Chapter 1

Introduction to Location Theory and Models

1.1 INTRODUCTION

If you ask what to look for in buying a house, any realtor will tell you that there are three things that are important: location, location, and location. The theory behind this answer is that the community in which you elect to live and the location within that community are likely to affect your quality of life at least as much as the amenities within your house. For example, if you live within walking distance of the local elementary school, your children will not need to be bused to school. If you live near a community center, you may be able to avoid involvement in car pools taking children to and from activities. If your house is too close to a factory, noise, traffic, and pollution from the factory may degrade your quality of life.

Location decisions also arise in a variety of public and private sector problems. For example, state governments need to determine locations for bases for emergency highway patrol vehicles. Similarly, local governments must locate fire stations and ambulances. In all three of these cases, poor locations can increase the likelihood of property damage and/or loss of life. In the private sector, industry must locate offices, production and assembly plants, distribution centers, and retail outlets. Poor location decisions in this environment lead to increased costs and decreased competitiveness.

In short, the success or failure of both private and public sector facilities depends in part on the locations chosen for those facilities. This book presents methods for finding desirable or optimal facility locations.

Network and Discrete Location: Models, Algorithms, and Applications, Second Edition.
Mark S. Daskin.

We should emphasize from the beginning that the word "optimal" is used in a *mathematical* sense. That is, we will define quantifiable objectives that depend on the locations of the facilities. We will then identify algorithms (rigorous procedures) for finding optimal or at least good facility locations.

Two factors limit the broader optimality of the sites suggested by the optimization models discussed in this text. First, in many cases, nonquantifiable objectives and concerns will influence siting decisions to a great extent. Often, the qualitative factors that influence siting decisions are critically important. Thus, to the extent that the procedures discussed in this text ignore qualitative concerns and factors, the sites identified by the mathematical algorithms are *optimal* only in a narrow sense of the word. Second, the performance of a system is affected by many factors of which location is only one. For example, the ability of an ambulance service to save lives (an objective that many would attribute to such a service) depends not only on the proximity of the ambulance bases to the calls for service (which can be measured and optimized to some extent) but also on such factors as the training and skill of the paramedics, the public's knowledge of emergency medical procedures and when it is appropriate to call for an ambulance, the existence of a 911 emergency line, and the protocols and technologies employed by the paramedics.

In the face of (1) exogenous qualitative concerns that influence siting decisions and (2) nonlocation factors that affect the performance of facilities, one might legitimately ask, "Why bother developing mathematical location models?" There are a number of answers to this question. First, while location is not the only factor influencing the success or failure of an enterprise, it is critical in many cases. Poorly sited ambulances will lead to an increased average response time with the associated increase in mortality, that is, more deaths. Second, while exogenous qualitative factors will influence siting decisions, mathematical models allow us to quantify the degradation in the quantifiable objectives that comes from recognizing the qualitative concerns. Thus, if it is important to locate an ambulance in one district for political reasons, the increase in average response time (or maximum response time) resulting from the imposition of this political constraint can be quantified. Third, the modeling process (identifying objectives and constraints and collecting data) often improves the decisions that are made even if the models are never run. Fourth, there are nonlocation problems in which models identical to those discussed in later chapters arise. For example, the problem of selecting tools in a flexible manufacturing context is mathematically

similar to that of locating ambulances in a city (Daskin, Jones, and Lowe, 1990).

Section 1.2 outlines a number of key questions that are addressed in location models. Section 1.3 extends the discussion of ambulance location problems and introduces some of the terminology used in location modeling. In this section, we also describe qualitatively another location problem—that of locating landfill sites for solid wastes. Section 1.4 identifies a number of key dimensions along which facility location models can be constructed. Section 1.5 outlines a taxonomy of location models based largely on the underlying topology of the space in which the demands and facilities are embedded. After outlining the taxonomy in Section 1.5.1, Section 1.5.2 develops a simple analytic location model from which insight about key tradeoffs in location problems can be derived. Finally, Section 1.6 summarizes the chapter.

1.2 KEY QUESTIONS ADDRESSED BY LOCATION MODELS

Mathematical location models are designed to address a number of questions including:

(a) How many facilities should be sited?

(b) Where should each facility be located?

(c) How large should each facility be?

(d) How should demand for the facilities' services be allocated to the facilities?

The answers to these questions depend intimately on the context in which the location problem is being solved and on the objectives underlying the location problem. In some cases, such as ambulance siting problems, we will want to locate the facilities as near as possible to the demand sites. In locating radioactive waste repositories, we will want to be in a geologically stable region and would like to be as far as possible from major population centers.

The number of facilities to be located as well as the size of the individual facilities is often a function of the service/cost tradeoffs. In many cases, not only does the quality of service improve as the number of facilities located increases but the cost of providing the service also increases. For example, having more ambulances is generally preferable to having fewer ambulances, since the likelihood of having an

available ambulance near a call for service increases with the number of vehicles deployed. In addition, there are no significant economies of scale in ambulance operations, that is, having a single site with multiple ambulances is not much cheaper than having the same number of ambulances at multiple locations. Thus, having a large number of single vehicle ambulance bases is likely to be preferable to having a small number of multivehicle bases. At some point, however, the quality of medical care provided by the paramedics *degrades* as more ambulances are added to the system. The reason for this is that there are not enough demands requiring a variety of medical skills to maintain the paramedics' training. In many manufacturing contexts, there are significant economies of scale, which drive the location decisions toward having a smaller number of large facilities.

Facility location models are also concerned with the allocation of demands to facilities. In some cases, it is important that demands at a site not be split between facilities. For example, in some retailing operations, a retail store must be supplied by a single warehouse. For administrative reasons, the store's supply cannot be split between different warehouses. In other cases, such as ambulance services, demands can be served by any available facility. Facility location models must reflect these different demand allocation policies and must then allocate demands (or fractions of the total demand in a region) to different facilities. In many cases, demands will be allocated to the nearest (available) facility; in other cases, doing so may not be optimal.

1.3 EXAMPLE PROBLEM DESCRIPTIONS

In this section, we outline a number of different facility location contexts and qualitatively define some of the classical location problems.

1.3.1 Ambulance Location

As indicated above, poor ambulance locations can cost lives! To illustrate this point, a commonly cited statistic is that if a person's brain is denied oxygen for more than 4 min (e.g., as a result of a stroke or heart attack), the likelihood of the individual surviving to lead a normal life drops below 50%. This suggests that we would like to locate ambulances so that the maximum response time is well under 4 min. Thus, one objective might be to *minimize the number of ambulances needed so that all demand nodes are within a given number of minutes (the*

service standard) of the nearest ambulance. Such a model formulation is known as a *set covering model*. Demands are said to be *covered* if the nearest ambulance is located not more than X min away, where X is the service standard used in the model (e.g., 4 min). Set covering models have been used by a number of authors in locating ambulances and other emergency service vehicles (e.g., Toregas et al., 1971; Walker, 1974; Plane and Hendrick, 1977; Daskin and Stern, 1981; Jarvis, Stevenson, and Willemain, 1975).

