Maximal Accessibility Network Design in the Public Sector ©

Robert Aboolian
College of Business Administration, California State University San Marcos, San Marcos, California 92096, USA
raboolia@csusm.edu

Oded Berman
Rotman School of Management, University of Toronto, 105 St. George Street, Toronto, Ontario, Canada M5S 3E6
berman@rotman.utoronto.ca

Vedat Verter
Desautels Faculty of Management, McGill University, Montreal, Quebec H3A 1G5
vedat.verter@mcgill.ca

December 16, 2014

This paper focuses on designing facility networks in the public sector so as to maximize the number of people benefiting from their services. We develop an analytical framework for the maximal accessibility network design problem that involves determining the optimal number, locations and capacities of a network of public sector facilities. We assume that the time spent for receiving the service from a facility is a good proxy for its accessibility. We provide a generic model that incorporates both the congestion at the facilities and the customer-choice environment that underlies most of the services offered by the public sector. We devise an ϵ-optimal algorithm for the arising nonlinear integer program. The proposed algorithm performs well in tackling fairly large problem instances. Through a realistic example based on the hospital network of Toronto, Canada, we demonstrate the model's capability in providing policy insights.

Key words: Pubic Sector, Accessibility, Service System Design, Elastic Demand, Congestion, Nonlinear Integer Program.

1. Introduction

The modern governments' mandate is to maximize the societal benefit by acting as agents of the public, in contrast with the private firms' mission to maximize its shareholder benefits. "The *public sector* includes all government controlled entities such as ministries, departments, funds, organizations,

and business enterprises, which political authorities at all levels use to implement their social and economic policies" (Statistics Canada, 2008). In some sectors, such as energy and mining, government enterprises compete with the private sector, whereas in others such as, the construction of mega-hospitals, transportation and communication infra-structure, public-private partnerships are becoming more common. Nevertheless, in all these endeavours, the governments' overarching objective is to maximize the accessibility of their services to the general public, while maintaining their economic viability.

The healthcare sector is a good example of industries where the government-controlled entities and private enterprises co-exist. For example, while private mammography centres offer screening services for profit, state/provincial governments provide subsidized breast cancer screening programs (in some cases, by recruiting some of the screening centres to form their network) so as to protect the largest number of women within the designated age group. Another example is long-term care where the establishment of a network of long-term care facilities that is capable of caring for all the elderly in need is typically within government mandate, although private facilities do exist. The education sector exhibits very similar characteristics. Notwithstanding the abundance of private universities, colleges, schools and daycare facilities, modern governments are obliged to provide publicly-funded education opportunities at all levels, so that education is accessible to everyone regardless of their household income.

The policy makers and regulators constantly face the problem of (re-)designing a public service so as to maximize the number of people who benefit from the program. This is usually a better option than simply increasing the budget allocated to that public service, which often inherits the systemic obstacles that mitigate performance improvement. Empirical evidence suggests that improved accessibility would lead to increased participation in public services. For example, Zimmerman (1997) found out that the convenience of access to a facility is a very important factor in the customers' decision to have prostate cancer screening. We define the maximal accessibility network design problem as follows: Determine the optimal number, locations and capacities of a network of facilities so as to maximize the number of people who can benefit from the service being provided. In this context, the governmental budget allocation manifests itself as a limit on the total amount of service capacity that can be distributed across the jurisdiction. In this paper, we view each service facility as a single server and aim at optimizing its overall capacity. Note that there are alternative ways of configuring the resource levels to provide the same capacity at a facility. This tactical capacity planning problem, however, is out of the scope of our work.

Unless participation is mandated by the government (e.g., separation of the recyclables and household waste), each individual is free to choose whether or not to use the services of a publicly-funded

program. The people who participate often patronize the facility with highest accessibility. In representing this *customer-choice* environment, we assume that the time spent for receiving the service from a facility is a good proxy for its accessibility. This is certainly a simplification, since we ignore other potentially important factors such as the variation among facilities pertaining to the quality of service. Assuming such differences are insignificant, the travel time to the facility and the waiting (plus service) time at the facility comprises that site's attractiveness for an individual. An important feature of this problem is the *congestion* at the facilities caused by the uncertainty in demand and the limited capacity. Note that the expected total time customers spend in the system (waiting and receiving service) depends on the level of congestion, and hence each individual's facility choice is indeed affected by the preferences of the other members of the public. Consequently, the total demand at a facility is elastic with respect to its accessibility to the people who reside in its vicinity.

The contributions of this paper are two-fold: First, we formalize and analyze the basic network design problem pertaining to the public sector. Although there is some literature that focuses on some of the features of the maximal accessibility network design problem (which will be discussed in the next section), we are not aware of a study that includes all the features examined in this paper. Second, we devise an ϵ-optimal solution method for the problem. Given the nonlinear nature of the arising models for this problem family, an overwhelming majority of the relevant literature is confined to heuristic solution approaches.

We assume that a set of potential operational facilities that can provide the new service (mammography service) are available (e..g. hospitals). Therefore the location decisions we consider is to select those facilities that will offer the new service.

There are two common ways to model flexible capacity of a queuing system. One way which takes the micro view is to assume a multiple parallel servers each with service rate 1 and the control of the system is the number of servers $N = \mu$. The other way which takes the macro view and which we follow in the paper is to assume a single server in each facility with a flexible service rate μ which is the control variable. The discrete-capacity model may be more suitable to a single facility (for example a pump at a gas station). The continuous model is more appropriate for complex system such as mammography service in a hospital, where capacity is not only the mammography machines but also doctors, nurses and examining rooms.

Zhang et al. (2010) and Aboolian et al. (2012) constitute the most relevant papers to our work. Here, we point out the differentiating characteristics of this paper so as to better highlight our contributions. Also aiming at maximizing the total number of people who receive service in a customer-choice environment, Zhang et al. (2010) adopts the micro viewpoint of optimizing the number of servers

(in particular, the number of mammography machines in each screening center) at each facility. In contrast, we view the macro perspective i.e., using a *service rate*. In Zheng et al. (2010), the problem is formulated as a bi-level problem where the allocation of customers to facilities is determined in the lower level and the location of facilities and their capacity level are determined in the upper level. Bi-level programming approach could be shown to be efficient when the capacity alternatives are relatively small, but as these alternatives increases, given the exponential increase in combinations of capacity levels even for a given facility location set in the upper level problem, this approach gradually loses its efficiency. In our problem, where we consider the capacity as a continuous variable with unlimited alternatives to choose from, bi-level programming becomes almost impossible to use. Therefore, in our paper the problem is solved as an exact (single level) mixed integer problem (MIP). While Aboolian et al. (2012) also optimize the service rate at each facility in one version of the problems studied, their aim is to design a service network so as to maximize the total profit assuming that the customer-facility allocations can be made by a central planner. The problem in Aboolian et al. (2012) is also formulated as a bi-level problem.

The remainder of the paper is organized as follows. Section 2 provides an overview of the most relevant literature and positions this paper. Section 3 presents the model we propose for the problem, whereas an ϵ-optimal procedure is highlighted in Section 4. Section 5 reports on the computational performance of the solution algorithm and presents a realistic illustrative example. Our concluding remarks are in Section 6.

2. Overview of the Literature

Two threads of research are immediately relevant to the work presented in this paper: (i) maximal covering location problem, and (ii) design of service facility networks with congestion. Schilling et al. (1993) provide a comprehensive review of the early work on the covering problems in facility location, and a very recent review can be found in Farahani et al. (2012). Concerning the incorporation of congestion in service system design, Berman and Krass (2002a) provides an early survey of the literature.

In location theory, each customer within a predefined distance (or time) of a facility is considered *covered* by that facility. The *maximal covering location problem* involves determining the optimal sites for a predetermined number of facilities so as to maximize the total number of people covered. It is important that everyone within the threshold distance is considered covered, and hence the reduction in accessibility, as the distance to the facility increases, is not represented by the concept of coverage.

As an implementation of partial coverage, Verter and Lapierre (2002) used a linear decay function in modeling participation in preventive healthcare programs. Berman and Krass (2002b) presented the gradual coverage decay function using a step function. Berman et al. (2003) presented the gradual coverage decay model with two pre-specified threshold distances, where a customer is considered fully covered within the first threshold, partially covered between the two thresholds and "not covered" otherwise.

Note that these three models represent the customer's access to the facilities, and not necessarily to the service being offered, since they do not incorporate the congestion at the facilities. Zhang et al. (2009) extended Verter and Lapierre (2002) by representing the accessibility of a service as the sum of the travel and waiting times, the latter due to congestion caused by the pre-specified service capacities at each site.

