Chapter 17
Stochastic Location Models with Congestion

Oded Berman and Dmitry Krass

Abstract In this chapter we describe facility location models where consumers generate streams of stochastic demands for service, and service times are stochastic. This combination leads to congestion, where some of the arriving demands cannot be served immediately and must either wait in queue or be lost to the system. These models have applications that range from emergency service systems (fire, ambulance, police) to networks of public and private facilities. One key issue is whether customers travel to facilities to obtain service, or mobile servers travel to customer locations (e.g., in case of police cars). For the most part, we focus on models with static (fixed) servers, as the underlying queueing systems are more tractable and thus a richer set of analytical results is available. After describing the main components of the system (customers, facilities, and the objective function), we focus on the customer-facility interaction, developing a classification of models based on the how customer demand is allocated to facilities and whether the demand is elastic or not. We use our description of system components and customer-response classification to organize the rich variety of models considered in the literature into four thematic groups that share common assumptions and structural properties. For each group we review the solution approaches and outline the main difficulties. We conclude with a review of some important open problems. We specifically outline the advances and new approaches that have been developed since the previous edition of this volume.

17.1 Introduction

The class of facility location models that is the main focus of the current chapter make the following key assumptions:

1. Customers generate a *stochastic* stream of demands, typically assumed to be a Poisson process, or, more generally a renewal process.

O. Berman (✉) · D. Krass
Rotman School of Management, University of Toronto, Toronto, ON, Canada
e-mail: berman@rotman.utoronto.ca; krass@rotman.utoronto.ca

G. Laporte et al. (eds.), *Location Science*,
https://doi.org/10.1007/978-3-030-32177-2_17

2. Facilities, contain resources (often called "servers") that have *limited* capacity and *stochastic service times*.
3. Customer-facility interactions happen as the result of *customers traveling to facilities* to seek service, i.e., our primary focus is on the "fixed" or "immobile" server models (in the "mobile server" case, servers travel to customers to provide service).
4. Due to stochastic arrivals of customer demands at the facilities, stochastic service times, and limited capacities, facilities will experience periods of *congestion* where not all arriving demands can be served immediately. Customers that arrive when the system is busy may either enter a queue or leave without getting service. This behavior will result in either *queues*, or *lost demands*, or both.

Applications of these models range from public service facilities such as hospitals, medical clinics and government offices, to private facilities such as retail stores or repair shops.

We note that these assumptions specifically exclude a number of interesting and important classes of related location models, some of these are treated in other chapters in the current volume; we refer the reader specifically to Chap. 8 for an in-depth discussion of the issues outlined below.

First, there are many models that incorporate capacity limitations in a deterministic, rather than stochastic, manner. These include models seeking to ensure that there is sufficient average capacity to provide adequate service, models that try to design a system that should perform well even under stochastic conditions by equalizing loads between facilities, and models that handle possible congestion indirectly by requiring certain reserve capacity at the facilities. All of these can be regarded as deterministic approximations of the underlying stochastic system. While this deterministic approach leads to large technical simplifications and, as a result, much easier computations, the roughness of the approximation is usually impossible to estimate *a priori*. This may lead to systems with poor levels of customer service (at some of the facilities), and is typically not appropriate in cases where understanding and controlling potential congestion is important.

Second, there are some models where facilities are modeled as reliability, rather than queueing, systems, i.e., a facility may "fail" with certain probability in some periods, at which point it cannot provide service to customers (who are typically assumed to try to seek service from non-failed facilities)—these and related models are discussed in Chap. 22. Such models do incorporate stochastic demands explicitly. Moreover, "failure" periods may be regarded as representing periods of congestion at the facilities when new customer arrivals are blocked. Thus, these models are closer to the systems we study. However, the key difference is that "reliability" models treat the blockage probability as exogenous to the system (a typical assumption is that each facility may fail with certain probability at any time, where such probability is a system parameter), while models where facilities are represented as queues treat the probability of blockage as endogenous, i.e., a direct outcome of other decisions such as capacity allocation and customer-facility interactions. Thus, reliability models can only be regarded as approximations for

the systems we are interested in. We refer to Snyder (2006) as well as to Chaps. 8 and 22 in this volume for a review of reliability and related models.

Third, there is an important class of models where servers are assumed to be "mobile", i.e., servers travel to customers rather than customers traveling to facilities. Examples of the underlying systems include emergency services (fire, ambulance, police) as well as repairmen making house calls. These models are close "cousins" of the fixed-server models we are interested in as they include most of the same components: stochastic demand streams, stochastic service times, congestion/queuing behavior. However, these models also include additional significant levels of complexity, such as dynamic dispatching and routing of servers, repositioning servers between facilities, re-routing a sever before completion of the call, etc. The underlying queuing models are analytically intractable, even if the facility locations are assumed fixed, leading to various approximation-based approaches. In contrast, the queuing systems underlying models with fixed servers are often (though not always) analytically tractable, allowing for, theoretically, more precise solutions in many cases. We refer the reader to a survey by Berman and Krass (2002) and to a more recent survey on emergency systems planning by Ignolfsson (2013) for more details on models with mobile servers. We note that the technical distinction between models with fixed and mobile servers does not lie in the server mobility per se, but rather in how the underlying queuing network is modeled (in fact, some of the models described in this chapter have been applied in mobile server contexts). We will provide more precision for this distinction below, once the underlying technical framework is properly introduced.

The field of *Stochastic Location models with Congestion and Immobile Servers* (SLCIS), the main focus of this chapter, has seen a rather explosive growth over a relatively recent time period. As noted in Berman and Krass (2002), by the early 2000s, only a handful of papers on SLCIS could be found. However, by 2006 over 20 contributions were listed in the comprehensive review by Boffey et al. (2006) (we are only counting the papers that meet the assumptions for SLCIS models discussed earlier). In the last 8 years, this number has roughly doubled. It is our intent to review the current state of the field, as well as to systematize the many variants of SLCIS models that have been proposed.

We note that much of the recent work has been on models with elastic demand—i.e., where the intensity of customer demands depends on the quality of the service provided by the facilities. In this regard it is important to mention a review by Brandeau et al. (1995) that describes early foundation for much of this work.

As with most other location models, one could focus on cost minimization or on net revenue (profit) maximization. Cost minimization is more appropriate when the revenues are either not well-defined (e.g., in the case of public health facilities), or are assumed to be exogenous to the model (e.g., when customer demand levels and prices are fixed). While most SLCIS models in the literature are formulated with the cost minimization objective, profit optimization is more general and is much more natural when demand is elastic. Therefore, we will assume this objective type in our general formulation in the following section.

Several interesting new ideas have been introduced to SLCIS models since the previous edition of this volume. These are highlighted in the present version.

The remainder of this chapter is organized as follows. We start by describing the main model components in Sect. 17.2. A crucial part of any SLCIS model is the set of assumptions made about how customers and facilities interact, specifically how customer demand is "allocated" to facilities and how much of the potentially available demand is "captured". These issues are explored in detail in Sect. 17.3, where we also introduce a classification of SLCIS models based on the types of customer response. All model components come together in Sect. 17.4 where we formulate a "general" SLCIS model and review the main features that are typically included in various sub-classes. In Sect. 17.5 we provide an overview of SLCIS models discussed in the literature, providing a unifying structure organized around four main "themes". We also discuss the key challenges that arise for different model classes and computational approaches that have been developed. In the last section we discuss conclusions and suggestions for future research.

17.2 Key Model Components

In this section we specify the key model components that allow us to identify the main classes of SLCIS models. These classes and the relevant solution approaches will be described in the following sections. As noted earlier, SLCIS models describe the system consisting of customers, facilities and their interactions. We start by describing each of these components in more detail.

17.2.1 Customers

Customers are assumed to be located in a set J, with customer location $j \in J$ capable of generating a demand stream with maximum intensity of $\lambda_j^{\max}$ per unit time. In the vast majority of models described in the literature, J is assumed to be a discrete set, often conceptualized as the set of nodes of some underlying network $G = (J, A)$, where A is the set of links. Other common alternatives in location (but not in SLCIS) literature include J being a sub-region of the real plane R^2, or consisting of both links and nodes of a network G. The most general SLCIS setting we are aware of is given in Baron et al. (2008), where J is a bounded sub-space of R^N and can contain a mixture of discrete points and continuous regions. To keep the presentation as transparent as possible, we will retain the common assumption that J is discrete and $n = |J|$ is the number of customer demand points, which we will frequently refer to as "nodes".

Let u_j represents the *utility* derived by customers at node $j \in J$ from the services offered by the facilities. The demand stream generated by j is assumed to be a

Poisson process with rate $\lambda(u_j) \in [0, \lambda_j^{\max}]$. We will postpone the description of utility functions until Sect. 17.3.1, since other system components need to be defined first. However, we can already identify two different classes of SLCIS models: the *elastic demand* models, where $\lambda(u_j)$ is a non-constant function, i.e., $\lambda(u_j) \neq \lambda_j^{\max}$ for some values of u_j, and the *inelastic demand* models where the demand rate is assumed to be constant and equal to $\lambda_j^{\max}$. As a shorthand, we will use $\lambda_j = \lambda(u_j)$ to represent the demand rate of customer node $j \in J$. The inter-arrival times of the demand processes generated by different customer locations are assumed to be independent.

We should also note that while it is tempting to relax the Poisson assumption for the demand process, this must be done with care as the facilities see aggregate demands from different customer locations, i.e., a superposition of the demand processes. In order to apply standard queueing results to the facilities, the demand process seen by each facility must be a renewal process. While the superposition of Poisson processes is Poisson, which is obviously a renewal process, in general, the superposition of renewal processes is not a renewal process. This quickly leads to a loss of tractability for the models. Thus, except for some trivial extensions, the Poisson assumption for demand streams appears unavoidable; one interesting exception occurs when customer demand space is continuous, rather than finite, in which case facilities see Poisson arrivals under much looser conditions—see Baron et al. (2008) for the development and required assumptions. However, there is no problem (at least from the analytical point of view) in assuming that the demand process at each node $j \in J$ is not time-homogenous, i.e., that the demand rate is a function of time. To simplify the presentation, we will stick with the time-homogenous assumption.

An important implicit assumption in all SLCIS models we are aware of is that all customer nodes generate "identical" demands (possibly, within certain priority classes), i.e., that the streams of demand are indistinguishable with respect to the originating node once they reach the facility.

17.2.2 Facilities

Customer demands are serviced by the *facilities* that contain *service resources* (or "servers"). All aspects related to the facilities, including their number, locations, and the amount/ types of resources allocated to them can potentially be treated as decision variables in the model. In describing the system dynamics below we will initially treat the values of these variables as given, but will relax this assumption when describing model formulations later.

We will assume that facility locations must belong to some set I and that at most $m \geq 0$ facilities can be located; we will use $i \in I$, to represent the location (site) of facility i. By far, the most common assumption in SLCIS literature is that set I is discrete, i.e., that all potential locations for the facilities have already been enumerated. In this case, we can assume without loss of generality that $I \subset J$

(since any point in I not containing customers can be treated as a customer demand point with the maximum demand rate equal to 0). Other options, include $I \subset R^2$, leading to *continuous* SLCIS models (see, for example, Brimberg and Mehrez 1997 and Brimberg et al. 1997), or $I \subset J \cup A$ for a network $G = (J, A)$, leading to *network* SLCIS models (see, e.g., Berman et al. 2014). Unless stated otherwise, we will generally assume I to be discrete.

To take advantage of the discreteness of I we will follow the typical convention in location modeling and define $y_i \in \{0, 1\}$ to be a binary indicator variable with the value 1 if a facility is open at site $i \in I$, and 0 otherwise. To ensure that the total number of open facilities does not exceed m we require:

$$\sum_{i \in I} y_i \leq m. \tag{17.1}$$

If a facility is opened at $i \in I$ (i.e. $y_i = 1$), it must be allocated some service capacity $\mu_i > 0$, which can be thought of as the average processing rate. We will assume that $\mu_i = 0$ whenever $y_i = 0$, which can be enforced by

$$\mu_i \leq \mu^{\max} y_i, \quad i \in I, \tag{17.2}$$

where $\mu^{\max}$ is the maximum possible processing capacity that can be assigned to a facility.

As noted in Baron et al. (2008), there are two standard approaches to represent facility capacity in queuing environment: as a "single-server" facility where the capacity level can take on any value in some interval $\mu_i \in [0, \mu^{\max}]$, or as a "multi-server" facility housing $\kappa_i \geq 0$ parallel servers each with fixed capacity μ^0, where $\kappa_i \in \{0, \ldots, k\}$ is an integer, $\mu_i = \kappa_i \mu^0$ is the processing capacity of facility i, and k is the maximum number of servers that can be stationed at a facility (with $\mu^{max} = k\mu_0$).

While there are some important differences between the single-server and multi-server models (these will be touched on later) our bias is to favor the single-server representation. It is more transparent, typically leads to cleaner analytical results, and seems more practical as well: a typical facility will house a variety of processing resources and discrete "servers" may be hard to identify. For example, a medical clinic will often house doctors, nurses, examination rooms, X-ray machines, etc. While it is sensible for a planner to think of processing capacity of a clinic in terms of patients per hour (and how this processing capacity changes when certain resources are added or removed), it is harder to think of the clinic containing κ distinct servers (are these doctors? nurses? rooms?). Thus, unless stated otherwise, each facility will be assumed to house a single "server" with capacity μ.

We note that even in settings where μ is a continuous decision variable, it is sometimes useful to discretize it. This is because, as will be seen shortly, μ appears in many non-linear expressions for service levels and waiting times; discretization is a common trick used to linearize the corresponding expressions—this idea was first explored in Vidyarthi and Jayaswal (2014). When discretization is used, it is

assumed that the capacity μ_i of each facility i must satisfy

$$\mu_i \in \{\mu^1, \ldots, \mu^L\},$$

where $\mu^l, l = 1, \ldots, L$ represent a discrete set of options for service levels. Defining binary decision variables z_{il} which take on a value of 1 is $\mu_i = \mu^l$ and 0 otherwise, we can now write:

$$\mu_i = \sum_{l=1}^{L} z_{il}\mu^l, \ i \in I \tag{17.3}$$

$$\sum_{l=1}^{L} z_{il} = 1, \ \ i \in I, \tag{17.4}$$

The service times at each facility are assumed to be stochastic. More specifically, following Baron et al. (2008), we assume First Come First Serve (FCFS) service discipline and that service requirements (which can be thought of as the amount of work required to process one customer request) are independent and identically distributed random variables with a cumulative distribution function (CDF) $\mathscr{F}_S(w)$, and a well-defined moment generating function (MGF) $G_S(\eta)$. We also assume that the mean service time $E[S] = 1$. This assumption is made with no loss of generality as it simply rescales service times. Note that in this framework, since μ_i represents the service rate of facility i, the mean service time is $1/\mu_i$ and it is not hard to show that the distribution of service times is given by $F_S(\mu_i w)$ with MGF $G_S(\eta/\mu_i)$.

We define x_{ij} to be the *demand allocation* decision variables, specifying what portion of demand from customer node $j \in J$ is directed to facility $i \in I$. The key underlying assumption is that once the decisions about the number of facilities, their locations y_i and the service capacities μ_i for $i \in I$ are made, the demand allocations x_{ij} can be determined; the exact mechanism for determining demand allocations depends on the underlying assumptions about system dynamics and is described later. Mathematically, we assume that x_{ij} satisfies the following set of constraints

$$\sum_{i \in I} x_{ij} \leq 1, \ j \in J \tag{17.5}$$

$$x_{ij} \leq y_i, \ i \in I, \ j \in J \tag{17.6}$$

$$x_{ij} \in \{0, 1\}, \ i \in I, \ j \in J \tag{17.7}$$

These constraints are quite standard in location models: (17.5) ensures that at most 100% of customer demand from j is allocated to the facilities, (17.6) prevents allocating a customer to an unopened facility. Constraint (17.7) enforces the binary assumption for the allocation variables x_{ij}, with the value of 1 if the demand stream generated by customer node j is directed to facility i, and 0 otherwise.

The integrality of x_{ij} reflects the "single sourcing" assumption made in most SLCIS models, requiring each customer point to be assigned to at most one facility. An alternative is to allow "multi-sourcing", in which case x_{ij} is allowed to be continuous, by replacing (17.7) with its linear relaxation. We also note that constraints (17.5)–(17.7) represent "minimal" requirements on x_{ij}; they are often supplemented by other constraints describing the mechanisms by which allocation of customers to facilities is made.

We allow for the possibility that the demand from j is not assigned to any facility, i.e., $\sum_{i \in I} x_{ij} = 0$, which we interpret as the case of *lost demand*, i.e. demand that could have been captured but was lost, usually due to insufficient system capacity. The amount of lost demand is typically controlled via a penalty cost or constraints—we will return to these when we discuss specific model formulations below.