One of the common problems associated with the set covering model is that the solution is likely to call for locating more vehicles than the community can afford. If we deployed one less vehicle and relocated the remaining vehicles to maximize the number of demands that can be served within the given service standard (e.g., 4 min), the fraction of the demands that would not be serviceable within the service standard would generally be far less than $1/N$, where N is the number of ambulances called for by the set covering model. In other words, the last few ambulances add relatively little to the fraction of demands that can be served within the service standard but add significantly to the cost of the ambulance service. This suggests an alternate objective: *maximize the number of demands that can be covered within a specified service standard using a given number of vehicles*. Such a model is known as a *maximum covering model*. In practice, the fleet size that is input into such a model is often varied from 1 up to the number required for full coverage as indicated by the set covering model. This allows us to trace out the tradeoff between additional vehicles and coverage. Such a curve is shown in Figure 1.1. In this example, ten vehicles are needed to cover all demands. In other words, the solution to the set covering model for this problem is ten vehicles. The maximum covering model would then be solved for one through nine vehicles in the system. Notice that the incremental coverage *decreases* as additional vehicles are added to the system. Maximum covering models and their variants have also been used in analyzing ambulance systems and related emergency services (e.g., Daskin, 1982, 1983; Eaton et al., 1985; Church and ReVelle, 1974; Belardo et al., 1984).

In some cases, logical choices for the service standard might not be readily available. The choice of 4 min in the discussion of the covering models was predicated on the observation that irreversible brain damage is likely to occur if the brain is denied oxygen for more than 4 min. However, this does not necessarily imply that 4 min is the appropriate service standard. A shorter service standard could be justified by the observation that the clock for brain damage begins at the onset of the medical incident (the stroke or heart attack that denies the brain

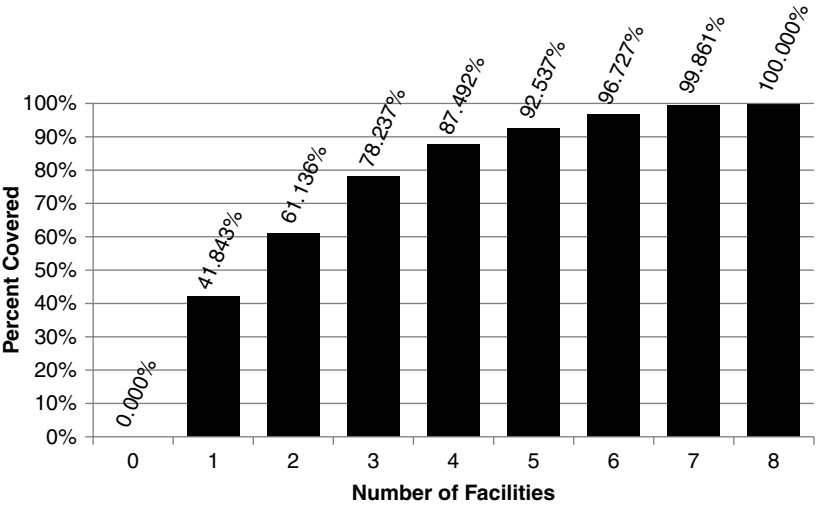

Figure 1.1 Typical tradeoff in maximum covering model.[1]

oxygen), while the clock for the response time begins only once the vehicle begins to roll out of its base. There is often a long time (several minutes) between when a medical incident arises and when a vehicle begins traveling to the scene. This additional time is consumed by the time required to recognize the need for an ambulance, the time to notify the dispatcher of the need, and the time required by the dispatcher to assign the call to a vehicle and to notify the vehicle's crew of the call. On the other hand, a longer service standard may be dictated by budgetary considerations. Requiring all demands to be served within 4 min may be too costly. It would be cheaper to have all demands served within 5 min or some longer time period. This suggests yet another model and another objective function: minimize the maximum response time (the time between a demand site and the nearest ambulance) using a given number (P) of vehicles. Such a model is referred to as the *P-center problem.*

Covering and center problems focus on the worst-case behavior of the system, for example, the maximum response time. In practice, there is often a tradeoff between minimizing the maximum response time and minimizing the *average* response time. This suggests yet a fourth model and objective that might be used in locating ambulances:

[1] The tradeoff curve shown is for the CITY1990.GRT data set with a coverage distance of 410 miles. The results shown are optimal values.

minimize the average response time (the time between a demand site and the nearest ambulance) using a given number (P) of vehicles. This model is called the *P-median problem* (Hakimi, 1964, 1965).

While we have outlined a number of different objectives that might be used in locating ambulances, it is important that we realize explicitly some of the factors that have been ignored in this discussion. The importance of doing so was enunciated particularly well by Jacobsen (1990, p. 205) who pointed out that formulating a problem incorrectly (e.g., failing to account for important problem factors) is likely to be far more important than whether or not you obtain an optimal or suboptimal solution to a particular problem formulation. Thus, while the focus of this book is generally on finding optimal solutions (or near optimal solutions) to specific mathematical statements of facility location problems, we must always ask whether the model being solved adequately represents the real problem being analyzed.

In the context of ambulance location, at least three facets of the real-world problem have been ignored in the discussion above. First, the models outlined above ignore the stochastic (or random) nature of demands and the fact that the nearest vehicle might not be available when called upon to serve a demand. A variety of approaches have been adopted to address this problem including: extending the deterministic models outlined above (Aly and White, 1978; Weaver and Church, 1983b, 1984; Daskin, 1982, 1983); incorporating queuing theory into location models (Larson, 1974; Fitzsimmons, 1973); and simulation approaches (Swoveland et al., 1973). Once the inputs to the model are recognized as being random variables, the outputs are likely to be random variables as well. Thus, we might no longer be interested only in the *average* response time (as in the *P*-median model) but also in the distribution of response times. Also, just as the demands are stochastic, so are the travel times. Models with stochastic travel times have also been developed (Weaver and Church, 1983a; Mirchandani and Odoni, 1979; Daskin and Haghani, 1984; Daskin, 1987).

Second, there is a need to balance the workload of the different vehicles. This stems from the needs (1) to preserve morale among the emergency medical service employees and (2) to maintain the skill level of all paramedics at some minimal level by ensuring that they are all exposed to a minimum number of medical emergencies of differing types.

