Foundations and Trends® in Technology,
Information and Operations Management

Integrated Modeling for Location Analysis

Suggested Citation: Ho-Yin Mak and Zuo-Jun Max Shen (2016), "Integrated Modeling for Location Analysis", Foundations and Trends® in Technology, Information and Operations Management: Vol. 9, No. 1-2, pp 1–152. DOI: 10.1561/0200000037.

Ho-Yin Mak
Saïd Business School, University of Oxford
UK
ho-yin.mak@sbs.ox.ac.uk

Zuo-Jun Max Shen
University of California at Berkeley
USA
maxshen@berkeley.edu

now
the essence of knowledge
Boston — Delft

Contents

Integrated Modeling for Location Analysis

Ho-Yin Mak[1] and Zuo-Jun Max Shen[2]

[1] *Saïd Business School, University of Oxford, UK*
ho-yin.mak@sbs.ox.ac.uk
[2] *University of California at Berkeley, USA*
maxshen@berkeley.edu

ABSTRACT

Delivery of products and services relies on well-managed operations. In designing large-scaled supply chain and service systems, locations of key facilities are a critical decision, as these facilities form the backbone of operations of these systems. For example, a key to effective supply chain management is the deployment of a structurally well-designed facility network, consisting of plants, warehouses, retail stores, etc. The aim of the study of facility location is to develop analytical methodologies to inform the planning decisions for evaluating and selecting siting plans for these facilities that ensure both convenient provision of (or access to) products and services by customers and users, as well as efficient operations (i.e., low operating costs).

Facility location and network design has long been an integral topic of study in operations management. In this literature, one may observe that earlier works mainly focused on a strategic view of accessibility and operational costs, using performance metrics based on strategic distances between the chosen facilities and customers or suppliers. This traditional approach often neglects the impacts of future tactical and operational activities to be conducted in the network, and optimizes objectives that do not fully reflect the long-term performance of the facility network. In attempt to rectify this shortcoming, researchers have proposed an integrated modeling approach that enhances the classical models

Ho-Yin Mak and Zuo-Jun Max Shen (2016), "Integrated Modeling for Location Analysis", Foundations and Trends® in Technology, Information and Operations Management: Vol. 9, No. 1-2, pp 1–152. DOI: 10.1561/0200000037.

by jointly considering strategic, tactical and operational activities in facility systems. By integrating tactical and operational characteristics of facility networks into strategic design decisions, the integrated approach offers a more balanced perspective on the strategic trade-offs in network design.

As shown in a series of recent research, this integrated modeling approach can potentially deliver new insights into facility location problems in a variety of contexts, e.g., supply chain network design, deployment of health care facilities, and design of storage systems for renewable power. In this monograph, we perform a review of some important concepts in this emerging stream of literature. Motivated by supply chain design applications, we first discuss the basic modeling concepts, including both mathematical programming-based and analytical approaches for modeling. While simulation-optimization approaches can be used for analyzing location problems, they are not covered in the scope of this monograph. We also review techniques adopted in the literature to analyze and solve these classes of location models. This is aimed to serve as a reference for readers (especially students) who like to develop their own models but are less familiar with this line of research. Furthermore, we review a number of applications of this line of research, covering both applications in supply chain contexts and other emerging domains, such as sustainable transportation, energy and health care.

1

Introduction

Facility location is one of the most crucial strategic planning decisions for governments, firms and non-profit organizations alike. A popular saying in real estate is that the three most important attributes of a property are its *location, location, location*. In marketing, *place* (i.e., location) is considered as one of the four building blocks ("four P's") of a marketing strategy, along with price, product and promotion. For firms selling tangible products and services alike, strategic location planning is often the basis of firms' competitive advantages. For retail stores and service facilities, good location planning allows customers to access the firm's offering with low access or inconvenience costs (e.g., in the form of travel cost or time), and thereby enhances customers' willingness to pay and the firm's revenue. For back-end support facilities such as distribution centers and warehouses, a carefully located network of facilities serves as the backbone for efficient logistics operations. In the public sector, choice of locations of public service facilities (or equipment, as mobile facilities), such as hospitals, fire stations and ambulances, plays a critical role in determining the level of service provided to the public, such as response times to emergency calls.

While facility location can deliver positive strategic value to firms (or organizations), it also poses significant planning risk in a large variety of applications due to the often hefty resource commitment involved. For example, in selecting a distribution center location, the firm needs to acquire (purchase or engage in long-term lease) a piece of land, construct (or acquire) a building, and equip the building with necessary labor and equipment. As a result of this heavy commitment, facilities are very costly to be relocated or closed after they start operating. This high cost of recourse calls for foresight in planning, particularly in forecasting the future operating environment (e.g., demand and costs) and in understanding the long-term operating characteristics of, as well as possible interactions between, the facilities.

Strategically, various considerations underpin the choice of facility locations. Proximity to markets and/or suppliers, operational efficiency of logistics operations (e.g., to replenish stock at the chosen retail store locations), availability of skilled labor or natural resources, access to free or low-tariff trade zones, presence of favorable tax or regulatory policies, political stability of the region, etc., are examples of important factors to evaluate when planning for a network of facilities. As Daskin (2011) (Chapter 1) suggests, these include quantifiable and non-quantifiable factors, and the focus of developing mathematical models is on the former.

Among those quantifiable factors of consideration, the trade-off between service level and cost pertains in the majority of location planning scenarios. Service level refers to the accessibility of facilities by their users, and is typically determined by factors such as response time, and the costs and inconvenience of access. Note that these factors are decreasing in the distances between users and facilities; that is, service level is typically improved as a denser facility network is deployed. Therefore, to maximize user accessibility, a ubiquitous location strategy, where users never need to travel long distances to access the nearest facilities, could be desirable. Examples of this strategy include those adopted by Seven-Eleven in certain densely populated (especially Asian) large cities, or Starbucks in major North American cities. While such strategies make facilities extremely accessible, the obvious downside is

the higher operational costs due to the lack of economies of scale in operating each facility.

On the other hand, operating costs of the facility network depend on a multitude of factors. In the majority of planning scenarios, the overall costs consist of fixed and variable components. It is particularly important to note that many facility types (e.g., factories, hospitals, transportation terminals) employ expensive equipment and thus the fixed component of costs is typically sizable. Therefore, the dominant factor in the strategic consideration of operating costs is often economies of scale. That is, operating costs can often be reduced by deploying a network with fewer (i.e., sparser) facilities each handling a larger volume of demand. An example is the "four corners" strategy commonly adopted by North American retailers that operate small numbers of distribution centers, typically near the major East and West Coast ports, to serve demand from the entire continent.

Naturally, the goals of improving service level (which calls for locating more facilities) and reducing operating costs (locating fewer facilities) are in conflict. The early literature focuses on developing optimization models that attempt to balance these goals in different planning contexts. We shall review some of the classical models in the next section.

1.1 Brief Review of Classical Location Theory

In this section, we briefly review some of the most common location models used in practice. Typically, the planner is faced with the problem of locating a number of facilities to serve a discrete set of spatially dispersed customers. Many classical facility location models are formulated to deliberately characterize the trade-off between access distance and costs. Access distance refers to a measure of the distance between customers and the facilities that they patronize, and reflects the design quality of service. Two popular measures of access distance employed in the literature are demand-weighted distance and coverage distance. We shall review these concepts and some of the associated optimization models below.

1.1.1 Demand-Weighted Distance Models

Demand-weighted distance is a popular metric for access distance considered in facility location models. Particularly, it considers the weighted average of distances between individual customer locations and their respective assigned (or patronized) facilities. Often, the weights are selected to be proportional to the volumes of demand (e.g., number of potential consumers, forecasted sales volumes, etc.) at the customer sites. The consideration of such weights allows the decision maker to prioritize service provision to customers in the sense that facilities tend to be located closer to more important customers with larger weights.

In the case where the costs of serving a customer location are bilinear in the location's demand volume and access distance, demand-weighted distance reflects the system-wide operations costs of serving all customers with the assigned facilities. One example is a supply chain setting in which facilities are distribution centers (DCs) and customers are retail stores. Demand-weighted distance, in this case, provides a proxy for the total transportation costs under direct shipments, such that the costs of shipping to one retailer location are approximately given by the shipment volume (demand) times the shipment distance. Below, we briefly review the classical location models that incorporate demand-weighted distance objective.

The P-median problem, originally formulated by Hakimi (1964, 1965) is concerned with minimizing the demand-weighted distance of serving a set of customers by locating a given number (P) of facilities. Note that in graph theory terminology, the *absolute median* of a network is a point from which the sum of weighted distances to all nodes of the network is the smallest. Thus, the problem of finding the set of P locations that minimize the total demand-weighted distance is referred to as the P-median problem. To formulate the problem, we define the following notation:

Sets
I = set of customers;
J = set of candidate facility locations.

Demand and Cost Parameters
μ_i = demand volume at customer location i, $i \in I$;
d_{ij} = distance between locations i and j, $i \in I, j \in J$
P = number (budget) of facilities to be located.

Decision Variables
$X_j = 1$ if facility is opened at location $j \in J$, 0 otherwise;
$Y_{ij} = 1$ if facility at $j \in J$ is assigned to serve customer location $i \in I$.

The problem is to select, out of the candidate set J, some P facilities, and assign them to serve customers in set I. These decisions are indicated by the X_j and Y_{ij} binary decision variables, respectively. In the P-Median problem formulation provided below, the objective is to minimize the total distance between customers and their assigned facilities, weighted by demand (1.1). The constraints stipulate that each customer location must be assigned to one facility (1.2), that such assignment can only be made if said facility is opened (1.3), and that the number of facilities opened equals P (1.4).

$$
\begin{aligned}
[P\text{-Median}] \quad \min \quad & \sum_{i \in I} \sum_{j \in J} \mu_i d_{ij} Y_{ij} && (1.1)\\
\text{s.t.} \quad & \sum_{j \in J} Y_{ij} = 1 \text{ for } i \in I && (1.2)\\
& Y_{ij} - X_j \leq 0 \text{ for } i \in I, j \in J && (1.3)\\
& \sum_{j \in J} X_j = P && (1.4)\\
& X_j \in \{0,1\} \text{ for } j \in J\\
& Y_{ij} \in \{0,1\} \text{ for } i \in I, j \in J.
\end{aligned}
$$

For various properties and solution heuristics of the P-median problem, one may refer to, e.g., the recent review by Daskin and Maass (2015).

A closely-related model is the uncapacitated fixed charge facility location model, which is often also referred to as the uncapacitated facility location (UFL) model. In this model, the hard budget constraint (1.4) is relaxed; instead, opening a facility at site $j \in J$ incurs a fixed cost of f_j. By considering an objective function that combines the fixed costs of opening facilities and the distance-based costs of serving customers,

the UFL model may provide a more flexible characterization of the trade-off between the budget of locating facilities and access distance. Let ρ be the unit cost of serving one unit of customer demand per unit distance between the customer and the assigned facility (e.g., unit shipping cost). Then, the uncapacitated fixed charge location model can be formulated as follows:

$$[\text{UFL}]: \quad \min \quad \sum_{j\in J} f_j X_j + \rho \sum_{i\in I}\sum_{j\in J} \mu_i d_{ij} Y_{ij} \qquad (1.5)$$
$$\text{s.t.} \quad (1.2), (1.3).$$

It is also noted that both the P-median and UFL models do not consider capacity of facilities (e.g., available land area for warehouses). Let C_j be the maximum demand volume that can be handled by a facility at $j \in J$. The capacitated fixed charge facility location model (CFL) is formulated by adding the following capacity constraint, which limits the volume of customer demand that can be assigned to a facility, to the UFL model:

$$\sum_{i\in I} \mu_i Y_{ij} \le C_j \text{ for } j \in J. \qquad (1.6)$$

In generalizing the UFL to the CFL model, one consideration of note is the modeling of single versus multiple sourcing. In the UFL model, one may note that the constraints that Y_{ij} must take on binary values can be relaxed without loss. This is because, given binary values of X_j, the remaining problem in the Y_{ij} variables is a bipartite assignment problem, which is a special case of the minimum cost flow problem. Thus, the basic feasible solutions (in $\mathbf{Y}$) are naturally integer-valued (see, for example, Section 11.4 of Ahuja *et al.* (1993) for more detailed discussions). This suggests that, under the UFL setting, it is always optimal to serve all demand from a customer site to the same facility, i.e., use single sourcing. In fact, it can be observed that it is always optimal to assign all demand at a customer location to the nearest open facility. In the CFL model, however, due to the additional capacity constraint (1.6), such closest assignment may not necessarily be feasible. Then, the distinction between single and multiple sourcing becomes relevant. If the application allows demand volume at the same customer site to be split in proportions (given by Y_{ij}) among multiple facilities,

one may relax the binary constraints on Y_{ij} to simply $0 \leq Y_{ij} \leq 1$, which potentially improves the objective value.

1.1.2 Coverage Distance

The demand-weighted distance objective provides an average-case view (over the set of customers) of the facility network, by considering the aggregate service measure (measured by access distance) provided to all customers, weighted by demand sizes. This may not be the most appropriate objective in applications where the worst-case service provision to customers is of primary concern. For example, for emergency medical services, the planning objective is often to maximize the volume or proportion of potential demand that can be served within a prescribed time guarantee, rather than the average response time to requests. Similar considerations arise in retail settings, where stores can attract customers located within certain distances. In these applications, the primary concern in planning is whether or not a facility is available within a certain critical distance, which is referred to as the coverage distance, to each customer.

To reflect whether a customer is located within the coverage distance, denoted by d_C, of a facility, we define the binary parameter $a_{ij} = \mathbf{1}(d_{ij} \leq d_C)$, where $\mathbf{1}(\cdot)$ denotes the indicator function. Then, we can formulate the set covering location model (Toregas *et al.*, 1971), which aims to locate the minimum number of facilities to cover all customers within the coverage distance.

$$\text{[Set Covering Location]}: \quad \min \quad \sum_{j \in J} f_j X_j \tag{1.7}$$

$$\text{s.t.} \quad \sum_{j \in J} a_{ij} X_j \geq 1 \text{ for } i \in I \tag{1.8}$$

$$X_j \in \{0, 1\} \text{ for } j \in J.$$

The objective (1.7) is to minimize the number (or more generally, opening costs) of facilities required to satisfy constraints (1.8) that require at least one facility to be opened within the coverage radius from each customer location. The set covering location problem has important applications in the public sector. For example, the location

of facilities such as hospitals, emergency medical services, police and fire stations, and schools, all should incorporate the access radius as a primary criterion in planning.

More generally, the set covering problem is one of selecting an optimal (minimum-cost) set of subsets of a collection of elements under the constraint that all elements have to be covered in at least one selected subset. In the facility location context, the elements refer to customer locations, and each feasible subset of elements is defined as the group of customers within the coverage distance of each candidate facility location. Thus, selecting among these subsets of customers is equivalent to selecting among candidate locations. Furthermore, the constraint that all elements are included in selected subsets is interpreted as requiring all customers to be covered within the prescribed coverage distance from some selected facilities.

The general formulation for set covering (Roth, 1969) is provided as follows. Let I be the set of elements to be covered, and $N \subseteq 2^I$ be a collection of feasible subsets of I. Then, for each member $R \in N$, we define the binary decision variable Z_R to indicate whether the set R is selected or not, with the cost associated given by c_R. Then, the general set covering problem can be formulated as:

$$[\text{General Set Covering}]: \quad \sum_{R \in N} c_R Z_R \tag{1.9}$$

$$\text{s.t.} \quad \sum_{R \in N : i \in R} Z_R \geq 1 \text{ for } i \in I \tag{1.10}$$

$$Z_R \in \{0, 1\} \text{ for } R \in N.$$

Interestingly, the general set covering problem arises in the solution procedure of some class of integrated location models with weighted-distance objectives. We shall revisit this in Section 3.

One limitation of the set covering location problem is its strict requirement that all customers must be covered, which was appropriate in the original context studied by Toregas *et al.* (1971) of locating emergency service facilities. While this requirement is often necessary for public sector services, we note that it is often the case that the marginal demand coverage for increasing the number of facilities is decreasing. Thus, in settings involving planners in the private sector,

it is often beneficial to leave out certain customers that are too costly to cover. One may then consider the maximum covering problem that maximizes demand coverage subject to a given budget to locate facilities, formulated as follows:

$$\text{[Max Covering Location]}: \quad \max \quad \sum_{i \in I} \mu_i U_i \tag{1.11}$$

$$\text{s.t.} \quad \sum_{j \in J} a_{ij} X_j \geq U_i \text{ for } i \in I \tag{1.12}$$

$$\sum_{j \in J} X_j \leq P \tag{1.13}$$

$$U_i, X_j \in \{0, 1\} \text{ for } i \in I, j \in J.$$

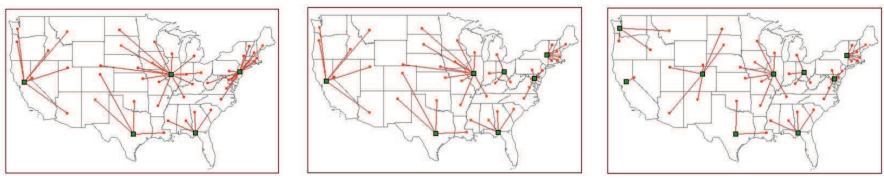
In the above, the objective (1.11) is to maximize the volume of demand being covered by the network of facilities, where binary decision variable U_i indicates whether customer location i is covered. Constraints (1.12) are similar to (1.8) in the set covering problem, but allow the flexibility of not covering certain customer locations, in which case they do not contribute to the objective ($U_i = 0$). Constraint (1.13) limits the number of facilities to the budgeted number (P).

1.1.3 Motivation for Integrated Modeling

The location models discussed so far focus on the fundamental trade-off between facility location costs and access distance. Despite the strategic importance of this trade-off, we may observe in a variety of applications that this alone is inadequate to capture other important strategic considerations in location design. Here, we provide an illustration based on a supply chain design setting.

Consider the problem of deploying DCs to serve a geographical market (e.g., the contiguous US). For illustration, we use the 49-node data set provided by Daskin (2011). The 49 nodes, which serve as both customer locations and candidate facility locations are the state capitals of the 48 contiguous states and Washington DC. The demand rates at each of these customer nodes are assumed to be proportional to the state populations and the shipping costs are proportional to great

circle distances[1] between the cities. Following the classical modeling approach, one might determine the locations based on the UFL model, by considering location costs f_j as the (annualized) construction and operating costs of the DCs and ρd_{ij} as the shipping cost per unit demand between two locations i and j. To illustrate the trade-off between location and transportation costs, we vary the weight $\rho = 1, 1.5, 2$ on the unit transportation cost and compare the optimal DC locations, as mapped in Figure 1.1. Intuitively, a higher transportation cost weight leads to locating more DCs, as higher unit transportation costs favors reducing shipping distances from DCs to customers by increasing the density of DCs. In general, the relative magnitudes of the location cost and transportation cost weights determine the degree of *consolidation* of the supply chain network.



(a) $\rho = 1$ (b) $\rho = 1.5$ (c) $\rho = 2$

Figure 1.1: UFL Solutions Under Different Transportation Costs

However, one may notice that the aforementioned consideration does not fully capture the consolidation-deconsolidation trade-off in strategic distribution network design. In supply chain management, it is well known that facility costs, transportation costs and inventory costs are the three major cost components driving network design decisions (e.g., Chopra and Meindl, 2007). The conventional models focus on the former two, but do not account for inventory costs. To illustrate why this can be a problem, we extend the example by comparing the UFL setting with two other alternative settings that consider inventory costs.

[1]The great circle distance is the shortest distance between two points on a sphere. It is often used as a proxy for the straight-line distance between two cities, adjusted for the Earth's surface curvature.

Figure 1.2 (a) shows the solution to the UFL problem for the dataset (with $\rho = 1$), which consists of five opened facilities in Sacramento (CA), Austin (TX), Tallahassee (FL), Springfield (IL) and Trenton (NJ). We refer to the UFL problem as Setting 1 and the corresponding optimal solution (set of chosen facilities) as Solution 1. To account for inventory costs, consider a setting (Setting 2) in which demand is random (with mean and standard deviation proportional to state population). Each facility, once located, needs to carry enough safety stock to ensure a 95% Type-1 service level. Under this alternative model, we may solve a stochastic optimization model to obtain the optimal solution (Solution 2) illustrated in Figure 1.2 (b). One can observe that there are now only three DCs instead of five.

(a) UFL Solution (b) Service Level (c) Transshipments

Figure 1.2: Maps of Location Plans under Different Model Settings

One may naturally wonder why locating three DCs rather than five (at different locations) would be optimal as one considers safety stock holding costs. One major reason is the effect of risk pooling (Eppen, 1979). In particular, safety stock can be reduced by pooling larger volumes of demand at smaller number of DCs. This "statistical" economies of scale effect tilts the optimal balance in the consolidation-deconsolidation trade-off and causes the optimal number of DCs to be reduced. A more detailed discussion of such effects will be provided in later chapters. To make things even more interesting, we consider another alternative setting (Setting 3) in which facilities may transship inventory among themselves to cope with random demand. Furthermore, instead of satisfying a Type-1 service level, the safety stock level is chosen to minimize a newsvendor-type cost function including holding, shortage, and transshipment costs. The resulting optimal solution (Solution 3) is provided in Figure 1.2 (c). Interestingly, not only is it optimal to

locate five rather than three DCs, but the locations are also slightly different from Solution 1; in particular, Montegomery (AL) is selected instead of Tallahassee (FL). This is, in part, due to the consideration of transshipment operations. First, with the possibility of sharing inventory through transshipments, it is possible to share inventory and achieve pooling benefits without the need to deliberately consolidate the network of DCs (i.e., reducing to four DCs in the case of no transshipments). Second, the specific locations of the five DCs is the outcome of the trade-off between minimizing transportation costs to customers and transshipment costs. The former encourages DCs to be located closer to centers of customer clusters, and the latter encourages DCs to be placed closer to each other. With the additional consideration of the transshipment effect, the choice of Montegomery (AL) allows the set of DCs to be, on average, more centrally located within the country.

Table 1.1: Percentage Performance Gaps of the Three Solutions Under the Three Settings

	Solution 1	Solution 2	Solution 3
Setting 1	0.00%	2.15%	0.23%
Setting 2	6.29%	0.00%	6.19%
Setting 3	1.95%	5.90%	0.00%

Note that the optimal strategy in one setting is suboptimal in the others. In Table 1.1, we compare the performance of each of the three solutions under each of the three settings. In particular, we report the percentage cost increase of each solution over the optimal one in the same setting. We observe that both Solutions 1 (6.3% worse than optimal) and 3 (6.2% worse than optimal), which suggest opening five DCs, perform substantially worse in Setting 2 than the optimal solution (Solution 2). This suggests that failure to account for the risk-pooling effect leads to significant cost increases. On the other hand, Solution 2 also performs relatively poorly under Setting 3, suggesting that failure to account for transshipment opportunities at the network design stage also leads to cost inefficiencies. Finally, although Solution 1 differs from Solution 3 by the location of just one DC, it performs about 2% worse

under Setting 3. This further highlights that importance of selecting the right set (on top of the right number) of facilities for the problem setting on hand.

From this simple illustrative example, we can see that conventional models that consider generic, distance-only objectives (e.g., the UFL model) may fail to capture important design characteristics arising from specific operations of certain facility types, leading to significantly sub-optimal network designs. This potential shortcoming can be overcome by enhancing the models with an integrated view of both the conceptual cost-distance trade-off and the operating characteristics of the specific facility types. This monograph is dedicated to reviewing the recent developments of this line of research.

1.2 Aims and Scope

While we have briefly introduced the classical facility location models in Section 1.1, we do not attempt to provide a comprehensive review of this extensive literature. Our focus will be on integrated models that incorporate operational features of facilities beyond distance-focused considerations. For more comprehensive reviews and discussions of the properties and solution strategies for classical models, as well as various extensions, applications and modeling discussions, one may refer to the excellent texts by Daskin (2011), Drezner (1995), Hamacher and Drezner (2002), and Laporte *et al.* (2015). Likewise, while many of the applications we shall discuss make use of important results in research streams such as inventory theory to model operational features of facilities, we also do not intend to provide a full review of these areas beyond what is required to develop the integrated facility location models. Interested readers may refer to, for example, Zipkin (2000), for more complete discussion and references.

The study of integrated facility location modeling has a long history. In the 1980's, works by Daskin (1983), Eaton *et al.* (1985), and ReVelle and Hogan (1989) consider the operational characteristics of mobile facilities such as ambulances and the optimal deployment strategies taking into account congestion probabilities. However, it was until the 2000's when this research area sustained very rapid growth. Part of

the reason was the significant computational challenges associated with solving the integrated models, which had been difficult to overcome before the recent advancements in computational power of computers as well as optimization (particularly, stochastic and nonlinear integer programming) theory. With the rapid growth, a myriad of modeling approaches, solution methodologies and application areas have been proposed by researchers and practitioners. The aim of this monograph is to provide a timely review of some of these important developments. With the exploding growth and huge volume of related research, our review is inevitably restricted in scope and cannot be comprehensive. As our aim is to review major modeling approaches, solution methodology and some promising current and future research directions, some application areas are inevitably omitted. For other recent reviews, we refer interested readers to Shen (2007) and Mak and Shen (2011). It is also notable that simulation-optimization techniques, designed for ranking and selection problems where performances of alternatives can be evaluated via simulation, are also a promising approach to the class of problems that we consider, since the operational performance of facilities can be simulated in detail. Our focus will be mainly on mathematical programming and analytical modeling approaches, and refer interested readers to Fu (2002), Hong and Nelson (2009), and Luo *et al.* (2015) (and the references therein) for this alternative methodology.

In this monograph, we provide discussion on four aspects of the research stream. In Chapter 2, we discuss several popular modeling approaches employed by researchers to model integrated location problems, such as nonlinear integer programming, stochastic programming and continuous approximation. In Chapter 3, we provide a brief account of some promising solution methodologies, including decomposition methods and conic optimization methods. Then, in Chapters 4 and 5, we draw from the broad range of applications of the integrated modeling framework in the classical supply chain design context and several other emerging application domains, respectively. Finally, we conclude the volume and discuss some promising future research directions in Chapter 6.

1.3 Notation

Throughout the monograph, we use boldface letters to represent matrices or vectors of variables denoted by the same letters. For example, $\mathbf{Y}$ is the matrix with components being the Y_{ij}'s. Furthermore, $\mathbf{x}'$ denotes the transpose of column vector $\mathbf{x}$, and $\mathbf{x}'\mathbf{y}$ denotes the inner product of column vectors $\mathbf{x}$ and $\mathbf{y}$.

2

Integrated Modeling Approaches

In this section, we first provide a review and some discussion on some major modeling approaches commonly employed to formulate integrated models for facility location. Following the previous discussion on integer programming formulations for classical location models, the aim of this section is to provide readers with some ideas about how further techniques may be required to capture various operational features of facilities. In each section, we illustrate the key modeling approach with a specific example from the literature. With this discussion, we hope that (especially beginner) readers will have some preliminary ideas of how to formulate models for their problems of interest. Afterwards, in Chapter 3, we shall discuss how these models often require the use of certain computational methodologies.

2.1 Nonlinear Integer Programming

As discussed previously, the distance-concerned models, such as those introduced in Section 1.1, may fall short of capturing the important operational characteristics of facilities. We first discuss a general extension of the UFL model that reflect the costs incurred in the tactical and

operational phases of managing the network of facilities. In particular, we may use a function $G(\mathbf{Y})$ to denote the (non-distance-dependent) tactical and operational costs for the network of facilities, given demand assignments according to $\mathbf{Y}$. Then, we may generalize the objective function of the UFL as:

$$\min \sum_{j \in J} f_j X_j + \sum_{i \in I} \sum_{j \in J} \rho \mu_i d_{ij} Y_{ij} + G\left(\mathbf{Y}\right). \tag{2.1}$$

The key feature in formulation is the function $G(\cdot)$. Operational characteristics of facilities can be reflected by selecting appropriate functional forms. As shall be discussed later, much of the literature focuses on functional forms that are separable by j, i.e., $G\left(\mathbf{Y}\right) = \sum_{j \in J} G_j(\mathbf{Y_j})$. Separability indicates that the facilities do not interact in their tactical and operational decisions, beyond dividing up the market in the strategic phase (i.e., with the values of $\mathbf{Y}$ determined). For cases where facility operations do interact, such as in the case discussed in Section 1.1.3 where facilities perform lateral transshipments, the function $G(\cdot)$ will not be separable.

By selecting appropriate forms of the $G(\cdot)$ function, operational characteristics such as economies (or diseconomies) of scale in demand volume can be modeled. Throughout this monograph, we shall discuss various models that follow from selecting different functional forms of the $G(\cdot)$ function. To begin, we consider the supply chain design (SCD) model proposed by Shen *et al.* (2003), which was among the first integrated facility location model for supply chain design in the literature. In this model, the inbound transportation costs, as well as cycle and safety inventory holding costs incurred at DCs are considered and modeled in the $G(\cdot)$ function. From classical inventory theory (e.g., Zipkin, 2000; Axsäter, 2007), these cost factors exhibit economies of scale (through economic ordering cycles and risk pooling), in the sense that the marginal cost to serve incremental demand volume tends to decrease. To capture this, Shen *et al.* (2003) develop a model in which the function $G(\cdot)$ is equivalent to a concave function in the demand volume assigned to each DC.

We begin the discussion by modifying and defining the following parameters:

Demand Parameters

μ_i = Mean (daily) demand at retailer i;
σ_i^2 = Variance of (daily) demand at retailer i;
α = Type 1 service level, i.e., target probability of having no stock-outs;
z_α = Standard normal z-score corresponding to α, where $P(z \leq z_\alpha) = \alpha$;
L = replenishment lead time for DC.

Cost Parameters

β = Weight factor associated with transportation cost;
θ = Weight factor associated with inventory cost;
F_j = Fixed (administrative) cost of placing an order at warehouse j;
g_j = Fixed shipping cost from supplier to warehouse j;
$\bar{a}_j$ = Variable shipping cost (per unit) from supplier to warehouse j;
h = Inventory holding cost per unit per year;
χ = Number of days in a year.

Shen *et al.* (2003) consider the case where demand volumes at retailers follow independent normal distributions. In their model, in addition to the fixed facility location costs and DC-retailer shipping costs as considered in the UFL model, they incorporate the costs of holding and replenishing inventory at the DCs in the objective function. In particular, they consider each DC j to replenish inventory following continuous review (r, Q) policies, with reorder points determined based on Type 1 service level α. For tractability, they adopt the approximation proposed by Axsäter (1996) of determining the cycle order quantities using the EOQ as described below.

The mean daily demand handled by DC j is given by $D_j = \sum_{i \in I} \mu_i Y_{ij}$. Suppose there are n replenishment cycles (where n is to be optimized) in a year. Then, the order quantity per cycle is, on average, $\chi D_j/n$ and the shipment cost per replenishment order is $g_j + \bar{a}_j \chi D_j/n$, which consists of the fixed and variable components. The average cycle inventory can be approximated by $\chi D_j/(2n)$. The holding and replenishment costs involved in an ordering cycle is then:

$$F_j n + \beta(g_j + \bar{a}_j \chi D_j/n)n + \theta h \chi D_j/(2n).$$

Optimizing over n, we obtain the following as the optimal number of order cycles and the corresponding cycle holding and replenishment costs:

$$\begin{aligned} n^* &= \sqrt{\frac{\theta h \chi D_j}{2(F_j + \beta g_j)}} \\ C^*_{\text{cycle}} &= \sqrt{2\theta h \chi D_j (F_j + \beta g_j)} + \beta \bar{a}_j \chi D_j. \end{aligned} \tag{2.2}$$

Furthermore, to maintain a Type-1 service level of α, the required safety stock is given by z_α times the standard deviation of demand during lead time, i.e., $\sqrt{L \sum_{i \in I} \sigma_i^2 Y_{ij}}$. Thus, the annual holding cost for the safety stock at warehouse j is then given by:

$$\theta h z_\alpha \sqrt{L \sum_{i \in I} \sigma_i^2 Y_{ij}}. \tag{2.3}$$

Incorporating the above inventory-related costs to the UFL model, we obtain the following optimization model:

$$\begin{aligned} [SCD:] \min \quad & \sum_{j \in J} \Bigg[f_j X_j + \sum_{i \in I} (d_{ij} + a_j) \beta \chi \mu_i Y_{ij} \\ & + \sqrt{2\theta h (F_j + \beta g_j)} \sqrt{\sum_{i \in I} \chi \mu_i Y_{ij}} + \theta h z_\alpha \sqrt{\sum_{i \in I} L \sigma_i^2 Y_{ij}} \Bigg] \\ = \quad & \sum_{j \in J} \Bigg[f_j X_j + \sum_{i \in I} \hat{d}_{ij} Y_{ij} + K_j \sqrt{\sum_{i \in I} \mu_i Y_{ij}} \\ & + q \sqrt{\sum_{i \in I} \sigma_i^2 Y_{ij}} \Bigg] \qquad (2.4) \\ \text{s.t.} \quad & \sum_{j \in J} Y_{ij} = 1 \text{ for } i \in I \qquad (2.5) \\ & Y_{ij} - X_j \leq 0 \text{ for } i \in I, j \in J \qquad (2.6) \\ & X_j \in \{0, 1\} \text{ for } j \in J \qquad (2.7) \\ & Y_{ij} \in \{0, 1\} \text{ for } i \in I, j \in J \qquad (2.8) \end{aligned}$$

where:

$$\begin{aligned} \hat{d}_{ij} &= (d_{ij} + a_j)\beta\chi\mu_i \\ K_j &= \sqrt{2\theta h(F_j + \beta g_j)} \\ q &= \theta h z_\alpha \sqrt{L}. \end{aligned}$$

Recall in our discussion of the CFL and UFL models that, whenever the capacity constraint is not violated, it is optimal to determine the Y_{ij} variables by assigning each retailer to the nearest open DC. However, this property does not necessarily hold for the [SCD] model, as pointed out by Shen *et al.* (2003) and Daskin *et al.* (2002). In particular, due to the consideration of inventory costs, it may be optimal to assign a retailer to a DC not necessarily nearest to it such that better inventory pooling is achieved and the overall costs can be reduced. They identify examples that retailers may not even be assigned to DCs at the same locations (i.e., when $d_{ij} = 0$). Interestingly, they also show that the nearest-assignment property necessarily holds at the optimal solution when the mean-variance ratio of demand is equal for all retailers (i.e., $\sigma_i/\mu_i = \gamma$ for all $i \in I$). This condition is valid, for example, when demand arises from Poisson processes, in which case $\gamma = 1$. Under this condition, the objective function of the [SCD] model can be simplified as:

$$\min \sum_{j\in J} \left[f_j X_j + \sum_{i\in I} \hat{d}_{ij} Y_{ij} + \hat{K}_j \sqrt{\sum_{i\in I} \mu_i Y_{ij}} \right] \tag{2.9}$$

where $\hat{K}_j = K_j + q\sqrt{\gamma}$.

Formulations in the form of (2.4) or (2.9) involve nonlinear (square root) terms. In general, as the operational characteristics of facilities vary across applications, it is not guaranteed that the $G(\cdot)$ function is convex (or can be represented using convex optimization formulations). Therefore, such formulations may not generally be solvable by standard integer linear or convex programming branch-and-bound methods available in commercial solver packages (such as CPLEX or Gurobi). In the literature, one main research focus is on developing efficient solution algorithms for these problems by exploiting problem-specific structural properties. As shall be discussed in Section 3.1, efficient approaches for

solving the [SCD] problem exploit the concavity property of the square root terms in 3.1. In Section 3.2, we also discuss how the [SCD] problem can be transformed into mixed-integer second order cone programs (MISOCP), whose continuous relaxations are convex conic optimization problems, and how to accelerate solution algorithms with specific cutting planes.

2.2 Stochastic Programming

In various applications, such as in supply chain network design, tactical planning for facilities are often subject to substantial uncertainty. The prime example is inventory planning for DCs subject to demand uncertainty. In light of such uncertainties, operating flexibility is crucial for the facility network. Such flexibility introduce mutual dependence in the tactical-phase operating characteristics among facilities, and thus, require non-separable forms of function $G(\cdot)$ to be modeled appropriately. Furthermore, to reflect planning uncertainty, it is common to model the function $G(\cdot)$ using a stochastic programming representation. This approach is particularly useful when the operational characteristics of facilities involve repeated re-optimization of certain decisions (e.g., routing of shipments) over discrete time epochs in the future, where uncertainty in the operational environment persists.

In this section, we illustrate the stochastic programming modeling approach based on the example discussed in Section 1.1.3 in which facilities hold inventory to meet stochastic demand, and are allowed to share inventory by performing lateral transshipments to better avoid shortages. This application fits the stochastic programming approach very well, because the operational phase of the problem involves repeated optimization of inventory, transshipment and demand allocation decisions. Over the long run, a stochastic program with an expected cost objective could be used to model the long-run average operational performance of the facility network.