For each facility i we define the set $N_i = \{j \in J | x_{ij} = 1\}$, which represents the *service region* of facility i (clearly $N_i = \emptyset$ when $y_i = 0$). Observe that once λ_i and x_{ij} are known, the demand rate facing an open facility i is a Poisson process with rate

$$\Lambda_i = \sum_{j \in N_i} \lambda_j = \sum_{j \in J} \lambda_j x_{ij}. \tag{17.8}$$

As mentioned earlier, the Poisson property results from the fact that superposition of Poisson processes is also a Poisson process. Moreover, the demand streams faced by different facilities are independent of each other. Thus, each facility $i \in I$ acts as a stand-alone queueing system with Poisson arrivals and general service times, i.e., a $M/G/1$ (or $M/G/\kappa_i$) queue with service rate μ_i.

System stability (i.e., ensuring that queue lengths are finite) requires that

$$\Lambda_i \leq \mu_i, i \in I, \tag{17.9}$$

which acts as a constraint on capacity assignment decisions. In addition, the framework defined above allows us to express the key performance characteristics of the facilities, such as the steady-state system waiting time $W_i = W(\Lambda_i, \mu_i)$ (this includes both queueing and service times), and the steady-state number of customers in the system $L_i = L_i(\Lambda_i, \mu_i)$, both of which are random variables whose distributions can, in principle, be obtained. We will come back to these quantities when we discuss system costs and service-level constraints in the next section.

It may be also useful to require that each facility face some minimum demand rate Λ^{min} in order to ensure that it can be operated economically; sometimes these minimum demand rates are imposed by regulators for public service facilities (see, e.g., Zhang et al. 2010). These constraints take the form

$$\Lambda_i \geq \Lambda^{min} y_i, i \in I. \tag{17.10}$$

We note that many models make additional assumptions regarding the operations of facilities. For example, the assumption that the distribution of service times is exponential is quite common (though likely not very realistic in many real-life systems; e.g., see the discussion in Boffey et al. 2006). Some authors (e.g., Boffey et al. 2010) assume limited buffer space at the facilities. We will delay the discussion of these additional aspects until Sect. 17.5. For the moment we regard each facility as an infinite-buffer $M/G/1$ or $M/G/\kappa$ queue.

Remark The fact that each facility (once location, capacity and customer allocation decisions are made) can be viewed as an independent queueing system is the main characteristic distinguishing *immobile* from *mobile* server models; in mobile server models the systems operated by different facilities cannot be decoupled. This is because in these models the typical assumption is that server assignments are dynamic, i.e., depend on the state of the system: a server from a given facility may service demands from customers at point j under some conditions, but not under others. This leads to a system which is not, in general, separable, and where servers located at different facilities must be treated as distinguishable. Such queueing networks are analytically intractable even when all location, capacity and allocation decisions are made. Thus, all modeling approaches involve strong approximations and/or descriptive/simulation components (e.g., the Hypercube model proposed by Larson (1974) is frequently used as the modeling foundation).

In contrast, SLCIS models decompose into a set of queues with Poisson arrivals—systems for which strong analytical results (both exact and approximate) are available. We emphasize that this tractability relies on the static nature of customer-to-facility allocations: the demand allocations are determined once and then remain in force for all states of the system. Thus, SLCIS models where customers decide which facility to visit based on the current state of the system (e.g., based on posted information about current waiting times), or where other dynamic customer allocation mechanisms may be present, are likely to be closer (in terms of tractability and solution approaches) to models with mobile servers. On the other hand, models with mobile servers where static and non-intersecting service regions are assumed for all facilities (effectively assuming away dynamic customer reallocation) are quite similar to SLCIS models; many of the mobile server models reviewed in Berman and Krass (2002) fall into this group. Thus, instead of differentiating stochastic location models with mobile vs. immobile servers, it is more useful to differentiate models with dynamic vs. static assignments.

17.2.3 Costs, Revenues, and Constraints

To complete the description of the system it remains to specify two components: (1) the mechanisms by which customers are "allocated" to the facilities, expressed by the variables x_{ij} (which would also determine the actual demand rates λ_j, $j \in J$), and (2) the overall system costs and constraints assuring acceptable service levels.

We will postpone the discussion of (1) until Sect. 17.3, focusing on the costs and constraints in the current section and treating values of the key location, allocation, capacity assignment and demand level decisions $\{y_i, x_{ij}, \mu_i, \lambda_i\}, i \in I, j \in J$ as fixed. Following the common modeling practice, all costs below are assumed to be per unit time.

17.2.3.1 Travel Cost and Coverage Constraints

We assume that for each customer $j \in J$ and potential facility location $i \in I$ a distance metric $d(i, j)$ is defined, satisfying the regular properties of distance. The travel cost function $TC(d), d \geq 0$, representing the cost of traveling distance d is assumed to be non-decreasing and non-negative. This yields the System Travel Cost per time unit of

$$STC = \sum_{j \in J} \sum_{i \in I} TC(d(i, j))\lambda_j x_{ij}, \tag{17.11}$$

where we assume that constraint (17.6) ensures that customers are only assigned to open facilities. This expression merely states that the system travel cost is the sum of travel costs of all customers to their assigned facilities. We note that a frequent assumption is that the travel cost is a linear function of distance. More generally, since both J and I are discrete, one could simply redefine the distance measure to be $d'(i, j) = TC(d(i, j))$ for all $j \in J$, $i \in I$ and use this new measure in place of the original one. Thus, after suitably redefining distances and without loss of generality, we can write

$$STC = \beta \sum_{j \in J} \sum_{i \in I} d(i, j)\lambda_j x_{ij}, \tag{17.12}$$

where $\beta > 0$ is a parameter relating the travel cost to other terms in the objective function (the meaning of this parameter is discussed in Sect. 17.3). We will use this linear form in place of (17.11) from this point on.

A possible concern with the previous expression is that the short travel cost of one customer will be added to the long travel cost of another, resulting in the total quantity that may look reasonable, but will still provide poor service to some customers. To assure that no customer faces an unreasonably long travel distance, one can impose *coverage constraints*:

$$\sum_{i \in I} d(i, j)x_{ij} \leq R \quad \text{for all } j \in J, \tag{17.13}$$

where $R > 0$ is the "coverage radius", i.e., the maximum allowed travel distance for a customer to be "covered" by a facility (this constraint should be interpreted as referring to the "adjusted" distance measure that incorporates the travel cost,

as discussed above). We note that most SLCIS models will include either (17.12) or (17.13); while, in principle, both can be used in the same model, such usage is rare.

17.2.3.2 Congestion Costs and Service Level Constraints

While travel-related costs are present in all classes of location models covered in the current volume, the congestion-related costs and constraints are, of course, a defining feature of the stochastic location models with congestion, in particular of SLCIS models. As discussed earlier, the two common performance measures in a queueing system operated by each open facility $i \in I$ are the system waiting time W_i (recall that this includes the service time; a closely related measure is W_i^q which only covers the waiting time in queue) and the number of customers in the system L_i, which are random variables with certain steady-state distributions. The most common way to define congestion costs is in terms of expectations of these quantities, $\overline{W}_i$ and $\bar{L}_i$, respectively. Since the two are related by Little's Law, we will focus on the former (which is also more commonly used). For an $M/G/1$ queue, the expression for the mean waiting time in the system $\overline{W}$ can be found in any standard reference on queuing (see, e.g., Gross and Harris 1985, p. 255):

$$\overline{W} = \overline{W}^q + \frac{1}{\mu} = \frac{1+\gamma^2}{2}\frac{\rho}{1-\rho}\frac{1}{\mu} + \frac{1}{\mu} \tag{17.14}$$

where $\overline{W}^q$ is the expected time in queue, $\rho = \lambda/\mu$ is the utilization ratio and γ^2 is the squared coefficient of variation for service times, given by $\gamma^2 = \sigma^2\mu^2$, where σ^2 is the variance of service times. Each term in the expression for $\overline{W}^q$ has an intuitive interpretation. Recall that we are assuming Poisson arrivals, which have coefficient of variation equal to 1, and thus the term $\frac{1+\gamma^2}{2}$ represents the average squared coefficient of variation for arrival and service processes, often called the "variability factor" (for exponential service this term equals to 1). The second term, $\frac{\rho}{1-\rho}$ can be interpreted by recalling that ρ is the probability that the server is busy and thus $(1-\rho)$ is the probability that an arriving demand goes straight into service. The ratio can thus be interpreted as the length of the busy period measured in units of the length of the free period. The last term is simply the average service time per customer, sometimes known as the "scale effect" to recognize that as more capacity is assigned to the system, the average service time per customer declines. Thus

$$\overline{W}^q = [\text{Variability Factor}]\left[\frac{\text{Prob system busy}}{\text{Prob system free}}\right][\text{Scale Effect}]. \tag{17.15}$$

The expression for $\overline{W}$ simply adds the expected service time to the above.

Remark As noted earlier, two popular ways to represent the queueing system at a given facility are as either single-server $M/G/1$ queue with capacity μ, where μ is

a decision variable, or as a multi-server $M/G/\kappa$ system where each of the κ servers has capacity μ^0 and κ is the decision variable. If we set $\kappa\mu^0 = \mu$, i.e., require both systems to have the same processing capacity, we can ask to what extent are these systems "equivalent"? Can the simpler $M/G/1$ system be used as an approximation of harder-to-analyze $M/G/\kappa$ one?

First note that the coefficient of utilization ρ is the same when $\mu = \kappa\mu^0$. While no closed-form expression for $\overline{W}$ is known for the multi-server $M/G/\kappa$ case, a popular approximation (see e.g., Hopp and Spearman 2000, p. 273) is:

$$\overline{W} = \overline{W}^q + \frac{1}{\mu^0} = \frac{1+\gamma^2}{2}\frac{\rho^{\sqrt{2(\kappa+1)}-1}}{1-\rho}\frac{1}{\kappa\mu^0} + \frac{1}{\mu^0}, \tag{17.16}$$

which is very similar to (17.14): focusing on the expression for $\overline{W}^q$, we see that the only difference is that ρ in the numerator of (17.14) is replaced with $\rho^{\sqrt{2(\kappa+1)}-1}$ in (17.16). In fact, the latter approximates the probability that all servers are busy in the $M/G/\kappa$ system. Thus, each term in the intuitive interpretation (17.15) of $\overline{W}^q$ has the same interpretation for both systems. The only difference in the expected waiting times is that $M/G/1$ system is busy more frequently (since $1 > \rho > \rho^{\sqrt{2(\kappa+1)}-1}$), thus yielding larger values of $\overline{W}^q$. On one hand, the relative difference in $\overline{W}^q$ can be quite large (it approaches 100% as $\rho \to 0$). On the other hand, this difference should be small when ρ is close to 1 and waiting times in both systems are significant, while when ρ is small, the waiting times in both systems are quite small and the large relative difference may not be of practical significance. Thus, as a rough approximation, $M/G/1$ system can be used in place of $M/G/\kappa$ when the expected waiting times are of primary interest.

However, when the primary measure of interest is the expected total time in the system $\overline{W}$, one has to be more careful. When the system is highly utilized, i.e., ρ is close to 1, the main determinant of $\overline{W}$ is the waiting time and the previous argument applies. However, when the system utilization is lower, the expected service time will play a large role. Since it is $1/\mu^0$ for $M/G/\kappa$ and $1/\mu = \kappa/\mu^0$ for $M/G/1$, the former system will process customers κ times faster than the latter, and the approximation is no longer appropriate. Thus, with respect to $\overline{W}$, the approximation can only be justified in the heavy utilization case.

Turning our attention back to the $M/G/1$ system, we would like to rewrite (17.14) in terms of decision variables in our model. This is not difficult to do, and with a little algebraic manipulation we obtain the following expression for the expected waiting time at an open facility $i \in I$:

$$\overline{W}_i = \overline{W}_i^q + \frac{1}{\mu_i} = \frac{(1+\gamma^2)\Lambda_i}{2\mu_i(\mu_i - \Lambda_i)} + \frac{1}{\mu_i} \tag{17.17}$$

with Λ_i given by (17.8). We assume that $\overline{W}_i = 0$ if there is no facility at i.

One important question is how to treat the term γ^2 in the preceding expression. The "traditional" approach, adopted by all models described in the previous edition of the current text, has been to treat γ^2 as an intrinsic model parameter, rather than a decision variable, i.e., to assume that the coefficient of variation of service times is fixed in advance. While this is certainly the case when a specific distribution of service times is assumed (e.g., in $M/M/1$ queues $\gamma^2 = 1$), there is, in principle, no reason why this should not be a decision parameter in the system. For example, if the decision on how much capacity to install in facility i also deals with *what kind* of capacity to install, then the coefficient of variation γ could well be affected: service systems with higher level of automation may have lower γ, while more manual processes may have higher γ (of course the resulting values may be different at different facilities, so γ_i notation would have to be used). Another case where γ may be a decision variable is when customers at different nodes have different service time variabilities, in which case the allocation decisions x_{ij} may well influence not only Λ_i, but also the variability of service times γ_i. Nevertheless, the treatment of this parameter as exogenous, rather than a decision variable is quite common in SLCIS models; moreover its value is typically assumed to be identical at all facilities, which is reflected in our usage of γ without a subscript.

Several recent papers have relaxed the assumption that γ^2 is a fixed model parameter. One approach is to assume a one-to-one relationship between coefficient of variation of service times γ_i and service capacity μ_i at facility i, replacing γ^2 with $\gamma^2(\mu_i)$ in the previous expression. This idea is explored in Ahmadi-Javid et al. (2018), where γ_i is assumed to be a linear function of μ_i.

If the discretization of service times described by (17.3) and (17.4) is used, a very general relationship between μ and γ can be modeled. Recall that this approach assumes there are L discrete choices of capacity level. It is quite natural to assume that each choice $l \in L$ defines a pair (μ^l, γ^l) (in fact, two different choices could have identical capacity but different variability values). The coefficient of variation at facility i can now be written as

$$\gamma_i = \sum_{l=1}^{L} z_{il}\gamma^l, \tag{17.18}$$

where the decision variables $z_{il}, i \in I, l \in \{1, \ldots, L\}$ represent the choice of capacity level, as before. Now, for each fixed arrival rate Λ_i and capacity level l at facility i we can pre-compute the values of $\overline{W}_i^l(\Lambda_i)$ and write

$$\overline{W}_i(\Lambda_i) = \sum_{l=1}^{L} \overline{W}_i^l(\Lambda_i) z_{il},$$

which is linear in the decision variable. If, in addition we assume that Λ_i is discrete (which is natural in many contexts), we can further simplify the previous expression, while allowing for different coefficients of variation at different facilities (at the cost,

of course, of the approximation inherent in the discretization approach). Variations of this approach are used in Ahmadi-Javid and Hoisenpour (2018), Azizi et al. (2017), and Schön and Saini (2018).

Another observation regarding (17.17) is that $\overline{W}_i$ (and $\overline{W}_i^q$) is decreasing in μ_i, increasing in Λ_i and convex with respect to both μ_i and Λ_i whenever system stability conditions (17.9) hold. These properties are exploited in several SLCIS models that follow.

Let $WC(w)$ represent the "waiting cost", i.e. the cost incurred by customers waiting w units of time in the system (here, and hereafter, we assume that waits include service times; an equivalent treatment can be developed by focusing on waiting times in queue only, i.e. W^q). As with the travel costs, we assume that $WC(w)$ is non-negative and non-decreasing, noting that many models make the simplifying assumption that the waiting cost is proportional to w. The total expected waiting cost in the system can now be expressed as

$$SWC = \sum_{j\in J}\sum_{i\in I} WC(\overline{W}_i)x_{ij}. \tag{17.19}$$

In view of non-linear dependence of the expected waiting time $\overline{W}_i$ on the decision variables, SWC is a non-linear function even when the waiting cost is assumed to be linear.

We note that since the waiting cost is only incurred by customers who are assigned to some facility, we should also add a penalty term for customers that are not assigned to any facility (i.e., not served)—otherwise the model may have an incentive to not assign customers even if service capacity is available. The "lost demand" customers may be represented in the revenue term described later (i.e., they are treated as an opportunity cost of lost revenue). Alternatively they can be represented by a term $p\sum_{j\in J}\left(1-\sum_{i\in I}x_{ij}\right)$ which may be added to the SWC expression above, where p represents the penalty for not servicing a customer.

There are two potential issues with using (17.19) as the *sole* measure of service quality (in terms of waiting times) at the facilities. First, as with the system travel cost, a small value of SWC does not necessarily ensure that all customers are receiving adequate service—a small expected waiting time at one facility may "hide" a large expected waiting time at another. Thus, one may want to add the constraints (these are traditionally stated in terms of waiting time, rather than system time; we follow this tradition):

$$\overline{W}_i^q \le EW, \quad i \in I, \tag{17.20}$$

where EW represents the acceptable maximum waiting time at any facility.

