IIE Transactions
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/uiie20

A bilevel model for preventive healthcare facility network design with congestion

Yue Zhang [a], Oded Berman [b], Patrice Marcotte [c] & Vedat Verter [d]

[a] College of Business Administration, University of Toledo, 2801 W. Bancroft Street, Toledo, OH, 43606, USA

[b] Rotman School of Management, University of Toronto, 105 St. George Street, Toronto, Ontario, M5S 3E6, Canada

[c] DIRO, Universite de Montreal, C.P. 6128, Succursale Centre-Ville, Montreal, Quebec, H3C 3J7, Canada

[d] Desautels Faculty of Management, McGill University, 1001 Sherbrooke Street West, Montreal, Quebec, H3A 1G5, Canada

Available online: 09 Oct 2010

To cite this article: Yue Zhang, Oded Berman, Patrice Marcotte & Vedat Verter (2010): A bilevel model for preventive healthcare facility network design with congestion, IIE Transactions, 42:12, 865-880

To link to this article: http://dx.doi.org/10.1080/0740817X.2010.491500

IIE Transactions (2010) **42**, 865–880

ISSN: 0740-817X print / 1545-8830 online
DOI: 10.1080/0740817X.2010.491500

A bilevel model for preventive healthcare facility network design with congestion

YUE ZHANG[1], ODED BERMAN[2], PATRICE MARCOTTE[3] and VEDAT VERTER[4,*]

[1] *College of Business Administration, University of Toledo, 2801 W. Bancroft Street, Toledo, OH 43606, USA*
E-mail: yue.zhang@utoledo.edu
[2] *Rotman School of Management, University of Toronto, 105 St. George Street, Toronto, Ontario, Canada, M5S 3E6*
E-mail: Berman@rotman.utoronto.ca
[3] *DIRO, Universite de Montreal, C.P. 6128, Succursale Centre-Ville, Montreal, Quebec, Canada, H3C 3J7*
E-mail: marcotte@iro.umontreal.ca
[4] *Desautels Faculty of Management, McGill University, 1001 Sherbrooke Street West, Montreal, Quebec, Canada, H3A 1G5*
E-mail: vedat.verter@mcgill.ca

Received April 2009 and accepted April 2010

Preventive healthcare aims at reducing the likelihood and severity of potentially life-threatening illnesses by protection and early detection. The level of participation in preventive healthcare programs is a critical determinant in terms of their effectiveness and efficiency. This article presents a methodology for designing a network of preventive healthcare facilities so as to improve its accessibility to potential clients and thus maximize participation in preventive healthcare programs. The problem is formulated as a mathematical program with equilibrium constraints; i.e., a bilevel non-linear optimization model. The lower level problem which determines the allocation of clients to facilities is formulated as a variational inequality; the upper level is a facility location and capacity allocation problem. The developed solution approach is based on the location–allocation framework. The variational inequality is formulated as a convex optimization problem, which can be solved by the gradient projection method; a Tabu search procedure is developed to solve the upper level problem. Computational experiments show that large-sized instances can be solved in a reasonable time. The model is used to analyze an illustrative case, a network of mammography centers in Montreal, and a number of interesting results and managerial insights are discussed, especially about capacity pooling.

Keywords: Preventive healthcare, network design, bilevel, congestion, equilibrium, variational inequality

1. Introduction

Preventive healthcare programs can save lives and contribute to a better quality of life by reducing the need for radical treatments, such as surgery or chemotherapy. Flu shots, blood tests, mammograms, and anti-smoking advice are among the most well-known preventive services. The substantial savings in the costs of diagnosis and therapy as well as the lower capital investment associated with preventive healthcare programs have been recognized for three decades (Walker, 1977). For example, studies show that mammograms taken on a regular basis have the potential to reduce deaths from breast cancer for women between the ages of 50 and 69 by up to 40% (Health Canada, 2005). Gornick *et al.* (2004) found that 36% of breast cancer patients without a mammogram received the diagnosis of late stage cancer, whereas this rate was 20% for the patient group who had mammograms.

In contrast with sick people who need urgent medical attention, the potential clientele of preventive healthcare often do not feel the necessity to receive these services and may not participate in preventive programs offered in their region. For instance, by 2003 every province in Canada had an organized program offering biennial mammography screening to asymptomatic women between the ages of 50 and 69 with no previous history of breast cancer. Although the proportion of women participating in the screening has increased over time, reaching 34% nationally in 2002, none of the organized programs have achieved the nationally established target of 70% participation (Public Health Agency of Canada, 2006). The achievement of desired participation level continues to be a challenge to many preventive healthcare programs. According to many empirical studies, such as McNoe *et al.* (1996), Zimmerman (1997), and Facione (1999), the ease of

*Corresponding author

access to the facilities is a crucial factor that influences people's decision to participate.

Recognizing the significance of accessibility, this article presents a methodology for designing a network of preventive healthcare facilities so as to maximize participation to preventive programs. The number of facilities to be established, the location of each facility, and the capacity at each facility are the main determinants of the configuration of the healthcare facility network. We represent the capacity at each facility by the number of service teams and use the total time needed to receive preventive services at a facility as a proxy for its accessibility. The total time comprises the time spent during transportation to the facility as well as the time spent at the facility while waiting and receiving services. Typically, the level of participation in a preventive program decreases with the expected total time to receive the services. In order to ensure service quality, however, the preventive healthcare facilities cannot be operated unless the size of their clientele fulfills a minimum workload requirement.

The methodology presented in this article incorporates the differentiating features of preventive healthcare. First, the number of people who seek the services at each facility is not controlled by the policy maker: preventive healthcare is a *user choice environment* in terms of the allocation of clients to facilities. Second, if the wait times are extensive due to congestion at the facilities, then clients' willingness to participate in a preventive program could decrease significantly. That is, the demand for preventive care is *elastic* with respect to the configuration of the facility network. The third distinguishing factor of preventive healthcare is the apparent link between *volume and quality of care*; i.e., the previously mentioned minimum workload requirement. In Canada, for example, a radiologist is required to interpret at least 4000 mammograms annually.