As in all situations, we must ask whether facility location is really the correct problem. The quality of medical care delivered to the public and the likelihood of people surviving major medical incidents (e.g., auto crashes, assaults with deadly weapons, heart attacks, and strokes)

depend on many factors in addition to the location of ambulances in the community. The installation of a 911 emergency phone line can reduce the time needed to contact an ambulance dispatcher. This reduction in time may be greater (as compared with its cost) than that achievable by any relocation of ambulances. Improving the quality of hospital emergency room care may also go a long way toward reducing fatalities. It may be more cost effective to spend public funds on such improvements than it would be on relocating vehicles or adding ambulances. Instituting a community-wide CPR (cardiopulmonary resuscitation) education program might also be a cost-effective way of saving lives.[2]

The analysis assumed that all calls are equally important. In fact, this is not the case. Calls are often differentiated into critical (life-threatening) and noncritical calls. Also, some patients can be cared for at the scene of the incident, while others require transport to the hospital. In addition, the models outlined above fail to recognize the temporal variation in the overall intensity of calls (typically Friday night is the busiest time of the week) and the temporal variation in the spatial distribution of calls (more incidents will be reported from business districts during working hours than during the early morning hours). This temporal variation in demand suggests that having fixed sites may not be optimal; using relocatable ambulances may be preferable (Carson and Batta, 1990).

Once we distinguish between the severity of different demands for ambulance services, we recognize it may be advantageous to institute a multitiered system in which paramedics with differing levels of training are deployed (along with vehicles with correspondingly different levels of equipment). It may not be cost effective to have all paramedics trained at the highest level and to have all vehicles capable of responding to all types of medical emergencies. We may be able to deploy more vehicles and paramedics by (i) disaggregating calls based on their severity and (ii) allowing response units and personnel to be specialized for certain types of calls. In a multitiered system, dispatching rules become even more complicated. Not only might it not be advantageous to dispatch the nearest available vehicle (since doing so might leave large portions of the service area uncovered), but we must now decide which type of vehicle and crew to dispatch to each event.

[2] The notion of asking whether location solutions are the best way to attack a problem extends well beyond emergency services. For example, in considering problems of energy management, one solution might be to install additional power generation facilities. Another solution might be to manage the peak demand for power better.

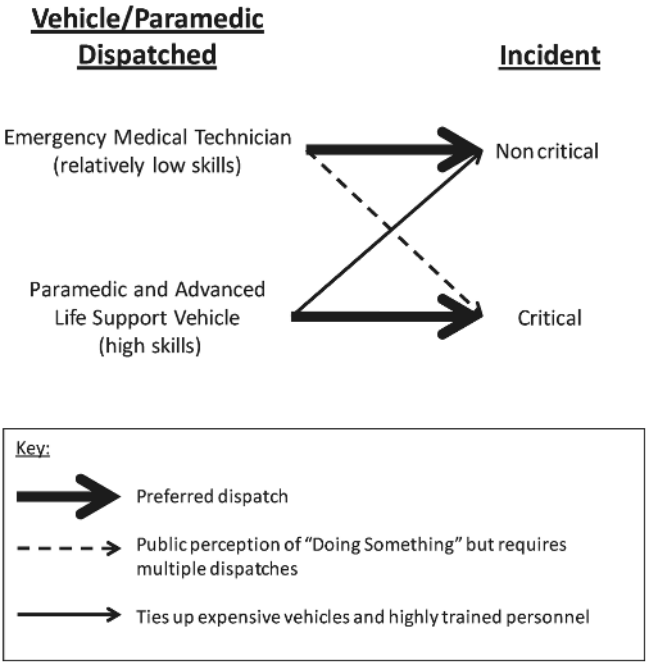

Figure 1.2 Dispatching options in a multitiered system.

The possibilities are shown schematically in Figure 1.2. If possible, we should dispatch an EMT (emergency medical technician) to a noncritical call and a paramedic in an advanced life support (ALS) vehicle to a critical incident. However, we might also want to dispatch an EMT to a critical call if the vehicle is likely to get to the scene before the ALS vehicle. Doing so would give the public the impression that something is being done about the emergency. This policy, however, ties up extra resources since two vehicles would then be dispatched under these conditions. Similarly, if an ALS vehicle is very near a noncritical incident, we might elect to dispatch the ALS vehicle. This policy has the advantage of getting medical assistance to the scene quickly, but has the disadvantage of tying up an expensive ALS vehicle and highly trained crew that might be needed elsewhere while they are busy serving the noncritical incident.

Finally, we note that the models we have briefly outlined, in particular the set covering model, the maximum covering model, and the *P*-center model, have all been used extensively in a broad range of public sector facility location problems including the location of libraries, schools, clinics, hospitals, and bus stops.

1.3.2 Siting Landfills for Hazardous Wastes

We turn briefly to another problem, that of locating landfills for disposal of hazardous wastes. First, any such site must be deemed geologically stable and suitable. Given this condition, the choice between sites might be dictated by a number of objectives. First, we would like to be as close as possible to the waste generation sites to minimize the transport costs as well as the exposure of the public to the hazardous wastes while they are en route to the disposal site. Minimizing the average (or total) shipping distance over some period of time results in a *P*-median formulation. However, many of the waste generation sites may be close to heavily populated regions. In this case, we would like the disposal sites to be far from populated areas. This suggests the use of a *maxisum or maxian model* in which we attempt *to locate a given number (P) of facilities to maximize the (population weighted) distance between population centers and the nearest sites* (Church and Garfinkel, 1978; Minieka, 1983). Clearly, this objective and the *P*-median objective conflict. The presence of conflicting objectives is common in facility location problems.

Both models take the number of facilities as given. In practice, we need to balance the initial capital investment costs against the ongoing operating costs. Thus, we might like to minimize the sum of the fixed site preparation costs and the discounted present cost of the stream of operating costs (e.g., on-site operating costs and transport costs). This leads to what is known as a *fixed charge facility location problem.*

Finally, we would like to reduce the inequities across communities. No community wants to be the dumping site for the rest of the state or the rest of the country. Thus, we might like to spread the risk or disbenefit around to the extent that it is possible to do so (e.g., Ratick and White, 1988; Erkut and Neuman, 1992; Wyman and Kuby, 1994). This has recently become an issue not only in the location of disposal sites but also in the routing of materials from generation sites to disposal facilities (Lindner-Dutton, Batta, and Karwan, 1991; ReVelle, Cohon, and Shobrys, 1991; List and Mirchandani, 1991; List et al., 1991).

1.3.3 Summary

In summary, modeling location problems requires an understanding of the real-world operations that are to be reflected in the model. Models need not reflect *every* aspect of the real-world operations. In fact, parsimonious models are generally better than complex inscrutable models.

The ability to know what must be incorporated into a model and what can safely be treated as exogenous is both an art and a science. As illustrated above, location problems often involve multiple conflicting objectives. The purpose of modeling is to identify the tradeoffs between the objectives while capturing as much of the richness of the real-world problem as is necessary to ensure the credibility of the modeler and model itself. Finally, we must always ask whether improving facility locations is the most cost-effective way of improving the system under study.