Second, the *expected* waiting time may not be sufficient to express the desired service quality; we may wish to ensure that most customers experience no waiting at all or that the probability of "long" waits is sufficiently low. For this we need to

consider a constraint of the form

$$P(W_i^q > T) \leq \alpha_T, \quad i \in I, \tag{17.21}$$

where $P(\cdot)$ is the steady-state distribution of W_i^q, $T > 0$ is the specified threshold for the waiting times, and $\alpha_T \in (0, 1)$ is the maximum acceptable probability of waits longer than T at any facility. For example, α_0 represents the maximum acceptable proportion of customers that must wait for service at any facility.

Both (17.20) and (17.21) above are examples of *Service level Constraints* (SCs) that are quite common in SLCIS models. Since (17.20) refers to the expected behavior of the system, while (17.21) refers to the probability of occurrence of certain (undesirable) events, we will refer to the former as the "Mean SC" and the latter as the "Probabilistic SC". While the Mean SC is easily expressed in terms of the decision variables by substituting (17.17) into (17.20), the Probabilistic SC requires an expression for the steady-state distribution of the waiting time, which is not generally available. One option is to make additional assumptions about the distribution of service times (e.g., assuming $M/M/1$ or $M/E_k/1$ queues at the facilities) since steady-state distributions of waiting times have been derived for many common systems. Another option is to use an approximation. The one we follow here is based on Baron et al. (2008). Assume that the service constraints (17.21) are specified and let

$$V(T, \alpha_T) = -\frac{ln(\alpha_T)}{T};$$

observe that since $ln(\alpha_T) < 0$, this is a positive constant that is decreasing in α_T and in T. Then (under certain mild technical assumptions), constraint (17.21) is satisfied whenever

$$G_S(\frac{V(T, \alpha_T)}{\mu_i})(\Lambda_i - 1) \leq V(T, \alpha_T), \tag{17.22}$$

where $G_S(\cdot)$ is the MGF of service times defined earlier. Recall that $G_S(\eta)$ is an increasing function for $\eta > 0$, implying that the left-hand side of (17.22) is decreasing in μ_i. This is quite intuitive: when T or α_T are decreased, the probabilistic SC becomes tighter, requiring more capacity at the facility. In fact, as $V(T, \alpha_T)$ becomes larger, satisfying (17.22) requires more capacity μ_i.

This leads to a general view of service constraints: for any arrival rate Λ_i at facility $i \in I$ one can define a minimum capacity level $\bar{\mu}(\Lambda_i)$ such that SC holds if and only if

$$\mu_i \geq \bar{\mu}(\Lambda_i), \tag{17.23}$$

where $\bar{\mu}(\Lambda_i)$ is computed (perhaps numerically) from (17.20), (17.21), or (17.22). Of course, an equivalent view is to specify a function $\bar{\Lambda}(\mu)$, which is just an inverse

of $\bar{\mu}(\Lambda)$, so that SC holds whenever

$$\Lambda_i \leq \bar{\Lambda}(\mu_i), \tag{17.24}$$

i.e., for a given capacity level μ_i there is a maximal arrival rate $\bar{\Lambda}(\mu_i)$ for which an adequate service level can be provided by facility i. This view extends to other definitions of SCs (e.g., instead of using waiting time one could use L or another service level measure)—the only thing that changes is the way functions $\bar{\mu}(\Lambda)$ and $\bar{\Lambda}(\mu)$ are computed.

We note that system stability conditions imply that $\bar{\mu}(\Lambda) > \Lambda$ (equivalently $\bar{\Lambda}(\mu) < \mu$) and the difference $\bar{\mu}(\Lambda) - \Lambda$ may be interpreted as the amount of the "capacity cushion" (capacity in excess of the minimal possible level) needed to ensure adequate service given the arrival rate Λ. For many systems and many specifications of service level constraints it has been shown that this amount grows proportionately to $\sqrt{\Lambda}$, i.e.

$$\bar{\mu}(\Lambda) \approx \Lambda + Q\sqrt{\Lambda} \tag{17.25}$$

for some constant Q (see, e.g., the discussion in Castillo et al. 2009). The derivations in Whitt (1992) suggest that, under many conditions, a good approximation for Q is provided by

$$\sqrt{2}Q \approx \sqrt{\gamma^2 + 1}P(W > 0).$$

Thus, $\sqrt{2}Q/\sqrt{\gamma^2 + 1}$ is approximately equal to the probability of waiting, a natural service level measure. To summarize, when the probability of waiting is used as the service-level measure, the constraint

$$P(W_i > 0) \leq \alpha_0, \ i \in I$$

holds if

$$\mu_i \geq \bar{\mu}(\Lambda_i) \approx \Lambda_i + \left[\sqrt{\frac{\gamma^2 + 1}{2}}\alpha_0\right]\sqrt{\Lambda_i}, \ i \in I. \tag{17.26}$$

Similar expressions can be derived with for service level measures where the threshold for waiting time is set above 0.

As noted earlier, incidence of long waits can be controlled through service level constraints and/or explicit waiting cost terms in the objective function. While, in principle, both can be used in the same SLCIS model, it is far more common to use one or the other. In models where only service level constraints are used, these constraints will be tight in an optimal solution (since capacity is costly). If, in addition, the demand is assumed to be inelastic, Λ_i is a linear function of the

decision variables x_{ij}. In this case a significant simplification is achieved by using the previous expression: setting the SC as an equality, we can eliminate decision variables μ_i from the model, replacing them with the right-hand side of (17.26).

17.2.3.3 Facility Costs

We assume that the decision to open a facility at $i \in I$ incurs two types of costs: the *fixed cost* FC_i, which depends on the characteristics of the location i, and the *variable cost* $VC(\mu_i)$, which depends on the amount of capacity μ_i allocated to the facility. The function $VC(\mu)$ is assumed to be non-decreasing and non-negative with $VC(0) = 0$; concavity of $VC(\mu)$ is a frequently made assumption, reflecting economies of scale. With these definitions, the System Facility Cost is defined as follows:

$$SFC = \sum_{i \in I} FC_i y_i + \sum_{i \in I} VC(\mu_i) \tag{17.27}$$

17.2.3.4 Revenues and Overall Objectives

We assume that each customer that is served brings in a revenue r to the system (for public service applications, we can treat r as a "system benefit" parameter). The total expected revenue can now be expressed as

$$SR = r \sum_{i \in I} \Lambda_i = r \sum_{j \in J} \lambda_j \sum_{i \in I} x_{ij}. \tag{17.28}$$

In principle, parameter r can be treated as a decision variable—the price charged by the decision-maker for service. However, in the majority of SLCIS literature this term is treated as an exogenous parameter (Tong 2011 and Berman et al. 2014 being the exceptions). Since treating prices as decision variables introduces significant new complications, we will generally treat r as constant in the model.

We also observe that when demand is inelastic (i.e., $\lambda_j = \lambda_j^{\max}$ for all $j \in J$) and when the constraints require that all customers must be served (i.e., $\sum_{i \in I} x_{ij} = 1$, $j \in J$), it is easy to see that $SR = r \sum_{j \in J} \lambda_j^{\max}$, which is a constant. In this case, the revenue term in the objective can be dropped, leading to a pure cost minimization case. Even in models where some customers may not be served, but the demand is inelastic, it is common to use cost minimization with a penalty term, which can be interpreted as opportunity cost for unserved customers.

To summarize, the overall objective for a general SLCIC model is given by

$$\text{maximize}\,[SR - STC - SWC - SFC],$$

where the respective components are defined by (17.28), (17.12), (17.19), and (17.27). We note that in most specific models described in the literature, only a subset of the terms above is present, the rest being implicitly controlled by constraints (e.g., in the presence of service level constraints, the SWC term is often dropped).

Most of the terms above depend on demand allocations x_{ij} and demand rates λ_j, which have not yet been described. This is the subject of the following section.

17.3 Customer Response: Demand Levels and Allocations

In this section we discuss the mechanism determining the allocation of customer demand to facilities, represented by x_{ij} variables, and the amount of demand λ_j generated by customers at $j \in J$.

In location modeling two approaches for allocating customer demand to facilities are generally considered: *directed choice*, where the same decision-maker determining the number and locations of the facilities also has the power to assign customers to the facilities in a way that will optimize the model objective, and *user choice* where customers self-assign to facilities based on maximization of their own utility functions which may not be aligned with the overall model objective. For example, a common customer utility function is the travel distance. Thus, in a user choice environment, each customer will select the closest facility, while in the directed choice case a customer may be assigned to a further facility even when a closer one is open (if such assignment reduces the overall facility cost).

The same framework can be applied to the SLCIS models. However it may be more useful to also classify the models in terms of the assumed customer reaction to the service offered by the facilities. We differentiate four classes of models:

Type NR: Models with no customer reaction: customers do not control the demand allocations and the demand rates are fixed (directed choice with inelastic demand)

Type AR: Models with allocation-only reaction: customers select utility-maximizing facilities, but the demand rates are fixed (user choice with inelastic demand)

Type DR: Models with demand rate-only reaction: customer do not control the demand allocations but do determine the demand rates (directed choice with elastic demand)

Type FR: Models with full customer reaction: customers control both, the allocation of demand (by selecting the utility-maximizing facilities) and the demand rates (user choice with elastic demand).

This classification is summarized on Table 17.1.

The *NR models* correspond to the standard directed choice assumptions in the literature: the values of the assignment variables x_{ij} are entirely controlled by the decision-maker and must only satisfy the basic constraints (17.5)–(17.7). One may

Table 17.1 Model classification by customer response

	Demand allocation	
	Decision-maker	Customer
Inelastic demand	NR	AR
Elastic demand	DR	FR

also interpret such models as describing a "social optimum" (also known as "first best solution" in economics)—the customers will accept whatever assignments are needed to optimize the overall system objective, even if that means that some of them may have to travel to more distant and more congested facilities than the ones available in their immediate neighborhood. On the other hand, since the objective function combines the costs borne by the decision-maker (facility costs SFC) with those borne by the customers (travel cost STC and waiting cost SWC), the interests of both parties should be "balanced" in the solution. Customer demand is assumed to be inelastic, with $\lambda_j = \lambda_j^{\max}$ for all $j \in J$. Since customer utility has no effect in this model, there is no need to define it. We note that x_{ij} are usually assumed to be binary in NR models (though it is easy to construct examples showing that higher objective values may be possible with fractional assignments). This is due to the concern that enforcing fractional demand allocations is likely impractical in most contexts. Thus, in NR models only the "minimal" constraints (17.5)–(17.7) need to be imposed on demand allocations: the decision-maker is free to choose any allocation that satisfies these constraints.

The other three model types assume some form of customer reaction in the form of utility-maximizing behavior. The description of the utility mechanism is provided next.

17.3.1 Customer Utility Functions

Recall that u_j is the utility derived by customer $j \in J$ from the service provided by the facilities. Note that there are two costs borne by the customer: travel and waiting. Suppose a customer experiences travel distance d (as before we assume that distances have been redefined to represent travel costs) and expected system waiting time. Let the utility $U(d, w)$ be a non-increasing function of d and w. To relate u_j to $U(d, w)$ we assume that the total utility derived by customer j is only affected by the facility this customer actually visits. Since $\sum_j x_{ij} \leq 1, x_{ij} \in \{0, 1\}$, this leads to

$$u_j = \sum_{i \in I} U(d(i, j), \overline{W}_i) x_{ij}, \tag{17.29}$$

Note that this definition remains valid even when the single-sourcing assumption is relaxed. In this case, $x_{ij} \in [0, 1]$ represents the proportion of time facility i is used by customer j, and u_j can be interpreted as the resulting *expected* utility. Observe

also that if a customer does not receive service from any facility, $x_{ij} = 0$ for all $i \in I$ and $u_j = 0$.

Perhaps the most natural specification for the utility function $U(d, w)$ is the linear form

$$U^L(d, w) = -(\tau_d d + \tau_w w), \tag{17.30}$$

where $\tau_d, \tau_w > 0$ are the relative weights on travel distance and waiting time, respectively. When $\tau_w = 1$, the parameter τ_d can be interpreted as the average travel speed, so that $\tau_d d$ is the average travel time, and the right-hand side of (17.30) represents the negative of the total expected time spent by the customer in the system (until the end of service).

There are two other common specifications of $U(d, w)$. The simpler one is

$$U^D(d, w) = -\tau_d d, \tag{17.31}$$

i.e., customer's utility is simply proportional to the traveling distance (representing the travel cost) and is independent of the waiting time. This is a very popular specification form appearing (often implicitly) in numerous SLCIS models. While the lack of dependence on w may seem counterintuitive, it is usually justified by assuming that customers do not have advance knowledge of waiting times at the facilities and thus must make their decisions based on travel times only (though in a steady-state system some learning about expected waiting times should, presumably, occur). Alternative justification is that the waiting costs are dominated by the travel costs. Perhaps more importantly, as will be seen below, specification (17.31) avoids many technical complications that occur when a more general utility structure is used and can thus be treated as an approximation.

Another natural specification is the log-linear form

$$U^E(d, w) = exp(-\tau_d d - \tau_w w), \tag{17.32}$$

which is quite similar to (17.30) with the advantage of the utility being non-negative, convex and bounded by 1. Note that $U^E(d, w) = 1$ when $d = w = 0$, i.e., when the customer incurs neither travel nor waiting cost, and $U^E(d, w) \to 0$ as $d, w \to \infty$. This makes it convenient to interpret $U^E(d, w)$ as *the proportion of maximum available demand realized from customer j if this customer is faced with travel distance d and expected wait w*. This interpretation will be useful when describing elastic demand models below.

Finally, we note that a utility function can be defined in terms of service measures other than the expected waiting time—one can use the probability of waiting $P(W^q > 0)$, or any other performance measure of the queuing system operated at the facilities. The specifications of the utility can also be generalized to incorporate other decision variables, such as the price charged by the facility operator for service (see Berman et al. (2014) for an example).

17.3.2 SLCIS Models with Customer Reaction

Once a utility function is specified, it should be possible to specify the customer reaction as well. At a first glance, this seems fairly straightforward: a SLCIS model with customer reaction can be viewed as a Stackelberg Game, where the decision-maker first specifies the number, locations and capacities of the facilities (i.e., values of m, y_i and μ_i for $i \in I$) and then each customer selects a utility-maximizing strategy, i.e. allocates their demand to the utility-maximizing facility. Unfortunately, as we will see shortly, this may lead to situations where no equilibrium solution (i.e., set of choices for all customers) exists.

One fundamental issue is the implicit assumption that faced with the same set of alternatives (here, set of open facilities and processing capacities) customers always make the same choice. There is a rich body of research in marketing and economics that suggests that this may not be the case. A related question is how well can the customers measure their own utility? After all, if the utility function includes waiting times, a stochastic element is automatically present in measuring $U(d, w)$. Other stochastic elements, including uncertainties about travel times or even the non-waiting time aspects of the quality of the service interaction at the facility may also be present. Game Theory and Marketing literature have defined two notions of utility: deterministic and stochastic, with the associated large bodies of research. SLCIS literature have also adopted these two different notions of utility, leading to distinct classes of models.

As discussed below, in order to ensure the existence of equilibrium in deterministic utility models one has to allow for fractional choice, where the customers allocate their purchases among many (possibly all) facilities. Thus, the random choice element naturally enters in the deterministic utility setting, with the allocation vector derived from the equilibrium conditions. This set of models is discussed next.

An alternative approach, discussed in Sect. 17.3.2.4 is to assume a Proportional Allocation (PA) mechanism, where customers allocate their demand among the available facilities proportionally to the utility derived from each facility. The main advantage of this approach is that the allocation vector is specified from the start in closed form, leading to a simpler structure. Moreover, if one assumes a stochastic utility setting together with some additional assumptions, the (PA) mechanism naturally arises, providing additional axiomatic justification to this model class.

17.3.2.1 Customer Reaction Models with Deterministic Utility 1: Models with Allocation-Only Reaction (AR)

Here we assume that, once the facility locations and service capacities are determined by the decision-maker, the customer allocates their demand so as to maximize their deterministic utility function $U(d, w)$. Moreover, AR models assume that the demand rate of each customer node is fixed *a priori*, with $\lambda_j = \lambda_j^{\max}$ for all $j \in J$. For concreteness, we will assume the linear specification of the utility function

$U^L(d, w)$ given by (17.30), though much of the discussion extends to alternative specifications as well.

Even in this relatively simple setting complications quickly arise. This has to do, primarily, with the fact that customer utility is a function of the waiting time $\overline{W}_i$, which is not directly controlled by the decision-maker, but rather arises as a result of joint actions of the decision-maker and *all* customers: the former determines facility locations and capacities μ_i, while the latter determine the demand rates Λ_i. This gives rise to traffic equilibrium conditions, where the actions of one customer (adjusting their demand rate λ_j and/or demand allocation x_{ij}) change the waiting times at the facilities and thus affect the utilities of all other customers. Thus, not only is there a bi-level game being played between the decision-maker and the customers, but there is also a simultaneous non-cooperative game being played between the customers themselves. Moreover, the response functions in the latter are rather complicated, which may lead to lack of equilibria (if customers are restricted to simple strategies), or to multiple equilibria, not to mention serious difficulties in computing these equilibria. We discuss these issues briefly below, referring the interested reader to more general references on spatial equilibria, e.g., Nagurney (1999).