To the best of our knowledge, there are only two earlier papers that study the preventive healthcare facility network design problem. Verter and Lapierre (2002) used travel distance as a proxy for the accessibility of a facility by assuming that people would only go to the closest facility to seek preventive services. Their model maximizes participation by determining the number of facilities to be established as well as the location of each facility. Recently, Zhang *et al.* (2009) modeled each facility as an $M/M/1$ queue to incorporate the limited service capacity of preventive healthcare facilities and the resulting congestion. They used the total (travel, waiting, and service) time as a proxy for accessibility and assumed that clients at the same population zone would patronize the same facility; i.e., the one with the minimum expected total time. This assumption may not be realistic in the context of preventive care, and it also prevents Zhang *et al.* (2009) from identifying an equilibrium allocation of clients to facilities for many of the problem instances they studied.

This article extends the state of the art in preventive healthcare facility network design by: (i) determining the optimal number of servers at each facility as well as the number and locations of the facilities; and (ii) incorporating the possibility that people from the same population zone can patronize different facilities, which usually guarantees the existence of an allocation equilibrium. These require a completely different modeling approach (and solution methodology) than the earlier papers. In representing the problem, we adopt a bilevel formulation, where the lower level model captures the user choice nature of the allocation decisions, whereas the number, locations, and capacities of the facilities are determined at the upper level. The lower level finds a user equilibrium recognizing that the potential clients choose the facility that minimizes their expected total time. At equilibrium, everyone is content with the facility they patronize i.e., people from the same population zone expect to spend the same total time even if they go to different facilities. We formulate the lower level problem via a variational inequality. To the best of our knowledge, this is the first article that incorporates a user equilibrium in a facility network design model with congestion. Our results demonstrate that this new model framework for the considerations on user choice and capacity reallocation leads to a significant improvement in the accuracy of the solution as well as in the level of accessibility and participation.

We also note that, consistent with Zhang *et al.* (2009), the article primarily aims at a non-appointment or "walk-in" system, which applies to many routine preventive healthcare services (e.g., flu shots, anti-smoking advice, drug and alcohol use, nutrition, etc.) and runs in many countries or areas. Since we could not find a more suitable case with available data, the network of mammography centers in Montreal is used here as an illustrative example to show how our model and methodology can be applied and to discuss a number of managerial insights.

The remainder of this article is organized as follows. The next section provides an overview of the relevant literature. Section 3 describes the problem and formulates it as a bilevel programming model. A location–allocation framework is proposed in Section 4 to solve the problem. Section 5 presents computational results for the model. An illustrative case, based on the network of mammography centers in Montreal, is studied in Section 6. In the final section, conclusions and future research directions are provided.

2. Related literature

Although the design of healthcare facilities has been studied for a long time, few models incorporate the distinguishing features of preventive healthcare, which have been discussed in Zhang *et al.* (2009). The recent review by Daskin and Dean (2004) on the location of healthcare facilities, for example, makes no reference to preventive care. Similarly, more general literature reviews by Berman and Krass (2002) and Marianov and Serra (2002), which focus on public facility location problems with stochastic demand

and congestion in the context of fixed versus mobile servers, do not cite any articles on preventive healthcare. As mentioned earlier, Verter and Lapierre (2002) and Zhang *et al.* (2009) are the only papers that study the network design of preventive healthcare facilities.

A number of facility location and capacity allocation papers, based on a Mathematical Program with Equilibrium Constraints (MPEC) structure, are related to ours. Chao *et al.* (2003) developed a system optimization model for a resource allocation problem in a multiple facility system; i.e., allocating capacity and demand to facilities by a central authority in order to maximize average customer waiting time. They proved that the optimal solution of their model is incentive compatible for customers. However, their model does not consider travel distance, thus resulting in a solution structure of "one large and many small." Marianov (2003) presented a model for locating multiple-server facilities to maximize the overall demand. The demand at a node that is elastic to travel time and congestion is represented by the number of clients in the system. Similar to Chao *et al.* (2003), this is also a system optimization model and the assignment of demand to facilities is determined by a central authority, while the demand at each node is determined by an equilibrium constraint. Marianov *et al.* (2005) extended Marianov (2003) by allowing allocation of servers to facilities. Marianov *et al.* (2008) proposed a competitive facility location problem, in which clients choose a facility based on travel time and waiting time. User choice is probabilistic by using a logit formulation, and an equilibrium allocation of clients to facilities can be identified by solving a set of non-linear equations.

We note that the user choice behavior addressed in our model is different from the ones discussed above. Instead of assigning clients to facilities or letting clients choose where to shop probabilistically, our model not only allows clients to select the facility with the minimum expected total time but also allows the possibility that clients from the same population zone go to different facilities (they spend the same total time even if they go to different facilities). To accomplish this, we develop a completely different modeling approach and solution methodology.

The user equilibrium model has been extensively studied in the Spatial Price Equilibrium Problem (SPEP) and the Traffic Network Equilibrium Problem (TNEP). The SPEP seeks to determine the commodity supply prices demand prices and trade flows satisfying the equilibrium condition that demand price is equal to supply price plus transportation cost. Studies in this area include Samuelson (1952), Takayama and Judge (1964, 1971), Marcotte *et al.* (1992), etc. These models have also been used to study equilibrium problems in agriculture, energy markets, and finance, such as in Judge and Takayama (1973), Nagurney (1992), and Nolte (2008). In a congested transportation network, the TNEP aims at finding clients' travel paths with minimal cost from their origins to destinations. In particular, the network design problem associated with the traffic equilibrium is to optimize the capacities of the network links, so as to balance the transportation, investment, and maintenance costs of a network subject to congestion. This problem has been studied by Dafermos (1968) and Abdulaal and LeBlanc (1979). Magnanti and Wong (1984), Marcotte (1986), and Florian and Hearn (1995). Interested readers may also refer to Nagurney (1999) for a detailed review of user equilibrium models.

Despite the similar bilevel structure, the current article differs from the mentioned TNEP models in several aspects. First, congestion in the TNEP occurs on links, whereas in our problem it occurs at facilities where waiting times are modeled based on queueing theory. Second, the upper level of our problem is to select an undetermined number of locations and to make the capacity decisions at these locations, which is a combinatorial problem by itself. Third, our bilevel problem cannot be easily simplified to a single-level problem due to a minimum workload requirement.

