Socially Optimal Location of Facilities with Fixed Servers, Stochastic Demand and Congestion

Ignacio Castillo • Armann Ingolfsson • Thaddeus Sim

Department of Finance and Management Science, University of Alberta School of Business, Edmonton, Alberta, Canada T6G 2R6

ignacio.castillo@ualberta.ca • armann.ingolfsson@ualberta.ca • thaddeus.sim@ualberta.ca

We present two capacity choice scenarios for the socially optimal location of facilities with fixed servers, stochastic demand and congestion. Walk-in health clinics, motor vehicle inspection stations, automobile emissions testing stations, and internal service systems are motivating examples of such facilities. The choice of locations for such facilities influences not only distances for users traveling to the facilities but also user waiting times at the facilities. In contrast to most previous research, we explicitly embed both customer travel and delay costs in the objective function and solve the location-allocation problem as well as choose service capacities for each open facility simultaneously. The choice of capacity for a facility that is viewed as a queueing system could mean choosing a service rate for the 'servers' (scenario 1) or choosing the number of servers (scenario 2). We are able to express the optimal service rate in closed form in scenario 1 and the (asymptotically) optimal number of servers in closed form in scenario 2. This allows us to eliminate both the number of servers and the service rates from the optimization problems, leading to tractable mixed-integer nonlinear programs. Our computational results show that both problems can be solved efficiently using widely available optimization software.
(Service Facility System Design; Capacity Choice; Stochastic Facility Location; Social Optimum)

1. Introduction and Motivation

The choice of locations for service facilities influences not only distances for users traveling to the facilities but also, together with capacity decisions and customers' choices, how long users can expect to wait at the facilities before receiving service. Customer choice processes are complicated, but in some situations it is reasonable to assume that each customer will choose the facility that minimizes the sum of travel and waiting costs. One such situation is walk-in health clinics. In two separate studies by Newman (1984) and Jan et al. (2000), survey respondents identified waiting times at the clinics as one of the most important factors that influence their preference for which location to patronize. Travel time, although important, was not as significant a factor as waiting time. Other motivating examples of situations where customers can reasonably be expected to choose the facility that minimizes travel and waiting costs are motor vehicle inspection stations, automobile emissions testing stations, and internal service systems (such as tool cribs in manufacturing plants or photocopiers in offices).

When users are free to choose which facility to patronize (referred to as 'user choice'), they may make choices that in the aggregate are suboptimal compared to a *social optimum*, where the facility choice is dictated to users (referred to as 'directed choice') so as to minimize the sum of travel and waiting costs for all users. When it is not possible to dictate facility choice to users, it may be possible to influence user choices through 'tolls' or differential service fees. When there is complete flexibility in setting such tolls, they can be used to induce users to act in a way that results in a social optimum (Grossman Jr. and Brandeau 2002). The optimization problems we develop are applicable in such situations, as well as in directed choice situations.

Most models in the location theory literature are deterministic. Models with stochastic demand and congestion provide greater realism by incorporating two sources of uncertainty: the exact timing of customer demand and the time it takes to serve individual customers at service facilities. Berman and Krass (2002) contains a recent review of such models. The models in the literature are conveniently classified according to two characteristics: (1) whether they are covering-type or median-type and (2) whether the servers are mobile or fixed.

Development of models with mobile servers, stochastic demand and congestion has been motivated primarily by the need to locate emergency service facilities, such as ambulance or fire stations, where servers (ambulances or fire engines) travel to the site of the emergency. In contrast, our optimization problems fall in the 'fixed servers' category, where users travel to facilities to receive service. Recent models with fixed servers that incorporate stochastic demand and congestion include the covering-type models by Marianov and Serra (1998, 2001) and Marianov and Rios (2000). These models are formulated as probabilistic maximal covering location-allocation problems with a constraint on the probability that either waiting time or queue length does not exceed a specified threshold. Modeling each facility as an $M/M/s$ or $M/D/s$ queueing system makes it possible to linearize these 'congestion constraints,' leading to relatively tractable mixed-integer linear programs. The models solved in Marianov and Serra (1998) and Marianov and Rios (2000) assume that each facility will have the same number of servers, but Marianov and Serra (2001) formulate models that allow the number of servers at each facility to be a decision variable.

Classical facility location theory and the location models we have cited, both with mobile and fixed servers, assume directed choice, with customers typically being allocated to the closest facility via a binary variable; that is, the entire demand from a particular node is allocated to a single facility. Usually, in directed choice models, congestion at the service facilities is controlled by incorporating service level constraints. Such constraints set limits on either the distribution of the number of users waiting to receive service or the distribution of user waiting times. The limits

could be on the expected values or on tail probabilities of these distributions. Such service level constraints tend to equalize the level of congestion at the different facilities, measured either in terms of the number of users waiting or the waiting time. Note that if the number of servers varies between facilities, then a constraint that equalizes the number of waiting customers at different facilities will *not* equalize waiting times at different facilities. For example, a facility with two servers and an average of 10 waiting users will have shorter waiting times on average than a facility with one server and an average of 10 waiting users, all else being equal.

We believe that the 'closest facility' assumption is typically not appropriate for fixed server models, where customers travel to facilities. When users choose facilities, Heinhold (1975, 1978) and Grossman Jr. and Brandeau (2002) argue that customers will make their choice based both on travel time to a facility and expected waiting time at the facility. Heinhold (1975, 1978) report on empirical studies of vehicle inspection stations and health clinics whose results were consistent with each user choosing a facility to minimize the sum of her travel and waiting times. Consequently, a user may choose not to travel to the closest facility but to one that is further away and less congested. If one accepts this premise, then it seems more appropriate to incorporate waiting times in the objective function and allow average waiting times to vary between facilities, if this reduces the overall time spent by all users either traveling or waiting for service, as opposed to including service-level constraints that tend to equalize waiting times between facilities.

In this paper, we propose two median-type models with fixed servers. Our interest is in optimization problems that jointly select facility locations, facility capacities, and user allocation to facilities to minimize the system-wide cost. We account for stochastic demand originating at the nodes of a network and random service times at facilities. In contrast to most previous research (Amiri (1998) and Wang et al. (2001) are exceptions), we explicitly embed both customer travel and delay costs in the objective function and solve the location-allocation problem as well as choose service capacities for each open facility simultaneously. We envision our models being used at the system design stage. Consequently, we make modeling choices to reduce the data needs of the model and to focus on the interactions between location decisions, capacity choices, and demand allocation, rather than the detailed modeling of any of these three sets of decisions.

The choice of capacity for a facility that is viewed as a queueing system could mean choosing a service rate for the 'servers', choosing the number of servers, or both. The possible values for the service rate could either consist of a continuous range of values (e.g., when the speed of a machine can be set to any value within some range) or a discrete set (e.g., when there is a choice between several different machines with different service rates or when several employees form a team that

operates as a single server).

We present and solve optimization problems for the following two capacity choice scenarios.

Scenario 1. Situations where the number of parallel servers at each open facility and their common service rate can be any positive number. In such situations it is optimal to operate each open facility with a single server.

Scenario 2. Situations where the number of parallel servers at each open facility can be any positive integer but the service rate is fixed. We employ approximations that allow the number of servers to be a continuous variable.

In scenario 1, where each facility operates with a single server, we assume that the cost of providing service is proportional to the service rate of that server. In scenario 2, we assume the cost of providing service to be proportional to the number of servers. Both of these cost structures are captured by assuming the cost of providing service at a facility to be proportional to the product of the number of servers and the service rate for that facility and this is the assumption we use in presenting the general system we are considering.

We are able to express the optimal service rate in closed form in scenario 1 and the (asymptotically) optimal number of servers in closed form in scenario 2. This allows us to eliminate both the service rates and the number of servers from the optimization problems we formulate leading to tractable mixed-integer nonlinear programs that can be solved efficiently using standard optimization software. Moreover, we are able to show that the problems for scenarios 1 and 2 are structurally identical and share the same useful properties. This implies that facilities with multiple servers can be modeled and solved without any increase in computational effort compared to single server facilities.

The rest of the paper is organized as follows. Section 2 gives a detailed description of the general system we consider. Section 3 discusses the single server scenario in which the optimal location-allocation and the service rate for each open facility are determined simultaneously. In Section 4, we discuss the multiple servers scenario in which the optimal location-allocation and the number of parallel servers to be housed at each facility are determined simultaneously. In Section 5, we illustrate the application of our optimization problems to two previously published examples. Finally, conclusions and suggestions for future work are given in Section 6.

2. General Problem Description

We consider a general system with M candidate facility locations. Each open facility j will house s_j parallel servers with common service rate $\mu_j,\ j = 1, \ldots, M$. Service times at each facility are assumed to be i.i.d. and exponentially distributed random variables with mean $1/\mu_j$. The servers are fixed and hence the customers must travel to an open facility to receive service.

We assume customers are aggregated to C locations. At each location, customers generate service demands according to an independent Poisson process with rate $\lambda_i,\ i = 1, \ldots, C$. When a customer from location i requires service, we assume she chooses to go to facility j with probability Y_{ij}, where $0 \leq Y_{ij} \leq 1$ and $\sum_{j=1}^{M} Y_{ij} = 1$, independent of the demand generation processes and the choices made by other customers. The expected demand rate for facility j is therefore $\Lambda_j = \sum_{i=1}^{C} Y_{ij}\lambda_i$. Naturally, we have $\sum_{j=1}^{M} \Lambda_j = \sum_{i=1}^{C} \lambda_i$. We will require that $\Lambda_j < s_j\mu_j,\ j = 1, \ldots, M$ to ensure queue stability. With our assumptions, each facility operates as an $M/M/s_j$ queue with arrival rate Λ_j and service rate μ_j.

Our objective is to minimize total long-term expected costs. Our decision variables are a location vector $\boldsymbol{X}^\star = \{X_j^\star\}$, where

$$X_j^\star = \begin{cases} 1 & \text{if we open a facility at candidate facility location } j \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

a server vector $\boldsymbol{S}^\star = \{s_j^\star\}$, a service rate vector $\boldsymbol{U}^\star = \{\mu_j^\star\}$, and a service demand allocation matrix $\boldsymbol{Y}^\star = \{Y_{ij}^\star\}$.

This is referred to as social optimization in the queueing design and control literature (see Gross and Harris (1998)) or finding a system optimum in the transport equilibrium literature (see Beckmann et al. (1956) and Sheffi (1985)); that is, we minimize the cost of the entire system rather than finding a 'user equilibrium,' where each user minimizes her individual costs, given the choices made by other users. System optimum (and user equilibrium) formulations for transportation networks model delays due to congestion on links in the transportation network and assume the capacity of each link is given. In contrast, our formulations model congestion at the nodes (i.e., facilities) of the network and incorporate the choice of capacity for each facility.