Consider first the original "single-sourcing" assumption, i.e. that a customer will only patronize a single facility. Utility maximization implies that if $x_{ij} = 1$ for some $i \in I$ and $j \in J$, then

$$U^L\left(d(i, j), \overline{W}_i\right) \geq U^L\left(d(k, j), \overline{W}_k\right) \quad \text{for all } k \in I \text{ with } y_k = 1,$$

which, assuming for simplicity that $\tau_w = \tau_d = 1$ in (17.30), is equivalent to

$$d(i, j) + \overline{W}_i \leq d(k, j) + \overline{W}_k \quad \text{if } y_k = 1, k \in I.$$

Recalling that Λ_i is given by (17.8) and $\overline{W}_i$ by (17.17), this leads to the following equilibrium conditions that must be satisfied by allocations x_{ij}:

$$d(i, j) + \overline{W}_i \leq [d(k, j) + \overline{W}_k]y_k + M(1 - x_{ij}), \quad i, k \in I, j \in J \tag{17.33}$$

$$\overline{W}_i = \frac{(1 + \gamma^2)\Lambda_i}{2\mu_i(\mu_i - \Lambda_i)} + \frac{y_i}{\mu_i + M(1 - y_i)}, \quad i \in I \tag{17.34}$$

$$\Lambda_i = \sum_{j \in J} \lambda_j^{\max} x_{ij}, \quad j \in J \tag{17.35}$$

$$\sum_{i \in I} x_{ij} \leq 1, \quad j \in J \tag{17.36}$$

$$x_{ij} \leq y_i, \quad i \in I, j \in J \tag{17.37}$$

$$x_{ij} \in \{0, 1\}, \tag{17.38}$$

where M is a suitably large constant. We assume that some finite limit can be imposed on the expected waiting time $\overline{W}_i$ at any facility and that $M \geq d(i, j) + \overline{W}_i$ for all j and i.

Of course a trivial solution to this system is to have $x_{ij} = 0$ for $j \in J, i \in I$ (which also implies $\overline{W}_i = 0$ for all $i \in I$), i.e., to have complete loss of all customer demand. Clearly, we are interested in non-trivial solutions where at least some customers choose to obtain service. On the other hand, the system may not have enough capacity to serve all customers. We therefore make the following definition.

Definition 17.1 A subset of customer nodes $J' \subset J$ is **serviceable** if

$$\sum_{j \in J'} \lambda_j^{\max} \leq \sum_{i \in I} \mu_i.$$

A subset J' is **fully served** if $\sum_{i \in I} x_{ij} = 1$ for all $j \in J'$, i.e. if (17.36) holds as equality for all $j \in J'$.

This definition simply assures that there is sufficient capacity to serve any serviceable subset. We are interested solutions where at least some serviceable subsets of J are fully served. Unfortunately, the system (17.33)–(17.38) may have no such solutions.

Example 17.1 Consider a network with one customer node j and two facility nodes $0, 1$ both of which contain facilities, i.e., $y_0 = y_1 = 1$. Assume further that $\mu_0 = \mu_1 > \lambda_j^{\max}$, and thus $J = \{j\}$ is serviceable. Assume $d(j, 0) = d(j, 1)$. Then, since $W_i = 0$ if $x_{ij} = 0$ and $W_i > 0$ when $x_{ij} = 1$ for $i = 0, 1$, there is no feasible solution to the system (17.33)–(17.38). Indeed, if customers at j select facility i, it creates non-zero waiting time at that facility, making the other facility a utility-maximizing choice. Other similar examples of non-existence of equilibria with binary allocation vectors are easy to construct. □

The underlying reason for the phenomena illustrated above is that single-sourcing strategies create discontinuities (a facility receives either all of customer's demand, or none of it), while the existence of equilibria typically requires continuity of the underlying functions. Indeed, intuitively it is clear that in the previous example equilibrium allocations are achieved if the customers at j visit each facility with equal frequency. This, of course, requires the relaxation of the single-sourcing assumption, allowing x_{ij} to take on fractional values, which are interpreted as visit frequencies. In addition to replacing (17.38) with its linear relaxation, the equilibrium-defining inequality (17.33) has to be adjusted as follows.

Recall the definition of u_j given by (17.29), which is now interpreted as the expected utility for customers at $j \in J$ given a fractional allocations vector x_{ij}, $j \in J, i \in I$ (we emphasize that the waiting times are affected by the allocations of all customers, not just the ones at j). We seek allocations under which no customer can improve their utility by making unilateral changes. It follows that the equilibrium

utilities u_j^*, $j \in J$ must satisfy

$$d(i,j) + \overline{W}_i \begin{cases} = -u_j^* \text{ if } x_{ij} > 0; \\ \geq -u_j^* \text{ if } x_{ij} = 0 \end{cases} \tag{17.39}$$

(recall that we are assuming linear utilities which are equal to the negative of total travel and waiting times). These conditions can be represented by replacing (17.33) with the following complementarity conditions:

$$d(i,j) + \overline{W}_i \geq v_j, \ \ j \in J, i \in I \tag{17.40}$$

$$(d(i,j) + \overline{W}_i - v_j)x_{ij} = 0, \ \ j \in J, i \in I \tag{17.41}$$

$$v_j \geq 0. \tag{17.42}$$

Note that for a feasible solution we must have $v_j = -u_j^*$, indicating that the new decision variable represents the equilibrium "disutility" for customers at $j \in J$. We will refer to a solution of the system (17.34)–(17.42) as *Customer Flow Equilibrium*.

The following result follows directly from Theorem 5.4 of Ashtiani and Magnanti (1981) by continuity of $U\left(d(i,j), \overline{W}_i(\boldsymbol{x})\right)$ for all $j \in J, i \in I$, where $\boldsymbol{x}$ is a fractional allocation vector with components x_{ij}.

Theorem 17.1 *For any values of $y_i \in \{0,1\}$ and $\mu_i \geq 0$ such that $\mu_i \leq My_i$, if a subset $J' \subset J$ is serviceable, then there exists at least one customer flow equilibrium x_{ij}, $j \in J, i \in I$ under which J' is fully served.*

In particular, if the system has the capacity to service all of customer demand, i.e., J is serviceable, at least one customer flow equilibrium must exist under which all customers are served.

The discussion and the result above is quite general: in particular, they extend models with elastic demand (i.e., models of type FR discussed below). Additionally, in place of the expected waiting time for an $M/G/1$ queue, a general measure of "congestion" can be used with the only requirements that it is strictly increasing, twice differentiable, non-negative and convex (recall that all capacity decisions are considered to be fixed in this section). These requirements are clearly satisfied by most performance measures for queueing systems, including multi-server and limited-buffer queues. We refer the reader to Brandeau et al. (1995) for a discussion of these more general settings.

It is important to realize that the customer flow equilibrium may not be unique. In fact, as illustrated in the following example, there may be multiple allocation vectors satisfying the equilibrium conditions for a particular fully served subset of customer nodes.

Example 17.2 Consider adding a second identical customer node j' to the system in Example 17.1. Now, if customers at both nodes are assigned to different facilities: $x_{ij} = 1, x_{(1-i)j} = 0, x_{ij'} = 0, x_{(1-i)j'} = 1$ for $j = 0, 1$, we have two different

equilibria. In fact, there may be infinitely many equilibria: any assignment satisfying

$$x_{ij} = \alpha, x_{(1-i)j} = 1 - \alpha, x_{ij'} = 1 - \alpha, x_{ij'} = \alpha, \quad \alpha \in [0, 1]$$

is also an equilibrium. □

In principle, different equilibrium allocation vectors may lead to different values of the objective function in the underlying SLCIS model, creating uncertainty as to which solution will actually arise. However, all equilibria are "similar" in certain key aspects, as shown in the following theorem based on the result in Brandeau and Chiu (1994):

Theorem 17.2 *For any two customer flow equilibria under which the same subset $J' \subset J$ is fully served, the values of $\Lambda_i \;\; i \in I$ (total demand seen at each facility) and $u_j, \;\; j \in J$ (equilibrium utility of each customer) are the same.*

This theorem implies that, under a sensible specification of the objective function, where the total travel and waiting cost for each customer node is a function of u_j, all equilibria will give rise to the same values of the objective.

While the previous results show that AR models with multi-sourcing demand allocations are well-posed, there is an important issue concerning computational tractability of system (17.34)–(17.42). Even for fixed facility locations and capacities, solving the customer flow equilibrium conditions is far from easy. While the system is a linear complementarity problem with respect to variables v_j, $\overline{W}_i$ and x_{ij}, the waiting time is, in general, non-linear with respect to the capacity decision μ_i, resulting in a non-linear complementarity problem, which is often computationally challenging.

While certain numerical approaches (described in Nagurney 1999) do exist, they are computationally heavy even for moderate-size problems (see Tong 2011). Often, to get reasonable algorithmic efficiency one has to make simplifying assumptions about the system. For example, assuming $M/M/1$ allows for a variable substitution $\mu_i = \lambda_i + 1/\overline{W}_i$, where the waiting times, rather than capacities, are used as decision variables. This turns the equilibrium conditions into a linear complementarity problem, making the system much more solvable. Zhang et al. (2010) were able to compute equilibria for such a system with $|J| \approx 500$ and $|I| \approx 40$ (note that their model also had elastic demands, which likely increased computational complexity). However, computing the equilibrium is only a subproblem of an SLCIS model. Thus embedding even a simplified computation in an overall exact optimization procedure is very computationally challenging. Hence both of the papers cited above resort to search heuristics for the upper level (location and capacity allocation decisions).

An interesting recent development was presented in Aboolian et al. (2016) who show that for the $M/M/1$ system traffic equilibrium constraints can be linearized through the introduction of additional binary variables $z_{ij} = 1$ if $x_{ij} > 0$ (i.e. customer j makes some use of facility i) and $z_{ij} = 0$ if $x_{ij} = 0$. It is not clear if this approach can be extended to non-$M/M/1$ settings.

In view of the difficulties involved in using the customer flow equilibrium approach above, it is natural to think of model simplifications. We mention two such approaches. One is to keep the single-sourcing assumption in spite of the possible non-existence of equilibria (see Zhang et al. 2009). The reason this may be reasonable is that, as mentioned earlier, non-existence is a result of discontinuity—when re-assignment of a single customer alters the waiting times at the facility for the remaining customers. It is reasonable to assume that for realistic problem instances, this should not be an issue: as the number of customers and customer nodes grows, no single assignment should exert a significant impact on waiting times at the facilities. Thus, asymptotically, single-sourcing equilibria should emerge. Indeed, Zhang et al. (2009) did not report issues with non-existence of equilibria when solving realistic-size problem instances for mammography clinics in Montreal, Canada. The obvious advantage of the single-sourcing approach is that the system (17.33)–(17.38) is much easier to solve and can be embedded as part of constraints in a larger SLCIS model.

The second approach is to use distance-only utilities $U^D(d)$ given by (17.31). Since these are independent of waiting times, the existence of customer flow equilibria is no longer an issue; utility-maximizing behavior by customers merely implies that once facility locations are specified, each customer travels to the closest facility, replacing (17.33) with

$$d(i, j) \leq d(k, j)y_k + M(1 - x_{ij}), \quad i, k \in I, j \in J, \tag{17.43}$$

which leads to significant simplifications (obviously, single-sourcing assumption can be retained here as well).

Another alternative, which bypasses some of the difficulties discussed above, is to use stochastic utility model, which is discussed in Sect. 17.3.2.4.

17.3.2.2 Customer Reaction Models with Deterministic Utility 2: Models with Demand-Only Reaction (DR)

In this model class, the decision-maker has the control of the demand allocation vector $\boldsymbol{x}$, however, the demand $\lambda_j = \lambda(u_j)$ for customer node $j \in J$ is assumed to be a function of the utility u_j realized by customers at j. Following Brandeau et al. (1995) we assume that

$$\lambda_j = \lambda_j^{\max} h(u_j),$$

where, as defined earlier, $\lambda_j^{\max}$ is the maximum possible demand rate at node j and $h(u) \in [0, 1]$ is a strictly decreasing, twice differentiable function with $h(0) = 1$ and $h(u) \to 0$ as $u \to u_j^{min}$, where u_j^{min} is the lower bound on the utility for customers at j (e.g., if utilities are scaled to be non-negative, then we can set $u_j^{min} = 0$). Thus,

$h(u_j)$ can be interpreted as the percentage of the maximum available demand at j that is "captured" by the system; it is often called the "participation rate".

Recall that by (17.29), the utility u_j is a function of the waiting time and travel distance faced by customers at j. As in the case of NR models, we will assume that x_{ij} is binary, motivated by the same considerations: when customer demand allocations are dictated by the decision-maker, rather than by an equilibrium condition of the previous section, enforcing fractional assignments is typically unrealistic. Thus, assuming all customers at j will be served (as will be shown below, this assumption holds automatically in DR models), $x_{ij} = 1$ for exactly one $i = i(j) \in I$. Then, the demand from customer j that is captured in response to the offered travel distance of $d(i(j), j)$ and waiting time $\overline{W}_{i(j)}$ is given by the composition of the decay function and the utility functions by:

$$\lambda_j(d(i(j), j), \overline{W}_{i(j)}) = \lambda_j^{\max} h(U(d(i(j), j), \overline{W}_{i(j)})), \quad j \in J. \tag{17.44}$$

One example of a functional forms that satisfy the required assumptions is the identity function $h(u) = u$ together with the exponential utility U^E given by (17.32), leading to the popular "exponential decay" demand specification:

$$\lambda_j(d(i(j), j), \overline{W}_{i(j)}) = \lambda_j^{\max} \exp(-\tau_d d(i(j), j) - \tau_w \overline{W}_{i(j)}), \quad j \in J. \tag{17.45}$$

While this expression is assumed in several published DR models, most of the results below apply to more general functional forms as well. Observe that (17.44) implicitly defines an equilibrium condition: the left-hand side depends on the waiting time $\overline{W}_{i(j)}$ at facility $i(j)$, which is a function of demand $\Lambda_{i(j)} = \sum_{j \in J} \lambda_j x_{i(j),j}$ seen by this facility. Thus, (17.44) should be seen as a system of $|J|$ equations that must be solved to yield the actual demand rates; this system decouples into subsystems consisting of all customers $j \in J$ assigned to facility i (i.e., with $i(j) = i$) for each open facility (i.e., $y_i = 1$). Thus, even though the allocation variables x_{ij} are fixed (or, rather, set by the decision-maker) for DR models, the issues related to existence and uniqueness of equilibria must be dealt with. The following result is based on Berman et al. (2014), where it is established for the case where price r is also a decision variable.

Theorem 17.3 *For any given facility location, capacity, and demand allocations y_i, μ_i, x_{ij} for $i \in I, j \in J$, there exist a unique equilibrium arrival rates $\lambda_j(d(i(j), j), \overline{W}_{i(j)})$ and waiting times $\overline{W}_i$.*

Note that, unlike the case for AR models, this result holds with binary demand allocations x_{ij} (it obviously extends to the fractional allocations as well). As illustrated in Aboolian et al. (2012), as well as in Berman and Kaplan (1987), computation of the equilibrium demand is relatively simple in this case, based on the fixed-point iteration approach.

An interesting feature of the DR model is that it is self-regulating: as waiting times become longer at the facilities, customer demand is automatically reduced. Thus, the system stability is assured by (17.44) without the need for explicit

constraints (17.9). Moreover, even though customer assignments are "dictated" by the decision-maker through the specification of x_{ij}, assigning a customer to a more distant or more congested facility leads to a lower demand λ_j, with the resulting loss of revenue. Thus, the model assures that the objectives of the decision-maker and customers are aligned, while avoiding the complexities of full traffic equilibrium treatment (another way to interpret the DR model is that the hard constraint requiring each customer to be assigned to their utility-maximizing facility is replaced with a soft constraint, allowing violations of such assignments at a cost). In fact, Aboolian et al. (2012) report (based on computational experiments) that at optimum all customers are almost always assigned to their utility-maximizing facility, though rare exceptions do occur.