1.4 KEY DIMENSIONS OF LOCATION PROBLEMS AND MODELS[3]

Location problems and models may be classified in a number of ways. The classification may be based on the topography that is used (e.g., planar problems versus discrete location problems, problems on trees versus those on general graphs, and problems using different distance metrics) or the number of facilities to be located. Problems may also be classified based on the nature of the inputs (e.g., whether they are static or dynamic, known with certainty or only known in a probabilistic sense). Models may further be classified based on whether single or multiple products or demands must be accommodated by the facilities being located, whether there is one objective or multiple objectives, whether the beneficiaries and investors are the same or different actors, whether the facilities are of unlimited capacity or are capacitated, as well as a variety of other classification criteria. This section identifies key dimensions or characteristics of facility location problems and models.

1.4.1 Planar Versus Network Versus Discrete Location Models

One of the key differences between location models is in the way in which demands and candidate facility locations are represented. In *planar location models*, demands occur anywhere on a plane. We often represent demands using a spatially distributed probability distribution [which gives the likelihood of demands arising at any given (X, Y)

[3] Similar taxonomies have been developed by Brandeau and Chiu (1989) and Krarup and Pruzan (1990).

coordinate]. In such problems, facilities may be located anywhere on the plane. This modeling approach is to be contrasted with *network location models*, in which demands and travel between demand sites and facilities are assumed to occur only on a network or graph composed of nodes and links. Often, we assume that demands occur only at the nodes of the network, though some network location models have permitted demands to be generated anywhere on the links of the network. In network location models, facilities can be located only on the nodes or links of the network. One of the key questions we will be interested in considering is: when is location only on the nodes of the network optimal? The presence of an underlying network often facilitates the development of solution algorithms. *Discrete location models* allow for the use of arbitrary distances between nodes. As such, the structure of the underlying network is lost. However, by removing the restriction that the distances between nodes are obtained from an underlying network, the more general class of discrete location models allows a broader range of problems to be modeled. Discrete location problems are generally formulated as mixed integer programming problems as discussed below. For a further discussion of the differences between these three types of models, the reader is referred to Chhajed, Francis, and Lowe (1993).

The focus of this book is on network and discrete location models. Handler and Mirchandani (1979) and Mirchandani and Francis (1990) provide excellent overviews of network location models, while planar location models are discussed in Hurter and Martinich (1989) and Love, Morris, and Wesolowsky (1988).

1.4.2 Tree Problems Versus General Graph Problems

Within the class of network location models, we often distinguish between problems that arise on *trees* and those that must be formulated on a more general (fully connected) *graph*. Figure 1.3 illustrates a number of different trees and general networks.

A tree is a network in which there is at most one path from any node to any other node. In other words, a tree is an acyclic graph or a graph with no cycles. In general, we will focus our attention on spanning trees (trees in which there is exactly one path between any node and any other node). If such a tree has N nodes, it will have $N - 1$ links.

Our interest in trees as opposed to more general graphs results from two considerations. First, many real-life problems can be

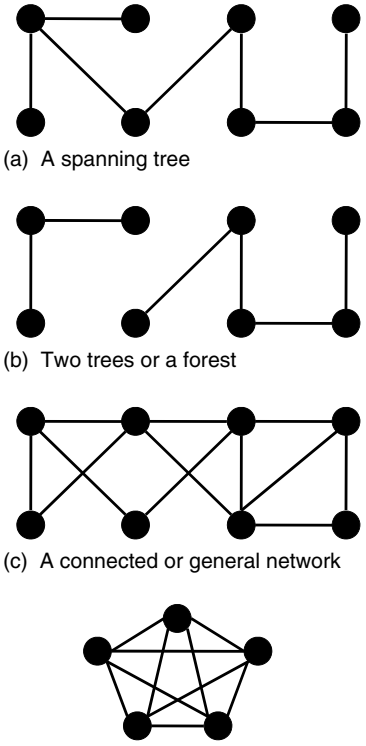

Figure 1.3 Example trees and graphs.

represented quite well as trees. For example, the links depicting major highways within a region often form a tree as long as we ignore the cycles formed by beltways surrounding major urban areas. Also, major parts of power transmission and telecommunication networks—particularly the portions used for the delivery of local services—are essentially trees. Second, given a verbal or mathematical statement of a location problem, it is often the case that we can solve the problem easily on a tree while solving it on a more general network is exceptionally difficult. In Chapter 3, we formalize the notions of easy and difficult problems using complexity theory.

1.4.3 Distance Metrics

Location models are also often characterized by the distance metric (the method of measuring distances) that is used. For network location models, we will generally use the shortest distance between any pair of

points using links in the network. In Chapter 2, we discuss algorithms for finding the shortest paths between points in a network. In planar location problems, one of three distance metrics is typically employed:

(a) Manhattan or right-angle distance metric

$$d\left[(x_i, y_i); (x_j, y_j)\right] = |x_i - x_j| + |y_i - y_j|$$

(b) Euclidean or straight-line distance metric

$$d\left[(x_i, y_i); (x_j, y_j)\right] = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$$

(c) ℓ_p distance metric

$$d\left[(x_i, y_i); (x_j, y_j)\right] = \left\{(|x_i - x_j|)^p + (|y_i - y_j|)^p\right\}^{1/p}$$

where $d\left[(x_i, y_i); (x_j, y_j)\right]$ is the distance between the *i*th and *j*th points and (x_i, y_i) gives the coordinates of the *i*th point. A bit of thought will show that the ℓ_2 metric (the ℓ_p metric when $p = 2$) is the same as the Euclidean distance between two points and that the ℓ_1 metric is equivalent to the Manhattan or right-angle distance. What is the ℓ_∞ metric? This is left as an exercise for the reader.

1.4.4 Number of Facilities to Locate

Another way of characterizing facility location problems is by the number of facilities to be located. In some problems (e.g., the *P*-median, *P*-center, and maximum covering problems), the number of facilities to locate is *exogenously* specified. In other cases (e.g., the set covering problem and the fixed charge facility location problem), the number of facilities is *endogenous* to the problem and is a model output. For those problem statements in which the number of facilities to locate is exogenously specified, we also distinguish between *single-facility* location problems and those in which *multiple facilities* are to be sited. Often, single-facility location problems are dramatically easier than are their multifacility counterparts.

1.4.5 Static Versus Dynamic Location Problems

Most of the location models that we will consider will be *static* problems. In static models, the inputs do not depend on time; typically, we will use a single "representative" set of inputs and solve the problem for a single "representative" period.

