continued need for afloat positioning of wartime material on ocean going vessels. In essence, afloat prepositioning allows for mobile supply points that can be moved in international waters towards areas where demands may occur. This model examined the selection of ship locations and ports for off-loading material at the regional level, and also looked at a lower hierarchy for determining the subsequent inland movement of the material of selected ports. Like many effectiveness based military models, the model aimed to maximize the primary objective, service (minimize delivery times), and ignored cost, the stochastic nature of demands, and the various dynamic time-phased issues with moving munitions over long periods of time.

Bell (2003) built upon Johnstone's work, and used facility location modeling techniques to develop a strategy for constructing a global network of munitions locations for the US Air Force. Bell (2003) developed a combined minimal cost and maximum coverage solution to address the multi-item facility location problem. Bell accomplished this by selecting warehouse locations from a discrete set of potential locations, while simultaneously determining multi-item inventory levels at the selected locations. Due to the size of the problem, simulated annealing was used to find near optimal solutions to the problem. Using this approach, it became evident that both cost and coverage objectives could be improved compared to the initial solution provided by Air Force and efficient frontiers of solutions were mapped in order to more fully understand the tradeoffs in cost and coverage.

RAND Corporation researchers have also been involved in the reconsideration of military basing decisions. Amouzegar et al. (2004a) conducted a study of locations for future forward support bases that rely upon a wide range of war reserve materials stored around the globe. The study considered both land and sea-based locations around the globe and also considered co-locating U.S. Army and U.S. Air Force support facilities. The developed multi-item optimization model considered various capacities at different locations and aims to minimize global system costs while meeting dynamic time-based service requirements for varying conflict scenarios. The model includes actual transportation constraints and distances and considers the allocation of demand points to selected locations (location-allocation).

Additionally, this military application of location modeling was one of the first to include the vulnerability and defense of the basing options in the model. Aspects of this stream of research and the resulting resource allocation modeling were also detailed in Amouzegar et al. (2004b). Additionally, the same location modeling approach was later reapplied to an expanded set of scenarios and military planning considerations (Amouzegar et al. 2006).

A separate study by Killingsworth et al. (2008) considered the selection and costs of modifying an existing supply chain network for a military helicopter system in use by the U.S. Army. This analysis considered the repair capabilities of the current system across the globe and considered multiple scenarios for altering the network and selecting a new repair center location in the network. The problem was formulated as a discrete minimal cost network flow problem and the cost savings and payback period, for investment in a new repair facility, were considered. This effort identified significant annual cost savings ($9–18 million), with payback periods ranging from 2 to 6 years.

The Canadian military also faces the challenge of rapidly deploying military capabilities to virtually any location around the globe with minimal time and costs. Ghanmi and Shaw (2008) developed a multi-objective model minimizing cost and response time for deploying Canadian forces around the globe. Like their American colleagues, the Canadians also considered the pre-positioning of military assets at forward locations in advance of demands, while also trying to reduce problems with military airlift. Ghanmi (2011), further developed a discrete facility location model to select "support hub" locations with the multiple objectives of cost and capability. Formulated as a mixed integer non-linear program, the work of Ghanmi (2011) is based in the facility location problem literature; however it uses a unique objective function to achieve the maximum relative cost avoidance when determining the optimal number and location of hubs in the final solution.

17.3.3.2 Regional Facility Networks

In addition to assessing global networks in military applications of location modeling, a number of regionally focused models have appeared in the literature in the last decade as well. Primarily, these applications have addressed the need to build more-detailed repair and distribution networks to be controlled by a single military commander. An interesting feature of many of the regionally focused models is that they simultaneously solve routing or inventory problems as part of the facility location problem.

Gue (2003) addresses building a network of discrete supply locations to support the small and highly mobile forces of the U.S. Marine Corps. Gue (2003) addresses multiple time periods, and seeks to minimize the actual quantity of inventory that is positioned on seaborne supply points that are mobile and are consistent with their seaborne logistics support. The problem is a dynamic mixed period facility location and multiple commodity flow problem, and it is formulated as a mixed integer program. Decision variables for this model include the candidate list of locations in a

region, the inventories to position at land and sea locations, and the inventory flow of items to combat locations. The model is tested against demands created from actual combat scenarios expected by the U.S. Marine Corps. In addition, the analysis allows the military users to consider tradeoffs between the use of land and sea based supply locations, and the impact of the distance between the seaborne supply points and the shoreline. As a military application, Gue (2003) represents an early dynamic location model, with the analysis of the tradeoffs between mobile (sea-based) and fixed (land-based) supply locations in the same regional network.

Farahani and Asgari (2007) use a multi-objective set covering approach to determine the best location for military distribution centers from a discrete set of current warehouse locations while minimizing cost and improving the utility (quality) of selected locations. The utility objective is unique in that it has twenty-four separate attributes. To address this Farahani and Asgari (2007) adapt multiple attribute decision-making (*MADM*) techniques to assess this objective within the model. Additionally, the model is multi-echelon in nature considering the location of supply points (factories), warehouses, and supported locations. In doing so, the model aims to create sub-groups of supported locations that can only be supplied by one warehouse; while still allowing direct shipments from the factory that by-pass the warehouse (Farahani and Asgari 2007, p. 1841). Results are presented using an efficient frontier of the two objectives as minimal cost and maximum utility objectives are in conflict, and tradeoffs must be considered. The results of the research show that the military was able to reduce their critical coverage distance in comparison to existing military practice, and were able to improve the balance of workload among locations.

More recently, Rappold and Van Roo (2009) build a non-linear integer model for a multi-echelon engine repair network for the US Air Force. They combine facility location and inventory allocation problems and attempt to minimize the cost of the repair system and simultaneously determine the correct level and allocation of inventory across locations for the entire system. Using a two-step optimization approach, this study initially selects the minimal cost locations for the repair network, and then determines the finite repair capacities to set for each location in order to balance supply and demand in the system. Although the study is formulated for a small region, the work could be extended to global networks, within general complexity limits. Rappold and Van Roo's (2009) work is an extension of multi-echelon inventory problems in the military first studied by Sherbrooke (1966) and Muckstadt (1973), and it makes unique contributions in that the demand for the system is stochastic and finite capacities are put on the total inventory in the system. This application provides a tool for military decision makers to improve acquisition decisions for purchasing optimal numbers of expensive inventory items (such as aircraft engines) to stock at a network of repair locations.