We assume that there is a fixed cost of C_f for each open facility, a cost of traveling per unit time of C_t for each customer, a cost of queueing and service delay per unit time of C_q for each customer, and a cost of providing service per unit of capacity of C_c. By 'unit of capacity' we mean

having one server available for $1/\mu_j$ time units. We denote the system-wide expected cost by

$$z(\boldsymbol{Y},\boldsymbol{X},\boldsymbol{S},\boldsymbol{U}) = C_f \sum_{j=1}^{M} X_j + C_t \sum_{j=1}^{M}\sum_{i=1}^{C} Y_{ij}\lambda_i t_{ij} + C_q \sum_{j=1}^{M} L(\Lambda_j,\mu_j,s_j) + C_c \sum_{j=1}^{M} s_j\mu_j, \qquad (2)$$

where t_{ij} denotes the expected travel time between customer location i and candidate facility location j, $L(\Lambda_j,\mu_j,s_j)$ denotes the average number of users in open facility j (waiting to be served or being served), and C_f, C_t, C_q, and C_c are positive cost coefficients.

Estimating the cost coefficients may be difficult. An approach, when no cost estimates are available, is to assume $C_t = C_q = 100$ and then solve the problem parametrically for several values of a function of C_f and $C_c\mu$ (e.g., $C_f/(C_c\mu)$), which can be interpreted as comparing the cost of providing service to that of a user (in this instance, we assume the service rate is the same for all open facilities). If the facilities are such that users typically make a special trip to receive the service, then one could set $C_t = 200$ to account for round trip travel. In more data rich situations, it may be appropriate to use different cost coefficients for users at different locations and servers at different facilities. The objective function can be generalized (with the definitions of the cost coefficients extended in the obvious way) as follows to allow for this:

$$z(\boldsymbol{Y},\boldsymbol{X},\boldsymbol{S},\boldsymbol{U}) = \sum_{j=1}^{M} C_{fj} X_j + \sum_{j=1}^{M}\sum_{i=1}^{C} C_{ti} Y_{ij}\lambda_i t_{ij} + \sum_{j=1}^{M} L(\Lambda_j,\mu_j,s_j)\frac{1}{\Lambda_j}\sum_{i=1}^{C} C_{qi}\lambda_i + \sum_{j=1}^{M} C_{cj} s_j\mu_j. \quad (3)$$

Observe that when the service rates at each facility are fixed, the cost of providing service, at each facility, becomes proportional to the number of servers. In this situation, one can interpret $C_{cj}\mu_j$ to be the labor cost per time unit (if the servers are people) at facility j.

Instead of including a fixed cost of opening each facility, our model allows for the possibility of pre-specifying the maximum number of facilities to be opened, denote P. Solving the problem parametrically in P may provide valuable insight at the system design stage.

Our location problem can now be formulated as follows:

$$(P0): \quad \min \quad C_f \sum_{j=1}^{M} X_j + C_t \sum_{j=1}^{M} \sum_{i=1}^{C} Y_{ij} \lambda_i t_{ij} + C_q \sum_{j=1}^{M} L(\Lambda_j, \mu_j, s_j) + C_c \sum_{j=1}^{M} s_j \mu_j \tag{4}$$

$$\text{s.t.} \quad \sum_{j=1}^{M} Y_{ij} = 1, \ i = 1, \ldots, C, \tag{5}$$

$$Y_{ij} \leq X_j, \ i = 1, \ldots, C, \ j = 1, \ldots, M, \tag{6}$$

$$s_j \leq \kappa_1 X_j, \ j = 1, \ldots, M, \tag{7}$$

$$\mu_j \leq \kappa_2 X_j, \ j = 1, \ldots, M, \tag{8}$$

$$\sum_{j=1}^{M} X_j \leq P, \tag{9}$$

$$\sum_{i=1}^{C} Y_{ij} \lambda_i - s_j \mu_j < 0, \ j = 1, \ldots, M, \tag{10}$$

$$Y_{ij} \geq 0, \ i = 1, \ldots, C, \ j = 1, \ldots, M, \tag{11}$$

$$s_j \in \{0, 1, 2, \ldots\}, \ j = 1, \ldots, M, \tag{12}$$

$$\mu_j \geq 0, \ j = 1, \ldots, M, \tag{13}$$

$$X_j \in \{0, 1\}, \ j = 1, \ldots, M, \tag{14}$$

where $\kappa_1 > 0$ and $\kappa_2 > 0$ are large numbers that guarantee feasibility. Constraint (5) ensures that the service demands for each customer location are met; constraints (6), (7), and (8) ensure that service demands, servers, and service rates, respectively, are allocated to open facilities; constraint (9) ensures that at most P facilities are opened; and constraint (10) ensures that queue stability conditions are not violated. Location theory models commonly require the demand allocation variables Y_{ij} to be binary. In our two capacity choice scenarios, we allow these variables to take any value between zero and one, but as Proposition 4 will show, the optimization problems will have an optimal solution where $\boldsymbol{Y}$ is integer (i.e., zero or one).

3. Scenario 1: Single-Server Facilities

We now specialize problem $(P0)$ to a situation where it is possible to vary the service rate μ_j at each facility over $[0, \infty)$. Amiri (1998) and Wang et al. (2001) formulated similar problems. However, these authors incorporated the 'closest facility' assumption, which we believe does not accurately model the motivating examples we are considering. Under the varying service rate assumption, it is optimal to have a single server at each open facility, as stated formally in the following proposition.

Proposition 1 *Given a service demand allocation matrix* $\boldsymbol{Y}$ *and a location vector* $\boldsymbol{X}$*, the optimal number of fixed servers at each open facility* j *is* $s_j^\star = 1$*.*

(The proofs of propositions 1 and 2 are given in the Appendix.) We use $z_{M/M/1}(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{S}, \boldsymbol{U})$ to denote the objective function in this section.

Stidham Jr. (1970) proves the optimality of having a single server for considerably more general systems than $M/M/s_j$. The critical assumptions are that it is possible to vary the service rate μ_j and that the cost of providing service is proportional to $\mu_j s_j$.

Thus, the capacity decision for each open facility reduces to choosing an optimal service rate $\mu_j^\star$ for an $M/M/1$ queue. The optimization problem in this section could be used as an approximation for two situations where the assumptions do not strictly hold. First, when n_j employees work together as a team (i.e., the n_j employees serve one customer together, and finish one customer before beginning to serve the next) with service rate $n_j\mu^0$, then the number of employees to have at facility j could be approximated as $n_j = \mu_j^\star/\mu^0$, rounded up or down. Second, if each facility operates as a queue with s_j parallel servers and fixed service rate μ^0, and the facility experiences heavy traffic (i.e., $\rho_j \approx 1$), then such a facility will operate almost identically to a single server queue with service rate $s_j\mu^0$. Hence, the optimal number of servers can be approximated as $s_j = \mu_j^\star/\mu^0$. In our optimization problem one cannot predict ahead of time whether all facilities will experience heavy traffic (it depends on how users are allocated to facilities). However, if C_f and/or C_c dominate the cost structure, then it is likely that the optimal solution will have facilities experiencing heavy traffic.

Substituting $s_j^\star = 1,\ j = 1, \ldots, M$, we can express the system-wide cost as follows:

$$\begin{aligned} z_{M/M/1}(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{S^\star}, \boldsymbol{U}) &= C_f \sum_{j=1}^{M} X_j + C_t \sum_{j=1}^{M}\sum_{i=1}^{C} Y_{ij}\lambda_i t_{ij} + C_q \sum_{j=1}^{M} L(\Lambda_j, \mu_j, 1) + C_c \sum_{j=1}^{M} \mu_j \\ &= C_f \sum_{j=1}^{M} X_j + C_t \sum_{j=1}^{M}\sum_{i=1}^{C} Y_{ij}\lambda_i t_{ij} + C_q \sum_{j=1}^{M} \frac{\Lambda_j}{\mu_j - \Lambda_j} + C_c \sum_{j=1}^{M} \mu_j. \end{aligned} \tag{15}$$

We now eliminate the service rates from the optimization problem using Proposition 2.

Proposition 2 *Given a service demand allocation matrix* $\boldsymbol{Y}$ *and a location vector* $\boldsymbol{X}$*, the optimal service rate* $\mu_j^\star$ *at each facility* j *is*

$$\mu_j^\star = \Lambda_j + \sqrt{\frac{C_q}{C_c}}\sqrt{\Lambda_j}, \tag{16}$$

where $\Lambda_j = \sum_{i=1}^{C} Y_{ij}\lambda_i$*.*

This result is well known (see Gross and Harris (1998)). The expression for $\mu_j^\star$ can be interpreted to consist of the sum of the minimum capacity Λ_j needed for queue stability and a strictly positive 'safety capacity' $\sqrt{C_q/C_c}\sqrt{\Lambda_j}$. In words, a service rate of at least Λ_j is required to guarantee queue stability; however, safety capacity must be added to the minimum as a protection against stochastic variability. The safety capacity is proportional to $\sqrt{\Lambda_j}$ and the proportionality coefficient $\sqrt{C_q/C_c}$ is given. This proportionality coefficient is relatively robust to the values of the cost coefficients due to the square root function; that is, it is not necessary to estimate C_q and C_c precisely to obtain plausible results. Moreover, with C_q and C_c fixed, an n-fold increase in Λ_j requires only a $\sqrt{n}$-fold increase in the safety capacity, which constitutes significant economies of scale.

By substituting $\mu_j^\star$ in the objective function and simplifying we can express the system-wide expected cost as follows:

$$z_{M/M/1}(\boldsymbol{Y},\boldsymbol{X},\boldsymbol{S}^\star,\boldsymbol{U}^\star) = C_f\sum_{j=1}^{M} X_j + C_c\sum_{i=1}^{C}\lambda_i + C_t\sum_{j=1}^{M}\sum_{i=1}^{C} Y_{ij}\lambda_i t_{ij} + 2\sqrt{C_qC_c}\sum_{j=1}^{M}\sqrt{\sum_{i=1}^{C} Y_{ij}\lambda_i}. \quad (17)$$

Note that the second term is constant and can be eliminated from the objective function.