3. The model

In this section, we model the problem of preventive healthcare facility network design by a bilevel formulation. The objective of the problem is to maximize the level of participation, by locating an undetermined number of facilities at the population zones over the network and allocating a given number of servers to open facilities. We incorporate the congestion at the facilities in the model and assume that clients patronize the facility with the minimum expected total time, which comprises the travel time to the facility and the expected time spent at the facility for possibly waiting and receiving the service (system waiting time). We also assume that the number of clients at each population zone who request service is a linearly decreasing function of the expected total time. In order to ensure service quality, facilities cannot be established unless their client base exceeds a minimum workload requirement.

An overwhelming majority of the healthcare budgets around the globe are spent on responding to acute problems, urgent needs of patients, and pressing concerns. Most preventive healthcare programs aim at improving the efficiency of the regional healthcare systems with limited resources rather than making major investments into increasing the resource infrastructure. Therefore, the policies to increase the number of people receiving preventive services has been an integral part of many reform programs. For example, the Montreal breast cancer screening program studied in Section 6 involved accrediting some of the existing mammography facilities in the city rather than building new ones. The accreditation criterion is incorporated in the model via the minimum workload requirement, which ensures a certain quality of care at each accredited preventive care facility. As originally noted by Verter and Lapierre (2002), a radiologist at each accredited facility needs to read a certain number of mammograms in order

not to lose their skills. In 1998, this minimum number was established as 4000 mammograms per year by the Quebec Ministry of Health. During the same time period, the U.S. Food and Drug Administration required a radiologist to interpret at least 960 mammograms and a radiology technician to perform at least 200 mammograms in 24 months to retain their accreditation (U.S. Food and Drug Administration, 1999). Note that when there is more than one mammography machine at a facility, it is very common to have a single radiologist on staff who reads all the collected images (which applies to all centers in Montreal). Hence, the minimum workload requirement applies to each facility rather than each server in the model. Although we do not consider fixed setup costs in the article, we investigate the optimal resource allocation strategy. To this end, the total number of servers to be allocated to open facilities is fixed. Therefore, the number of facilities that could be opened is limited, mainly due to the minimum workload requirement as well as the total number of available servers.

Let $G = (N, L)$ be a network with a set of nodes $N(|N| = n)$ and a set of links L. The nodes represent the neighborhoods of a city or the population zones, and the links are the main transportation arteries. The fraction of clients residing at node i is denoted by h_i. We assume that the number of clients who require preventive care over the entire network is Poisson distributed with a rate of λ per unit of time, and thus from each node i at a rate λh_i. We assume that there is a finite set of potential locations $M(|M| = m)$ in G for facilities. Let $S \subset M$ be a set of open facilities.

We assume that there are $Q_{\max}$ available servers and each can provide an average of μ services per unit of time, and one or more servers can be allocated to each open facility. As in Gunes *et al.* (2004), we assume that the service time is exponentially distributed.

We define three sets of decision variables:

$$y_j = \begin{cases} 1 \text{ if facility is located at node } j, \\ 0 \text{ otherwise;} \end{cases}$$

$s_j =$ number of servers at facility j;

$x_{ij} =$ fraction of clients from population node i who request service from facility j.

Denote by a_j the arrival rate of clients at facility j; that is,

$$a_j = \lambda \sum_{i=1}^{n} h_i x_{ij}, \quad j \in S. \tag{1}$$

Note that a facility cannot be established at node j unless a_j exceeds a minimum workload requirement denoted by $R_{\min}$. This requirement may apply to the whole facility or to each server, depending on the type of preventive care being offered. Here, we assume that a minimum workload is required for each facility.

We also denote by $\bar{T}_{ij}$ the average total time that clients from node i spend in order to receive service at facility j. The average total time $\bar{T}_{ij}$ is composed of two components: (i) the travel time from node i to facility j through the shortest path denoted by t_{ij}; and (ii) the average time clients spend at the facility possibly waiting and receiving service, which we denote by $\bar{W}(a_j, s_j)$; that is,

$$\bar{T}_{ij} = t_{ij} + \bar{W}(a_j, s_j), \quad i \in N, \quad j \in S. \tag{2}$$

Representing facility j as an $M/M/s_j$ queue, the general formula for the mean waiting time is (Kleinrock, 1975):

$$\bar{W}(a_j, s_j) = \frac{C(s_j, u_j)}{s_j} \frac{1}{\mu(1-\rho_j)} + \frac{1}{\mu}, \qquad j \in S, \tag{3}$$

where

$$u_j = \frac{a_j}{\mu}, \quad \rho_j = \frac{a_j}{s_j \mu},$$

$$C(s_j, u_j) = \frac{1 - K(u_j)}{1 - \rho_j K(u_j)},$$

$$K(u_j) = \frac{\sum_{l=0}^{s_j - 1} {u_j}^l / l!}{\sum_{l=0}^{s_j} {u_j}^l / l!}.$$

Denote the total participation rate (fraction) at node i by p_i; that is,

$$p_i = \sum_{j \in S} x_{ij}, \quad i \in N. \tag{4}$$

As we assume that clients choose the facility with the minimum expected total time, denote by T_i this shortest time incurred by clients at node i. We assume that the total participation rate p_i at node i is a decreasing function of the shortest time T_i (participation function). For simplicity, we assume that it is a linear function with an intercept A_i and a slope γ (although it can be generalized to other decreasing functions, such as an exponential function); that is,

$$p_i(T_i) = A_i - \gamma T_i, \quad i \in N. \tag{5}$$

Denote by $T_i(p_i)$ the inverse participation function; that is,

$$T_i(p_i) = \frac{A_i - p_i}{\gamma}, \quad i \in N. \tag{6}$$

In fact, $T_i(p_i)$ represents a threshold time; i.e., the total time clients at node i are willing to incur in order to participate in the service, while the actual time incurred by clients at node i to facility j is $\bar{T}_{ij}$.