To model stochasticity, we use the random variable $\tilde{D}_i(\omega)$ (with mean μ_i and standard deviation σ_i) to denote the random demand at customer location i, where $\omega \in \Omega$ denote a specific realization or "scenario" of random events governing demand outcomes. To meet random demand,

we define S_j as the inventory level held at DC j. We define decision variables $V_{jk}(\omega)$ to denote the amount of inventory transshiped from facility j to facility k under scenario ω. For transshipments from j to k, a transportation cost of τ_{jk} per unit is incurred. If demand cannot be fully satisfied in a scenario, there may possibly be a shortage at DC j, denoted by the decision variable $B_j(\omega)$. Each unit of shortage (assumed to be backordered for simplicity) will incur a penalty of p. Then, the cost components related to inventory and transshipments can be formulated as follows:

$$
\begin{aligned}
G(\mathbf{Y}) = \min \quad & E_\omega \left[\sum_{j \in J} \left(\sum_{k \in J \setminus \{j\}} \tau_{jk} V_{jk}(\omega) + p B_j(\omega) \right) \right] \\
& + \sum_{j \in J} r S_j \qquad (2.10) \\
\text{s.t.} \quad & S_j + \sum_{k \in J \setminus \{j\}} [V_{kj}(\omega) - V_{jk}(\omega)] \\
& = \sum_{i \in I} D_i(\omega) Y_{ij} - B_j(\omega) \text{ for } j \in J, \omega \in \Omega \quad (2.11) \\
& V_{jk}(\omega) \geq 0 \text{ for } j \in J, k \in J \setminus \{j\}, \omega \in \Omega \\
& B_j(\omega) \geq 0 \text{ for } j \in J, \omega \in \Omega.
\end{aligned}
$$

In the above, we consider transshipment operations for a single period, which enables the problem to be formulated as a two-stage stochastic program. The objective (whose optimal value yields $G(\cdot)$) is to minimize expected costs of transshipments and shortage, plus the cost of carrying inventory at facilities (2.10). The constraints (2.11) impose material flow balance at facilities, by requiring the stocking levels plus net inbound transshipments to equal realized demand less shortages. Note that under the assumption that shortages are backordered, the expected DC-to-customer shipping costs correspond to the (mean) demand-weighted distance, and thus can be reflected by the same $\sum_{i \in I} \hat{d}_{ij} Y_{ij}$ term considered in the [SCD] problem.

Remark 2.1. In inventory theory (e.g. , Robinson, 1990), it is known that the optimal inventory control policies of transshipment problems exhibit stationarity under stationary problem parameters and certain

regulatory conditions. Therefore, one can think of the above single-period formulation as modeling the long-run average costs for a periodic-review system operated over a longer time horizon.

Substituting the above formulation (2.10) in (2.1), one can obtain an integer stochastic linear programming formulation for the supply chain design problem with transshipments. This two-stage model can be interpreted as having a strategic phase, in which the network design is to be determined subject to demand uncertainty, and a tactical phase, in which transshipments and shipping decisions are made in response to evolving demand information. Due to high dimensionality, this class of integer stochastic programs are often difficult to solve. We shall discuss some possible solution strategies, such as decomposition methods, for similar problems in Section 4.5.

The modeling approaches discussed thus far focus on mathematical programming methods. As shall be discussed in future chapters, the main focus of the analysis of such models is to obtain numerical solutions computationally. In the next section, we shall review a complementary modeling approach that is more amenable to analytical studies.

2.3 Continuous Approximation

Fundamentally, the optimal location strategy is governed by trade-offs among various strategic forces, including the distance-cost trade-off discussed in Chapter 1. Understanding of the underlying trade-offs must be developed to assist managers to choose the best strategies under different scenarios. The models discussed so far are all formulated as (stochastic, linear or nonlinear) integer programming problems. The fact that they are extensions of the classical models discussed in Section 1.1 imply that these problems are generally NP-hard. Much of the literature has focused on computational approaches, and on designing efficient solution methods in particular. As to be discussed later, many of these algorithmic studies have made possible the investigation of interesting case studies regarding various applications based on specific data sets. This approach is best suited for producing a detailed and implementable design that is (close to) optimal, given a set of input

data specific to the application. However, one limitation is that these integer programs typically do not exhibit analytical tractability, and thus qualitative managerial insights can rarely be obtained by analyzing the models' properties.

To remedy the analytical challenges, researchers have adopted an alternative continuous approximation (CA) approach that is more amenable to analytical (as opposed to computational) studies. The modeling philosophy behind this approach can be summarized as follows (Daganzo, 2005): Using concise data summaries and analytical models in place of detailed data and numerical algorithms, it is possible to formulate closed-form representations of design and operational cost of facility networks. These models characterize the most important characteristics of the location design problem, while abstracting out implementational details. The approximations are parsimoniously developed to ensure analytical tractability and, at the same time, to capture the underlying trade-offs inherent to the problem. Then, managerial insights can be drawn by analyzing the mathematical structures of the models.

The key starting point of the CA modeling approach is to consider a generic problem in a large and spatially homogeneous region, on which customers are evenly spread with some density. That is, we start with a problem that abstracts away spatial heterogeneity, because such level of detail is only required for implementational purposes and often does not alter the fundamental trade-offs for many problems. Naturally, because the region is homogeneous, the optimal facility locations will also be evenly spaced. This is the key to the approximation approach, as this allows us to optimize the *density* of facilities rather than specific locations. Therefore, CA models capture scale effects of the facility network, but ignores spatial heterogeneity. As to be discussed later, this limitation can be overcome numerically for cases where problem parameters vary slowly over space.

To illustrate the CA modeling approach and its power in generating qualitative insights, we consider the following alternative modeling scheme for the [SCD] problem formulated in Section 2.1. Instead of a network consisting of retailer and candidate facility locations, we

consider an infinite homogeneous plane on which retailers are uniformly located with a density of δ retailers per unit area. As detailed in Daganzo (2005), one can consider either the case where retailers are located at constant distance apart or one where they are randomly located with uniform density. Each retailer faces independent stochastic demand following a Poisson process over time, with rate μ. Because the problem is homogeneous, we can consider DCs to be located at grid points of the plane (Figure 2.1). A uniform location scheme is optimal if retailers are also located at grid points (with a different density), and will be optimal in the expectation sense for the case where retailers are randomly located.

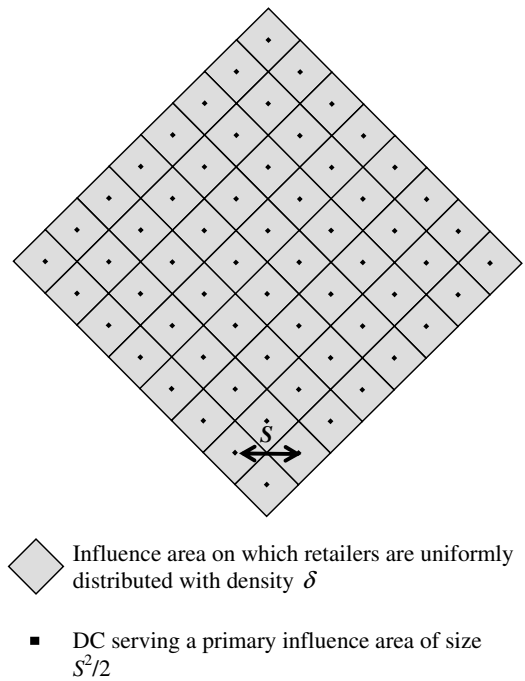


Figure 2.1: Segment of Service Region with Square-shaped Primary Influence Areas

We also assume that shipping distances are measured using the $\ell - 1$ metric, that is, the distance between two points on the plane, with coordinates (x_1, y_1) and (x_2, y_2) respectively, is considered to be $|x_1 - x_2| + |y_1 - y_2|$. This is also known as the Manhattan distance

metric and is useful for modeling travel distances in cities, where the metric measures distance over a grid-shaped road network. Under this metric, it can be shown that the optimal allocation of customers to facilities, such that customer-facility distances are minimized, follow the rotated square-shaped layout shown in Figure 2.1. Following this layout, the separation between every two adjacent DCs is $S/\sqrt{2}$ miles, meaning that each DC has a square-shaped influence area of $S^2/2$ square-miles. This greatly simplifies the problem, as we only need to optimize the density of DCs instead of their specific locations. As a result, we are able to formulate the problem with a single decision variable, S.

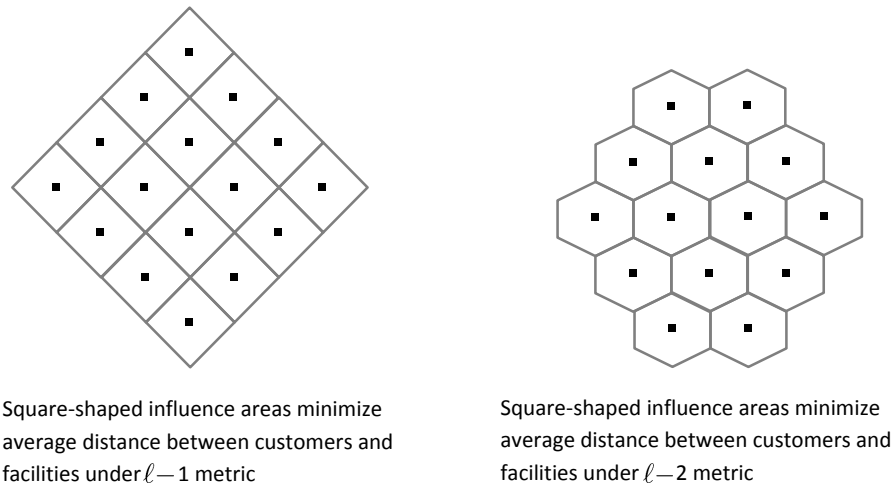


Figure 2.2: Optimal Partition of Influence Areas under $\ell - 1$ and $\ell - 2$ Metric

Remark 2.2. One may note that, under the $\ell - 2$ metric, i.e., where the distance between two points is given by the Euclidean distance $\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$, a hexagonal (instead of square) partition of the plane is optimal (e.g. Cui *et al.*, 2010). See Figure 2.2 for an illustration. Newell (1973) shows that different uniform partitioning schemes lead to cost models that only differ by constant multiples, and uses numerical examples to point out that exact shapes do not matter much as long as they are relatively round shaped. For example, the average travel distances under hexagonal, diamond and square

arrangements only differ by a few percent. Rather than the exact shape, Newell suggests that the researcher should focus on optimizing the scale or the size of the influence areas. In this analysis, details such as the exact shapes of influence areas are not critical issues in high-level planning and therefore simplifying assumptions can be made about them.

The problem is to determine the density of facilities (equivalently, the value of S), to minimize the continuous analog of objective function (2.9), which consists of three components: fixed location costs, transportation costs and inventory costs. Note that, because the service region in question is an infinitely large plane, the total costs will be infinitely large. Therefore, it is conventional to consider average costs per retailer (or equivalently, per unit area). As shall be discussed later, this convention makes it easier to extend the models for spatially heterogeneous cases. We briefly discuss how these components are modeled below.

We first note that, as each DC serves an influence area of $S^2/2$, there are (on average) $S^2\delta/2$ retailers served by each DC. Conversely, the fixed cost incurred by locating each DC, denoted by f (because all locations have the same cost, the index j is dropped), is shared among $S^2\delta/2$ retailers on average. Thus, the average fixed location cost per retailer is given by $\frac{2f}{\delta S^2}$. Second, it can be shown that the average $\ell - 1$ distance between any point in a square-shaped influence area and the DC in the center of the square is equal to $S/3$. Per unit time, an average of μ units are shipped from a DC to a retailer. Therefore, using $\hat{d}$ to denote the unit transportation cost, the average transportation cost incurred by a retailer per unit time is given by $\hat{d}\mu S/3$. Finally, following (2.9), the inventory cost incurred at each DC is given by a constant K times to the square root of the mean (and thus variance under Poisson demand) of demand volume handled, i.e., is equal to $K\sqrt{\delta S^2 \mu/2}$ given that the DC serves $\delta S^2/2$ retailers. Dividing this cost among the retailers, the average inventory cost per retailer is given by $\frac{K}{S}\sqrt{2\mu/\delta}$. Combining the above, the total supply chain costs as a function of S, denoted by $\mathbf{C}_{SCD}(S)$, become:

$$\mathbf{C}_{SCD}(S) = \frac{2f_1}{\delta S^2} + \frac{\hat{d}\mu S}{3} + \frac{K}{S}\sqrt{2\mu/\delta} \tag{2.12}$$

and the continuous approximation version of problem $[SCD]$ is formulated as $\min_{S>0} \mathbf{C}_{SCD}(S)$. It can be directly observed that $\mathbf{C}_{SCD}(S)$ is a convex function, and thus can be minimized by solving for the first order condition, which yields the following.

Proposition 2.1. The unique minimizer of (2.12) is given by the unique positive real root to:

$$S^3 - \frac{hz_\alpha\sqrt{\mu}}{2k_1 v}S - \frac{f_1 V}{4k_1\delta v} = 0.$$

Recall that the motivation for employing the CA methodology is to gain qualitative understanding on the underlying trade-offs of the facility network design problem. From equation (2.12), the basic trade-offs are evident. The fixed location cost term and inventory cost term are decreasing (and convex) in S, meaning that these cost factors create forces that push for consolidation (larger S or lower density of DCs). The reason is that both cost components exhibit economies of scale: the average location cost per retailer can be lowered if the DC serves more retailers; and the inventory costs can be reduced by pooling demand, both through saving on inbound transportation cost under the EOQ setting and through reduction of safety stock with risk pooling. On the other hand, the transportation cost term is increasing (linearly) in S, creating a force pushing for deconsolidation (smaller S). From this trade-off, we can immediately observe the effect of taking inventory costs into consideration. Without the inventory cost term, the optimal value of S will be smaller, i.e., failure to account for inventory cost considerations would lead to under-consolidation of the DC network. This observation was made by Shen *et al.* (2003) based on extensive computational studies using the [SCD] model. With the CA framework, the same conclusion can be obtained using intuitive algebraic arguments.

More generally, we evaluate the effect of the problem parameters on the optimal solution characterized in Proposition 2.1. The case where inventory costs are neglected is the special case where $K = 0$.

Corollary 2.1. The effects of input parameters on the minimizer of (2.12) are summarized in Table 2.1.

Table 2.1: Effect of Input Parameters on Optimal Solution

Increase in Parameter	Separation S
Retailer Density δ	$\downarrow$
Mean Retailer Demand μ	$\downarrow$
Fixed Cost f	$\uparrow$
Inventory Cost Rate K	$\uparrow$
Transportation Cost Rate $\hat{d}$	$\downarrow$

We have used the CA counterpart of [SCD] as an example to highlight the potential of the CA approach to greatly simplify the problem while retaining the structural characteristics of the problem for qualitative analysis. The resulting model can be solved and analyzed using elementary calculus. However, in certain applications, it requires special techniques to develop such tractable models. A very useful tool, known as dimensional analysis, will be discussed in Section 3.3.

Despite its modeling power, the CA approach suffers from the downside of not directly prescribing implementable solutions. Generally, this approach is more suited for theoretical analysis of strategic forces underpinning a network design problem, rather than data-driven implementation exercises. The reason that the results from CA models are not directly implementable is that the analysis focuses on uniform regions of homogeneous characteristics (e.g., geographical density of retailers and demand per retailer). Nevertheless, we remark that for problem instances where cost and demand parameters vary slowly spatially, it is possible to obtain high quality solutions for nonhomogeneous problems by making use of the same cost model. The basic idea is to *approximate* the local cost at each point in the region by assuming that this point belongs to an infinite homogeneous region with uniform problem parameter values equal to the local ones. By applying this approximation point-wise, we obtain a spatial cost function defined over the region of interest. Then the average cost over the region can be approximated by the spatial average of this cost function.

For example, consider a region R, where the demand rate per retailer $\mu(x)$ and the density of retailers $\delta(x)$ are functions of the location of

point $x \in R$. Then the local optimal DC separation $S^*(x)$ at point x can be approximated by substituting $\mu = \mu(x)$ and $\delta = \delta(x)$ into the condition in Proposition 2.1. Similarly, $\mathbf{C}(S^*(x))$ approximates the local cost at point x. Integrating this point-wise approximate cost function over the region R with density function $\delta(x)$, the total cost is *approximated* by:

$$\text{Total cost for region } R \approx \int_{x \in R} \mathbf{C}(S^*(x))\delta(x)dx. \tag{2.13}$$

Daganzo (2005) and Ouyang and Daganzo (2006) explain in detail why this approximation approach gives rise to high quality solutions to complicated logistics problems. In particular, the quality of the approximation is very high when the location-specific parameters ($\mu(x)$ and $\delta(x)$ for example) vary slowly with location x. Note that, although the uniform problem does not admit a simple closed-form solution, the integration in (2.13) can be easily implemented numerically.

With such an approximation scheme, the local solution at any point can be approximated point-wise using the optimal solution from an infinite homogeneous problem with parameter values equal to the local ones. Therefore, the relationship between the local solution and local parameter values at any point in a nonhomogeneous region is similar to that between the solution and parameter values in an infinite homogeneous problem. This suggests that any comparative statics results can be interpreted as the impacts of different local environmental conditions on the local optimal supply chain design, which further adds to the value of the qualitative insights obtained based on analyzing the homogeneous model.

2.4 Discussion

The nonlinear integer programming modeling approach follows naturally from classical integer programming models for facility location. As one incorporates operational features, such as inventory carrying dynamics, into the model, nonlinear terms are introduced. In most cases, this makes the problem less tractable. Traditionally, nonlinear terms are often avoided by many modelers due to their intractability

by standard methods (e.g., branch and bound methods with simplex subroutines). With advances in solution approaches to mixed integer nonlinear problems, whether a good solution can be obtained efficiently depends greatly on the functional form of the nonlinear terms introduced in the objective function and constraints. Thus, when formulating this class of models, the modeler ought to be cautious about the ensuing solution approach (some of which will be reviewed in Chapter 3) and computational tractability. Certain functional forms (e.g., convex, concave, submodular), in the right models, would allow efficient problem-specific algorithms to be developed by exploiting special structure. Other functional forms (e.g., involving conic constraints) could be solved readily with standardized solvers. Often, developing effective and efficient models involves evaluating the trade-off between the accuracy of capturing modeling features versus computational efficiency. Overall, the nonlinear modeling approach constitutes a key building block for developing rich models for facility location. As shall be seen in Chapter 5, this approach is promising in modeling many recent applications in different emerging domain areas.

Stochastic programming methods are important for modeling the risk aspect of strategic facility location problems. With advances in the understanding and modeling of risk measures, stochastic modeling of facility location problems is reaching beyond the risk neutral (expected value objective) view presented in Section 2.2 to incorporate other risk preferences of the decision maker. The use of convex risk measures, such as the conditional value-at-risk (which shall be discussed in Section 4.2.2), often allows risk preferences to be captured while preserving computational tractability of the problem. Often, to incorporate such risk measures, it is possible that the resulting model becomes nonlinear. Therefore, as discussed above for the case of nonlinear integer programming, the modeler should be cautious about the computational tractability aspect when attempting to build a rich risk preference model.

Compared with the mathematical programming techniques discussed, the CA approach has received less attention in the operations management community until fairly recently. This may have to do with the

strong development of research in operations economics (e.g., Cachon, 2012), which focuses on developing qualitative managerial insights into operations problems using economic modeling. For facility location problems, mathematical programming approaches are less amenable to the analysis of problem properties with managerial implications, because the combinatorial nature of integer programming problems, in most cases, limits researchers to purely computational analyses due to the discrete nature of decisions. CA models allows modelers to relax the confounding combinatorial problem and investigate instead an approximation in which the objective function and constraints are characterized by smooth functions of a concise number of continuous decision variables. We believe that advances in this approach open up new opportunities in developing insights into important location and network design problems, many of which were previously considered to be analytically intractable. Overall, this development helps bridge the divergence of the more computation-focused "operations research" and the more insights-driven "operations economics" branches of research in the context of facility location. Recently, an increasing number of papers using this modeling approach has appeared in top journals, including Cui *et al.* (2010), Lim *et al.* (2013), Cachon (2014), Lim *et al.* (2016), and Belavina *et al.* (2016). We believe that this is a promising direction to pursue.

3

Solution Techniques

After reviewing some popular modeling strategies employed in integrated location models in Chapter 2, we shall provide some discussion on common solution methodologies required to solve these problems. One will see that development of efficient solution schemes relies critically on the mathematical structure of the optimization formulation. Therefore, our aim of providing the discussions in Chapters 2 and 3 is to help readers identify strategies to come up with formulations amenable to the design of efficient solution approaches.

3.1 Decomposition Methods

As extensions of the classical facility location problems, the integrated location problems are typically NP-hard. For example, the [UFL] problem is a special case of the [SCD] formulation (2.9) with $\hat{K}_j = 0$ for all $j \in J$. Therefore, it is natural to attack integrated problems with similar solution approaches that exhibit proven success toward the classical problems. In particular, decompositions methods, such as Lagrangian relaxation and branch and price, are popular methods to develop efficient solution algorithms for classical methods, and have been widely

adopted to tackle integrated problems. In this section, we discuss these approaches using problem (2.9) as an example. The branch and price and Lagrangian relaxation algorithms for this problem are developed by Shen *et al.* (2003) and Daskin *et al.* (2002), respectively.

3.1.1 Branch and Price

The branch and price approach for the [SCD] problem is built on the observation that the problem can be reformulated as a general set covering problem (1.9). In particular, the set of elements to be covered is the set of customer locations I, and the collection of feasible subsets is given by $N = 2^I$. The cost associated with each subset $R \in N$ is given by the cost of serving the subset of customers with the facility that yields the lowest fixed location, shipping and inventory costs, that is, $c_R = \min_{j \in J} \hat{c}_{R,j}$ where

$$\hat{c}_{R,j} = f_j + \sum_{i \in R} \hat{d}_{ij} + \hat{K}_j \sqrt{\sum_{i \in R} \mu_i}.$$

Note that the set covering formulation (1.9) is a linear integer program. Thus, the reformulation gets rid of nonlinearity in the objective function (2.9), at the expense of introducing an exponential number of variables Z_R, as the set N is exponential in size. For integer programs with polynomial number of constraints and exponential number of variables, branch and price (e.g., Barnhart *et al.*, 1998) is often an effective solution strategy. This approach builds on the standard branch and bound algorithm by solving the continuous relaxation of the integer problem at each branch and bound node (which is a linear program with an exponential number of variables) using a column generation procedure, which we discuss below.

The column generation procedure tackles the continuous relaxation problem iteratively, first by including only a small subset of variables and then subsequently adding more. Let [SCR] denote the continuous relaxation of (1.9), obtained by replacing the constraints $Z_R \in \{0, 1\}$ with $0 \leq Z_R \leq 1$ for $R \in N$. We first begin with a manageable subset $N' \subset N$ and exclude all $R \in N \setminus N'$ from the formulation. The cost

parameters c_R are computed for each $R \in N'$. In the column generation procedure, we solve the following restricted formulation with a subset $N' \subseteq N$ of variables, referred to as the *master problem*, while updating the subset N' iteratively:

$$[\text{Master Problem}]: \quad \sum_{R \in N'} c_R Z_R \tag{3.1}$$

$$\text{s.t.} \quad \sum_{R \in N': i \in R} Z_R \geq 1 \text{ for } i \in I \tag{3.2}$$

$$Z_R \in \{0, 1\} \text{ for } R \in N. \tag{3.3}$$

Remark 3.1. To ensure feasibility of the master problem, the initial set N' must include elements that collectively include all $i \in I$. One simple way to guarantee so is by including all $|I|$ singleton subsets of I in N'.

Note that the optimal objective value of the master problem is an upper bound on that of [SCR], because the former is equivalent to the latter with the additional constraints $Z_R = 0$ for $R \in N \setminus N'$. Let the optimal dual solution values corresponding to (3.2) be denoted by π_i for each $i \in I$. Note that the optimal solution to the master problem can be viewed as a basic feasible solution to the original continuous relaxation problem [SCR], at which Z_R for $R \in N \setminus N'$ are not in the basis. To obtain a solution with lower cost, one may identify nonbasic variable(s) with negative reduced cost(s) to enter the basis. Note that such nonbasic variables must be among those Z_R with $R \in N \setminus N'$, because otherwise the current solution would not be optimal for the master problem. Therefore, if the reduced cost for any variable z_R, $R \in N \setminus N'$, is negative, then including said variable in the master problem can possibly improve the optimal solution. On the other hand, if all variables have nonnegative reduced costs, then the current solution is optimal to [SCR].

Each iteration of the column generation algorithm proceeds by identifying the variables (columns) with the lowest reduced costs, and adding the corresponding subsets into N' to expand the master problem. The reduced cost of the variable Z_R is given by $c_R - \sum_{i \in R} \pi_i$. The

problem of identifying the subsets R corresponding to the minimum reduced costs, referred to as the *pricing problem*, can be formulated as follows:

$$[\text{Pricing Problem}]: \quad \min_{R \subset I} \min_{j \in J} f_j + \sum_{i \in R} (\hat{d}_{ij} - \pi_i) + \hat{K}_j \sqrt{\sum_{i \in R} \mu_i}. \quad (3.4)$$

Note that the pricing problem is decomposable by J. That is, one can solve the problem

$$\min_{R \subset I} f_j + \sum_{i \in R} (\hat{d}_{ij} - \pi_i) + \hat{K}_j \sqrt{\sum_{i \in R} \mu_i} \quad (3.5)$$

for each $j \in J$ separately, and identify the j that yields the lowest objective value. If this minimum value is negative, the variable Z_R corresponding to the optimal subset R in (3.4) can be added to the master problem. Alternatively, because any column with negative reduced cost can potentially lead to cost improvements, one may add multiple Z_R variables corresponding to those R in (3.4) that yield negative reduced costs.

Next, we focus on solving (3.5) efficiently. We first provide an nonlinear integer programming reformulation as follows:

$$\min_{\mathbf{y} \in \{0,1\}^{|I|}} \sum_{i \in I} \hat{b}_i y_i + g(\sum_{i \in I} \hat{c}_i y_i) \quad (3.6)$$

where $\hat{b}_i = \hat{d}_{ij} - \pi_i$, $\hat{c}_i = \hat{K}_j^2 \mu_i$ and $g(x) = \sqrt{x}$. Note that choosing a set R in (3.5) is equivalent to setting $y_i = 1$ for $i \in R$ and $y_i = 0$ for $i \notin R$.

We note that problem (3.6) exhibits a knapsack-like trade-off. In particular, a negative value of $\hat{b}_i$ (the gain of including element i) is offset by the increase in cost due to the square root term (which can be interpreted as a soft knapsack capacity constraint). Similar to the (continuous relaxation of the) knapsack problem, there exists a simple sorting algorithm that solves (3.6), as characterized by the following results. To begin, we re-order the indices i in the set I by sorting its elements in increasing order of $\hat{b}_i/\hat{c}_i$, such that $\hat{b}_1/\hat{c}_1 \leq \cdots \leq \hat{b}_m/\hat{c}_m < 0 \leq \hat{b}_{m+1}/\hat{c}_{m+1} \leq \cdots \leq \hat{b}_{|I|}/\hat{c}_{|I|}$. Then, the following results (Ozsen *et al.*, 2008; Mak and Shen, 2009) hold.

Proposition 3.1. For any increasing function $g(\cdot)$, there exists an optimal solution to the continuous relaxation of (3.6), denoted by $(y_1^*, \cdots, y_{|I|}^*)$, that satisfies:

1. $y_i^* = 0$ for $i = m+1, \cdots, |I|$;
2. $0 < y_i^* < 1$ for at most one $i \in I$;
3. If $y_k^* > 0$ for some $1 \leq k \leq m$, then $y_i^* = 1$ for $1 \leq i \leq k-1$.

Proof. Part 1 follows directly from the fact that $g(\cdot)$ is increasing. Thus, increasing the value of y_i from 0 to $\epsilon > 0$ will increase both the $\sum_{i\in I} \hat{b}_i y_i$ and $g\left(\sum_{i\in I} \hat{c}_i y_i\right)$ terms, resulting in a worse objective value (in the minimization sense).

To prove Part 2, suppose to the contrary that $(y_1', \cdots, y_{|I|}')$ is an optimal solution where $0 < y_k' < 1$ and $0 < y_l' < 1$ where $k < l$. From Part 1, $1 \leq k < l \leq m$. Let z' denote the objective value associated with this solution. Then, one can define another solution, $(y_1'', \cdots, y_{|I|}'')$, as follows:

$$y_i'' = \begin{cases} y_i' & \text{if } i \neq l, k \\ y_k' + \epsilon & \text{if } i = l \\ Y_l' - \frac{\hat{c}_k}{\hat{c}_l}\epsilon & \text{if } i = l \end{cases} \tag{3.7}$$

where $\epsilon = \min\left\{1 - y_k', \frac{\hat{c}_l}{\hat{c}_k} y_l'\right\}$ which implies that $(y_1'', \cdots, y_{|I|}'')$ is feasible. Denote the objective value of the new solution by Z''. Then,

$$\begin{aligned} Z'' - Z' &= \epsilon\left(\hat{b}_k - \hat{b}_l \frac{\hat{c}_k}{\hat{c}_l}\right) + g\left(\sum_{i\in I} \hat{c}_i y_i' + \hat{c}_k \epsilon - \epsilon \hat{c}_l \frac{\hat{c}_k}{\hat{c}_l}\right) - g\left(\sum_{i\in I} \hat{c}_i y_i'\right) \\ &= \epsilon\left(\hat{b}_k - \hat{b}_l \frac{\hat{c}_k}{\hat{c}_l}\right) \\ &\leq \epsilon\left(\hat{b}_k - \frac{\hat{b}_k}{\hat{c}_k}\hat{c}_k\right) = 0. \end{aligned} \tag{3.8}$$

The inequality (3.8) holds because $k < l$, i.e., $\hat{b}_k/\hat{c}_k \leq \hat{b}_l/\hat{c}_l$. The above implies that Y_j'' is optimal. Furthermore, because $\epsilon = \min\left\{1 - y_k', \frac{\hat{c}_l}{\hat{c}_k} y_l'\right\}$, the following holds.

1. If $\epsilon = 1 - y_k'$, $y_k'' = 1, 0 < y_l'' < 1$
2. If $\epsilon = \frac{\hat{c}_l}{\hat{c}_k} y_l'$, $y_l'' = 0, 0 < y_k'' < 1$.

In both cases, the number of variables with strictly fractional values is reduced by one without increasing the objective value. Part 2 then follows from repeating the same argument until only one fractional value remains in the solution.

Part 3 can be proved by a similar contradiction argument similarly as Part 2. □

Note that Proposition 3.1 holds for any increasing function $g(\cdot)$, not only the square root form. Ozsen *et al.* (2008) and Ozsen *et al.* (2009) and Mak and Shen (2009) have applied variants of this result to solve subproblems with other nonlinear functions.

Furthermore, for cases where the function $g(\cdot)$ is increasing and concave, such as the case for the square root function, the following result holds:

Corollary 3.1. If the function $g(\cdot)$ is increasing and concave, there exists an optimal solution to the continuous relaxation of (3.6) at which y_i takes on integer values for all $i \in I$, and the continuous relaxation is tight.

Proof. This result follows from the fact that the continuous relaxation of (3.6) is a concave minimization problem over a polyhedron, which admits an optimal solution at a basic feasible solution. Note that the polyhedron $\{\mathbf{y}|y_i \geq 0, y_i \leq 1 \text{ for } i \in I\}$ is defined by $2|I|$ constraints over $|I|$ variables. Hence, at any basic feasible solution, $|I|$ of the $2|I|$ constraints hold at equality. Furthermore, for each i, at most one of the constraints $y_i \geq 0$ or $y_i \leq 1$, but not both, can hold at equality. Therefore, $|I|$ of the $2|I|$ constraints holding at equality implies that one of the two constraints corresponding to each i holds at equality, all all $i \in I$, i.e., y_i takes on integer values for all $i \in I$. □

Utilizing the above results, the pricing problem can be solved by enumerating m solutions, obtained by setting $y_1 = \cdots = y_k = 1$ and $y_{k+1}, \cdots y_{|I|} = 0$ for $k = 1, \cdots, m$:

Algorithm 1. The following algorithm solves the pricing problem (3.4), when $g(\cdot)$ is increasing and concave.

- *Step 0:* Initialize $m = 0$, $k = 1$, $L^* = 0$, $k^* = 0$.
- *Step 1:* Sort items in set I such that $\hat{b}_1/\hat{c}_1 \leq \hat{b}_2/\hat{c}_2 \leq \cdots \hat{b}_{|I|}/\hat{c}_{|I|}$. Set $m = \sup\{i|1 \leq i \leq |I|, \hat{b}_i < 0\}$. If $m = 0$, go to Step 4; otherwise, go to Step 2.
- *Step 2:* Compute $L(k) = \sum_{i=1}^{k} \hat{b}_i + g\left(\sum_{i=1}^{k} \hat{c}_i\right)$. If $L(k) < L$, set $L^* = L(k)$ and $k^* = k$.
- *Step 3:* If $k = m$, go to step 4; otherwise, increment $k \to k + 1$ and go to Step 2.
- *Step 4:* Return optimal solution $\mathbf{y}$ by setting $y_i = 1$ for $i = 1, \cdots, k^*$ and $y_i = 0$ for $i = k^* + 1, \cdots, |I|$, and objective value L^*.

The complexity of the algorithm is $O(|I| \log |I|)$, which is the complexity of the step of sorting elements in I in increasing order of $\hat{b}/\hat{c}$. Based on this subroutine to solve the pricing problem, one may then proceed with the standard branch-and-price algorithm (e.g., Barnhart *et al.*, 1998) to solve the set covering formulation of the [SCD] problem. In particular, one performs branch and bound by fixing $X_j = 1$ or 0 iteratively. At each node of the branch and bound tree, instead of solving the continuous relaxation of the original problem (as in the standard branch and bound algorithm), the continuous relaxation of the master problem (with some X_j's fixed) is solved by the column generation procedure described above.

3.1.2 Lagrangian Relaxation

A popular alternative decomposition method for solving large scale (linear or nonlinear) integer programming problems is Lagrangian relaxation. It is built on the theory of Lagrangian duality and offers, for problems with certain "block" structures, an effective means for computing upper and lower bounds on the optimal objective value. The general idea can be described qualitatively as follows. Suppose there is a large-scaled constrained optimization problem with a linked-block structure, i.e., decision variables and constraints can be separable except for a small number of linking constraints. Without the presence of such

linking constraints, the problems would be solvable for each block separately. Thus, a natural approximation approach is to relax these hard linking constraints to soft constraints, i.e., impose penalty for violations. The Lagrangian relaxation algorithm involves iteratively selecting the value of such penalty parameters (referred to as Lagrangian multipliers or Lagrangian dual variables associated with the linking constraints), while taking advantage of separability to solve the relaxed problem. For a more detailed tutorial on the Langrangian relaxation approach, the interested reader may refer to Fisher (1985), for example. For the [SCD] problem, Daskin *et al.* (2002) propose a Lagrangian relaxation algorithm that is closely related to the branch-and-price algorithm proposed by Shen *et al.* (2003).

The procedure begins by taking the Lagrangian dual of the [SCD] problem by relaxing constraints (2.5) and imposing corresponding Lagrangian multipliers π_i: $\max_{\boldsymbol{\pi}} L(\boldsymbol{\pi})$, where

$$
\begin{aligned}
L(\boldsymbol{\pi}) = \min \quad & \sum_{j\in J}\sum_{j\in J}\left[f_j X_j + \sum_{i\in I}\hat{d}_{ij}Y_{ij} + \hat{K}_j\sqrt{\sum_{i\in I}\mu_i Y_{ij}}\right] \\
& + \sum_{i\in I}\pi_i\left[1 - \sum_{j\in J}Y_{ij}\right] \\
= \min \quad & \sum_{j\in J}\sum_{j\in J}\left[f_j X_j + \sum_{i\in I}(\hat{d}_{ij} - \pi_i)Y_{ij} + \hat{K}_j\sqrt{\sum_{i\in I}\mu_i Y_{ij}}\right] \\
& + \sum_{i\in I}\pi_i \qquad (3.9) \\
\text{s.t.} \quad & Y_{ij} - X_j \le 0 \text{ for } i \in I, j \in J \\
& X_j \in \{0,1\} \text{ for } j \in J \\
& Y_{ij} \in \{0,1\} \text{ for } i \in I, j \in J.
\end{aligned}
$$

The Lagrangian dual can be interpreted as an approximation of the original problem obtained by relaxing a set of difficult constraints (2.5), without which the problem becomes easier to solve (in particular, decomposable by j), and replacing them with penalty imposed for violation. In this particular problem, (2.5) can be viewed as a set of *complicating constraints* because the problem exhibits a block structure, i.e., the

objective coefficients and remaining constraints forming independent blocks corresponding to each $j \in J$ once constraints (2.5) are removed. Therefore, these constraints are chosen to be relaxed, and the penalty terms $\sum_{i \in I} \pi_i \left[1 - \sum_{j \in J} Y_{ij}\right]$ are imposed in the objective function. It can be shown that, for any value of $\boldsymbol{\pi}$, the optimal objective value of the relaxed problem, $L(\boldsymbol{\pi})$, is a lower bound on that of the original [SCD] problem. Therefore, the best (tightest) lower bound can be obtained by maximizing the Lagrangian function over the penalty coefficients $\boldsymbol{\pi}$.

Three difficulties remain in developing an efficient algorithm for obtaining good solutions based on the Lagrangian dual problem. First, given any $\boldsymbol{\pi}$, an efficient subroutine for solving (3.9) is needed. Second, the values of $\boldsymbol{\pi}$ need to be optimized to obtain the tightest lower bound. Third, because solutions to the Lagrangian dual problem may not necessarily be feasible in the original problem due to relaxing (2.5), one needs to construct feasible solutions (upper bound solutions) based on information from the Lagrangian dual solution. We shall discuss each of these issues below.

To solve (3.9), one first observes that this problem is separable by $j \in J$, i.e., can be solved for each j separately, thanks to the relaxation of complicating constraints. Then, for each j, the subproblem becomes:

$$\begin{aligned} \min \quad & \sum_{j \in J} \left[f_j X_j + \sum_{i \in I} (\hat{d}_{ij} - \pi_i) Y_{ij} + \hat{K}_j \sqrt{\sum_{i \in I} \mu_i Y_{ij}} \right] \\ \text{s.t.} \quad & Y_{ij} - X_j \leq 0 \text{ for } i \in I \\ & X_j \in \{0,1\}, Y_{ij} \in \{0,1\} \text{ for } i \in I. \end{aligned} \tag{3.10}$$

To solve (3.10), one may compare the resulting objective values from fixing $X_j = 0$ and 1 and solving for the Y_{ij} variables optimally. In the former case, $Y_{ij} = 0$ for all $i \in I$, and the objective value is 0. In the latter case, observe that the subproblem of optimizing Y_{ij} is equivalent to the pricing problem in the column generation procedure, (3.4). Therefore, one can invoke the results of Proposition 3.1 and Corollary 3.1 to solve the subproblem using Algorithm 1. If the optimal objective value of (3.4) is smaller than $-f_j$, then it is optimal to set $X_j = 1$ (and Y_{ij} equal to the corresponding optimal values in (3.4)) in (3.10); otherwise, $X_j = Y_{ij} = 0$ for all $i \in I$ is the optimal solution.