In general, the incorporation of congestion at the facilities under stochastic demand has been studied within the context of the *service network design problem*. The problem aims at determining the optimal configuration of the service facilities i.e., their number, locations as well as capacities, taking into account the trade off between the total cost of offering the service and the service quality. There are a multiplicity of measures for service quality, including the average waiting time per customer and the average number of customers waiting for service. There are two common ways of incorporating service quality in a service system design model: (i) Including a measure of service quality as an additional cost term in the objective function and minimizing the total overall cost (Elhedhli, 2006; Berman and Drezner, 2006; Aboolian et al., 2008; Castillo et al., 2009), and (ii) Including an additional constraint in the model to ensure that service quality remains above a certain threshold (Marianov and Serra, 1998; Marianov and Serra, 2002; Berman et al., 2006; Baron et al., 2008). In this paper, we adopt the second approach, using the objective of maximizing accessibility. All the service system design models cited above assume inelastic demand, whereas we make an explicit attempt to incorporate demand elasticity in this paper.

In a related paper, Marianov (2003) presents a model to site a predetermined number of multi-server facilities so as to maximize the demand served under elastic demand conditions. This paper differs from our approach in three ways: (i) the number of servers at each facility is given, (ii) the customer allocations are done by a central decision maker, and (iii) the congestion at the facilities is represented by the number of customers. In contrast, the customer choice model we propose in the next section optimizes the capacity and captures demand elasticity with respect to the total time spent by the customer in receiving the service. Marianov et al. (2005) (by taking the micro approach discussed in Section 1 to model flexible capacity) extend the earlier model to also determine the number of servers

allocated to each facility and represent the customers sensitivity to the waiting time at a facility. Recall that our model is focused on optimizing the service rate at each facility (the macro approach). In addition they used heuristic concentration to solve small-scale hypothetical problem instances.

Finally, it is important to note that other measures of accessibility have been used. Two such measures are the floating catchment area (Luo and Qi 2009) and a modified catchment method that combines the floating catchment area and the gravity model (Gu et al. 2010). These two measures are mainly based on travel times and do not consider the service and waiting times.

3. Maximal Accessibility Network Design Problem

We consider a finite set $M = \{1, \ldots, m\}$ of potential facility locations, a finite set $N = \{1, \ldots, n\}$ of population zones, and a travel time metric t_{ij} for i, $j \in M \cup N$. Without loss of generality, we assume $M \subset N$ and N represents nodes of a network, in which case t_{ij} is the shortest travel time between i and j. The facilities to be located in M provide a pre-specified set of public services.

Let $S \subset M$ be a set of facility locations (we call the facility located at $j \in M$, facility j). We assume that the people at $i \in N$ generate a stream of Poisson demands with homogeneous rate $\lambda_i \geq 0$, where λ_i, the demand rate of node i, is determined as follows. Let $\lambda_i^{max} \geq 0$ denote the maximum demand rate that can be generated by node i — this can be thought of as the total number of people at i who could potentially be interested in the services offered by the facilities in S. Suppose that y_{ij} is the fraction of the population of node $i \in N$ who request service from facility $j \in S$. Then $\lambda_{ij} = \lambda_i^{\max} y_{ij}$ is the actual demand rate from i seen by facility j, such that

$$\lambda_i = \sum_{j \in S} \lambda_{ij} = \lambda_i^{\max} \sum_{j \in S} y_{ij}, \tag{1}$$

and $\Lambda_j -$ the total demand that facility j faces is given by

$$\Lambda_j = \sum_{i \in N} \lambda_{ij} = \sum_{i \in N} \lambda_i^{\max} y_{ij}. \tag{2}$$

We assume that the service at each facility j is exponentially distributed with service rate $\mu_j \geq 0$ and the system is an $M/M/1$ queueing system (μ_j's are decision variables in our model). For the $M/M/1$ system, W_j, the expected total time customers spend in facility j (which includes the expected waiting and service times), can be computed as follows:

$$W_j = W(\Lambda_j, \mu_j) = \frac{1}{\mu_j - \Lambda_j}, \quad \Lambda_j < \mu_j. \tag{3}$$

Define $\tau_{ij} = t_{ij} + W_j$ to be the expected total time that customers from node i that receive service at

facility j spend (from the time that travel starts until the time that the customer leaves the facility). Let f_i be the fraction of λ_i^{max} that is realized. Then

$$f_i = \sum_{j \in S} y_{ij}, \tag{4}$$

and

$$\lambda_i = f_i \lambda_i^{\max}. \tag{5}$$

As we assume that customers select the facility with the shortest expected total time, we denote by $\hat{T}_i$ the shortest total time incurred by customers at node i. We also assume that this will happen only if the shortest expected total time does not exceed a certain threshold denoted by η_i. Let $T_i = \min\{\hat{T}_i, \eta_i\}$.

We consider f_i to be elastic with respect to T_i, such that f_i is a function of T_i: $\mathcal{F}(T_i) \in [0, f_i^{\max}]$, where $f_i^{\max} \leq 1$ is the maximum participation fraction of the demand rate from node i. Although $\mathcal{F}(T_i)$ can be generalized to any decreasing functions in $[0, f_i^{\max}]$, for simplicity, we consider the following bounded linear function (considered also in Zhang et al. (2010)):

$$f_i = \mathcal{F}(T_i) = (f_i^{\max} - \alpha T_i) \text{ for } i \in N, \tag{6}$$

where α represents the sensitivity level of demand to T_i. We note that $T_i \leq \eta_i = \frac{f_i^{\max}}{\alpha}$ and we can write T_i as a function of f: $T_i(f_i) \in \left[0, \frac{f_i^{\max}}{\alpha}\right]$ such that

$$T_i(f_i) = \frac{(f_i^{\max} - f_i)}{\alpha} \text{ for } i \in N, \tag{7}$$

In fact, for a given f_i, $T_i(\ f_i)$ is the total time threshold that $100 f_i\%$ of customers at node i are willing to spend for receiving service and transit time, while the actual total time that customers from node i spend at facility $j \in S$ and in transit is τ_{ij}.

Remark 1 *We assume that the decision maker is risk neutral like in most queuing models and therefore there should be no problem to using the expected time customers stay in the system for the elasticity and the threshold.*

Given the set of facility locations S and service rates μ_j, $j \in S$, the customers when choosing which facility to patronize would like to minimize their expected total time. The eventual choice of facilities is a user-equilibrium problem, where at equilibrium, customers do not want to change their choices.

As in Zhang et al. (2010), this equilibrium condition can be stated as: given S and μ_j, $j \in S$,

$$\tau_{ij} = t_{ij} + W(\Lambda_j^*, \mu_j) \begin{cases} = T_i(f_i^*) & \text{if } y_{ij}^* > 0 \\ > T_i(\ f_i^*) & \text{if } y_{ij}^* = 0 \end{cases} \text{ for } i \in N,\ j \in S. \tag{8}$$

Note that $T_i(f_i^*)$ can be interpreted as the equilibrium disutility for clients from node $i \in N$. Thus, node $i \in N$ will not be patronized from node $j \in S$ when τ_{ij} is greater than the equilibrium disutility.

To further explain the equilibrium condition for a customer-choice system; let us focus on the case where y_{ij} are binary variables. In this case if $y_{ij} = 1$, for $i \in N$ and $i \in S$, we must have

$$t_{ij} + W(\Lambda_j^*, \mu_j) \leq t_{ik} + W(\Lambda_k^*, \mu_k) \text{ for } k \in N, \tag{9}$$

which is equivalent to:

$$t_{ij} + W(\Lambda_j^*, \mu_j) \leq t_{ik} + W(\Lambda_k^*, \mu_k) + K(1 - y_{ij}) \text{ for } k \in N, \tag{10}$$

where K is sufficiently large to ensure that the inequality holds when $y_{ij} = 0$. Since we seek allocations under which no customer can do better by making unilateral move, we need to ensure that τ_{ij} is the same for all $j \in S$ where y_{ij} is positive, and hence (8) follows. Since y_{ij} can indeed be a fraction, then to find y_{ij}^* for $i \in N$, $j \in S$ in (8), we need solve the following nonlinear complementarity problem:

$$\begin{aligned} t_{ij} + W(\Lambda_j^*, \mu_j) - T_i(f_i^*) &\geq 0 \text{ for } i \in N, j \in S \\ y_{ij}(t_{ij} + W(^*_j, \mu_j) - T_i(f_i^*)) &= 0 \text{ for } i \in N, j \in S \\ y_{ij} &\geq 0 \text{ for } i \in N, j \in S. \end{aligned} \tag{11}$$

Given that from (2) and (3) we have $W(\Lambda_j^*, \mu_j) = \frac{1}{\mu_j - \sum_{r \in N} \lambda_r^{\max} y_{rj}}$ for $j \in S$ and from (4) and (7) we have $T_i(f_i) = \frac{f_i^{\max} - \sum_{k \in S} y_{ik}}{\alpha}$ for $i \in N$, nonlinear complementarity problem (11) can be rewritten as:

$$\begin{aligned} t_{ij} + \frac{1}{\mu_j - \sum_{r \in N} \lambda_r^{\max} y_{rj}} - \frac{f_i^{\max} - \sum_{k \in S} y_{ik}}{\alpha} &\geq 0 \text{ for } i \in N,\ j \in S \\ y_{ij}\left(t_{ij} + \frac{1}{\mu_j - \sum_{r \in N} \lambda_r^{\max} y_{rj}} - \frac{f_i^{\max} - \sum_{k \in S} y_{ik}}{\alpha}\right) &= 0 \text{ for } i \in N,\ j \in S \\ y_{ij} &\geq 0 \text{ for } i \in N,\ j \in S. \end{aligned} \tag{12}$$

Equations (12) can be regarded as the equilibrium condition.