Dell et al. (2008) consider manpower issues related to military units and personnel to be assigned to hundreds of U.S. Army locations around the United States. The Base Realignment and Closure (*BRAC*) commission in 2005 sponsored this research, and it resulted in the closure of 13 active duty Army installations and literally hundreds of Army Reserve facilities around the nation. Dell et al. (2008) model

aims to minimize costs for the US Army while also considering the military value (service) achieved by the optimal solution, similar to previous work by Ewing et al. (2006). The model is able to open, close or modify the capacity of current facilities, while simultaneously determining how to allocate Army units across the selected locations, and uncovered potential savings to the military of $7.6 billion over two decades.

Similar savings was sought in a study by Ferrer (2010), who considered improving the location of spare aircraft engines for the US Air Force's F-16 fighter aircraft within the continental United States. The application of the *p*-median approach by Ferrer (2010) included a separate analysis for the two different engine types used on F-16 aircraft (Pratt & Whitney and *G&E*) and identified the best set of locations for storing each type of engine that met the distance constraint of the model. An open architecture approach was considered that combined the two types of engines into one inventory pool, assuming interchangeability of the engines. Ferrer's (2010) analysis identified potential direct savings by the US military of $58 million and a 48 % reduction in safety stock.

Burks et al. (2010) studied the Location Pickup and Delivery Problem with Time Windows (*LPDPTW*) that simultaneously considered the location of depots for a number of vehicles that would provide time definite delivery and pickup of materials throughout a geographic region. Simultaneously, vehicle assignments to depot locations and allocation of customers to those depots was determined. The study was motivated by the US Army's wartime experiences in the Persian Gulf region and the need to minimize its logistics costs in future operations while maintaining wartime capability. Additionally, the model minimized delivery "lateness", and system-wide costs. The problem is also organized based on a set of tree layers (acyclic directed graphs) and therefore analyzes and captures the hierarchical nature of the Army's distribution network (Division, Brigade, and Battalion support levels). Making it even more challenging, Burks et al. (2010) considered multi-commodity, temporal, and facility capacity issues in addition to time factors. Solutions for this complex hybrid problem were generated using an adaptive Tabu search methodology, based upon Glover and Laguna (1997), which provides dynamic updates within the search heuristic to ensure both intensification and diversification during the search of the feasible region. The approach was demonstrated on six different scenarios to provide a realistic range of demands that might occur in actual military operations.

Toyoglu et al. (2011) also considered a dynamic environment, addressing a hierarchical location routing problem based on the daily distribution of ammunition to combat military units in the field of battle. This detailed problem is formulated as a multi-layer, dynamic location routing problem. It considered a heterogeneous fleet of vehicles with capacitated as well as fixed and mobile supply locations, Within their model, Toyoglu et al. (2011) further consider exposure of delivery vehicles to enemy fire and modify driving distances to take this into account. Additionally, several unique aspects of this paper make it stand out. First, the model not only includes three echelons of facility locations, but it also simultaneously considers location decisions at two of these echelons; a feature rarely seen in the literature. In addition, the model allows for multiple sourcing where more than one military supply point

or transfer point can fill the demand at a particular combat location. Lastly, hard time windows are imposed, making it even more realistic for military combat operations where success and life and death can be determined by the on-time delivery of ammunition to front line combat units.

Overall, as a group, the military applications of military modeling in recent years (including these regional applications) have begun to answer the call for studies of Hughes (1984) to be more focused on wartime rather than peacetime objectives. Recent military location studies are more focused on wartime combat objectives and are also more integrated with the related routing and inventory decisions being made by military commanders. Additionally, dynamic and hierarchical problems and multi-sourced problems are adding to both the complexity and realism of the military applications of location analysis.

17.3.3.3 Coverage Models in the Military

The primary objective of most studies in Sect. 17.3 so far has been minimizing costs or distances (facility location problem or *p*-median type formulations) in the model. While coverage or improved service has appeared as a secondary objective or constraint in those models they have generally not been the focus. The next section considers those military applications that have focused primarily on the coverage objective and have used models from the classic location literature such as the maximal location covering problem, set location covering problem or *p*-center problem.

Chan et al. (2008) determine the optimal location for signal receiving stations used for military search and rescue operations. In addition to selecting locations, the problem also allocates receiving equipment to each location and determines the listening frequencies to use. Because coverage areas during search and rescue need to be able to triangulate the position of a distress signal, simple single coverage is insufficient, and all areas require triple coverage. The study uses a multi-objective linear integer program (*MOLIP*) to approximate the original non-linear formulation of the problem, and a network flow approach is used to generate solutions for two cases studies provided by the US Department of Defense. Analysis of solutions to this problem finds significant improvement (four standard deviations) in comparison to randomly generated baseline solutions.

Bell et al. (2011) studied basing locations for military fighter aircraft used for homeland defense. This research was motivated to examine which military airfields the U.S. should consider critical for homeland defense against airborne threats, like those that struck on 11 Sept 2001. This study was intended to help the U.S. Department of Defense and Department of Homeland Security make positioning decisions and to ensure that key locations were not closed by *BRAC*. The first stage of the modeling approach was based upon on the set covering problem, where aircraft launch times and speeds where also used to determine critical coverage distances in the model. This stage identified the minimum number and location of sites for positioning military response aircraft to cover a set of sixty-six critical locations.

Sensitivity analysis during this stage made it evident that alternate optimal solutions existed to the covering problem, and therefore the second stage of the model used a *p*-median approach to find the minimum aggregate distance solution from among those alternate optimal solutions to the location set covering problem. Results of the study were briefed to decision makers across the US Air Force; however, the names of locations and impact on actual decision making cannot be reported due to the sensitive (classified) nature of the problem.