Remark 3.2. It is not coincidence that both the column generation and Lagrangian relaxation procedures give rise to the same subproblem. In fact, these two methods are known to be equivalent and lead to the same decomposition reformulation as well as bounds (see, for example, Vanderbeck and Savelsbergh, 2006).

The next algorithmic issue is to optimize over the values of $\boldsymbol{\pi}$. We take note of the following:

Lemma 3.2. The Lagrangian function, $L(\boldsymbol{\pi})$, is concave and piecewise linear in $\boldsymbol{\pi}$.

Proof. For each given solution $(\mathbf{X}, \mathbf{Y})$, the objective value (3.9) is linear in $\boldsymbol{\pi}$. With $(\mathbf{X}, \mathbf{Y})$ restricted to binary values, there are a finite number of feasible solutions. Therefore, the value of $L(\boldsymbol{\pi})$ is given by the pointwise minimum of a finite number of linear functions, which yields a piecewise linear concave function. □

Lemma 3.2 implies that the Lagrangian dual is a concave-maximization problem. However, the objective function $L(\boldsymbol{\pi})$ is nondifferentiable. For nondifferentiable concave maximization (or convex minimization) problems, subgradient methods are often effective. These methods are extensions of gradient methods for differentiable convex minimization problems, based on the notion of subgradients, the generalization of gradients for nondifferentiable functions.

Definition 3.1. Let $\mathbf{C} \subseteq \mathbb{R}^n$ be a convex set and $f : \mathbf{C} \to \mathbb{R}$ be a concave function. A vector $\mathbf{g}$ is a subgradient of f at $\mathbf{x} \in \mathbf{C}$ if, for every $\mathbf{y} \in \mathbf{C}$, it satisfies:

$$f(\mathbf{x}) + \mathbf{g}'(\mathbf{y} - \mathbf{x}) \leq f(\mathbf{y}). \tag{3.11}$$

Geometrically, condition (3.11) states that the tangent line at $\mathbf{x}$ defined by the subgradient lies above the function $f(\cdot)$ over its domain. It is clear that the subgradient is a generalization of the gradient. In particular, where $f(\mathbf{x})$ is differentiable at $\mathbf{x}$, then the gradient is the unique subgradient at $\mathbf{x}$. In general, when $f(\mathbf{x})$ is nondifferentiable, the subgradient is not unique. The set of all subgradients of f at $\mathbf{x}$ is known

as the *subdifferential* of f, denoted by $\partial f(\mathbf{x})$. It is easy to check that the subdifferential of a concave function is a closed, convex set.

Based on the notion of subgradients, one can generalize the steepest descent (ascent) methods for smooth concave function maximization to nondifferentiable functions by replacing gradients with subgradients in identifying descent directions. For theoretical aspects of subgradient methods, we refer readers to discussions in textbooks such as Bazaraa *et al.* (2004) (Chapter 8.9). Here, we provide an application-driven outline of the procedure for maximizing $L(\boldsymbol{\pi})$.

We begin with the following result that identifies a subgradient for the function $L(\boldsymbol{\pi})$.

Lemma 3.3. $\mathbf{g} = [(1 - \sum_{j\in J} Y^*_{1j}), \cdots (1 - \sum_{j\in J} Y^*_{|I|j})]'$ is a subgredient at $\boldsymbol{\pi}$ for $L(\boldsymbol{\pi})$, where $(\mathbf{X}^*, \mathbf{Y}^*)$ is the solution to the inner problem (3.9) given $\boldsymbol{\pi}$.

Proof. By definition 3.1, we need to show that $\mathbf{g}$ satisfies condition (3.11) for any $\hat{\boldsymbol{\pi}} \in \mathbf{R}^{|I|}$.

$$\begin{aligned} L(\hat{\boldsymbol{\pi}}) &= \min \sum_{j\in J}\sum_{j\in J}\left[f_j X_j + \sum_{i\in I}\hat{d}_{ij}Y_{ij} + \hat{K}_j\sqrt{\sum_{i\in I}\mu_i Y_{ij}}\right] \\ &\quad + \sum_{i\in I}\hat{\pi}_i\left[1 - \sum_{j\in J} Y^*_{ij}\right] \\ &\leq \sum_{j\in J}\sum_{j\in J}\left[f_j X^*_j + \sum_{i\in I}\hat{d}_{ij}Y^*_{ij} + \hat{K}_j\sqrt{\sum_{i\in I}\mu_i Y^*_{ij}}\right] \\ &\quad + \sum_{i\in I}\hat{\pi}_i\left[1 - \sum_{j\in J} Y^*_{ij}\right] \\ &= L(\boldsymbol{\pi}) + \sum_{i\in I}(\hat{\pi}_i - \pi_i)\left[1 - \sum_{j\in J} Y^*_{ij}\right]. \end{aligned}$$

In the above, the inequality holds because $(\mathbf{X}^*, \mathbf{Y}^*)$ is a feasible, but not necessarily optimal, solution to the inner problem (3.9) given $\hat{\boldsymbol{\pi}}$ □

With the above result, we present the following subgradient algorithm.

Algorithm 2. The subgradient algorithm for maximizing $L(\boldsymbol{\pi})$ can be stated as follows:

- Step 0: Initialize $\boldsymbol{\pi}^1$ as any starting value, e.g., $\mathbf{0}$, $UB = \infty$, $LB = \infty$. Set iteration counter $n = 1$.
- Step 1: Solve (3.9) with $\boldsymbol{\pi} = \boldsymbol{\pi}^n$. Let $(\mathbf{X}^n, \mathbf{Y}^n)$ denote the optimal solution. If $L(\boldsymbol{\pi}^n) > LB$, update $LB = L(\boldsymbol{\pi}^n)$.
- Step 2: Construct feasible solution $(\hat{\mathbf{X}}^n, \hat{\mathbf{Y}}^n)$ by repairing $(\mathbf{X}^n, \mathbf{Y}^n)$ such that relaxed constraints (2.5) are satisfied. Compute objective value, denoted by z^n, corresponding to $(\hat{\mathbf{X}}^n, \hat{\mathbf{Y}}^n)$. If $z^n < UB$, update $UB = z^n$ and incumbent solution $(\hat{\mathbf{X}}, \hat{\mathbf{Y}}) = (\hat{\mathbf{X}}^n, \hat{\mathbf{Y}}^n)$. If $(UB - LB)/UB < \epsilon$ for pre-specified tolerance level ϵ, go to Step 4. Otherwise, go to Step 3.
- Step 3: Compute subgradient $\mathbf{g} = [(1 - \sum_{j \in J} Y^*_{1j}), \cdots, (1 - \sum_{j \in J} Y^*_{|I|j})]'$. Update $\boldsymbol{\pi}_{n+1} = \boldsymbol{\pi}_n + \delta^n \mathbf{g}/||g||_2$. Increment $n \rightarrow n+1$. Go to Step 1.
- Step 4: Terminate algorithm and return incumbent solution $(\hat{\mathbf{X}}^n, \hat{\mathbf{Y}}^n)$.

To implement Algorithm 2, a few details remain to be filled in. First, one need to determine step sizes δ^n for updating $\boldsymbol{\pi}$. Theoretically (Bazaraa *et al.*, 2004, Theorem 8.9.2), the algorithm is guaranteed to converge to the global optimal solution for step sizes satisfying $\{\delta^n\} \rightarrow 0^+$ and $\sum_{n=0}^{\infty} \delta^n = \infty$. However, not all step size sequences satisfying these two conditions work efficiently. For example, the step size sequence $\delta^n = 1/n$ is known to lead to slow convergence in practice.

An alternative is to use $\delta^n = \lambda^n(L^* - L(\boldsymbol{\pi}^n))$, where L^* is the optimal value $L(\boldsymbol{\pi}^*)$. However, because this optimal value is not known *a priori* in practice, a practical choice is to replace it with some upper bound $\bar{L}$. In Step 3, $\bar{L}$ can be chosen to be UB, the best upper bound identified so far. Furthermore, $\{\lambda^n\}$ is a decreasing sequence that approaches zero. An approach that typically works efficiently in practice is to begin with $\lambda^1 = 2$ and, if LB has not been updated in Step 1 for a certain number of iterations (e.g., 20), set $\lambda^{n+1} = \lambda^n/2$. This procedure allows for

refining the search region once the algorithm has not identified improved solutions (in $\boldsymbol{\pi}$) for a certain number of iterations. Furthermore, the termination condition $\lambda^n < \hat{\epsilon}$ for some tolerance $\hat{\epsilon}$ can be included. In this algorithm, one point to caution about is that, if UB is too far from the true optimal value, the algorithm is not guaranteed to converge. In case of such difficulties, an alternative (albeit more complex) step size rule that guarantees convergence without requiring knowledge of the optimal value is the variable target method proposed by Sherali *et al.* (2000).

Next, we discuss the issue of how to identify good upper bounds for updating subgradients and associated feasible solutions to the original problem. Recall that, at iteration n, the solution $(\mathbf{X}^n, \mathbf{Y}^n)$ to the relaxed problem may not satisfy constraints (2.5). In particular, it is possible that for some $i \in I$, either $\sum_{j \in J} Y_{ij}^n = 0$ or ≥ 2. To repair this solution, we consider each $i \in I$ one by one, and compute the incremental cost (based on the assignments so far) of assigning i to each DC that is open in the current solution (i.e., $X_j^n = 1$). The assignment with the lowest incremental cost is chosen. Note that this greedy procedure does not necessarily identify the optimal assignment given the set of DCs to be opened. Therefore, one can also consider improvements via exchange heuristics. See Daskin *et al.* (2002) for details.

Another subroutine that helps improve efficiency of the algorithm is variable fixing. In Step 1, let v_j^n denote the optimal value in subproblem (3.4) for the current iteration.

Proposition 3.2. The following variable fixing rules hold:

- (Node exclusion rule:) If $X_j^n = 0$ and $LB + f_j + v_j^n > UB$, then $X_j = 0$ at the optimal solution and this can be imposed as a constraint in subsequent iterations;
- (Node inclusion rule:) If $X_j^n = 1$ and $LB - (f_j + v_j^n) > UB$, then $X_j = 1$ at the optimal solution and this can be imposed as a constraint in subsequent iterations.

Proof. We first prove the node exclusion rule. If $X_j = 1$ at the optimal solution, the constraint $X_j = 1$ can be added to the original problem

without affecting the optimal objective value, i.e., UB remains a valid upper bound. However, with this additional constraint, the resulting lower bound at iteration n would become $LB + f_j + v_j^n$ instead of LB (because DC j must be selected due to the additional constraint), which contradicts the validity of UB as an upper bound. Therefore, X_j must equal 0 at the optimal solution.

The node inclusion rule can be proved similarly. □

Finally, we note that strong duality does not generally hold for the Lagrangian dual. Therefore, the best upper and lower bounds (UB and LB) obtained from Algorithm 2 need not be equal. To close the gap, one may embed the algorithm in a branch and bound procedure (Daskin *et al.*, 2002).

It is notable that the efficacy of decomposition approaches depends heavily upon the mathematical structure of the optimization formulations. Thus, ideally, the modeler should keep in mind the potential implications on solution efficiency when formulating the mathematical model from the outset. This is also true for other solution approaches as well, including the conic programming method that we shall review in the next section.

3.2 Conic Programming

Conic programming is an important branch of convex optimization. A wide array of problems with applications in operations research as well as various engineering disciplines can be modeled with special classes of conic programs, such as second-order cone programs (SOCPs) (see, e.g., Boyd and Vandenberghe, 2009). SOCPs are a generalization of convex quadratic programs and linear programs (LPs) and can be very efficiently solved with interior point algorithms. In this section, we use the [SCD] problem as an illustration of how SOCP techniques can help develop efficient solution approaches for integrated facility location models.

We first define SOCP problems as follows (e.g., Lobo *et al.*, 1998).

Definition 3.2. An SOCP is an optimization problem in the form:

$$\begin{aligned} \min \quad & \mathbf{f}^T\mathbf{x} \\ \text{s.t.} \quad & \|\mathbf{A}_i\mathbf{x} + \mathbf{b}_i\| \leq \mathbf{c}_i^T\mathbf{x} + d_i \text{ for } i = 1, \cdots, N, \end{aligned}$$

where $\mathbf{x} \in \mathbb{R}^n$ denotes the vector of decision variables, $f \in \mathbb{R}^n, \mathbf{A}_i \in \mathbb{R}^{(n_i-1)\times n}, \mathbf{b}_i \in \mathbb{R}^{n_i-1}, \mathbf{c}_i \in \mathbb{R}^n, d_i \in \mathbb{R}$, and $\|\cdot\|$ denotes the Euclidean norm.

Consider formulation (2.4), in which the objective contains nonlinear (square root) terms, while all constraints are linear. We shall show that this nonlinear formulation can be reexpressed with SOCP constraints. First, note that one can replace the objective with:

$$\min \sum_{j\in J} \left[f_j X_j + \sum_{i\in I} \hat{d}_{ij} Y_{ij} + K_j U_j + q V_j \right], \tag{3.12}$$

and add the following constraints:

$$\sqrt{\sum_{i\in I} \mu_i Y_{ij}} \leq U_j \text{ for } j \in J \tag{3.13}$$

$$\sqrt{\sum_{i\in I} \sigma_i^2 Y_{ij}} \leq V_j \text{ for } j \in J. \tag{3.14}$$

Next, we recall that Y_{ij} is constrained to take binary values. This implies that, in any feasible solution, $Y_{ij} = Y_{ij}^2$. Therefore, constraints (3.13, 3.14) are equivalent to:

$$\sqrt{\sum_{i\in I} \mu_i Y_{ij}^2} \leq U_j \text{ for } j \in J \tag{3.15}$$

$$\sqrt{\sum_{i\in I} \sigma_i^2 Y_{ij}^2} \leq V_j \text{ for } j \in J. \tag{3.16}$$

Note that, because $\mu_i > 0$ and $\sigma_i^2 \geq 0$, the left hand side terms in (3.15, 3.16) are Euclidean norms, and therefore these constraints are in SOCP form. Then, adding constraints (2.5-2.8), we obtain a mixed integer second-order cone program (MISOCP). This class of problems can be solved optimally with branch and bound routines, in which continuous relaxations solved with interior point algorithms. Practical

implementations of such routines are available in commercial solver packages such as CPLEX.

Atamtürk *et al.* (2012) use a similar transformation to formulate a supply chain design model with disruption risk considerations as a MISOCP. To speed up computations, they adopt a class of extended polymatroid inequalities that are valid for constraints in the form (3.15, 3.16) in a branch-and-cut procedure. Extended polymatroid inequalities are closely related with the concepts of submodular functions and extended polymatroids. We begin the discussion with the following definitions.

Definition 3.3. A set function $g : 2^I \rightarrow \mathbb{R}$ is submodular if, for all $S_1, S_1 \subseteq I$, $g(S_1 \cup S_2) + g(S_1 \cap S_2) \leq g(S_1) + g(S_2)$.
A set $EP_g \in \mathbb{R}^I$ is an extended polymatroid associated with g, for a submodular function g, if $EP_g = \{\mathbf{x} \in \mathbb{R}^I \,|\, \sum_{i \in I} x_i \leq g(S) \text{ for all } S \subseteq I\}$.

Atamtürk and Narayanan (2008) prove the following proposition regarding a class of valid inequalities associated with extended polymatroids.

Proposition 3.3. [Proposition 1 of Atamtürk and Narayanan (2008)] For the lower convex envelope of a submodular function g, given by $conv\{(\mathbf{y}, t) \in \{0,1\}^I \times \mathbb{R}, g(\mathbf{y}) \leq t\}$, an inequality in the form $\sum_{i \in I} \pi_i y_i \leq t$ if and only if $\pi \in EP_g$.

These inequalities are known as *extended polymatroid inequalities*. We then discuss how this result can be applied to the case of constraints (3.15). The case for (3.16) is analogous. First, we define $g(\mathbf{Y}_j) = \sqrt{\sum_{i \in I} \mu_i Y_{ij}^2}$. It can be shown that $g(\cdot)$ is a submodular function. Then, Proposition 3.3 suggests that the extended polymatroid inequalities $\sum_{i \in I} \pi_i y_i \leq t$ for $\pi \in EP_g$ are valid for the lower convex envelope of $g(\cdot)$, and thus for the MISOCP formulation of the $[SCD]$ problem. However, as the set EP_g is itself a polyhedron, it is impossible to enumerate all possible π values. Therefore, these inequalities are added to the formulation iteratively with a cut generation procedure. In particular, in each iteration of the branch-and-bound procedure where

a continuous relaxation of the MISOCP formulation is solved, one can identify a violated extended polymatroid inequality, or prove that no such inequality is violated, by solving a separation problem as stated below (Atamtürk *et al.*, 2012).

Proposition 3.4. For given $(\mathbf{Y}_j^*, U^*)$, let ζ^* and π^* be the optimal objective value and solution to the separation problem $\max_{\pi \in EP_g} \sum_{i \in I} \pi_i Y_{ij}^*$, respectively. If $\zeta^* > U^*$, then the extended polymatroid inequality $\sum_{i \in I} \pi_i^* Y_{ij} \leq U$ cuts off $(\mathbf{Y}_j^*, U^*)$; otherwise, all extended polymatroid inequalities are satisfied at $(\mathbf{Y}_j^*, U^*)$.

Note that the separation problem is equivalent to maximizing a linear function over an extended polymatroid. This can be efficiently solved with the greedy algorithm (Edmonds, 1970). Details of this implementation can be found in Appendix B of Atamtürk *et al.* (2012).

Both the decomposition methods and conic programming methods discussed in Sections 3.1 and 3.2 aim at providing computationally efficient routines to obtain numerical solutions to problem instances. The technique to be discussed in the next section, on the other hand, provides an important tool for simplifying and obtaining tractable approximations to analytical models.

3.3 Dimensional Analysis

To complement the techniques presented in the previous sections for solving computational facility location models, we further discuss a powerful technique for analyzing continuous approximation models (discussed in Section 2.3), known as dimensional analysis. This technique is often employed in the development and selection of models in physical sciences and statistics. In particular, when developing possible models of relationships between physical quantities (variables and parameters), dimensional analysis helps eliminate models (relationships) that would be violated when variables and parameters are re-scaled.

3.3.1 π-Theorem and EOQ Example

The core idea of dimensional analysis is that physically meaningful relationships must be invariant to rescaling of parameter dimensions. For example, if a certain physical relationship holds regarding the density of DCs and the distribution cost of the supply chain network, then the same relationship should remain to hold if one redefines distance in kilometers instead of in miles. This idea is formalized in Buckingham's celebrated π-Theorem (e.g., Bridgman, 1922):

Theorem 3.4. Consider a problem that can be described in (dependent and independent) variables $q_1, \cdots, q_n$, where a dimensionally homogeneous relationship in the form $f(q_1, \cdots, q_n) \equiv 0$ holds. Then, the equation can be restated as $F(\pi_1, \cdots, \pi_p)$, where $\pi_k, k = 1, \cdots, p$ are dimensionless parameters in the form $\pi_k = q_1^{a_{1k}} q_2^{a_{2k}} \cdots q_n^{a_{nk}}$.

The π-Theorem formalizes the rescaling idea, i.e., any physically meaningful relationship between parameters of arbitrary dimensions can be reexpressed as one between dimensionless parameters constructed from the original parameters that is invariant under scaling. For more discussion of the π-Theorem (in the context of fluid mechanics), one may refer to Sonin (2001). The key to applying this theorem is how to identify the dimensionless parameters, also known as π-groups. To this end, consider a problem with n parameters defined over k dimensions. As an illustration, we consider the economic ordering quantity (EOQ) model, in which we want to identify the relationship between four parameters: the demand rate D (in items/time or items$^1 \times$ time^{-1}), holding cost rate h (in dollars$^1 \times$ item$^{-1} \times$ time^{-1}), fixed replenishment cost K (in dollars^1) and the optimal (replenishment and holding) cost rate C (in dollars$^1 \times$ time^{-1}). These four parameters are defined over three dimensions: items, time and dollars. Therefore, in this example, $n = 4$ and $k = 3$. It is well known that, when the replenishment cycles are optimized, we have $C = \sqrt{2KDh}$. What we shall show below is that, without going through the analysis of inventory dynamics or even defining the objective function of the EOQ problem, we can already derive the structure of this relationship through dimensional analysis.

To summarize the relationships between parameters with their respective dimensions, we consider the k by n matrix $\mathbf{M}$, referred to as the *dimensional matrix*, whose rows correspond to dimensions and columns correspond to parameters. Its (i, j)-th component is the power of the i-th dimension in the j-th parameter. For example, in the EOQ case, we have

$$\mathbf{M} = \begin{bmatrix} 1 & -1 & 0 & 0 \\ -1 & -1 & 0 & -1 \\ 0 & 1 & 1 & 1 \end{bmatrix}.$$

Consider the first column, which corresponds to the demand rate parameter. Because demand rate has dimensions of items$^1 \times$ time^{-1}, the entry in the first row (corresponding to items) is 1, the entry in the second row (time) is -1, and the entry in the third row (dollars) is 0. Similarly, the second column corresponding to holding cost rate, which has dimensions dollars$^1 \times$ item$^{-1} \times$ time^{-1}, has entries of -1 in the first and second rows (items and time) and 1 in the third (dollars). For the third column, which corresponds to the replenishment cost K (in dollars^1), only the third row entry (dollars) is 1 and others are 0. Finally, the fourth column, which corresponds to the optimal cost rate C (dollars$^1 \times$ time^{-1}), has entries of 0 for the first row (items), -1 for the second row (time) and 1 for the third (dollars).

The significance of the dimensional matrix is that the π-groups can be identified by solving the equation

$$\mathbf{M} \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} = \mathbf{0}. \tag{3.17}$$

That is, they correspond to vectors in the null space of $\mathbf{M}$. This is because, by definition, π-groups are dimensionless and have no units. In order to construct such a group, the original problem parameters should be multiplied in a way that their units cancel out. For the EOQ example, suppose vector $[a_1, a_2, a_3, a_4]^T$ satisfies (3.17). By the construction of the matrix $\mathbf{M}$, (3.17) implies that the parameter group $D^{a_1} h^{a_2} K^{a_3} C^{a_4}$ is dimensionless (i.e., has dimensions dollars$^0 \times$ item$^0 \times$ time^0), that is, their units cancel out each other. Thus, the dimensionless π-groups

can be obtained by solving the equation (3.17). From the rank-nullity theorem, the following holds:

Theorem 3.5. The number of multiplicatively-independent π-groups is given by $n - rank(\mathbf{M})$.

For the EOQ example, $rank(M) = 3$, and thus there can be only one independent π-group, corresponding to $(a_1, a_2, a_3, a_4) = (-1, -1, -1, 2)$. Therefore, π-group is given by $\pi_1 = D^{-1}h^{-1}K^{-1}C^2 = C^2/(KDh)$. Then, by the π-theorem, the relationship between the four parameters must hold in the form of $f(C^2/(KDh)) \equiv 0$, i.e., $C^2/(KDh) \equiv \gamma$ or $C \equiv \sqrt{\gamma KDh}$ for some (dimensionless) constant γ. This is consistent with the classical result of the EOQ model that $C \equiv \sqrt{2KDh}$ under the optimal solution. Furthermore, recall that we did not make use of any knowledge of the EOQ problem itself or any relationships between the parameters, beyond their units. In fact, the analysis does not even rely on the definition of C being the cost rate under the optimal replenishment policy (or any definition of the objective function). Hence, any replenishment policy that depends only on the other three parameters would yield an average cost rate in the same form, with possibly different values of γ.

One can also perform a similar analysis to uncover the relationship between the three input parameters D, h, K and the optimal order quantity Q (in items). Replacing C with Q, we update the fourth column of the $\mathbf{M}$ matrix to obtain:

$$\mathbf{M} = \begin{bmatrix} 1 & -1 & 0 & 1 \\ -1 & -1 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{bmatrix}.$$

Again, $rank(M) = 3$ and there is only one independent π-group, which corresponds to $(a_1, a_2, a_3, a_4) = (-1, 1, -1, 2)$. Hence, $\pi_1 = hQ^2/(KD)$. Then, the π-theorem implies that $hQ^2/(KD) \equiv \gamma$ and $Q = \sqrt{\gamma KD/h}$ for some constant γ. Again, without considering the explicit cost structure of the EOQ model, we are able to identify the relationship between the optimal order quantity and the other input parameters. This relationship is again consistent with the optimal solution of the EOQ model, in which $\gamma = 2$. Furthermore, as is the case for the optimal cost

rate C, this dimensional analysis does not make use of the assumption that Q is the optimal replenishment quantity. Therefore, any consistent ordering policy that depends only on D, h and K would yield (average) order quantities proportional to $\sqrt{KD/h}$.

The EOQ example discussed above illustrates the potential power of dimensional analysis. For concisely defined problems with small number of parameters, the number of π-groups to consider is small, following Theorem 3.5. Then, it is possible to identify relationships between input parameters and parameters of interest (such as order quantity and cost rate in the EOQ example) that greatly enhance tractability of the problem. In Section 3.3.2, we discuss how this idea can be utilized to model routing costs for inventory replenishments at DCs.

3.3.2 Example: Retail Store Density Model

In this section, we illustrate the technique of dimensional analysis using a retail network design model adopted from Cachon (2014). In this work, the author studies the interactions between carbon emissions of transportation and the density of retail stores. In particular, operating the retail network incurs two types of travel and their associated carbon emissions: customers' travel between home and store, and truck travel to replenish the stores' inventories. The underlying trade-off of interest is one of choosing to operate a dense network of stores, which reduces customer travel at the expense of additional truck travel, versus a sparse network, which saves on truck travel while increasing customer travel distances. To capture this trade-off, Cachon (2014) proposes a centralized model that determines the optimal store density that minimizes the sum of the two travel costs, in addition to the facility cost. Based on this model, he investigates the impact of various emission-curbing measures on the optimal trade-off.

Consider a region of area a in which n stores are located with a fixed density (e.g., located at grid points). Customers are uniformly spaced within the region and the population size is normalized to one. Let $d_c(n)$ and $d_t(n)$ be the aggregate customer and truck travel distances, and c_c and c_t be the associated unit costs (which includes fuel costs and possibly carbon taxes), respectively. Furthermore, let

$t_s(n)$ be the store space required and c_s be the facility cost per unit store space. Then, the problem of optimizing the density (or, for a fixed area, the number) of stores can be formulated as $\min_n Z(n)$, where $Z(n) = c_s t_s(n) + c_c d_c(n) + c_t d_t(n)$.

We assume that aggregate customer demand per unit area in the region follows a normal distribution with mean λ and variance σ^2. For tractability, Cachon (2014) proposes an approximation for $t_s(n)$ by assuming that inventory is controlled using a carefully selected base stock policy such that stock-outs (backorders) are rare and replenishments typically utilize a full truck (approximately). In particular, under such assumptions, the expected inventory level at a facility at the end of a period is approximately given by $z\hat{\sigma}$, where z is a safety stock factor and $\hat{\sigma}$ is the standard deviation of demand handled by each facility, that is, the expected inventory level is approximately equal to the safety stock level. This is because the expected replenishment quantity is approximately equal to the expected demand per replenishment cycle. Because a facility covers a/n units of area, $\hat{\sigma} = \sqrt{a/n}\sigma$. Therefore, for n facilities, the total store space required is $\phi_s\sqrt{an}$, where $\phi_s = z\sigma$.

For the transportation cost terms, Cachon (2014) shows that there exists constants ϕ_c and ϕ_t such that the following hold, by using geometrical derivations:

$$d_c(n) = \phi_c\sqrt{a/n} \tag{3.18}$$

$$d_t(n) = \phi_t\sqrt{an}. \tag{3.19}$$

We illustrate alternative derivations of the same results using dimensional analysis. We first focus on $d_c(n)$, the aggregate customer travel distance. With n stores located, to minimize customer travel distance, it is obvious that each customer will patronize the respective nearest store. Effectively, the region will be partitioned into n subregions, each covering the customer locations from which a store is the closest out of all n options. Such a partitioning scheme is known as a Voronoi diagram, which can be efficiently computed given the geometry of the region. However, to obtain tractable expressions for the travel distance, we further make the assumptions that the region is of a regular shape and is large enough. These assumptions imply that the subregions will

be regular polygons (e.g., squares, triangles, hexagons), each of size a/n, centered on the stores. Then, $d_c(n)$ is given by the average distance from any point in a regular polygon (customer location) to its center (the store).

We follow the steps for dimensional analysis illustrated in Section 3.3.1. Let $\hat{a} = a/n$, the area of the regular polygon in question. The relationship between the area of the polygon and the average distance can be described in the dimension of miles, and two parameters, $\hat{a}$ and $d_c(n)$, with dimensions of miles^2 and miles, respectively. Therefore, one can consider the dimensional matrix $\mathbf{M} = [2 \quad 1]$. As $rank(\mathbf{M}) = 1$, we can form only one independent dimensionless group, i.e., $\hat{a}/(d_c(n))^2$. Then, the π-Theorem implies that there exists constant γ such that $\hat{a}/(d_c(n))^2 \equiv \gamma$, or $d_c(n) \equiv \sqrt{\hat{a}/\gamma}$. Substituting $\hat{a} = a/n$ and defining $\phi_c = 1/\sqrt{\gamma}$, we obtain (3.18).

Next, we consider the truck travel distances. Cachon (2014) assumes that the stores are replenished by a single truck that visits the n stores in one tour, and thus considers d_t to be the optimal length of a traveling salesman tour. We note that, for the case where d_t is given by the optimal travel distance under a vehicle routing problem (i.e., trucks have limited capacity and cannot visit all n stores in one tour), one can utilize the results of, for example, Daganzo (1984) to obtain a similar formula as (3.19). We will discuss how to obtain an approximate formula for the optimal traveling salesman tour based on dimensional analysis (Daganzo, 2005).

We first assume that the region in question is square-shaped (which can be relaxed). The traveling salesman problem can be characterized by one dimension, miles, and three parameters: a (in miles^2), n (dimensionless) and L^* (in miles), defined as the optimal traveling salesman tour connecting the n stores. The dimensional matrix can be written as $\mathbf{M} = [2 \quad 0 \quad 1]^T$. The rank of the matrix is one, and thus we obtain two π-groups: n and $d_t/\sqrt{a}$. Theorem 3.4 then stipulates that $L^*/\sqrt{a} \equiv f(n)$, or equivalently, $d_t \equiv \sqrt{a} f(n)$, with some function $f(\cdot)$ to be identified.

The combinatorial nature of the traveling salesman problem makes the identification of $f(\cdot)$ challenging. Interestingly, for the asymptotic case where n is very large, one can obtain structural results that yield a tractable formulation.

Proposition 3.5. For any positive and even integer m, $f(n) \leq 4m + mf(n/m^2)$.

Proof. For any given instance of store locations, one can form a feasible (but suboptimal) traveling salesman tour for the entire region by the following heuristic: (1) for some even integer $m \geq 2$, partition the original square-shaped region evenly into $m \times m$ square subregions; (2) solve for the optimal traveling saleman tour for points within each subregion; and (3) join the subtours obtained in each subregion to form one single tour that visits all stores. Note that step (3) can be done as follows. First, form pairs of adjacent subregions. Then, for each pair of adjacent subregions, remove one link each from the two subtours and add two links to join the corresponding stores such that the subtours are joined. Note that, as m is an even number, one can form a sequence of adjacent pairs such that repeating this procedure will yield one complete tour.

Because the region is spatially homogeneous, each of the subregions, with area a/m^2, have an average number of stores of n/m^2. Therefore, the expected length of each subtour in Step (2) will be given by $\frac{\sqrt{a}}{m} f\left(\frac{n}{m^2}\right)$. Further, the expected extra distance incurred by each extra link added in Step (3), which is bounded above by the distance between two randomly chosen points in two adjacent subregions, cannot exceed two times the length of each subregion, i.e., $2\sqrt{a}/m$. Note also that $2m^2$ such links are added in total. Thus, combining Steps (1-3), the above heuristic yields a feasible tour with expected length not exceeding $4m\sqrt{a} + m\sqrt{a} f\left(\frac{n}{m^2}\right)$. Since this provides an upper bound on the optimal length $d_t/\sqrt{a} \equiv f(n)$, we may obtain the desired result by rearranging terms. □

Proposition 3.5 shows that $f(\cdot)$ is bounded above by $mf(n/m^2)$ plus a term in $O(m)$. Using a similar argument, one can also obtain a lower bound in the form of $mf(n/m^2)$ minus a term in $O(m)$. Asymptotically, as N goes to infinity, the $O(m)$ terms are dominated, and we observe that

$$\lim_{n\to\infty} \frac{f(n)}{mf(n/m^2)} = 1.$$

Because this holds true for all m, $f(n) = O(\sqrt{n})$. One can then use the approximation $f(n) \approx k\sqrt{n}$, or equivalently, $d_t = k\sqrt{an}$, for sufficiently large n, where k is a constant that depends on the metric. Daganzo (2005) further discusses that the same argument holds for regions of non-square shapes, by partitioning the (possibly) irregular region into (approximately) square subregions, and shows that the constant k does not depend on the shape of the region. Using this approximation, we can then approximate the truck travel distance by $\phi_t\sqrt{an}$ in (3.19) by letting $\phi_t = k$.

Combining the cost terms, the overall cost function is given by $C(n) = \phi_c c_c\sqrt{an} + (\phi_t c_t + \phi_s c_s)\sqrt{an}$, and the optimal density of stores (i.e., stores per unit area) to operate is given by

$$n^*/a = \frac{\phi_c c_c}{\phi_t c_t + \phi_s c_s}.$$

Cachon (2014) considers that the cost parameters c_t, c_c and c_s to consist of both operating cost and emission cost components. Then, using this model, he investigates the effect of different compositions of the two types of costs on the optimal store density. For example, if emission costs are very high (e.g., due to a high carbon tax), intuition would suggest the operation of a denser network of stores, because trucks are more fuel-efficient than cars, and increasing store density causes substitution of car travel with additional truck travel. However, the model shows that this is not always the case, because such substitution is not a linear effect. In particular, one mile of truck travel displaces a smaller and a smaller car travel distance as store density increases, as customer (home-to-store straight line) travel distances decrease faster than truck travel (traveling salesman) distances as store density increases. Furthermore, Cachon (2014) calibrates the model to parameter inputs based on actual operating costs and emission values of vehicle models in the market and identify several key insights. In particular, attempting to optimize supply chain designs based on cost metrics solely without incorporating emission considerations may substantially increase emissions (compared with the minimum-emissions scenario). Furthermore, carbon tax and improvements in truck fuel efficiency are less effective measures in reducing overall emissions than the improvement in fuel efficiency of cars.

3.4 Discussion

For many facility location models, particularly ones in the form of mixed integer nonlinear programs, decomposition methods help exploit special structures (e.g., concave or submodular objective functions) by breaking down the problem into smaller pieces, which helps circumvent confounding factors such as linking constraints. The efficiency of such approaches is very problem specific, i.e., they work exceptionally well for certain problems but not for some others. In contrast, for the class of problems that can be formulated in the conic form, conic programming approaches make use of general purpose solvers that work for any problem in the class. As general-purpose solution algorithms for conic programs continue to advance, the computational efficiency of the conic approach will further improve. Thus, as a rough rule of thumb, where there are alternative formulations (approximations) for the same problem, ones in the conic form are often preferable. On the other hand, decomposition methods could be effective for other problems that possess special structures (e.g., general concave increasing objective functions) that cannot be captured in the conic form.

Dimensional analysis, while fairly commonly employed in certain engineering disciplines, is relatively little known within the operations management community from our observation. For compactly formulated analytical models, this technique allows one to greatly focus the class of functional forms to consider. We believe that this technique is not only effective in developing analytical continuous approximation models, but also promising in capturing operational characteristics for computational models (in the formulation of nonlinear objective terms or constraints) as well.

4

Applications in Supply Chain Settings

In Chapters 2 and 3, we have outlined a number of popular modeling and solution techniques employed in the study of integrated facility location problems in the literature. In this chapter and the next, we further discuss examples of applications of these techniques in different problem contexts. A rich literature covers applications in the supply chain domain, in which problems of designing networks of warehouses and DCs are tackled. It is notable that, while the business planning context is common, the models developed to account for different problem characteristics, such as risk, capacity and multiple-commodity considerations, can be substantially different. As problem complexity grows with the richness of features incorporated in the model, the modeler should carefully select the most imperative features to the specific application while maintaining problem tractability.

4.1 Capacitated Distribution Center Location for Traditional Supply Chains

The supply chain network design model [SCD] discussed in Chapter 2 is applibe maidcable in various distribution network planning settings. However, the model is *uncapacitated*, i.e., it does not take into account

the potential limitations in terms of physical space or material handling volume of candidate facility locations. Often, although a hard capacity constraint is not present, locating a facility that handles larger demand volume would entail higher investment costs (e.g., due to additional equipment). Modeling-wise, varying investment requirements associated with different capacity levels can be reflected by defining multiple entities in the candidate location set J corresponding to the same physical site, each having a different capacity limit and a different fixed location cost f_j.

Recall that the [SCD] model is an extension of the classical [UFL] model. In Section 1, we discussed that the conventional approach to incorporate facility capacity is to consider the capacitated extension [CFL] by adding constraint (1.6). Note that this constraint imposes a limit on the total demand volume that can be assigned to a facility, and thus reflects capacity at a strategic planning perspective, i.e., that customers must be re-assigned to other facilities whenever the capacity limit is to be exceeded. Under the integrated modeling approach, one can consider an additional, more flexible treatment of capacity considerations at the tactical level by adjusting the inventory control policy. In particular, when the capacity constraint arises from physical space limit, one can attempt to accommodate demand from more demand sources by operating under smaller replenishment lot sizes to reduce peak inventory levels, instead of opening additional facilities. This possibility is studied by Ozsen *et al.* (2008) and Ozsen *et al.* (2009). We discuss this modeling extension in this section below.