The behavior of DR model involves an interesting feedback loop: as the service offered by the facilities is improved (by locating the facilities closer to customer nodes, or allocating more capacities to the facilities), the customers respond by generating more demand (positive feedback), which leads to increased congestion at the facilities, leading to reduced demand (negative feedback). Thus one could legitimately ask whether models with elastic demand may lead to counter-intuitive results where service improvements result in a net loss of demand. Fortunately, this is not the case as shown in the following result from Berman et al. (2014):

Theorem 17.4 *For $j \in J$, let $\lambda_j(d_j, w_j)$ be the equilibrium demand rate when the travel time is d_j and the expected waiting time is w_j. Then λ_j is non-increasing in d_j and w_j (strictly decreasing when the utility function is strictly decreasing in the corresponding parameter).*

Thus, with a reasonably behaved utility function, when the service offered to customers at $j \in J$ is improved in terms of either travel distance or waiting time, or both, the demand rate increases, leading to higher revenue for the decision-maker (for this customer node). Since nodes that are currently not served (i.e., with $\sum_i x_{ij} = 0$) can be treated as having the travel distance that is so high that the demand rate is negligibly close to 0, the decision to serve these nodes by assigning them to *any* open facility can be treated as reducing the travel distance. This leads to the following result:

Corollary 17.1 *In the elastic demand case, there exists an optimal solution to SLCIS where every demand node is served.*

17.3.2.3 Customer Reaction Models with Deterministic Utility 3: Full Response Models (FR)

In this model class, the customer response to facility location and capacity allocation decisions includes both the level and the allocation of demand. Thus, the equilibrium values of x_{ij} and λ_j are described by a system that includes flow equilibrium conditions (17.40)–(17.42), as well as the elastic demand equilibrium (17.44). The

existence and uniqueness of equilibria are assured by the following corollary:

Corollary 17.2 *The equilibrium existence and uniqueness results of Theorems 17.1 and 17.2 extend to the FR model class.*

The reader can refer to Brandeau et al. (1995) for further details; note that the uniqueness result has the same limitations as for the AR models (i.e., uniqueness can only be guaranteed with respect to the values of the objective, provided the objective function is suitably defined). Also, just as in AR models, this corollary requires fractional allocation vectors x_{ij}.

The computation of equilibrium solutions presents even more challenges than for AR models. One approach to deal with this complexity is by using the DR model as an approximation—as noted above, computational experiments suggest that optimal solutions to DR and AR models often coincide. Another approach, which is becoming more popular, is to use an alternative specification of demand allocation vectors described in the following section.

17.3.2.4 Proportional Allocation (PA) Models

As discussed above, the PA modelling framework is based on the assumption that customers allocate their demand among many (possibly all) facilities in proportion to the utility derived from these facilities. Essentially, each customer node $j \in J$ is viewed as a "market" with facilities competing for shares of this market.

The simplified structure, where customer demand allocations appear in closed form and can be analyzed for additional insights, together with several attractive mathematical properties have attracted significant recent interest to this model class, with several new approaches appearing since the first edition of this book.

These models have their theoretical origins in the MCI model of Cooper and Nakanishi (1988). As discussed below, they are also closely linked to stochastic utility theory. In the competitive location literature these models have appeared under many names, including "competitive interaction models", "Huff-type models", "gravity models", "multinomial logit models", "market-share models". While there are minor specification differences between these, the basic structure remains the same; we refer the reader to Chap. 14, as well as the review by Berman et al. (2009a).

Since SLCIS models of AR and FR type can be regarded as bi-level games played between the decision-maker and the customers, proportional allocation mechanism can be applied to the SLCIS context as well. This mechanism specifies the solution to the non-cooperative game played between customers once the decision-maker's strategy is specified as follows: for customers at $j \in J$ and (open) facility at $i \in I$, the demand allocations are given by

$$x_{ij} = \frac{U(d(i,j), \overline{W}_i) y_i}{\sum_{k \in I} U(d(k,j), \overline{W}_k) y_k}, \tag{17.46}$$

where the numerator represents the utility derived from facility i by customers at j, and the denominator is the total utility derived by customers at j from all open facilities. Note that if there are any pre-existing competitive facilities that may attract customer demand, they should be included as additional term $\sum_{k\in C} U(d(k,j),\overline{W}_k)$ in the denominator, where C is the set of competitive facilities. To simplify the exposition, we will assume no competitive facilities in the remainder of the current section.

Note that with the specification (17.46), it is easy to see that $\sum_{k\in I} x_{kj} = 1$ for all customers j, implying that all of customer's visits will be captured by the open facilities. In case where none of the open facilities provide adequate service (e.g., all are too far away to be considered), this may be unrealistic. A common modification is the inclusion of "outside option", i.e., the option for the consumer not to use the service offered by the facilities at all. Suppose the utility of this option for customers at j is given by U_{0j}. Then by adding this term to the denominator of the expression above we obtain

$$x_{ij} = \frac{U(d(i,j),\overline{W}_i)y_i}{U_{0j} + \sum_{k\in I} U(d(k,j),\overline{W}_k)y_k}, \tag{17.47}$$

where the outside option is modeled as a pre-existing competitive facility providing utility constant U_{0j}. Observe that in this case $\sum_{k\in I} x_{ij} < 1$.

In both cases, the demand allocations are fractional, and the demand rate from j attracted by facility i is given by $\lambda_j x_{ij}$. For deterministic utility models we drew a distinction between FR and AR models depending on whether λ_j is elastic or not. A similar distinction can, in principle, be drawn for PA models, with $\lambda_j = \lambda_j^{max}$ for AR models and λ_j being elastic with respect total utility derived by customer j from all facilities: $U_j = \sum_i U(d(i,j),\overline{W}_i)x_{ij}$. While PA-FR models of this type have been considered in deterministic location literature (see, e.g., Aboolian et al. 2007, 2012), we are not aware of any SLCIS models of this type. Thus, all current PA models follow the AR assumption that available customer demand at each node is equal to λ_j^{max}.

Note, however, that when specification (17.47) is used, the resulting model automatically retains some aspects of elastic demand. This because the total captured demand from customers j is given by $\lambda_j^{max}(1 - \frac{1}{U_{0j}+U_j(I)})$, where $U_j(I) = \sum_{k\in I} U(d(k,j),\overline{W}_k)y_k$ is the total utility derived by customers at j from the service offered by all open facilities. Thus, as the value of offered service declines, the amount of captured demand declines as well—exhibiting similar behavior as when the demand is specified explicitly. The fact that this elasticity of demand is represented by a single model parameter U_{0j} makes the model (as well as the parameter estimation) simpler, accounting for the popularity of this representation. On the other hand, it should be obvious that explicit demand specification via (17.45) provides much more modeling flexibility.

To complete the specification of the proportional allocation model one needs to select a particular utility function. The popular Multinomial Logit (MNL)

specification (McFadden 1974) employs exponential utilities, leading to

$$x_{ij} = \frac{\exp(-\tau_d d(i,j) - \tau_w \overline{W}_i) y_i}{U_{0j} + \sum_{k \in I} \exp(-\tau_d d(k,j) - \tau_w \overline{W}_k) y_k}, \tag{17.48}$$

where weights τ_d, τ_w, as well as the outside option parameter U_{0j} can be estimated from the available consumer demand allocation data using the MNL methodology.

Two interesting observations can be made with respect to the MNL model. First, it can be derived axiomatically from the stochastic utility theory. The following discussion is based on McFadden (2005)—please refer there for further details. If one assumes that customer utility is given by

$$U^s_{ij} = U^L(d(i,j), \overline{W}_i) + \epsilon_{ij},$$

where $U^L(d,w)$ is the linear utility function given by (17.30) and ϵ_{ij} is a Gumbel random variable, then under further assumption that Independence of Irrelevant Alternatives holds, Eq. (17.48) can be shown to be a unique equilibrium demand allocation vector. This important result, due to McFadden (1974), provides a link between stochastic utility and proportional allocation models. Indeed, the (MNL) model is extremely popular in econometrics and marketing literature, being the dominant model in brand choice and related fields. On the other hand, Independence of Irrelevant Alternatives assumption is routinely observed to be broken, leading to many generalizations of stochastic utility models; see McFadden (2005) for further discussion.

The second observation for the (MNL) model is that, under very mild conditions, the user equilibrium conditions (17.33) can be regarded as the limiting case of the (MNL) model above. Assume that the weights τ_d, τ_w are scaled by same parameter θ. It is shown in Fisk (1980) that the (MNL) allocation (17.48) approaches the user equilibrium solution (17.39) as $\theta \to \infty$. This result holds as long as the waiting times at the facility are continuous and non-decreasing in the total demand seen by the facility. Thus, the (MNL) model can be viewed as a proper generalization of the user equilibrium model with exponential utilities. This, together with its attractive analytical properties described below, accounts for the popularity of this model in some of the recent SLCIS papers.

The key advantage of the proportional allocation approach is that the values of x_{ij} are directly computable from (17.46) or (17.48) without having to solve the cumbersome flow equilibrium equations. Nevertheless, it is important to recognize that an equilibrium condition is implicit in the definition above, even in the case of models with inelastic demand: the expressions for x_{ij} above are functions of waiting times $\overline{W}_i$, which, in turn, are functions of x_{ij}. Thus, (17.46) together with waiting time specification (17.17) and facility-level demand specification (17.8) form a system of non-linear equations. A solution to this system represents the equilibrium demand allocations and waiting times. The issues of existence and uniqueness of the equilibrium were examined in some detail by Lee and Cohen (1985). The existence

follows directly from standard fixed-point results and the continuity of x_{ij} in (17.46) and is based on Theorem 1 in Lee and Cohen (1985):

Theorem 17.5 *There exists an equilibrium solution* $(x_{ij}, \overline{W}_i, \lambda_j), i \in I, j \in J$ *to the proportional allocation model.*

Lee and Cohen (1985) also examine uniqueness and stability of equilibria, where stability refers to whether a system where customers start with some arbitrary demand allocations, evaluate their utilities and then re-allocate according to (17.46) will naturally reach an equilibrium. They derive sufficient conditions for both uniqueness and stability.

Theorem 17.6 *For proportional allocation models the equilibrium is unique and stable*

Some of the key results stated above also extend to PA models of FR type (i.e., elastic demand), though sometimes certain additional conditions are required. However, as noted earlier, no SLCIS models of this type have been described in the literature (though AR models with outside option partially fill this gap).

17.4 General SLCIS Model Specification

In this section we summarize the discussion in the preceding sections. Putting all the modeling components together allows us to provide the following formulation for the General SLCIS with M/G/1 queues at facilities:

maximize $Z =$

$$r \sum_{j \in J} \lambda_j \sum_{i \in I} x_{ij} \tag{17.49}$$

$$- \sum_{j \in J} \sum_{i \in I} \beta d(i, j) \lambda_j x_{ij} \tag{17.50}$$

$$- \sum_{j \in J} \sum_{i \in I} WC(\overline{W}_i) x_{ij} \tag{17.51}$$

$$- \sum_{i \in I} FC_i y_i - \sum_{i \in I} VC(\mu_i) \tag{17.52}$$

$$\overline{W}_i = \frac{(1+\gamma^2)\Lambda_i}{2\mu_i(\mu_i - \Lambda_i)} + \frac{y_i}{\mu_i + M(1-y_i)}, \qquad i \in I \tag{17.53}$$

$$[\ \lambda_j \text{ specification for DR and FR models }] \tag{17.54}$$

$$[\ x_{ij} \text{ specification for AR, FR, and PA models }] \tag{17.55}$$

$$[\text{ Coverage Constraints }] \tag{17.56}$$

$$[\text{ SC Constraints }] \tag{17.57}$$

$$\sum_{i \in I} y_i \leq m \tag{17.58}$$

$$\Lambda_i = \sum_{j \in J} \lambda_j x_{ij}, \qquad i \in I \tag{17.59}$$

$$\sum_{i \in I} x_{ij} \leq 1, \qquad j \in J \tag{17.60}$$

$$x_{ij} \leq y_i, \qquad i \in I, j \in J \tag{17.61}$$

$$\mu_i \geq \Lambda_i \qquad i \in I, j \in J \tag{17.62}$$

$$x_{ij} \geq 0; \ \mu_i \geq 0; \ y_i \in \{0, 1\}. \tag{17.63}$$

The objective function (17.49)–(17.52) represents the total profit which includes the revenue, travel, congestion, and facility fixed and capacity costs, respectively. Constraints (17.53) define the expected waiting time for M/G/1 queues. These can be substituted with constraints defining other relevant congestion measures, different queueing mechanisms or both. Specifications (17.54) are only relevant for elastic demand models of type DR and FR type; when the demand rate is assumed to be inelastic, one should omit these and set $\lambda_j = \lambda_j^{\max}$. Similarly, specifications (17.55) are only relevant for user-choice models of AR and FR type. Constraints (17.58)–(17.62) enforce the basic interconnections between the decisions variables and are typically present in some form in all models.

To the best of our knowledge, no published work contains all components listed in the general formulation above. The specific SLCIS models considered in the literature typically include only some of the terms in the objective function, differ in terms of the queueing assumptions and performance measures, as well as in which (if any) of the specifications (17.54)–(17.57) to include. The models also differ in terms of the decision variables. While variables y_i and x_{ij} are present in all models we are familiar with (though x_{ij} may be restricted to binary values only), most models will assume that the number of facilities is m and not a decision variable. Many models also assume that all facilities have identical capacity μ, thus dropping the decision variables μ_i as well.

It is clear that the variety of SLCIS models one can define by mixing and matching different parts of the general formulation above is almost unlimited. In the next section we try to bring some structure to the models considered in the literature

by grouping them around some common themes and describing the key challenges and solution techniques that have been developed for them.

17.5 SLCIS Models in the Literature: Overview and Classification

Our primary focus (with a few exceptions) is on relatively recent SLCIS models that have appeared since the survey of Boffey et al. (2006).

As noted earlier, the published SLCIS models constitute a rather bewildering pattern of different assumptions, constraints and response mechanisms. However, several common themes do emerge, allowing us to identify five common types of models: Coverage-Type (CT), Service-Objective (SO), Balanced-objective (BO), Explicit Customer Response (ECR), and Proportional Allocation (PA) models. These are described in more detail in the following sections. The relevant references are summarized on Tables 17.2–17.6. These tables have the following format: the first column identifies the reference by the list of authors/year of publication; the next two columns identify the Model Class by customer response type, as well as by the utility function used, if applicable. The following three columns indicate the main underlying system assumptions: the nature of the queuing system, and whether the number of facilities and the number of servers are flexible or not. The next two columns identify the presence of coverage and service level constrains. The following five columns indicate the presence of the corresponding terms in the objective function. The last two columns briefly describe the solution approach and any additional comments.

17.5.1 Coverage-Type (CT) Models

These models, listed on Table 17.2, aim to design the system that provides *adequate* service to customers, where adequacy is usually defined through travel distance and congestion delays, which are controlled through coverage and service level constraints, respectively. The defining feature of this model class is the presence of general coverage constraints (17.56), for instance constraints (17.13). The CT models include Baron et al. (2008), Berman et al. (2006), Kakhki and Moghadas (2010), Marianov and Serra (1998). These models were among the very first SLCIS models to be considered, dating back to Marianov and Serra (1998), and stem directly from similar models for systems with mobile servers (see Berman and Krass (2002) for an extensive discussion).