From the discussion above, the number of servers and the service rates can be removed as decision variables. Constraints (7), (8) and (10) are eliminated from the general problem ($P0$) giving:

$$\begin{aligned} (P1): \quad \min \quad & C_f\sum_{j=1}^{M} X_j + C_t\sum_{j=1}^{M}\sum_{i=1}^{C} Y_{ij}\lambda_i t_{ij} + 2\sqrt{C_qC_c}\sum_{j=1}^{M}\sqrt{\sum_{i=1}^{C} Y_{ij}\lambda_i} \quad (18)\\ \text{s.t.} \quad & (5),\ (6),\ (9),\ (11),\ (14). \end{aligned}$$

Problem ($P1$) is a mixed-integer nonlinear program with linear constraints and a nonlinear objective function. It has CM continuous decision variables, M binary decision variables, and $C(M+1)+1$ linear constraints. Amiri (1998) formulated a similar problem and presented a Lagrangian relaxation heuristic to solve it. Wang et al. (2001) also formulated a similar problem and developed a greedy adding heuristic to solve it. We note that it is probably possible to devise additional heuristic solution procedures. However, the objective function in our optimization problem has a useful *pseudolinearity* property that guarantees that nonlinear programming algorithms combined with branch and bound will reach global optimality (Bazaraa et al. 1993; Martos 1975).

Proposition 3 *$z_{M/M/1}(\boldsymbol{Y},\boldsymbol{X},\boldsymbol{S}^\star,\boldsymbol{U}^\star)$ is pseudolinear; that is, it is both pseudoconvex and pseudoconcave.*

Proof. The functions $\sum_{j=1}^{M} X_j$ and $\sum_{i=1}^{C} Y_{ij}\lambda_i t_{ij}$ are linear and are therefore pseudolinear as well. Since addition and multiplication with a positive constant preserves pseudocovexity and pseudoconcavity, it remains only to show that $\sqrt{\sum_{i=1}^{C} Y_{ij}\lambda_i}$ is pseudolinear.

Martos (1975) shows that if $f : I\!R_n \rightarrow I\!R$ is a pseudoconvex (pseudoconcave) function and $g : I\!R \rightarrow I\!R$ is an increasing function, then the composite function $h(\cdot) = g(f(\cdot))$ is pseudoconvex (pseudoconcave). Since $\sum_{i=1}^{C} Y_{ij}\lambda_i$ is a linear function (and therefore pseudolinear) and $\sqrt{\cdot}$ is an increasing function, $\sqrt{\sum_{i=1}^{C} Y_{ij}\lambda_i}$ is pseudolinear. □

Pseudoconvexity implies quasiconvexity and strict quasiconvexity and, similarly, pseudoconcavity implies quasiconcavity and strict quasiconcavity. (See Bazaraa et al. (1993) or Martos (1975) for definitions of these properties.) Since $z_{M/M/1}(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{S}^\star, \boldsymbol{U}^\star)$ is pseudolinear, it has all of these properties. When the feasible region is bounded, as it is for $(P1)$, quasiconvexity and quasiconcavity of the objective function imply that the function has a minimum at an extreme point of the feasible region. Furthermore, strict quasiconvexity and strict quasiconcavity imply that a local minimum is a global minimum. Therefore, an optimal solution to $(P1)$ can be found by searching among the extreme points of the feasible region until a point satisfying the Karush-Kuhn-Tucker conditions is reached. We next show that $\boldsymbol{Y}^\star$ is integer at such a point.

Proposition 4 *Problem* $(P1)$ *has a globally optimal solution that is integer in* $\boldsymbol{Y}^\star$.

Proof. Since $z_{M/M/1}(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{S}^\star, \boldsymbol{U}^\star)$ is strictly quasiconvex and strictly quasiconcave and has a global minimum solution at an extreme point of the feasible region, it suffices to show that the constraint matrix corresponding to constraints (5) and (6) is totally unimodular. The non-negativity constraint (11) has no effect on total unimodularity and constraints (9) and (14) pose no restrictions on $\boldsymbol{Y}$.

Let the constraint matrix be $\boldsymbol{A}' = \begin{pmatrix} \boldsymbol{A} \\ \boldsymbol{I} \end{pmatrix}$, where $\boldsymbol{A}$ is the matrix for (5) and $\boldsymbol{I}$ is the identity matrix for (6). Total unimodularity of a matrix $\boldsymbol{C}$ implies that both $(\boldsymbol{C}, \boldsymbol{I})$ and $\boldsymbol{C}'$ are totally unimodular and therefore it suffices to show that $\boldsymbol{A}$ is totally unimodular.

If the columns of $\boldsymbol{A}$, corresponding to the variables Y_{ij}, are ordered lexicographically by (i, j), then $\boldsymbol{A}$ will have the consecutive ones property (see Fulkerson and Gross (1965)). Therefore, $\boldsymbol{A}$ is totally unimodular. □

If one imagines the facility locations $\boldsymbol{X}$ being fixed, then constraints (6) simply eliminate some of the Y_{ij} variables, namely the ones corresponding to demand allocation to facilities that have not been opened. Constraints (5) are identical to the supply constraints in a transportation problem

and are therefore totally unimodular. The only other constraint set in $(P0)$ involving $\boldsymbol{Y}$ is (10), which ensures queue stability at each facility. If this constraint set were present in $(P1)$, then it would destroy the unimodularity property, but the elimination of $\boldsymbol{U}$ and $\boldsymbol{S}$ as decision variables removed the need to include this constraint set.

When it is necessary to require that each customer location is allocated entirely to a single open facility (i.e., when $\boldsymbol{Y}$ must be integer), then this integrality property allows for substantial reduction in computational effort. In addition, allowing $\boldsymbol{Y}$ to take fractional values in problem $(P1)$ permits sensitivity analysis given an optimal location vector $\boldsymbol{X}^\star$. In particular, it is possible to test how sensitive the global optimal service demand allocation matrix $\boldsymbol{Y}^\star$ is to the accuracy of the cost coefficients.

We now turn our attention to the multi-server capacity choice scenario. To our knowledge similar problems have not been proposed previously in the literature.

4. Scenario 2: Multi-Server Facilities

In this section, we specialize problem $(P0)$ for the situation where the service rate μ_j for each facility is fixed but the number of servers s_j can be varied. For clarity, we denote the objective function by $z_{M/M/s}(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{S})$ in this section, leaving the service rate vector $\boldsymbol{U}$ out of the argument list since it is assumed fixed.

We will use approximations for the expected number of customers and the optimal number of servers $s_j^\star$ in an $M/M/s_j$ system. These approximations will allow us to eliminate the number of servers from problem $(P0)$ in much the same way as in the single server facilities scenario.

We use approximations for $M/M/s_j$ systems for two reasons. First, exact expressions for performance measures for such systems are only defined for integer values of s_j. When solving continuous relaxations of $(P0)$ it is useful to be able to treat the s_j's as continuous variables. Second, we know of no exact results that would allow us to express the optimal number of servers at each facility in closed form. The approximations we use result in expressions that are very similar to the expressions for single server facilities.

The approximations are based on an asymptotic analysis of the $M/M/s$ queue developed by Halfin and Whitt (1981) and extended by Borst et al. (2002), where both the arrival rate and the number of servers tend to infinity simultaneously in a particular fashion. This analysis results in approximations that are remarkably accurate even when the arrival rate or the number of servers are relatively small.

Let $r_j = \Lambda_j/\mu_j$ be the offered load to facility j. Borst et al. (2002) provide the following approximation for the optimal number of servers (under cost assumptions that are more general than ours):

$$s_j^\star = r_j + y_j^\star \sqrt{r_j}, \tag{19}$$

where $y_j^\star$ is a constant. They also provide approximations for the probability of queueing delay and the expected number of users waiting or being served, assuming the optimal number of servers is used, as follows:

$$P(y_j^\star) = \left\{ 1 + \frac{y_j^\star \Phi(y_j^\star)}{\phi(y_j^\star)} \right\}^{-1}, \tag{20}$$

$$L(\Lambda_j, \mu_j, s_j^\star) = r_j + \frac{P(y_j^\star)}{y_j^\star} \sqrt{r_j}, \tag{21}$$

where $\Phi(y_j^\star)$ and $\phi(y_j^\star)$ are the CDF and the PDF for a standard Normal random variable, respectively. The constant $y_j^\star$ is determined from

$$y_j^\star = \arg\min_{y>0} \left\{ y + \frac{c_j P(y)}{y} \right\}, \tag{22}$$

where $c_j = C_q/(C_c \mu_j)$. Borst et al. (2002) prove their results for the case where the cost of providing service is convex and strictly increasing in the number of servers and the cost of waiting is strictly increasing in the time each customer spends waiting. Our linear cost assumptions are a special case of these assumptions.

The analysis in Borst et al. (2002) proves that the expression (19) for the number of servers is optimal in the limit when the arrival rate approaches infinity. However, they also demonstrate through numerical experiments that the result of this expression, rounded to an integer, usually identifies the optimal number of servers correctly even when the arrival rate is comparatively low. Our computational results in Section 5 confirm this.

As in the previous section, the expression (19) for $s_j^\star$ can be interpreted to consist of the sum of the minimum capacity r_j needed for queue stability and a strictly positive 'safety capacity' $y_j^\star \sqrt{r_j}$. The quantity $y_j^\star$ is a constant that depends only on the cost ratio c_j and can be determined by solving a single-variable convex minimization problem before solving the location-allocation problem. Like the square root function, the proportionality coefficient $y_j^\star$ grows relatively slowly as a function of the c_j and is therefore relatively robust to the values of the cost coefficients C_q and C_c.

By substituting the expression for $L(\Lambda_j, \mu_j, s_j^\star)$ in the objective function we can express the system-wide expected cost as follows:

$$z_{M/M/s_j}(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{S}^\star) = C_f \sum_{j=1}^{M} X_j + C_c \sum_{j=1}^{M} \Lambda_j + \sum_{j=1}^{M} \sum_{i=1}^{C} Y_{ij} \lambda_i \left(C_t t_{ij} + \frac{C_q}{\mu_j} \right) + \sum_{j=1}^{M} \left\{ \frac{C_q P(y_j^\star)}{y_j^\star} + C_c y_j^\star \mu_j \right\} \sqrt{r_j}. \tag{23}$$

Note that the second summation is constant and can be eliminated from the objective function.

From the discussion above, it is evident that the number of parallel servers can be removed as decision variables and constraints (7), (8) and (10) can be eliminated from the general problem, leading to:

$$(P2): \quad \min \quad C_f \sum_{j=1}^{M} X_j + \sum_{j=1}^{M} \sum_{i=1}^{C} Y_{ij} \lambda_i \left(C_t t_{ij} + \frac{C_q}{\mu_j} \right) + \sum_{j=1}^{M} \left\{ \frac{C_q P(y_j^\star)}{y_j^\star} + C_c y_j^\star \mu_j \right\} \sqrt{r_j} \tag{24}$$
$$\text{s.t.} \quad (5),\ (6),\ (9),\ (11),\ (14).$$

Problem $(P2)$, like problem $(P1)$, is a mixed-integer nonlinear program with CM continuous decision variables, M binary decision variables, and $C(M+1)+1$ linear constraints.