4.1.1 Tactical Modeling of Facility Capacity

Our discussion builds on the same notation defined in Chapter 2 for the discussion of the [SCD] model. Recall that, in the [SCD] model, DCs are considered to replenish inventory under continuous review (r, Q) policies. In the classical (r, Q) model, the reorder point r is given by the sum of the safety stock and the expected demand during lead time. Furthermore, note that the maximum possible inventory level carried by the DC is reached in the scenario that no demand is realized during the lead time (see Figure 4.1), i.e., the maximum level is given by

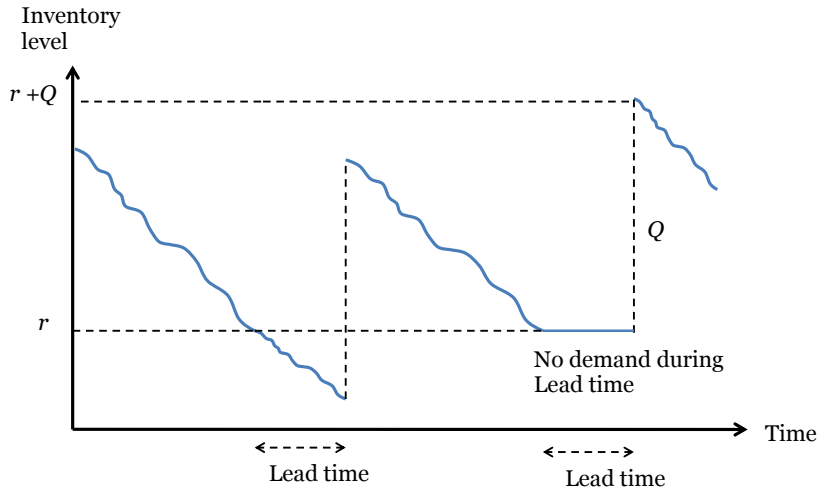


Figure 4.1: Maximum Inventory Level under Continuous Review (r, Q) Policy

$r + Q$. Therefore, for a facility j ordering a lot size Q_j per replenishment cycle subject to a storage space capacity of C_j, one can formulate the following capacity constraint:

$$Q_j + \sum_{i \in I} L\mu_i Y_{ij} + z_\alpha \sqrt{\sum_{i \in I} L\sigma_i^2 Y_{ij}} \leq C_j, \tag{4.1}$$

where, as defined in Section 2.1, L denotes the replenishment lead time, μ_i and σ_i denote the mean and standard deviation of demand at location i, and Y_{ij} is the binary decision variable indicating whether demand at customer location i is served by the facility at location j.

In (4.1), the second and third terms on the left hand side are the expected lead time demand and safety stock levels, respectively. Adding this constraint (for all $j \in J$) to [SCD] may entail additional computational complexity, as it involves nonlinear terms. Furthermore, as Q_j has to be chosen subject to the capacity constraint, it is no longer possible to directly employ the EOQ approximation as is done in the original [SCD] model. In particular, the cycle stock holding and inbound replenishment costs in (2.4) (the third term) must be formulated explicitly as:

$$G_j(Q_j, D_j) = F_j \frac{D_j}{Q_j} + \beta \left(g_j \frac{D_j}{Q_j} + a_j D_j \right) + \theta \frac{hQ_j}{2}, \tag{4.2}$$

where $D_j = \sum_{i \in I} \mu_i Y_{ij}$. Incorporating the new cycle stock expression, one can formulate the capacitated supply chain design problem as:

$$[\text{CSCD}] : \min \sum_{j \in J} \left[f_j X_j + \sum_{i \in I} \hat{d}_{ij} Y_{ij} + G_j \left(Q_j, \sum_{i \in I} \mu_i Y_{ij} \right) + q \sqrt{\sum_{i \in I} \sigma_i^2 Y_{ij}} \right] \tag{4.3}$$

s.t. (2.5-2.8) and (4.1). Atamtürk *et al.* (2012) observe that, under a single sourcing arrangement where $Y_{ij} \in \{0, 1\}$, one can transform both the nonlinear objective and capacity constraints into SOCP form, and the resulting [CSCD] problem can be written as an MISOCP. However, for cases with multiple sourcing or for larger-sized instances, one may adopt the Lagrangian relaxation decomposition procedure proposed by Ozsen *et al.* (2008) and Ozsen *et al.* (2009), as we discuss below.

4.1.2 Lagrangian Relaxation for Capacitated Problem

In this section, we shall illustrate the Lagrangian relaxation procedure for the case of multiple sourcing (Ozsen *et al.*, 2009). In particular, in the presence of capacity limits, one can possibly split the demand at a retailer among multiple DCs to allow for more flexible demand assignments. Modeling-wise, multiple sourcing can be captured by relaxing the integrality constraints in (2.8) to interval constraints $0 \leq Y_{ij} \leq 1$, for $i \in I, j \in J$. The Y_{ij} values can then be interpreted as the probability that an order from i is allocated to DC j, following some randomized order allocation procedure.

We next discuss how the [CSCD] problem with multiple sourcing can be solved with Lagrangian relaxation. The single sourcing variant of the problem can also be solved by using a similar procedure. To simplify the notation, we first define

$$W_j(D_j) = \left\{ \min_{Q_j > 0} G_j(Q_j, D_j), \text{ s.t. } (4.1) \right\}.$$

Following the Lagrangian relaxation procedure discussed in Section 3.1.2 for the [SCD] problem, we may relax constraints (2.5) and assign

Lagrangian multipliers π_i. Then, given the values of π_i, we face the following subproblem for each $j \in J$:

$$\begin{aligned} \min \quad & f_j X_j + \sum_{i \in I} (\hat{d}_{ij} - \pi_i) Y_{ij} + W_j \left(\sum_{i \in I} \mu_i Y_{ij} \right) \\ \text{s.t.} \quad & Y_{ij} - X_j \leq 0 \text{ for } i \in I \\ & X_j \in \{0,1\}, 0 \leq Y_{ij} \leq 1 \text{ for } i \in I. \end{aligned} \tag{4.4}$$

If $X_j = 0$, the objective value is 0. If $X_j = 1$, the problem reduces to:

$$\min_{0 \leq Y_{ij} \leq 1} \sum_{i \in I} (\hat{d}_{ij} - \pi_i) Y_{ij} + W_j \left(\sum_{i \in I} \mu_i Y_{ij} \right). \tag{4.5}$$

Letting $\hat{b}_i = \hat{d}_{ij} - \pi_i$ and $\hat{c} = \mu_i$ (the subscipts j are suppressed for brevity), we obtain a problem that is equivalent to the continuous relaxation of problem (3.6) considered in Section 3.1.1, but with the function $g(\cdot)$ (the square root function) replaced with $W_j(\cdot)$. Recall that the key result, Proposition 3.1, applies for all increasing function $g(\cdot)$, and thus holds for problem (4.5) as well. Proposition 3.1 suggests that at the optimal solution to (4.5), the subset Y_{ij} variables taking strictly positive values can be identified through a sorting procedure, according to the $\hat{b}_i/\hat{c}_i$ ratio. Furthermore, there is at most one i where $0 < Y_{ij} < 1$.

However, unlike the case for the subproblem of [SCD], the nonlinear function $g(\cdot)$ is now non-concave. Therefore, it is possible for Y_{ij} to take a value strictly between 0 and 1 for some i. Ozsen *et al.* (2008) provide the following result to identify possible fractional solutions in such scenarios, by analyzing the Karush-Kuhn-Tucker (KKT) necessary conditions for optimality of problem (4.5).

Proposition 4.1. [Theorem 2 of Ozsen *et al.* (2008)] Assume that $\mathbf{Y}^*$ is an optimal solution to (4.5). If Y^*_{kj} takes on a strictly fractional value, then the following holds for retailer k:

$$\hat{b}_k + \mu_k \frac{\partial}{\partial \sum_{i \in I} \mu_i Y^*_{ij}} W_j \left(\sum_{i \in I} \mu_i Y^*_{ij} \right) = 0. \tag{4.6}$$

Combining Propositions 3.1 and 4.1, Ozsen *et al.* (2008) propose the following algorithm.

Algorithm 3. The algorithm for solving problem (4.5) can be stated as follows:

- Step 0: Let $I^+ = \{i \in I | \hat{b}_i \geq 0\}$ and $I^- = I \setminus I^+$. Let $Y_{ij} = 0$ for all $i \in I^+$.
- Step 1: Sort retailers in set I^- such that $\hat{b}_1/\hat{c}_1 \leq \hat{b}_2/\hat{c}_2 \leq \hat{b}_m/\hat{c}_m$, where $m = |I^-|$.
- Step 2: For $k = 1, \cdots, m$, let $D_{j,k} = \sum_{i=1}^{k} \mu_i$ ($D_{j,0} = 0$). Define $\Delta_{jk} = \{\hat{D}_{j,k} \in \mathbb{R} | \hat{b_k} + \mu_k \frac{\partial W_j(\hat{D}_{j,k})}{\partial \hat{D}_{j,k}} = 0, D_{j,k-1} < \hat{D}_{j,k} < D_{j,k}\}$, i.e., Δ_{jk} corresponds to the set of solutions to (4.6).
  - If Δ_{jk} is empty, let $S_{j,k} = \sum_{i=1}^{k} \hat{b}_i + W_j(D_{j,k})$. This corresponds to the objective value of the candidate solution of setting $Y_{ij} = 1$ for $i = 1, \cdots, k$.
  - If Δ_{jk} is non-empty, let $\hat{S}_{j,k} = \min_{d \in \Delta_{jk}} \left\{ \sum_{i=1}^{k-1} \hat{b}_i + \hat{b}_k (\hat{D}_{j,k} - D_{j,k-1})/\mu_k + W_j(\hat{D}_{j,k}) \right\}$ and $D^*_{j,k}$ be the corresponding $\hat{D}_{j,k}$ value; and let $S_{j,k} = \min \left\{ \hat{S}_{j,k}, \sum_{i=1}^{k} \hat{b}_i + W_j(D_{j,k}) \right\}$. This corresponds to the objective value of the candidate solution of setting $Y_{ij} = 1$ for $i = 1, \cdots, k$, or $Y_{ij} = 1$ for $i = 1, \cdots, k-1$ and Y_{kj} to its best fractional value, whichever is better.
- Step 3: Let $k^* = argmin_{k=1,\cdots,m} S_{j,k}$. Then, the optimal solution is given by:

$$Y_{ij} = \begin{cases} 1, & \text{for } i < k^* \\ (D^*_{j,i} - D_{j,i-1})/\mu_i & \text{for } i = k^* \\ 0, & \text{for } i > k^* \text{ or } i \in I^+. \end{cases} \tag{4.7}$$

Embedding Algorithm 3 in the subgradient procedure (Algorithm 2), one can solve the [CSCD] problem efficiently. Ozsen *et al.* (2008) and Ozsen *et al.* (2009) discuss effective heuristics to construct upper bound (i.e., feasible) solutions based on solutions to the subproblem, for both the single sourcing and multiple sourcing versions of the problem. Note that subproblem (4.5) yields lower bounds for both variants, because

it relaxes the integrality constraints on the Y_{ij} variables in the single sourcing formulation. To obtain stronger lower bounds, it is also possible to perform a branch and bound procedure for the subproblem with integrality constraints, which involves using Algorithm 3 as a subroutine to solve its continuous relaxation.

Using this solution procedure, Ozsen *et al.* (2008) and Ozsen *et al.* (2009) find that high-quality solutions to the [CSCD] problem can be obtained efficiently. Furthermore, comparing between the two cases, they find that multiple sourcing can lead to significant cost savings because of the extra flexibility to allocate demand. They also find that it is typically the case that only a small number of retailers will be multi-sourced in the optimal solution. These findings provide design guidelines for supply chain planning scenarios in practice.

The capacitated network design model discussed in this section considers strategic capacity planning in light of future uncertainty in demand. The examples to be discussed in the next section further incorporates the consideration of risk more explicitly in the formulation of network design models.

4.2 Supply Chain Design under Uncertainty

Network design constitutes a key decision in strategic supply chain planning. Due to the capital intensive investments in facilities such as DCs, these decisions are costly, or impossible, to reverse. Especially when entering new markets or introducing new products, planners must pay extra attention into uncertainties in future demand, costs and other parameters. For classical facility location problems, Snyder (2006) provides an excellent review of modeling techniques for incorporating planning uncertainty, and the associated solution methodologies. In recent years, there has been a growing stream of work on incorporating planning uncertainty in integrated supply chain design models. In this section, we discuss some of the modeling concepts.

As discussed in Section 2.2, one popular approach for modeling optimization problems under uncertainty is to employ stochastic programming formulations. The scenario-based modeling approach, with its statistical foundation and managerial significance, has been often

employed to guide supply chain planning decisions. In this section, we formulate and discuss scenario-based extensions to the [SCD] model with different objectives that correspond to different risk preferences of the decision maker.

4.2.1 Risk Neutral Objective

The first cut approach to model optimal decisions under random environments is often to consider the average outcome, i.e., the expected value of the objective function. This corresponds to the risk neutral decision criterion, under which the decision maker is indifferent between a risky (random) outcome and a deterministic outcome as long as the expected value of the former is equal to the latter. This is often an appropriate decision criterion when the future payoff consists of many independent realizations of the random parameters over a long period of time, such that the cumulative average performance converges to the expected value. For example, in supply chain settings where demand for products are uncertain, the risk neutral objective is appropriate if the network has a long planned life cycle and the demand rates realized in each future period of operation are forecasted to follow a stationary probability distribution.

Snyder *et al.* (2007) consider a stochastic version of the [SCD] problem in which there are two decision phases. In the first phase, DC locations are chosen out of the candidate set J. At this stage, planning parameters, such as demand rates and shipping costs, are uncertain. We define a set S of mutually-exclusive and exhaustive scenarios of future outcomes regarding these operating parameters. The values of the demand and shipping cost parameters (μ_i and $\hat{d}_{ij}$ in the [SCD] objective function (2.9)) under scenario s ($\in S$) are denoted by μ_{is} and $\hat{d}_{ijs}$, respectively. We also assume that scenario s occurs with known probability p_s (where $\sum_{s \in S} p_s = 1$). The strategic facility location decisions (indicated by decision variables X_j) are made prior to observing the scenario realizations.

Then, after the DC locations are fixed, the problem proceeds to its second stage where one of the scenarios in S is realized and the decision maker needs to determine the tactical decisions of supply

chain operations. In particular, the assignment of retailer demand to DCs, indicated by decision variables Y_{ijs}, are determined conditional on scenario s being realized. These decisions determine the inventory- and transportation-related costs of the supply chain network. Then, the stochastic supply chain design problem can be formulated as follows.

$$\begin{aligned}
[\text{S-SCD}]: \quad & \min \sum_{j\in J}\left[f_j X_j + \sum_{s\in S} p_s \left(\sum_{i\in I} \hat{d}_{ijs} Y_{ijs} \right.\right. \\
& \left.\left. + \hat{K}_j \sqrt{\sum_{i\in I} \mu_{is} Y_{ijs}}\right)\right] && (4.8) \\
\text{s.t.} \quad & \sum_{j\in J} Y_{ijs} = 1 \text{ for } i \in I, s \in S && (4.9) \\
& Y_{ijs} - X_j \leq 0 \text{ for } i \in I, j \in J, s \in S \\
& X_j \in \{0,1\} \text{ for } j \in J \\
& Y_{ijs} \in \{0,1\} \text{ for } i \in I, j \in J, s \in S.
\end{aligned}$$

Note that the objective (4.8) is to minimize expected costs under the distribution of future scenarios. To solve this problem, Snyder *et al.* (2007) propose a similar Lagrangian relaxation algorithm as the one discussed in Section 3.1.2. In particular, if one relaxes (4.8) by imposing Lagrangian multipliers π_{is} for each $i \in I$ and $s \in S$, the problem can be decomposed by both i and s. This gives rise to $|I| \times |S|$ subproblems in the form of (3.10), which can be solved using the sorting procedure. Following this approach, Snyder *et al.* (2007) find that the scenario-based [S-SCD] problem can be solved efficiently.

4.2.2 Risk Averse Objective

While the [S-SCD] model is conceptually intuitive and computationally efficient to solve, it assumes risk neutrality, i.e., that the decision maker is indifferent between risky and risk-free outcomes, as long as the expected costs are the same. Conceptually, in a stochastic optimization problem, we make first-stage or "here-and-now" decisions ($\mathbf{X}$) when key parameters (denoted by $\boldsymbol{\theta}$) are not yet revealed and are known to follow some distribution $\mathcal{F}$. Then, as the values of $\boldsymbol{\theta}$ are revealed, we obtain an outcome (e.g., cost) of $L(\mathbf{X}, \boldsymbol{\theta})$, possibly given by the optimal objective

value of some recourse optimization problem (e.g., by optimizing the Y_{ijs} variables in [S-SCD]). The risk neutral [S-SCD] model optimizes the facility location decisions by considering the mean of the outcome, i.e., $E_{\mathcal{F}}[L(\mathbf{X}, \boldsymbol{\theta})]$. As such, the decision maker disregards the possible dispersion (or risk) of the random variable $L(\mathbf{X}, \boldsymbol{\theta})$. In practice, however, managers are often risk averse, and prefer outcomes with smaller degrees of risk. Conceptually, this corresponds to a future cost distribution that has lower variability, even at the expense of a larger expectation.

The appropriate choice of a risk-aware objective, however, is not trivial. The concept of risk measures is often useful. A risk measure is defined to be a deterministic, i.e., risk-free, quantity (payoff or cost) that the decision maker is willing to trade the risky position (the random outcome $L(\mathbf{X}, \boldsymbol{\theta})$) for. Various risk measures, their properties and applicability in various decision contexts are broadly studied in disciplines such as decision analysis, economics and finance (e.g., Artzner *et al.*, 1999). In the following, we discuss some popular risk measures (objectives) often employed in location analysis and how they can be incorporated in integrated supply chain design formulations.

To guard against adverse outcomes, the minimax objective is often employed in robust optimization (e.g., Serra and Marianov, 1998, for the P-median problem). This objective optimizes the worst-case (i.e., largest) realized cost out of all possible outcomes. That is, it considers $\sup_{\boldsymbol{\theta}} L(\mathbf{X}, \boldsymbol{\theta})$. Applying this to the [SCD] problem, one can replace objective function (4.8) in the [S-SCD] model by the following, which selects the largest (worst) cost out of all scenarios in S:

$$\min \sum_{j \in J} f_j X_k + W$$

$$\text{s.t.} \quad W \geq \sum_{i \in I} \hat{d}_{ijs} Y_{ijs} + \hat{K}_j \sqrt{\sum_{i \in I} \mu_{is} Y_{ijs}} \text{ for } s \in S.$$

Overall, the major advantage of the minimax approach is that the absolute worst case is optimized. Then, under all scenarios, the cost is bounded above by the tightest possible bound. This approach caters to decision makers who are pessimistic and want to plan for the absolute worst case. However, being too pessimistic is also the major problem of the minimax approach. Optimizing the worst outcome tends to generate

a solution in which the performance is uniform over all scenarios. The average case performance may be sacrificed in order to avoid the worst case scenario that may be extremely unlikely to occur. Therefore, the key problem with the minimax objective is that it weighs all possible scenarios equally. This is the most conservative approach taken, when there is not enough available data to define meaningful probability distributions on the outcome. However, even in the absence of detailed probabilistic information, the firm would always prefer to perform better under certain scenarios considered as important and be willing to sacrifice the performance under other less important ones. In fact, optimizing the weighted performance across scenarios is equivalent to assuming a probability measure on the defined scenarios and optimizing an expected utility function.

Several different approaches have been taken to avoid focusing on the absolute worst case. For example, the α-reliable framework defines a subset of scenarios, called the α-reliable set, and minimizes the worst cost or regret within this set. The α-reliable set is defined endogenously such that the total probability that the stochastic cost falls below the objective value considered is at least α, a level specified by the user. Using our previous notation, this objective corresponds to the measure

$$v_\alpha(\mathbf{X}) = \inf v, \text{ s.t. } P(L(\mathbf{X}, \boldsymbol{\theta}) \leq v) \geq \alpha.$$

A reason for doing this is that the worst possible case has a low probability of occurring (less than $1 - \alpha$) and we do not want such a highly unlikely event to alter the decision by too much. An example is that an airport is never designed for the absolute peak demand (e.g., the Sunday following Thanksgiving) because the cost of doing so is too high and such high capacity is rarely ever needed. Note that the α-reliable objective is equivalent to the α-quantile, i.e., the Value at Risk (VaR).

Daskin *et al.* (1997) formulate a new version of the P-median problem minimizing the maximum regret within the α-reliable set and evaluate the trade-off between having a higher value of α (probability that the regret will not exceed the optimal objective value) and having lower maximum regret. Applied to the [SCD] problem, one can formulate the α-reliable objective by replacing objective function (4.8) in the [S-SCD] model with the following:

$$\begin{aligned}
\min \quad & \sum_{j\in J} f_j X_k + W \\
\text{s.t.} \quad & W \geq \sum_{i\in I} \hat{d}_{ijs} Y_{ijs} + \hat{K}_j \sqrt{\sum_{i\in I} \mu_{is} Y_{ijs}} \\
& -M(1 - Z_s) \text{ for } s \in S \qquad (4.10) \\
& \sum_{s\in S} p_s Z_s \geq \alpha \qquad (4.11) \\
& Z_s \in \{0, 1\} \text{ for } s \in S. \qquad (4.12)
\end{aligned}$$

In the above, M is defined to be a constant larger than the maximum (second-stage) cost under any scenario $s \in S$. Binary decision variables Z_s are defined to indicate whether scenario s is included in the α-reliable set or not. Constraints (4.10) ensure that the value of W corresponds to the second-stage cost under the worst scenario included in the α-reliable set, and constraint (4.11) makes sure that the aggregate probability of scenarios included in the set exceeds α. Note that this new objective relies on the big-M formulation, which is known to be computationally inefficient in practice, due to their resulting in weak continuous relaxations in branch-and-bound. Generally, VaR objectives are known to be less tractable to optimize, because they do not preserve convexity, i.e., for $L(\mathbf{X}, \boldsymbol{\theta})$ convex in $\mathbf{X}$, $v_\alpha(\mathbf{X})$ is not convex in $\mathbf{X}$ in general. Another shortcoming of this risk measure is that it does not consider anything beyond the α-quantile and therefore does not distinguish between a long tail and a short tail.

To address these limitations, one can consider a risk measure known as conditional value at risk (CVaR). The α-CVaR is the conditional expectation of costs above the α-VaR (Rockafellar and Uryasev, 2000, 2002), defined as the expected excess costs associated with outcomes outside the α-reliable set. Mathematically, the α-CVaR is defined as:

$$v_\alpha(\mathbf{X}) + \frac{1}{1-\alpha} \int \max\{L(\mathbf{X}, \boldsymbol{\theta}) - v_\alpha(\mathbf{X})\} d\mathcal{F}(\boldsymbol{\theta}).$$

CVaR is known to be a coherent risk measure, i.e., it satisfies the axioms of monotonicity, translational invariance, subadditivity and positive homogeneity (for details, refer to, e.g., Artzner *et al.*, 1999). CVaR is particularly amenable to optimization, as it preserves convexity.

Furthermore, it can be computed in the following tractable form:

$$\inf_{v,\mathbf{X}} v + \frac{1}{1-\alpha} E_{\mathcal{F}}[L(\mathbf{X}, \boldsymbol{\theta}) - v_\alpha(\mathbf{X})]^+. \tag{4.13}$$

Applying this objective, Chen *et al.* (2006) formulate the α-reliable mean excess regret version of the P-median model and compare it with the Daskin *et al.* (1997) model. With the CVaR objective, we know that with probability of at least α, the outcome will be less than the optimal objective value, and with probability $1-\alpha$, the conditional expectation of the outcome is equal to the optimal objective value. They show that the CVaR model is much easier to solve than the α-reliable model when the problem contains a large number of scenarios.

In the [SCD] context, the CVaR objective can be formulated by replacing objective function (4.8) in the [S-SCD] model by the following:

$$\min \quad v + \frac{1}{1-\alpha} \sum_{s \in S} p_s U_s \tag{4.14}$$

$$\text{s.t.} \quad U_s \geq \sum_{i \in I} \hat{d}_{ijs} Y_{ijs} + \hat{K}_j \sqrt{\sum_{i \in I} \mu_{is} Y_{ijs}} - v \text{ for } s \in S \tag{4.15}$$

$$U_s \geq 0 \text{ for } s \in S. \tag{4.16}$$

The above formulation follows directly from (4.13) and does not introduce any additional binary variables to the problem. Overall, the CVaR concept provides an attractive modeling option that provides the versatility to capture different degrees of risk aversion with computationally tractable formulations.

The models discussed in this section are formulated based on a scenario-based approach. In the next section, we shall discuss that such ideas for formulations, when interpreted appropriately, can also be used to model supply chain networks that distribute multiple commodity types.

4.3 Multiple-Commodity Supply Chain Design

The supply chain design models that we have discussed thus far consider the tactical and operational aspects of managing one single commodity. This modeling approach serves as a reasonable approximation where

products processed by the facilities are similar in characteristics, such that they can be modeled collectively as a generic product following the aggregate planning principle (e.g., Nahmias and Cheng, 2009). In modern supply chains, however, product proliferation has caused product lines to lengthen significantly. For example, retailers such as WalMart and Amazon handle millions of products. Using Amazon as an example, its network of fulfillment centers carry inventory for a myriad of stock keeping units (SKUs). While many of these are fast moving products sold at high volumes, there are a large number of SKUs falling in the *long tail* with low sales volumes. It is a sensible strategy to control inventory for these products in different ways. In particular, for fast moving SKUs, Amazon can afford to stock them closer to the markets, by carrying them at all fulfillment centers. However, for low moving SKUs that are individually infrequently ordered, a more centralized stocking strategy, which maximizes risk pooling, is a sensible choice.

Incorporating these differentiated supply chain strategies for multiple product classes necessitates careful network design modeling. In this section, we discuss two models for this purpose. First, we propose an extension to the [SCD] model designed to handle a general number of commodities. Second, we review a recently proposed model that captures a small number (two) of product classes that exhibit different characteristics with respect to scale economies.

4.3.1 Multiple Commodities in [SCD] Model

We first consider a multiple-commodity extension of the [SCD] model, proposed by Shen (2005). Let S denote a set of commodities. For the demand and shipping cost parameters in (2.9), we append the subscript s corresponding to each commodity $s \in S$. We consider that a facility, once opened (indicated by decision variables X_j for $j \in J$), can be used to handle one or multiple commodities. We use decision variables Y_{ijs} to indicate whether demand for commodity s at retailer i is served by DC j $(= 1)$ or not $(= 0)$.

Then, the multiple-commodity supply chain design model can be formulated as follows

$$
\begin{aligned}
\text{[M-SCD]}: \quad & \min \sum_{j \in J} \left[f_j X_j + \sum_{s \in S} \left(\sum_{i \in I} \hat{d}_{ijs} Y_{ijs} + \hat{K}_j \sqrt{\sum_{i \in I} \mu_{is} Y_{ijs}} \right) \right] \\
\text{s.t.} \quad & \sum_{j \in J} Y_{ijs} = 1 \text{ for } i \in I, s \in S \\
& Y_{ijs} - X_j \leq 0 \text{ for } i \in I, j \in J, s \in S \\
& X_j \in \{0, 1\} \text{ for } j \in J \\
& Y_{ijs} \in \{0, 1\} \text{ for } i \in I, j \in J, s \in S.
\end{aligned}
$$

Note that the [M-SCD] formulation is structurally equivalent to the [S-SCD] problem discussed in Section 4.2.1, in that commodities in the former corresponds to scenarios in the latter. In both models, the objective is to minimize the (weighted) sum of costs realized over the set of commodities or scenarios. Because the two models are structurally identical, the Lagrangian relaxation solution approach developed for [S-SCD] will also work efficiently for [M-SCD].

4.3.2 Multiple Commodities with Different Scale Economies

The [M-SCD] model, while general enough to handle an arbitrary number of commodity types, assumes that economies of scale is exhibited in carrying inventory for each commodity (as a result of inbound transportation consolidation and safety stock risk pooling). In practice, however, it is noted that both economies and diseconomies of scale can often be exhibited at different levels of throughput at facilities (Lu *et al.*, 2014). In particular, at high throughput levels, congestion occurs and the marginal cost of handling additional demand volume tends to increase. Therefore, the inventory cost function for a facility, instead of being a concave (square root) function, can be "inverse S-shaped," i.e., consists of a concave region at low demand volumes, and then a convex region at higher volumes. Lu *et al.* (2014) propose a supply chain design model with a single commodity, where operating costs at facilities can be reflected with inverse S-shaped cost functions, and an associated column generation solution algorithm.

In a multiple-commodity setting, different scale economies characteristics of different commodities pose additional planning difficulties.

To address this, Shu *et al.* (2014) propose a model that addresses the supply
scale economies characteristics. We discuss this model for the special case of two commodities by using the same notation as defined before:

$$\begin{aligned}
\min \quad & \sum_{j\in J}\Bigg[f_j X_j + \sum_{i\in I}\hat{d}_{ij1}Y_{ij} + \sum_{i\in I}\hat{d}_{ij2}Y_{ij} \\
& + g_{j1}\left(\sum_{i\in I}\mu_{i1}Y_{ij}\right) + + g_{j2}\left(\sum_{i\in I}\mu_{i2}Y_{ij}\right)\Bigg] \qquad (4.17)\\
\text{s.t.} \quad & \sum_{j\in J} Y_{ij} = 1 \text{ for } i \in I \\
& Y_{ij} - X_j \leq 0 \text{ for } i \in I, j \in J \\
& X_j \in \{0,1\} \text{ for } j \in J \\
& Y_{ij} \in \{0,1\} \text{ for } i \in I, j \in J.
\end{aligned}$$

In the objective function (4.17), the functions $g_{j1}(\cdot)$ and $g_{j2}(\cdot)$ are operational cost functions for the two commodities being handled at facility j, and are allowed to take any increasing, concave-convex (i.e., nonlinear) shape. Shu *et al.* (2014) consider that demand for all commodities at the same retailer to be assigned to the same DC, i.e., Y_{ij} (instead of Y_{ijs} for $s = 1, 2$). In practice, such a constraint may help control operational complexity and possibly reduce shipping costs by shipment consolidation of demand for the two commodities. They propose a column generation method to solve the problem. Following the standard column generation procedure discussed in Section 3.1.1, one needs to solve the following pricing subproblem for each $j \in J$, similar to (3.6) (again, suppressing subscripts j):

$$\min_{\mathbf{y}\in\{0,1\}^{|I|}} \sum_{i\in I}\hat{b}_i y_i + g_1\Big(\sum_{i\in I}\mu_{i1}y_i\Big) + g_2\Big(\sum_{i\in I}\mu_{i2}y_i\Big) \qquad (4.18)$$

where $\hat{b}_i = \hat{d}_{ij} - \pi_i$, $\hat{c}_{i1} = \mu_{i1}$ and $\hat{c}_{i2} = \mu_{i2}$.

Recall that, for (3.6) in which the objective function contains only one nonlinear term, Proposition 3.1 suggests that the optimal solution to the continuous relaxation can be obtained by first ranking the set I by the $\hat{b}_i/\hat{c}_i$ ratio, and that at most one y_i can take on a fractional value. However, in (4.18), the same result does not hold when there

are two nonlinear terms $g_1(\cdot)$ and $g_2(\cdot)$. Shu *et al.* (2014) show that, interestingly, the following result holds:

Proposition 4.2. Given any increasing functions $g_1(\cdot)$ and $g_2(\cdot)$, for any optimal solution to the continuous relaxation of (4.18), denoted by $(y_1^*, \cdots, y_{|I|}^*)$, there exist α, β such that:

1. $\alpha \frac{\hat{c}_{i1}}{\hat{b}_i} + \beta \frac{\hat{c}_{i2}}{\hat{b}_i} \leq 1$ for any i where $y_i^* = 1$;
2. $\alpha \frac{\hat{c}_{i1}}{\hat{b}_i} + \beta \frac{\hat{c}_{i2}}{\hat{b}_i} = 1$ for any i where $y_i^* \in (0, 1)$;
3. $\alpha \frac{\hat{c}_{i1}}{\hat{b}_i} + \beta \frac{\hat{c}_{i2}}{\hat{b}_i} \geq 1$ for any i where $y_i^* = 0$.

Furthermore, there exist at most two elements i such that $y_i^* \in (0, 1)$.

To see the connection between Propositions 4.2 and 3.1 (which holds where the objective function does not include $g_2(\cdot)$), consider the case where $\hat{c}_{i2} \equiv 0$, i.e., $g_2(\cdot)$ is a constant. In this case, Proposition 4.2 suggests that the optimal solution can be obtained by first sorting the set I in increasing order of the $\frac{\hat{c}_{i1}}{\hat{b}_i}$ ratio; and subsequently partitioning the sorted set into three subsets, where y_i is set to 1, some fractional value, and 0, respectively. This result is consistent with Proposition 3.1.

The proof of Proposition 4.2 is quite involved and can be found in Shu *et al.* (2014). Based on this proposition, Shu *et al.* (2014) further propose a geometric argument that solves the continuous relaxation of (4.18) in polynomial time. In particular, one first maps the elements in I to points $(\frac{\hat{c}_{i1}}{\hat{b}_i}, \frac{\hat{c}_{i2}}{\hat{b}_i})$ on a two-dimensional plane. Then, conditions 1-3 in Proposition 4.2 refer to partitioning the plane into two half-planes with the line $\alpha x + \beta y = 1$ for some α, β. All $O(|I|^2)$ possible partitions can be enumerated considering that there are either zero, one or two elements in I where y_i^* take on fractional values (i.e., the mapped points lie exactly on the line by condition 2). For each partition, the possible fractional values of y_i^* can be obtained via solving for the KKT conditions of the problem while fixing all other variables to 0 or 1 as appropriate. Using this procedure, the continuous relaxation of (4.18) can be solved efficiently. This can be embedded as a subroutine of the column generation algorithm (possibly with branch and bound to

enforce integrality for the pricing problem) to solve the original problem (4.17).

The discussion of multiple-commodity models, particularly in Section 4.3.1, illustrates how scenario-based models could be used for modeling different product types, in addition to modeling uncertainty. However, such models also come with potential limitations. The next section discusses a problem in which scenarios are a less effective approach to modeling disruption-related uncertainties in supply chain operations, and other formulation techniques are better suited for the purpose.

4.4 Supply Chain Design with Disruption Considerations

Driven by catastrophic events including September 11 and Hurricane Katrina, much research interest has been drawn on devising strategies that mitigate the risks of disruptions in supply chains. The presence of disruption risks are known to substantially alter the desirability of supply chain strategies. A case in point is the impact of the 2011 Japanese earthquake and tsunami on Toyota's supply chain. With all its accolades for its just-in-time strategy under normal circumstances, Toyota finds that its lean supply network structure (e.g., its rule of sourcing from only two suppliers for critical parts) leaves it with little excess capacity to quickly revamp production after the disruption and thus hampers recoverability (Wall Street Journal, 2011b). Researchers have cautioned that disruption risks are fundamentally different than demand uncertainty (Snyder and Shen, 2006) and production yield uncertainty (Chopra *et al.*, 2007), despite the fact that all three exacerbate the demand-supply mismatch, and must be safeguarded against using different strategies. Tomlin (2006, 2009) discusses the use of different mitigation and contingency strategies to protect the supply chain against disruptions. Excellent reviews of operations management research on the topic of supply chain disruptions can be found in Vakharia and Yenipazarli (2009) and Snyder *et al.* (2015).

Research on strategic facility location under the threat of disruptions has emerged in the literature over the past decade. Due to the all-or-nothing (i.e., a facility is either running or disrupted) nature of disruptions, planning uncertainty has to be modeled differently than for

other types of recurrent uncertainty, which often can be characterized using scenario planning (as discussed in Section 4.2.1). Particularly, because each facility can either be disrupted or not, the number of disruption scenarios can be up to $2^{|J|}$ for $|J|$ candidate facility locations under consideration, leading to computationally intractable formulations. To circumvent this difficulty, researchers make use of special problem structures to formulate tractable models. Snyder and Daskin (2005) propose a formulation for the P-Median problem when facilities are subject to independent disruptions with equal probabilities, together with an efficient Lagrangian relaxation solution algorithm. Shen *et al.* (2011) consider the UFL variant of the model and propose heuristics and approximation algorithms. Lim *et al.* (2010) and Lim *et al.* (2013) study an extension in which facility location and fortification decisions are made jointly, and propose integer programming and continuous approximation models for the problem, respectively. Cui *et al.* (2010) generalize the model of Snyder and Daskin (2005) to allow unequal disruption probabilities and devise an efficient branch-and-bound algorithm for its Lagrangian relaxation subproblem based on a supermodular property. Li and Ouyang (2010) use a continuous approximation approach to study the effect of spatially-correlated disruptions under a similar UFL setting. Note that these papers consider direct extensions of the UFL and P-median models and do not consider inventory costs.

There have been several works that incorporate inventory cost considerations in supply chain design under the threat of disruptions. The common assumption in this literature is that disruptions lead to complete shutdowns of facilities. This all-or-nothing capacity uncertainty motivates Qi *et al.* (2010) study a supply chain design problem in which supply disruptions may occur at either the supplier or the facilities (where inventory is held). Mak and Shen (2012) consider a stochastic optimization model for designing a supply chain network with dynamic sourcing, an arrangement common for online retailers (to be discussed in Section 4.5). Under the threat of disruptions, the dynamic sourcing arrangement enables both risk pooling (inventory sharing among facilities) and risk diversification (limiting disruption losses by placing smaller quantities of inventory over more facilities). Chen *et al.* (2011)

consider an extension of the [SCD] model in which retailer demand is to be reassigned to other facilities if the primary one is disrupted. As the former two studies have been covered in previous reviews (Mak and Shen, 2011), we will focus on reviewing the latter paper in this section.