We next turn our attention to the objective of maximizing accessibility. Define a binary decision variable $x_j, j \in M$ to be 1 if an already located facility at j decides to open a service with a service rate of $\mu_j > 0$ and 0 otherwise, and let $\mathbf{x}$ represent the m-dimensional location vector. As defined earlier, the decision variable y_{ij} is the fraction of the population of node $i \in N$ who request service from facility $j \in M$. The customer allocation is then represented by a $(m \times n)$-dimensional matrix $\mathbf{y}$,

and given the service rate vector $\boldsymbol{\mu}$, the total number of people who would benefit from the public service is

$$Z(\mathbf{x},\mathbf{y},\boldsymbol{\mu}) = \sum_{j\in M}\sum_{i\in N} \lambda_i^{max} y_{ij}. \tag{13}$$

Define $W^{\max}$ to be the maximum waiting time that is allowed at each facility such that $W_j \leq W^{\max}$ for $j \in S$. From (3), for $j \in S$

$$\mu_j \geq \sum_{i\in N} \lambda_i^{\max} y_{ij} + \frac{1}{W^{\max}}.$$

Define $\mu^{\min}$ and $\mu^{\max}$ ($\mu^{\max} > \frac{1}{W^{\max}}$ and $\mu^{\max} \geq 2\mu^{\min}$) to be the minimum and maximum possible service rate at each facility, respectively, such that $\mu^{\min} \leq \mu_j \leq \mu^{\max}$ for facility $j \in S$. Define $C^{\max}$ to be the available capacity to assign to all open facilities, such that $\sum_{j\in S} \mu_j = C^{\max}$. We also require $\mu^{\min} \leq C^{\max}$.

We can now state the mathematical programming formulation of the Maximal Accessibility Network Design Problem (MANDP) as follows:

$$\max Z(\mathbf{x},\mathbf{y},\boldsymbol{\mu}) = \textstyle\sum_{j\in M}\sum_{i\in N} \lambda_i^{max} y_{ij}$$

subject to

$$\sum_{j\in M} y_{ij} \leq 1, \quad i \in N \tag{14}$$

$$y_{ij} \leq x_j, \quad i \in N, j \in M \tag{15}$$

$$\mu_j - \sum_{i\in N} \lambda_i^{\max} y_{ij} - \frac{x_j}{W^{\max}} \geq 0, \quad j \in M \tag{16}$$

$$x_j \mu^{\min} \leq \mu_j \leq x_j \mu^{\max},\ j \in M \tag{17}$$

$$\sum_{j\in M} \mu_j = C^{\max} \tag{18}$$

$$t_{ij} + \frac{1}{\varepsilon(1-x_j) + \mu_j - \sum_{r\in N} \lambda_r^{\max} y_{rj}} - \frac{f_i^{\max} - \sum_{k\in M} y_{ik}}{\alpha} \geq 0,\ i \in N,\ j \in M \tag{19}$$

$$y_{ij}\left(t_{ij} + \frac{1}{\varepsilon(1-x_j) + \mu_j - \sum_{r\in N} \lambda_r^{\max} y_{rj}} - \frac{f_i^{\max} - \sum_{k\in M} y_{ik}}{\alpha}\right) = 0,\ i \in N,\ j \in M \tag{20}$$

$$y_{ij} \geq 0, x_j \in \{0,1\}, \mu_j \geq 0,\ j \in M,\ i \in N. \tag{21}$$

Constraints (14) ensure that the number of people served at each population zone cannot exceed its population, whereas constraints (15) guarantee that service can be received from only open facilities. Constraints (16, 17) ensure that the waiting time and the service rate at each facility remain within their pre-specified levels, respectively. Constraint (18) makes sure that the total available service capacity is distributed to the open facilities. Constraints (19, 20) are the equilibrium conditions, where ε is a very small number and the term $\varepsilon(1-x_j)$ is to avoid division by zero when $x_j = 0$ forcing the associated y_{ij} and μ_j variables to zero.

We note that in MANDP we do not include a budget constraint on locating facilities since we assume that the potential locations are already up and running and there is no fixed cost associated with allocating servers to them. Nevertheless, if there are fixed costs for locating facilities our methodology to solve the problem can still be applied.

The formulation above belongs to the *customer choice* class of models since it is based on the assumption that customers always travel to the facility with minimum expected time (rather than to a facility chosen by a central authority). The main difficulty in solving the problem is, clearly, the equilibrium constraints (19, 20), which are nonlinear. In Section 4, we will show how MANDP can be linearized to find ϵ-optimal solutions efficiently.

4. An ϵ-optimal Solution Method for MANDP

In this section we outline an efficient approach to solve MANDP as an MIP. In Section 4.1 we first formulate the problem of finding the optimal customer allocation $\mathbf{y}^*(\mathbf{x}, \boldsymbol{\mu}(\mathbf{x}))$ given location vector $\mathbf{x}$ and server capacity vector $\boldsymbol{\mu}(\mathbf{x})$. We then formulate the problem of jointly finding the optimal server allocation $\boldsymbol{\mu}^*(\mathbf{x})$ and optimal customer allocation $\mathbf{y}^*(\mathbf{x}, \boldsymbol{\mu}^*(\mathbf{x}))$ given a location vector $\mathbf{x}$. Since the resulting problem is nonlinear, in Section 4.2 we describe how to linearize this problem. Finally, in Section 4.3 we show how MANDP can be presented as an efficient MIP. The idea we use in Section 4.2 is to replace the decision variables $\{\mu_j\}$ by the waiting times $\{W_j\}$ and then to approximate functions of $\{W_j\}$ using the TLA method developed in Aboolian et al. (2007).

4.1. Formulation of Subproblems OCA and OSACA

We first focus on a subproblem aiming at finding Optimal Customer Allocation (OCA) given a set of open facilities each with a pre-specified capacity. To find the customer allocation we just need to solve the nonlinear complementarity problem (12) given $S = \{j \in M \| x_j = 1\}$, the set of facility nodes under vector $\mathbf{x}$, and $\boldsymbol{\mu}(\mathbf{x})$.

Let z_{ij} be a binary decision variable which is equal to one if any fraction of customers at $i \in N$ visits facility $j \in S$ (i.e. if $y_{ij} > 0$), and zero otherwise (if $y_{ij} = 0$). Then clearly $y_{ij} \leq z_{ij},\ i \in N, j \in S$. To obtain the equilibrium solution to the nonlinear complementarity problem (12), we propose the following model for OCA (recall that $\mathbf{x}$ and $\boldsymbol{\mu}(\mathbf{x})$ are known):

$$\max Z_{\mathbf{x},\ \boldsymbol{\mu}(\mathbf{x})}(\mathbf{y}) = \sum_{i \in N} \lambda_i^{max} \left(\sum_{j \in S} y_{ij} \right)$$

subject to

$$y_{ij} - z_{ij} \leq 0,\ i \in N, j \in S \tag{22}$$

$$\sum_{k\in S} y_{ik} - \left(f_i^{\max} - \alpha t_{ij} - \frac{\alpha}{\mu_j - \sum_{r\in N} \lambda_r^{\max} y_{rj}} \right) \geq 0, \quad i \in N, j \in S \tag{23}$$

$$\sum_{k\in S} y_{ik} - \left(f_i^{\max} - \alpha t_{ij} - \frac{\alpha}{\mu_j - \sum_{r\in N} \lambda_r^{\max} y_{rj}} \right) - \mathcal{L}_i(1 - z_{ij}) \leq 0, \quad i \in N, j \in S \tag{24}$$

$$f_i^{\max} - \alpha t_{ij} - \frac{\alpha}{\mu_j - \sum_{r\in N} \lambda_r^{\max} y_{rj}} + \mathcal{L}_i(1 - z_{ij}) \geq 0, \quad i \in N, j \in S \tag{25}$$

$$y_{ij} \geq 0, z_{ij} \in \{0,1\}, j \in S, i \in N \tag{26}$$

where $\mathcal{L}_i$ is a large enough number. Later we show how to compute $\mathcal{L}_i$. Consider the following exhaustive cases:

Case 1: $(y_{ij} > 0)$. In this case, from (22) we have $z_{ij} = 1$ and from (23) and (24) we obtain $\sum_{k\in S} y_{ik} = f_i^{\max} - \alpha t_{ij} - \frac{\alpha}{\mu_j - \sum_{r\in N} \lambda_r^{\max} y_{rj}}$. Hence $y_{ij} \left(t_{ij} + \frac{1}{\mu_j - \sum_{r\in N} \lambda_r^{\max} y_{rj}} - \frac{f_i^{\max} - \sum_{k\in S} y_{ik}}{\alpha} \right) = 0$.