As noted above in the discussion of ambulance systems, inputs are rarely static. Thus, while most location *models* are static, most location *problems* are *dynamic* in that the inputs (and consequently the outputs as well) depend on time. Inputs that may depend on time include demands, costs, and available and preexisting candidate facility locations. In dynamic problems, the models must explicitly include multiple periods of time. Different periods might allow us (i) to capture hourly differences in the mean number of demands for service, (ii) to reflect differences between the spatial patterns of demands on weekdays and weekends, or (iii) to account for increases in demands or costs over a period of years.

In dynamic problems, we are concerned not only with the question of *where* to locate facilities but also with the question of *when* to invest in new facilities or to close existing facilities. In some models of dynamic location problems, once a facility is opened it is assumed to be available for all future periods. In other models, facilities may be opened, closed, or moved throughout the planning horizon (Ballou, 1968; Sweeney and Tatham, 1976; Van Roy and Erlenkotter, 1982; Wesolowsky, 1973; Wesolowsky and Truscott, 1975).

While most researchers and planners have a good idea of what is meant by a static location model, there is considerably less agreement about what is meant by a dynamic location model. One approach might be to identify a single set of locations that perform well with respect to a number of spatially different demand patterns that occur at different times. Such a problem statement might arise in locating fire stations that need to respond well to demands during working hours as well as on weekends. This approach might also be appropriate in locating facilities to serve demands that vary in a cyclic manner (e.g., Osleeb and Ratick, 1990). A second approach to the dynamic location problem would be that of identifying the optimal evolution of facility locations over time. Such a model would be appropriate for a firm that needs to locate warehouses to supply its customers and that plans to expand from a set of regional retail outlets to a national chain. In some cases, it is best to find an optimal first period decision as opposed to a plan for all future time periods (Daskin, Hopp, and Medina, 1992). Finally, an alternate definition of the dynamic location problem would be that of positioning vehicles in real time to respond to minute-by-minute

changes in the fleet of available (nonbusy) vehicles. This problem in particular has been analyzed by Kolesar and Walker (1974) using set covering models.[4]

1.4.6 Deterministic Versus Probabilistic Models

Just as the inputs to models may be either static or dynamic, so too the inputs may be deterministic (certain) or probabilistic (subject to uncertainty). In dealing with location problems over time, many of the inputs are likely to be uncertain. For example, future calls for ambulance services are not known with certainty. Instead they must be predicted and, as such, are subject to uncertainty. This book focuses on deterministic models, though in some cases we can readily generalize the algorithms or model formulations to include some probabilistic components. Louveaux (1993) reviews stochastic location models.

1.4.7 Single- Versus Multiple-Product Models

The models outlined above have all implicitly assumed that we are dealing with a single homogenous product or service and that all demands are identical. Most location models make this assumption. However, in practice, it is often important to distinguish between different products or services all of which will be served by the same set of facilities. For example, it may be important to distinguish between critical and noncritical calls.

In some cases, products are distinguished by having different origins and destinations. For example, a single set of transshipment facilities may be used by an automobile manufacturer in shipping finished vehicles from assembly plants to dealers. At such transshipment points, vehicles are offloaded from railcars and loaded onto trucks for final delivery to customers (typically, dealers). Each assembly plant/customer combination would represent a different product. In other words, we would need to distinguish between Cadillac Sevilles going from a Cadillac assembly plant to a Cadillac dealer in San Diego and Chevrolet Corvettes going from a Chevrolet plant to a dealer in Los Angeles, even though both vehicles might use the same transshipment point in southern California.

[4] Ratick et al. (1987) review dynamic location models. They distinguish between models in which facilities remain in the siting plan once they are opened and models that allow facilities to be opened and closed throughout the planning horizon.

1.4.8 Private Versus Public Sector Problems

In private sector problems, the investment costs and benefits are typically measured in monetary units. Furthermore, the costs and benefits are generally incident on the same actors: the firm, its management, and its investors all of whom share common objectives and goals. All this makes cost/benefit analysis relatively easy.

In public sector location problems, many nonmonetary cost and benefits must also be considered. For example, in locating hazardous waste repositories, there are a number of environmental costs that may be difficult to translate into monetary units. In locating emergency services, the dollar value of the lives saved as a result of shorter travel times may be exceedingly difficult to assess. In siting public schools, the benefits may be measured in terms of the number of students who graduate from high school. In public sector problems, not only are costs and benefits often incommensurable, but there are often multiple benefit measures (as discussed in Section 1.4.9). In addition, while the costs of public sector projects may be borne by the public at large, the benefits are often concentrated on fewer people. Thus, investment in public schools directly benefits school-aged children and their parents. Such investments do not directly benefit other members of society, such as the elderly. Finally, public sector investments are often complicated by the political process in which beneficiaries of one investment may agree to support projects from which they do not directly benefit. Thus, groups representing the elderly may agree to support additional funding of public schools provided other groups' support enhanced health care legislation.[5]

1.4.9 Single- Versus Multiple-Objective Problems and Models

Most models capture a single objective; however, most problems are inherently multiobjective in nature. Since one of the purposes of location modeling is to help identify tradeoffs, single-objective models must often be run with a range of input parameters (e.g., running a P-median problem with a number of values of P to trace the tradeoff between average distance to a facility and the number of facilities sited). Alternatively, multiple models need to be employed (e.g., Eaton and Daskin, 1980).

[5] ReVelle, Marks, and Liebman (1970) were among the first to distinguish between public and private sector location problems. Ghosh and Harche (1993) provide a recent review of location models used in private sector decision making.

1.4.10 Elastic Versus Inelastic Demand

Most models treat demand as given and independent of the level of service. In fact, demand in almost all cases depends on the level of service provided. This, in turn, depends on the facility locations and the types and sizes of facilities used. In some cases, demand is likely to be relatively inelastic (independent of the level of service). For example, if someone needs an ambulance, he or she is unlikely to inquire about the cost. An individual is also unlikely to bother to find out the expected arrival time of the vehicle and to identify alternative means of getting to the hospital if the expected response time is too long. On the other hand, consumers' choices of where to shop depend critically on the amenities within the shopping center, the location of the center, and the number and variety of stores in the shopping center. Despite the fact that demand in most real-world location problems exhibits some degree of elasticity with respect to service (which depends in part on the location decisions), we will generally treat demand as inelastic. Recent work by Perl and Ho (1990) has examined some of the implications of elastic demand on public facility location models. Kuby (1989) formulates a model that maximizes the number of firms that can coexist in a market. His model also incorporates elastic demand.

1.4.11 Capacitated Versus Uncapacitated Facilities

Many facility location models (e.g., standard set covering, maximum covering, *P*-median, and *P*-center models) treat facilities as having unlimited capacity. Other models impose explicit capacity limits on facilities. In still other cases, the size of a facility is a model output.