As mentioned earlier, given facility locations y_j and the number of servers at each open facility s_j, the lower level problem involves the clients' facility choices so as to minimize their expected total time. This is a user equilibrium problem, and at equilibrium, no client wants to change her choice. This equilibrium condition can be stated as: given S and s_j, $j \in S$, for all pairs of (i, j), $i \in N$, $j \in S$:

$$\bar{T}_{ij} = t_{ij} + \bar{W}(a_j^*, s_j) \begin{cases} = T_i(p_i^*) & \text{if } x_{ij}^* > 0, \\ \geq T_i(p_i^*) & \text{if } x_{ij}^* = 0, \end{cases} \tag{7}$$

Equation (7) states that if there is a flow of clients from node i to facility j, then the actual time incurred by clients at node i to facility j must be equal to the threshold time (the longest time that clients would accept to go to the facility); and if the actual time exceeds the threshold time, there is no flow.

To find x_{ij}^* in Equation (7), we have to solve the following non-linear complementarity problem:

$$\begin{aligned} &t_{ij} + \bar{W}(a_j, s_j) - T_i(p_i) \geq 0, && i \in N, \quad j \in S, \\ &x_{ij}[t_{ij} + \bar{W}(a_j, s_j) - T_i(p_i)] = 0, && i \in N, \quad j \in S, \\ &x_{ij} \geq 0, && i \in N, \quad j \in S. \end{aligned} \tag{8}$$

Note that we expect that $p_i \leq A_i, i \in N$, and the stability of the queue $a_j < s_j\mu$, $j \in S$, can be naturally satisfied in Equation (8).

Alternatively, using vector–matrix notation, we can rewrite the complementarity problem (8) as a variational inequality problem. Group y_j, s_j, x_{ij}, t_{ij}, a_j, p_i, $\bar{W}(a_j, s_j)$, and $T_i(p_i)$, $i \in N$, $j \in M$, respectively, into column vectors $\mathbf{y} \subset R^m$, $\mathbf{s} \subset R^m$, $\mathbf{x} \subset R^{mn}$, $\mathbf{t} \subset R^{mn}$, $\mathbf{a} \subset R^m$, $\mathbf{p} \subset R^n$, $\bar{\mathbf{W}}(\mathbf{a}, \mathbf{s}) \subset R^m$, and $\mathbf{T}(\mathbf{p}) \subset R^n$. Then, given $\mathbf{y}$ and $\mathbf{s}$, the variational inequality problem is to find a vector $\mathbf{x}^* \in \mathbf{X}(\mathbf{y}) \subset R^{mn}$ such that:

$$\begin{aligned} &\langle \mathbf{t}, \mathbf{x}^* - \mathbf{x} \rangle + \langle \bar{\mathbf{W}}(\mathbf{a}^*, \mathbf{s}), \mathbf{a}^* - \mathbf{a} \rangle - \langle \mathbf{T}(\mathbf{p}^*), \quad \mathbf{p}^* - \mathbf{p} \rangle \leq 0, \\ &\quad \forall \mathbf{x} \in \mathbf{X}(\mathbf{y}), \end{aligned} \tag{9}$$

where the feasible set $\mathbf{X}(\mathbf{y})$ is defined as

$$\begin{aligned} \mathbf{X}(\mathbf{y}) &= \{\mathbf{x} : x_{ij} \geq 0, j \in S; x_{ij} = 0, j \in M - S\}, \\ S &= \{j : y_j = 1\}, \end{aligned}$$

and $\langle \cdot, \cdot \rangle$ represents the inner product (i.e., for two column vectors $\mathbf{x}$ and $\mathbf{y}$, $\langle \mathbf{x}, \mathbf{y} \rangle = \sum_i x_i y_i$). We note that since both $\mathbf{a}$ and $\mathbf{p}$ can be expressed by $\mathbf{x}$ as in Equations (1) and (4), whose elements x_{ij} indeed are the only variables of the variational inequality problem (9). For the ease of exposition, we keep $\mathbf{a}$ and $\mathbf{p}$ in the following expressions.

Theorem 1 proves that given $\mathbf{y}$ and $\mathbf{s}$, the solution to the variational inequality problem (9) is the equilibrium of this problem (7).

Theorem 1. *Given* $\mathbf{y}$ *and* $\mathbf{s}$, *the flow pattern* $\mathbf{x}^* \in \mathbf{X}(\mathbf{y})$ *is in equilibrium if, and only if, it satisfies the variational inequality problem* (9).

The proof for Theorem 1 is given in Appendix 1.

Therefore, the bilevel formulation of our preventive healthcare facility network design problem is

$$\max_{\mathbf{y},\mathbf{s},\mathbf{x}} \lambda \sum_{i=1}^{n} h_i \sum_{j=1}^{m} x_{ij}, \tag{10}$$

subject to

$$s_j \geq y_j, \qquad j \in M, \tag{11}$$

$$\sum_{j=1}^{m} s_j = Q_{\max}, \tag{12}$$

$$\lambda \sum_{i=1}^{n} h_i x_{ij} \geq R_{\min} y_j, \quad j \in M \tag{13}$$

$$y_j \in \{0, 1\}, \quad s_j = \text{Integer}, \quad j \in M, \tag{14}$$

$$x_{ij} \geq 0, \quad i \in N, \quad j \in M, \tag{15}$$

$$t'_{ij} = t_{ij} + B(1 - y_j), \quad i \in N, \quad j \in M, \tag{16}$$

$$\begin{aligned} &\langle \mathbf{t}', \mathbf{x} - \mathbf{x}' \rangle + \langle \bar{\mathbf{W}}(\mathbf{a}, \mathbf{s}), \mathbf{a} - \mathbf{a}' \rangle - \langle \mathbf{T}(\mathbf{p}), \mathbf{p} - \mathbf{p}' \rangle \leq 0, \\ &\quad \forall \mathbf{x}' \geq 0, \end{aligned} \tag{17}$$

where in Equation (17), $\mathbf{t}' \subset R^{mn}$ is a column vector associated with $t'_{ij}, i \in N, j \in M$, and $\mathbf{x}' \subset R^{mn}, \mathbf{a}' \subset R^m, \mathbf{p}' \subset R^n$ are column vectors, in which

$$\begin{aligned} a'_j &= \lambda \sum_{i=1}^{n} h_i x'_{ij}, \quad j \in M, \\ p'_i &= \sum_{j=1}^{m} x'_{ij}, \quad i \in N, \end{aligned}$$

and $\bar{W}(a_j, s_j)$ and $T_i(p_i)$ are expressed in Equation (3) and Equation (6), respectively.