Case 2: $(y_{ij} = 0)$. In this case, whether $z_{ij} = 1$ and then $\sum_{k\in S} y_{ik} = f_i^{\max} - \alpha t_{ij} - \frac{\alpha}{\mu_j - \sum_{r\in N} \lambda_r^{\max} y_{rj}}$, or $z_{ij} = 0$, and then, from (23) and (24), $\sum_{k\in S} y_{ik} \geq f_i^{\max} - \alpha t_{ij} - \frac{\alpha}{\mu_j - \sum_{r\in N} \lambda_r^{\max} y_{rj}}$, we obtain $y_{ij} \left(t_{ij} + \frac{1}{\mu_j - \sum_{r\in N} \lambda_r^{\max} y_{rj}} - \frac{f_i^{\max} - \sum_{k\in S} y_{ik}}{\alpha} \right) = 0$.

Therefore, constraints (22)–(24) ensure that the second set of equilibrium condition in (12) hold and since constraints (23) are the first part of the equilibrium condition in (12), we can conclude that constraints (22)–(24) ensure that the equilibrium condition in (12) holds.

Constraints (25) ensure that if $f_i(T_i) = (f_i^{\max} - \alpha \tau_{ij}) < 0$, then z_{ij} in (25) is forced to be equal to 0 and therefore, from (22), $y_{ij} = 0$. If $z_{ij} = 1$, from (25) $\tau_{ij} = T_i(f_i) \leq \frac{f_i^{\max}}{\alpha}$. In case $\sum_{k\in S} z_{ik} = 0$, then from (22) we have $\sum_{k\in S} y_{ik} = 0$ and from (23) $f_i^{\max} - \alpha \tau_{ij} = f_i^{\max} - \alpha t_{ij} - \frac{\alpha}{\mu_j - \sum_{r\in N} \lambda_r^{\max} y_{rj}} \leq 0$ for $j \in S$, therefore, $\tau_{ij} = T_i(f_i) \geq \frac{f_i^{\max}}{\alpha}$.

Note that if $\sum_{k\in S} y_{ik} > 0$ then there exists a $j \in S$ such that $z_{ij} = 1$ and from (23) and (24) we obtain $\sum_{k\in S} y_{ik} = f_i^{\max} - \alpha(t_{ij} + W_j) \leq 1$. Therefore, constraints (14) are redundant in OCA.

Since when $f_i^{\max} - \alpha t_{ij} \leq 0, z_{ij} = 0$, we can replace in OCA and MANDP, the parameter N with N_j for any $j \in M$, where N_j denotes the set of all customer nodes for which $f_i^{\max} - \alpha t_{ij} > 0$.

As a next step we consider the subproblem aiming at Optimal Server Assignment and Customer Allocation (OSACA). To this end, we assume that only a location vector $\mathbf{x}$ is given, and consider the problem of determining also the corresponding optimal capacity assignment vector $\boldsymbol{\mu}^*(\mathbf{x})$ in addition to the optimal customer allocation $\mathbf{y}^*(\mathbf{x}, \boldsymbol{\mu}^*(\mathbf{x}))$. Since the service rate at each facility is a decision variable we should include the maximum waiting time, the service rate and the total service capacity constraints. Let S be the set of open facilities corresponding to $\mathbf{x}$. Then, the mathematical programming formulation of OSACA for S is:

$$\max Z_{\mathbf{x}}(\mathbf{y}) = \sum_{j\in S} \sum_{i\in N_j} \lambda_i^{max} y_{ij}$$

subject to

(22)-(25),

$$\mu_j \geq \sum_{i \in N} \lambda_i^{\max} y_{ij} + \frac{1}{W^{\max}}, \quad j \in S \tag{27}$$

$$\mu^{\min} \leq \mu_j \leq \mu^{\max}, \quad j \in S \tag{28}$$

$$\sum_{j \in S} \mu_j = C^{\max}, \tag{29}$$

$$y_{ij} \geq 0, z_{ij} \in \{0,1\}, \ \mu_j \geq 0, j \in S, i \in N_j.$$

Recall that constraints (27) ensure that $W_j \leq W^{\max}$ from all $j \in S$ and together with constraints (28) and (29) are the relevant constraints from MANDP that determine the size of μ_j, $j \in S$. Thus, the only differences between problem OSACA and problem OCA are constraints (27)-(29) and the new decision variables $\mu_j \geq 0, j \in S$. Here it can be easily verified that we can set $\mathcal{L}_i = \max_{j \in S}\{\alpha t_{ij} + \alpha W^{\max}\}$ since it ensures that when $z_{ij} = 0$, there is no conflict between (23) and (24) (i.e. $\sum_{k \in S} y_{ik} \geq f_i^{\max} - \alpha t_{ij} - \frac{\alpha}{\mu_j - \sum_{r \in N} \lambda_r^{\max} y_{rj}}$, and $\sum_{k \in S} y_{ik} \leq a$ (a is a number that is greater than or equal to $f_i^{\max}$). We note that problem OSACA is still a very hard problem to solve because of the nonlinearities in constraints (23)–(25).

In the next section, we will convert OSACA into an MIP.

4.2. Formulating subproblem OSACA as an MIP

The nonlinearity in OSACA is due to the expected waiting times $\{W_j\}$. As mentioned earlier, the main idea here is to use $\{W_j\}$ instead of $\{\mu_j\}$ as decision variables and then to approximate functions of $\{W_j\}$. Since $W_j = \frac{1}{\mu_j - \sum_{i \in N_j} \lambda_i^{\max} y_{ij}}$,

$$\mu_j = \frac{1}{W_j} + \sum_{i \in N_j} \lambda_i^{\max} y_{ij} \text{ for } j \in S. \tag{30}$$

Since from (30) $\sum_{i \in N_j} \lambda_i^{\max} y_{ij} = \mu_j - \frac{1}{W_j}$ for $j \in S$ and from (29) $\sum_{j \in S} \mu_j = C^{\max}$, we can reformulate the objective function

$$Z_{\mathbf{x}}(\mathbf{y}) = \sum_{j \in S} \sum_{i \in N_j} \lambda_i^{max} y_{ij} = \sum_{j \in S} \left(\mu_j - \frac{1}{W_j} \right) = \sum_{j \in S} \mu_j - \sum_{j \in S} \frac{1}{W_j}$$

as our new objective function:

$$Z_{\mathbf{x}}(\mathbf{W}) = C^{\max} + \sum_{j \in S} \frac{-1}{W_j}. \tag{31}$$

Note that $\frac{-1}{W_j}$ is concave increasing in W_j for $j \in S$, and therefore we can use the TLA approximation technique proposed in Aboolian et al. (2007) and summarized in Appendix A to create an ϵ-approximator (piece-wise linear approximation with maximum relative error ϵ) for $\frac{-1}{W_j}$.