The problem Problems $(P2)$ and $(P1)$ have the same constraints and the objective function for both problems can be written as

$$C_f \sum_{j=1}^{M} X_j + \sum_{j=1}^{M} \sum_{i=1}^{C} Y_{ij} D_{ij} + \sum_{j=1}^{M} E_j \sqrt{\sum_{i=1}^{C} Y_{ij} \lambda_i} \tag{25}$$

for appropriately defined coefficients D_{ij} and E_j. Consequently, $(P2)$ is structurally identical to $(P1)$ and shares all of its properties, including pseudolinearity of the objective function and the existence of a global minimum that is integer in $\boldsymbol{Y}^\star$. This implies that facilities with multiple servers can be modeled and solved without any increase in computational effort compared to single server facilities. This is in contrast to the model in Marianov and Serra (2001) which allows the number of servers at each facility to be a decision variable but at considerable cost in model complexity.

When an optimal solution to $(P2)$ has been found, one can calculate the offered loads r_j at each facility and the approximately optimal number of servers $s_j^\star$ using (19). However, this may result in a non-integer number of servers. To obtain an integer number of servers, one can either round up or down. One could evaluate an exact version of the objective function (one that uses an exact expression for the $M/M/s_j$ queue) to choose between rounding up or down. The resulting feasible solution provides an upper bound on the optimum system-wide cost.

The approximate expression (19) for the optimal number of servers is likely to be least accurate when the number of servers is small. When the optimal solution to $(P2)$ involves a small number of servers, one could perform a more exhaustive search. For example, for facilities where $s_j^\star \leq 5$, one could choose s_j as the value in $\{0, 1, 2, \ldots, 10\}$ that gives the lowest overall cost, evaluated using an exact version of the objective function.

Since problem $(P2)$ involves approximations, it would be useful to obtain bounds on the optimum cost for an optimization problem that does not employ these approximations. An upper bound can be obtained by constructing a feasible solution by rounding the number of servers at each facility to an integer value, as we have described. The following procedure can be used to generate a lower bound on the optimum cost. First, solve $(P2)$ with $C_q = C_c = 0$. This eliminates the nonlinear portion of the objective function and provides a lower bound $z_{M/M/s}^{f,t}$ on the system-wide fixed and travel cost. Second, to obtain a lower bound on the system-wide waiting and capacity cost, suppose that all users are served by the facility with the highest service rate $\mu_{\max} = \max_{j=1,2,\ldots,M}\{\mu_j\}$. One could simply use grid search to minimize the sum of waiting and capacity costs, employing an exact expression for the expected number in system; that is, we find the $s^\star$ that minimizes

$$z_{M/M/s}^{q,c}(s) = C_q L\left(\sum_{i=1}^{C}\lambda_i, \mu_{\max}, s\right) + C_c s\mu_{\max}. \tag{26}$$

The optimal number of servers can be found quickly since L is convex in s, which follows from the result in Dyer and Proll (1977) and Little's Law. One can use (19) to provide a starting point for the search. Due to economies of scale, the sum of the system-wide waiting and capacity costs would be minimized by assigning all users to the facility with the highest service rate, if this were possible. Consequently, the sum $z_{M/M/s}^{f,t} + z_{M/M/s}^{q,c}(s^\star)$ provides a lower bound on the optimum cost.

5. Numerical Results

We applied $(P1)$ and $(P2)$ to two published examples. We used the optimization package MINLP (Fletcher and Leyffer 1994) accessed via the NEOS Server (Czyzyk et al. 1998). MINLP implements a branch and bound algorithm searching a tree whose nodes correspond to continuous nonlinearly constrained programs. The continuous problems are solved using a sequential quadratic programming solver which is suitable for solving large nonlinearly constrained problems. In view of propositions 3 and 4, MINLP (or any other mixed-integer nonlinear programming solver) is guaranteed to reach global optimality with integer demand allocation variables when solving $(P1)$ or $(P2)$.

MINLP is a gradient-based solver. The partial derivative with respect to Y_{ij} of the term in the objective function that involves $\sqrt{\Lambda_j}$ is $\lambda_i \left(2\sqrt{\Lambda_j}\right)^{-1}$, which approaches infinity when Λ_j (the demand allocated to facility j) approaches zero. This might cause numerical problems for gradient-based solvers depending on how the numerical derivatives are computed. To prevent this, one can add constraints or bounds that eliminate solutions with $\Lambda_j \approx 0$ and $X_j = 1$. Such constraints or bounds eliminate solutions that open a facility but allocate no demand to it, but other than that they do not alter the feasible region of $(P1)$ or $(P2)$. We took this approach to ensure the stability of MINLP. Alternatively, one could use solvers that do not require gradient information. The cutting plane procedure in Westerlund and Pörn (2002) is one such solver.

The first example, from Hillier and Lieberman (2001), involves the location of tool cribs in a plant and the second example, from Marianov and Serra (1998), involves the location of walk-in health clinics in a geographical region.

5.1 Locating tool cribs

Hillier and Lieberman (2001) give the following plant design example. A factory, whose shape is shown in Figure 1, requires one or more tool cribs to store tools required by mechanics who work in the factory. Each tool crib is staffed by one or more clerks. A mechanic, when requiring tools, travels to one of the tool cribs, waits for service if all clerks at that tool crib are busy, obtains the tools, and travels back to her original location. The process repeats itself when the mechanic returns the tools.

The time for a clerk to provide service to a mechanic follows an exponential distribution with rate $\mu = 120$ mechanics per hour. Arrivals of mechanics needing to either obtain or return tools is modeled as a spatial Poisson process with a uniform rate over the L-shaped factory area, with a total arrival rate of 120 mechanics per hour. A clerk costs the company \$20 per hour and the value of a mechanic's output (when busy) is \$48 per hour. Consequently, we set $C_q =$ \$48 per hour, $C_t =$ \$96 per hour (to account for travel to and from a tool crib), and $C_c =$ \$20$/\mu =$ \$0.167 per unit of capacity (recall that we define the total cost of service to be proportional to the total service capacity $\sum_{j=1}^{M} \mu_j s_j$). The fixed cost of each open tool crib is $C_f =$ \$16 per hour.

The required decisions are which of the three potential tool crib locations to use (see Figure 1), how many clerks to have at each tool crib, and how to divide the factory area between the tool cribs that are opened so as to minimize the system-wide expected cost.

Hillier and Lieberman (2001) solve the problem by enumerating the possible solutions. They reduce the number of possible solutions by assuming that all tool cribs have the same number

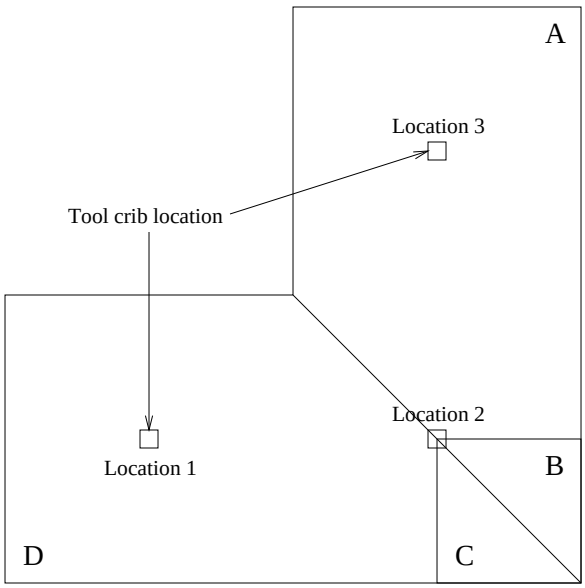

Figure 1: Plant layout, with possible tool crib locations.

of clerks and will receive the same proportion of the total demand in the factory – assumptions that are quite intuitive given the factory shape and uniform demand distribution in this example. However, even in this simple and idealized example, equal allocation of users to open facilities may not be optimal, as we now explain.

Suppose it has been decided to use locations 1 and 3, with one server at each facility and further suppose that the service rates μ_1 and μ_3 can be set to any desired values; that is, our first capacity choice scenario. If the only concern is to minimize total travel time for the mechanics (the first term in the objective function of $(P1)$), then an optimal solution would be to divide the factory equally, with a diagonal dividing line as shown in Figure 1. All points on the dividing line are equidistant to locations 1 and 3, but so are all points in the square consisting of triangles B and C shown in Figure 1. (See the Appendix for proof.) As a result, there are infinitely many alternate travel time minimizing solutions, corresponding to all possible ways of dividing the aforementioned square between locations 1 and 3. Consider now the minimization of the second term in the objective function of $(P1)$. For this example, if we define F to be the fraction of the total factory area that is allocated to location 1, this reduces to minimizing $\sqrt{F}+\sqrt{1-F}$ over $F \in [0,1]$. It is easily verified that this expression is minimized at $F = 0$ and $F = 1$ and maximized at $F = 0.5$. In words, any deviation from allocating an equal number of users to the two facilities reduces the total waiting and server cost. As a result, under the objective function used in problem $(P1)$, a solution that allocates all points that are equidistant to both facilities to the same tool crib will have lower cost

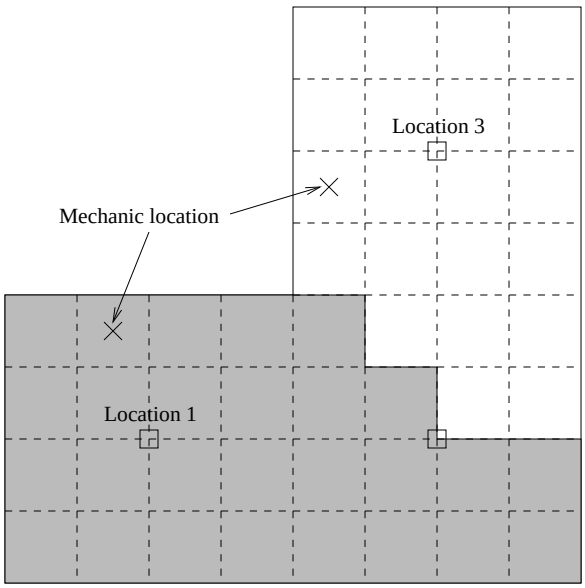

Figure 2: The size $n = 4$ instance of the tool crib example. With $P = 2$, it is optimal to locate tool cribs at locations 1 and 3. Mechanics that are allocated to location 1, assuming locations 1 and 3 have tool cribs, are shown as shaded.

than a solution that allocates equal areas to the two tool cribs.