The objective (10) is to maximize the level of total participation. Constraints (11) ensure that at least one server will be assigned to each open facility. Constraint (12) limits the total available servers to $Q_{\max}$. Constraints (13) guarantee the minimum workload requirement at each open facility. Constraints (16), where B represents a big number, stipulate that clients can obtain service only from open facilities; i.e., the travel time to a location where there is no open facility is set to a large number. Constraints (16) and (17) together are equivalent to Equation (9). Again, the constraints $p_i \leq A_i, i \in N$, and the stability of the queue $a_j < s_j\mu$, $j \in M$, can be naturally satisfied in Equation (17).

Since the variational inequality in the model is highly non-linear and the decision variables of the upper level problem are binary or integer, the entire problem is extremely difficult to solve. Thus, the focus of the study is on developing an efficient heuristic.

4. Solution methodology

Our solution approach is based on the location–allocation framework:

Allocation (Alloc P): the lower level problem: Given a set of facility locations and the associated capacities, identify equilibrium flows of clients to the facilities.
Location (Loc P): the upper level problem: Determine the best set of locations and and the associated capacities.

In the algorithm, (Alloc P) serves as a sub-routine for (Loc P) for any set of locations, with (Alloc P) the number of participants at each location and the objective function value can be determined. The next section provides an exact solution algorithm for (Alloc P) for $s_j \leq 2$. Section 4.2 presents an approximation method to solve the allocation problem for $s_j > 2$. A Tabu search procedure for (Loc P) is developed in Section 4.3.

4.1. *Allocation algorithm for $s_j \leq 2$*

For the ease of exposition, we first introduce a solution algorithm for $s_j = 1$, i.e., we drop capacity decisions, and then we extend the algorithm to solve the up-to-two-server case; i.e., $s_j \leq 2$.

When $s_j = 1$, each open facility becomes an $M/M/1$ queueing system. Thus, the mean waiting time in Equation (3) reduces to

$$\bar{W}(a_j) = \frac{1}{\mu - a_j}, \quad j \in S, \tag{18}$$

and the variational inequality (17) reduces to

$$\langle \mathbf{t}', \mathbf{x} - \mathbf{x}' \rangle + \langle \bar{\mathbf{W}}(\mathbf{a}), \mathbf{a} - \mathbf{a}' \rangle - \langle \mathbf{T}(\mathbf{p}), \mathbf{p} - \mathbf{p}' \rangle \leq 0, \quad \forall \mathbf{x}' \geq 0, \tag{19}$$

where $\bar{W}(a_j)$ and $T_i(p_i)$ are expressed in Equation (18) and Equation (6), respectively.

According to Nagurney (1999), there are several approaches to solve a variational inequality. One of them is to reformulate it as an optimization problem, which can be solved by a variety of optimization methods. According to Theorem 1.1 in Nagurney (1999), for a general variational inequality problem VI($\mathbf{F}(\mathbf{x})$,$\mathbf{X}$), if $\mathbf{F}(\mathbf{x})$ is continuously differentiable on $\mathbf{X}$ and the Jacobian matrix $\nabla \mathbf{F}(\mathbf{x})$ is symmetric and positive semi-definite, then solving this VI($\mathbf{F}(\mathbf{x})$,$\mathbf{X}$) is equivalent to solving a convex optimization problem defined as

$$\begin{aligned} &\min f(\mathbf{x}), \\ &\text{subject to} \\ &\mathbf{x} \in \mathbf{X}, \end{aligned} \tag{20}$$

where $\nabla f(\mathbf{x}) = \mathbf{F}(\mathbf{x})$.

We employ this method to solve our user equilibrium problem (Alloc P). For the sake of brevity, let $z_{ij} = \lambda h_i x_{ij}$, representing the actual number of clients traveling from node i to facility j. Group z_{ij} into a column vector $\mathbf{z} \subset R^{mn}$. Replacing $\mathbf{a}$, $\mathbf{p}$, and $\mathbf{x}$ by $\mathbf{z}$, Equation (19) can be reformulated as

$$\langle \mathbf{F}(\mathbf{z}), \mathbf{z} - \mathbf{z}' \rangle \leq 0, \quad \forall \mathbf{z}' \in \mathbf{Z}, \tag{21}$$

where $\mathbf{Z} = \{\mathbf{z} : \mathbf{z} \geq 0\}$ and

$$F_{ij}(\mathbf{z}) = \frac{1}{\mu - \sum_{i=1}^{n} z_{ij}} + t'_{ij} - \frac{A_i}{\gamma} + \frac{\sum_{j=1}^{m} z_{ij}}{\lambda h_i \gamma}, \quad i \in N, \quad j \in M. \tag{22}$$

It is obvious that $\mathbf{F}(\mathbf{z})$ is continuously differentiable on $\mathbf{Z}$, and in Theorem 2, we show that the Jacobian matrix $\nabla \mathbf{F}(\mathbf{z})$ can satisfy the above conditions.

Theorem 1. *The Jacobian matrix $\nabla \mathbf{F}(\mathbf{z})$ is symmetric and positive semi-definite.*

The proof for Theorem 2 is given in Appendix 1.

Therefore, we can reformulate our variational inequality problem as a convex optimization problem (recall that $\nabla f(\mathbf{z}) = \mathbf{F}(\mathbf{z})$):

$$\min f(\mathbf{z}) = -\sum_{j=1}^{m} \ln\left(\mu - \sum_{i=1}^{n} z_{ij}\right) + \sum_{i=1}^{n}\sum_{j=1}^{m}\left(t'_{ij} - \frac{A_i}{\gamma}\right) z_{ij} + \sum_{i=1}^{n} \frac{1}{2\lambda h_i \gamma}\left(\sum_{j=1}^{m} z_{ij}\right)^2,$$

subject to

$$z_{ij} \geq 0, \quad i \in N, \quad j \in M. \tag{23}$$

Since this is almost an unconstrained problem, we use the gradient projection method (Kelley, 1999) to solve this convex optimization problem. According to Theorem 1.4 in Nagurney (1999), the existence of equilibrium can be guaranteed. Since $\nabla \mathbf{F}(\mathbf{z})$ may not be positive definite, the equilibrium in general may not be unique. One simple example with multiple equilibrium solutions is a network with two client zones and two facilities, where all the parameters are symmetric. However, in practice, almost all of the cases we face have precisely one equilibrium.