Denote $G(W_j) = \frac{-1}{W_j}$ for $W_j \in \left[\frac{1}{\mu^{\max}}, W^{\max}\right]$ and $G\epsilon(W_j)$ to be the ϵ-approximator for $G(W_j)$, such that

$$G(W_j) \leq G\epsilon(W_j) \leq (1+\epsilon)G(W_j).$$

Let $L_j\epsilon$ be the number of linear segments in $G\epsilon(W_j)$, with the endpoints of segment l given by c_j^l and c_j^{l+1} for $l \in \{1, ..., L_j\epsilon\}$. Let b_j^l be the slope of segment l, and $a_j^l = c_j^{l+1} - c_j^l$ be the length of this segment (projected onto the W_j axis). The function $G\epsilon(W_j)$ can be represented as follows for $W_j \in [\frac{1}{\mu^{\max}}, W^{\max}]$:

$$G\epsilon(W_j) = -\mu^{\max} + \sum_{l=1}^{L_j\epsilon(W_j)} a_j^l b_j^l e_j^l, \tag{32}$$

where in (32) we take into consideration (see (30)) that $G(W_j)$, which is a negative function, starts at $-\mu^{\max}$ and $L_j\epsilon(W_j) = \max\{l : c_j^l \leq W_j\}$ (note that $L_j\epsilon = L_j\epsilon(W^{\max})$), and $e_j^l = 1$ if $l < L_j\epsilon(W_j)$ and $e_j^l = \frac{W_j - c_j^l}{a_j^l}$ if $l = L_j\epsilon(W_j)$. Note that $W_j = \sum_{l=1}^{L_j\epsilon} a_j^l e_j^l + \frac{1}{\mu^{\max}}$.

We now introduce the linearized $\varphi - optimal$ model for OSACA. Let,

$$Z_{\mathbf{x}}\varphi(\mathbf{W}) = C^{\max} + \sum_{j\in S}\left(\sum_{l=1}^{L_j\epsilon} a_j^l b_j^l e_j^l - \mu^{\max}\right) = \sum_{j\in S}\sum_{l=1}^{L_j\epsilon} a_j^l b_j^l e_j^l + C^{\max} - \sum_{j\in S}\mu^{\max}. \tag{33}$$

The problem is:

$$\max Z_{\mathbf{x}}\varphi(\mathbf{W}) = \sum_{j\in S}\sum_{l=1}^{L_j\epsilon} a_j^l b_j^l e_j^l - \sum_{j\in S}\mu^{\max} + C^{\max}$$

subject to

$$\begin{aligned}
y_{ij} - z_{ij} &\leq 0, \ i \in N_j, j \in S \\
\sum_{k\in S} y_{ik} - (f_i^{\max} - \alpha t_{ij} - \alpha W_j) &\geq 0, \quad i \in N_j, j \in S \\
\sum_{k\in S} y_{ik} - (f_i^{\max} - \alpha t_{ij} - \alpha W_j) - \mathcal{L}_i(1 - z_{ij}) &\leq 0, \quad i \in N_j, j \in S \\
f_i^{\max} - \alpha t_{ij} - \alpha W_j + \mathcal{L}_i(1 - z_{ij}) &\geq 0, \quad i \in N_j, \ j \in S \\
W_j - \sum_{l=1}^{L_j\epsilon} a_j^l e_j^l &= \frac{1}{\mu^{\max}}, \ j \in S \\
W_j &\leq W^{\max}, j \in S
\end{aligned} \tag{34}$$

$$\sum_{l=1}^{L_j\epsilon} a_j^l b_j^l e_j^l - \sum_{i\in N_j} \lambda_i^{\max} y_{ij} \geq 0, j \in S \tag{35}$$

$$\sum_{l=1}^{L_j\epsilon} a_j^l b_j^l e_j^l - \sum_{i\in N_j} \lambda_i^{\max} y_{ij} \le \mu^{\max} - \mu^{\min},\ j \in S \tag{36}$$

$$\sum_{j\in S}\sum_{i\in N_j} \lambda_i^{\max} y_{ij} - \sum_{j\in S}\sum_{l=1}^{L_j\epsilon} a_j^l b_j^l e_j^l = C^{\max} - \sum_{j\in S} \mu^{\max} \tag{37}$$

$$0 \le e_j^l \le 1,\ \ j \in S,\ l \in \{1, ..., L_j\epsilon\}$$

$$y_{ij} \ge 0, z_{ij} \in \{0,1\},\ W_j \ge 0,\ \ j \in S, i \in N_j.$$

We note that the linearized version computes W_j^* instead of $\mu_j^*,\ j \in S$, but once $\{y_{ij}^*\}$ and $\{N_j^*\}$ are known we can simply compute $\mu_j^* = \frac{1}{W_j^*} + \sum_{i\in N_j} \lambda_i^{\max} y_{ij}^*,\ j \in S$. Constraints (35-36) ensures that $\mu^{\min} \le \mu_j \le \mu^{\max},\ j \in S$. Constraint (37) ensures that $\sum_{j\in S} \mu_j = C^{\max}$.

It can be easily verified that the $Z_{\mathbf{x}}^*(\mathbf{y}) \le Z_{\mathbf{x}}^{\varphi^*}(\mathbf{W}) \le (1+\varphi) Z_{\mathbf{x}}^*(\mathbf{y})$, where

$$\varphi = \begin{cases} \le \epsilon & \text{if } \sum_{j\in S} \mu_j \le 2 \sum_{j\in S} \sum_{i\in N_j} \lambda_i^{\max} y_{ij} \\ = -\epsilon + \frac{\sum_{j\in S} \mu_j}{\sum_{j\in S} \sum_{i\in N_j} \lambda_i^{\max} y_{ij}} \epsilon & \text{if } \sum_{j\in S} \mu_j > 2 \sum_{j\in S} \sum_{i\in N_j} \lambda_i^{\max} y_{ij} \end{cases}$$

and for a small enough ϵ (e.g. $\epsilon = 0.001$) the error φ will be negligible and we can consider the solution to the φ -optimal model as the optimal solution for OSACA.

4.3. The MIP Formulation of MANDP

Next we introduce the linearized φ-optimal model for MANDP by extending the linearized φ-optimal model for OSACA to include the location of facilities as well.

We now use the binary variables x_j defined earlier and since S is not given as in OSACA we consider all the potential customers in M. The problem is

$$\max Z\varphi(\mathbf{W}, \mathbf{x}) = \sum_{j\in M}\sum_{l=1}^{L_j\epsilon} a_j^l b_j^l e_j^l - \mu^{\max} \sum_{j\in M} x_j + C^{\max}$$

subject to

$$y_{ij} - z_{ij} \le 0,\ \ i \in N_j, j \in M$$

$$z_{ij} - x_j \le 0,\ \ i \in N_j, j \in M$$

$$W_j \le W^{\max} x_j,\ \ j \in M$$

$$\sum_{k\in M} y_{ik} - (f_i^{\max} - \alpha t_{ij} - \alpha W_j) + 1 - x_j \ge 0,\ \ \ i \in N_j, j \in M \tag{38}$$

$$\sum_{k\in M} y_{ik} - (f_i^{\max} - \alpha t_{ij} - \alpha W_j) - \mathcal{L}_i(1 - z_{ij}) \le 0,\ \ \ i \in N_j, j \in M$$

$$f_i^{\max} - \alpha t_{ij} - \alpha W_j + \mathcal{L}_i(1 - z_{ij}) \ge 0,\ \ \ i \in N_j, j \in M$$

$$W_j - \frac{x_j}{\mu^{\max}} - \sum_{l=1}^{L_j\epsilon} a_j^l e_j^l = 0,\ \ j \in M$$

$$\sum_{l=1}^{L_j\epsilon} a_j^l b_j^l e_j^l - \sum_{i\in N_j} \lambda_i^{\max} y_{ij} \geq 0,\ j \in M$$

$$\sum_{l=1}^{L_j\epsilon} a_j^l b_j^l e_j^l - \sum_{i\in N_j} \lambda_i^{\max} y_{ij} \leq \mu^{\max} - \mu^{\min},\ j \in M$$

$$\sum_{j\in M}\sum_{i\in N_j} \lambda_i^{\max} y_{ij} - \sum_{j\in M}\sum_{l=1}^{L_j\epsilon} a_j^l b_j^l e_j^l + \sum_{j\in M} \mu^{\max} x_j = C^{\max}$$

$$0 \leq e_j^l \leq 1,\ \ j \in M,\ l \in \{1, ..., L_j\epsilon\}$$

$$y_{ij} \geq 0,\ W_j \geq 0, x_j \in \{0,1\}, z_{ij} \in \{0,1\},\ j \in M, i \in N_j.$$

In this formulation we have a new constraint ($z_{ij} \leq x_j \quad i \in N_j, j \in M$). To ensure that y_{ij} cannot be positive unless there is a facility located at j, and since (23) is not feasible when $x_j = 0$, an additional term is added in (38) to ensure that $\sum_{k\in M} y_{ik} \geq b$ (where b is a negative number) when $x_j = 0$. Again, it can be easily verified that $Z^*(\mathbf{x}, \mathbf{y}) \leq Z\varphi(\mathbf{W}, \mathbf{x}) \leq (1+\varphi)Z^*(\mathbf{x}, \mathbf{y})$, where

$$\varphi = \begin{cases} \leq \epsilon & \text{if } \sum_{j\in M}\mu_j \leq 2\sum_{j\in M}\sum_{i\in N_j}\lambda_i^{\max} y_{ij} \\ = -\epsilon + \frac{\sum_{j\in M}\mu_j}{\sum_{j\in M}\sum_{i\in N_j}\lambda_i^{\max} y_{ij}}\epsilon & \text{if } \sum_{j\in M}\mu_j > 2\sum_{j\in M}\sum_{i\in N_j}\lambda_i^{\max} y_{ij} \end{cases}$$

and for a small enough ϵ (e.g. $\epsilon = 0.001$) the error φ will be negligible and we can consider the solution to the φ-optimal model as the optimal solution for the original model of MANDP.