To apply $(P1)$ and $(P2)$ to this example, we must aggregate the users into nodes, as illustrated in Figure 2. The factory layout consists of three adjoining squares. We define a size n instance of the example to be one where each of the three squares is divided into n^2 smaller squares. We aggregate the demand in each small square into a point at the centre of the square and call it a mechanic. For instance, a size 4 problem has a total of 48 mechanics. Each mechanic generates service demand following a Poisson process with rate $\lambda_i = 120/(3n^2)$. The mechanics' travel distances are calculated using the l_1 (rectilinear) norm and are converted into travel time by assuming a constant travel speed.

We began by solving $(P1)$ with a specified maximum number of facilities P but leaving out the fixed facility cost. Figure 2 shows the optimal solution to a size 4 instance of $(P1)$ with $P = 2$. As expected, it is optimal to open locations 1 and 3 and all mechanics that are equidistant from locations 1 and 3 are assigned to the same facility (facility 1 in this case). Two of the equidistant 'mechanics' in this instance correspond to a square with half its area closer to location 1 and half its area closer to location 3. One can view the allocation of these mechanics to facility 1 (as opposed to equal division between the two facilities) as an aggregation error. As n increases, this type of aggregation error is reduced. Figure 3 shows how the optimal objective function value for $(P1)$ changes with the instance size n. For benchmark purposes, we also show the objective function

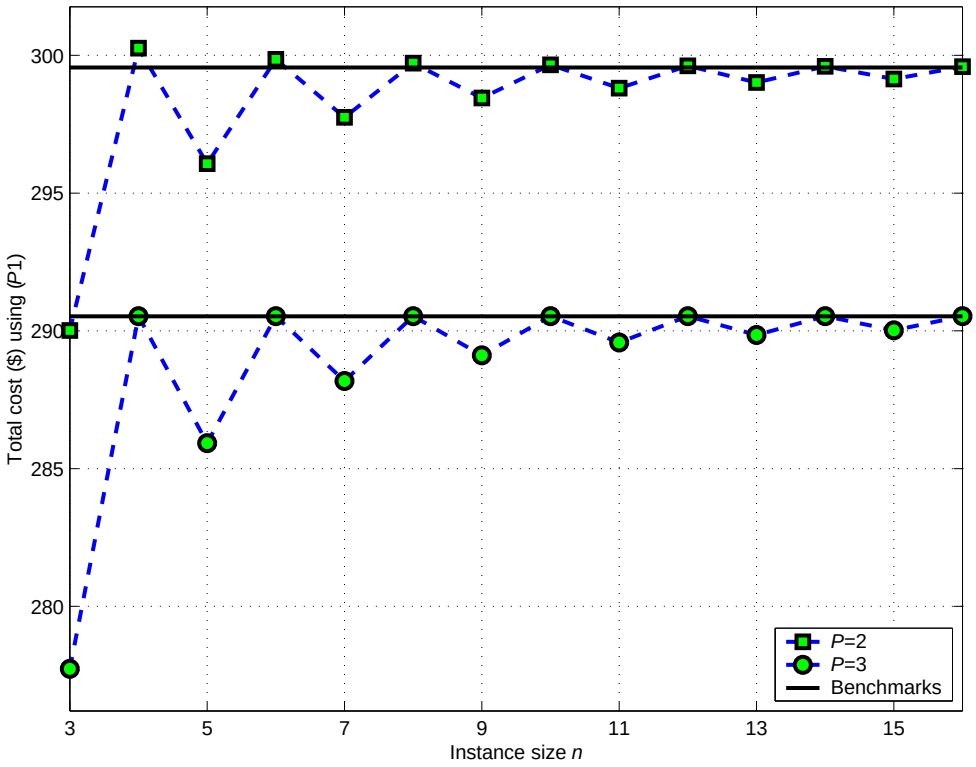

Figure 3: Optimal objective function value for the tool crib example using ($P1$) as a function of instance size n.

value obtained by allocating region A to location 3 and regions B, C,and D to location 1 and calculating expected travel times as shown in Hillier and Lieberman (2001) (see the Appendix). We see that the optimal objective function value for ($P1$) is within 1% of the benchmark value for $n \geq 6$ and within 0.2% for n even.

Figure 3 also shows how the optimal objective function value for ($P1$) changes with the instance size n supposing that all three locations can be used; that is, using $P = 3$ as the maximum number of facilities to be opened but excluding the facility cost from the objective function. Again, we compare the objective function values to a benchmark, obtained by assuming that all three facilities are open and each facility is allocated the demand from one of the three adjoining squares that make up the factory layout. We see that for $P = 3$ the optimal objective function value for ($P1$) is within 1% of the benchmark value for $n \geq 6$ and is exact for n even.

Next, we applied ($P2$) to this example with the fixed facility cost included in the objective function. Observe that in this instance, the total offered load is $120/120 = 1$ and the utilization of each facility equals 33%. The low offered load and light utilization might lead one to expect that our model ($P2$), which uses the asymptotic approximation (19), will be unreliable in identifying

the optimal solution. This was not the case. The optimal solution that we obtained for $n = 4$ and $P = 3$ was to open all three facilities with $s_j^\star = 0.991,\ j = 1, \ldots, M$ and to divide the users equally between the facilities. By rounding $s_j^\star$ to one, we obtain the solution that Hillier and Lieberman (2001) identified as optimal. We solved this problem with $n = 16$ as well and obtained the same solution. The ability of the formulation $(P2)$ to identify the optimal solution even in this small and lightly loaded system is consistent with the numerical experiments in Borst et al. (2002). We expect our formulation to be even more reliable in finding the optimal solution for systems with higher offered load.

The instances we solved for this example had 3 binary variables, 81 to 2,304 continuous variables, and 109 to 3,703 constraints. Solution times on the NEOS Server increased approximately quadratically with the number of continuous variables, ranging from 0.04 to 144.70 seconds. As stated in Section 3, when it is necessary to require that each customer location is allocated entirely to a single open facility (i.e., $\boldsymbol{Y}$ is integer), the integrality property of our optimization problems allows for substantial reduction in computational effort. Figure 4 shows how the solution time for $(P1)$ with $P = 2$ and $C_f = 0$ changes with the instance size n when replacing constraint (11) with $Y_{ij} \in \{0, 1\},\ i = 1, \ldots, C,\ j = 1, \ldots, M$. Solution times increased exponentially when the demand allocation variables $\boldsymbol{Y}$ are required to be integers. NEOS was unable to solve problem instances $n = 15$ and 16 with $\boldsymbol{Y}$ integer.

Another strength of our optimization problems is that they can be applied when the simplifying assumptions of uniform user demand, equal allocation of users to facilities, and equal capacity for every facility are not realistic. We illustrate this with the second example we consider.

5.2 Locating walk-in health clinics

Marianov and Serra (1998) describe an example where walk-in health clinics are to be located on a 30-node network. Table 1 provides the coordinates (in miles) and population of each node. Figure 5 illustrates the topology of the network. Travel distances between the nodes of the network are calculated using the l_2 (Euclidean) norm and converted to travel times using a constant travel speed of 20 miles per hour.

All nodes in the network are potential clinic locations. The servers are physicians and each open clinic can have one or more physicians. Each physician provides service following an exponential distribution with rate $\mu = 3$ patients per hour.

We scaled the population values used by Marianov and Serra (1998) to be representative of a small North American city with a total population of about 100,000 and a population density of

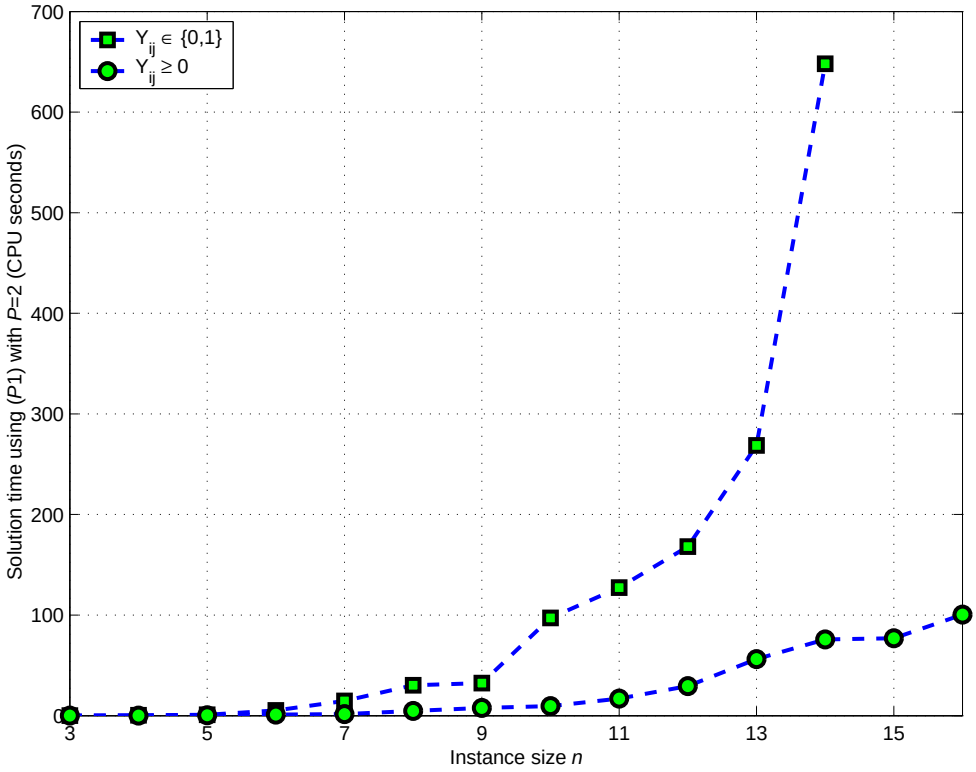

Figure 4: CPU seconds for the tool crib example using $(P1)$ with $P = 2$ as a function of instance size n.