4.4.1 Model with Backup Assignments

The main idea behind the model formulation of Chen *et al.* (2011) is the concept of multiple-level backup assignments in response to facility disruptions. In particular, a retailer can be assigned to one facility each at R ($\leq |J|$) different levels. The level-r facility will serve the retailer's demand if the all facilities assigned at levels $1, \cdots, r-1$ are disrupted; i.e., the level-one facility is the primary service facility and the level-r facility serves as the $(r-1)$-st backup. This notion of multi-level facility assignments was initially proposed by Snyder and Daskin (2005) for the reliable P-Median problem.

Suppose any facility $j \in J$ can be disrupted independently and with equal probability q. Let binary decision variable $Y_{ijr} = 1$ if facility j serves retailer i at level r, and 0 otherwise. Then, if $Y_{ijr} = 1$, j serves demand at i with a probability of $q^{r-1}(1-q)$, which is the probability that the level-$1, \cdots, r-1$ facilities are all disrupted while j is working. As a result, the expected demand rate at facility j is given by $\sum_{i \in I} \sum_{j=1}^{R} \mu_i (1-q) q^{r-1} Y_{ijr}$.

Chen *et al.* (2011) consider demand to be deterministic, and consider an EOQ-type cost structure. In particular, the fixed and variable replenishment costs at DC j are given by g_j and $\bar{a}_j$, respectively, and the holding cost rate is h. The ordering quantity is Q_j is to be determined. The annual inventory and inbound logistics costs at DC j can be expressed as:

$$\begin{aligned} C_j(Q_j) &= \frac{g_j \sum_{i \in I} \sum_{j=1}^{R} \mu_i (1-q) q^{r-1} Y_{ijr}}{Q_j} + \frac{h Q_j}{2} \\ &\quad + \bar{a}_j \sum_{i \in I} \sum_{j=1}^{R} \mu_i (1-q) q^{r-1} Y_{ijr}. \end{aligned}$$

Optimizing over Q_j, we obtain the optimal costs as:

$$\begin{aligned} C_j(Q_j^*) &= \sqrt{2g_j h \sum_{i\in I}\sum_{j=1}^{R} \mu_i(1-q)q^{r-1}Y_{ijr}} \\ &\quad +\bar{a}_j \sum_{i\in I}\sum_{j=1}^{R} \mu_i(1-q)q^{r-1}Y_{ijr}. \end{aligned}$$

Then, Chen *et al.* (2011) formulate the reliable joint inventory-location problem [RJIL] as follows:

$$\begin{aligned} \text{[RJIL]}: \min \quad & \sum_{j\in J}\left[f_j X_j + \sum_{i\in I}\sum_{j=1}^{R} d_{ij}\mu_i(1-q)q^{r-1}Y_{ijr} + C_j(Q_j^*)\right] \\ & +\pi \sum_{i\in I}\mu_i q^R \\ = \quad & \sum_{j\in J}\left[f_j X_j + \sum_{i\in I}\sum_{j=1}^{R}\beta_{ijr}Y_{ijr} + \sqrt{\sum_{i\in I}\sum_{j=1}^{R}\alpha_{ijr}Y_{ijr}}\right] \\ & +\pi \sum_{i\in I}\mu_i q^R \qquad (4.19) \\ \text{s.t.} \quad & \sum_{j\in J} Y_{ijr} = 1 \text{ for } i\in I, r = 1,\cdots,R \qquad (4.20) \\ & \sum_{r=1}^{R} Y_{ijr} \le X_j \text{ for } i\in I, j\in J, r = 1,\cdots,R \qquad (4.21) \\ & X_j, Y_{ijr} \in \{0,1\} \text{ for } i\in I, j\in J, r = 1,\cdots,R \end{aligned}$$

where $\alpha_{ijr} = (d_{ij} + \bar{a}_j)\mu_i(1-q)q^{r-1}$, $\beta_{ijr} = 2g_j h\mu_i(1-q)q^{r-1}$.

In the above formulation, the four terms in the objective function (4.19) include the fixed location costs of DCs, outbound transportation costs, inventory and inbound transportation costs, and penalty costs for not meeting demand, i.e., when all R assigned facilities have failed and demand cannot be met (incurring a penality of π per unit of demand). The constraints (4.20) and (4.21) stipulate that each retailer is assigned to one facility at each of the R levels, and that said facilities are opened, respectively.

4.4.2 Solution Approach

The formulation of [RJIL] consists of a nonlinear objective function (with square root terms), linear constraints, and binary decision variables. Invoking the fact that $Y_{ijr} = Y_{ijr}^2$, it is possible to express the square root terms in the objective function with additional variables and SOCP constraints. It is then possible to solve the problem directly using integer conic programming solvers. Alternatively, one can utilize Lagrangian relaxation methods, which is the route taken by Chen *et al.* (2011). We briefly review their approach below.

Obtaining a Lower Bound

Similar to the case with solving the [SCD] problem, the Lagrangian relaxation algorithm for [RJIL] proceeds with relaxing constraints (4.20) and imposing Lagrangian multipliers π_i. Then, the resulting problem is decomposable by $j \in J$:

$$\min_{\mathbf{X},\mathbf{Y}\in\{0,1\}} f_j X_j + \sum_{i\in I}\sum_{j=1}^{R}(\beta_{ijr} - \pi_i)Y_{ijr} + \sqrt{\sum_{i\in I}\sum_{j=1}^{R}\alpha_{ijr}Y_{ijr}} \quad (4.22)$$

$$\text{s.t. } (4.21).$$

To solve the above subproblem, one can compare the cases of setting $X_j = 0$ or $X_j = 1$. In the former case, the resulting objective is zero. In the latter, one needs to further solve the following problem to determine the values of Y_{ijr}:

$$\min_{\mathbf{Y}\in\{0,1\}} \sum_{i\in I}\sum_{j=1}^{R}\gamma_{ijr}Y_{ijr} + \sqrt{\sum_{i\in I}\sum_{j=1}^{R}\alpha_{ijr}Y_{ijr}} \quad (4.23)$$

$$\text{s.t.} \sum_{r=1}^{R+1} Y_{ijr} = 1 \text{ for } i \in I, \quad (4.24)$$

where $\gamma_{ijr} = \beta_{ijr} - \pi_i$. Note that, a slack variable $Y_{ij(R+1)}$ is added to express constraint (4.24) in equality form.

Subproblem (4.23) has an objective function that is structurally similar to (3.4). However, the same solution algorithm cannot be directly adapted due to the additional constraints (4.24) stipulating that the

facility can be assigned to each i at most one of the R levels (the case of no assignment is indicated by assigning to the dummy $R+1$-st level). For this subproblem, Chen *et al.* (2011) devise a polynomial-time algorithm based on the interval partition argument to be outlined below. For ease of exposition, we drop the subscript j in the remainder of this section as we focus on solving subproblem (4.23) for a given $j \in J$.

First, assume that $(\alpha_{ir}, \gamma_{ir}) \neq (\alpha_{ir'}, \gamma_{ir'})$ for $r \neq r'$ (without loss of generality, because otherwise, we can consider either r or r' and remove the other, without affecting the optimal objective value) and let $\mathbf{R} = \{1, \cdots, R+1\}$. Under (4.24), one of Y_{ir} is to be set to one for each i. Suppose the values of Y_{kr} have been fixed for $k \in I \setminus \{i\}$ and let $w_i = \sum_{k \in I \setminus \{i\}} \sum_{r=1}^{R} \alpha_{kr} Y_{kr}$. Then, the marginal contribution to the objective value by setting $Y_{ir} = 1$ is given by $M_{ir}(w_i) = \sqrt{w_i + \alpha_{ir}} - \sqrt{w_i} + \gamma_{ir}$.

For fixed w_i, the optimal r^* where $Y_{ir^*} = 1$ is given by: $\rho_i(w_i) = \{r \in \mathbf{R} | M_{ir'}(w_i) \geq M_{ir}(w_i), \forall r' \in \mathbf{R}\}$. Graphically, this can be obtained by plotting $M_{ir}(w_i)$ against w_i for all r, and selecting the r corresponding to the lowest curve at each point w_i. Instead of enumerating over $\mathbf{R}$ to identify the lowest curve, note that one only needs to consider only a reduced set, as explained below. Note that the set $\rho_i(w_i)$ will contain multiple elements if the lowest curves intersect at w_i. Note also that $\rho_i(w_i)$ is a subset of $N_i = \{r \in \mathbf{R} | M_{ir}(0) < M_{ir'}(0)$ or $\gamma_{ir} < \gamma_{ir'}, \forall r' \in \mathbf{R} \setminus \{r\}\}$, and it is sufficient to consider only elements in N_i rather than the full set $\mathbf{R}$. This is because, for the curve associated with r to be the lowest, it holds for every $r' \in \mathbf{R} \setminus \{r\}$ that either $M_{ir}(w) \leq M_{ir'}(w)$ for all $w \geq 0$, or the curves $M_{ir}(w)$ and $M_{ir'}(w)$ cross at some $\hat{w} > 0$. In the latter case, by continuity of $M_{ir}(\cdot)$ and $M_{ir'}(\cdot)$, the two curves cross only if (i) $M_{ir}(0) < M_{ir'}(0)$ and $M_{ir}(\infty) = \gamma_{ir} > M_{ir'}(\infty) = \gamma_{ir'}$, or (ii) $M_{ir}(0) > M_{ir'}(0)$ and $M_{ir}(\infty) = \gamma_{ir} < M_{ir'}(\infty) = \gamma_{ir'}$. For $r, r' \in N_i$, $M_{ir}(w_i)$ and $M_{ir'}(w_i)$ intersect at

$$\bar{w}^i_{rr'} = \frac{(\alpha_{ir} - \alpha_{ir'})^2}{4(\gamma_{ir} - \gamma_{ir'})^2} + \frac{(\gamma_{ir} - \gamma_{ir'})^2}{4} - \frac{\alpha_{ir} + \alpha_{ir'}}{2} > 0.$$

Let $|N_i| = n$. If $n > 1$, then one can sort the elements in N_i into an ordered sequence $r(i,1), r(i,2), \cdots, r(i,n)$, such that $\gamma_{i,r(i,k)} > \gamma_{i,r(i,k+1)}$ for $1 \leq k \leq n-1$. Note that, because $(\gamma_{ir} - \gamma_{ir'})(M_{ir}(0) - M_{ir'}(0)) < 0$ for any $r, r' \in N_i$, this ordering implies $M_{i,r(i,k)} < M_{i,r(i,k+1)}$ and

$\alpha_{i,r(i,k)} < \alpha_{i,r(i,k+1)}$. Then, we define the following breakpoints of w_i values that help characterize the optimal r to select:

$$\begin{aligned} w_1^{i-} &= 0, \quad w_n^{i+} = \infty \\ w_k^{i-} &= \max_{k'=1,\cdots,k-1} \bar{w}^i_{r(i,k),r(i,k')} \text{ for } 2 \le k \le n \\ w_k^{i+} &= \min_{k'=k+1,\cdots,n} \bar{w}^i_{r(i,k),r(i,k')} \text{ for } 1 \le k \le n-1. \end{aligned}$$

Recall that $\bar{w}^i_{rr'}$ indicates the w_i value at which the functions $M_{ir}(w_i)$ and $M_{ir'}(w_i)$ intersect. Therefore, w_k^{i-} and w_k^{i+} indicate the highest and lowest points along the w_i axis at which $M_{ir(i,k)}(w_i)$ intersects with $M_{ir'}(w_i)$ for r' among those ranked (in increasing ordering of $M_{ir}(0)$) below and above $r(i,k)$, respectively. Note that the intervals $[w_k^{i-}, w_k^{i+}]$, $k = 1, \cdots, n$ for a non-overlapping partition of $[0, \infty)$. This partition helps identify the optimal r for different values of w_i.

Proposition 4.3. [Proposition 2 in Chen *et al.* (2011)] For all $i \in I$ and $w_i \ge 0$, $\rho_i(w_i) = \{r(i,k) | w_i \in [w_k^{i-}, w_k^{i+}]\}$.

Proof. First, we show that $w_i \in [w_k^{i-}, w_k^{i+}]$ implies that $M_{ir(i,k)} \le M_{ir(i,k')}$ for all $k' \le k$. For $k' > k$, it holds that $w_i \le w_k^{i+} = \min_{l=k+1,\cdots,n} \bar{w}^i_{r(i,k),r(i,l)} \le \bar{w}^i_{r(i,k),r(i,k')}$. As $k' > k$, $M_{i,r(i,k)}(w) < M_{i,r(i,k')}(w)$ at $w = 0$, and the two (continuous) functions intersect once at $\bar{w}^i_{r(i,k),r(i,k')}$. Therefore, $M_{i,r(i,k)}(w_i) \ge M_{i,r(i,k')}(w_i)$. Using a similar argument, one can also show that the same holds for $k' < k$.

Next, to prove the converse claim, we need to show that $w_i \notin [w_k^{i-}, w_k^{i+}]$ implies that $M_{ir(i,k)} > M_{ir(i,k')}$ for some $k' \ne k$. Consider some $k' < k$ and note again that $w_i \le w_{k'}^{i+} = \min_{l=k'+1,\cdots,n} \bar{w}^i_{r(i,k'),r(i,l)} \le \bar{w}^i_{r(i,k'),r(i,k)}$. Because $w_i \notin [w_k^{i-}, w_k^{i+}]$, the above inequalities cannot both hold at equality, and therefore $w_i < \bar{w}^i_{r(i,k'),r(i,k)}$. Since $M_{i,r(i,k)}(0) > M_{i,r(i,k')}(0)$ and the two functions intersect once only at $\bar{w}^i_{r(i,k'),r(i,k)}$, the above imply that $M_{i,r(i,k')}(w_i) < M_{i,r(i,k)}(w_i)$. One can follow a similar argument for $k' > k$ to complete the proof. □

Proposition 4.3 suggests that it is optimal to set $Y_{i,r(i,k)} = 1$ if and only if $w_i \in [w_k^{i-}, w_k^{i+}]$. This condition is further equivalent to $w = \sum_{i \in I} \sum_{r=1}^{R+1} \alpha_{ir} Y_{ir} \in [\hat{w}_k^{i-}, \hat{w}_k^{i+}]$, where $\hat{w}_k^{i-} = w_k^{i-} + \alpha_{i,r(i,k)}$ and

$\hat{w}_k^{i-} = w_k^{i+} + \alpha_{i,r(i,k)}$. Note that because $\alpha_{i,r(i,k)}$ is increasing in k and $w_k^{i+} \leq w_{k+1}^{i-}$, the intervals $[\hat{w}_k^{i-}, \hat{w}_k^{i+}]$, $k = 1, \cdots, n$, are mutually disjoint but their union is a strict subset of $[0, \infty)$. Let $\omega_i = \cup_{k=1}^{n}\{[\hat{w}_k^{i-}, \hat{w}_k^{i+}]\}$.

To optimize the subproblem (4.23), we need to identify the value of $w = \sum_{i\in I}\sum_{r=1}^{R+1}\alpha_{ir}Y_{ir}$ where $Y_{ir} = 1$ if and only if $w \in [\hat{w}_k^{i-}, \hat{w}_k^{i+}]$ concurrently for all $i \in I$. That is, an optimal solution correspond to a w value that satisfies $w \in \cap_{i\in I}\omega_i$. Note that by definition of ω_i, $\cap_{i\in I}\omega_i$ is a union of a polynomial number of disjoint intervals. By enumerating these intervals and constructing the solution (i.e., the Y_{ir} values) accordingly, one can evaluate all such candidate solutions to find the one with minimum cost. This can be done using the following algorithm.

Algorithm 4. The following algorithm solves (4.23):

- Step 1: For each $i \in I$, compute the non-dominated subset N_i of $\{1, \cdots, R+1\}$, and the values of $\hat{w}_k^{i-}, \hat{w}_k^{i+}$ for all $k = 1, \cdots, |N_i|$.
- Step 2: Sort pairs $\mathbf{W} = \{\hat{w}_k^{i-}, 0\}_{i\in I, k=1,\cdots,|N_i|}\cup\{\hat{w}_k^{i+}, 1\}_{i\in I, k=1,\cdots,|N_i|}$ into $\{(s_1, c_1), (s_2, c_2), \cdots, (s_{|\mathbf{W}|}, c_{|\mathbf{W}|})\}$ such that $s_1 \leq s_2 \leq \cdots \leq s_{|\mathbf{W}|}$. Initialize $k = 1$ and the incumbent objective value $V = 0$. Initialize incumbent solution $\mathbf{Y}$ by letting $Y_{ir} = 0$ for $r = 1, \cdots, R$ and $Y_{i,R+1} = 1$ for all $i \in I$.
- Step 3a: Repeat incrementing l by 1 until $c_l = 0$ and $c_{l+1} = 1$. If $l = |\mathbf{W}|$, return solution $\mathbf{Y}$ and objective value $\mathbf{0}$; Otherwise, go to Step 3b.
- Step 3b: For each $i \in I$, identify $k(i) \in \{1, \cdots, |N_i|\}$ that satisfies $[s_l, s_{l+1}) \subseteq [\hat{w}_{k(i)}^{i-}, \hat{w}_{k(i)}^{i+}]$. If $k(i)$ exists for all i, set $\hat{Y}_{i,r(i,k(i))} = 1$ and $\hat{Y}_{i,r'} = 0$ for $r \neq r(i, k(i))$ and go to Step 3c; otherwise, go to Step 3a.
- Step 3c: Compute $w = \sum_{i\in I}\sum_{r=1}^{R+1}\alpha_{ir}\hat{Y}_{ir}$. If $w \in [s_l, s_{l+1})$, then evaluate the objective $\hat{V}$ associated with solution $\hat{\mathbf{Y}}$. If $\hat{V} < V$, set $V = \hat{V}$ and $\mathbf{Y} = \hat{\mathbf{Y}}$. Go to Step 3a.

In Algorithm 4, Step 2 identifies all intersections of $[\hat{w}_k^{i-}, \hat{w}_k^{i-}]$ for all pairs of i and k, and sorts them in increasing order. Then, Step 3b

checks if each of these intersections indeed intersects $[\hat{w}_k^{i-}, \hat{w}_k^{i-}]$ for some k in all other $i \in I$. If it does, it is a candidate solution and its objective value is evaluated in Step 3c. After all candidate solutions are evaluated, the best one (i.e., with the lowest objective value) is returned.

After solving (4.23) for j, let V_j denote the corresponding objective value. To solve (4.22), it is optimal to set $X_j = 1$ and $\mathbf{Y}$ according to the solution identified in Algorithm 4 if $V_j + f_j < 0$; and $X_j = 0$ otherwise. Collectively for all $j \in J$, this procedure identifies a lower bound for the [RJIL] problem.

4.4.3 Obtaining an Upper bound

The solution constructed in the lower bound procedure may violate constraints (4.20). In particular, it is possible that (i) $\sum_{j \in J} Y_{ijr} > 0$ (i.e., a retailer is assigned to multiple DCs at the same level) or (ii) $\sum_{j \in J} Y_{ijr} = 0$ (i.e., a retailer is not assigned to any DC at some level), for some r and i. To obtain a feasible solution (which generates an upper bound on the optimal objective value), Chen *et al.* (2011) propose two heuristic procedures to repair the lower bound solution to ensure feasibility.

The first heuristic checks across all i and r for the two types of infeasibility scenarios. In particular, it loops through all $i \in I$ and $r = 1, \cdots, R$, and whenever type (i) infeasibility is detected, the solution is modified by keeping only one facility among those assigned that minimizes costs, assuming all other assignments remain unchanged (i.e., in a greedy manner). After repeating the same for all i and r, any open DC that is not assigned to any retailer is closed. Then, the procedure loops through all i and r again to identify type (ii) scenarios. Whenever such a case is found, it compares the costs of assigning said retailer i to each open facility at level r assuming all other assignments are fixed, and completes the assignment with the lowest cost.

The second heuristic only utilizes the values of $\mathbf{X}$ from the lower bound solution. To determine the values of $\mathbf{Y}$ while guaranteeing feasibility with respect to constraints (4.20), the procedure simply assigns each retailer to the r-th closest open facility at level r for each $r = 1, \cdots, R$. Both heuristics are easy to implement and computationally inexpensive.

With these upper bound identification heuristics, one may then proceed with the standard subgradient algorithm to solve the original [RJIL] problem. These heuristics are found to be effective by Chen *et al.* (2011). For a 49-city data set, they find that the Lagrangian relaxation algorithm is consistently able to identify feasible (upper bound) solutions that are within less than one percent from the optimal solution with computation times of less than one minute.

One potential limitation of the [RJIL] model is that it considers the long-run average inventory cost under the continuous review model. In practice, when facilities are disrupted, demand reallocation among the working facilities is typically utilized as a short-term measure. That is, the consideration of long-run average inventory cost is not necessarily the most accurate approximation. One potential approach to overcome this limitation is to consider a periodic review inventory model in which demand can be reallocated, rather than a continuous review one. We shall discuss a model in the coming section that utilizes this idea in a slightly different application context.

4.5 Fulfillment Center Location for Online Retailers

Recent years have seen tremendous growth of online retail businesses (e.g., Forbes, 2013). Online retail stores and platforms such as Amazon and Alibaba are heavily investing in logistics infrastructure to enhance distribution capabilities (e.g., Forbes, 2015). With the increasing focus on delivery response times with the emergence of same-day delivery businesses, rapid and cost-efficient distribution has become a key competitive focus of online retailers.

Different from many traditional distribution networks serving brick-and-mortar retail stores, one key operating characteristic of online retail distribution systems is that the allocation of demand (orders) is often performed on a per-order basis. That is, orders from the same geographical market for the same item may be fulfilled by different DCs, depending on the delivery time requirements of the orders and inventory levels at different DCs (e.g., Xu *et al.*, 2009; Acimovic and Graves, 2014), in a way that minimizes total fulfillment costs. Note that this mode of operations allows inventory pooling among multiple

DCs without having to physically centralize operations. Compared with traditional distribution models in which customer orders are fulfilled using single sourcing (modeling-wise, where the Y_{ij} variables are binary), this new mode of fulfillment, which leads to a dynamic multiple sourcing arrangement, provides an opportunity to improve both dimensions of customer service (response time and distance-to-market) and operating costs in the fundamental trade-off. This opportunity arises from the fact that statistical economies of scale of inventory pooling can now be achieved while operating a denser facility network because inventory can be shared among facilities.

To quantify the contrast between these online distribution systems and traditional ones, additional modeling effort is required to develop decision models for the design of such networks, as the basic models (such as [SCD]) typically consider static demand allocation, often with single sourcing, and do not capture the dynamic fulfillment feature of online retail. In this section, we discuss a stochastic programming model proposed by Mak (2012) for this type of problems.

4.5.1 Network Design with Dynamic Fulfillment

In this section, we present the problem formulation for the supply chain design problem for online retailers with dynamic fulfillment, proposed by Mak (2012). The problem consists of two phases. In the first stage (the design phase), the network structure, i.e., location of DCs and the allocation of retailers to DCs, are to be determined. We use the binary variable X_j to indicate whether a DC is located at site $j \in J$, in which case a fixed cost of f_j is incurred. Besides, we use Z_{ij} to indicate whether facility $j \in J$ is connected to demand location $i \in I$, in which case a fixed cost of c_{ij} is incurred. This cost term is used for modeling the additional costs due to flexible shipment patterns, and may reflect physical costs such as the use of more flexible but more expensive transportation options, and the additional inconvenience costs due to operational complexities.

Then, in the second (management) phase, the inventory and shipment decisions are made given the network structure determined in the design stage. We consider that the DCs operate under periodic

review in the management phase. The key operating characteristic of dynamic fulfillment is that the allocation of orders to DCs is dynamically optimized based on the realized demand and inventory levels to minimize overall fulfillment costs. At the beginning of a period, before demand is realized, the inventory level at DC j is replenished up to y_j (with zero lead time). We use $\omega \in \Omega$ to denote demand scenarios with Ω being the sample space. For each scenario ω, we use $D_i(\omega)$ denote the realized demand at customer location i. After scenario $\omega \in \Omega$ is observed, shipment quantities, denoted by $w_{ij}(\omega)$ are optimized to minimize the total shipping, lost sales shortage and inventory holding costs (the unit costs are denoted by d_{ij}, p_i h_j, respectively). It can be shown that, when demand and cost parameters are stationary (and independent) over time, a stationary base stock policy is optimal. Thus, it is sufficient to consider the expected cost of a single period in the management phase, which will be equivalent the long-term average costs for an infinite-horizon problem.

The problem of online retailer supply chain design problem [ORSCD] can be formulated as the following stochastic integer program:

$$
\begin{aligned}
[\text{ORSCD}]: \min \quad & \sum_{j\in J} f_j X_j + \sum_{i\in I}\sum_{j\in J} c_{ij} Z_{ij} + \sum_{j\in J} h_j y_j + E\left[\sum_{i\in I} p_i D_i(\omega) \right. \\
& \left. \sum_{i\in I}\sum_{j\in J} (d_{ij} - p_i - h_j) w_{ij}(\omega)\right] && (4.25)\\
\text{s.t.} \quad & X_j \geq Z_{ij} \text{ for } i \in I, j \in J && (4.26)\\
& \sum_{i\in I} w_{ij}(\omega) \leq y_j \text{ for } j \in J, \omega \in \Omega && (4.27)\\
& \sum_{j\in J} w_{ij}(\omega) \leq D_i(\omega) \text{ for } i \in I, \omega \in \Omega && (4.28)\\
& w_{ij}(\omega) \leq D_i(\omega) Z_{ij} \text{ for } i \in I, j \in J, \omega \in \Omega && (4.29)\\
& X_j \in \{0,1\} \text{ for } j \in J \\
& Z_{ij} \in \{0,1\} \text{ for } i \in I, j \in J \\
& y_j \geq 0 \text{ for } j \in J \\
& w_{ij}(\omega) \geq 0 \text{ for } i \in I, j \in J, \omega \in \Omega.
\end{aligned}
$$

In the problem formulation, the objective (4.25) is to minimize the design-phase fixed costs of locating facilities and constructing arcs plus the expected value of the inventory, shipping and shortage costs. The firm first selects a subset of candidate sites at which to locate DCs and assigns retailers to these DCs. Constraints (4.26) require that a retailer cannot be connected to a DC unless the latter is opened. The opened DCs order from the supplier with ample capacity and receive shipments with zero lead time. After replenishments arrive, demand realizations at retailers (i.e., ω) are observed. Given the information on realized demand at retailers and inventory availability at DCs, the flows from DCs to retailers are optimized to minimize total cost of the current period minus the salvage value of the leftover inventory at the end of the horizon, subject to the supply (4.27) and demand (4.28) flow balance constraints. Finally, constraints (4.29) stipulate that flows are only allowed on arcs that are constructed in the network design phase.

In the next section, we will discuss the solution procedure proposed by Mak (2012), which allows the problem to be solved efficiently.

4.5.2 Lagrangian Relaxation Algorithm

The Lagrangian relaxation algorithm proposed by Mak (2012) makes use of the network recourse decomposition (NRD) technique proposed by Powell and Cheung (1994). The main idea behind the technique is to find a piecewise linear (in the inventory levels y) and separable (by candidate DC j) approximation to the recourse function value. This is done by Lagrangian relaxation and decomposition of the recourse problem into subproblems that can be solved efficiently.

Most solution techniques proposed in the literature (e.g., for the [S-SCD] problem) for solving stochastic facility location problems involve discretization of the sample space into discrete and disjoint scenarios (see Snyder, 2006, for a review). Such an approach is desirable when the modeler can identify a relatively small number of scenarios with managerial significance. When scenarios are used to model random parameters that follow discrete distributions (or sampled from continuous distributions), the number of scenarios required and the resulting problem size can be enormous. For example, if there are 50 retailers each with 10

possible levels of demand, 10^{50} scenarios (a practically infinite number) are required to completely characterize the multi-variate distribution. As a heuristic approach, one may consider only a manageable sample of (say 1000) scenarios and formulate a deterministic equivalent problem. An alternative is to employ the NRD method to be discussed below, which allows us to approximate the recourse function value efficiently even without the use of scenarios. This approach proves a promising technique for obtaining approximate solutions to this class of stochastic programming problems.

By relaxing constraints (4.28) and (4.29), and imposing penalty multipliers ζ_i and η_{ij}, respectively, the resulting problem is decomposable by candidate DC locations j. Note that we use the same multiplier vectors η and ζ for all $\omega \in \Omega$. While this may yield weaker lower bounds than allowing the multiplier to vary with ω, we shall see that doing so enables us to solve the subproblems efficiently. Computational results from Powell and Cheung (1994) show that the approximations obtained with the NRD technique provide good lower bounds to the recourse function values. For a particular $j \in J$, the Lagrangian subproblem is given by:

$$\begin{aligned} V_j(\eta,\zeta) = \min \quad & \sum_{i\in I}(c_{ij} - \eta_{ij}E[D_i(\omega)])Z_{ij} \\ & + E\left[\sum_{i\in I}\hat{d}_{ij}w_{ij}(\omega) + h_j y_j\right] \qquad (4.30) \\ \text{s.t.} \quad & 0 \le w_{ij}(\omega) \le D_i(\omega), \text{ for each } i \in I, \omega \in \Omega \\ \text{where:} \quad & \hat{d}_{ij} = d_{ij} - p_i + \zeta_i + \eta_{ij}. \end{aligned}$$

Given nonnegative Lagrangian multipliers η and ζ, a lower bound to [ORSCD] is given by:

$$L(\eta,\zeta) = \sum_{j\in J}\min\{V_j(\eta,\zeta) + f_j, 0\} + \sum_{i\in I}\{(p_i - \zeta_i)E[D_i(\omega)]\}. \quad (4.31)$$

The subproblem (4.30) is a newsvendor problem with multiple customers and heterogeneous underage costs. When demands are integer-valued, it can be efficiently solved by the procedure proposed by Cheung and Powell (1996). It is possible to apply similar reasoning for continuous distributions. In particular, if the demands are normally distributed, we can obtain the optimal objective value to (4.30) as follows.

Proposition 4.4. Let $\hat{I} = \{i \in I | \hat{d}_{ij} \leq h_j\}$ and let $m = |I|$. Sort the retailers in set $\hat{I}$ in increasing order of $\hat{d}_{ij}$, i.e., $\hat{d}_{1j} \leq \hat{d}_{2j} \leq ... \leq \hat{d}_{mj} \leq h_j$. Furthermore, let $\hat{d}_{m+1} = h_j$. Then, the optimal objective value of (4.30) is given by:

$$V_j(\eta, \zeta) = \sum_{i \in I} \min\{c_{ij} - \eta_{ij} E[D_i(\omega)], 0\} + Q_j(y_j^*) \tag{4.32}$$

where:

$$\begin{aligned} Q_j(y_j) &= h_j y_j - \sum_{i=1}^{m} (\hat{d}_{i+1,j} - \hat{d}_{ij}) \int_0^{y_j} \bar{\Phi}\left(\frac{s - \mu(i)}{\sigma(i)}\right) ds \\ \mu(i) &= E\left[\sum_{k=1}^{i} D_k\right] \\ \sigma(i) &= \sqrt{Var\left(\sum_{k=1}^{i} D_k\right)} \end{aligned}$$

and y_j^* is the solution to the equation:

$$h_j = \sum_{i=1}^{m} (\hat{d}_{i+1,j} - \hat{d}_{ij}) \bar{\Phi}\left(\frac{y_j^* - \mu(i)}{\sigma(i)}\right). \tag{4.33}$$

Proof. Let $\rho(s, i)$ be the probability that the s-th unit of stock at DC j is shipped to retailer i and $\hat{\rho}(s)$ be the probability that the unit is not shipped out. Then, the expected recourse function, i.e., the value of the expectation in (4.30) given $y_j = S$, is given by:

$$Q_j(y_j) = \int_0^{y_j} \sum_{i \in I} \left[\hat{d}_{ij} \rho(s, i) + h_j \hat{\rho}(s)\right] ds.$$

The optimal value is then given by $\min_{y_j \geq 0} Q_j(y_j)$. Note that for $i \in I$ where $\hat{d}_{ij} > h_j, \rho(s, i) = 0$ for all s. Sort the remaining retailers (denoted by $\hat{I}$) in increasing order of $\hat{d}_{ij}$, i.e., $\hat{d}_{1j} \leq \hat{d}_{2j} \leq ... \leq \hat{d}_{mj} \leq h_j$. Then, the probabilities are given by:

$$\begin{aligned} \rho(s, i) &= P\left(\sum_{k=1}^{i} D_k \geq s\right) - P\left(\sum_{k=1}^{i-1} D_k \geq s\right) \\ \hat{\rho}(s) &= P\left(\sum_{k=1}^{m} D_k < s\right). \end{aligned}$$

As $\mathbf{D}$ is normally distributed, the probabilities are given by:

$$\begin{aligned} P\left(\sum_{k=1}^{i} D_k \geq s\right) &= 1 - \Phi\left(\frac{s-\mu(i)}{\sigma(i)}\right) = \bar{\Phi}\left(\frac{s-\mu(i)}{\sigma(i)}\right) \\ \text{where: } \mu(i) &= E\left[\sum_{k=1}^{i} D_k\right] \\ \sigma(i) &= \sqrt{Var\left(\sum_{k=1}^{i} D_k\right)}. \end{aligned}$$

Knowing the above and letting $\hat{d}_{m+1} = h_j$, we may express $Q_j(y_j)$ as follows:

$$Q_j(y_j) = h_j y_j - \sum_{i=1}^{m} (\hat{d}_{i+1,j} - \hat{d}_{ij}) \int_0^{y_j} \bar{\Phi}\left(\frac{s-\mu(i)}{\sigma(i)}\right) ds.$$

It is easy to show that $Q_j(y_j)$ is convex and the optimal y_j^* can be found by solving for the first-ion, which order condition (4.33).

Since the (w, y) terms and the Z terms in (4.30) are not linked, they can be optimized separately. Let the optimal inventory level y_j found by solving (4.33) be y_j^*, then, the optimal value of the expected value term in (4.30) is given by $Q_j(y_j^*)$. Moreover, the optimal Z values can be easily found by setting Z_{ij} to 1 if its cost coefficient is negative and 0 otherwise, which yields (4.32). □

Note that equation (4.33) is a generalization of the critical fractile condition for the standard newsvendor problem. It can be solved efficiently using numerical methods (e.g., Newton's method). Using Proposition 4.4, it is possible to obtain a lower bound to [ORSCD] given a set of dual multipliers (η, ζ). An important point to note is that, using our Lagrangian relaxation procedure, each iteration involves solving one subproblem (4.30) for each j separately. In practice, for decentralized decision making, different locations $j \in J$ can be managed by separate divisions of the firm, or even by separate companies, e.g., if the firm outsources its distribution operations in different locations to different third-party logistics service providers. Based on our decomposition approach, we make the following observation:

Remark 4.1. The Lagrangian relxation algorithm requires decision makers managing locations in J to solve problem (4.30). In each iteration, each location j reports to a central *coordinator* the optimal values of $E[w_{ij}(\omega)]$. The coordinator will then update the multipliers ζ_i and communicate the new values to the locations.

Remark 4.1 suggests that location managers only need to communicate with the central coordinator by exchanging partial information, but not with each other. This provides an effective coordination mechanism for decentralized supply chains, in which individual location managers do not necessarily have the incentive to cooperate, and in many cases, may even compete with each other. The job of the coordinator is to update the dual multipliers ζ_i, which can be interpreted as the prices per unit of expected shipment to retailer i. An interesting point is that not only the inventory control decisions can be decentralized, but the network design decisions as well. Furthermore, note that as the dual multipliers η_{ij} are associated with constraints (4.29) that are specific to each location j, updating of multipliers η_{ij} can be done by each location j independently.

For the coordinator's problem of updating average unit prices ζ_i, the standard subgradient procedure (e.g., Algorithm 2) is not applicable, as it requires a procedure to estimate tight upper bounds given information from the lower bound solutions. For our problem, the recourse function value, i.e., the expected value term in [ORSCD], needs to be evaluated in order to find a feasible solution or an upper bound. While approximate methods for doing so exist in the literature (e.g., Cheung and Powell, 1996; Harrison and Van Mieghem, 1999), it is not practical to embed any of these as a subroutine and run it for large number of iterations (in the order of thousands) since doing so would significantly increase computation time. Therefore, Mak (2012) uses an alternative subgradient procedure that guarantees convergence without requiring an upper bound, known as the variable target value method (see Sherali *et al.* (2000) for details). To obtain a cost estimate of a feasible solution (not necessarily an upper bound), we fix the network structure to the lower bound solution (i.e., X and Z variables) and solve the resulting restricted problem. This can be done by a distributionally-

robust optimization approximation (see, for example, Goh and Sim, 2010).

4.5.3 Discussion

In the solution algorithm, we impose state-independent Lagrangian multipliers, i.e., they take on the same values across all demand and disruption realizations. While the exact recourse value for our stochastic program can be obtained by using state-dependent multipliers, the number of such multipliers to be determined would equal the number of possible states of the system which is infinite with a continuous demand distribution. Therefore, we use state-independent Lagrangian multipliers for practical reasons. Similar techniques based on Lagrangian relaxation of weakly-coupled recourse problems have been utilized in the literature to approximate recourse functions of stochastic programming problems. Recent papers by Kunnumkal and Topaloglu (2008) and Topaloglu and Kunnumkal (2006) apply similar techniques to approximate the value functions in stochastic dynamic programming. Such techniques provide computationally efficient approximations for the recourse function values that would otherwise require computationally expensive discretization methods. Furthermore, under the decentralized decision making regime discussed in Remark 4.1, this allows the coordination prices ζ_i to be deterministic, and thus reducing the information sharing requirement between the individual locations and the central coordinator.