Please note that $L_j\varepsilon$ which is the number of line segments used to approximate the inverse waiting time at facility j is a function of ε and $\mu^{\max}$. Each line segments corresponds to one continuous variable e_j^l. Table 1 shows the number of line segments required for each potential facility location for each combination of ε and $\mu^{\max}$.

5. Computational Results

In this section, we present a set of computational experiments to demonstrate the performance of the proposed algorithm as well as a realistic illustrative example.

5.1. The Efficiency of the Algorithm

We consider the 40 p-median problems in Beasley (1990) as a basis of our computational experiments. These problems range from $n = 100$ to $n = 900$ population zones. The values of p provided by Beasley (1990) are irrelevant in the context of this paper. For $100 \leq n \leq 500$ and for $600 \leq n \leq 900$, we generate problem instances by respectively increasing the number of alternative facility locations $m = 20, 40, 60, \ldots$ and $m = 10, 20, 30 \ldots$ until the resulting instance cannot be solved within a pre-specified time. For example, Beasley Problem #6 with $n = 200$ cannot be solved within 3,600 seconds

$\mu^{\max}\backslash\varepsilon$	0.05	0.01	0.005	0.001
50	10	21	29	63
100	11	24	34	74
150	12	26	36	80
200	13	27	38	85
250	13	29	40	88
300	14	29	41	91
350	14	30	42	94
400	14	31	43	96
450	14	31	44	98
500	15	32	45	99

Table 1 The number of line segments required for each potential facility location due to model linearization

for $m \geq 160$, and hence our experiment set contains only seven instances for this problem. For each Beasley problem, we selected the m potential locations as nodes $k * \lfloor (n/m) \rfloor$, where $k = 1, 2, ..., m$. For example, for Beasley Problem #1 ($n = 100$), when $m = 20$ the chosen alternative facility locations are nodes $5, 10, 15, 20, \ldots, 100$.

Our aim is to evaluate the performance of the algorithm in terms of the CPU time, for which we set the limit as 1 hour per instance. A total of 136 problem instances were solved to optimality on a computer with Intel Core S2 Duo 2.67 Ghz with 4 GB ram running Windows Vista. The program was coded in C^{++} and all resulting MIPs were solved using CPLEX 12.6. The other model parameters were set as follows: total service capacity $C^{\max} = \frac{n}{2}$, service rate limits $\mu^{\min} = 5$, $\mu^{\max} = \frac{n}{10}$, maximum wait time $W^{\max} = 1$, maximum demand rate $\lambda_i^{max} = 1$, maximum participation fraction $f_i^{\max} = 1$, slope of the participation function $\alpha = .4$ and maximum approximation error $\varepsilon = 0.001$. For $100 \leq n \leq 600$, $n = 700$ and $800 \leq n \leq 900$, we solved 5, 4, and 3 instances respectively. During the experiment we recorded the true gap and for $\varepsilon = 0.001$, we observed $0.00024 \leq gap \leq 0.00512$ with an average gap $= 0.00156$.

Table 2 depicts the average CPU times for the smaller Beasley problems ($100 \leq n \leq 600$) in our set of experiments, whereas Table 3 reports on the larger problems that were solved. For a given (n, m), the number of problem instances solved to optimality are denoted by [.], unless all the instances were solved to optimality. As expected, when the number of population zones n is fixed, the problem becomes more computationally challenging as the number of alternative locations m increases (see Table 2). This is not necessarily true, however, when n is increased for fixed m. For example Table

2 shows that the average CPU time for the $n = 200, m = 80$ problems is 14.41 seconds, whereas the smaller $n = 100, m = 80$ problems require 5.71 seconds on the average. Table 3 shows that the proposed algorithm performs well in solving fairly large problem instances. For example, we were able to solve two of the $n = 800$ and $m = 40$ instances within an average of about 6.5 minutes.

$n\backslash m$	20	40	60	80	100
100	0.24	0.48	5.24	5.71	9.71
200	0.35	2.50	4.66	14.41	22.31
300	0.23	5.10	14.47	32.74	151.73
400	0.47	20.91	190.71	556.19 [4]	754.73 [2]
500	0.67	35.45	323.45 [4]	[-]	[-]

Table 2 Average CPU times (sec) for the smaller instances from Beasley (1990)

$n\backslash m$	10	20	30	40
600	0.59	2.72	34.25	291.53 [4]
700	0.38	2.76	28.90	141.12
800	0.96	57.34	74.25	390.05 [2]
900	6.74	68.22 [2]	31.84 [1]	[-]

Table 3 Average CPU times (sec) for the larger instances from Beasley (1990)

5.2. An Illustrative Case Study

In this section, we present a realistic case that is based on the network of 22 hospitals in the City of Toronto, Canada. Berman et al. (2007) is the source for the 96-node network model we use here in representing the hospital system, which serves a population of over 2.6 million according to the 2011 census. Each node represents a forward sortation area (FSA) defined by the first 3 digits of the Canadian postal code. Berman et al. (2007) placed the nodes at the FSA centroids and established a link between any two nodes if the corresponding FSAs share a boundary. Using Euclidean distances among the connected nodes, they computed the shortest distance between all node pairs. To obtain the travel time between each node pair, we divide the shortest distance between the nodes by the average speed of travel.

As an illustrative case, we study the development of a network of clinics as part of a preventive care program offered by the government (e.g., cancer screening, vaccination, counseling). There are two

hospitals in three FSAs: M3N, M6M and M6S, and a single hospital in 16 FSAs. Thus, we assume that each of the 19 hospital sites constitutes an alternative location for the clinics to be established. The demand of each node is the total number of residential and business dwellings. There are 1.12 million dwellings in the area represented by our model. For FSAs on the border of the City of Toronto, the number of dwellings in the neighbouring FSAs are added to represent the fact that some of the customers of these hospitals do come from outside the city. Without loss of generality, we assume at most one annual visit per dwelling and set the λ_i^{max} values accordingly.

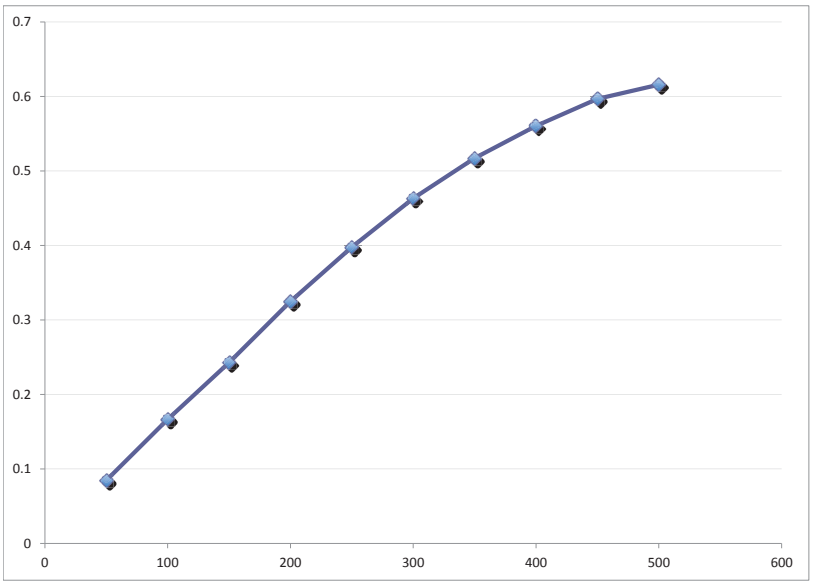

Figure 1 **The impact of total available system capacity on participation**

In developing the network of clinics, the government needs to decide which existing hospitals should be housing a new clinic, (x_j), and the allocation of the total investment in building the available service capacity at each clinic, (μ_j). Given the strategic nature of the MANDP, the optimal capacity levels prescribe the target number of patients that can be seen in a given time window. The most appropriate number of physicians, nurses, technicians and equipment needed to achieve this target is left to the detailed operational design phase pertaining to each clinic. The public reaction to the offered services is represented in the model by the participation rates (y_{ij}).