We attempt to generalize the above solution methodology to solve the general case. Unfortunately, due to the complex mean waiting time formula for an $M/M/C$ queue (3), we can only prove that the above solution methodology can be applied when $s_j \leq 2$. This is primarily because the mean waiting time formula for an $M/M/2$ queue can be written as

$$\bar{W}(a_j) = \frac{1}{2\mu - a_j} + \frac{1}{2\mu + a_j}, \tag{24}$$

which means that there are two additive items, each of which has the same structure as the one in Equation (18). Therefore, $\mathbf{F}(\mathbf{z})$ in Equation (21) is still continuously differentiable, and its Jacobian matrix $\nabla \mathbf{F}(\mathbf{z})$ is still symmetric and positive semi-definite. For the sake of brevity, we omit the proof, which is quite easy to prove. In contrast, when $s_j > 2$, the Jacobian matrix $\nabla \mathbf{F}(\mathbf{z})$ may not be symmetric and positive semi-definite. As for the case with $s_j = 1$, the variational inequality (17) for $s_j \leq 2$ can be transformed to a convex optimization problem and solved by the gradient projection method.

4.2. *Allocation algorithm for $s_j > 2$*

As mentioned earlier, the complex mean waiting time formula for an $M/M/C$ queue (3) makes the method hard to

be directly generalized to solve the variational inequality when $s_j > 2$. One direction to overcome this difficulty is to replace this exact formula by an approximation. We investigated several approximations, such as in Kontogiorgis and Tibbs (2005), and we found that all of them are all too complicated to be applied here.

Motivated by the mean waiting time formula for an $M/M/2$ queue (24), we develop an approximation for the mean waiting time formula for an $M/M/C$ queue as follows:

$$\tilde{W}(\tau, C) = \frac{a(\mu, C)}{C\mu - \tau} + \frac{b(\mu, C)}{C\mu + d(\mu, C)\tau}, \qquad (25)$$

where τ denotes the arrival rate at a facility, and

$$a(\mu, C), b(\mu, C), d(\mu, C) = \arg\min \sum_{0<\tau_i<C\mu} [\bar{W}(\tau_i, C) - \tilde{W}(\tau_i, C)]^2, \quad C \geq 3. \qquad (26)$$

The idea of this method is to select a number of discrete points τ_i between zero and $C\mu$ and then to choose $a(\mu, C)$ $b(\mu, C)$, and $d(\mu, C)$ to minimize the squared sum of the errors between the exact values and the approximate values at these discrete points, just as in a least square estimation.

Note that this approximation may not be very accurate when the arrival rate at a certain facility τ is very close to zero or the total capacity $C\mu$. However, for our facility network design problem, since at equilibrium τ is typically not close to the two ends, this approximation formula works very well. Several instances with the performance of the approximation are shown in Appendix 2, in which the approximation error is usually within 2%. The performance of the approximation can be improved a lot if the range of τ_i is chosen around the equilibrium arrival rate, which usually can be roughly estimated in advance.

Clearly, this simple yet accurate approximation formula has the same structure as the one for an $M/M/2$ queue, and thus the solution methodology developed in Section 4.1 is applicable.

4.3. *Location algorithm*

We develop a Tabu search procedure to solve the upper level problem (Loc P). Tabu search (Glover, 1986) is one of the most successful metaheuristics, and it is designed so as to avoid local optima and instead explore other regions of the solution space.

Suppose that M is the set of potential locations for a facility and $s_{\max}$ is the maximum number of servers allowed to be allocated to each facility. Define $SM = \{M, \ldots, M\}$ as the set of potential locations for an individual server; i.e., cloning the potential facility set $s_{\max}$ times. In other words, we divide each potential facility into $s_{\max}$ pseudo-facilities. We call a pseudo-facility open if a server is allocated to it.

The Tabu search heuristic starts from a given initial solution with $Q_{\max}$ servers. Then, each iteration of the Tabu search focuses on a "neighborhood" of the current solution. We define two types of neighborhood moves: "Remove" and "Add." The main reason why we do not use "Exchange" is that a very large number of neighborhood moves need to be evaluated in each iteration for this type. "Remove" results from removing a pseudo-facility from the set of open pseudo-facilities, which leads to the smallest decrease in the overall participation rate; "Add" results from adding an additional pseudo-facility to the set of open pseudo-facilities, which leads to the largest improvement in the overall participation rate. In each iteration, "Remove" and "Add" are executed successively. The procedure repeats until no feasible solution that improves the objective function can be identified within a given number of iterations denoted by N_{ite}. Eventually, the heuristic outputs the best feasible solution found so far.

When moves are selected, Tabu restrictions are used to prevent moving back to previously investigated solutions. In this article, we define a Tabu list in which each value is associated with a pseudo-facility to represent its Tabu status. Once removed or added to the set of open pseudo-facilities, a node is classified as Tabu with a length equal to T_{len}, which represents the number of iterations in which the node typically will not be selected for removing or adding. However, even for a Tabu node, it can still be selected for adding, if an aspiration criteria is satisfied. We use the typical criterion which states that if a move produces a feasible solution that is better than the best known feasible solution, then the Tabu status is disregarded and the move is executed.

5. Computational experiments

The purpose of this section is to examine the computational performance of the solution approach. The gradient projection method for solving the convex optimization problem and the Tabu search algorithm for solving the upper level problem were coded in C. All runs were performed on a Pentium IV PC with 3.2 GHz of CPU and 1 GB of RAM.

For the computational experiments, the number of potential facilities (m) was set to 10, 20 and 40, and the number of population zones (n) was set to 100, 200, and 400. In total, there were nine problem sets. In each problem set, ten instances were generated. For each instance, the demand rate at each zone (λh_i) was randomly generated in the interval $[0, 2.4(m/n)]$ per hour. The travel times were randomly generated in the interval [0, 1.25] hours, and the following parameter values were used.