The dynamic fulfillment operations entail interesting considerations in the network design strategy. For instance, under the threat of disruptions of DCs (see Section 4.4), dynamic fulfillment enables both risk pooling (inventory sharing among demand sources) and risk diversification (placement of smaller inventory volumes at more locations to better contain disruption damage) simultaneously. Mak and Shen (2012) evaluate the optimal design of the supply chain network under such a scenario. Furthermore, to evaluate the strategic implications of dynamic shipment configurations on optimal network design, Lim *et al.* (2016) employ a continuous approximation approach to analyze the strategic factors and identify several insights to the contrary of previous studies. We discuss this next in Section 4.6.

4.6 Analytical Study on Effects of Inventory Sharing on Network Configuration

In recent years, firms have taken advantage of advancements of information technology to improve supply chain operations via enhancing operational flexibility. The dynamic fulfillment strategy for online retailers discussed in Section 4.5 is exemplary of this movement (see, for example, the case of Amazon discussed in Xu *et al.* (2009)). In offline businesses, similar examples are also becoming commonplace. IBM utilizes a "neighborhood" stocking strategy, in which an order may be allocated to any stocking location within a certain radius of the customer location, for its service parts logistics. This strategy is estimated to save $5 million per year in costs, without jeopardizing customer service (Gresh and Kelton, 2003). For its import supply chain, Dell performs diversion of in-transit inventory to respond to dynamically evolving demand forecasts (Foreman *et al.*, 2010).

As discussed in Section 4.5, this class of dynamic and adaptive network operations give rise to makeovers in fundamental network design strategies. Particularly, these enhance the agility of supply chain networks, i.e., their ability to quickly adapt to changes in operating environments. Throughout Section 4, one common planning uncertainty considered in most of the models we have discussed is in customer demand volumes. In such contexts, a key element of supply chain agility is responsively allocating inventory to match with orders so as to fulfill demand with as high probability and as low cost as possible. Classical findings in the literature of inventory theory state that inventory sharing (i.e., risk pooling, see Eppen (1979)) is an effective means to maintain customer service while lowering inventory holding costs in supply chain networks. Lim *et al.* (2016) perform an analytical study to evaluate how agile modes of supply chain operations enhance the inventory sharing capabilities of supply chain networks, and, as a result, how they impact the optimal network design. We shall discuss their model and major findings in this section.

Particularly, Lim *et al.* (2016) focus on the interactions between the optimal network design and two modes of inventory sharing. The first is the conventional *physical pooling* strategy of using a centralized stocking

location (a DC) to fulfill demand from multiple sources. The second is the dynamic fulfillment strategy, as discussed in Section 4.5, which allows inventory to be shared among multiple DCs with *informational pooling*. In terms of network design strategy, physical pooling is enhanced when the network is more consolidated, i.e., there are fewer facilities each handling a larger demand volume; whereas informational pooling is more depended upon when the network is deconsolidated, i.e., there are more facilities each handling a smaller demand volume. While it is intuitive that both modes of inventory pooling help reduce inventory costs, the balance between the two requires careful balance of various cost factors, such as fixed facility investment (and operating) costs and transportation costs.

4.6.1 Problem Description and Model

Lim *et al.* (2016) consider a firm deciding to locate a number of DCs to serve a number of geographically-dispersed customer demand points in a large service region. Following the CA approach outlined in Section 2.3, we consider the demand points to be uniformly distributed over the plane with homogeneous density δ. Each demand point has a normally-distributed random demand with mean μ and standard deviation σ. As a result of spatial homogeneity, it is sufficient to consider the density of DCs, or equivalently, the distance X between adjacent DCs, rather than the specific locations. We consider distances under the $\ell - 2$ metric. For a demand point, we refer to the nearest DC to it as its primary DC, and reciprocally, we say that the demand point is in said DC's primary influence area. Each DC holds an inventory level given by the mean plus Z_1 times the standard deviation of the aggregated demand of customers within its influence area (i.e., Z_1 is the safety stock factor). Physical pooling of inventory occurs at each DC, as the inventory held at each DC is primarily used to serve the aggregated demand from all demand points within the primary influence area.

Besides physical pooling, the DCs share inventory with each other using dynamic fulfillment. In particular, if a DC stocks out, the excess demand in its primary influence area can be re-routed to nearby DCs. We consider a neighborhood sharing strategy, in which every M^2 DCs,

in a square-shaped configuration (see Figure 4.2), form an inventory sharing group within which such re-routing is possible. To model such operations, we consider that DCs replenish inventory under periodic review with zero lead time. When demand is realized, a linear program is solved within each inventory sharing group to minimize the overall costs of shipping from DCs to demand points and shortage penalty. Similar to the case discussed in Section 4.5, we consider the expected costs in a single period that reflects the long-run average performance of the network. Lim *et al.* (2016) derive the major cost components of operating the network as follows. Note that the number of DCs to be located is a decision variable, and the overall service region is assumed to be very large. Thus, the model considers the average costs per demand point.

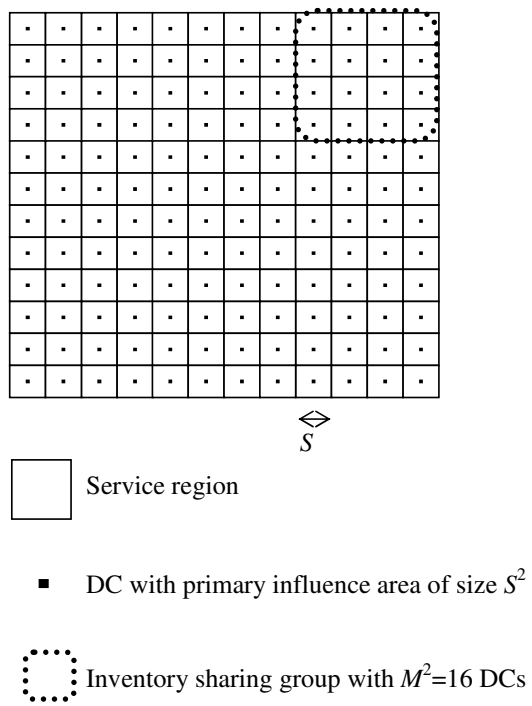


Figure 4.2: Segment of Service Region with Square-shaped Primary Influence Areas and Inventory Sharing Groups

Cost Components

The first major cost component is the fixed location and operation costs of DCs. We consider the firm to incur an annualized cost of f to locate (finance) and operate each DC. Because there are δX^2 demand points in the primary influence area of a DC, the average fixed location costs per demand point is given by $\frac{f}{\delta X^2}$. Note also that it is straightforward to include variable costs of DC operations, but Lim *et al.* (2016) find that such terms would not affect the analysis and thus we omit them for brevity throughout this discussion.

Second, we discuss the inventory related costs, which include the holding costs and shortage penalty costs. First, because we consider a periodic review model under a base stock policy, we omit the costs of holding cycle stocks as they only depend on average demand volume but not the decision variables, and focus on safety stock costs. With a safety stock factor of Z_1, each DC holds a safety stock level of $Z_1\sqrt{\delta X^2\sigma^2}$, incurring a holding cost of $hZ_1\sqrt{\delta}\sigma X$. Pro-rated by δX^2 demand points, the average holding cost per demand point is $\frac{h\sigma Z_1}{\sqrt{\delta}X}$.

Under the dynamic fulfillment arrangement, demand can be fulfilled from any DC within the same inventory sharing group. Conversely, shortage (backorders assumed) is incurred only when all DCs in the same inventory sharing group stock out concurrently. Therefore, the expected amount of backorders of an inventory sharing group is given by $E[D_2 - M^2\delta X^2\mu - Z_2\sqrt{\delta}M\sigma X]^+$, where D_2 denotes the random aggregate demand within the inventory sharing group. With normally-distributed demand, the expected shortage is given by $(Z_2)M\sqrt{\delta}\sigma X$, where $L(z)$ is the standard normal loss function $L(z) = \phi(z) - z\bar{\Phi}(z)$; and $\bar{\Phi}(\cdot)$ and $\phi(\cdot)$ are the complementary cumulative distribution function and density function of the standard normal distribution, respectively. Dividing among the δM^2X^2 demand points in the group, the average expected backorder costs per demand point is given by $\frac{p\sigma L(Z_2)}{\sqrt{\delta}XM} = \frac{p\sigma Z_1}{\sqrt{\delta}X}\frac{L(Z_2)}{Z_2}$.

Next, we discuss the modeling of distribution costs. Unlike models with static sourcing (i.e., in which the DC that serves a customer is fixed as a long-term decision), the dynamic fulfillment model considers multiple DCs within the same inventory sharing group as candidates

to serve a demand point. Under our two-stage setting, the shipments from DCs to demand points within an inventory sharing group can be optimized after demand quantities are realized, by solving a linear program. Lim *et al.* (2016) utilize dimensional analysis to develop an analytical expression for the expected distribution costs, i.e., the expected optimal objective value of said linear program, as we discuss below.

Consider an inventory sharing group with M^2 DCs. For tractability, Lim *et al.* (2016) partition the influence area of each DC into K subzones of equal areas and define an aggregate demand point at the centroid of each subzone to represent the aggregated demand of demand arising in the subzone. As a result, there are KM^2 aggregate demand points, each with a normally distributed demand with mean $\hat{\mu} = \frac{\delta X^2}{K}\mu$ and standard deviation $\hat{\sigma} = \frac{\sqrt{\delta}X}{\sqrt{K}}\sigma$. Lim *et al.* (2016) consider $K = 4$, and also find that the choice of more refined partitions (i.e., larger K) would not lead to significant changes in the resulting expected cost expression.

Upon realization of demand at the aggregate demand points, the problem of optimizing shipment quantities (denoted by w_{ij} from DC j to demand point i) and backorder levels (denoted by b_j for DC j) can be formulated as follows:

$$\begin{aligned}
\min_{w,b} \quad & \sum_{j=1}^{M^2}\left[\sum_{i=1}^{KM^2} d_{ij}w_{ij} + \bar{d}b_j\right] \\
\text{s.t.} \quad & \sum_{i=1}^{KM^2} w_{ij} \le y_j + b_j \ , \text{ for } j = 1, ..., M^2 \\
& \sum_{j=1}^{M^2} w_{ij} \ge e_i, \text{ for } i = 1, ..., KM^2 \\
& w_{ij} \ge 0 \ , \text{ for } i, j = 1, ..., M^2.
\end{aligned}$$

In the above, $y_j = \delta X^2\mu + Z_1 X\sqrt{\delta}\sigma$ denotes the inventory level held at DC j, e_i denotes the realized demand at demand point i, d_{ij} denotes the distance between demand point i and DC j, and $\bar{d}$ is constant larger than the maximum distance between any DC and any demand point (such that backorders are more costly than any shipments within the group). We are interested in the total transportation distance (in item-

miles), excluding the backorder penalty. Let $T = \sum_{i=1}^{KM^2} \sum_{j=1}^{M^2} d_{ij} w_{ij}^*$, where w^* denotes the optimal values of the w variables in the above linear program. Then, the expected shipping distances are given by $\bar{T} = \frac{E[T]}{M^2}$, taken over the probability distribution of demand. Note that $\bar{T}$ has dimension item-miles.

Lim *et al.* (2016) apply the following dimensional analysis argument. First, the problem involves two dimensions: items and miles, together with six variables: the independent variables $\hat{\mu}$ (in items), $\hat{\sigma}$ (in items), X (in miles), Z_1 (dimensionless), M (dimensionless), and the dependent variable $\bar{T}$ (in item-miles). Hence, we obtain the following dimensional matrix:

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{bmatrix}$$

with $rank(\mathbf{A}) = 2$. By Theorem 3.5, there can be four independent π-groups. Lim *et al.* (2016) consider the following groups: $\frac{\bar{T}}{\hat{\mu}X}$, M, Z_1, and $\frac{\hat{\sigma}}{\hat{\mu}}$. By the π-theorem, the relationship between the π-groups can be written in the form of:

$$\frac{\bar{T}}{\hat{\mu}X} = \mathfrak{f}\left(M, Z_1, \frac{\hat{\sigma}}{\hat{\mu}}\right), \quad \text{or equivalently} \quad \bar{T} = \hat{\mu}X\mathfrak{f}\left(M, Z_1, \frac{\hat{\sigma}}{\hat{\mu}}\right).$$

Lim *et al.* (2016) identify the unspecified function $\mathfrak{f}(\cdot)$ approximately by regression over extensive simulation data:

$$\bar{T} = 1.43\hat{\mu}X + 0.329\hat{\sigma}L(Z_1)X\log_2 M. \tag{4.34}$$

The choice of this approximation is justified as follows. First, in the backorder setting, the average realized demand in the influnece area of a DC ($4\hat{\mu}$) will need to be shipped for a distace no shorter than that from the primary DC (which can be shown to be $0.357X$). This corresponds to the first term in (4.34). Second, after attempting to fulfill as much demand as possible from the primary DCs, there will be some shortage at certain demand points (as well as surplus inventory at certain DCs). The amount of shortage (per aggregate demand point) is proportional to $\hat{\sigma}L(Z_1)$. The average shipping distance to fulfill such shortages corresponds to the expected cost of a transportation problem with random demand and supplies, which has been proved by Daganzo and Smilowitz (2004) to increase in $n\log_2 n$ as n, the

number of demand/supply points (M^2 in this case), goes to infinity. Dividing among the DCs in the group, the average (per DC) expected transportation cost to fill unfulfilled demand is then proportional to $\hat{\sigma} L(Z_1) \log_2 M$.

Based on this approximation, Lim *et al.* (2016) propose the following approximation:

$$\begin{aligned}&\text{Shipping distance per inventory sharing group}\\ &\quad = k_1 \delta \mu X^3 + k_2 \sqrt{\delta} \sigma X^2 L(Z_1) \log_2 M,\end{aligned}$$

where $k_1 = 0.357$ and $k_2 = 0.082$ are dimensionless constants. Observing that the two terms reflect the shipping costs from primary and secondary DCs, respectively, Lim *et al.* (2016) allow the unit costs for these two types of shipments (u and v, respectively) to be possibly different. This allows the possibility of capturing additional costs of shipping from secondary DCs, such as additional administrative costs or expedited shippnig costs to recover for the increased shipping distance.

Optimization Problem Formulation and Analysis

Integrating the cost components discussed above, we may formulate the supply chain design problem as:

$$\begin{aligned}&\min_{X,Z_1,Z_2} \mathfrak{C}(X, Z_1, Z_2)\\ = &\min_{X,Z_1,Z_2} \frac{f}{\delta X^2} + \frac{h\sigma Z_1}{\sqrt{\delta} X} + \frac{p\sigma Z_1}{\sqrt{\delta} X} \frac{L(Z_2)}{Z_2} + u\mu k_1 X\\ &+ \frac{v k_2 \sigma}{\sqrt{\delta}} L(Z_1) \log_2 \left(\frac{Z_2}{Z_1}\right).\end{aligned} \tag{4.35}$$

The first term constitutes the fixed location costs. The second and third reflect the inventory-related (i.e., holding and backorder penalty) costs. The fourth and fifth terms correspond to the shipping costs. The cost function involves three decision variables that characterize the supply chain design strategy: the separation between adjacent DCs X, which determines the *density* of DCs, the local safety stock factor Z_1, which determines the intensity of physical pooling, and the group safety stock factor Z_2, which determines the intensity of informational pooling. Therefore, by analyzing the properties of the optimal solution

(X^*, Z_1^*, Z_2^*), one can obtain insights into the interactions between the strategic choices of network density and pooling arrangements.

We first discuss a result characterizing the optimal choice of pooling modes given fixed X.

Lemma 4.1. The optimal solution to (4.35), for given X, is given by the unique solution to the following equations:

$$h = vk_2 X \log_2 M \bar{\Phi}(Z_1) + p\bar{\Phi}(Z_2), \tag{4.36}$$

$$Z_1 = R^{-1}\left(\frac{\ln 2\, p}{vk_2 X}\frac{\phi(Z_2)}{Z_2}\right), \tag{4.37}$$

where $R(z) = L(z)/z$.

It is interesting to note that the first optimal condition (4.36) exhibits a similar structure as the critical fractile condition in the newsvendor problem. In particular, the optimal stocking levels at DCs should balance the opportunity costs of over-stocking and under-stocking at the margin. The marginal over-stocking cost is given by the additional holding cost h for the marginal unit of safety stock. The marginal under-stocking cost is given by the expression on right hand side of (4.36): when the primary DC stocks out (with probability $\bar{\Phi}(Z_1)$), the marginal cost incurred is given by the unit shipping cost from a secondary DC ($vK_2 X \log_2 M$); and when the entire inventory sharing group stocks out (with probability $\bar{\Phi}(Z_2)$), the marginal cost incurred is the backorder penalty. The second optimality condition (4.37) characterizes the optimal balance between safety stock factors Z_1 and Z_2. It can be shown that, under this condition, the optimal inventory sharing group size M $(= Z_2/Z_1)$ is decreasing in Z_1. This indicates that as each DC stocks more safety stock, which enhances physical pooling, it is optimal to reduce the intensity of informational pooling by downsizing the inventory sharing groups. In other words, the two inventory sharing modes work as substitutes.

Using Lemma 4.1 and further optimizing over X, Lim *et al.* (2016) characterize the optimal solution as follows.

Proposition 4.5. The optimal solution to (4.35) is given by the unique solution to (4.36), (4.37), and

$$\mu u k_1 = \frac{2f}{\delta X^3} + \frac{\sigma Z_1}{\sqrt{\delta} X^2}\left(h + pR(Z_2)\right). \tag{4.38}$$

The third optimality condition (4.38) highlights the fundamental trade-off in the choice of DC network density. In particular, it is optimal to strike a balance between transportation costs, reflected by the term on the left of the equality sign, and the facility costs and inventory-related (including backorder penalty) costs, reflected by the terms on the right hand side. Further, as in the conventional models, it is notable that transportation costs considerations favor a higher density (i.e., shorter shipping distances to customers), while facility cost considerations favor the opposite, due to economies of scale. The new perspective in the model is the consideration of inventory and shortage costs involving the two inventory sharing modes. It is not immediately obvious whether the optimal balance between the two inventory sharing modes would favor a denser or less dense network configuration. Further scrutiny into this issue leads to the following observation.

Proposition 4.6. The impacts of the two modes of inventory sharing on network design are as follows:

$$\frac{\partial^2}{\partial X \partial Z_1}\mathfrak{C}(X, Z_1, Z_2) \leq 0 \quad \text{and} \quad \frac{\partial^2}{\partial X \partial Z_2}\mathfrak{C}(X, Z_1, Z_2) \geq 0.$$

In particular, the first item in Proposition 4.6 suggests that fixing Z_2, a larger value of Z_1, i.e., higher intensity of physical pooling, favors a larger X, i.e., lower network density. This is due to the fact that, consistent to findings from conventional models, physical pooling works best when each DC handles a larger influence area such that more demand sources are pooled. The second item suggests that the opposite holds for Z_2. That is, a higher intensity of informational pooling favors a higher network density. This is caused by the fact that the informational pooling can be performed at lower (shipping) costs if the DCs are spaced closer together, such that customers are, on average, closer to *secondary* DCs. Overall, Proposition 4.6 illustrates the opposing effects of the two inventory sharing modes on the optimal network configuration. The interactions between the two help explain some interesting contrasts between the optimal network design strategy when both inventory sharing modes are active and in conventional models where only physical pooling is present, as we discuss below.

Contrasts with Conventional Results

Based on the analysis of the model (4.35), Lim *et al.* (2016) find that well-known results from the conventional literature assuming physical pooling may fail to hold. In particular, they raise two results from "conventional wisdom" and show how they break down under the new model. The first conventional wisdom is that

If the inventory service level becomes higher, it is optimal to decrease the density of DCs.

This result was found by Shen *et al.* (2003) in the study of the [SCD] model, in which physical pooling causes an economies of scale effect such that it is beneficial to centralize larger demand volumes at a smaller number of DCs. As service level increases, the weight of such benefits are enhanced, and thus the optimal network density decreases. The same qualitative finding is also confirmed in subsequent research (e.g., Naseraldin and Herer, 2008). The same holds for the CA model discussed in Section 2.3 (see Table 2.1). Interestingly, Lim *et al.* (2016) show that this managerial insight does not hold when informational pooling is also opositpresent.

Proposition 4.7. When the penalty cost parameter p increases, the optimal value of X in (4.35) decreases.

Proposition 4.7 shows that, when the network design is jointly optimized in the presence of both forms of inventory sharing, the aforementioned conventional wisdom is reversed. Note that, as service level is endogenously determined in the model, the proposition is stated with regard to an exogenous increase in penalty cost p. While this result appears counterintuitive, it can be explained by the interactions between the two sharing modes as characterized by Proposition 4.6. In particular, while an increase in penalty cost encourages inventory sharing, the net effect on optimal density depends on which sharing mode takes command, as they have opposite effects on optimal density. Proposition 4.7 suggests that the effect of informational pooling dominates that of physical pooling, resulting in a net increase in optimal network density.

Next, we discuss a classical insight regarding transportation costs.

If the unit transportation cost for shipping from DCs to customers becomes higher, it is optimal to decrease the density of DCs.

This claim holds for the UFL problems and its extendions, including the [SCD] problem and its variants. With an increase in unit transportation cost, the benefits of locating more DCs to reduce shipping distances are enhanced, overcoming the additional location costs. In the case of inventory-location problems with physical pooling, the same directional effect occurs despite the associated increase in inventory costs due to the pooling of smaller demand volumes at individual DCs. Interestingly, we find that the effect of such change in the presence of informational pooling is less straightforward.

Proposition 4.8. When u increases, the optimal X value in (4.35) decreases. In addition, when v increases, the optimal X value in (4.35) increases.

Proposition 4.8 suggests that the effects of changes in unit transportation cost parameters u and v, corresponding to shipments from primary and secondary DCs, respectively, take opposite directions. For shipments from primary DCs, an increase in unit transportation cost (u) causes the optimal network density to increase, consistent with conventional results. However, for shipments from secondary DCs enabled under informational pooling, an increase in unit transportation cost (v) leads to a decrease of optimal network density, i.e., increase in shipping distances. This is due to the interactions between the two inventory sharing modes - when informational pooling becomes costly to perform due to increased shipping costs, it is optimal to strategically enhance physical pooling, favoring the location of fewer but larger DCs.

Furthermore, the analysis reviewed in Section 4.6 indicate that the form of inventory sharing can have profound effects on the optimal supply chain design strategy. Lim *et al.* (2016) argue that there is a need for refinement of conventional understanding developed based on classical models to account for emerging supply chain operations in the new era. With the rapid integration of supply chain management with

developments such as data-driven e-commerce and big data analytics, various strategic questions remain to be addressed. We concur that this is a promising area for future research.

4.7 Discussion

The integrated modeling approach originated from supply chain modeling. Since then, most applications can be found in the supply chain domain. While many basic problems can be studied, there remains substantial opportunity for further research. Much of the existing studies, such as the ones reviewed in Sections 4.1 and 4.3, consider an expanded breadth of facility operational features, such as capacity and multi-commodity issues. Studies such as the models reviewed in Sections 4.4 and 4.5, on the other hand, extend the standard models along the time dimension by considering two-stage stochastic optimization problems. We believe that one promising area for future research is to further extend this line of problems to multiple time periods, to capture the strategic dynamic expansion and diffusion of facility networks under stochastic demand. Traditionally, dynamic facility location problems have been mainly studied in deterministic settings, mainly because of the intractability of stochastic multiple stage problems. With the recent developments in methodologies such as approximate dynamic programming (e.g., Powell, 2007), it is now promising to study the optimal path of facility deployment over time. Using such methodologies, an interesting technical aspect would be to formulate approximate value functions for future deployment and operations of facility networks as a function of facilities currently (and previously) located. Careful application of modeling techniques that have been employed to model various operational features (e.g., the use of nonlinear models) could potentially be useful.

5

Applications in Emerging Areas

In recent years, researchers have expanded the scope of integrated modeling to applications beyond the supply chain domain. Location planning problems arising in the health care, transportation, energy, retail and service sectors all exhibit distinctive problem characteristics that require careful modeling and analysis. Thanks to rapid recent development in modeling and optimization techniques, state-of-the-art models have been proposed and studied in various cutting edge problems. In this chapter, we provide a review of some examples of recent work in these emerging research areas. These examples cover the topics of sustainable transportation, renewable energy, retail strategy and health care, all of which are themes that have drawn considerable interest in the broader operations research/ management science community. Through this, we hope to convey the message that facility location research, despite being a traditional area of study, still has important contributions to make (as well as challenges to tackle) going forward in some of the cutting edge areas.

5.1 Infrastructure Planning for Electric Vehicles

In this section, we review an application of the integrated modeling methodology in designing the support infrastructure system for electric vehicles (EVs) with battery switching. With rising concerns over energy security and greenhouse gas emissions, EVs have been seen as one of the most commericially viable green alternatives to conventional gasoline cars. Both the public and private sectors have poured in massive investments to promote adoption of EVs. For example, the China and US Federal governments have committed $15 billion and $2.4 billion, respectively, in grants to support EV technology, manufacturing and infrastructure developments. The Nissan-Renault Alliance, an active player in EV development, is reported to have spent over $4 billion in EV-related research and development.

One of the most critical drawbacks of the current EV technology is the short autonomous driving range of the vehicles. Nissan Leaf, the best-selling EV model worldwide, has a range of about 105 miles before it needs to be recharged. Combined with the relatively long recharging times, in the order of 20-30 minutes for a 80% charge using the fastest commercial chargers today, the range limitation poses significant inconvenience to potential buyers of EVs. To match the convenience of gasoline cars, which can be conveniently refueled at gas stations within five minutes, the technology of battery swapping was proposed by firms such as Better Place (an EV-infrastructure operator) and Tesla (an EV manufacturer). Using this technology, an EV can be refueled by mechanically replacing the battery, which only takes less than two minutes, at specialized facilities known as battery swapping stations. The swapped-out battery can then be recharged at the battery swapping station and later be used for another EV.

However, to fully match the convenience of gasoline cars, the network of battery swapping stations must provide a high level of accessibility, comparable with the very dense network of gas stations. This entails the important planning problem for firms such as Better Place and Tesla in determining locations for the battery swapping stations. These facilities are highly capital intensive, costing up to $2 million apiece to deploy, due to the high cost of the battery swapping equipment. Furthermore,

as each battery may cost $8,000 to $10,000 to purchase, the cost of capital involved to equip all swapping stations with sufficient batteries is very significant.

The planning problem of locating battery swapping stations in a transportation network encompasses two major factors. First, EV adoption is still at an infacy stage, and thus adoption and usage patterns are highly uncertain. This calls for planning approaches that provide robust solutions in light of the inherent planning uncertainty, particularly where there is no or little data available regarding EV driving patterns of the region in question. Second, the operating characteristics of charging stations, particularly with regard to inventory requirements of spare batteries and their charging needs, carry subtle impact on the selection of locations. Taking into account both problem features, Mak *et al.* (2013) propose an integrated optimization model to select a set of locations that minimizes facility (land, installation and operating) and expected inventory (for holding batteries at stations) costs.

5.1.1 Basic Model

We shall review the swapping station location model proposed by Mak *et al.* (2013). Consider a transportation network consisting of a set P of travel paths. To cover travel needs, the swapping stations must provide coverage of segments of travel paths longer than half the range of the EV, such that round trips can be completed on such segments. Let Q be the collection of such subsegments (referred to as "subpaths"); we let $b_{pq} = 1$ if subpath q ($\in Q$) is a part of path p ($\in P$), and 0 otherwise. Consider a set of candidate facility locations J along the transportation network. Let $a_{jq} = 1$ if candidate location j ($\in J$) is located along subpath q ($\in Q$), such that a swapping station at j can cover EV swapping demand traveling along q. Let f_j denote the fixed cost of locating a swapping station at j, and h be the cost of holding a battery in stock at a swapping station. Furthermore, let g_j be the maximum capacity of batteries that can be stocked at location j. This capacity limit is given by the minimum of the physical capacity on storage space, and the power load capacity that limits the number of batteries that can be recharged in parallel at the same location.

We define the binary decision variable X_j to indicate whether a station is located at j ($X_j = 1$) or not ($= 0$). Further, we use $Z_{jq} = 1$ ($Y_{jp} = 1$) to indicate that a swapping station at j covers swapping demand along subpath q (path p). Furthermore, let $I_j(\mathbf{Y})$ denote the required number of spare batteries to be stocked at station j to serve the collection of paths as indicated by the decision variables $\mathbf{Y}$. As demand is uncertain, we consider $I_j(\mathbf{Y})$ to be a random variable, the explicit formulation of which will be discussed later. Then, the cost-minimization model for battery swapping station location (abbreviated as [BSL]) can be formulated as follows:

$$[\text{BSL}] : \min \sum_{j \in J} (f_j X_j + hE\left[I_j(\mathbf{Y})\right]) \tag{5.1}$$

$$\begin{aligned}
\text{s.t.} \quad Y_{jp} &\geq b_{pq} Z_{jq} \text{ for } j \in J, p \in P, q \in Q && (5.2)\\
\sum_{j \in J} a_{jq} Z_{jq} &\geq 1 \text{ for } q \in Q && (5.3)\\
Y_{jp} &\leq X_j \text{ for } j \in J, p \in P && (5.4)\\
P\left(I_j(\mathbf{Y}) \leq g_j\right) &\geq 1 - \epsilon_g \text{ for } j \in J && (5.5)\\
X_j &\in \{0,1\} \text{ for } j \in J\\
Y_{jp} &\in \{0,1\} \text{ for } j \in J, p \in P\\
Z_{jq} &\in \{0,1\} \text{ for } j \in J, q \in Q. && (5.6)
\end{aligned}$$

The objective function (5.1) is to minimize the cost of locating swapping stations and the expected cost of equiping them with sufficient batteries to meet swapping demand. Constraints (5.2) require that, if subpath q is to be covered by a station at j, then the station at j serves swapping demand for all travel paths p that include q as a subsegment. Constraints (5.3) require that all subpaths be covered by some station, and (5.4) stipulate that a station must be located at j if the location is used to cover swapping demand on any path p. Chance constraints (5.5) require that the number of batteries to be required at a swapping station at j to not exceed capacity g_j with high enough probability $1 - \epsilon_g$. Note that, because swapping demand is uncertain, the number of batteries required is a random variable, and thus constraints (5.5) take the form of chance constraints. In the next subsection, we discuss how

these chance constraints and the expected inventory cost term in (5.1) can be expressed in tractable formulations, by modeling the operating and charging characteristics of swapping stations.

5.1.2 Operating Characteristics of Swapping Stations

Model [BSL] is an extension of the set covering problem discussed in Chapter 1. In particular, if the battery inventory cost terms in (5.1) and the battery charging capacity constraints (5.5) are removed, the decision variables $\mathbf{Y}$ will become redundant, and the problem reduces to one of determining the minimum cost to locate facilities to ensure that all subpaths in set Q are covered. However, as the costs of carrying batteries at swapping stations are very significant, it is imperative to fully capture their implications on optimal location design by delving into the operational characteristics of battery swapping.

Battery Swapping Considerations

We first obtain the expected inventory cost in (5.1) by developing a model for battery swapping operations. Suppose the arrivals of demand for swapping on a travel path p at some observation point follows a Poisson process with rate λ_p (which may be uncertain to the planner). We also consider the arrival processes associated with different paths to be independent. As swapping demand on different paths are directed to some swapping station, the arrival process of EVs requiring swaps will be the superposition of the Poisson processes associated with the individual paths, and is thus a Poisson process. In particular, for a station at j, the arrival rate of swapping demand is given by $\lambda_j = \sum_{p \in P} \lambda_p Y_{jp}$.

When a battery is unloaded at a swapping station upon a swap, it is recharged and then later swapped onto another incoming EV. When the station holds an inventory of multiple batteries, we assume that batteries are reused in first-in, first-out (FIFO) order for simplicity. Under the FIFO order, an unloaded battery can be recharged until all other batteries at the station are used, and the next swap request arrives. That is, if there are I_j batteries at station j, a battery can be recharged for an amount of time equal to the sum of I_j consecutive

interarrival periods. We consider a service requirement that at least α (> 0.5) proportion of swapping requests are fulfilled by batteries that have been recharged for at least t time units. Note that this service requirement is equivalent to requiring that there are fewer than I_j EV arrivals within t time periods with probability of at least α.

As the arrival process of EVs at station j is Poisson with rate λ_j, the number of EV arrivals within a time period of t follows a Poisson distribution with mean $\lambda_j t$. Using the normal approximation for the Possion distribution, the number of batteries needed to satisfy the service requirement is given by:

$$I_j = t\lambda_j + \Phi^{-1}(\alpha)\sqrt{t\lambda_j} = t\sum_{p\in P}\lambda_p Y_{jp} + \Phi^{-1}(\alpha)\sqrt{t\sum_{p\in P}\lambda_p Y_{jp}}, \quad (5.7)$$

where $\Phi^{-1}(\cdot)$ is the inverse cumulative distribution function of the standard normal distribution.

If λ_p is precisely known to the planner, then (5.7) can be directly substituted in (5.1) (without the expectation operator as the quantity is deterministic) and (5.5) (where the chance constraint would become deterministic). Then, following the treatment in Section 3.2, one can replace Y_{jp} in (5.7) with Y_{jp}^2 to convert the resulting objective function (5.1) and capacity constraint (5.5) in SOCP form.

Planning Uncertainty

In practice, however, it is not always reasonable to consider demand rates to be known precisely. EV infrastructure planning suffers from a well-known *chicken-and-egg* problem: consumers do not adopt (in mass scale) unless the infrastructure is in place; while the planner suffers from planning uncertainty and may even be unwilling to invest in infrastructure without observing the demand pattern. In this section, we discuss how to tackle uncertainty in the expectation and probability terms of (5.1) and (5.5), respectively.

The overaching difficulty in the EV planning context is the lack of data. In particular, the key planning uncertainty lies in the values of the swapping demand rates λ_p, which depend on usage patterns as well as adoption levels of EVs. While travel survey data, e.g., the California

Household Travel Survey (California Department of Transportation, 2010), can be used to forecast usage patterns of EVs, one has to bear in mind that such data are collected on conventional vehicle types which may exhibit different usage patterns than EVs. Similarly, forecasts of adoption levels, at the relatively early stage of market penetration, may exhibit significant errors. Therefore, we attempt to develop decision models that support decisions that are *robust* with respect to both planning uncertainty and possible mis-specification, or *ambiguity*, of data. The modeling framework is drawn from the distributionally-robust optimization literature (e.g., Goh and Sim, 2010).

We begin with the following *model of uncertainty*. Let λ_p be given by a linear function of a number of mutually independent random factors, $\tilde{z}_l, l = 1, ..., L$, known as the *primitive uncertainties*, i.e.,

$$\lambda_p = \sum_{l=1}^{L} \hat{\lambda}_{pl} \tilde{z}_l, \tag{5.8}$$

where $\hat{\lambda}_{pl}, l = 0, ..., L$ are known constants. Furthermore, we assume that the precise distribution of $\tilde{\mathbf{z}} = (\tilde{z}_1, ..., \tilde{z}_L)$ is not known; instead, only the means, supports, and variances of $\tilde{z}_l, l = 1, \cdots, L$ are given. The family of possible joint distributions with the given descriptive statistics is given by $\mathbb{F}$, which is assumed to be non-empty. These primitive uncertainties can represent factors such as regional EV adoption factors, path specific flow rates, etc. One may estimate these descriptive statistics based on data from household travel surveys, pilot studies, expert judgment, or a combination thereof. The distributionally-robust optimization framework allows for ambiguity in the specification of random parameters $\tilde{\mathbf{z}}$, by optimizing over the worst expectation (or probability) over the family $\mathbb{F}$. One advantage of employing this framework is the less onerous data burden on estimating these descriptive statistics compared with fitting the full distribution of uncertain parameters.

Following the distributionally-robust optimization framework, in (5.1) and (5.5), we consider $\sup_{F \in \mathbb{F}} E_F [I_j(\mathbf{Y})]$ and $\inf_{F \in \mathbb{F}} P (I_j(\mathbf{Y}) \leq g_j)$, respectively. For the former, Mak *et al.* (2013) derive the exact value by obtaining the worst-case distribution explicitly. However, the exact worst-case expectation is not computationally tractable when embedded

into the optimization formulation. Therefore, they provide the following bounds:

Proposition 5.1. [Proposition 2 of Mak *et al.* (2013)] Suppose $\hat{\lambda}_{pl}$ is nonnegative for each $p \in P, l = 1, ..., L$. Moreover, $\tilde{z}_l$ has nonnegative support for each $l = 1, ..., L$. Let $a_l = \frac{\bar{z}_l}{\mu_l}$ and $b_l = \frac{\sigma_l}{\mu_l}$ for $l = 1, ..., L$. Let $a = \min_{l=1,...,L}\{a_l\}$, $b = \max_{l=1,...,L}\{b_l\}$, $a' = \max_{l=1,...,L}\{a_l\}$ and $b' = \min_{l=1,...,L}\{b_l\}$. Then, the following upper bound holds:

$$\sup_{\mathbb{P}\in\mathbb{F}} E_{\mathbb{P}}\left[\sqrt{\sum_{p\in P}\lambda_p Y_{jp}}\right] \leq \bar{\Psi}\sqrt{\sum_{p\in P}\sum_{l=1}^{L}\hat{\lambda}_{pl}\mu_l Y_{jp}^2}.$$

If $a \geq b^2 + 1$, the following lower bound holds:

$$\sup_{\mathbb{P}\in\mathbb{F}} E_{\mathbb{P}}\left[\sqrt{\sum_{p\in P}\lambda_p Y_{jp}}\right] \geq \underline{\Psi}\sqrt{\sum_{p\in P}\sum_{l=1}^{L}\hat{\lambda}_{pl}\mu_l Y_{jp}^2},$$

where:

$$\underline{\Psi} = \sqrt{a} - \frac{a-1}{\sqrt{a} + \sqrt{1 - \frac{b^2}{a-1}}}, \quad \bar{\Psi} = \sqrt{a'} - \frac{a'-1}{\sqrt{a'} + \sqrt{1 - \frac{b'^2}{L(a'-1)}}}.$$

Using the upper bound provided in Proposition 5.1, one can provide a conservative (with respect to ambiguity in demand rates) estimate of expected battery cost. By evaluating its value against the lower bound, Mak *et al.* (2013) find that the upper bound is very tight under practical parameter settings. Integrating this battery cost expression into [BSL], we may replace the objective function (5.1) with

$$\min \sum_{j\in J}(f_j X_j + hV_j)$$

and add the constraints

$$V_j \geq \sum_{p\in P}\sum_{l=1}^{L} t\hat{\lambda}_{pl}\mu_l Y_{jp} + \Phi^{-1}(\alpha)\bar{\Psi}\sqrt{\sum_{p\in P}\sum_{l=1}^{L} t\hat{\lambda}_{pl}\mu_l Y_{jp}^2}\ , \text{ for each } j \in J. \tag{5.9}$$

Note that (5.9) is in computationally-tractable SOCP form. Therefore, if one can transform (5.5) into linear or SOCP constraints, the

whole problem [BSL] can be transformed into a computationally-efficient MISOCP. Mak *et al.* (2013) show that this indeed can be done, by making use of the following result:

Proposition 5.2. [Proposition 4 of Mak *et al.* (2013)] The chance constraints (5.5) are equivalent to:

$$\inf_{\mathbb{P}\in\mathbb{F}} P_{\mathbb{P}}\left(\sum_{p\in P}\sum_{l=1}^{L}\hat{\lambda}_{pl}\tilde{z}_l Y_{jp} \le \hat{g}_j\right) \ge 1-\epsilon_g \text{ for each } j\in J,$$

$$\text{where } \hat{g}_j = \frac{\left(\sqrt{g_j + \Phi^{-1}(\alpha)^2/4} - \Phi^{-1}(\alpha)/2\right)^2}{t} \text{ is a constant.} \quad (5.10)$$

Proposition 5.2 transforms a chance constraint on a nonlinear expression into one on a linear expression. Then, using the results of Chen *et al.* (2007), one can obtain tight, distributionally-robust bounds on the new chance constraints (5.10) in SOCP form. Therefore, combining Propositions 5.1 and 5.2, we can tightly approximate [BSL] in MISOCP form. Mak *et al.* (2013) show that this formulation is computationally tractable, and thereby investigate various infrastructure design questions regarding charging and battery technological development and standardization of batteries for different car models.