Table 1: Coordinates and populations for the 30-node network.

i	x	y	p_i	i	x	y	p_i
1	3.2	3.1	12,980	16	3.0	5.1	2,011
2	2.9	3.2	11,335	17	1.9	4.7	1,828
3	2.7	3.6	10,238	18	1.7	3.3	1,828
4	2.9	2.9	7,130	19	2.2	4.0	1,645
5	3.2	2.9	6,399	20	2.5	1.4	1,645
6	2.6	2.5	3,839	21	2.9	1.2	1,645
7	2.4	3.3	3,656	22	2.4	4.8	1,463
8	3.0	3.5	3,473	23	1.7	4.2	1,463
9	2.9	2.7	3,108	24	6.0	2.6	1,463
10	2.9	2.1	3,108	25	1.9	2.1	1,463
11	3.3	2.8	2,925	26	1.0	3.2	1,280
12	1.7	5.3	2,742	27	3.4	5.6	1,097
13	3.4	3.0	2,559	28	1.2	4.7	1,097
14	2.5	6.0	2,194	29	1.9	3.8	1,097
15	2.1	2.8	2,194	30	2.7	4.1	1,097

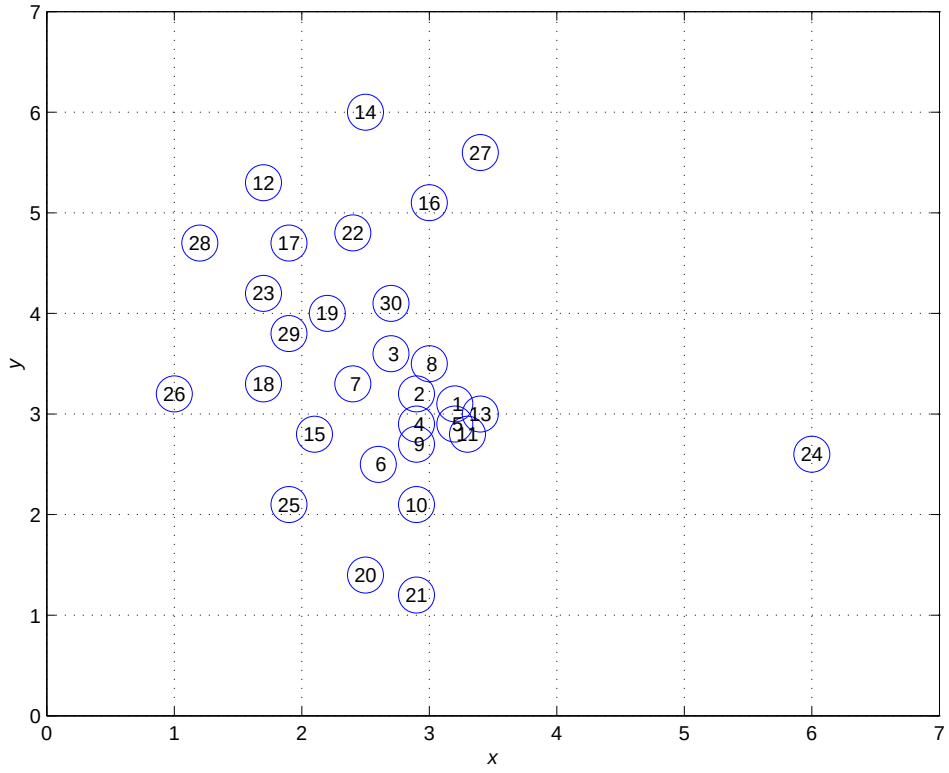

Figure 5: 30-node network for the walk-in health clinics example.

approximately 4,000 per square mile. We assumed an average frequency of health clinic visits of 4 per person per year and that the health clinics are open 8 hours a day, 250 days a year, in order to estimate the hourly arrival rates from each node.

Marianov and Serra (1998) solve a (mixed-integer linear) probabilistic maximal covering location-allocation problem with a constraint on the probability that either system time or queue length does not exceed a specified threshold. They assumed that each clinic serves only patients located within a certain distance (as is typical with covering-type problems). In their model, the number of servers at each facility must be specified ahead of time. In contrast, we solve the location-allocation problem as well as choose the number of servers for each open clinic simultaneously.

First, we solved this problem using ($P2$) with maximum number of facilities $P = 10$, fixed cost $C_f = 0$, travel cost $C_t = 200$ (accounting for round trip travel), and waiting cost $C_q = 100$. We varied the server cost C_c from $30/\mu$ to $240/\mu$. Recall that $C_c\mu$ can be interpreted as the cost per time unit per server. Therefore, we varied the cost of a server's time from 30% to 240% of the cost of a patient's time. These problem instances have 30 binary variables, 900 continuous variables, and 931 constraints. The solution times ranged from 3.64 to 4.11 seconds on the NEOS Server. In contrast, Marianov and Serra (1998) reported solution times of 10 to 40 minutes for their model using CPLEX 3.0 on a cluster of eight DEC 3000-700 AXP workstations.

When the server cost C_c increases, the optimal number of open facilities for problem ($P2$) will stay the same or decrease, all else being equal. Figure 6 confirms this. Indeed, we found that for $C_c = 240/\mu$, it was optimal to have a single health clinic located at node 2, with $s_2^\star = 71.50$ servers. We then evaluated an exact version of the objective function for problem ($P2$) (one that uses an exact expression for the $M/M/s_j$ queue) to choose between rounding up or down. The resulting feasible solution has 72 servers and provides an upper bound on the optimum system-wide cost. Also, when the server cost C_c increases, the optimal objective function value for problem ($P2$) must increase, all else being equal. The results in Figure 7 confirm that this happens as well. Moreover, Figure 7 also shows the lower bounds on the optimum costs. Note that the gap between the optimum cost and the lower bound decreases as the number of facilities being opened decreases, and disappears whenever a single facility is opened.

Figure 8 illustrates the optimal allocation of nodes to facilities when with $C_f = 0$, $C_t = 200$, $C_q = 100$, and $C_c = 105/\mu$. The open facilities are shown as shaded geometric shapes and the other nodes are identified with the same unshaded geometric shape as the facility to which they are assigned. The districts defined by the allocation variables are contiguous and appear plausible in the sense of approximating user choice. In order to validate the accuracy of the approximations

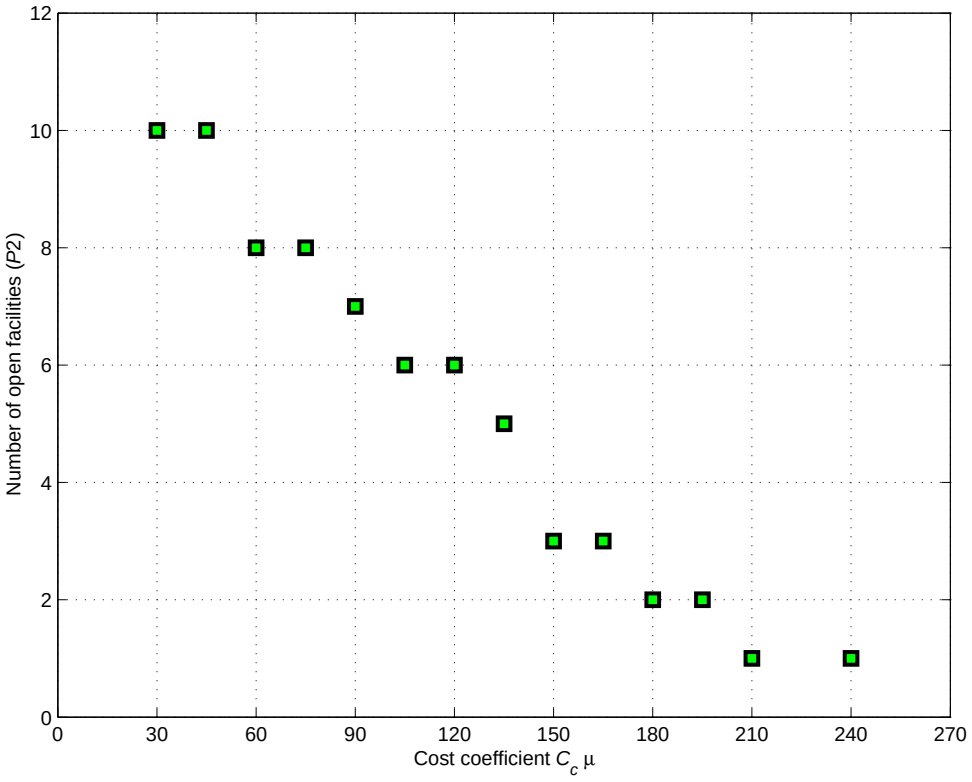

Figure 6: Optimal number of open facilities for the walk-in health clinics example using ($P2$) with $C_f = 0$, $C_t = 200$, and $C_q = 100$ as a function of $C_c\mu$.

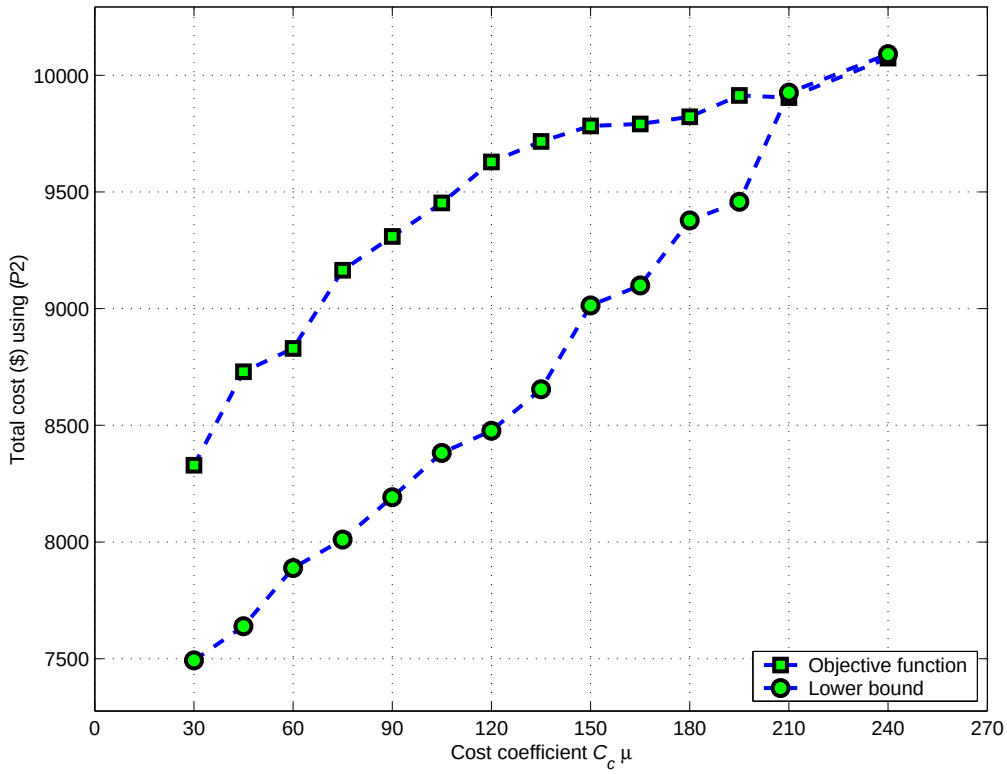

Figure 7: Optimal objective function value and lower bound for the walk-in health clinics example using $(P2)$ with $C_f = 0$, $C_t = 200$, and $C_q = 100$ as a function of $C_c\mu$.