Following this study, Plastria (2012) offered a third stage to Bell et al. (2011) work, based on a *p*-center-sum approach (Hansen et al. 1994). Plastria (2012) points out that original formulation allows the critical distances to differ based on different aircraft speeds and levels of protection. Plastria (2012) also creates a new critical "protection" measure substitutable for the critical distance measure to account for these differences in the model.

In another interesting military application, Tanerguclu et al. (2012) used location modeling techniques and *GIS* data to improve the position of air defense weapons and radar systems. This problem maximizes the expected coverage of potential air attack routes (demands) by air defense systems while considering the probability of a defense system successfully shooting down an airborne threat. A spatial decision support system was developed to help quickly make decisions about weapon locations. The problem is complicated because vertical geography can obscure a defense systems view of the sky it is designed to protect. This is why the link to *GIS* data and their topographical mapping capabilities are important and unique to this problem. Tanerguclu et al. (2012) base their maximal expected coverage model upon Daskin (1983), however it differs in two important ways. First there are two different types of facilities being located (weapons and radars) and a demand location is only considered covered if it is visible and within the range of both type of facilities. Second, the model is hierarchical in nature and multiple weapon locations may be assigned to more limited set of radar locations. Tanerguclu et al. (2012) uses *LINGO* optimization software to generate solutions and dramatically improved upon the paper maps solutions previously used by military commanders.

Finally, in one of the most recent military applications of coverage models is by Dillenburger et al. (2013), who consider dropping equipment for military and humanitarian aid missions by air. They seek to find the optimal location of dropping equipment based on minimizing collateral damage and maximizing recovery of equipment. Collateral damage is a concern for airdrop operations; especially in humanitarian aid operations where dropping large amounts of supplies by air into densely populated areas can create higher levels of damage on an already beleaguered population. Therefore, the model in this research assigns weights to areas that should be avoided. A number of search algorithms for generating solutions to the problem are tested by Dillenburger et al. (2013), and these algorithms are variations of either a differential evolution or response surface algorithms. In order to aid military decision-making, the paper presents the results of 14 different potential military scenarios, which lead to a number of suggested guidelines for military planners conducting airdrop operations.

17.3.3.4 Mobile or Dynamic Supply Locations

In addition to the more traditional location models, several military applications have considered mobile demand or supply locations in their model in order to more accurately replicate military operations. The facility location research of Dawson et al. (2007) considers military security teams positioning in order to protect periodic repair operations on the military's nuclear intercontinental ballistic missile (*ICBM*) fleet. The security (supply) locations that provide protective coverage for a repair (demand) location in the problem are mobile and are repositioned as needed. In the problem the ability of the military to repair an *ICBM* at a demand location is dependent on the current location of security teams in the missile field. Dawson et al. (2007) use a combined *p*-center and *p*-median approach for the problem and employ a two-stage heuristic method to generate solutions. Their first stage finds the *p*-center solution that minimizes the maximum distance to any demand location, and the second stage attempts to find the minimum aggregate distance solution that maintains the maximum distance found in the first stage *p*-center solutions.

Extending the work of Dawson et al. (2007), Overholts et al. (2009) studies the *ICBM* maintenance problem and presents a different modeling approach that changes the position of security coverage each day to allow for scheduling of repair of nuclear missiles over multiple time periods. One of the authors of this chapter took part in this study, and therefore we are able to provide a more detailed look at this particular work.

The work of Overholts et al. (2009) is particularly interesting as it considers the dramatic changes in military operations that have occurred in the last 15 years and shows how these changed constraints may lead to different analytical approaches to problems. For example, before 2003 maintenance schedules for *ICBMs* located in the vast missile fields of Wyoming, Montana and the Dakotas were developed by hand. Though not guaranteed to be overly efficient, these schedules were feasible due to plentiful security personnel to cover operations. However by 2004, security requirements for nuclear facilities increased, simultaneously with decreasing security personnel availability due to budgets cuts and deployments to the Middle East. This resulted in challenges for maintenance schedulers who needed to build a schedule that could support increased maintenance requirements despite fewer security personnel required to escort the maintenance crews. Further, complicating this problem is the fact that not all missile maintenance actions are equivalent, therefore the research team had to work with the Air Force technicians in Wyoming to develop a new weighting scheme with 18 separate categories to give priority in the model to the most important repair work. Though constrained by a need to develop a model useable by non-academics (necessitating an *EXCEL* model) Overholts et al. (2009) developed a two-stage model to first select security alert locations, and then build a daily maintenance schedule. The resulting first stage of the model is based on the maximal covering problem formulation of Church and ReVelle (1974) and aims to cover the maximum amount of weighted demand by locating security teams at different discrete locations throughout the missile field. The second stage of the

model, also based upon Church and ReVelle (1974), attempts to assign or schedule the highest amount of weighted maintenance actions during the day at different missile locations. In order to test its ability, the new approach is compared to the actual maintenance schedules that were utilized by the US Air Force in Wyoming during a 26-day period in 2005, and significant improvements in maintenance effectiveness are observed with this modeling approach. Following the completion of the analytical study, the results of the study were presented to Air Force leaders, and transfer of the modeling capabilities to Air Force personnel was accomplished. As an applications effort, this study highlights the need for close communication and involvement of practitioners in the actual location modeling work, and the eventual transfer of control of the newly developed analytical method back to the decision makers who will use it.

A recent study by Qiu and Sharkey (2013) addressed an integrated location and inventory planning problem motivated by mobile military sea-based supply ships. The objectives of this problem include trying to minimize costs subject to minimum service levels, measured by centrality of the mobile supply locations (ships) among customers in different time periods. Complicating the problem, the demand points (ships) are mobile across a time horizon, with variable inventory demand in each discrete time period. Qiu and Sharkey (2013) employ dynamic programming and tested on a number of problems based on actual support of humanitarian crises in Haiti in 2010. Overall, Qiu & Sharkey provide one of the newest military applications that is grounded heavily in classic literature related to dynamic facility location problems, and that highlights the mobile nature of supply locations in many military problems.