1.4.12 Nearest Facility Versus General Demand Allocation Models

As discussed above, the allocation of demand to facilities is a critical issue in location modeling. Often, demands are assigned to the nearest facility provided that facility has the capacity to serve the demand. In capacitated problems, this may result in the need to split the demand at a site between several facilities. If this is not permissible in a particular problem setting, explicit constraints must be included in the model (typically, in the form of integer variables) to force all of the demand at a particular location to be assigned to a single facility. This may

result in some demands being assigned to facilities other than the closest site. In still other cases, models must recognize that a fraction of the demand at a site will be served by the nearest facility, and the remainder of the demand will be served by more remote facilities when the nearest facility is busy.

1.4.13 Hierarchical Versus Single-Level Models

In many systems, a hierarchy of facilities exists with flows between the facilities that are being located. For example, in a national health care system, rural health centers are likely to refer patients to clinics, which, in turn, may refer patients to community hospitals. In some such systems, services provided at the lower level (e.g., the rural health center) are offered at higher levels; in other cases, these services are not replicated. Narula (1986) refers to these as *successively inclusive* and *successively exclusive facility hierarchies*, respectively. Also, in some systems, patients may elect to go to the facility of their choice; in others, they must begin service at the lowest level facility in the hierarchy and be referred up from there. In such hierarchical location problems, the locations of the different facilities interact significantly through the flows between the facilities. Facility interactions also arise in many facility layout problems (Francis, McGinnis, and White, 1992).

1.4.14 Desirable Versus Undesirable Facilities

In most location problems, we are interested in locating desirable facilities. In other words, value increases, in some sense, the closer the facilities are to the people or goods being served. Ambulances, fire stations, schools, hospitals, post offices, warehouses, and production plants are all considered desirable facilities in this sense.[6] Some facilities, however, are considered undesirable in the sense that most people want them located as far away as possible. Typically, such facilities are either noxious (posing a health or welfare hazard to people) or obnoxious (posing a threat to people's lifestyles) facilities (Erkut and

[6] While these facilities are considered desirable in a general sense, it is clear that many people might not want to buy a house immediately adjacent to a fire station, for example, since the disruption associated with the fire engines responding to calls for service may outweigh the benefit of being near the station. Nevertheless, it is generally better to be near a fire station than to be far from a station. Similar issues might arise in the location of other generally desirable facilities.

Neuman, 1989). Hazardous waste sites, landfills, incinerators, missile silos, and prisons generally fall into this category. In the location of undesirable facilities, it is often useful to distinguish between cases in which we are only concerned with the distance between facilities, as might be the case in locating nuclear missile silos, and those in which we are concerned with the distance between the facilities being located and population centers, as might be the case in locating landfills. In almost all practical location contexts involving the location of undesirable facilities of any kind, multiple conflicting objectives are likely to come into play. Thus, while we would like landfills to be located far from population centers, we also want to minimize the costs of transporting material from the waste generation sites to the landfill, as discussed above in Section 1.3.2. Unfortunately, much of the waste that is deposited in landfills is generated in highly populated areas. Thus, in locating landfills, the tradeoff between minimizing transportation costs and minimizing the number of people affected by the landfills needs to be identified.

1.5 A TAXONOMY OF LOCATION MODELS

In this section, we outline a taxonomy of location models, which is based largely on the modeling assumptions about the spatial configuration of the demands being served and facilities providing the service.

1.5.1 Typology of Location Models

Figure 1.4 illustrates this taxonomy (and is adapted from Daskin (2010)).

Analytic location models assume that demands are distributed in some manner over space. For example, we might typically assume that demands are uniformly distributed over a square of diamond-shaped region. By uniformly distributed, we mean that the density of demand is constant over the shape of the region. One way to think about this is that the demand region is a piece of bread and the demand density is the thickness of peanut butter that is spread (perfectly with exactly even thickness) over the piece of bread. Candidate facilities can be located anywhere in the service region. Clearly, these sorts of models make very strong assumptions. The population in the United States, for example, is anything but uniformly distributed across the country. Within the contiguous United States, New Jersey has a density of

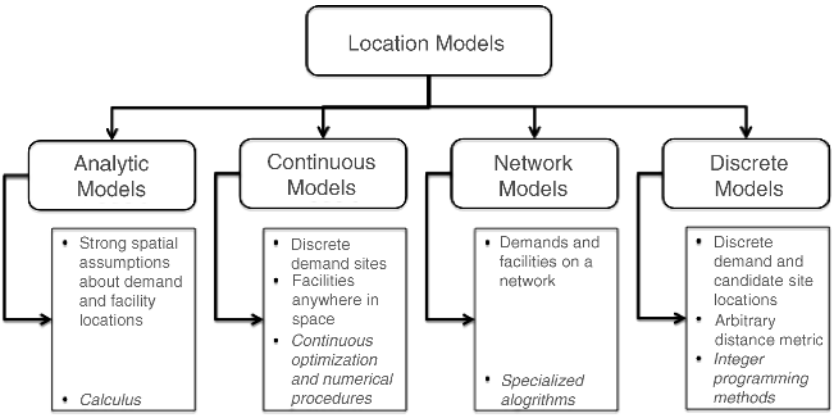

Figure 1.4 Alternative taxonomy of location models.

approximately 1200 people per square mile, while Wyoming has a density of about 6 people per square mile, at the other extreme. Nevertheless, these methods can provide some insights into the structure of solutions to location models. Section 1.5.2 illustrates this sort of modeling.

Despite its name, **continuous location models** assume that the demands occur at discrete sites. The demand level at these sites is known *a priori*. The candidate facilities can be located anywhere in a region. The Weber location model is typical of models in this class. The Weber model finds the center of gravity of the demand points. The easiest way to think about the Weber problem is to imagine the demands as being weights that are suspended below a plywood board by frictionless pulleys at each of the demand points. The weight is proportional to the demand at that point. The strings on which the demands are suspended are tied to a tiny ring. The point at which the ring comes to rest is the location of the Weber point or the solution to the Weber problem. This problem is typically solved using the Weiszfeld procedure as described by Drezner et al. (2001).

Network location models treat demands and facilities as being located on a network composed of nodes and links. The U.S. Interstate Highway system is a typical network that might be utilized for this sort of analysis. All demands and all facilities must be located on the network. Typically, demands are located at the nodes of the network. Facilities can be located either on the nodes or on the links of the network. One question that is often asked is whether at least one optimal solution to the problem at hand consists of locating facilities only on the nodes of the network. As we will see, for some objectives and

problems, the answer to this question is *yes*, while for other problems, the answer is *no*. Research in this field is often aimed at finding very efficient algorithms to solve special instances of network location problems (see Chapter 3 for a formal definition of efficiency). The algorithm of Goldman (1971) discussed in Chapter 6 is illustrative of this sort of modeling.