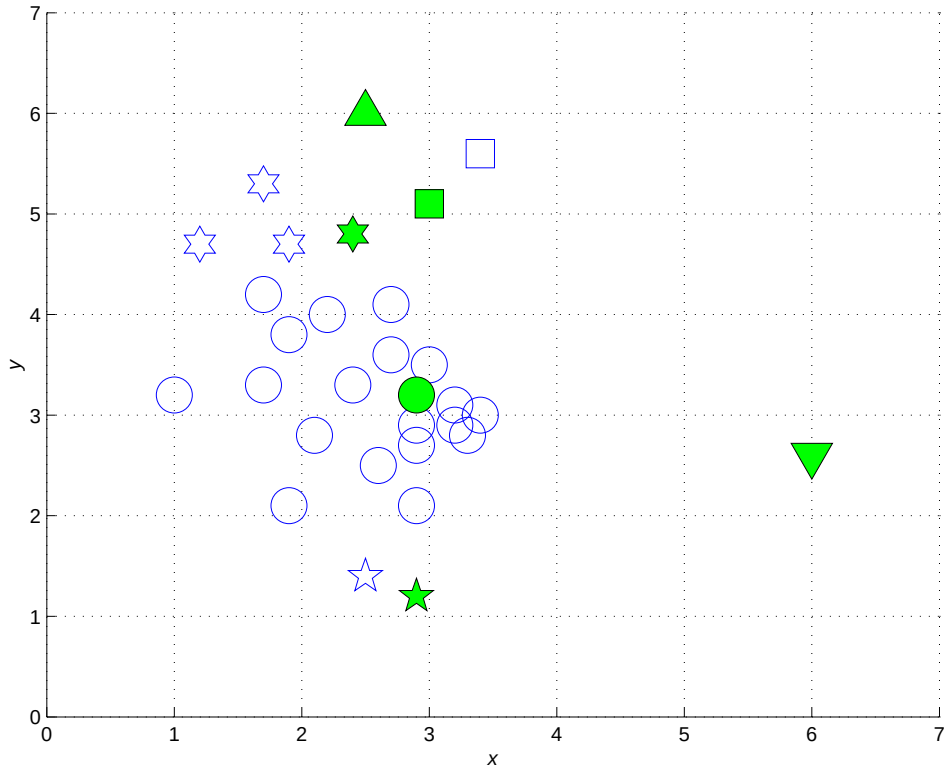

Figure 8: Optimal service demand allocation for the walk-in health clinics example with $C_f = 0$, $C_t = 200$, $C_q = 100$, and $C_c = 105/\mu$. The open facilities are shown as shaded geometric shapes. Every other unshaded node is shown with the geometric shape of the facility it is assigned to.

when the number of servers are relatively small, we evaluated an exact version of the objective function for problem $(P2)$ and performed an exhaustive search on the number of servers for the optimal allocation of nodes to facilities given in Figure 8 with $C_f = 0$, $C_t = 200$, $C_q = 100$, and $C_c = 105/\mu$. Table 2 confirms that the approximations are remarkably accurate even when the number of servers at a facility is relatively small. Note that node 2, the open facility for the largest district in Figure 8 (shaded circle), is the one chosen as the optimal single facility solution for $C_f = 0$, $C_t = 200$, $C_q = 100$, and $C_c = 240/\mu$.

Second, we solved this problem using $(P2)$ with maximum number of facilities $P = 10$, travel cost $C_t = 200$, waiting cost $C_q = 100$, and server cost $C_c = 45/\mu$. We now varied the fixed cost C_f from 0 to 270. As well, these problem instances have 30 binary variables, 900 continuous variables, and 931 constraints. The solution times ranged from 3.54 to 4.08 seconds on the NEOS Server.

When the fixed cost C_f increases, the optimal number of open facilities for problem $(P2)$ will stay the same or decrease, all else being equal. Figure 9 confirms this. Indeed, we found that for $C_t = 200$, $C_q = 100$, $C_c = 45/\mu$, and $C_f = 270$, it was optimal to have a single health clinic located

Table 2: Validation of approximations for the optimal allocation of nodes to facilities given in Figure 8 with $C_f = 0$, $C_t = 200$, $C_q = 100$, and $C_c = 105/\mu$.

i	Λ_j	$r_j = \Lambda_j/\mu_j$	Approximate $s_j^\star$	Exact $s_j^\star$
2	165.63	55.21	61.35	61
14	4.39	1.46	2.46	3
16	6.21	2.07	3.26	3
21	6.58	2.19	3.42	4
22	14.26	4.75	6.56	7
24	2.93	0.98	1.79	2

at node 2, with $s_2^\star = 75.75$ servers. We then evaluated an exact version of the objective function for problem $(P2)$ to choose between rounding up or down. The resulting feasible solution has 76 servers. Also, when the fixed cost C_f increases, all else being equal, the optimal objective function value for problem $(P2)$ must increase until it is more beneficial to open less facilities, resulting in (a nonlinear) 'saw-tooth' profile. The results in Figure 10 confirm that this happens as well.

6. Conclusions and Future Research

Our work was motivated by the observation that locations of service facilities determine not only the distance a user needs to travel to the closest facility but also, in conjunction with facility capacities and customer choices, how long users will wait for service. Walk-in health clinics, motor vehicle inspection stations, automobile emissions testing stations, and internal service systems are motivating examples of facilities where users may experience waiting times of similar magnitude as travel times.

It seems plausible that users will have some information about the relative level of congestion at different facilities and will therefore decide which facility to use based both on travel time and expected waiting time. Indeed, empirical research has supported this hypothesis. Given that both expected waiting times and travel times are important, it seems appropriate to incorporate both in the objective function and allow average waiting times to vary between facilities, if this reduces the overall time spent by all users either traveling or waiting for service. This is in contrast to models that include service-level constraints that tend to equalize waiting times between facilities.

Our optimization problems are tractable mixed-integer nonlinear programs for the socially optimal location of facilities with fixed servers. We account for stochastic demand originating at the nodes of a network and random service times at facilities. In contrast to most previous research, we include both customer travel and delay costs in the objective function and allow the models

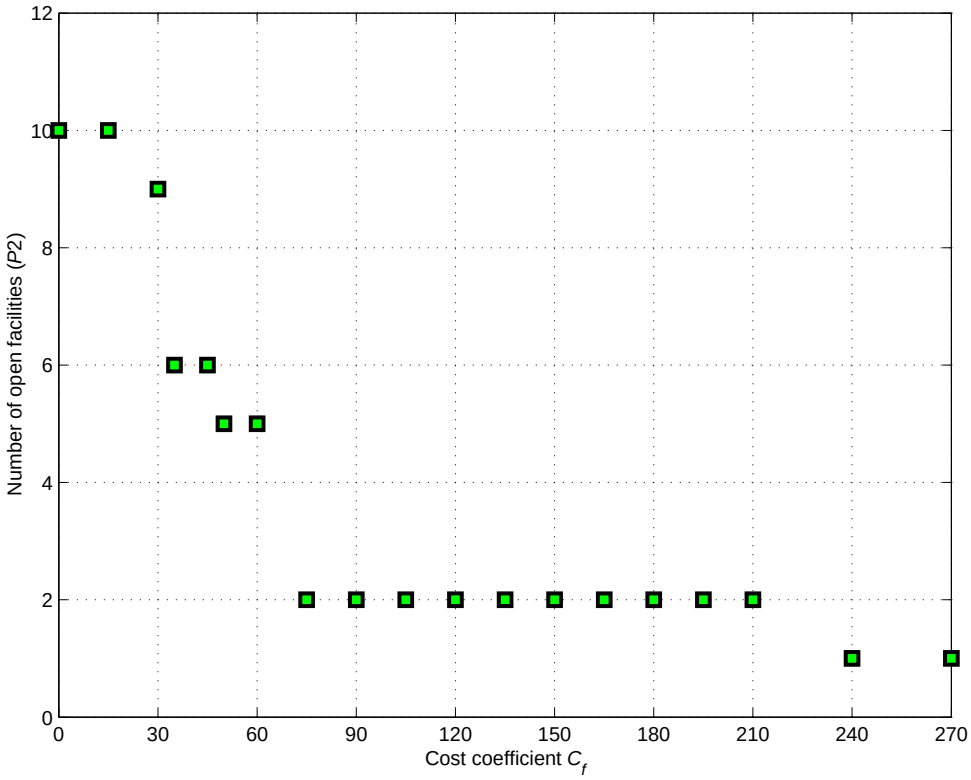

Figure 9: Optimal number of open facilities for the walk-in health clinics example using ($P2$) with $C_t = 200$, $C_q = 100$, and $C_c = 45/\mu$ as a function of C_f.