17.3.3.5 Target Detection in a Spatial Area

Another area of military operations that has benefitted from operations research involves spatial and location decision- making, includes finding and detecting military targets over a large geographic area. The roots of this military application go back to World War II, when mathematical models were first used to aid in the detection of aircraft, submarines and other enemy threats in a combat zone. It is important to note that this line of research continues today, and given the globalization of threats, is as important as ever. Flint et al. (2003) conduct a study on how to coordinate the search for enemy targets with multiple unmanned aerial vehicles (*UAVs*). The stochastic nature of target detection in a bounded spatial area is addressed in Flint et al. (2003) using dynamic programming methods to formulate solutions to this problem. Kierstead and DelBalzo (2003) also consider finding moving targets while searching for enemy submarines. Although general techniques for the stationary detection problem have been around since the 1940s, Kierstead and DelBalzo (2003) consider dynamic targets in an environment where space and time are continuous. Applying a genetic algorithm Kierstead and DelBalzo (2003) find solutions that provide improvement of up to 46 % over standard search techniques.

More recently, Kress and Royset (2007) provided a location oriented search problem, to help special operations teams find the maximal number of targets they can serve using *UAV*'s to help them conduct the local search. The problem is based on the location-routing problem where the team locations are determined prior to the search routing of the *UAV*s. However, this model includes additional constraints such as the topography of the region, communications problems between vehicles, and the location of potential threats in the region that are not included in most *LRP* formulations. The results of the model were tested in actual US Army field operations in 2006, and resulting in up to a 50 % improvement in target detection compared to manual solutions generated by experienced military field commanders.

There are a growing number of studies related to searching for targets using spatial modeling techniques, and these studies are closely related to facility location modeling. This is one area where location modelers, especially those interested in dynamic and stochastic techniques, might find interesting problems to study that have direct military applications.

17.3.4 Military Location Modeling with Dispersion and Security

Military applications of location modeling sometimes also need to consider enemy threats to resources and personnel. Although some of the models reviewed up to this point included these constraints in their models, work has been done that considers dispersion and security the primary focus of the research. While security and protection are always a focus of military commanders, this objective has received more attention since 9–11. McGarvey and Cavilier (2003) study facility placement under constraints that include "forbidden" regions. This problem was motivated in part by the need of military commanders to avoid opposing military forces and to keep hazardous materials out of civilian areas. Additionally, the work of Church and Scaparra (2007 extends the *r*-interdiction median problem to include fortification of the most important locations in a network that can potentially be disrupted or attacked by outside forces. Similarly, the work of Scaparra and Church (2008) uses a bi-level mixed integer program for critical infrastructure protection planning, while Liberatore et al. (2011) presents a model to optimally allocate defensive resources among existing facilities in order to minimize the impact of terrorist attacks. This stream of modeling research, which includes facility network interdiction and protection of locations has widespread application. Dong et al. (2010) considers multiple coverage layers of demand locations faced with limited protection resources in a network. The model seeks to maximize the maximum coverage that protected facilities provide to demand locations, when the unprotected locations in the network are destroyed.

Salmeron et al. (2009) considers the military planning of distribution operations when certain locations in the transportation network are threatened. Using a stochastic mixed integer model, Salmeron et al. (2009) specifically looks at ocean ship scheduling and routing when some of the seaports in the network are vulnerable to attack by the enemy, such as the deployment of material during Operation Desert

Storm in 1990–1991. The model is primarily a dynamic routing problem that considers both multiple ship and inventory types that must flow from supply to demand locations in the network. However, attacks or disruptions may occur stochastically at a location in the network, and the objective function of the model aims to minimize the resulting total disruption in the network. Although, levels of protection are not yet considered in this problem, it may form the basis for continued work on hybrid location routing problems that consider enemy attacks and needed protection at critical nodes in the network.

17.3.5 Competitive Location Modeling: Attackers and Defenders

Military commanders not only consider how to defend their own networks, but also consider how to compromise a competitor's network. For example, Leinart et al. (2002) developed a vertex cut-set algorithm on the transformation of a graph representing a network to attack and intentionally disrupt the command, control and communications network of an opposing enemy force. In particular, they considered the best way to disrupt the facilities, radio towers, equipment, satellites, that support the flow of information needed by the opposing forces commanders. The objective of the model is to disrupt a set of target locations that possesses the maximum value (cost vs. benefit) for the military operation.

Other military applications take into account dynamic actions by both attackers and defenders to compete against each other in a geographical (spatial) area. For example, research by Brown (2005) creates a decision tool for the prepositioning of missile defense assets by the US military and its allies in Japan (defenders), taking into account the potential attack of locations in Japan by ballistic missiles in North Korea (attackers). This problem is an instance of a Stackelberg game, which the authors represent as a bi-level integer linear program (e.g. Moore and Bard 1990). The optimization model consists of an inner objective where the attackers attempt to maximize the damage at enemy locations and the outer objective where the defenders attempt to minimize the maximum amount of damage the attacker can achieve. This bi-level program is converted to a mixed integer linear program for solution purposes, and solved using *CPLEX* 9.0. The location aspects of this problem are interesting, as the model optimizes the positioning of a limited number of defender weapons at a candidate set of locations in Japan, based on the positioning decisions of the attacker who locates his weapons in North Korea. The model is further complicated by the inclusion of multiple weapon types with different coverage ranges, as well as considering the various probabilities that a defender's missiles actually successfully intercept an inbound attacker missile. Additionally, the study analyzes the value of secrecy in six different scenarios where the attacker and defender have more or less information about the location of the other side's weapons prior to the actual attack. This military application was successfully presented to the US Navy and US Strategic Command and was integrated into their actual planning and operations for ballistic missile defense.