Finally, **discrete location models** make no particular assumptions about the demand and facility locations. We are simply given the locations or coordinates of the demand nodes and the candidate locations. The "distances" between the demand and candidate locations do not need to adhere to any particular formula. For example, the airline fares between different cities in the United States do not seem to be related very well to the distances between the cities. These models are often formulated as integer programming models and solved using exact or heuristic (approximate) methods. Many of the models discussed in this book fall into this category of location models.

1.5.2 A Simple Analytic Model

While most of this book is devoted to network and discrete location models, this section presents a simple analytic location model. The problem that we address in this section is the analytic analog of the fixed charge location problem outlined in Chapter 7. The service network we consider is a diamond with the travel directions at 45° to the sides of the diamond as shown in Figure 1.5. Demands are assumed to be uniformly distributed over the region with a demand density of ρ demands per unit area. If the service region has an area a, then we can show that (1) it is optimal to locate a single facility to serve the region at the center of the service area and that the average distance between the facility and a randomly selected demand is given by $(2/3)\sqrt{a/2}$. If we were to divide the service area into N equally sized diamond-shaped subregions and we were to locate a facility at the center of each

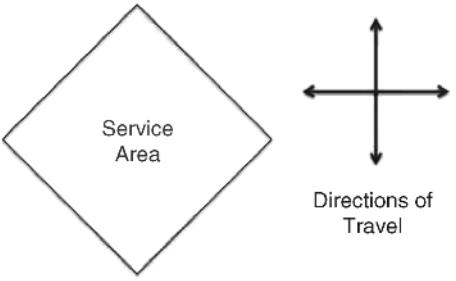

Figure 1.5 Service area and directions of travel for a simple analytic location model.

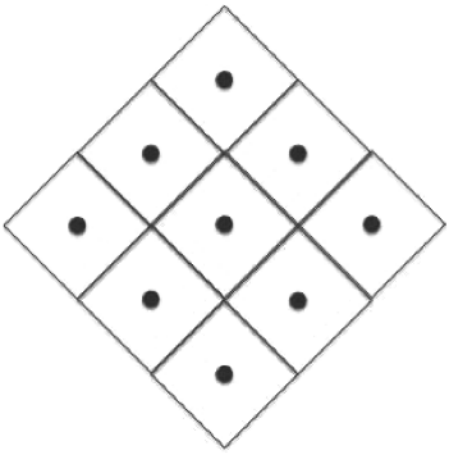

Figure 1.6 Example service region subdivided into nine subregions.

subregion, the average distance between a randomly selected demand and the nearest facility would be $(2/3)\sqrt{a/2N}$. Figure 1.6 illustrates a service region divided into $N = 9$ subregions.

Associated with each facility that we locate is a fixed cost of f. There is also a unit transportation cost c per demand per mile. The key problem that we face is determining the optimal number of facilities to locate so that we minimize the sum of the facility and transport costs. As we increase the number of facilities, the total fixed facility costs will increase (linearly), but the transport costs will decrease (with the square root of the number of facilities located. This tradeoff is shown in Figure 1.7 for a region of 100 square miles, a unit transport cost of 1, a demand density of 25 demands per square mile per unit time, and a fixed facility cost of 225 per facility per unit time. In this case, the optimum number of facilities to locate is nine resulting in a total cost of $5953.

In general, the total cost as a function of the number of facilities that we locate is given by

$$\mathrm{TC}(N) = fN + c\rho a\left(\frac{2}{3}\sqrt{\frac{a}{2N}}\right) \tag{1.1}$$

The first term of (1.1) represents the fixed facility costs, while the second represents the transport costs. The term in parentheses is the average distance, ρa represents the total number of demands, and c converts the rest of the term to monetary units. Ignoring the need for the number of facilities to be an integer (and ideally a squared number so that we can evenly divide the service region into equally sized diamonds), we can find the optimal number of facilities to locate by

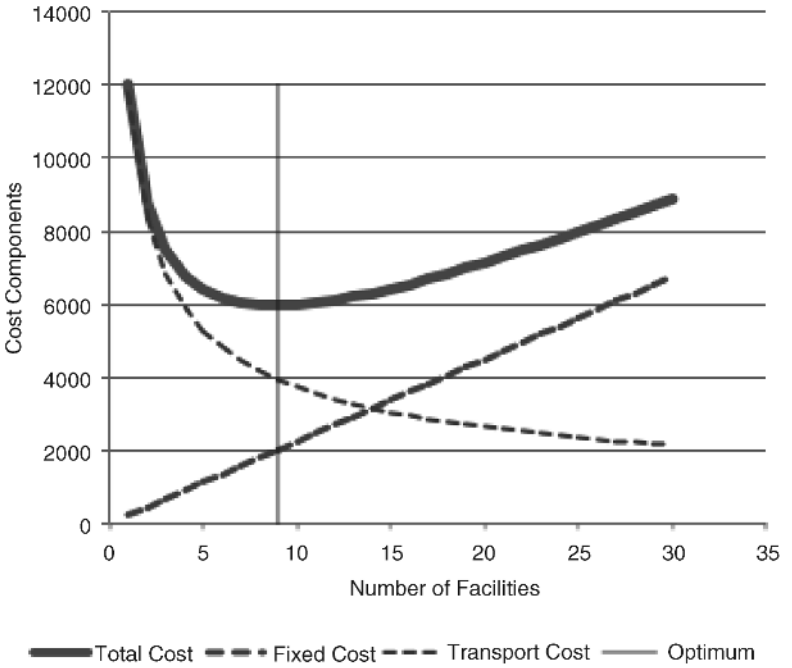

Figure 1.7 Typical cost components in a simple analytic model.

taking the derivative of (1.1) with respect to N and equating the derivative to 0. We do this in (1.2) below.

$$\frac{\mathrm{dTC}(N)}{\mathrm{d}N} = f - c\rho a \frac{1}{3}\sqrt{\frac{a}{2}}N^{-1.5} = 0 \tag{1.2}$$

Solving for N, we obtain

$$N^* = a \cdot \left(\frac{c\rho}{3\sqrt{2}f}\right)^{2/3} \tag{1.3}$$

If we substitute the optimal number of facilities given by (1.3) into the total cost function (1.1), we obtain the optimal total cost as shown in (1.4):

$$\begin{aligned} \mathrm{TC}(N^*) &= fN^* + c\rho a\left(\frac{2}{3}\sqrt{\frac{a}{2N^*}}\right) \\ &= af^{1/3}(c\rho)^{2/3}\left\{\left(\frac{1}{3\sqrt{2}}\right)^{2/3} + 2\left(\frac{1}{3\sqrt{2}}\right)^{2/3}\right\} \\ &= 1.145af^{1/3}(c\rho)^{2/3} \end{aligned} \tag{1.4}$$

The second line of (1.4) breaks the total cost into two terms: the first is derived from the fixed facility costs and the second from the transport costs. At the optimal number of facilities, the transport costs are twice the fixed facility costs. This is also shown graphically in Figure 1.7.