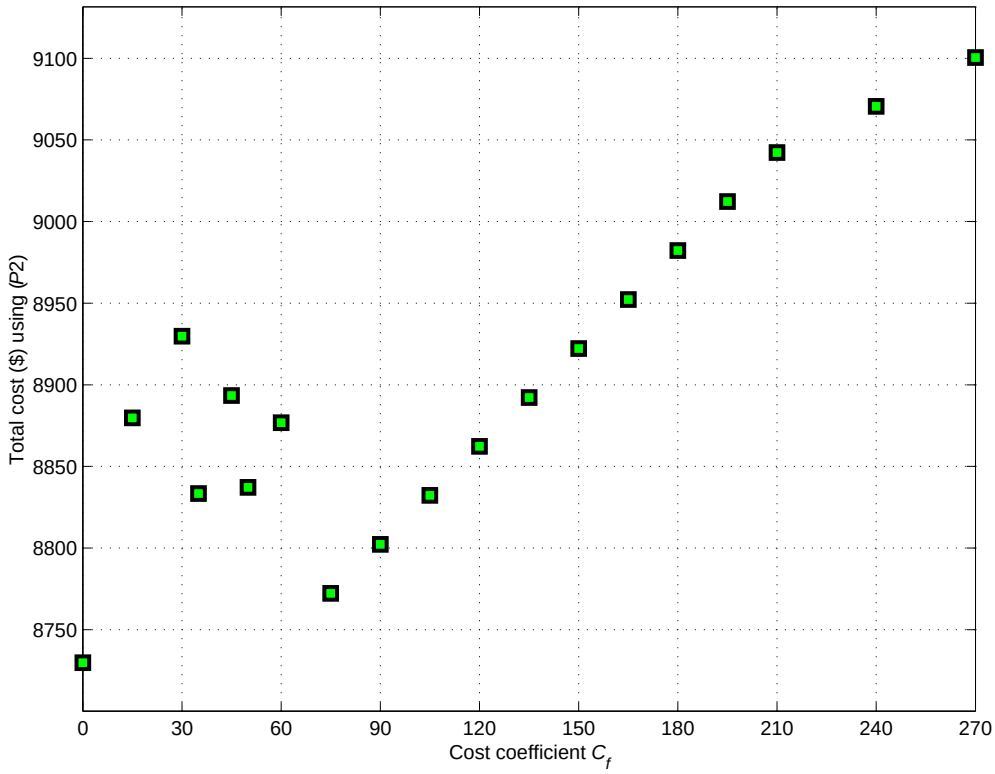

Figure 10: Optimal objective function value for the walk-in health clinics example using ($P2$) with $C_t = 200$, $C_q = 100$, and $C_c = 45/\mu$ as a function of C_f.

to solve the location-allocation problem as well as choose service capacities for each open facility simultaneously. We considered two capacity choice scenarios. The two scenarios result in structurally identical optimization problems that share useful properties such as pseudolinearity of the objective function and the existence of a global minimum that is integer in the demand allocation variables. Furthermore, the capacity for an open facility, which is viewed as a queueing system, is relatively robust to the values of the cost coefficients; that is, it is not necessary to estimate the coefficients precisely to obtain near-optimal results.

Several useful extensions to our optimization problems could be explored. In data rich situations, it may be appropriate to generalize the cost structure and use functions that better approximate the empirical evidence, yet retain the tractability of the problems. It is possible, for instance, to consider convex and strictly increasing functions for the cost of providing service, strictly increasing functions for the cost of waiting, and concave functions for the cost of traveling to reflect economies of scale, and still solve the problems to optimality with the resulting allocation variables being integer. Moreover, in situations where it is not practical to use tolls to induce a social optimum, it would be useful to explicitly model user choice and find equilibrium solutions.

The design of a system of geographically separate telephone call centers that are managed as one large 'virtual call center' is one setting where our scenario 2 model could be applicable. When call routing in such a system is determined centrally, the directed choice assumption is appropriate. 'Travel costs' would correspond to phone charges from a particular customer location to a particular call center and waiting costs would include telephone charges for customers waiting on hold, at least if customers call a toll-free number.

Finally, it might be possible to relate our optimization problems to the integration of location and inventory decisions. Open facilities could be viewed as production facilities that produce in either a make-to-order or make-to-stock fashion. Integrated models would simultaneously determine capacities, safety stock levels, and locations for production facilities. We are currently working on these integrated models.

Appendix

Proof of Proposition 1: The objective function $z_{M/M/1}(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{S}, \boldsymbol{U})$ is separable in the s_j's. The cost contribution of the number of servers at open facility j is

$$z_{s_j} = C_q L(\Lambda_j, \mu_j, s_j) + C_c s_j \mu_j. \tag{A1}$$

The last term is independent of s_j for a fixed $\rho_j = \Lambda_j/(s_j\mu_j)$. Stidham Jr. (1970) shows that for a given value of ρ_j, increasing s_j increases $L(\Lambda_j, \mu_j, s_j)$, so $s_j = 1$ minimizes z_{s_j}. □

Proof of Proposition 2: Given a service demand allocation matrix $\boldsymbol{Y}$ and a location vector $\boldsymbol{X}$, $z_{M/M/1}(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{S^\star}, \boldsymbol{U})$ is separable in the μ_j's. Thus the optimal service rate $\mu_j^\star$ at each facility j can be obtained by minimizing

$$z_{\mu_j} = C_q \frac{\Lambda_j}{\mu_j - \Lambda_j} + C_c\mu_j. \tag{A2}$$

The cost contribution z_{μ_j} can be verified (by inspecting its second derivative with respect to μ_j) to be a convex function of μ_j. Solving $\partial z_{\mu_j}/\partial\mu_j = 0$ yields $\mu_j^\star = \Lambda_j + \sqrt{C_q/C_c}\sqrt{\Lambda_j}$. □

Locating Tool Cribs Example: Consider an arbitrary point with coordinates (x, y) in the square consisting of triangles B and C in Figure 1. Denoting the coordinates of locations 1 and 3 by (x_1, y_1) and (x_3, y_3), the arbitrary point is equidistant from the two locations if the following equality holds:

$$|x - x_1| + |y - y_1| = |x - x_3| + |y - y_3|. \tag{A3}$$

For this particular plant layout and potential tool crib locations, we have that $x \geq x_3 > x_1$ and $y \leq y_1 < y_3$, and therefore $|x - x_i| = x - x_i$ and $|y - y_i| = y_i - y$ for $i = 1, 3$. Substituting in (A3) and simplifying, we get $y_1 - x_1 = y_3 - x_3$. In this example, $x_1 = y_1$ and $x_3 = y_3$, so (A3) holds.

Assume now that all users in region A are assigned to location 3 and all users in regions B, C, and D are assigned to location 1. Let T be the one-way travel time for a mechanic whose location is uniformly distributed over the factory area. Hillier and Lieberman (2001) calculate $\mathrm{E}[T] = 0.0139$ if regions A and B are assigned to location 3 and regions C and D to location 1. Since all of region B is equidistant from locations 1 and 3, allocating it to location 1 instead of location 3 does not change $\mathrm{E}[T]$. The system wide travel cost is therefore $\$96 \times \sum_{i=1}^{C} \lambda_i \mathrm{E}[T] = \$160$.

The probability of a demand for service coming from region A is the ratio of the area of A and the entire plant, or 11/24. The arrival rates to the two locations are therefore $\Lambda_3 = (11/24) \times 120 = 55$ per hour and $\Lambda_1 = 120 - \Lambda_3 = 65$ per hour, and the second term in the objective function of $(P1)$ is

$$\$2\sqrt{C_qC_c}\left(\sqrt{\Lambda_1} + \sqrt{\Lambda_3}\right) = \$2\sqrt{48 \times 20/120}\left(\sqrt{65} + \sqrt{55}\right) = \$87.56. \tag{A4}$$

Adding a constant term \$20 (the first term in (17)) and the facility cost $2 \times \$16$, we get a total of \$299.56. This is the benchmark value for $P = 2$ shown in Figure 3. The benchmark value of \$290.53 for $P = 3$ is calculated using the same procedure. □

References

Amiri, A. 1998. The design of service systems with queueing time cost, workload capacities and backup service. *Eur. J. Oper. Res.* **104** 201–217.

Bazaraa, M.S., H.D. Sherali, C.M. Shetty. 1993. *Nonlinear Programming: Theory and Algorithms.* John Wiley & Sons, New York, NY.

Beckmann, M., McGuire, C. B., Winsten, C. B. 1956. *Studies in the Economics of Transportation.* Yale University Press, New Haven, CT.

Berman, O., D. Krass. 2002. Facility location problems with stochastic demands and congestion. Z. Dresner, H.W. Hamacher, eds. *Facility Location: Applications and Theory.* Springer-Verlag, Berlin. 329–371.

Borst, S., A. Mandelbaum, M.I. Reiman. 2002. Dimensioning large call centers. Working paper. Technion, Haifa, Israel.

Czyzyk, J., M. Mesnier, J. Moré. 1998. The NEOS Server. *IEEE Comput. Sci. Eng.* **5** 68–75.

Dyer, M.E., L.G. Proll. 1977. On the validity of marginal analysis for allocating servers in $M/M/c$ queues. *Manage. Sci.* **23** 1019–1022.

Fletcher, R., S. Leyffer. 1994. Solving mixed integer nonlinear programs by outer approximation. *Math. Program.* **66** 327–349.

Fulkerson, D.R., O.A. Gross. 1965. Incidence matrices and interval graphs. *Pac. J. Math.* **15** 835–855.

Gross, D., C.M. Harris. 1998. *Fundamentals of Queueing Theory.* John Wiley & Sons, New York, NY.

Grossman Jr., T.A., M.L. Brandeau. 2002. Optimal pricing for service facilities with self-optimizing customers. *Eur. J. Oper. Res.* **141** 39–57.

Halfin, S., W. Whitt. 1981. Heavy-traffic limits for queues with many exponential servers. *Oper. Res.* **29** 567–588.

Heinhold, M. 1975. Ein operations-research-ansatz zur bestimmung von standorten und einzugsbereichen von krankenversorgungsbetrieben. *Z. Oper. Res.* **19** B103–B118.

Heinhold, M. 1978. An operational research approach to allocation of clients to a certain class of service institutions. *J. Oper. Res. Soc.* **29** 273–276.

Hillier, F.S., G.J. Lieberman. 2001. *Introduction to Operations Research.* McGraw-Hill, New York, NY.

Jan, S., G. Mooney, M. Ryan, K. Bruggemann, K, Alexander. 2000. The use of conjoint analysis to elicit community preferences in public health research: A case study of hospital services in South Australia. *Aust. N.Z. J. Publ. Heal.* **24** 64–70.

Marianov, V., M. Rios. 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., D. Serra. 1998. Probabilistic maximal covering location-allocation for congested systems. *J. Regional Sci.* **38** 401–424.

Marianov, V., D. Serra. 2001. Hierarchical location-allocation models for congested systems. *Eur. J. Oper. Res.* **135** 196–209.

Martos, B. 1975. *Nonlinear Programming: Theory and Methods.* North-Holland, Amsterdam.

Newman, R. G. 1984. A Conjoint Analysis in Outpatient Clinic Preferences. *J. Health Care Market.* **4** 41–49.

Sheffi, Y. 1985. *Urban Transportation Networks: Equilibrium Analysis with Mathematical Programming Methods.* Prentice-Hall, Englewood Cliffs, NJ.

Stidham Jr., S. 1970. On the optimality of single-server queuing systems. *Oper. Res.* **18** 708–732.

Wang, Q., R. Batta, C.M. Rump. 2001. A new class of facility location models with congestion. Working paper. University at Buffalo, Buffalo, NY.

Westerlund, T., R. Pörn. 2002. Solving pseudo-convex mixed integer optimization problems by cutting plane techniques. *Optim. Eng.* **3** 253–280.