Murphy et al. (2010) also consider the dynamic missile defense topic, but for U.S. homeland defense. This model considers the simultaneous moves of both attackers and defenders and also considers the positioning of a limited set of resources to protect a larger number of critical locations dispersed across a large geographic area. The formulation of the model follows von Neumann and Morgenstern (2004), where the attacker attempts to maximize expected gains and the defender attempts to minimize expected losses. Murphy et al. (2010) present two computational methods for reducing the size of the problem to make it more manageable, and they include in their analysis the ability to reduce costs in the system through improved strategic solutions. This military application highlights the difference between Cold-War (large detectable bombers) and post-Cold War (small stealthy cruise missiles) threats. The paper analyzes the effectiveness and cost of different military defense strategies for defense scenarios ranging from 5 to 40 US cities. One of the main contributions of this work is to point out the difficulty involved with assigning "value" to different military targets. It also highlights that the solutions selected are sensitive to different risk tolerances of the decision maker. For instance, using a risk-tolerant Nash approach, the effectiveness of the models solutions may be improved in comparison to more static and risk-averse approaches that only consider the most critical locations.

Finally, in another military application that included competitive actions, Brown et al. (2011) use an attacker-defender model to study a problem faced by the US Navy, Coast Guard and Department of Homeland Defense as they attempt to defend seaports from terrorist vessel attacks. This problem analyzes both the path that vessels take as they travel the ocean in an attempt to attack vital seaport operations, and the defenders capability to detect those attacking vessels through the optimal positioning of its own defending vessels and on-shore radar sites. The problem is stochastic and the objective is to minimize the probability that one or more attack vessels will evade detection by the defender. An interesting aspect of this problem is the geographic structure of individual ports is included with restrictions on port navigation and observation based on the physical layout of the port area. Solutions to this model show that despite the attacker having complete knowledge of the defenders vessels and facilities, the defender can still detect nearly 100 % of the attacking vessels in scenarios representing defense of the ports in Hong Kong and Bahrain.

The military applications in this section show that facility and resource location decisions are imbedded in the attacker defender models being used by militaries to defend against terrorist attacks and potential missile attacks in the modern post-cold war error. These models contain aspects of coverage and protection of locations and often contain stochastic aspects including the probability of detecting an adversary or the probability of a location surviving an enemy attack. Overall, these models represent some of the most complex and advanced wartime application of location and spatial modeling.

17.4 Military Applications Features

The articles that appear in Table 17.1 reveal several trends in military location analysis. To better understand these applications, a typology of features of military applications is presented, including a list of some dominant trends and attributes seen in location models in the literature to date.

17.4.1 Military Applications Typology

1. *Multiple Objective*. The majority (twenty five out of thirty six) of the location models reviewed pursued more than one objective. Different combinations of cost, coverage, dispersion, and security objectives were commonly present. While many classic problems address a single objective it is evident that modern military location problems are more complex, and often require a multi-objective approach.
2. *Global Attributes*. Many of the applications reviewed in the literature have constraints or attributes related to the global nature of military problems. Both land and sea-based facility locations were common in the applications. Additionally, the complexity of facility location problems for the military is higher than ever, considering cross border transportation bottlenecks, multiple transportation modes, short customer response requirements vs. long distances, unknown construction costs in foreign countries, and potential cross-loading of materials being moved between facilities in location routing applications. The global environment present-day militaries operate within are far more complex that of the middle 20th century when an assumption that an occupied area would be fully controlled by the occupying military. The modern military locations look more like islands within a given geography, rather than points within a secure border.
3. *High Stock-Out Costs*. In the military applications reviewed, it was clear that meeting demand from a supply facility was critically important, and doing so was assured through the outright prohibition of stock-outs, or the establishment of artificially high stock-out cost when not filling demand. This makes sense for combat support models, given stock-outs can result in failed missions, lost lives and undermining political objectives of the nation. Additionally, the location of facilities was also impacted by time sensitive nature of demand requirements from military forces in the field.
4. *High Demand Variability*. Military operations are often characterized by high demand variability. Sometimes there are long periods of zero demand, followed by spikes in demand caused by combat operations or readiness training. Such demand is atypical in commercial operations, and therefore military models need to consider forecasts of demand at locations where previous demands have never occurred. Such demand has been called “lumpy” by supply chain academics and can exhibit high variability in both spatial and temporal dimensions. As such,

thirteen out of the thirty-six articles reviewed employed stochastic models to consider these highly variable demand patterns.

5. *Minimize Logistics Footprint*. A common objective in modern military operations is to limit the amount of equipment and logistics support moved to a location to support a military conflict. The weight and scope of such logistics support is often called the "footprint" and a common objective is to minimize this. As military applications seek to select facilities for peacetime operations, they must also consider the footprint of logistics support that would have to be deployed abroad to other locations in the case of a new conflict. This is especially important as military cargo is often bulky, hazardous, and restricted by political factors from being transported through certain locations when en route to a conflict (e.g. the French prohibiting over flight to *NATO* aircraft during the bombing of Libya in the 1980s).
6. *Competition*. A key feature of many military applications is that enemy forces often compete against and attempt to destroy or disrupt the location network created by military decision makers. Therefore, military decision makers must be concerned with not only where the facilities are best located for operations, but also how they will be protected and secured. This may involve considering the location of enemy forces. Surprisingly, only a limited number of the studies described in this chapter (eight out of thirty-six) included security or protection features as part of their location model.
7. *Dynamic*. Military applications must constantly deal with changing demand and supply characteristics within their models. As political interests and threats change in the global environment, the dynamics of how to locate, adjust, or close facilities across multiple time periods becomes of interest. Similarly, in a more tactical combat environment, military applications that look at the dynamic movements of friendly and enemy forces through a geographic area impacts the decisions related to the placement of facilities, equipment and other resources. Twelve out of the thirty six papers reviewed in this chapter had dynamic aspects included in their location model.
8. *Hybrid and Hierarchical Models*. As the ability to solve more complex models has increased, added realism has been added to military applications, including combining location decisions with the closely related transportation routing and inventory management decisions common to a supply chain network. Our research sees a heavy emphasis on hybrid models that include routing and inventory optimization being accomplished simultaneously with military location decisions. Often these models consider location decisions at multiple layers in a hierarchical network where units and commanders have a span of control in their own operations that is tied to higher-level decisions made at the regional or even global network level. Eleven of the thirty-six military applications reviewed here contained some sort of hierarchical relationship in their location model.