Problem parameters:

(a) the service rate at each facility: $\mu = 2.5$ clients/hour;
(b) the intercept of the demand decay function: $A_i = 1$;

Table 1. The computational performance

			CPU time (seconds)	
Number of facilities	*Number of zones*	*Improvement (%)*	*Zhang et al. (2009)*	*Current*
10	100	7.360	0.496	32.015
	200	10.140	1.369	45.437
	400	6.402	2.541	68.534
20	100	15.379	2.564	346.959
	200	19.101	11.975	547.529
	400	19.947	23.516	916.544
40	100	22.353	16.344	2194.481
	200	23.529	39.764	4629.255
	400	22.686	72.894	16 355.832

(c) the slope of the demand decay function: $\gamma = 0.4$;
(d) the minimum workload requirement: $R_{\min} = 1.2$ clients/hour.

Tabu search parameters:

(a) the stopping criteria: $N_{\text{ite}} = 50$;
(b) the Tabu length: $T_{\text{len}} = 5$.

To better illustrate the performance of the proposed algorithm, we used Zhang *et al.* (2009) as a benchmark. Note that they assume that all participants from a population zone go to the same facility and use a heuristic procedure for client–facility allocation. For comparability, the number of total available servers $Q_{\max}$ in our model was determined by solving the model with $s_j = 1$ and relaxing constraints (12) for each instance. We compare the current model for the up-to-two-server case ($s_j \leq 2$) with Zhang *et al.* (2009) in terms of total participation level and computational time requirement. We note that the number of open facilities is not the same in the two models and the new model allows assigning more than one server to a facility. Therefore, the presented comparison favors the new model to some extent.

Table 1 reports the average "Improvement" in total participation and the "CPU time" of the ten instances for each problem set. In general, the total participation levels achieved by the current model are larger than that of Zhang *et al.* (2009). Note that the average improvement is increasing with the number of potential facilities, surpassing 20% for the largest problem instances. This improvement comes with a price of increased computational time. Table 1 shows that the average CPU time of the current model increases with the number of potential facilities and the number of zones, especially the former. The running time of our model is significantly longer than that of Zhang *et al.* (2009). In particular, we found that our exact allocation procedure requires much longer running time than the allocation heuristic used in their paper. For example, for the largest-sized instances, the convergence time of each allocation procedure requires several seconds, over 200 times that of the allocation heuristic.

There are two main reasons for the improvement in the objective function value. First, we find that the current model often utilizes a larger number of servers, which increases the accessibility of the facilities and hence total participation. The allocation heuristic and the single-sourcing assumption in Zhang *et al.* (2009) lead to larger variation among the number of clients patronizing each facility. Consequently, the number of facilities that satisfy the minimum workload requirement and hence the number of servers that can be allocated to the entire network is less than the number in the current model. The second reason for the improvement is capacity pooling; i.e., allocating two servers at some of the open facilities. Interestingly, detailed analysis of our results indicates that the improvement due to capacity pooling does not monotonically increase with the number of potential facilities. This suggests that the improvement trend we observe in Table 1 is mainly due to the effectiveness of our exact allocation method in tackling the larger problem instances.

Since the above computational experiments are only based on one parameter set, it may not be sufficient to provide enough generality of the problem. Therefore, we conducted sensitivity analyses based on the three important problem parameters: μ, γ, and $R_{\min}$. Since we can observe from Table 1 that the number of zones does not play a significant role, only the problem sets with 100 zones were chosen for the sensitivity analyses. The value of each parameter in the base case was set to either increase by 50% or decrease by 50%, and the value of $Q_{\max}$ remained the same as in the base case for each instance. Our purpose was to see the average changes in the objective function value as well as in the CPU running time compared to the base case, which are shown in Tables 2 and 3, respectively.

Note that decreasing μ, increasing γ, and increasing $R_{\min}$ may make the problem infeasible. For each instance of the base case problem, since the value of $Q_{\max}$ is determined by

Table 2. The change in objective function value (in percentage) with respect to the three parameters

	μ		γ		R_{min}	
Number of facilities	*1.25*	*3.75*	*0.2*	*0.6*	*0.6*	*1.8*
10	70.609	109.791	120.952	79.653	101.123	100.000
20	55.433	117.493	119.591	82.336	100.239	99.663
40	57.008	117.442	118.385	84.448	100.068	99.809

solving the model with $s_j = 1$ and relaxing constraints (12), this always makes the base case problem feasible. However, in either of the studied three cases, it may not be possible to find enough open facilities that satisfy the minimum workload requirement, especially when $Q_{\max}$ is an odd number. Thus, the values shown in Tables 2 and 3 are calculated by excluding these infeasible instances.

Both tables suggest that decreasing μ by 50% has the largest effect. When there are sufficient potential clients, capacity becomes the primary constraint for participation. Thus, the average objective function value declines most. Moreover, an interesting finding is that the running time also declines significantly. We note that the number of iterations during the Tabu search does not change much, whereas the running time for each allocation procedure decreases significantly. In other words, the convergence to the allocation equilibrium by the gradient projection method becomes much faster. In contrast, although increasing γ by 50% (i.e., raising people's sensitivity to time) also leads to reduced participation, participation only declines around 20%.

Second, the two tables show that increasing μ by 50% and decreasing γ by 50% seem to result in similar effects. With regards to the objective function value, participation increases around 20% in both cases. For the former case, this is because the travel time remains the same, whereas the mean waiting time declines. For the latter case, although more potential clients are willing to participate, capacity becomes the constraint. In both cases, more CPU running time is required on average. Again, this is mainly due to the convergence speed of the allocation procedure.

Finally, both increasing and decreasing $R_{\min}$ by 50% do not have a significant impact on either the objective function value or the running time. The primary reason for this is that, in the base case solutions, most of the open facilities already have two servers. Thus, for each instance, increasing or decreasing $R_{\min}$ does not make the optimal solution change much, as long as the problem is still feasible.