In deriving the optimal number of facilities, we have clearly made a large number of restrictive assumptions. Fortunately, the total cost is not very sensitive to small changes in the number of facilities. In fact, if we actually locate $N = \alpha N^*$ facilities—in other words, the number of facilities is α times the optimal number—the ratio of the cost using the suboptimal number of facilities to the optimal number can be shown to be

$$\frac{\mathrm{TC}(N)}{\mathrm{TC}(N^*)} = \frac{\mathrm{TC}(\alpha N^*)}{\mathrm{TC}(N^*)} = \frac{\alpha + 2/\sqrt{\alpha}}{3} \tag{1.5}$$

Figure 1.8 plots this ratio against the value of α. The insensitivity of the total cost to variations in the number of facilities used is clear. Table 1.1 presents this information in a slightly different way. For a desired percentage difference or error between the actual and optimal cost, the table gives the allowable range in α, the ratio of the actual to optimal number of facilities. For example, as long as the number of facilities is within 75–131% of the optimal number, the cost will be within 2% of the optimal total cost.

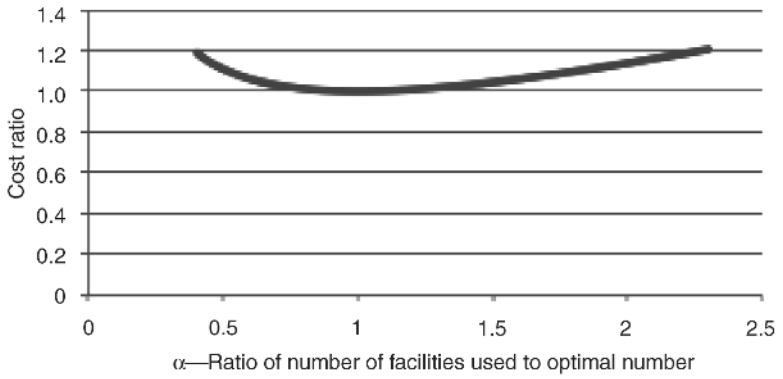

Figure 1.8 Ratio of actual to optimal cost versus ratio of actual to optimal number of facilities for the simple analytic model.

Table 1.1 Allowable Range of Alpha for Various Percentage Errors in the Optimal Cost for the Simple Analytic Model

%Error	Min Alpha	Max Alpha
0	1.000	1.000
1	0.808	1.212
2	0.751	1.316
5	0.631	1.530
10	0.516	1.811
25	0.345	2.476

1.6 SUMMARY

In this chapter, we have identified the key questions answered by facility location models. We have qualitatively introduced a number of classical facility location models through example problems. Finally, we have outlined a taxonomy of location models and problems. In the course of this discussion, we identified those areas that will be the primary focus of the remainder of this text. In particular, the text will focus on network and discrete location problems, ignoring planar, or continuous location problems and models.

Most network and discrete location problems of interest to us can be formulated as *linear programming problems* in which some of the variables are constrained to take on only integer values. Such problems are called *integer linear programming problems*. An understanding of linear programming is essential to the formulation and solution of many facility location problems. In addition, certain pure linear programming problems must be solved before most facility location problems can be attacked. For example, the problem of finding the shortest path from a facility to a demand node can be formulated as a linear programming problem. Often, shortest path distances are needed as inputs to facility location problems. Finally, once the facility locations are known, the problem of assigning demand nodes to facilities, particularly when the facilities have limited service capacities, can often be cast as another linear programming problem called the *transportation problem*. Chapter 2 reviews linear programming in general as well as a number of special linear programming problems that are intimately linked to facility location problems including the transportation problem and the shortest path problem.

EXERCISES

1.1 The ℓ_p distance metric was defined as follows:

$$\ell_p = d\left[(x_i, y_i), (x_j, y_j)\right] = \left\{\left(|x_i - x_j|\right)^p + \left(|y_i - y_j|\right)^p\right\}^{1/p}$$

If we let $\ell_\infty = \lim_{p\to\infty}\left\{\ell_p\right\}$, what is ℓ_∞ equal to?

Note: This distance metric is used in a number of industrial contexts. For example, it can be used to compute the time that it takes for an automated picker to move from one location to another in a warehouse when movements in both the X and Y directions can occur simultaneously, but the time to move between locations is governed by the larger of the two distances. Its three-dimensional extension has similar applications in robotics.

1.2 Use the real estate listings in your local newspaper to identify at least four or more houses in your city that are comparable in terms of the number of bedrooms and the number of bathrooms.

(a) What are the asking prices of the houses?

(b) What is the ratio of the largest asking price to the smallest asking price?

(c) What location factors might account for the differences in prices between the homes?

(d) What nonlocation factors might account for the price differences?

1.3 Identify at least two different objectives that public officials might have in locating new prisons.

1.4 With the ever-growing concerns about the environment, vehicle emission inspection policies are coming under increasing review.

(a) Discuss at least two different objectives that state officials would have in determining the locations of vehicle emission testing stations.

(b) Discuss nonlocational strategies that might be employed to increase public cooperation with emission testing laws.

(c) Discuss how the problem of locating vehicle emission testing stations fits into the location problem taxonomy outlined in Section 1.4.

1.5 The area of the contiguous United States is approximately 3.12 million square miles. The population (based on the 2010 census) of the continental United States was approximately 307 million people, resulting in a density of approximately 98.4 people per square mile. Assume that the cost of shipping an item 1 mile is \$0.01. For fixed costs ranging from \$20 to \$200 million (in increments of \$20 million) plot.

(a) The optimal number of facilities to use.

(b) The optimal total cost, using the model outlined in Section 1.5.2.

1.6 A national healthcare provider wants to establish clinics in a medium-sized city of 1,000,000 people. The area of the city is 1000 square miles. If the market penetration of the provider in the city is 20% of the population, the cost of a clinic is $250,000 per year, the cost per mile is $0.10, and each person is expected to make an average of four visits per year to the clinic(s),

(a) Find the optimal number of clinics for the provider to staff, using the model of Section 1.5.2.

(b) Compute the total cost of staffing this many clinics.

(c) Compute the average distance between a randomly selected patient and the nearest clinic.

(d) Note that the maximum distance (for the model of Section 1.5.2) is 1.5 times the average distance. Suppose the provider wants to ensure that no patient is more than 5 miles from the nearest clinic. How many clinics should the provider staff under these conditions? What is the new total cost of this configuration?

(e) Identify at least three problems associated with using the model of Section 1.5.2 in this context.