These eight areas are representative of the attributes of military applications in location modeling and provide a baseline for those initiating any new military location

analysis effort. Together they represent a unique typology that separates this application area from other areas, and highlights the complexity and challenges of solving location problems in a military context.

17.5 Conclusion and Future Opportunities

It is believed that the dynamic and changing nature of the global environment and military operations will continue to make this application area an important context for location analysts in the coming years. Although, the typology presented in Sect. 17.4 has listed several important features and trends in military location modeling, the research reviewed in this chapter also points to several areas that are underdeveloped and that can be expanded upon. In addition, we believe that the expanding nature of military missions in a changing world provides several emerging opportunities not yet included in published military applications. Below we describe several of the areas that offer opportunities for future research on military applications of location analysis.

- *Increasing Multiple Objectives*. Although many of the military applications reviewed in this chapter take a multiple objective approach, most look at only a combination of two objectives, and usually this includes cost and coverage. Opportunities still exist to include more objectives of protection and dispersion in military location applications. In addition, work that includes and balances three or more of these objectives simultaneously are much more realistic and complex problems and are currently missing from the literature.
- *Humanitarian Missions*. The past decade has seen an increased number of military operations related to providing humanitarian aid in the event of natural disasters around the globe. This instance brings unique challenges for location modelers who must rapidly design response supply chain networks when some or the entire existing facility infrastructure and transportation network may have been damaged by the disaster. Additionally, this means that the locations that are selected and the resources deployed have to maintain a self-sustaining capability since the underlying supply infrastructures in the disaster area may also be destroyed or degraded. This topic area offers unique opportunities for military and non-profit aid organizations to work together to make location decisions in the face of a crisis.
- *Homeland Defense*. Following the attacks on September 11th, 2001, the US military went through a major paradigm shift as it refocused its efforts on homeland defense. In the Cold-War homeland defense often included external detection capabilities in and around the coastal areas of the United States. Now, location studies for homeland defense must consider border protection, but must also consider the positioning of defense assets within the nation to protect important national resources. It is believed that this topic will grow in importance.
- *Urban Planning and Environmental Impacts*. Military location analysis and problem solving may often interact or even disrupt location modeling efforts

by urban planners. For example, Westervelt and White (2009) and Johnson et al. (2011), have shown that hazardous military operations can overlap with urban development areas and actually disrupt those efforts and the underlying environment. Therefore, there is an expanded need to look at joint planning between military and urban planners to understand how urban growth and the need for new public facilities such as hospitals, police stations, parks and housing areas may overlap into military locations where hazardous military materials and training operations already exist.

- *Reverse Logistics Network Decisions*. The military has a long history of designing multi-echelon repair networks for heavy equipment items such as aircraft, tanks and ships. However, as the need for environmental protection grows, the location of facilities that creates a more sustainable network is growing in importance. Military location modeling efforts that attempt to minimize fuel consumption levels and carbon emission levels for operations in the network will grow in importance as fuel prices and scarcity continue to grow in the future.
- *Resource Security*. As the global population and economies continue to grow over the coming decades, military location planners must consider the potential competition for limited natural resources needed by the nation. Although competition with the enemy has been considered at a tactical level in a number of military applications, future military applications should also consider the competition between opposing military supply chains seeking to secure valuable resources desired by others. It is believed that this resource security objective will continue to grow in importance in military applications as decision makers attempt to secure access to ever more precious resources such as minerals, fresh water, fuel, and strategic metals that are dispersed in a non-uniform patterns across the world.

In conclusion, we believe that many of the classic military problems still persist, though with added complexity factors previously unseen. Additionally, we believe new classes of problems, with objectives unlike the classic military problems, are emerging, and merit consideration by researchers interested in this area.

References

Amouzegar MA, Tripp RS, McGarvey RG, Chan EW, Roll, Jr RC (2004a) Supporting air and space expeditionary forces: analysis of combat support basing options. RAND Corporation, Santa Monica CA

Amouzegar MA, Tripp RS, Galway LA (2004b) Integrated logistics planning for the air and space expeditionary force. J Oper Res Soc 55(4):422–430

Amouzegar MA, McGarvey RG, Tripp RS, Luangkesorn L, Lang T, Roll RC Jr (2006) Evaluation of options for overseas combat support basing. MG-421. RAND Corporation, Santa Monica CA

Anderson EB (1998) Optimizing ammunition movement in support of US Pacific fleet's positioning plan. Master's Thesis, Naval Postgraduate School

Bell JE (2003) A simulated annealing approach for the composite facility location and resource allocation problem: a study of strategic position of US Air Force munitions. Doctoral Dissertation, Auburn Univ, AL

Bell JE, Griffis SE, Cunningham WA III, Eberlan JA (2011) Location optimization of strategic alert sites for homeland defense. Omega 39:151–158

Berman MB, Halliday JM, Carrillo MJ, Moore NY (1981) Combat benefits of a responsive logistics transportation system for the European theater, R-2860-AF. RAND Corporation, Santa Monica CA

Bonder S (2002) Army operations research-historical perspectives and lessons learned. Oper Res 50(1):25–34

Brown G, Carlyle M, Diehl D, Kline J, Wood K (2005) A two-sided optimization for theater ballistic missile defense. Oper Res 53(5):745–763

Brown G, Carlyle M, Abdul-Ghaffar A, Kline J (2011) A defender-attacker optimization of port radar surveillance. Nav Res Logist 58:223–235

Burks RE, Moore JT, Barnes JW, Bell JE (2010) Solving the theater distribution problem with tabu search. Mil Oper Res 15(4):5–26

Cooper L (1963) Location-Allocation Problems. Oper Res 11:331–343

Chan Y, Mahan JM, Chrissis JW, Drake DA, Wang D (2008) Hierarchical maximal-coverage location-allocation: case of generalized search-and-rescue. Comput Oper Res 35:1886–1904

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

Church RL, Scaparra MP (2007) Protecting critical assets: the r-interdiction median problem with fortification. Geogr Anal 39(2):129–146