We study the impact of increasing the overall system capacity, C^{max}, on the total number of people served, when the clinic location and capacity allocation decisions are optimized. We work with an hourly minimum service capacity (μ^{min}) of 10 patients and maximum service capacity (μ^{max}) of 200 patients. We assume 250 days/year, 8 hours/day for the clinics and average travel speed of 5 miles/hour for the patients. We also assume that 5% of the potential customers will not participate regardless

Location #	FSA #	# of hospitals	FSA code	$C^{\max}=150$ Facility located (1=yes)	Service rate assigned	Demand served	% utilization	$C^{\max}=450$ Facility located (1=yes)	Service rate assigned	Demand served	% utilization
1	11	1	M1P					1	34.97	25.75	73.63
2	16	1	M1W					1	39.87	31.03	77.83
3	20	1	M2K	1	18.19	15.93	87.58				
4	22	1	M2M	1	17.19	15.41	89.65	1	27.22	20.49	75.28
5	25	1	M2R					1	15.31	10.42	60.06
6	34	2	M3N	1	10.00	8.39	83.90	1	20.74	14.83	71.50
7	37	1	M4C					1	38.17	30.04	78.70
8	39	1	M4G					1	22.29	15.38	72.24
9	43	1	M4M					1	15.54	8.93	69.95
10	44	1	M4N					1	28.31	20.70	73.12
11	52	1	M4Y					1	18.38	13.18	71.71
12	54	1	M5B								
13	57	1	M5G	1	57.89	54.57	94.26	1	11.24	7.36	65.48
14	65	1	M5T					1	38.14	29.51	77.37
15	77	2	M6M					1	32.68	24.55	75.12
16	81	2	M6S					1	37.11	28.45	76.66
17	89	1	M9C	1	10.98	9.68	88.16	1	25.94	19.22	74.09
18	92	1	M9N	1	25.75	23.2	90.10	1	20.06	14.13	70.44
19	95	1	M9V	1	10.00	8.8	88.00	1	27.03	20.33	75.21
							average				average
Total				7	150	135.98	88.8%	17	450	334.3	72.8%

Table 4 Capacity allocation decisions as a function of the total available system capacity $C^{\max}$

of the level of accessibility i.e., $f_i^{max}=0.95$ and people are willing to spend about an hour and 45 minutes to get the service i.e., $\alpha=0.55$. We also use $W^{\max}=1$ hour and $\epsilon=0.001$ (the maximum error φ was recorded to be less than or equal to .001).

Figure 1 depicts the impact of increasing the total available system capacity on the total participation. The decreasing returns to scale in the capacity investment is expected. Note, however, that the total participation rate remains under 65% for very high levels of service capacity. The policy insight that can be drawn from this finding is that the government's ability to increase participation by improving access to preventive care is limited. In this illustrative case, there seems to be room for improvement through parallel investments into educational programs that would highlight the impor-

tance of preventive care. Consequently, the slope of the participation function, α, can be reduced to enable the government to provide service to a larger population at the same levels of accessibility.

Table 4 depicts a comparison between two overall system capacity scenarios: 150 patients/hour and 450 patients/hour. Under the former scenario, 7 clinics are established with an average utilization of 88.8%. Whereas, the latter scenario results in 17 open clinics i.e., only M5B and M2K are not in the optimal solution, with an average utilization of 72.8%. Note that the utilization can be as low as 60% in M2R. As expected, the standard deviation of the utilization increases in response to increased total system capacity, and hence the coefficient of variation of the utilization increases from 0.035 to 0.064. A careful analysis of the service rate assigned to M3N, M6M and M6S which currently host 2 hospitals each, under the $C^{max} = 450$ scenario, reveals that having two hospitals in these FSAs needs to be examined in more detail (note that the service rates assigned to the clinics in these three FSAs are well within the range of the optimal service capacities of the other clinics). We believe that the reason for having only 7 facilities in the case of $C^{\max} = 150$ is that the service pooling effect results in more time saving (from reducing waiting time) than the increase in travel time. In other words, the pooling effect dominates.

Table 5 reports on our analysis pertaining to the evolution of the clinic network as the total system capacity, C^{max}, is gradually increased. Focusing on the number of open facilities, we identify three ranges: $C^{max} < 250$, $250 \leq C^{max} \leq 400$, and $C^{max} > 400$. Although the number of open clinics is not robust in the first range, it is interesting to note that it takes the values 13 or 14 in the second range and 17 or 18 in the third range. More importantly, nine of the 13 clinics established for $C^{max} = 250$ remain in the solution as the overall system capacity is increased to a level where all but one clinics are open. The observed robustness has two implications for the regulator: (i) the initial system capacity should not be set less than a certain level to avoid the range where the solution is not robust to changes in C^{max}, and (ii) Within the robust range, it is possible to gradually build up the clinic network in the event that there are budget limitations associated with the number of clinics that can be established at the outset.

6. Concluding Remarks

In this paper, we provide a mathematical formulation for the problem of maximizing access to public services by determining the configuration of a facility network so as to optimize the incorporation of the customers' choices. In addition to the siting decisions, we address the aggregate capacity decisions at each facility to be established. We present a procedure to linearize the resulting nonlinear integer program and identify an ϵ-optimal solution. The proposed approach proved effective in tackling fairly large-scale problem instances.

Alternative Location #	50	100	150	200	250	300	350	400	450	500	550
1				1	1	1	1	1	1	1	1
2					1	1	1	1	1	1	1
3			1	1			1	1			
4		1	1		1	1	1	1	1	1	1
5					1	1	1	1	1	1	1
6			1		1	1	1	1	1	1	1
7				1	1	1	1	1	1	1	1
8					1	1			1	1	1
9									1	1	1
10					1		1	1	1	1	1
11	1	1				1	1	1	1	1	1
12				1						1	1
13			1		1				1	1	1
14						1	1	1	1	1	1
15		1			1	1	1	1	1	1	1
16				1	1				1	1	1
17			1	1		1	1	1	1	1	1
18			1	1	1	1	1	1	1	1	1
19			1	1	1	1	1	1	1	1	1
Total open facilities	1	3	7	8	13	13	14	14	17	18	18

Table 5 **Facility location decisions as a function of the total available system capacity $C^{\max}$**

In the context of a realistic case based on the Toronto hospital network, we demonstrate the capability of the modeling framework to produce policy insights. For this instance, we were able to show that (i) the capability of the Ontario government to increase participation in its services by simply increasing accessibility is limited, (ii) the current clustering of the hospitals in downtown Toronto may not be the best capacity allocation strategy (we note that this is based on our model that ignores other considerations such as quality that may be significant), and (iii) a gradual capacity expansion strategy can be robust in public services, as long as the system is designed with an overall capacity that is above a threshold level. It is important to note that, in this context, the additional investment

required for increasing the overall system capacity needs to be traded off against the potential benefits (i.e., cost savings and improved quality of life) of the more invasive treatments that are avoided by the increased service.

Our model can be generalized or extended in a couple of ways. First, although as mentioned earlier the bi-level programming is not an efficient approach to solve our problem, it could be proven to be an efficient approach for the special case where only a limited number of service capacities are available. Second, our model can be extended to the case where a fixed cost is required to open a potential location. We could use our approach to solve this extended case, but it will result in an increase in the approximation error bound φ defined in section 4.3. Thus, the development of an efficient approach for dealing with fixed cost for facilities remains a challenge for future research.

Acknowledgments

We thank the editorial team for their comments, which helped us greatly in improving the write up. The first author's work was partially supported by a summer grant from the College of Business Administration at the California State University, San Marcos. The second and third authors' work is partially supported by individual Discovery Grants from the Natural Sciences and Engineering Research Council of Canada.

References

Aboolian R., O. Berman and D. Krass (2007), Competitive Facility Location Model with Concave Demand, *European Journal of Operational Research*, 181, 598-619.

Aboolian R., O. Berman and Z. Drezner (2008), Location-Allocation of Service Units on a Congested Network, *IIE Transactions*, 40, 422-433.

Aboolian, R., O. Berman and D. Krass (2012), Profit Maximizing Distributed Service System Design with Congestion and Elastic Demand, in *Transportation Science*, 46(2), 247-261.

Baron, O., O. Berman and D. Krass (2008), Facility Location with Stochastic Demand and Constraints on Waiting Time, *Manufacturing & Service Operations Management,* 10, 484-505.

Beasley J.E. (1990), OR Library-Distributing Test Problems by Electronic Mail, *Journal of the Operational Research Society*, 41, 1069-1072.

Berman O., and D. Krass (2002a), Facility Location Problems with Stochastic Demands and Congestion, Z. Drezner and H.W. Hamacher (eds.) *Location Analysis: Applications and Theory*, Springer Verlag, Berlin-Germany, 329-371.