CT models usually assume that it may not be possible to provide adequate service to all customers and thus demand losses may occur. The objective is typically to maximize the "captured" demand, i.e., the total demand of customers

Table 17.2 Coverage-type (CT) models

Reference	Cust. resp.	Utility function	Queueing model	Flexible # Facil.?	Flexible proc. rate μ?	Coverage constr./ Lost demand	Service constr.	Obj: revenue (captured demand)	Obj: travel time	Obj: congest. cost	Obj: facility fixed cost	Obj: server variable cost	Solution approach	Additional comments
Baron et al. (2008)	AR	Distance	GI/G/N, GI/G/1	Yes	Yes	Yes	Yes				Yes	Yes, General concave	Decompose the problem into several simpler sub-problems. Developed heuristic based on the equitable facility configurations	Both single and multiple server models considered
Berman et al. (2006)	AR	Distance	M/M/1/c	Yes	No	Yes	Yes: % blocked calls				Yes, Min Total # of facilities		Variety of heuristics including tabu search and random adaptive search	Demand is lost due to coverage and congestion constraints
Kakhki and Moghadas (2010)	NR	N/A	M/G/1	No	No	Yes	Yes	Yes, Max covered demand					Exact: Obtain semi-definite relaxation that will provide an UB	No testing or comp. results

(continued)

Table 17.2 (continued)

Reference	Cust. resp.	Utility function	Queueing model	Flexible # Facil.?	Flexible proc. rate μ?	Coverage constr./ Lost demand	Service constr.	Obj: revenue (captured demand)	Obj: travel time	Obj: congest. cost	Obj: facility fixed cost	Obj: server variable cost	Solution approach	Additional comments
Marianov and Serra (1998)	NR	N/A	M/M/1, M/M/K	No	No	Yes	Yes	Yes, Max covered demand					Exact: Linearized the SLC, leading to a linear MIP	
Yang (2018)	AR and NR	Based on fac. service capacity, not waits	M/M/k	Yes	Flexible number of servers k	Yes	Yes: on prob. of long waits	Yes, Max covered demand			Budget constraint on total fixed and variable cost		SLC approximated by an upper bound on demand seen by each facility, where the latter is approximated separately	Both NR and AR models considered

Table 17.3 Service-objective (SO) models

Reference	Cust. resp.	Utility function	Queueing model	Flexible # Facil.?	Flexible proc. rate μ?	Coverage constr./ Lost demand	Service constr.	Obj: revenue (captured demand)	Obj: travel time	Obj: congest. cost	Obj: facility fixed cost	Obj: server variable cost	Solution approach	Additional comments
Aboolian et al. (2009)	AR	Distance	M/M/k	No	Yes. Total # of servers bounded	No	No		Minmax	Minmax			Meta-heuristics (genetic, tabu)	
Berman and Drezner (2007)	AR	Distance	M/M/k	No	Yes. Total # of servers bounded	No	No		Yes	Yes			Descent, simulated annealing, tabu search and genetic heuristics	
Boffey et al. (2010)	NR	N/A	M/Er/1/c	No	No	No	Yes: # blocked		Yes				Turns into Capacitated p-median in M/M/1/N case, solved as MIP (this is for r = 1). For general Er, do a greedy-type heuristic	
Drezner and Drezner (2011)	AR/Prop. Alloc.	Distance, exp	M/M/k	No	No	No	No		Yes	Yes			Heuristic (descent, tabu search, simulated annealing, genetic)	

(continued)

Table 17.3 (continued)

Reference	Cust. resp.	Utility function	Queueing model	Flexible # Facil.?	Flexible proc. rate μ?	Coverage constr./ Lost demand	Service constr.	Obj: revenue (captured demand)	Obj: travel time	Obj: congest. cost	Obj: facility fixed cost	Obj: server variable cost	Solution approach	Additional comments
Hamaguchi and Nakade (2010)	AR	Distance	M/G/1	No	No	No	No		Ignored	Max prob $W < \tau$			Heuristic (greedy + tabu), service times computed exactly via Laplace transform	Maximize probability that waiting time is below t
Marianov and Serra (2011)	NR	N/A	M/M/c/K	No	No	No	Demand loss due to block-ages		Yes	Yes			Ant colony heuristic	Bi-objective; (1) Travel cost, (2) Congestion cost (with a coefficient for the number of customers in the system)
Marianov et al. (2009)	NR	N/A	M/Er/K/c	No	No	No	Yes: # blocked		Yes				Similar to Boffey et al. (2010) with SLC estimated via Erlang queues	
Wang et al. (2002)	AR	Distance	M/M/1	No	No (fixed mu)	No	Yes: max uti-lization rate bounded		Yes	Yes			Greedy, tabu search and Lagrangian relaxation heuristics	

Table 17.4 Balanced-objective (BO) models

Reference	Cust. resp.	Utility function	Queueing model	Flexible # Facil.?	Flexible proc. rate μ?	Coverage constr./ Lost demand	Service constr.	Obj: revenue (captured demand)	Obj: travel time	Obj: congest. cost	Obj: facility fixed cost	Obj: server variable cost	Solution approach	Additional comments
Hoisenpour and Ahmadi-Javid (2016)	NR	N/A	M/M/1 with interrupt.	Yes	Yes	No	No	Yes (fixed price)	Yes	Yes	Yes	Yes	Exact algorithm based on Lagrangian relaxation	Random service interruptions
Aboolian et al. (2008)	AR	Distance	M/M/k	Yes	Yes	No	No		Yes	Yes	Yes	Yes	Exact algorithm and heuristics	
Aboolian et al. (2018)	NR	N/A	M/M/1	Yes	Yes	No	Prob. of late delivery (Mod 1), penalty on late delivery per incident (Mod 2), or per unit time (Mod 3)		Yes	Yes	Yes	Yes	Semi-exact algorithm using tangent line approximation method	Capacity discretized and used to linearize the model
Abouee-Mehrizi et al. (2011)	AR/ Prop. Alloc.	Exp	M/M/1/ balking	No	Yes	No		Yes: demand loss due to balking			Yes	Yes	Tabu search procedure to determine the location of the facilities, exact algorithm to obtain the optimal service rate at each facility, and a heuristic algorithm to obtain the price	Max total profit with limited room capacity for waiting

(continued)

Table 17.4 (continued)

Reference	Cust. resp.	Utility function	Queueing model	Flexible # Facil.?	Flexible proc. rate μ?	Coverage constr./ Lost demand	Service constr.	Obj: revenue (captured demand)	Obj: travel time	Obj: congest. cost	Obj: facility fixed cost	Obj: server variable cost	Solution approach	Additional comments
Ahmadi-Javid and Hoisen-pour (2018)	NR	N/A	M/G/1	Yes	Yes	No	No		Yes	Yes	Yes	Yes	Exact algorithm based on Conic programming (standard CPLEX cuts)	Capacity discretized. Coefficient of variation part of capacity decision
Ahmadi-Javid et al. (2018)	NR	N/A	M/G/1	Yes	Yes	No	No		Yes	Yes	Yes	Yes	Exact algorithm based on Conic programming (both standard CPLEX cuts and additional valid cuts used)	Coeff of variation a function of service rate
Azizi et al. (2017)	NR	N/A	M/G/1	Yes	Yes	No	No		Yes	Yes	Yes	Yes	Hub and Spoke network. Exact solution via MIP	Capacity discretized. Coefficient of variation part of capacity decision
Castillo et al. (2009)	NR	N/A	MM1, MMk	Yes	Yes	No	No		Yes	Yes	Yes	Yes	Exact: eliminated capacity variables, obtaining concave objective, then used Lagrangian relaxation	
Elhedhli (2006)	NR	N/A	M/M/1	Yes	Yes	No	No		Yes	Yes	Yes		Exact: linearization approach which eliminates capacity variables and replaces non-linear term in the	

													objective with a family of linear constraints; used column generation	
Kim (2013)	NR	N/A	G/G/1	No	No	No	Yes: total wait		Yes	Yes	Yes		Exact: uses clearing function f(mu,W), i.e. throughput at a facility with wait given by W; this allows for linearization of constraint, but f(,) is non-linear; used column generation	
Marianov and Rios (2000)	NR	N/A	M/M/1	Yes	No	No	Yes: prob queue below a threshold		Link constr. cost		Yes		Exact: linearized the SLC, then solved MIP	Application to the location of ATM switches
Pasandideh and Chambaria (2010)	NR	N/A	M/M/1/c	Yes (total location cost is bounded)	μ is fixed but buffer size is a decision variable	No	No		Yes (Obj 1)	Yes (Obj 1)		Obj 2: Min Ave % idle time per facility	Genetic heuristic	Bi-objective: (1) total waiting time, (2) total % idle at the facilities
Vidyarthi and Jayaswal (2014)	NR	N/A	M/G/1	Yes	Yes	No	No		Yes	Yes	Yes		Exact: linearized "nasty" term in obj function, leading to a convex problem with exp number of constraints. Similar to Elhedhli (2006)	
Wang et al. (2004)	AR	Distance	M/M/1	Yes	Yes	No	Yes: max utilization rate bounded		Yes	Yes	Yes	Yes	Greedy-type heuristics; shown to be optimal for some of the models considered	Present several different models, but most general one is of "social optimum" type

Table 17.5 Explicit customer response (ECR) models

Reference	Cust. resp.	Utility function	Queueing model	Flexible # Facil.?	Flexible proc. rate μ?	Coverage constr./ Lost demand	Service constr.	Obj: revenue (captured demand)	Obj: travel time	Obj: congest. cost	Obj: facility fixed cost	Obj: server variable cost	Solution approach	Additional comments
Aboolian et al. (2012)	DR	Exp. Demand/ lin. Utility	M/M/1, M/M/k	Yes	Yes	No	Yes: wait	Yes				Yes	Exact algorithm and heuristics	Explicit response. Max profit including a feedback loop between customer demand and congestion. Exogenous price
Aboolian et al. (2016)	FR	Exp. Demand/lin. Utility	M/M/1	Yes	Yes	No	No	Yes			Constraint specifying upper and lower capacity bound for each open facility		Exact algorithm and heuristics	Explicit response. Max profit (exogenous price). Linearization of traffic equil.
Berman and Drezner (2006)	DR	Distance	M/M/1	No	No	No	Yes: exp. wait times	Yes					Single facility: exact O(n3); Multi-facility: NLIP, heuristic algorithms (ascent algorithm, tabu, simulated annealing)	Explicit response
Berman and Kaplan (1987)	DR	Linear	M/M/1	No: m = 1	No	No	No	Yes					Exact algorithm (1-facility)	Explicit response. Single facility setting. Exog. price

Baron et al. (2007)	AR/ search	Distance	M/M/1/c	No	No	No	Demand loss to blockage, search costs	Demand lost to block-ages, search					Heuristics combined with an iterative calibration scheme to estimate the expected demand rate faced by the facilities	Explicit response. Comes closer to mobile server models due to dynamic search behavior by customers
Berman et al. (2014)	DR	Multipli-cative	G/G/1 and M/M/1	No: m = 1 or fixed	Yes	Yes	No	Yes			Yes	Yes	Exact algorithms (1-facility)/heuristic for m-facility	Explicit response. Demand elastic in travel, wait, and endogenous price
Tavakkoli-Moghaddam et al. (2009)	FR	Distance and price; linear	M/M/k/K	Yes	Yes (flex. Number of servers)	No	No (Service optimized by separate objectives)	Yes for one of the objec-tives		Indirectly: one objective is waiting time		Indirectly: one objective is ide time	Metaheuristics based on vibration damping optimization	Multiple objectives. Blockages not accounted for in customer utility of waiting times. Endogenous prices
Tong (2011)	FR	Multipli-cative	G/G/1	Yes	Yes	No	No	Yes			Yes	Yes	EXACT for 1-facility, exact and heuristic for m-facility case	Explicit response. Considers several models, including FR models with traffic equilibrium conditions. Endog. price
Zhang et al. (2009)	FR	Linear	M/M/1	Yes: min workload per facility	No	No	No	Yes					Location allocation heuristic (that includes an equilibrium facility and client allocation sets)	Explicit response. Max total participation (captured demand)
Zhang et al. (2010)	FR	Linear	M/M/k	Yes: min workload per facility	Yes	No	No	Yes					Bi-Level optimization model; lower level solved by variational inequalities and upper level heuristically	Explicit response. Max total participation (captured demand)

Table 17.6 Proportional allocation (PA) models

Reference	Cust. resp.	Utility function	Queueing model	Flexible # Facil.?	Flexible proc. rate μ?	Coverage constr./ Lost demand	Service constr.	Obj: revenue (captured demand)	Obj: travel time	Obj: congest. cost	Obj: facility fixed cost	Obj: server variable cost	Solution approach	Additional comments
Dan and Marcotte (2017)	AR	Linear w/ blockage penalty/ MNL	M/M/1/c	Yes	Yes	No	No	Demand lost due to blockage			Budget constraint on total fixed and variable cost		Bi-level optimization model. Semi-exact algorithm based on piecewise linear approximation of lower level. Heuristic	Corrects model of Marianov et al. (2008)
Marianov et al. (2008)	AR	Linear/ MNL	M/M/1/c	No	No	No	No	Demand lost due to blockage					Greedy-based Heuristics	Max captured demand (at own facilities); blockages do not affect customer utilities
Rabieyan and Seif-barghy (2010)	AR	Distance	M/M/1/ balking	No	No	Constraint on idle rate at facilities	No	Demand loss due to balking					Genetic, simulated annealing, and k-opt-type heuristic. The latter outperforms by up to 43% for larger instances	Max total benefit that incorporates travel distance and accounts for balking
Schön and Saini (2018)	AR	Linear	M/G/1	Yes	No	No	No	Yes	Yes		Yes	Yes	Exact MIP + Heuristic. Both capacity and promised service level are discretized. Various forms of PA modeled (incl. MNL)	Exogenous price
Zhang et al. (2012)	Model 1: PA/AR, Model 2: ECR/FR	Distance	M/M/k	Yes: min workload per facility	Yes	No	Yes: wait. Lost demand in Model 2	Lost demand in model 2					Probabilistic search algorithm and a genetic algorithm	Model 1: PA (MNL based on shortest time), model 2: ECR (shortest time)

who get adequate service. The travel and congestion costs are not included in the objective as these are controlled through the corresponding constraints. Earlier models were of type NR (directed choice); later models tended to be of type AR, but customer allocations were assumed to be only a function of travel distance, i.e., the underlying utility is given by (17.31), avoiding all complications related to equilibrium behaviors. It is interesting to note that even though demand is assumed to be inelastic, the assumption of demand losses can be viewed as (a rather crude) form of demand elasticity—corresponding to an implicit stepwise utility function, with customers using service only if coverage and service level constraints are met.

The typical formulation maximizes the objective consisting of (17.49) with revenue $r = 1$, reflecting the maximization of captured demand, subject to constraints (17.56)–(17.61). For models of type AR, one also adds constraints specifying the allocations. These enforce each customer to travel to the closest available facility. These constraints can be specified in various forms; see Berman et al. (2006) for a discussion.

It can be seen that this leads to a formulation which is a linear mixed-integer program (MIP), except for the service level constraints. However, as discussed in Sect. 17.2.3.2, under some conditions, the latter can be linearized. Recall that a general service level constraint can be recast as either (17.23), requiring adequate service capacity at each facility, or (17.24), placing an upper limit on the allowed arrival rate at each facility. When the capacities μ_i are decision variables, these reformulations remain non-linear. However, if one makes a simplifying assumption that all facilities have identical service rate μ (for multi-server facilities, this implies assuming identical number of servers at all facilities), non-linearities disappear. This is a common assumption in CT (and some other SLCIS) models: Berman et al. (2006), Kakhki and Moghadas (2010), Marianov and Serra (1998) assume identical and pre-specified service rates at the facilities. Under this assumption, (17.24) takes the form

$$\Lambda_i \leq \bar{\Lambda},$$

where the right-hand side is a constant which depends on the desired service level and is computable in advance. This shows the equivalence of a CT model with fixed service rates to the capacitated location problems. Such connection is discussed at length in Boffey et al. (2006).

The resulting linear MIP may, in principle, be solved exactly using off-the-shelf software, such as CPLEX. However, as pointed out in Berman et al. (2006), the formulation resulting from the addition of linearized service level constraints and the "closest assignment" constraints tends to be large and not very tight, causing computational difficulties for even moderately-sized instances. This has led Berman et al. (2006) and other authors to develop heuristic approaches.

We note an important result from Baron et al. (2008), who studied a very general version of the CT model, where both the number and the capacities of facilities are decision variables and the facility-related costs are quite general (in their version, all customer demand must be served and the objective is to minimize fixed and

variable location costs). They show that, under quite general conditions, the optimal facility configuration is one that ensures that each facility sees (approximately) the same demand, i.e., ideally, $\Lambda_i = \Lambda_k$ should hold for all open facilities $i, k \in I$ (identical demand may not be possible to achieve when customer demand originates from discrete nodes and single-sourcing assumption is made). Once the facility locations are decided, the optimal capacities μ_i can be computed through a separate optimization model.

This result provides an important insight for CT models: when the goal is to ensure "satisfactory" service experience, the optimal design should equalize loads on the facilities. This leads to an "Equitable Location Problem"—a deterministic problem where one seeks to locate a set of facilities so that the attracted demand is distributed as evenly as possible. Such problem was addressed in Baron et al. (2007), Berman et al. (2009b), and Suzuki and Drezner (2009).

While traditional applications of CT models (with or without congestion) is in emergency services, an interesting new theme is the location of recharging stations for electrical vehicles. Due to limited battery range, coverage constraints are crucial. On the other hand, user choice behavior must be taken into account as well. An AR-type SLCIS model with these features is developed in Yang (2018), where each station is modeled as an $M/M/K$ queue, with the number of stations and the number of servers at each station being decision variables. A service constraint limiting the probability of long waits is assumed. Users select facilities based on travel distance and capacities, not waits, which eschews the issues related to traffic equilibria (but the assumption does seem questionable). Due to non-equal capacities at the facilities and non-linearities inherent in the $M/M/K$ system, a heuristic approximation is developed to linearize the SC constraints.

17.5.2 Service-Objective (SO) Models

These models, listed on Table 17.3, seek to design a system that optimizes "customer service" using limited resources. Here "limited resources" means that the number of facilities to be located and the total available service capacity are specified through constraints, rather than through the objective function term (17.52). "Customer service" is typically defined as the combination of travel and congestion costs; thus the objective function typically includes terms (17.50) and (17.51). Since the congestion cost term (17.51) only measures the aggregate congestion, some authors (e.g., Boffey et al. 2010; Marianov et al. 2009; Marianov and Serra 2011; Wang et al. 2002) impose service level constraints to ensure that congestion is controlled at each facility. SO models assume inelastic demand, so the revenue term is missing in the objective as all available customer demand is assumed to be "covered" (even though some models do allow for demand losses due to congestion, these losses are controlled through service level constraints). Thus, all customers must be assigned to facilities and constraint (17.60) is specified as equality.