Cockayne EJ, Dreyer PA Jr, Hedetniemi SA, Hedetniemi ST (2004) Roman domination in graphs. Discret Math 278:11–22

Dantzig GB (1963) Linear programming and extensions. Princeton University Press, Princeton New Jersey

Dantzig GB, Fulkerson DR (1954). Minimizing the number of tankers to meet a fixed schedule. Nav Res Logist Q 1(3):217–222

Daskin MS (1983) A maximum expected covering location model: formulation, properties, and heuristic solution. Transp Sci 17(1):48–69

Daskin MS (2008) What you should know about location modeling. Nav Res Logist 55:283–294

Dawson MC, Bell JE, Weir JD (2007) A hybrid location method for missile security team positioning. J Bus Manag 13(1):5–19

Dell RF, Ewing PL, Tarantino WJ (2008) Optimally stationing army forces. Interfaces 38(6): 421–435

Dillenburger SP, Cochran JK, Cammarano VR (2013) Minimizing supply airdrop collateral damage risk. Socio-Econ Plan Sci 47:9–19

Dong L, Xu-Chen L, Xiang-Tao Y, Fei W (2010) A model for allocating protection resources in military logistics distribution system based on maximal covering problem. 2010 International Conference on Logistics Systems and Intelligent Management, Harbin China, 98–101

Drezner SM, Hillestad RJ (1984) Logistics models: evolution and future trends. In: Hughes WP Jr (ed) Military modeling. The Military Operations Research Society, Alexandria

Drezner Z, Wesolowsky GO (1978) Facility location on a sphere. J Oper Res Soc 29:997–1004

Eccles HE (1959) Logistics in the national defense. The Stackpole Co, Harrisburg

Ewing PL Jr, Tarantino W, Parnell GS (2006) Use of decision analysis in the army base realignment and closure (BRAC) 2005 military value analysis. Decis Anal 3(1):33–49

Farahani RZ, Asgari N (2007) Combination of MCDM and covering techniques in a hierarchical model for facility location: a case study. Eur J Oper Res 176:1839–1858

Ferrer G (2010) Open architecture, inventory pooling and maintenance modules. Int J Prod Econ 128:393–403

Flint M, Fernandez E, Polycarpou M (2003) Stochastic models of a cooperative autonomous UAV search problem. Mil Oper Res 8(4):13–32

Ford LR Jr, Fulkerson DR (1954) Maximal flow through a network, RM-1400-PR. RAND corporation, Santa Monica

Garfinkel RS, Rao MR (1971) The bottleneck transportation problem. Nav Res Logist Q 1(3): 465–472

Geisler MA, Wood MK (1951) Development of dynamic models for program planning. In: Koopmans TC, Cowles Commission Monographs (eds) Activity analysis of production and allocation, vol 13. Wiley, New York, pp 189–215

Geller A, George D, Tripp RS, Amouzegar MA, Roll RC Jr (2004) Supporting air and space expeditionary forces: analysis of maintenance forward support location operations. MG-151. RAND corp, Santa Monica CA

Ghanmi A (2011) Canadian forces global reach support hubs: facility location and aircraft routing models. J Oper Res Soc 62:638–650

Ghanmi A, Shaw RHAD (2008) Modelling and analysis of Canadian forces strategic lift and prepositioning options. J Oper Res Soc 59:1591–1602

Glover F, Laguna M (1997) Tabu search. Kluwer Academic Publishers, Norwell

Griffith SB (1963) Sun Tzu The art of war. Oxford University Press, Oxford

Gue KR (2003) A dynamic distribution model for combat logistics. Comput Oper Res 30:367–381

Hakimi SL (1964) Optimum locations of switching centers and the absolute centers and medians of a graph. Oper Res 12(3):450–459

Hansen P, Labbe M, Minoux M (1994) The *p*-center sum location problem. Cah du C.E.R.O. 36:203–220

Henning MA (2003) Defending the roman empire from multiple attacks. Discret Math 271: 101–115

Hillestad RJ (1982) Dyna-METRIC: dynamic multi-echelon technique for recoverable item control, R-2785-AF. RAND corporation, Santa Monica CA

Hillestad RJ, Carrillo MJ (1980) Models and techniques for recoverable item stockage when demand and the repair process are non-stationary, part I: performance measurement, N-1482-AF. RAND corporation, Santa Monica CA

Howard ME, Paret P (1989) Translated version of Clausewitz C (1832) On War. Princeton University Press, Princeton

Hughes WP Jr (1984) Military modeling. The Military Operations Research Society, Alexandria

Hughes WP Jr (2002) Navy operations research. Oper Res 50(1):103–111

Jacobs WW (1955) Military applications of linear programming. In: Antosiewicz HA (ed) Proceedings of the second symposium in linear programming, and Directorate of Management Analysis 2. National Bureau of Standards, Washington DC, pp 1–27

Johnson S, Wang G, Howard H, Anderson AB (2011) Identification of superfluous roads in terms of sustainable military land carrying capacity and environment. J Terramech 48(2):97–104

Johnstone DP (2002) Modeling the pre-positioning of air force precision guided munitions. Master's Thesis, Wright-Patterson AFB OH: Air Force Institute of Technology

Johnstone DP, Hill RR, Moore JT (2004) Mathematically modeling munitions prepositioning and movement. Math Comput Model 39(6):759–772

Kaplan S (1974) Application of programs with maximin objective functions to problems of optimal resource allocation. Oper Res 22:802–807

Kierstead DP, DelBalzo DR (2003) A genetic algorithm applied to planning search paths in complicated environments. Mil Oper Res 8(2):45–59

Killingsworth PS, Galway L, Kamiya E, Nichiporuk B, Ramey TL, Tripp RS, Wendt JC (2000) Flexbasing: achieving global presence for expeditionary aerospace forces. MR-1113-AF. RAND corporation, Santa Monica

Killingsworth WR, Berkowitz D, Burnett JE, Simpson JT (2008) The application of supply network optimization and location analysis to a DOD repair supply chain. Def Acquis Rev J 15(3): 277–291