The swapping station network design problem is an example of how similar modeling approaches developed for supply chain problems can be useful in emerging applications. In particular, the modeling of inventory costs for swapping stations is largely analogous to that for distribution centers discussed in earlier chapters. In the next section, we show another example where related modeling ideas can also be helpful in developing models for renewable energy planning as well.

5.2 Deployment of Energy Storage Devices in the Electric Grid

The generation of electricity accounts for about 1/3 of greenhouse gas emissions in US (US Environmental Protection Agency, 2014). With rising concerns over environmental footprint of economic activities, the electricity sector, together with the transportation sector (Section 5.1), are among the major domains for clean technology deployment. With

incentives policies such as feed-in tariffs (e.g., Alizamir *et al.*, 2016), Cohen *et al.*, 2015 and mandates such as renewable portfolio standards, substantial investments in renewable generation, such as solar and wind power, have been taking place worldwide. While these modes of power generation produce negligible greenhouse gas emissions in operation (i.e., after manufacturing and installation), they suffer from the intermittent nature of weather-dependent supply. Therefore, supply of power is highly stochastic and difficult to be matched with the demand profile, leading to costly operations such as curtailment of surplus supply (e.g., Wu and Kapuscinski, 2013) and running of costly (and emission-generating) spinning reserves to guard against low supply (e.g., Bitar *et al.*, 2012).

In operations management, the popular strategy to help mitigate supply-demand mismatch risks is to hold buffer inventory. In power systems, inventory can be held using energy storage (ES) devices such as batteries, flywheels (at smaller scale) or pumped hydroelectric storage (at larger scale). ES systems are typicall very costly to install and manage; therefore, their capacities and locations (both with respect to physical geography and the power grid topology), as well as the associated investments in transmission lines (i.e., connection to the grid) must be carefully planned for. Qi *et al.* (2015) study the problem of jointly planning the location and capacity of ES systems as well as the associated transmission line design for wind power. Their optimization model, which we shall review in this section, incorporates operational characteristics such as intermittence, spatial correlation and the variability pooling effect.

Qi *et al.* (2015) consider a set of wind farms distributed over a geographical region. Each wind farm generates power and is to transmit to an assigned junction site, at which the pooled power from multiple farms is further transmitted to a common substation. The problem is to optimally determine (1) the wind farm to junction site assignments, (2) the deployment (and capacities) of ES units at junction sites, and (3) the capacities of transmission lines connecting wind farms to junction sites, and from the latter to the substation. The objective is to minimize the sum of investment costs for ES and transmission capacity, and the lost revenue due to efficiency loss of charging and discharging of ES

units and curtailment when ES units have insufficient capacity to store or transmit generated power.

5.2.1 Uncapacitated Storage Problem

Following Qi *et al.* (2015), we begin the discussion by considering a simplified problem where ES units have unlimited capacity and in which their locations are to be optimized together with the capacities and topology of transmission lines. We first discuss the modeling of the operations of one single wind farm connected to an ES unit. This will serve as a building block for the network optimization problem involving multiple wind farms.

Consider a wind farm that produces a random amount of power[2], $\tilde{w}_t$ in period t, which is assumed to be uniformly distributed in $[\mu - \rho/2, \mu + \rho/2]$ (following Kim and Powell (2011)). In the uncapacitated problem, we consider relatively long planning periods, where δ is in the order of months or a year. The generated power is transmitted to the substation through a transmission line with capacity C. Any surplus power that cannot be transmitted is stored in the ES unit, with charging and discharging efficiency factors of α and β, respectively. In this simplified model, we first assume that the ES unit has infinite capacity. Therefore, the amount of energy loss in period t is given by $\tilde{l}_t = \max\{(\tilde{w}_t - C)(1 - \alpha\beta), 0\}$, with expectation of:

$$E[\tilde{l}_t] = \int_C^{\mu+\rho/2} (w - C)(1 - \alpha\beta)\frac{1}{\rho} dw = \frac{1 - \alpha\beta}{2\rho}(\mu + \rho/2 - C)^2.$$

At the planning stage, the transmission capacity C is to be determined. We consider a linear cost structure where the cost to invest in capacity C is given by qC. Then, the (annualized) investment and energy loss (with unit cost of p) cost is given by:

[2] As we consider periods of equal lengths δ, we use the terms power and energy interchangeably.

$$\begin{aligned} v_1(C) &= pE[\tilde{l}_t]/\delta + qC \\ &= \frac{p(1-\alpha\beta)}{2\rho\delta}C^2 + \left(q - \frac{p(1-\alpha\beta)}{\rho}(\mu+\rho/2)\right)C \\ &\quad + \frac{p(1-\alpha\beta)}{2\rho\delta}(\mu+\rho/2)^2. \end{aligned} \tag{5.11}$$

The transmission capacity is determined to minimize $v_1(C)$, subject to the constraint that $C \geq \mu$, i.e., the line must have sufficient capacity to transmit the average power output. Then, it is straightforward to see that the optimal capacity is given by $C^* = \max\{\mu, \mu + \rho(\frac{1}{2} - \theta)\}$, where $\theta = \frac{\delta q}{p(1-\alpha\beta)}$. Note that θ is a parameter that depends only on the characteristics of the ES unit but not on the wind farm. The optimal cost is given by:

$$v_1^* = \begin{cases} q\mu + \rho q(\frac{1}{2} - \frac{\theta}{2}), & \text{if } \theta < \frac{1}{2} \\ q\mu + \rho\frac{q}{8\theta}, & \text{otherwise.} \end{cases} \tag{5.12}$$

Using the above as a building block, we consider a more general problem with multiple wind farms. Specifically, we consider the problem of selecting among a set J of candidate junction sites to be used as junction sites to serve a set of wind farms I. Because ES systems are expensive, it is economical that they are deployed only in a subset of J, and each of them may serve multiple wind farms. That is, it is possible to use sites in J only as junction sites without building ES units (i.e., these sites will only re-route power to the substation without storage). Where appropriate, we append subscripts i and j to the notation introduced previously for the single wind farm problem to indicate the dependence on wind farm $i \in I$ and ES site $j \in J$. We use binary decision variables V_j and X_j to indicate whether site $j \in J$ is selected as a junction site without storage ($V_j = 1$) or not ($V_j = 0$) and whether an ES system is built at the site ($X_j = 1$) or not ($X_j = 0$). The construction costs for the transmission line connecting site j to the substation and for the ES unit are given by g_j and h_j, respectively. Furthermore, we use binary variables Y_{ij} to indicate whether ES site j is connected to wind farm i ($Y_{ij} = 1$) or not ($Y_{ij} = 0$). Similarly, we use Z_{ij} to indicate whether junction site (without ES) j is connected to wind farm i. To make such a connection in either case, a fixed cost of g_{ij} will be incurred to build the transmission line.

Power from wind farms connected to the same junction site j will be pooled, and the aggregate power is modeled by the random variable $w_{t,j}$ (for period t), which is assumed to follow a uniform distribution in $[\mu_j - \rho_j/2, \mu_j + \rho_j/2]$. With multiple wind farms, it is important to model the spatial correlation of wind power at different locations. Let $\mathbf{\Sigma}$ be the symmetric $|I| \times |I|$ matrix, whose entries are $\Sigma_{ik} = \rho_i \rho_k \sigma_{ik}$ where σ_{ik} is the correlation coefficient between the outputs of wind farms i and k[3]. Then, matching the first two moments, one can obtain:

$$\begin{aligned} \mu_j &= \sum_{i \in I} \mu_i Y_{ij} \\ \rho_j^2 &= \sum_{i \in I} \rho_i^2 Y_{ij} + \sum_{i,k \in I, i \neq k} \rho_i \rho_k \sigma_{ik} Y_{ij} Y_{kj} = \mathbf{Y}_j' \mathbf{\Sigma} \mathbf{Y}_j \end{aligned}$$

(Note that $Y_{ij} = Y_{ij}^2$).

In the single wind farm problem, we derived the optimal transmission and curtailment costs at a storage site:

$$v_j^* = \begin{cases} q_j \mu_j + \rho_j q_j (\frac{1}{2} - \frac{\theta_j}{2}), & \text{if } \theta_j < \frac{1}{2} \\ q_j \mu_j + \rho_j \frac{q_j}{8\theta_j}, & \text{otherwise.} \end{cases} \tag{5.13}$$

Recall that θ_j only depends on characteristics of site (and ES unit) j. Thus, one can partition J into subsets $J_1 = \{j \in J | \theta_j < \frac{1}{2}\}$ and $J_2 = J \setminus J_1$ before determining the assignments, i.e., the $\mathbf{Y}$ variables. Therefore, the overall optimization problem can be formulated as follows:

$$\begin{aligned} \min \quad & \sum_{j \in J} \left[h_j X_j + \sum_{i \in I} g_{ij}(Y_{ij} + Z_{ij}) + g_j(X_j + V_j) \right] \\ & + \sum_{j \in J} \left[\sum_{i \in I} q_j(\mu_i + \rho_i/2)(Y_{ij} + Z_{ij}) + q_j \left(\sum_{i \in I} \mu_i Z_{ij} + \frac{1}{2} \bar{P}_j \right) \right] \\ & + \sum_{j \in J_1} \left[q_j \sum_{i \in I} \mu_i Y_{ij} + q_j (\frac{1}{2} - \frac{\theta_j}{2}) \underline{P}_j \right] \\ & + \sum_{j \in J_2} \left[q_j \sum_{i \in I} \mu_i Y_{ij} + \frac{q_j}{8\theta_j} \underline{P}_j \right] \end{aligned} \tag{5.14}$$

[3]Notice that, because the standard deviation of $\tilde{w}_{t,i}$ is equal to $\rho_i/\sqrt{3}$, $\mathbf{\Sigma}$ is a scalar multiple of the covariance matrix of the random vector $\{\tilde{w}_{t,i}\}$.

$$\text{s.t.} \quad \sqrt{\mathbf{Y}_j' \boldsymbol{\Sigma} \mathbf{Y}_j} \leq \underline{P}_j \text{ for } j \in J \tag{5.15}$$

$$\sqrt{\mathbf{Z}_j' \boldsymbol{\Sigma} \mathbf{Z}_j} \leq \bar{P}_j \text{ for } j \in J \tag{5.16}$$

$$\sum_{j \in J} (Y_{ij} + Z_{ij}) = 1 \text{ for } i \in I \tag{5.17}$$

$$X_j + V_j \leq 1 \text{ for } j \in J \tag{5.18}$$

$$Y_{ij} \leq X_j, Z_{ij} \leq V_j \text{ for } i \in I, j \in J \tag{5.19}$$

$$X_j, V_j, Y_{ij}, Z_{ij} \in \{0, 1\} \text{ for } i \in I, j \in J.$$

In the above, the objective function (5.14) is to minimize the investment costs for ES systems and fixed construction costs for transmission lines (first line), the variable capacity costs for tranmission lines (second line), and the expected energy loss from charging and curtailment (third line). The constraints (5.15) and (5.16) ensure that the variables $\underline{P}_j$ and $\bar{P}_j$ give the values of ρ_j where site j is used as a junction site with and without ES system, respectively. Note that the objective function is increasing in both $\underline{P}_j$ and $\bar{P}_j$, and thus these inequality constraints hold at equality at the optimal solution. Constraints (5.17) ensure that each wind farm is connected to a junction site. Constraints (5.18) stipulate that a site can be selected as a junction site with or without ES unit, but not both. Finally, constraints (5.19) require that a junction site be deployed for it to serve any wind farm. Note that the overall formulation is an MISOCP, which can be solved by commercial solvers such as CPLEX.

5.2.2 Capacitated Storage Problem

In practice, ES units are expensive to install and thus it is typically undesirable to over-invest in capacity. Thus, the model described in Section 5.2.1, which assumes ample storage capacity of the ES units, may lead to suboptimal investment strategies. Qi *et al.* (2015) extend the uncapacitated model to account for ES capacity considerations. This formulation is developed based on an upper bound (i.e., a conservative approximation) on the expected energy overflow costs incurred when generated power cannot be stored due to insufficient storage capacity. Again, we first discuss the formulation based on a simple network with a single wind farm and one ES unit.

When power storage is capacitated, we need to take a more refined perspective on wind power output and charging/discharging operations. Based on the diurnal pattern of wind speed, Qi *et al.* (2015) consider refined time periods of length δ_b, in the order of hours. These time intervals are chosen such that the wind intensity of different periods can be assumed independent. Let $\tilde{s}_{b,\tau}$ denote the amount of energy stored in the ES unit at the end of period τ, during which $\tilde{w}_{b,\tau}$ units of power is produced, and S be the maximum storage capacity of the ES unit. Similarly as before, $\tilde{w}_{b,\tau}$ is assumed to be uniformly distributed in $[\mu_b - \rho_b/2, \mu_b + \rho_b/2]$. Depending on the level of charge at the start of period τ, given by $\tilde{s}_{\tau-1}$, the level of charge at the end of τ can be one of the following:

1. If $\tilde{w}_{b,\tau} > C$, i.e., energy production exceeds transmission capacity and excess energy needs to be stored in the ES unit, then $\tilde{s}_\tau = \min\{\tilde{s}_\tau + \alpha(\tilde{w}_{b,\tau} - C), S\}$;

2. If $\tilde{w}_{b,\tau} \leq C$, i.e., energy production is short of transmission capacity, and energy is discharged from the ES unit to fill the shortfall, then $\tilde{s}_\tau = \max\{\tilde{s}_\tau - \frac{1}{\beta}(C - \tilde{w}_{b,\tau}), 0\}$.

Therefore, $\tilde{s}_\tau$ is given by a piecewise linear function of the random variable $\tilde{w}_{b,\tau}$ as well as $\tilde{s}_{\tau-1}$. Considering the difficulty to identify the distribution of $\tilde{s}_\tau$, Qi *et al.* (2015) propose an upper bound on the expected amount of power overflow, $E[\tilde{o}_\tau] = E[\tilde{s}_\tau + \alpha(\tilde{w}_{b,\tau} - C) - S]^+$, and thus the resulting (opportunity) cost. This bound is derived based on the assumptions that (1) the capacity S is large relative to $\tilde{w}_{b,\tau}$ and C such that the $P(\tilde{w}_{b,\tau} = S \text{ and } \tilde{w}_{b,\tau+1} = 0) = P(\tilde{w}_{b,\tau} = 0 \text{ and } \tilde{w}_{b,\tau+1} = S) = 0$; and (2) the density function of $\tilde{s}_\tau$, $f_s(s)$ is decreasing in $s \in (0, S)$ when $C > E[\tilde{w}_\tau]$. Under these assumptions, Qi *et al.* (2015) prove that approximating $f_s(\cdot)$ with a uniform density, $\hat{f}_s(\cdot)$, over $[0, S]$ yields a closed form upper bound on the expected volume of power overflow.

Proposition 5.3. Assume $C \geq E[\tilde{w}_{b,\tau}]$. Then, the expected volume of overflow are bounded above by the following:

$$E_{\hat{f}_s}[\tilde{o}_\tau] \quad = \quad \frac{5\alpha}{24S}(\mu_b + \rho_b/2 - C\delta_b)^2.$$

Using Proposition 5.3, one can approximately solve for the optimal storage capacity by solving $\min_S pE_{\hat{f}_s}[\tilde{o}_t] + rS$, which yields $S^*(C) = \sqrt{\frac{5p\alpha}{24r\delta\delta_b}}(\mu_b + \rho_b/2 - C\delta_b)$ and the resulting cost $v_2(C) = \sqrt{\frac{5pr\alpha}{6\delta\delta_b}}(\mu_b + \rho_b/2 - C\delta_b)$. Combining this with the transmission overflow and construction cost $v_1(C)$ in (5.11), one can obtain the (approximate) optimal transmission line capacity as well as the resulting costs:

$$v_4(C) = \begin{cases} \frac{p(1-\alpha\beta)}{2\rho\delta}(\mu + \frac{\rho}{2} - C)^2 + qC + \sqrt{\frac{5pr\alpha}{6\delta\delta_b}}(\mu_b + \frac{\rho_b}{2} - C\delta_b) \\ \qquad\qquad \text{if } C \in \left[\frac{\mu_b}{\delta_b}, \frac{\mu_b+\rho_b/2}{\delta_b}\right) \\ \frac{p(1-\alpha\beta)}{2\rho\delta}(\mu + \frac{\rho}{2} - C)^2 + qC \quad \text{if } C \in \left[\frac{\mu_b+\rho_b/2}{\delta_b}, \mu + \frac{\rho}{2}\right]. \end{cases}$$

Note that $v_4(C)$ is convex in C, as it is the pointwise maximum of two convex (quadratic) functions. Then, the optimal C^*, as well as the resulting costs, can be obtained and expressed in closed form.

Unfortunately, this single capacitated wind farm model can not be readily incorporated in a multiple wind farm network optimization problem as in the uncapacitated case. This is because the value of $v_4(C^*)$, the capacity-related costs at a junction site, is dependent on characteristics of wind farms (unlike the case in (5.13) where θ_j does not depend on characteristics of individual wind farms), while assignment of wind farms to junction sites are determined endogenously. Therefore, it is not possible to jointly optimize the ES system and transmission network design directly. Qi *et al.* (2015) propose a heuristic based on solving the uncapacitated problem and adjusting for ES unit capacities subsequently. In particular, they propose first solving the uncapacitated problem optimally. Then, for each ES site chosen in the uncapacitated solution, one solves for the expected variable costs $v_4(C^*)$ of investing in the optimal ES capacity, and compare the resulting costs with that of investing in a transmission line with maximum capacity $(\mu_j + \rho_j/2)$. The option with lower costs is selected. Their computational results indicate that this heuristic performs very well.

A common theme exhibited in both the EV infrastructure planning and wind power storage examples is the inherent planning uncertainty

due to uncertain adoption of EVs or uncertain supply of wind power. However, these examples have not touched on the idea of deferring part of the location decisions until uncertainty is (partially) resolved by deploying facilities in phases. The next application to be discussed in the next section explores this possibility in the retail context.

5.3 Retail Expansion with Demand Learning

The supply chain design problems discussed so far involves the location of back-end or support facilities such as warehouses and DCs. On the other hand, location decisions are also crucial for the planning of front-end facilities, such as retail stores. Locations are one of the most influential factors in consumers' store choice because they usually prefer to go to the most conveniently located stores. Consequently, firms with a well-chosen set of store locations often thrive by developing sustainable competitiveness based on locational advantages. While the store location decision may offer such positive contributions, it also represents great risk since it involves a significant commitment of resources for a long period of time. Similar to the case of back-end facilities, poor store location decisions can negatively affect the firm's performance for an extended period of time due to the strategic nature of these decisions.

The risk of commitment in the retail industry intensifies where the firm enters a new market it is less familiar with, such as a foreign country. The inherent factors that determine demand, such as economic conditions and consumer behaviors, may not be well known to the firm at the stage of planning the network of stores. Typically, firms employ pilot testing and other means for marketing research to learn the market prior to entry. However, such studies typically do not completely resolve the uncertainties, and thus, firms still face substantial risk of deploying its network of stores suboptimally due to lack of complete demand knowledge. To mitigate such risks, firms often deploy stores in multple phases dynamically. This way, the firm learns from the operations of stores to learn the demand characteristics in earlier phases, and uses such knowledge to make better plans in later phases.

Interestingly, in practice, we observe that firms exhibit varying degrees of aggresiveness as to how quickly stores are deployed over time.

For example, Apple chose to expand cautiously in China because of the high uncertainty due to the already existing competitors (such as Lenovo and HP) and local copycat manufacturers (Wall Street Journal, 2011a). Since opening the first store in 2008, it has only recently added two new stores in 2011. In contrast, CVS Pharmacy is expanding aggressively in the Puerto Rican market, opening its first 9 stores all in 2010 (Providence Business News, 2010) and scheduled to open 13 additional stores by the end of 2012 (Puerto Rico Daily Sun, 2011).

In this section, we consider the retail store location problem where the uncertain market characteristic can be learned from store operations over time. The objective is to study how the firm should optimally devise its store deployment and expansion strategy taking into account such learning opportunities. The fundamental trade-off in designing the optimal expansion path is one between *active learning* and *deferred commitment*. In particular, should the firm deploy more stores at the early stage in order to learn the market faster, or should it defer a large portion of its investments until late stages when demand becomes more certain, so as to avoid the risk of making overly-aggressive investments? The choice between these strategies carries profound effects on the firm's long-term performance.

We adapt a model proposed by Bhatti *et al.* (2015), which focuses on an alternative fuel station service network planning problem, to consider the case of general retail network planning. This multiple-stage model captures the uncertainty of demand and the market learning effect. We consider the consumer adoption rate of the market to be uncertain, but can be learned over time. In each stage, we assume that the firm acquires more information on the adoption rate and has the option to locate additional facilities. To reflect the trade-off between active learning and deferred commitment, we let the amount of information acquired to be endogenously determined as a function of firm's action in the previous stages. Therefore, the firm can choose to either shorten the market learning time by aggressively investing upfront or defer the commitment at the cost of slowing down market learning.

5.3.1 Dynamic Retail Location Model with Demand Learning

We consider a market modeled as a network $G(V, A)$ with the set of vertices V and the set of arcs A. Let $V = \{I \cup J\}$ where $I = \{1, \cdots, m\}$ is the set of demand points and $J = \{1, \cdots, n\}$ is the set of candidate facility locations for stores. In each demand node i, we assume h_i number of potential consumers live. To capture the uncertainty in consumer adoption, we introduce a random variable θ and refer this to *consumer adoption rate*, assumed to be identical across the network of the target market (i.e., independent of i). We further assume each consumer adoption yields one demand per unit time. Therefore, θh_i is the product demand per unit time in each node i.

In the retail context, the covering objective can be used to model the relationship between the firm's location choice and consumers' decision to patronize stores, which determines demand. In particular, the tendency of consumers to patronize a store (i.e., demand coverage of the store) is decreasing in the distance between the demand node and the store. For a demand node $i \in I$, we consider a fraction of $g_i(d) \in [0, 1]$ of demand to be covered by its closest store located d units of distance from it. Similar to Berman and Krass (2002) and Berman *et al.* (2003), we consider the coverage function $g_i(d)$ to be non-increasing and convex, with $g_i(0) = 1$ for all $i \in I$. Then, with the set of open stores indicated by decision variables X, we can express the effective demand covered at node i by $\theta h_i g_i(d_i(X))$ where $d_i(X) = \min_{j \in X} d(i, j)$, which is the distance from the nearest opened store to i. Denoting the revenue per unit demand per unit time by r, the total revenue per unit time is $\sum_{i \in I} r\theta h_i g_i(d_i(X))$. We consider an infinite time horizon over which the discount rate α is applied.

We consider two decision epochs in the model. In the first, the firm selects a set of store locations (indicated by X^1) before learning the consumer adoption rate. Therefore, θ is known to be following a certain probability distribution. Upon deploying the set of stores, the firm accumulates demand knowledge through operating these stores. After a time of T, referred to as the *market learning time*, the uncertainty is resolved and the value of θ is known to the firm precisely. Then, the firm determines another set of additional stores (indicated by X^2) to locate,

and the retail network is operated over an (discounted) infinite horizon. It is further assumed that stores opened in the first phase cannot be closed or moved in the second phase, due to the high cost of doing so in practice. The key feature of the model is the endogenous learning time T, which is modeled as a function of first-stage decision X^1. Bhatti *et al.* (2015) consider the learning time as a function of first-stage demand coverage, defined as $c(X^1) = \sum_{i \in I} h_i g_i(d(X^1))$; particularly, they consider $T = \phi(c(X^1)) > 0$ where $\phi(\cdot)$ is a decreasing function. This setup captures the relationship that if more consumers are covered in the first stage, the market is learned at a faster rate. The two-stage decision model can be formulated as:

$$\max_{X^1 \subset J} \left\{ \mathbb{E}_\theta \left[\int_0^T e^{-\alpha t} \Big(\sum_{i \in I} r\theta h_i g_i(d_i(X^1)) - f(X^1) \Big) dt + \right. \right. \\ \left. \left. + e^{-\alpha T} V(X^2; X^1, \theta) \right] \right\} \tag{5.20}$$

where $V(X^2; X^1, \theta)$ is the optimal objective value of:

$$\max_{X^2 \subset J \setminus X^1} \left\{ \int_T^\infty e^{-\alpha t} \Big(\sum_{i \in I} r\theta h_i g_i(d_i(X^1 \cup X^2)) - f(X^1 \cup X^2) \Big) dt \right\}. \tag{5.21}$$

It is possible to re-express the above with a more compact formulation for the two-stage retail location problem with learning [TRLP-L]:

$$[\text{TRLP-L}]: \max_{\substack{X^1 \subset J \\ X^2 \subset J \setminus X^1}} \frac{1}{\alpha} \left\{ \mathbb{E}_\theta \left[(1 - e^{-\alpha T}) \sum_{i \in I} r\theta h_i g_i(d_i(X^1)) - f(X^1) \right. \right. \\ \left. \left. + e^{-\alpha T} \left(\sum_{i \in I} r\theta h_i g_i(d_i(X^1 \cup X^2)) - f(X^2) \right) \right] \right\}. \tag{5.22}$$

5.3.2 Solution Approach

Using the standard scenario-based formulation in two-stage stochastic programming, Bhatti *et al.* (2015) assume θ to be a discrete random variable with a set S of possible realizations (scenarios), each with probability p^s where $\sum_{s \in S} p^s = 1$. Then, [TRLP-L] can be reformulated

explicitly as:

$$[\text{P1}]: \quad \max_{\mathbf{X},\mathbf{Y},T} \sum_{s\in S} \frac{p^s}{\alpha} \Bigg[(1-e^{-\alpha T}) \sum_{i\in I}\sum_{j\in J} r\theta^s h_i g_i(d_{ij}) Y^1_{ij} - \sum_{j\in J} f_j X^1_j + e^{-\alpha T}\Bigg[\sum_{i\in I}\sum_{j\in J} r\theta^s h_i g_i(d_{ij}) Y^2_{ijs} - \sum_{j\in J} f_j X^2_{js}\Bigg]\Bigg] \tag{5.23}$$

$$\text{s.t.} \quad Y^1_{ij} \le X^1_j, \quad Y^2_{ijs} \le X^1_j + X^2_{js} \text{ for } i \in I, j \in J, s \in S \tag{5.24}$$

$$\sum_{j\in J} Y^1_{ij} = 1, \quad \sum_{j\in J} Y^2_{ijs} = 1 \text{ for } i \in I, s \in S \tag{5.25}$$

$$X^1_j + X^2_{js} \le 1 \text{ for } j \in J \text{ for } s \in S \tag{5.26}$$

$$T \ge \phi\Bigg(\sum_{i\in I}\sum_{j\in J} h_i g_i(d_{ij}) Y^1_{ij}\Bigg) \tag{5.27}$$

$$T \ge 0, \quad X^1_j, X^2_{js}, Y^1_{ij}, Y^2_{ijs} \in \{0,1\} \text{ for } i \in I, \forall j \in J, \forall s \in S.$$

The first and second lines of objective function (5.23) represent the time-discounted profit accumulated over the first and second stage, i.e., for the time intervals $[0, T]$ and (T, ∞), respectively. These consist of the operational revenue, less the cost of locating the sets of stores indicated by X^1 and X^2. The constraints (5.24) ensure that a facility must be opened if it is used to cover any customers. Constraints (5.25) stipulate that each customer location must be covered by one facility in each stage. Note that, if a demand node is too far from the facility covering it, the level of coverage can be zero. Constraints (5.26) make sure that a store is opened in at most one stage. Constraint (5.27), which we refer to as the coverage constraint, relate the learning time T with the first-stage coverage $c(X^1, Y^1) = \sum_{i\in I}\sum_{j\in J} h_i g_i(d_{ij}) Y^1_{ij}$.

The above formulation is a mixed integer nonlinear program. The major solution difficulty comes from the nonlinear (exponential) terms in the objective and the coverage constraint, which imposes the (generally nonlinear) relationship between the auxiliary variable T and the location and coverage variables (X^1, Y^1). Bhatti *et al.* (2015) propose a linearization-bounding procedure to solve the problem tractably while providing a performance bound. We outline the procedure below. For no-

tational brevity, we temporarily denote the first-stage demand coverage by just c and rewrite the coverage constraint as $T = \phi(c)$. First, we note that c will be bounded in the interval $[\underline{c}, \bar{c}]$ at the optimal solution. In particular, the upper bound $\bar{c}$ is the coverage corresponding to opening all stores in the first stage. The lower bound $\underline{c}$ can be obtained by first solving the deterministic counterpart of the problem where θ is fixed to the value θ_s for each scenario $s \in S$, and identifying the coverage level corresponding to opening set of common stores opened in all scenarios.

To linearize the formulation, we introduce another auxiliary decision variable $W = e^{-\alpha T}$. Then, one can observe that W is increasing in the coverage level c and replace the coverage constraint (5.27) with $W - e^{-\alpha\phi(c)} \leq 0$. Since the left hand side expression is nonlinear (and not necessarily convex), one may approximate the exponential term with a piecewise linear function of c, denoted by $\hat{W}(c)$, as follows. In particular, we partition the interval $[\underline{c}, \bar{c}]$ into K non-overlapping sub-intervals at break points $\underline{c}_k$ and $\bar{c}_k$. Then, within each interval, we use the linear function $\widehat{W}_k(c) = a_k + b_k c$ to approxmate the term $e^{-\alpha\phi(c)}$, while ensuring that $0 \leq \frac{e^{-\alpha\phi(c)} - \widehat{W}_k}{e^{-\alpha\phi(c)}} \leq \epsilon$ for all $c \in [\underline{c}_k, \bar{c}_k]$ where ϵ is a precribed tolerance level. Note that, as long as the term $e^{-\alpha\phi(c)}$ is continuous, an arbitrarily large tolerance level ϵ can be accommodated by increasing K, i.e., using more refined linear pieces to approximate the function. Furthermore, the piecewise linear function $\widehat{W}(c)$ is chosen such that it bounds the original expression from above, such that the approximation yields a conservative estimate of the overall objective value.

The key feature of this approximation scheme is that one can consider the k sub-intervals separately. In particular, one can impose the constraint $\underline{c}_k \leq c \leq \bar{c}_k$ and solve the K resulting subproblems independently. Let the optimal objective value for the k-th subproblem be $\hat{\Pi}_k$, which corresponds to the optimal profit while restricting the first-stage coverage to the interval $[\underline{c}_k, \bar{c}_k]$ (note that this can be negative if the subproblem is infeasible). The subproblem to be solved to obtain Π_k is formulated as follows.

$$\Pi_k = \max_{\mathbf{X},\mathbf{Y},\widehat{W}} \quad \sum_{s\in S}\frac{p^s}{\alpha}\left[(1-\widehat{W})\sum_{i\in I}\sum_{j\in J} r\theta^s h_i g_i(d_{ij}) Y^1_{ij} - \sum_{j\in J} f_j X^1_j\right. \tag{5.28}$$

$$\left. +\widehat{W}\left[\sum_{i\in I}\sum_{j\in J} r\theta^s h_i g_i(d_{ij}) Y^2_{ijs} - \sum_{j\in J} f_j X^2_{js}\right]\right]$$

$$\text{s.t.} \quad Y^1_{ij} \le X^1_j, \quad Y^2_{ijs} \le X^1_j + X^2_{js} \text{ for } i\in I, j\in J, s\in S$$

$$\sum_{j\in J} Y^1_{ij} = 1, \quad \sum_{j\in J} Y^2_{ijs} = 1 \text{ for } i\in I, s\in S$$

$$X^1_j + X^2_{js} \le 1 \text{ for } j\in J, s\in S$$

$$\widehat{W} = a_k + b_k c_k$$

$$c_k = \sum_{i\in I}\sum_{j\in J} h_i g_i(d_{ij}) Y^1_{ij}$$

$$c_k \in [\underline{c}_k, \bar{c}_k], \quad X^1_j, X^2_{js}, Y^1_{ij}, Y^2_{ijs} \in \{0,1\} \tag{5.29}$$

$$\text{for } i\in I, j\in J, s\in S.$$

Note that the objective function (5.28) still contains nonlinear terms as products of the variables Y^1_{ij} and Y^2_{ij}, and $\hat{W}$. These terms can be linearized using the standard procedure (e.g., Oral and Kettani, 1992) by observing that Y^1_{ij} and Y^2_{ij} are binary variables. Then, the final solution can be selected by taking the maximum of optimal profits in these subproblems, i.e., $\hat{\Pi} = \max_{k=1,\cdots,K} \hat{\Pi}_k$. Bhatti *et al.* (2015) prove the following result regarding the accuracy of such approximation.

Proposition 5.4. Let Π denote the optimal profit for problem [P1]. Then, $\hat{\Pi} \le \Pi$ and $\frac{\Pi - \hat{\Pi}}{\Pi} \le \epsilon$.

Proposition 5.4 states that the approximate profit obtained from the linearization approximation provides a lower bound on the true optimal profit, and the error can be made arbitrarily small by using more refined approximations (i.e., a larger K). The key advantage of this procedure is that the number of subproblems to solve increases linearly in K, and thus the computation times only mildly increases in the degree of approximation accuracy.

All of the applications discussed thus far have focused on the planning of for-profit facility networks. As a result, the objectives typically

involve maximizing profits or minimizing costs. The next application to be discussed provides an example of how design of public services may require different modeling considerations regarding both the planning objective and performance requirements.

5.4 Planning for Trauma Centers for Emergency Medical Services

In health care, the location of facilities is crucial to ensuring accessibility of service and cost efficiency. Moreover, the negative consequences of ill-informed facility location decisions in health care settings can extend beyond service quality and costs, leading to mortality and morbidity (Daskin and Dean, 2004). Among health care systems, emergency services are particularly reliant on careful location planning to perform their function, due to the high time sensitivity in demand. The US EMS (Emergency Medical Services) Act of 1973 requires that 95% of emergency service requests to be served within 30 minutes in rural areas and 10 minutes in urban areas Daskin and Dean (2004). In light of response time constraints, researchers have been working on location models for EMS based on covering distance models since the 1970's. For example, Toregas *et al.* (1971) and Church and Velle (1974) use the set covering and max covering models (discussed in Chapter 1) to study the problem of locating ambulances to cover EMS requests, respectively.

Response time in EMS is determined by not only travel distances, but also availability of ambulances. In particular, an ambulance may be unavailable upon request due to serving a prior request. To capture this, researchers have extended the covering models to incorporate the availability condition of ambulances. For example, Daskin (1983) formulates the maximum expected covering model, generalizing the max covering model to incorporate the objective of maximizing the expected number of available ambulances that can cover a demand node, assuming that each ambulance has a certain probability of being unavailable (busy). Eaton *et al.* (1985) report the implementation of such modeling in practice in Austin, TX. ReVelle and Hogan (1989) consider a chance-constrained problem in which the demand coverage within a prescribed time limit with a given probability is to be maximized. Marianov and ReVelle (1996) consider a queueing-based model that

relaxes the (strong) assumption that availability probabilities do not depend on facility siting decisions. Models for EMS facility location (see, for example, Brotcorne *et al.* (2003) for a review) involving busy ambulances (or servers) are often difficult to formulate or analyze due to the challenge of characterizing busy probabilities, which may endogenously depend on the choice of facilities. This challenge is much akin to the class of integrated facility location models discussed thus far.

With recent developments in theoretical and computational aspects of nonlinear integer optimization, it is now possible to tackle more complex EMS location planning problems driven by recent applications. Motivated by a EMS planning case in Korea, Cho *et al.* (2014) study a problem of simultaneously locating trauma centers (that treat emergency trauma patients) and helicopters. Similar to the case in prior models for ambulances, the modeling approach requires characterizing the busy probabilities of helicopters. An additional challenge is that the busy probabilities depend on the locations of both types of facilities with interacting operations. In particular, while the two types of facilities are complements (as helicopters transport patients to trauma centers) in operations, they also interact somewhat as substitutes as a demand location can be covered either directly by a trauma center (using ground ambulances) or by helicopters. We review the model formulation and solution approach in this section.