Berman, O. and D. Krass (2002b), The Generalized Maximal Covering Location Problem, *Computers & Operations Research*, 29, 563-591.

Berman, O., D. Krass, and Z. Drezner, (2003), The Gradual Covering Decay Location Problem on a Network. *European Journal of Operational Research*, 151, 474480.

Berman, O. and Z. Drezner (2006), Location of Congested Capacitated Facilities with Distance-Sensitive Demand, *IIE Transactions*, 38, 213-231.

Berman, O., D. Krass and J. Wang (2006), Locating Service Facilities to Reduce Loss Demand, *IIE Transactions,* 38, 933-946.

Berman, O., D. Krass, and M.B.C. Menezes, (2007), Facility Reliability Issues in Network p-Median Problems: Strategic Centralization and Co-location Effects, *Operations Research*, 55, 332-350.

Castillo, I., A. Ingolfsson and T. Sim (2009), Socially Optimal Location of Facilities with Fixed Servers, Stochastic Demand and Congestion, *Production and Operations Management,* 18, 721-736.

Elhedli, S. (2006), Service System Design with Immobile Servers, Stochastic Demand and Congestion, *Manufacturing & Service Operations Management,* 8, 92-97

Farahani, R.Z., N. Asgari, N. Heidari, M. Hosseininia and M. Goh (2012), Covering Problems in Facility Location: A Review, *Computers & Industrial Engineering*, 62, 368 - 407.

Gu W., X. Wang and S.E. McGregor (2010), Optimization of Preventive Health Care Facility Locations, *International Journal of Health Geographics*, 9(17), 1-16.

Luo W. and Y. Qe (2009), An Enhanced Two-Step Floating Catchment Area (E2SFCA) Method for Measuring Special Accessibility to Primary Care Physicians, *Health & Place*, 15, 1100-1107.

Marianov V. (2003), Location of Multiple-Server Congestible Facilities for Maximizing Expected Demand, when Services are Non-Essential, Annals of OR 123, 125 141.

Marianov V., M. Rios, F.J. Barros. (2005), Allocating servers to facilities, when demand is elastic to travel and waiting times, RAIRO Operations Research 39,143-162.

Marianov V. and D. Serra (1998), Probabilistic Maximal Covering Location–Allocation for Congested Systems, *Journal of Regional Science*, 38, 401-424.

Marianov, V., D. Serra (2002), Location-Allocation of Multiple-Server Service Centers with Constrained Queues or Waiting Times. *Annals of Operations Research* 111, 35-50

Schilling, D. A., V. Jayaraman and R. Barkhi (1993), A Review of Covering Problem in Facility Location. *Location Science,* 1, 25 - 55.

Statistics Canada (2008), *Guide to the Public Sector of Canada*, Catalogue no. 12-589-X

Verter V. and S. Lapierre (2002), Location of Preventive Health Care Facilities, *Annals of Operations Research*, 110, 123-132.

Zimmerman, S. (1997), Factors Influencing Hispanic Participation in Prostate Cancer Screening. *Oncology Nursing Forum*, 24, 499-504.

Zhang, Y., O. Berman, and V. Verter (2009), Incorporating Congestion in Preventive Healthcare Facility Network Design, *European Journal of Operational Research*, 198, 922-935.

Zhang, Y., O. Berman, P. Marcotte and V. Verter (2010), A Bilevel Model for Designing Preventive Healthcare Facility Networks, *IIE Transactions*, 42, 865-880.

Appendix A: The TLA Procedure

Let $f(t)$ be a concave, non-decreasing and twice-differentiable function for $t \in [0, \bar{\phi}]$ with $f(0) = 0$. We will construct a concave, piecewise linear function $f^{\epsilon}(t)$ such that $f^{\epsilon}(0) = 0$ and

$$f(t) \leq f^{\epsilon}(t) \leq (1+\epsilon) f(t) \text{ for } t \in [0, \bar{\phi}] \tag{A.1}$$

where $\epsilon \in (0,1)$ is the error bound.

To construct $f^{\epsilon}(t)$ the following notation is used:

l: is a line segment, $l = 1, \ldots, L$

b^l: is the slope of segment l

c^l: is the starting point of segment $l, c_{L+1} = \bar{\phi}$.

The TLA procedure is:

1. Set $f\epsilon(0) = 0$ and $c^1 = 0$.

2. Set the slope b^1 of the first segment equal to $f'(0)$ ($f'(0)$ is the derivative of $f(t)$ at $t = 0$).

3. The endpoint of the segment is a point $f\epsilon(c^2)$ on the ray originating at 0 and with the slope b^1 such that the relative error at c^2 is equal to ϵ.

4. The slope of segment l for $l = 2, \ldots, L$ is the slope of the ray originating at $(c^l, f\epsilon(c^l))$ that is tangent to the graph of $f(\cdot)$.

5. To find the endpoint of segment l for $l = 2, \ldots, L$, we repeat step 3 where c^2 is replaced by c^{l+1}, 0 is replaced by $f^{\epsilon}(c^{l+1})$ and b^1 is replaced by b^l.

In Figure A.1 we show $f(t)$ and $f^\epsilon(t)$ with three linear segments.

The relative error is the ratio $\frac{f^\epsilon(t)-f(t)}{f(t)}$. Note that since $f(0)=0$, the relative error at 0 is set to 0 and the first segment is tangent to the graph of $f(\cdot)$ at 0. The endpoint of the first segment c^2 is chosen so that $\frac{f^\epsilon(t)-f(c^2)}{f(c^2)}=\epsilon$. The procedure continues until f^ϵ has been defined for all points in $[0,\bar{\phi}]$.

It was shown in Aboolian, Berman and Krass (2007) that: (i) The TLA procedure converges in finitely many steps (finite L) to a piecewise linear function f^ϵ such that $f^\epsilon(0)=0$ and (A.1) holds for all $t\in[0,\bar{\phi}]$; and (ii) The number of linear segments in f^ϵ is minimized over all piecewise linear functions satisfying (A.1).

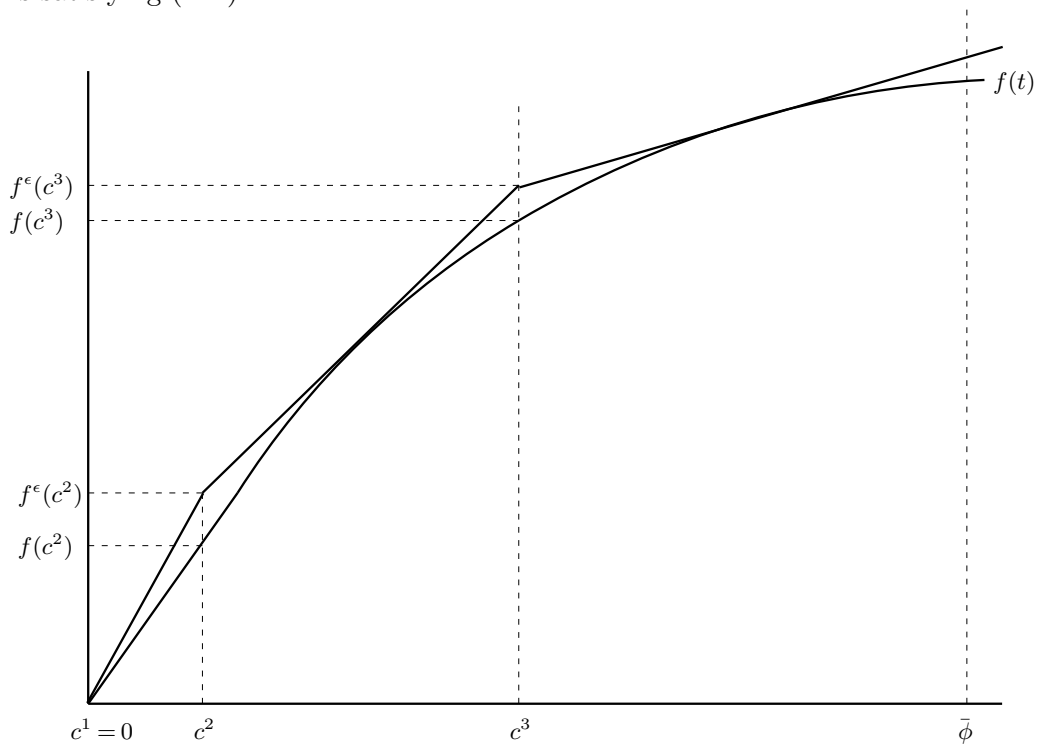

Figure A.1: $f(t)$ and the piecewise linear approximation $f^\epsilon(t)$ with three linear segments