Kress M, Royset JO (2007) Aerial search optimization model (ASOM) for UAVs in special operations. White paper report. Naval Postgraduate School, Monterey CA

Levin KD, Friedman Y (1982) Optimal deployment of logistic units in dynamic combat conditions. Eur J Oper Res 9:41–46

Liberatore F, Scaparra MP, Daskin MS (2011) Analysis of facility protection strategies against an uncertain number of attacks: the stochastic *r*-interdiction median problem with fortification. Comput Oper Res 38:357–366

Leinart JA, Deckro RF, Kloeber JM Jr., Jackson JA (2002) A network disruption modeling tool. Mil Oper Res 7(1):69–77
Litwhiler DW Jr, Aly AA (1979) Large region location problems. Comput Oper Res 6:1–12
Loerch A, Boland N, Johnson E, Nemhauser G (1996) Finding an optimal stationing policy for the US army in europe drawdown. Mil Oper Res 2(4):39–51
McGarvey RG, Cavalier TM (2003). A global optimal approach to facility location in the presence of forbidden regions. Comput Ind Eng 45(1):1–15
Melo MT, Nickel S, Saldanha-da-Gama F (2009) Facility location and supply chain management—a review. Eur J Oper Res 196:401–412
Moore JT, Bard JF (1990) The mixed integer linear bilevel programming problem. Oper Res 38:911–921
Morse PM, Kimball GE (1951) Methods of Operations Research. Peninsula Publishing, Los Altos California
Muckstadt JA (1973) A model for a multi-item, multi-echelon inventory system. Manag Sci 20(4):472–481
Muckstadt JA (1980) Comparative adequacy of steady state versus dynamic models for calculating stockage requirements, R-2636-AF. RAND corporation, Santa Monica
Murphy EM, Payne MD, van der Woude GW (2010) Strategy alternatives for homeland air and cruise missile defense. Risk Anal 30(10):1507–1519
Murty KG, Djang PA (1999) The US army national guard's mobile training simulators location and routing problem. Oper Res 47(2):175–182
Natrella JV (1955) New applications in linear programming. Comput Autom 4(11):22
Nguyen BU, Ng KYK (2000) Modeling Canadian search and rescue operations. Mil Oper Res 5(1):5–16
Overholts DL II, Bell JE, Arostegui MA (2009) A location analysis approach for military maintenance scheduling with geographically dispersed service areas. Omega 37:838–852
Plastria F (2012) A note towards improved homeland defense. Omega 40:244–248
Qiu J, Sharkey TC (2013) Integrated dynamic single-facility location and inventory planning problems. IIE Trans 45:883–895
Rappold JA, Van Roo BD (2009) Designing multi-echelon service parts networks with finite repair capacity. Eur J Oper Res 199:781–792
ReVelle CS, Rosing KE (2000) Defendens imperium romanum: a classical problem in military strategy. Am Math Mon 107(7):585–594
ReVelle CS, Marks D, Liebman JC (1970) An analysis of private and public sector location models. Manag Sci 16(11):692–707
Roth JP (1998) The logistics of the Roman Army at war (264 B.C.—A.D. 235). Brill, Leiden
Salmeron J, Wood K, Morton DP (2009) A stochastic program for optimizing military sealift subject to attack. Mil Oper Res 14(2):19–39
Scaparra MP, Church RL (2008) A bilevel mixed-integer program for critical infrastructure protection planning. Comput Oper Res 35:1905–1923
Segall RS (2000) Some quantitative methods for determining capacities and locations of military emergency medical facilities. Appl Math Model 24:365–389
Sentlinger BK (2000) Optimal positioning of naval precision guided munitions. Master's Thesis, Naval Postgraduate School
Sherbrooke CC (1966) METRIC: a multi-echelon technique for recoverable item control. Oper Res 16(1):122–141
Stewart I (1999) Defend the Roman Empire! Scientific American 281:136–138
Tanerguclu T, Maras H, Gencer C, Aygunes H (2012) A decision support system for locating weapon and radar positions in stationary point air defence. Inf Syst Front 14:423–444
Toregas C, Swain R, ReVelle C, Bergman L (1971) The location of emergency service facilities. Oper Res 19(6):1363–1373
Toyoglu H, Karasan OE, Kara BY (2011) Distribution network design on the battlefield. Nav Res Logist 58:188–209

Tripp RS, Galway L, Killingsworth PS, Peltz E, Ramey TL, Drew JG (1999) Supporting expeditionary aerospace forces: integrated strategic agile combat support planning framework. MR-1056-AF. RAND corporation, Santa Monica CA

Tripp RS, Galway L, Ramey TL, Amouzegar M, Peltz E (2000) Supporting expeditionary aerospace forces: a concept for evolving the agile combat support/mobility system of the future. MR-1179-AF. RAND corporation, Santa Monica CA

Underwood DK, Bell JE (1999) AEF munitions availability. Air Force J Logist 23(4):12–17

Westervelt JD, White MJ (2009) Identifying future locations for noise-generating military training opportunities within urbanizing landscapes. Mil Oper Res 14(1):53–60

Wood MK, Geisler MA (1951) Development of dynamic models for program planning. In: Koopmans TC (ed) Activity analysis of production and allocation. John Wiley & Sons, New York, pp 189–192

Woolsey G (1986) The fifth column: on the minimization of need for new facilities, or space wars, lack of presence, and Delphi. Interfaces 16:53–55

von Neumann J, Morgenstern O (2004) Theory of games and economic behavior. 60th anniversary edition. Princeton University Press, Princeton

Yamani A, Hodgson TJ, Martin-Vega LA (1990) Single aircraft mid-air refueling using spherical distances. Oper Res 38(5):792–800

Zangwill WI (1966) A deterministic multi-product, multi-facility production and inventory model. Oper Res 14(3):486–507

Index

H. A. Eiselt, V. Marianov (eds.), *Applications of Location Analysis,* International Series in Operations Research & Management Science 232, DOI 10.1007/978-3-319-20282-2