5.4.1 Location Model for Trauma Centers and Helicopters

Cho *et al.* (2014) consider the problem of locating K trauma centers and M heliports (helicopter depots) out of candidate sets J and H, respectively, to serve a set I of demand locations. The two types of facilities can be co-located, i.e., a heliport can be built on the roof of a trauma center or at a separate site and $H \supseteq J$. Each heliport (if located) is equipped with one helicopter. Demand occurs (independently) at each demand location $i \in I$ following a Poisson process with a rate of λ_i, and in such case, the patient is to be transported from i to one of the trauma centers, either using a ground ambulance or a helicopter.

To comply with response time regulations, sufficient trauma centers must be located such that patients can be transported to one of them

within 60 minutes upon demand occurence. This time-based requirement can be converted to distance-based requirements for the two modes (ground ambulance and helicopter) as follows. For ground ambulances, it is required that the road distance from the nearest ambulance station (the location of which is assumed to be known and fixed) to demand location i, d_i^r, plus the road distance from i to trauma center j, d_{ij}^r, must satisfy $d_i^r + d_{ij}^r \leq \bar{d}^r$. For helicopters, it is required that the Euclidean distance from the assigned heliport h to i, d_{hi}, and that from i to the nearest trauma center j, d_{ij}, satisfy $d_{hi} + d_{ij} \leq \bar{d}$. Note that while helicopters travel in straight line (thus the consideration of Euclidean distances), ambulances travel on ground along the road network. The values of $\bar{d}^g$ and $\bar{d}$ are determined based on the average operating speeds and loading times for ambulances and helicopters, respectively. For notational brevity, define

- $F^r = \{(i,j) | i \in I, j \in J, d_i^r + d_{ij}^r \leq \bar{d}^r\}$, i.e., the set of demand location-trauma center pairs that satisfy the ground ambulance response time limit;
- $F_i^r = \{j \in J | d_i^r + d_{ij}^r \leq \bar{d}^r\}$, i.e., the set of trauma centers that patients can be transported to within the time limit using ambulances;
- $F_j^r = \{i \in I | d_i^r + d_{ij}^r \leq \bar{d}^r\}$, i.e., the set of demand locations that a trauma center can serve using ambulances;
- $F = \{(i,j,h) | i \in I, j \in J, h \in H, d_{hi} + d_{ij} \leq \bar{d}\}$, i.e., the set of possible demand location-trauma center-heliport triplets where patients can be transported to the trauma center using helicopters within the time limit;
- $F_h = \{(i,j) | i \in I, j \in J, d_{hi} + d_{ij} \leq \bar{d}\}$, i.e., the set of demand location-trauma center pairs that can be served by heliport h;
- $F_i = \{(j,h) | j \in J, h \in H, d_{hi} + d_{ij} \leq \bar{d}\}$, i.e., the set of trauma center-heliport pairs that can jointly serve by demand location i;
- $F_j = \{(i,h) | i \in I, h \in H, d_{hi} + d_{ij} \leq \bar{d}\}$, i.e., the set of demand location-heliport pairs that can be served using trauma center j.

Let τ_{ijh} be the average time for a transportation and service cycle for a helicopter to travel from heliport h to retrieve a patient at i, transport him/her to trauma center j, and return to heliport h. Define X_j (W_h) as the binary decision variable indicating whether a trauma center (a heliport) is located at $j \in J$ ($h \in H$). Also, define Y^r_{ij} and Y_{ijh} be the continuous decision variables representing the average rates of patients transported from demand location $i \in I$ to trauma center $j \in J$ using ground ambulances and using a helicopter stationed at $h \in H$, respectively. Furthermore, we define the following auxilliary variables:

$$\lambda^r = \sum_{i \in I} \sum_{j \in F^r_i} Y^r_{ij} \tag{5.30}$$

$$\lambda_h = \sum_{(i,j) \in F_h} Y_{ijh} \text{ for } h \in H \tag{5.31}$$

$$\lambda_j = \sum_{i \in F^r_i} Y^r_{ij} + \sum_{(i,h) \in F_j} Y_{ijh} \text{ for } j \in J \tag{5.32}$$

$$r_h = \sum_{(i,j) \in F_h} \tau_{ijh} Y_{ijh} \text{ for } h \in H. \tag{5.33}$$

These auxilliary variables represent the total demand rate handled by ground ambulances, demand rate handled by heliport h, demand rate handled by trauma center j, and workload assigned to h, respectively.

Note that the above only specifies geographical coverage based on distance. Geographical coverage is only a necessary condition and is not sufficient for guaranteeing service without considering availability of helicopters. Between the two transportation modes, Cho *et al.* (2014) find that helicopters are typically the scarce resource in practice, and thus assume that there are ample ambulances available. To capture stochasticity in helicopter service, Cho *et al.* (2014) consider the expected demand coverage (a performance measure in line with Daskin (1983), for example) under random (Poisson) demand arrivals. In particular, with (Poisson) demand arrival rate of λ_h and workload of r_h, one can view the heliport h approximately as a single-server queue with average service time $\tau_h = r_h/\lambda_h$, by assuming that demand requests finding the helicopter busy will join a queue instead of being lost (or re-routed). The busy probability is thus r_h, and the expected demand covered by

heliport h per unit time is thus $\lambda_h(1-r_h)$. The objective is to maximize the total expected demand coverage by all heliports and trauma centers, given by $\sum_{h\in H}\lambda_h(1-r_h)+\lambda^r$.

Another important aspect of service quality is the waiting time for service at trauma centers. Cho *et al.* (2014) formulate a capacity constraint that limits λ_j by a limit $\mu_j\rho_j^{\omega,\xi}$ where $\rho_j^{\omega,\xi}$ is the maximum workload that can be assigned to a $M/M/k$ queueing system (representing the congested trauma center) such that the waiting time does not exceed threshold ω with a probability guarantee of ξ.

Combining the above considerations, the joint location problem for trauma centers and heliports can be formulated as the following mixed integer nonlinear program:

$$\max \quad \lambda^r + \sum_{h\in H}(1-r_h)\lambda_h = \lambda^r + \sum_{h\in H}\lambda_h - \lambda_h r_h \tag{5.34}$$

$$\text{s.t.} \quad (5.30-5.33) \tag{5.35}$$

$$W_j \le X_j \text{ for } j\in J \tag{5.36}$$

$$\sum_{j\in J} X_j \le K \tag{5.37}$$

$$\sum_{h\in H} W_h \le M \tag{5.38}$$

$$\sum_{j\in F_i^r} Y_{ij}^r + \sum_{(j,h)\in F_i} Y_{ijh} \le \lambda_i \text{ for } i\in I \tag{5.39}$$

$$\lambda_j \le \mu_j\rho_j^{\omega,\xi}X_j \text{ for } j\in J \tag{5.40}$$

$$r_h \le W_h \text{ for } h\in H \tag{5.41}$$

$$0\le Y_{ij}^r \le \lambda_i X_j \text{ for } (i,j)\in F^r \tag{5.42}$$

$$0\le Y_{ijh} \le X_j \text{ for } (i,j,h)\in F \tag{5.43}$$

$$X_j\in\{0,1\} \text{ for } j\in J$$

$$W_h\in\{0,1\} \text{ for } h\in H.$$

In the model, the objective (5.34) is to maximize expected demand coverage by all heliports and trauma centers (via ground ambulances). The constraints (5.36), (5.37) and (5.38) ensure that a heliport cannot be located at a trauma center site if the trauma center is not located, and that K trauma centers and M heliports are located in total, respectively.

Constraints (5.39) require that the demand coverage cannot exceed total demand arrivals. Constraints (5.40) impose the capacity limit on demand assignment to trauma centers such that their waiting times satisfy the probabilistic guarantee as discussed above. Finally, constraints (5.41), (5.42) and (5.43) stipulate that demand cannot be assigned to a facility of either type if said facility is not opened.

5.4.2 Solution Approach

The main challenge in solving the problem lies in that the objective function (5.34) carries the nonlinear terms $w_h = \lambda_h r_h$, which are neither convex nor concave. Cho *et al.* (2014) propose a linearization-bounding approach for solving the problem. To begin, note that it is possible to obtain a relaxation by using the McCormick envelope inequalities (McCormick, 1976):

$$w_h \geq \bar{\lambda}_h r_h + \bar{r}_h \lambda_h - \bar{\lambda}_h \bar{r}_h \tag{5.44}$$
$$w_h \geq \underline{\lambda}_h r_h + \underline{r}_h \lambda_h - \underline{\lambda}_h \underline{r}_h \tag{5.45}$$
$$w_h \leq \bar{\lambda}_h r_h + \underline{r}_h \lambda_h - \bar{\lambda}_h \underline{r}_h \tag{5.46}$$
$$w_h \leq \underline{\lambda}_h r_h + \bar{r}_h \lambda_h - \underline{\lambda}_h \bar{r}_h \tag{5.47}$$

where $\bar{r}_h, \underline{r}_h, \bar{\lambda}_h, \underline{\lambda}_h$ are upper and lower bounds on the possible values of r_h and λ_h, respectively. These inequalities bound the bilinear function from below (5.44-5.45) and above (5.46-5.47). Because the $\lambda_h r_h$ terms are to be minimized in the objective, we only need to impose the lower bounds (5.44-5.45) in the formulation. Moreover, observing that the lower bounds on r_h and λ_h are given by zero, (5.45) simply becomes $w_h \geq 0$. Therefore, one can obtain a relaxation of the problem (5.34) by replacing the objective with:

$$\begin{aligned} \max \quad & \lambda^r + \sum_{h \in H} (\lambda_h - w_h) && (5.48) \\ \text{s.t.} \quad & w_h \geq \bar{\lambda}_h r_h + \bar{r}_h \lambda_h - \bar{\lambda}_h \bar{r}_h \text{ for } h \in H \\ & w_h \geq 0 \text{ for } h \in H. \end{aligned}$$

Unfortunately, this linear relaxation turns out to be relatively weak. To improve the relaxation strength, (Cho *et al.*, 2014) propose to use

quadratic bounds instead of linear ones. In particular, they observe that $\tau_h = r_h/\lambda_h$ is the average service time of heliport h, which implies that $r_h\lambda_h = \tau_h\lambda_h^2$. Then, given lower and upper bounds $\underline{\tau}_h$ and $\bar{\tau}_h$ on τ_h, one can identify quadratic bounds on w_h in the form of $\underline{\tau}_h\lambda_h^2 \leq w_h \leq \bar{\tau}_h\lambda_h^2$. Again, as w_h is to be minimized, only the first inequality, which is convex, is needed. This leads to a relaxation of (5.34) in the form of a mixed integer quadratic program. Note that, because τ_h is a weighted average of τ_{ijh}, one can let $\bar{\tau}_h = \max_{i,j} \tau_{ijh}$ and $\underline{\tau}_h = \min_{i,j} \tau_{ijh}$. Cho *et al.* (2014) observe that the resulting quadratic lower bound is typically tighter than the linear McCormick bounds, especially at lower values of λ_h.

To further tighten the bound, one can observe that, for any fixed $\hat{\tau}_h$ that bounds the value of τ_h under the optimal solution, it holds that $\hat{\tau}_h$ and $\hat{\tau}_h\lambda_h^2$ provide valid lower bounds on the optimal values of r_h and w_h, respectively. To tighten the bound, one can divide the interval $[\underline{\lambda}_h, \bar{\lambda}_h]$ (not necessarily uniformly) into N subintervals $[\tau_{h,n}, \tau_{h,n+1}]$ for $n = 1, \cdots, N_h$. Then, for any feasible solution, the value of $\tau_h = r_h/\lambda_h$ will fall into exactly one of these intervals. Define new binary indicator variables z_{hn} which takes the value of one if the value of τ_h falls in the interval $[\tau_{h,n}, \tau_{h,n+1}]$, and zero otherwise, which can be imposed by adding the following logical constraints:

$$\lambda_h = \sum_{n=1}^{N_h} \lambda_{hn}, \quad r_h = \sum_{n=1}^{N_h} r_{hn} \text{ for } h \in H \tag{5.49}$$

$$0 \leq \lambda_{hn} \leq \bar{\lambda}_h z_{hn} \text{ for } n = 1 \cdots, N_h, h \in H \tag{5.50}$$

$$0 \leq r_{hn} \leq \bar{r}_h z_{hn} \text{ for } n = 1 \cdots, N_h, h \in H \tag{5.51}$$

$$\tau_{h,n}\lambda_{hn} \leq r_{hn} \leq \tau_{h,n+1}\lambda_{hn} \text{ for } n = 1, \cdots, N_h, h \in H \tag{5.52}$$

$$\sum_{n=1}^{N_h} x_{hn} = 1 \text{ for } h \in H \tag{5.53}$$

$$x_{hn} \in \{0, 1\} \text{ for } n = 1, \cdots, N_h, h \in H.$$

The above logical constraints form a multiple choice characterization that imposes one out of N possible lower bounds on each w_h term. In particular, (5.49) specify that the values of λ_h and r_h will be given by one of the λ_{hn} and r_{hn} values for the appropriate n. Constraints (5.50,

5.51) ensure that only the λ_{hn} and r_{hn} for the appropriate n will have a non-zero value. Constraints (5.52) make sure that the appropriate n be selected when r_h is given by $\hat{\tau}_h \lambda_h$ for some $\hat{\tau}_h \in [\tau_{h,n}, \tau_{h,n+1}]$. Finally, constraints (5.53) specify that one such interval be selected. With these logical constraints, the $\sum_{h \in H} w_h$ terms in objective (5.48) can be replaced with $\sum_{h \in H} \sum_{n=1}^{N} \tau_{h,n} \lambda_{hn}^2$.

It is obvious that the tightness of the resulting lower bound obtained from this multiple choice formulation depends on how finely the intervals $[\tau_{h,n}, \tau_{h,n+1}]$ are partitioned. Generally, a finer partition (loosely speaking, larger N_h) would lead to a better (more accurate) bound, while increase problem size and computation times. To obtain a good trade-off between accuracy and computation speed, Cho *et al.* (2014) propose a shifting quadratic envelopes algorithm, which solves the problem iteratively while updating the partitions. In the first iteration, $N_h = 1$ for all $h \in H$. Then, in each subsequent iteration, the interval in which the optimal λ_h lies is bisected evenly for each $h \in H$. The motivation is to start with a small number of intervals (and thus a modest-sized formulation) and refine the partition near the region where the true optimal value of λ_h is expected to lie in. Furthermore, to limit the problem size from increasing too much, the intervals that have low likelihood of containing the optimal values can be merged. Cho *et al.* (2014) find that, from a practical point of view, keeping three intervals below and one above the current λ_h solution provides a good trade-off between solution quality and computation speed.

Another aspect of the solution algorithm is to obtain feasible solutions (and associated upper bounds on the optimal objective value) based on the lower bound solutions. Cho *et al.* (2014) do so by solving a restricted version of the problem by fixing the variables $(X_j, W_h, \lambda^r, \lambda_h, Y_{ij}^r)$ based on their values in the lower bound solution and resolving for the remaining variables (Y_{ijh}, r_h). Note that the resulting problem, which determines the assignment of demand to heliports and determines the corresponding workload, is a linear program and can be solved efficiently. Combining these procedures, Cho *et al.* (2014) report that the overall algorithm performs very well computationally, compared with a Bender's decomposition benchmark.

5.5 Discussion

The scope of facility location problems has, in recent years, expanded well beyond the conventional supply chain domain. The studies reviewed in this section have covered emerging domain areas of sustainable transportation, renewable energy, retail, and health care. While these domain areas are seeing growing interest in the discipline of operations management in general, unique opportunities exist for researchers specialized in location modeling because of the strategic implications of spatial considerations in these potentially high-impact problems. For example, using a network design model calibrated to real data, Deo *et al.* (2015) reveal that redesigning the supply chain for infant HIV diagnosis could significantly improve the number of infected infants receiving treatments. We believe that further research along these lines can complement existing research focusing on other operations questions (e.g., capacity planning for renewable generation and storage). Particularly, incorporating the spatial factor in the former can enhance the breadth of analysis of the latter, such as the charging network analysis in the study of range and resale anxiety for EV adoption in Lim *et al.* (2016); the findings in the latter could also be used to capture various operational characteristics in building new models for the former.

6

Conclusion and Future Directions

Recent advances in optimization, particularly in the study of nonlinear, stochastic and robust integer problems, have greatly enhanced the expanded the scope of tractable problems and enabled rich problem features to be built into facility location models. This has allowed researchers and practitioners to enhance the classical facility location models by incorporating problem-specific facility characteristics into their models and obtain richer insights into location planning problems in practice. In this monograph, we have provided an introduction to this integrated modeling approach to facility location and a brief review of the vast and growing literature.

We believe that research in this area will continue to grow in the future. Particularly promising are new applications in supply chain planning and other emerging operations areas. In the supply chain domain, existing research (including the papers we have reviewed) has mainly focused on the planning of back-end distribution and storage facilities (i.e., DCs and warehouses). Recent developments in the industry, however, have made it interesting to consider the front-end aspects as well. We have discussed in Section 4.5 and 4.6 distribution networks for online retailers exhibit different characteristics than tra-

ditional brick-and-mortar ones. The emerging segments of rapid (e.g., same-day) delivery and online-to-offline (O2O) shopping in online retail have resulted in interesting new operations modes as well as business models that motivate novel research questions. For example, new businesses such as Google Shopping Express, Curbside and Instacart have been established to facilitate the new fulfillment needs and opportunities arising from the new trend. These new business (and operations) models exhibit very different characteristics compared with their conventional counterparts, and potentially provide many novel research questions regarding optimal distribution network planning that may be tackled with new classes of facility location modeling. Overall, we believe the rise of digital and information-driven supply chain operations will bring about tremendous research opportunities.

Intriguing developments exist in other domain areas as well. For example, sustainable transportation has become a focus of development of smart cities. This development also gives rise to interesting spatial network planning problems to be tackled with integrated facility location approaches. For example, He *et al.* (2016) and Kabra *et al.* (2015) study planning problems for car sharing and bike sharing problems, respectively. In both cases, service provision (the coverage of service regions and locations of bike stations) is to be optimized with consideration of service availability (having large enough fleets of cars and bikes to ensure availability). Likewise, in both the energy and health care sectors, high-impact network planning problems remain to be tackled with advanced modeling approaches.

We believe these are just a few of the many promising directions in which the literature will extend. We hope the review and discussion provided in this monograph may serve as a useful reference for researchers and practitioners in making their contributions to this growing research stream.

Acknowledgements

We would like to thank Charles Corbett (the editor) and the anonymous reviewer for their guidance and comments that has helped us prepare and revise this monograph. We are also grateful for helpful feedback provided by Wancheng Feng and Patrick Tsang.

References

Acimovic, J. and S. C. Graves. 2014. "Making better fulfillment decisions on the fly in an online retail environment". *Manufacturing & Service Operations Management*. 17(1): 34–51.

Ahuja, R. K., T. L. Magnanti, and J. B. Orlin. 1993. *Network flows: Theory, algorithms and applications*. Prentice Hall.

Alizamir, S., F. de Véricourt, and P. Sun. 2016. "Efficient feed-in-tariff policies for renewable energy technologies". *Operations Research*. 64(1): 52–66.

Artzner, P., F. Delbaen, J.-M. Eber, and D. Heath. 1999. "Coherent measures of risk". *Mathematical finance*. 9(3): 203–228.

Atamtürk, A., G. Berenguer, and Z.-J. Shen. 2012. "A conic integer programming approach to stochastic joint location-inventory problems". *Operations research*. 60(2): 366–381.

Atamtürk, A. and V. Narayanan. 2008. "Polymatroids and Mean-risk Minimization in Discrete Optimization". *Operations Research Letters*. 36(5): 618–622.

Axsäter, S. 1996. "Using the deterministic EOQ formula in stochastic inventory control". *Management Science*. 42(6): 830–834.

Axsäter, S. 2007. *Inventory control*. Springer.

Barnhart, C., E. L. Johnson, G. L. Nemhauser, M. W. Savelsbergh, and P. H. Vance. 1998. "Branch-and-price: Column generation for solving huge integer programs". *Operations research*. 46(3): 316–329.

Bazaraa, M., H. Sherali, and C. Shetty. 2004. *Nonlinear Programming: Theory and Algorithms, Second Edition*. Wiley.

Belavina, E., K. Girotra, and A. Kabra. 2016. "Online Grocery Retail: Revenue Models and Environmental Impact". *Management Science*.

Berman, O. and D. Krass. 2002. "The generalized maximal covering location problem". *Computers & Operations Research*. 29(6): 563–581.

Berman, O., D. Krass, and Z. Drezner. 2003. "The gradual covering decay location problem on a network". *European Journal of Operational Research*. 151(3): 474–480.

Bhatti, S. F., M. K. Lim, and H.-Y. Mak. 2015. "Alternative fuel station location model with demand learning". *Annals of Operations Research*: 1–23.

Bitar, E., K. Poolla, P. Khargonekar, R. Rajagopal, P. Varaiya, and F. Wu. 2012. "Selling random wind". In: *System Science (HICSS), 2012 45th Hawaii International Conference on*. IEEE. 1931–1937.

Boyd, S. and L. Vandenberghe. 2009. *Convex optimization*. Cambridge university press.

Bridgman, P. 1922. *Dimensional analysis*. Yale University Press.

Brotcorne, L., G. Laporte, and F. Semet. 2003. "Ambulance location and relocation models". *European journal of operational research*. 147(3): 451–463.

Cachon, G. P. 2012. "What is interesting in operations management?" *Manufacturing & Service Operations Management*. 14(2): 166–169.

Cachon, G. P. 2014. "Retail store density and the cost of greenhouse gas emissions". *Management Science*. 60(8): 1907–1925.

California Department of Transportation. 2010. "California Household Travel Survey". http://www.dot.ca.gov/hq/tsip/otfa/tab/chts_travelsurvey.html.

Chen, G., M. S. Daskin, Z.-J. M. Shen, and S. Uryasev. 2006. "The α-reliable mean-excess regret model for stochastic facility location modeling". *Naval Research Logistics (NRL)*. 53(7): 617–626.

Chen, Q., X. Li, and Y. Ouyang. 2011. "Joint inventory-location problem under the risk of probabilistic facility disruptions". *Transportation Research Part B: Methodological*. 45(7): 991–1003.

Chen, X., M. Sim, and P. Sun. 2007. "A robust optimization perspective on stochastic programming". *Operations Research*. 55(6): 1058–1071.

Cheung, R. K. and W. B. Powell. 1996. "An algorithm for multistage dynamic networks with random arc capacities, with an application to dynamic fleet management". *Operations Research*. 44(6): 951–963.

Cho, S.-H., H. Jang, T. Lee, and J. Turner. 2014. "Simultaneous location of trauma centers and helicopters for emergency medical service planning". *Operations Research*. 62(4): 751–771.

Chopra, S. and P. Meindl. 2007. *Supply Chain Management. Strategy, Planning & Operation*. Springer.

Chopra, S., G. Reinhardt, and U. Mohan. 2007. "The importance of decoupling recurrent and disruption risks in a supply chain". *Naval Research Logistics (NRL)*. 54(5): 544–555.

Church, R. and C. R. Velle. 1974. "The maximal covering location problem". *Papers in Regional Science*. 32(1): 101–118.

Cohen, M. C., R. Lobel, and G. Perakis. 2015. "The impact of demand uncertainty on consumer subsidies for green technology adoption". *Management Science*.

Cui, T., Y. Ouyang, and Z.-J. M. Shen. 2010. "Reliable facility location design under the risk of disruptions". *Operations Research*. 58(4-part-1): 998–1011.

Daganzo, C. F. 1984. "The distance traveled to visit N points with a maximum of C stops per vehicle: An analytic model and an application". *Transportation Science*. 18(4): 331–350.

Daganzo, C. F. 2005. *Logistics systems analysis*. Springer.

Daganzo, C. F. and K. R. Smilowitz. 2004. "Bounds and approximations for the transportation problem of linear programming and other scalable network problems". *Transportation science*. 38(3): 343–356.

Daskin, M. S. 1983. "A maximum expected covering location model: formulation, properties and heuristic solution". *Transportation Science*. 17(1): 48–70.

Daskin, M. S. 2011. *Network and discrete location: models, algorithms, and applications*. John Wiley & Sons.

Daskin, M. S., C. R. Coullard, and Z.-J. M. Shen. 2002. "An inventory-location model: Formulation, solution algorithm and computational results". *Annals of Operations Research*. 110(1-4): 83–106.

Daskin, M. S. and L. K. Dean. 2004. "Location of health care facilities". In: *Operations research and health care*. Springer. 43–76.

Daskin, M. S., S. M. Hesse, and C. S. Revelle. 1997. "α-reliable p-minimax regret: a new model for strategic facility location modeling". *Location Science*. 5(4): 227–246.

Daskin, M. S. and K. L. Maass. 2015. "The p-Median Problem". In: *Location Science*. Springer. 21–45.

Deo, S., J. Gallien, and J. O. Jónasson. 2015. "Improving HIV early infant diagnosis supply chains in sub-Saharan Africa: Models and application to Mozambique". *Tech. rep.* London Business School.

Drezner, Z. 1995. *Facility location: a survey of applications and methods*. Springer Verlag.

Eaton, D. J., M. S. Daskin, D. Simmons, B. Bulloch, and G. Jansma. 1985. "Determining emergency medical service vehicle deployment in Austin, Texas". *Interfaces*. 15(1): 96–108.

Edmonds, J. 1970. "Submodular functions, matroids, and certain polyhedra". *Combinatorial Structures and Their Applications*: 69–87.

Eppen, G. D. 1979. "Effects of centralization on expected costs in a multi-location newsboy problem". *Management Science*. 25(5): 498–501.

Fisher, M. L. 1985. "An applications oriented guide to Lagrangian relaxation". *Interfaces*. 15(2): 10–21.

Forbes. 2013. "Ecommerce is Growing Nicely while Mcommerce is on a Tear". http://www.forbes.com/sites/chuckjones/2013/10/02/ecommerce-is-growing-nicely-while-mcommerce-is-on-a-tear/.

Forbes. 2015. "Alibaba Starts Drone Delivery Test In Three-Day Program". http://www.forbes.com/sites/ywang/2015/02/03/alibaba-starts-drone-delivery-test-in-three-day-program/.

Foreman, J., J. Gallien, J. Alspaugh, F. Lopez, R. Bhatnagar, C. C. Teo, and C. Dubois. 2010. "Implementing supply-routing optimization in a make-to-order manufacturing network". *Manufacturing & Service Operations Management*. 12(4): 547–568.

Fu, M. C. 2002. "Optimization for simulation: Theory vs. practice". *INFORMS Journal on Computing*. 14(3): 192–215.

Goh, J. and M. Sim. 2010. "Distributionally robust optimization and its tractable approximations". *Operations research*. 58(4): 902–917.

Gresh, D. L. and E. I. Kelton. 2003. "Visualization, optimization, business strategy: a case study". In: *Visualization, 2003. VIS 2003. IEEE*. 531–538.

Hakimi, S. L. 1964. "Optimum locations of switching centers and the absolute centers and medians of a graph". *Operations Research*. 12(3): 450–459.

Hakimi, S. L. 1965. "Optimum distribution of switching centers in a communication network and some related graph theoretic problems". *Operations Research*. 13(3): 462–475.

Hamacher, H. W. and Z. Drezner. 2002. *Facility location: applications and theory*. Springer.

Harrison, J. M. and J. A. Van Mieghem. 1999. "Multi-resource investment strategies: Operational hedging under demand uncertainty". *European Journal of Operational Research*. 113(1): 17–29.

He, L., H.-Y. Mak, Y. Rong, and Z.-J. M. Shen. 2016. "Service Region Design for Urban Electric Vehicle Sharing Systems". *Tech. rep.* Working paper, National University of Singapore.

Hong, L. J. and B. L. Nelson. 2009. "A brief introduction to optimization via simulation". In: *Winter Simulation Conference*. Winter Simulation Conference. 75–85.

Kabra, A., E. Belavina, and K. Girotra. 2015. "Bike Share Systems: Accessibility and Availability". *Tech. rep.* INSEAD.

Kim, J. H. and W. B. Powell. 2011. "Optimal energy commitments with storage and intermittent supply". *Operations Research*. 59(6): 1347–1360.

Kunnumkal, S. and H. Topaloglu. 2008. "A refined deterministic linear program for the network revenue management problem with customer choice behavior". *Naval Research Logistics (NRL)*. 55(6): 563–580.

Laporte, G., S. Nickel, and F. S. da Gama. 2015. *Location science*. Vol. 145. Springer.

Li, X. and Y. Ouyang. 2010. "A continuum approximation approach to reliable facility location design under correlated probabilistic disruptions". *Transportation Research Part B: Methodological*. 44(4): 535–548.

Lim, M. K., A. Bassamboo, S. Chopra, and M. S. Daskin. 2013. "Facility location decisions with random disruptions and imperfect estimation". *Manufacturing & Service Operations Management*. 15(2): 239–249.

Lim, M. K., H.-Y. Mak, and Z.-J. Shen. 2016. "Agility and proximity considerations in supply chain design". *Management Science*.

Lim, M., M. S. Daskin, A. Bassamboo, and S. Chopra. 2010. "A facility reliability problem: formulation, properties, and algorithm". *Naval Research Logistics (NRL)*. 57(1): 58–70.

Lobo, M. S., L. Vandenberghe, S. Boyd, and H. Lebret. 1998. "Applications of second-order cone programming". *Linear Algebra and its Applications*. 284(1): 193–228.

Lu, D., F. Gzara, and S. Elhedhli. 2014. "Facility location with economies and diseconomies of scale: models and column generation heuristics". *IIE Transactions*. 46(6): 585–600.

Luo, J., L. J. Hong, B. L. Nelson, and Y. Wu. 2015. "Fully sequential procedures for large-scale ranking-and-selection problems in parallel computing environments". *Operations Research*. 63(5): 1177–1194.

Mak, H.-Y. 2012. "Supply Chain Network Design with Dynamic Inventory Sharing". *Working Paper, The Hong Kong University of Science and Technology*.

Mak, H.-Y., Y. Rong, and Z.-J. M. Shen. 2013. "Infrastructure planning for electric vehicles with battery swapping". *Management Science*. 59(7): 1557–1575.

Mak, H.-Y. and Z.-J. Shen. 2009. "A two-echelon inventory-location problem with service considerations". *Naval Research Logistics*. 56(8): 730–744.

Mak, H.-Y. and Z.-J. Shen. 2012. "Risk diversification and risk pooling in supply chain design". *IIE Transactions*. 44(8): 603–621.

Mak, H.-Y. and Z.-J. M. Shen. 2011. "Integrated Supply Chain Design Models". *Wiley Encyclopedia of Operations Research and Management Science*.

Marianov, V. and C. ReVelle. 1996. "The queueing maximal availability location problem: a model for the siting of emergency vehicles". *European Journal of Operational Research*. 93(1): 110–120.

McCormick, G. P. 1976. "Computability of global solutions to factorable nonconvex programs: Part I - Convex underestimating problems". *Mathematical programming*. 10(1): 147–175.

Nahmias, S. and Y. Cheng. 2009. *Production and operations analysis*. Vol. 5. McGraw-Hill New York.

Naseraldin, H. and Y. T. Herer. 2008. "Integrating the number and location of retail outlets on a line with replenishment decisions". *Management Science*. 54(9): 1666–1683.

Newell, G. F. 1973. "Scheduling, location, transportation, and continuum mechanics: some simple approximations to optimization problems". *SIAM Journal on Applied Mathematics*. 25(3): 346–360.

Oral, M. and O. Kettani. 1992. "A linearization procedure for quadratic and cubic mixed-integer problems". *Operations Research*. 40(1 supplement 1): S109–S116.

Ouyang, Y. and C. F. Daganzo. 2006. "Discretization and validation of the continuum approximation scheme for terminal system design". *Transportation Science*. 40(1): 89–98.

Ozsen, L., C. R. Coullard, and M. S. Daskin. 2008. "Capacitated warehouse location model with risk pooling". *Naval Research Logistics (NRL)*. 55(4): 295–312.

Ozsen, L., M. S. Daskin, and C. R. Coullard. 2009. "Facility location modeling and inventory management with multisourcing". *Transportation Science*. 43(4): 455–472.

Powell, W. B. 2007. *Approximate Dynamic Programming: Solving the curses of dimensionality*. John Wiley & Sons.

Powell, W. B. and R. K.-M. Cheung. 1994. "A network recourse decomposition method for dynamic networks with random arc capacities". *Networks*. 24(7): 369–384.

Providence Business News. 2010. "CVS to Open First 9 Stores in Puerto Rico". http://pbn.com/cvs-to-open-first-9-stores-in-puerto-rico,47959?print=1.

Puerto Rico Daily Sun. 2011. "CVS/Pharmacy to Open 13 Stores in Puerto Rico by 2012".

Qi, L., Z.-J. Shen, and L. V. Snyder. 2010. "The effect of supply disruptions on supply chain design decisions". *Transportation Science*. 44(2): 274–289.

Qi, W., Y. Liang, and Z.-J. M. Shen. 2015. "Joint Planning of Energy Storage and Transmission for Wind Energy Generation". *Operations Research*. 63(6): 1280–1293.

ReVelle, C. and K. Hogan. 1989. "The maximum availability location problem". *Transportation Science*. 23(3): 192–200.

Robinson, L. W. 1990. "Optimal and approximate policies in multiperiod, multilocation inventory models with transshipments". *Operations Research*. 38(2): 278–295.

Rockafellar, R. T. and S. Uryasev. 2000. "Optimization of conditional value-at-risk". *Journal of Risk*. 2: 21–42.

Rockafellar, R. T. and S. Uryasev. 2002. "Conditional value-at-risk for general loss distributions". *Journal of Banking & Finance*. 26(7): 1443–1471.

Roth, R. 1969. "Computer solutions to minimum-cover problems". *Operations Research*. 17(3): 455–465.

Serra, D. and V. Marianov. 1998. "The P-median problem in a changing network: the case of Barcelona". *Location Science*. 6(1): 383–394.

Shen, Z.-J. M. 2005. "A multi-commodity supply chain design problem". *IIE Transactions*. 37(8): 753–762.

Shen, Z.-J. M. 2007. "Integrated supply chain design models: a survey and future research directions". *Journal of Industrial and Management Optimization*. 3(1): 1.

Shen, Z.-J. M., C. Coullard, and M. S. Daskin. 2003. "A joint location-inventory model". *Transportation Science*. 37(1): 40–55.

Shen, Z.-J. M., R. L. Zhan, and J. Zhang. 2011. "The reliable facility location problem: Formulations, heuristics, and approximation algorithms". *INFORMS Journal on Computing*. 23(3): 470–482.

Sherali, H. D., G. Choi, and C. H. Tuncbilek. 2000. "A variable target value method for nondifferentiable optimization". *Operations Research Letters*. 26(1): 1–8.

Shu, J., M. Song, D. Xu, and K. Zhang. 2014. "A Column Generation Algorithm for Facility Location with General Facility Cost Functions". *Tech. rep.* University of Hong Kong.

Snyder, L. V. 2006. "Facility location under uncertainty: a review". *IIE Transactions*. 38(7): 547–564.

Snyder, L. V., Z. Atan, P. Peng, Y. Rong, A. J. Schmitt, and B. Sinsoysal. 2015. "OR/MS models for supply chain disruptions: A review". *IIE Transactions*: 1–21.

Snyder, L. V. and M. S. Daskin. 2005. "Reliability models for facility location: the expected failure cost case". *Transportation Science*. 39(3): 400–416.

Snyder, L. V., M. S. Daskin, and C.-P. Teo. 2007. "The stochastic location model with risk pooling". *European Journal of Operational Research*. 179(3): 1221–1238.

Snyder, L. V. and Z.-J. M. Shen. 2006. "Supply and demand uncertainty in multi-echelon supply chains". *Technical Report, Lehigh University*.

Sonin, A. A. 2001. "The physical basis of dimensional analysis". *Department of Mechanical Engineering, MIT, Cambridge, MA*.

Tomlin, B. 2006. "On the value of mitigation and contingency strategies for managing supply chain disruption risks". *Management Science*. 52(5): 639–657.

Tomlin, B. 2009. "Disruption-management strategies for short life-cycle products". *Naval Research Logistics (NRL)*. 56(4): 318–347.

Topaloglu, H. and S. Kunnumkal. 2006. "Approximate dynamic programming methods for an inventory allocation problem under uncertainty". *Naval Research Logistics (NRL)*. 53(8): 822–841.

Toregas, C., R. Swain, C. ReVelle, and L. Bergman. 1971. "The location of emergency service facilities". *Operations Research*. 19(6): 1363–1373.

US Environmental Protection Agency. 2014. "Sources of Greenhouse Gas Emissions". http://www.epa.gov/climatechange/ghgemissions/sources/electricity.html.

Vakharia, A. J. and A. Yenipazarli. 2009. *Managing supply chain disruptions*. Now Publishers Inc.

Vanderbeck, F. and M. W. Savelsbergh. 2006. "A generic view of Dantzig–Wolfe decomposition in mixed integer programming". *Operations Research Letters*. 34(3): 296–306.

Wall Street Journal. 2011a. "Apple Careful in China". http://www.wsj.com/articles/SB10001424053111903703604576588351297817510.

Wall Street Journal. 2011b. "Quake Still Rattles Suppliers". http://www.wsj.com/articles/SB10001424053111904563904576586040856135596.

Wu, O. Q. and R. Kapuscinski. 2013. "Curtailing intermittent generation in electrical systems". *Manufacturing & Service Operations Management*. 15(4): 578–595.

Xu, P. J., R. Allgor, and S. C. Graves. 2009. "Benefits of reevaluating real-time order fulfillment decisions". *Manufacturing & Service Operations Management*. 11(2): 340–355.

Zipkin, P. H. 2000. *Foundations of inventory management*. McGraw-Hill New York.