The models of this class are either of NR or AR type with distance-based utility function (customers travel to the closest open facility). An interesting exception is the use of AR model with proportional allocation and exponential utility (17.32) by Drezner and Drezner (2011) (though they do not comment on the existence and uniqueness of the equilibrium solution, it is in fact assured by the results cited earlier).

While the constraint set for SO models is quite similar to that of CT models (in fact, it is somewhat simpler since the coverage constraints and, in some cases, service level constraints are missing), inclusion of the congestion term in the objective leads to a non-linear model for which finding exact solutions is problematic. This difficulty is further compounded when the queues at the facilities are of multi-server type and/or have non-Markovian service times: in these cases exact closed-form expressions for the congestion-related performance measures are either not available, or are quite complex, requiring a separate procedure to evaluate the congestion levels for a each set of values of the facility location and customer allocation decision variables. For this reason, the proposed solution methods are all heuristic-based, typically employing meta-heuristic approaches such as tabu search, simulated annealing, and genetic algorithms.

SO models become significantly more complicated when capacities of facilities are allowed to be flexible (i.e., when μ_i are not assumed to be identical at all facilities). Most of the published models assume identical capacities, with Aboolian et al. (2009) and Berman and Drezner (2007) being notable exceptions.

17.5.3 Balanced-Objective (BO) Models

These models seek to design a system that "balances" the costs incurred by the two main "players" in the system: customers, who bear the travel and congestion costs, and the decision-maker who bears facility-related costs. They are listed on Table 17.4.

One may view BO models as seeking to achieve a "social optimum"; the objective functions in these models are similar to social welfare functions in economics, with the resulting models being similar to the "first best" models. Since the objective incorporates customer concerns, the models are typically of NR type: customers accept the directed assignments to optimize "social welfare", even if this leads to assignments that are suboptimal from individual customers' point of view (two references that incorporate customer response are Aboolian et al. 2008 and Abouee-Mehrizi et al. 2011). The demand is assumed to be inelastic. The coverage and service level constraints are typically absent, as service adequacy is addressed by the objective; the one exception appears to be Aboolian et al. (2018) where service constraint is present in one of the three proposed models.

The objective function typically includes the "customer-borne" cost terms (17.50)–(17.51) representing travel and congestion costs, as well as the "operator-borne" facility costs (17.52). Since most models do not assume any demand losses,

the revenue term (17.49) is not included; the exception being Abouee-Mehrizi et al. (2011), who model revenue losses due to balking and thus optimize the net profit. Two of the models in Aboolian et al. (2018) include penalty terms for late deliveries (i.e., delayed service), where the penalty is charged per instance or per amount of delay.

Most models in this class assume relatively simple queuing systems at the facilities with the two recent exceptions being Hoisenpour and Ahmadi-Javid (2016) who study a system with random service interruptions, and Azizi et al. (2017) who assume $M/G/1$-based hub-and-spoke system.

Other distinguishing features of most BO models are typically simple constraint sets and the inclusion of flexible capacity at the facilities as the decision variables. The main solution difficulty stems from the non-linearities inherent in the congestion (third) term of the objective function (17.51). There are several approaches for either making these terms less complex or linearizing them, leading to interesting exact algorithms. We describe two such approaches below.

The first is based on Castillo et al. (2009). They assume an $M/M/1$ queuing system at the facilities and use the average number of customers in the system $L_i(\Lambda_i, \mu_i)$ as the performance measure at facility i. For $M/M/1$ queue, this can be written as

$$L_i(\Lambda_i, \mu_i) = \frac{\Lambda_i}{\mu_i - \Lambda_i}. \tag{17.64}$$

All costs are assumed to be linear and uniform (i.e., identical for all facilities), leading to the following objective function:

$$\text{minimize } Z = \beta \sum_{j\in J} \sum_{i\in I} d(i,j)\lambda_j x_{ij} + WC \sum_{i\in I} L_i(\Lambda_i, \mu_i) + FC \sum_{i\in I} y_i + VC \sum_{i\in I} \mu_i, \tag{17.65}$$

where WC, FC, VC are the waiting cost, fixed cost and variable cost parameters respectively. This function is minimized subject to constraints (17.58), (17.60) specified as equality, as well as (17.59), (17.61) and (17.62).

Note that for any specified values of x_{ij} and y_i, the optimal capacity μ_i^* can be determined separately for each facility. Indeed, it is not difficult to show that

$$\mu_i^* = \Lambda_i + \sqrt{\frac{WC}{VC}\Lambda_i}.$$

Observe the similarity of this expression to (17.25) discussed earlier. It also has the same interpretation: the optimal capacity at facility i consists of the minimal level Λ_i, necessary to ensure system stability, and "capacity cushion" which grows with the square root of Λ_i and whose size depends on the ratio of waiting and capacity costs. Substituting the last expression into (17.65) and performing some

algebraic manipulations and noting that for NR models the total customer demand is an exogenous parameter, allows us to re-state the objective function as

$$\text{mininize } Z = \beta \sum_{j\in J}\sum_{i\in I} d(i,\, j)\lambda_j x_{ij} + 2\sqrt{WC \cdot VC}\sum_{i\in I}\sqrt{\sum_{j\in J}\lambda_j x_{ij}} + FC\sum_{i\in I} y_i,$$

subject to constraints (17.58), (17.61), and (17.60) specified as equality; the variables Λ_i and μ_i are no longer required.

This is a MIP with a single concave term in the objective. Several methods are available to obtain exact solutions for models of this type, which also arise in location-inventory models, competitive location models and other contexts. One approach, based on Lagrangian Relaxation, is described in Shen (2005); a variant of this is used in Castillo et al. (2009). Another approach, based on tangent-line approximation (TLA) of the concave term, is presented in Aboolian et al. (2007). The TLA leads to an $\epsilon-$optimal solution, where the maximum relative error from the exact solution is bounded by ϵ, with the value of this parameter set by the user (the smaller the ϵ, the higher the computational effort required; $\epsilon = 10\%, 5\%, 1\%$ are typical choices). Recently, Hoisenpour and Ahmadi-Javid (2016) apply Lagrangian Relaxation to a model with random service interruptions at the facilities.

It should be noted that in view of the discussion preceding (17.25), a similar "trick" for replacing the congestion cost term with a concave form should work for more general queueing systems as well, at least as an approximation.

The second approach for obtaining exact solutions to BO models is based on capacity discretization ideas described earlier. The following discussion follows Elhedhli (2006). Once again we start with the model whose objective function is given by (17.65) and assume an $M/M/1$ queue at each facility. Assume the processing capacity must be equal to one of $H + 1$ discrete values, i.e., that $\mu_i \in \{0, \mu^1, \mu^2, \ldots, \mu^H\}$ for all $i \in I$, where $\mu^1 < \mu^2 < \ldots < \mu^H$.

Treating the expected queue length L_i as a decision variable, we rewrite (17.64) as

$$\Lambda_i = \frac{L_i}{1 + L_i}\sum_{h=1}^{H}\mu^h z_{ih},\ i \in I, \tag{17.66}$$

where z_{ih}, as defined in (17.3 and 17.4) is a binary decision variable taking the value of 1 if $\mu_i = \mu^h$ and 0 otherwise. Now consider the function $f(L) = \frac{L}{1+L}$. It is concave, and can thus be represented as the minimum of tangent lines, yielding a linear form. This can be used to represent the expression (17.66) as an infinite set of linear constraints (note that the objective is already linear, in terms of the new variable L_i). The resulting MIP can be solved through a column generation approach. The reader should refer to Elhedhli (2006) for details. A similar approach is applied to hub-and-spoke SLCIS system in Azizi et al. (2017).

The capacity discretization approach with the resulting MIP with concave objective naturally lands itself to the TLA methodology mentioned above. This approach is applied, with promising computational results, to a set of balanced objective models with explicit (per occurrence or per delay length) penalties on service delays in Aboolian et al. (2018).

An interesting recent development in MIP literature is the efficient treatment of conic functions (particularly conic constraints)—see Atamtürk and Vishnu Narayanan (2011) for a general treatment and Atamtürk et al. (2012) for an application to a location-inventory problem. Some standard solvers, e.g., CPLEX, now provide automatic treatment of conic inequalities. The resulting methodology has seen recent applications in the SLCIS literature as well. Ahmadi-Javid and Hoisenpour (2018) consider a BO model with $M/G/1$ queues at the facilities, where capacity is discretized and each choice leads to a certain μ_i, γ_i pair. The initial MIP with non-linear objective is re-formulated as a conic program with a linear objective and conic constraint to which CPLEX solver can be directly applied. A further development along this lines is presented in Ahmadi-Javid et al. (2018) where instead of using discretization, an affine relationship is assumed between the coefficient of variation γ_i and facility capacity μ_i. Once again an original non-linear MIP is recast as a conic program, but in addition to now-standard CPLEX treatment, a number of additional valid cuts are developed. The latter lead to a strong improvement in computational efficiency.

In summary, the simpler structure of BO SLCIS models allows for effective exact approaches to be developed. Another interesting observation is that the "location-allocation" and "capacity determination" sub-problems often separate. As noted earlier, these models, being of type NR, may assign individual customers to rather distant facilities. However, since the travel cost is in the objective function, these "undesirable" assignments can be controlled by increasing the corresponding cost coefficients. The computational results in Castillo et al. (2009) suggest that when travel costs are "reasonably" high, the overwhelming majority of customers (over 99% in the instances solved) are assigned to the closest open facility in the optimal solution.

17.5.4 Explicit Customer Response (ECR) Models

ECR models specify an "explicit" customer response mechanism, i.e., they are of types AR, DR, or FR. These models are listed on Table 17.5. The demand in these models is generally elastic, though in a few cases elasticity is specified implicitly through demand losses due to blockages. The objective always includes the revenue term (17.49), and may also include the facility cost terms (17.52), unless the number of facilities and servers is given.

While this class of models has received much recent attention, the earliest publications date back to the very beginning of the SLCIS modeling: see Berman

and Kaplan (1987). Some of the seminal early work is described in Brandeau et al. (1995).

Many of the technical issues related to ECR models have been covered in Sect. 17.3.2. The problem of determining the optimal location for a single facility (Berman and Drezner 2006; Berman and Kaplan 1987; Tong 2011; Berman et al. 2014) can be solved exactly. However, the treatment of the multi-facility case is generally quite difficult since, as noted earlier, in addition to the non-linear objective function the underlying models include the feedback loop between the customer demand and congestion and/or the equilibrium conditions for facility-client allocations, or both. Thus, heuristic approaches are almost always employed for multi-facility models. These heuristics are usually two-level: at the lower level they incorporate subroutines for computing the equilibrium solutions (using non-linear optimization techniques) for a given location set. At the upper level they try improvement strategies to determine a good set of open facilities, often using meta-heuristics. As in the case of BO models, the determination of the optimal capacity at a facility can often be done through a separate exact optimization procedure, for a given location and customer-allocation scheme.

We illustrate the foregoing discussion with the approach loosely based on Aboolian et al. (2012), who proposed one of the few exact approaches available for ECR models (in fact, the approach outlined below is an improvement on the original methodology). The model is of DR type, i.e., customers accept directed assignments to facilities, responding by reducing their demand when travel and congestion costs increase. Both $M/M/K$ and $M/M/1$ queueing systems can be considered; we will focus on the latter for simplicity. The primary queuing performance measure is the expected waiting time $\overline{W}_i$ at each facility i. While a general concave utility function may be used, we employ the exponential utility (17.32) for transparency, with the elastic demand given by (17.45). The fixed and variable costs are assumed to be uniform, i.e., identical for all locations.

We start by observing that if customers at node $j \in J$ are assigned to facility i, the maximum demand is given by

$$\lambda_{ij}^{\max} = \lambda_j^{\max} \exp(-\tau_d d(i, j)),$$

quantities that can be pre-computed. The resulting model can be formulated as follows:

$$\text{maximize } Z = r \sum_{i \in I} \Lambda_i - FC \sum_{i \in I} y_i - VC \sum_{i \in I} \mu_i \tag{17.67}$$

$$s.t. \quad \overline{W}_i = \frac{y_i}{\mu_i - \Lambda_i} \qquad i \in I \tag{17.68}$$

$$\Lambda_i = \sum_{j \in J} \lambda_{ij}^{\max} \exp(-\tau_w \overline{W}_i) x_{ij} \qquad i \in I \tag{17.69}$$

(17.60), (17.61).

This reflects the typical structure of DR models: explicit specification of the waiting time and demand, in addition to regular constraints for location models. Note that system stability constraints (17.62) are omitted, since the demand automatically adjusts to the offered capacities.

The next observation is that once customer allocation variables x_{ij} are specified, both the optimal capacities at the facilities and the actual realized customer demands are easy to determine. In fact, the latter only depend on x_{ij} through the total *maximal* demand allocated to each facility:

$$\Lambda_i^{\max} = \sum_{j \in J} \lambda_{ij}^{\max} x_{ij}. \tag{17.70}$$

For each facility i we now solve the following univariate "capacity optimization" model:

$$\begin{aligned} &\text{maximize } r\Lambda_i - VC\mu_i \\ s.t. \quad &\Lambda_i = \Lambda_i^{\max} \exp(-\tau_w \frac{\Lambda_i}{\mu_i - \Lambda_i}) \\ &\mu_i \geq 0. \end{aligned}$$

Aboolian et al. (2012) show that the solution to this model is unique and can be found through a simple univariate search. Note that the solution yields both, the optimal capacity μ_i and the corresponding demand level Λ_i. It is convenient to represent these quantities as functions of the allocated maximum demand: $\mu(\Lambda_i^{\max})$, $\Lambda(\Lambda_i^{\max})$. Substituting these quantities into the original model (17.67)–(17.69) we obtain

$$\begin{aligned} \text{maximize } Z = &r \sum_{i \in I} \Lambda(\Lambda_i^{\max}) - FC \sum_{i \in I} y_i - VC \sum_{i \in I} \mu(\Lambda_i^{\max}) \\ &(17.60), (17.61), (17.70), \end{aligned}$$

where the only non-linearities occur in the objective function. By solving the capacity optimization model repeatedly over a range of possible values of $\Lambda_i^{\max}$, we can construct a piecewise linear approximation of the functions $\Lambda(\Lambda_i^{\max})$ and $\mu(\Lambda_i^{\max})$ to any desired level of tolerance. Using these approximations in the model above yields a linear MIP which can be solved using standard off-the-shelf software.

As noted earlier, the separation of capacity optimization and customer allocation problems is a common feature of ECR models and has been used by a number of authors. However, an important driver of the exact approach outlined above is that the model in Aboolian et al. (2012) is of DR type, i.e., directed assignment and single-sourcing are both assumed. The computational results presented in Aboolian et al. (2012) suggest that neither of these assumptions is very restrictive (echoing the results in Castillo et al. (2009) discussed earlier). It was observed that in the

vast majority of instances solved, customers were, in fact, assigned to facilities that minimize their sum of waiting and travel times, i.e., the facilities they would have selected under an FR model. Also, by splitting the original customer nodes into k copies each containing $1/k$ of the original demand, and allowing each of these new nodes to be assigned to a different facility, the impact of the single-sourcing assumption was examined. Again, it turned out that for the instances solved, the violation of this assumption was rare (all copies of the original node were assigned to the same facility in the vast majority of the cases) and when split assignments occurred, they did not have a large impact on the objective function. Intuitively, both effects can be explained by the fact that in DR models the incentives of customers and the decision-maker, while not identical, are well-aligned: by forcing customers to use a less convenient facility, the realized demand (and the revenue) are reduced. Thus, when designing the system, a design that maximizes customer utilities is often optimal, even though such maximization is not explicitly enforced in the model.

A notable recent advance for ECR models was made in Aboolian et al. (2016). They assumed $M/M/1$ system with the fixed costs and budget constraint replaced by the requirement that any open facility must have the capacity of at least μ^{min} and at most μ^{max} (a reasonable assumption in case of public service facilities). As described earlier, using waiting times W_j in place of capacities μ_j as decision variables and adding additional binary variables z_{ij} to represent whether customer i makes any use of facility j, they derive an MIP with the only non-linearity limited to $1/W_j$ terms. Since this is convex in W_j, the TLA methodology can be used to obtain a linear MIP which is ϵ-optimal for the original problem. They were able to solve fairly large problem instances (up to 900 customer nodes and up to 40 potential locations) to within (at most) 0.1% of optimality. However, as noted earlier, the approach may be quite fragile with respect to the $M/M/1$ assumption.

17.5.5 Proportional Allocation (PA) Models

As discussed earlier, these models incorporate explicit customer response to the service offered by the decision-maker; however the form of this response (allocation of customer's demand amongst the facilities) is pre-specified via Eq. (17.47). In the first edition of this volume these models were classified under the ECR type. However, with several interesting recent developments, these models now merit a separate category; they are listed on Table 17.6.

There are well-established methods for linearizing the fractional market share equation (17.47) when customer decisions are decoupled. However, as observed in Sect. 17.3.2.4, when customer's utility includes waiting time (or another measure of congestion at the facilities), the decisions become coupled and (17.47) defines a system of non-linear equations that make the resulting SLCIS computationally very challenging.

The $M/M/1$ system offers significant simplifications since it is possible to treat the waiting time, rather than capacity, as the decision variable. Zhang et al. (2012)

uses this approach to linearize the customer-level problem in their Model 1, while optimizing the decision-maker's level via heuristics.

A more general approach (at the cost of discretizing some key decisions) is developed in Schön and Saini (2018). For an $M/G/1$ system they use capacity discretization (which also allows them to model coefficients of variation as part of decision variables). In addition, they discretize offered service levels, i.e., wait times, at the facilities. All non-linearities in the model, such as both the numerator and denominator in (17.47), can now be discretized, and thus linearized through the introduction of additional integer variables. The resulting model is quite general—it can incorporate a variety of utility functions, as well as revenue and cost terms in the objective—is formulated as a linear MIP. However, the formulation is very large, and thus even relatively small instances cannot be solved to optimality by CPLEX. This leads to the development of several heuristic approaches.

A different approach, heavily rooted in economics literature, is taken by Dan and Marcotte (2017). Their starting point is the model of Marianov et al. (2008), the first published SLCIS model with PA mechanism. The facilities are modeled as limited buffer $M/M/1/b$ queues where b is the buffer size; customers are blocked from entering the facility when the queue size reaches b. The objective is to locate m facilities to maximize total captured demand, where customers have an option to choose either new or pre-existing "competitive" facilities. The model employs linear utilities (17.30) and MNL structure (17.48). A metaheuristic procedure, combining GRASP and Tabu Search, is proposed.

Dan and Marcotte (2017) point out and correct several deficiencies in this model: (1) the "captured demand" does not account for demand lost to blockages, (2) customer's utility function does not account for dis-utility due to blockages, leading to a perverse situation where a customer who obtains service after experiencing some waiting time has a lower utility than a customer who traveled the same distance but was then blocked from joining the queue, (3) the capacity μ_j was assumed to be identical at all facilities and was treated as an exogenous parameter. In addition, the new model of Dan and Marcotte (2017) introduces a budget constraint:

$$\sum_i (FC y_i + VC \mu_i) \leq B,$$

where B is the available budget, and other notation is consistent with the general model in Sect. 17.4. Note that the capacity decision is treated as a continuous variable (though the buffer size b is treated as an exogenous parameter with an identical value for all facilities).

The problem is first formulated as a bilevel model, with the upper level (leader) specifying the facility locations and capacities, with the objective of maximizing captured demand (both the objective and the constraints are linear), while the lower level (follower) allocating customer demand according to MNL mechanism and constraints relating wait times and blockage probabilities. In this initial form, the lower level is a fixed point equation, rather than an optimization problem. However, using the standard results from Fisk (1980), the lower level is converted to an

non-linear optimization problem, whose objective is shown to be convex. Next, a "semi-exact" solution procedure is developed, based on a similar procedure in Gilbert et al. (2015), using the following steps: (1) the lower-level objective is approximated with piecewise linear function and re-cast as an LP, (2) the optimality (i.e. duality) conditions for the LP are added as complementarity constraints to the upper level, resulting in a single-level integer program with complementarity constraints, (3) finally, similarly to Aboolian et al. (2016), the complementarity conditions are linearized through the addition of binary decision variables, resulting in a linear MIP. The resulting model yields an approximate solution to the original model due to the piece-wise linear approximation in step (1), however this approximation can be made arbitrarily precise by increasing the number of segments, hence the "semi-exact" nature of the algorithm. It should be noted that the resulting model tends to be quite large even when the original instance is of relatively small size, leading to computational difficulties. Thus a heuristic approach is proposed as well.

While these results may be quite fragile with respect to the $M/M/1$ assumption, they do indicate that capacity discretization is not the only way to approach PA-type models. They also point out that many methods developed in the transportation economics literature may be applicable to SLCIS models as well.

We finish the previous two sections with an important message from Zhang et al. (2012). In much of the literature, the difference between deterministic utility optimization of Sect. 17.3.2 and the proportional allocation is considered mainly on theoretical grounds, focusing on the difference between utility specifications, choice axioms, etc. Theoretical arguments can be made in favor of either approach. However, as shown in Zhang et al. (2012), these different mechanisms for modeling customer response may lead to very different optimal facility network designs, with wide-ranging implications: for example, it is shown that if PA choice model is assumed, while customers are actually following the utility optimization model (or vice-versa), many of the facilities will be over/ under-used, resulting in very different congestion patterns and network performance than what is predicted by the model. Thus, the choice of customer reaction model must be made based on empirical evidence of customer behavior in a given setting, rather than theoretical arguments for one or the other model.

17.6 Conclusions

In this chapter we have focused on a rather specialized sub-field of stochastic location models: problems with congestion and static customer assignments. However, as discussed above, this is a very active and growing field of research. We believe that the key drivers of this growth are that, on the one hand, SLCIS models do capture very important trade-offs and stochastic effects that must be taken into account when designing many real-life systems. On the other hand, these models retain enough structure to enable exact algorithmic approaches and managerial

insights that may not be available when more complex models (e.g., models with mobile servers or dynamic customer assignments) are considered.

The variety of SLCIS models considered in the literature is quite bewildering. We have systematized the models along two dimensions: by customer response and demand elasticity (leading to our NR/AR/DR/FR types), and by the key structural elements of the models (leading to our CT/SO/BO/ECR/PA model classes), as described in Sect. 17.5. We believe that this classification should be useful to future researchers in this field, both with respect to the importance of clearly spelling out the assumptions with respect to customer behavior and key model objectives, and with regards to realizing what key difficulties may arise for a given model type. We are pleased to note that several papers that were published after the first edition of this volume have adopted this classification.

We also hope that the proposed systematization will motivate the authors to ensure internal consistency of implicit assumptions in their models. This should help to avoid models where customer utilities are affected by travel times, but not waiting times, or by waiting times but not by blockages, etc. Of course, such simplifications may be necessary to make the model computationally tractable, but they should be explicitly spelled out and discussed.

Many open questions remain, as should be clear from the preceding sections. The assumptions made with respect to queueing behavior in many models are quite restrictive and could likely be generalized using the approximation approaches described in Sect. 17.2.3.2. The assumptions underlying NR models or AR models with distance-only utility are questionable and could lead to under-performance of the resulting system (especially with respect to the realized demand). The reliance of many authors on heuristic approaches without the ability to benchmark the resulting solutions versus the optimal ones is not comforting given the strategic nature of decisions underlying SLCIS models.

Some important strides towards deriving exact or semi-exact solution algorithms for models with realistic customer response mechanisms have been made since the first edition and are described above. These include (1) leveraging capacity discretization to incorporate variability of service times as endogenous parameter of the model, and also to develop clever linearization schemes; (2) adapting advances in conic programming to SLCIS models, and (3) pushing the boundary on the PA-type models. However, many ways to improve on the existing models remain to be explored. We hope that some of these improvements will be investigated in the next generation of SLCIS models. The importance of basing modeling choices on empirical evidence of customer behavior must also be emphasized.

Finally we would like to mention that many of the issues that have been explored in the SLCIS context (customer response, elastics demand) are still waiting to be addressed in the models with mobile servers/dynamic customer assignments. As noted earlier, these models involve a different level of complexity, with the underlying queueing systems being much less tractable. Nevertheless, the assumptions regarding customer behavior and response are very important and deserve further study.

References

Aboolian R, Berman O, Krass D (2007) Competitive facility location model with concave demand. Eur J Oper Res 181:598–619

Aboolian R, Berman O, Drezner Z (2008) Location and allocation of service units on a congested network. IIE Trans 40:422–433

Aboolian R, Berman O, Drezner Z (2009) The multiple server center location problem. Ann Oper Res 167:337–352

Aboolian R, Berman O, Krass D (2012) Profit maximizing distributed service system design with congestion and elastic demand. Transp Sci 46:247–261

Aboolian R, Berman O, Verter V (2016) Maximal accessibility network design in the public sector. Transp Sci 50(1):336–347

Aboolian R, Berman O, Wang J (2018) Responsive make-to-order supply chain network design. Working Paper

Abouee-Mehrizi H, Babri S, Berman O, Shavand H (2011) Optimizing capacity, pricing and location decisions on a congested network with balking. Math Method Oper Res 74:233–255

Ahmadi-Javid A, Hoseinpour P (2018), Convexification of queueing formulas by mixed-integer second-order cone programming: an application to a discrete location problem with congestion. Working Paper. arXiv:1710.05794

Ahmadi-Javid A, Berman O, Hoseinpour P (2018) Location and capacity planning of facilities with general service-time distributions using conic optimization. Working Paper, arXiv:1809.00080

Ashtiani H, Magnanti T (1981) Equilibria on a congested transportation network. SIAM J Algebra Discr Methods 2:213–226

Atamtürk A, Narayanan V (2011) Lifting for conic mixed-integer programming. Math Program 126(2):351–363

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

Azizi N, Vidyarthi N, Chauhan S (2017) Modelling and analysis of hub-and-spoke networks under stochastic demand and congestion. Ann Oper Res 264:1–40

Baron O, Berman O, Krass D, Wang Q (2007) The equitable location problem on the plane. Eur J Oper Res 183:578–590

Baron O, Berman O, Krass D (2008) Facility location with stochastic demand and constraints on waiting time. Manuf Serv Oper Manag 10:484–505

Berman O, Drezner Z (2006) Location of congested capacitated facilities with distance-sensitive demand. IIE Trans 38:213–221

Berman O, Drezner Z (2007) The multiple server location problem. J Oper Res Soc 58:91–99

Berman O, Kaplan E (1987) Facility location and capacity planning with delay-dependent demand. Int J Prod Res 25:1773–1780

Berman O, Krass D (2002) Facility location problems with stochastic demands and congestion. In: Drezner Z, Hamacher H (eds) Facility location: application and theory, Springer, Berlin, pp 329–371

Berman O, Krass D, Wang J (2006) Locating service facilities to reduce lost demand. IIE Trans 38:933–94

Berman O, Drezner T, Drezner Z, Krass D (2009a) Modeling competitive facility location problems: new approaches and results. In: Oskoorouchi M (ed) Tutorials in operations research, INFORMS, pp 156–181

Berman O, Drezner Z, Tamir A, Wesolowsky G (2009b) Optimal location with equitable loads. Ann Oper Res 167:308–326

Berman O, Krass D, Tong D (2014) Pricing, location and capacity planning on a network under congestion. Working Paper, University of Toronto

Boffey B, Galvão R, Espejo L (2006) A review of congestion models in the location of facilities with immobile servers. Eur J Oper Res 178:643–662

Boffey B, Galvão R, Marianov V (2010) Location of single-server immobile facilities subject to a loss constraint. J Oper Res Soc 61:987–999
Brandeau M, Chiu S (1994) Facility location in a user-optimizing environment with market externalities: analysis of customer equilibria and optimal public facility locations. Locat Sci 2:129–147
Brandeau M, Chiu S, Kumar S, Grossman T (1995) Location with market extermalities. In: Drezner Z (ed) Facility location, Springer, Berlin, pp 121–150
Brimberg J, Mehrez A (1997) A note on the allocation of queueing facilities in a continuous space using a minisum criterion. J Oper Res Soc 48:195–201
Brimberg J, Mehrez A, Wesolowsky G (1997) Allocation of queuing facilities using a minimax criterion. Locat Sci 5:89–101
Castillo I, Ignolfsson A, Sim T (2009) Social optimal location of facilities with fixed servers, stochastic demand and congestion. Prod Oper Manag 18:721–736
Cooper L, Nakinishi M (1988) Market share analysis. Kluwer Academic Publishers, Boston
Dan T, Marcotte P (2019) Competitive facility location with selfish users and queues. Oper Res https://doi.org/10.1287/opre.2018.1781
Drezner T, Drezner Z (2011) The gravity multiple server location problem. Comput Oper Res 38:694–701
Elhedhli S (2006) Service system design with immobile servers, stochastic demand, and congestion. Manuf Serv Oper Manag 8:92–97
Fisk C (1980) Some developments in equilibrium traffic assignment. Transp Res B-Meth 14(3):243–255
Gilbert M, Marcotte P, Savard G (2015) A numerical study of the logit network pricing problem. Transp Sci 49(3): 709–719
Gross D, Harris C (1985) Fundamentals of queueing theory, 2nd edn. John Wiley and Sons, New York
Hamaguchi T, Nakade K (2010) Optimal location of facilities on a network in which each facility is operating as an M/G/1 queue. J Serv Sci Manag 3:287–297
Hopp WJ, Spearman M (2000) Factory physics, 2nd edn. McGraw Hill, New York
Hoseinpour P, Ahmadi-Javid A (2016) A profit-maximization location-capacity model for designing a service system with risk of service interruptions. Transp Res E-Log 96:113–134
Ignolfsson A (2013) EMS planning and management. In: Zaric G (ed) Operations research and health care policy. Springer Science + Business Media, New York, pp 105–128
Kakhki H, Moghadas F (2010) A semidefinite relaxation for the queueing covering location problem with an M/G/1 system. In: Proceedings of the european workshop on mixed integer nonlinear programming, pp 231–236
Kim S (2013) A column generation heuristic for congested facility location problem with clearing functions. J Oper Res Soc 64:1780–1789
Larson R (1974) A hypercube queuing model for facility location and redistricting in urban emergency services. Comput Oper Res 1:67–95
Lee H, Cohen M (1985) Equilibrium analysis of disaggregate facility choice systems subject to congestion-elastic demand. Oper Res 33:293–311
Marianov V, Rios M (2000) A probabilistic quality of service constraint for a location model of switches in ATM communications networks. Ann Oper Res 96:237–243
Marianov V, Serra D (1998) Probabilistic maximal covering location-allocation for congested system. J Reg Sci 38:401–424
Marianov V, Serra D (2011) Location of multiple-server common service centers or facilities, for minimizing general congestion and travel cost functions. Int Reg Sci Rev 34:323–338
Marianov V, Rios M, Icaza MJ (2008) Facility location for market capture when users rank facilities by shorter travel and waiting times. Europ J Oper Res 191(1):32–44
Marianov V, Boffey T, Galvão R (2009) Optimal location of multi-server congestible facilities operating as M/Er/m/N queues. J Oper Res Soc 60:674–684
McFadden D (1974) Conditional logit analysis of quantitative choice behavior. In: Zarembka A (ed) Frontiers in econometrics. Academic, New York

McFadden DL (2005) Revealed stochastic preference: a synthesis. Econ Theory 26(2), 245–264
Nagurney A (1999) Network economics: a variational inequality approach. Kulwer Academic Publishers, Boston
Pasandideh S, Chambaria A (2010) A new model for location-allocation problem within queuing framework. J Ind Eng 6:53–61
Rabieyan R, Seifbarghy M (2010) Maximal benefit location problem for a congested system. J Ind Eng 5:73–83
Schön C, Saini P (2018) Market-oriented service network design when demand is sensitive to congestion. Transp Sci https://doi.org/10.1287/trsc.2017.0797
Shen Z-J (2005) Multi-commodity supply chain design problem. IIE Trans 37:753–762
Snyder LV (2006) Facility location under uncertainty: a review. IIE Trans 38:537–554
Suzuki A, Drezner Z (2009) The minimum equitable radius location problem with continuous demand. Eur J Oper Res 195:17–30
Tavakkoli-Moghaddam R, Vazifeh-Noshafagh S, Taleizadeh A, Hajipour V, Mahmoudi A (2009) Pricing and location decisions in multi-objective facility location problem with *M/M/m/k* queuing systems. Engr Opt 49(1):136–160
Tong D (2011) Optimal Pricing and Capacity Planning in Operations Managemnet. Ph.D. Thesis, University of Toronto
Vidyarthi N, Jayaswal S (2014) Efficient solution of a class of location-allocation problems with stochastic demand and congestion. Comp Oper Res 48:20–30
Wang Q, Batta R, Rump C (2002) Algorithms for a facility location problem with stochastic customer demand and immobile servers. Ann Oper Res 111:17–34
Wang Q, Batta R, Rump C (2004) Facility location models for immobile servers with stochastic demand. Nav Res Log 51:138–152
Whitt W (1992) Understanding the efficiency of multi-server service systems. Manag Sci 38:708–723
Yang W (2018) A user-choice model for locating congested fast charging stations. Transp Res E-log 110:189–213
Zhang Y, Berman O, Verter V (2009) Incorporating congestion in preventive healthcare facility network design. Eur J Oper Res 198:922–935
Zhang Y, Berman O, Marcotte P, Verter V (2010) A bilevel model for preventive healthcare facility network design with congestion. IIE Trans 42:865–880
Zhang Y, Berman O, Verter V (2012) The impact of client choice on preventitive healthcare network design. OR Spektrum 34(2):349–370