6. An illustrative case

In this section, we use the data from the network of mammography centers in Montreal to illustrate the application of our model and methodology. The problem is motivated by the decision of the Quebec Ministry of Health to subsidize mammograms for women between the ages of 50 and 69. The National Cancer Institute of Canada reports that breast cancer is the most common cancer diagnosed in Canadian women and is second only to lung cancer as the most common cause of cancer deaths among women (National Cancer Institue of Canada, 2006). Accordingly, screening recommendations in Canada include a mammogram of the breast and clinical breast examination in asymptomatic women aged between 50 and 69 years of age. As of 1996, there were 194 475 women in Montreal in this age group and 36 facilities with mammography machines. The Ministry made a policy decision to require a minimum of 4000 mammographies per year for facilities to be accredited. The problem is determine which facilities to be accredited so as to maximize participation.

There are 497 population zones in Montreal, which we use for representing spatial distribution of the potential clients. Following Zhang *et al.* (2009), we assume that the maximum participation rate A_i is 0.95 and the slope of the demand function γ is 0.55; i.e., women will spend a maximum of 1.75 hours for a mammography. Assuming 250 days per year and 8 hours per day, the number of potential clients in Montreal per hour $\lambda = 194\,475/250/8 = 97.24$, and the minimum workload requirement $R_{\min} = 4000/250/8 = 2$ clients per hour. We also assume that the

Table 3. The change in CPU time (in percentage) with respect to the three parameters

	μ		γ		R_{min}	
Number of facilities	*1.25*	*3.75*	*0.2*	*0.6*	*0.6*	*1.8*
10	13.463	198.059	168.115	64.249	114.045	99.190
20	21.535	144.065	142.216	86.977	137.381	102.868
40	9.588	217.651	285.966	101.681	111.774	109.794

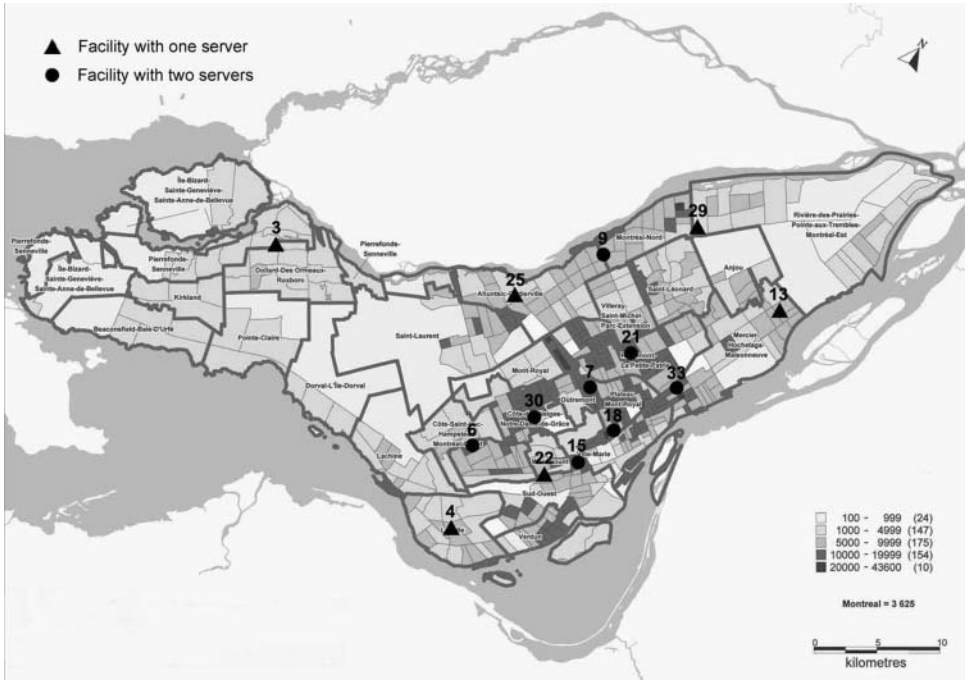

Fig. 1. The solution of the model with $s_j \leq 2$.

servers (machines) are homogeneous and set the capacity μ to five clients per hour.

Since a mammography unit machine is quite expensive, almost all of the current facilities in Montreal have only one or two machines. Therefore, we mainly focus on the case with $s_j \leq 2$ in this section, and we will briefly discuss the use of our approximation method for $s_j > 2$ at the end of the section. Also, when there is more than one machine, a single radiologist reads all the images in almost all cases. Hence, the minimum workload requirement should apply to each facility rather than each machine, and this is consistent with the formulation in Section 3.

As in Section 5, we determined the number of available servers $Q_{\max}$ by relaxing constraints (12) and solving the model with $s_j = 1$. We set $Q_{\max}$ equal to the resulting number of open facilities. The solution stipulates that 22 facilities should be accredited, and the overall participation would be 55.6%. Note that, with the same parameter setting, Zhang *et al*. (2009) suggested the accreditation of 21 facilities with an overall participation of 53.3%. That is, the accreditation of one more facility results in a 4.3% increase in the participation level over that for Zhang *et al*. (2009). This verifies our observation in the previous section that the use of the exact allocation method leads to less variation among the number of clients allocated to the accredited facilities and consequently more facilities can satisfy the minimum workload requirement.

We now turn to the model with $s_j \leq 2$ and $Q_{\max} = 22$. Figure 1 depicts the solution, which suggests that 14 facilities should be accredited and eight of these facilities should have two servers. The resulting participation is 58.8%, which is 5.8% higher than the solution of the model with $s_j = 1$. This can be interpreted as the benefit of capacity pooling at the mammography centers. Comparing the two solutions, we observe that some of the small facilities are merged so as to open larger ones; i.e., eight of the facilities in the $s_j = 1$ solution have to be closed so that their equipment can be utilized in the other facilities for capacity pooling. That is, increased total participation comes with a price of reduced spatial coverage under a predetermined number of servers.

To study this issue further, we conducted a parametric analysis on $Q_{\max}$. Table 4 and Fig. 2 depict the effect of increasing $Q_{\max}$ on the facility–server distribution and the participation level. In general, we observe that the larger the number of available servers, the more likely it is that the servers are centralized. Note that the number of facilities with two servers is increasing with $Q_{\max}$, whereas the

Table 4. Parametric analysis on $Q_{\max}$

	Number of facilities			
Number of servers	*Total*	*With one server*	*With two servers*	*Participation*
10	8	6	2	38.4
15	10	5	5	50.2
20	12	4	8	57.1
25	15	5	10	61.3
30	17	4	13	62.6
35	18	1	17	63.1
40	20	0	20	66.1

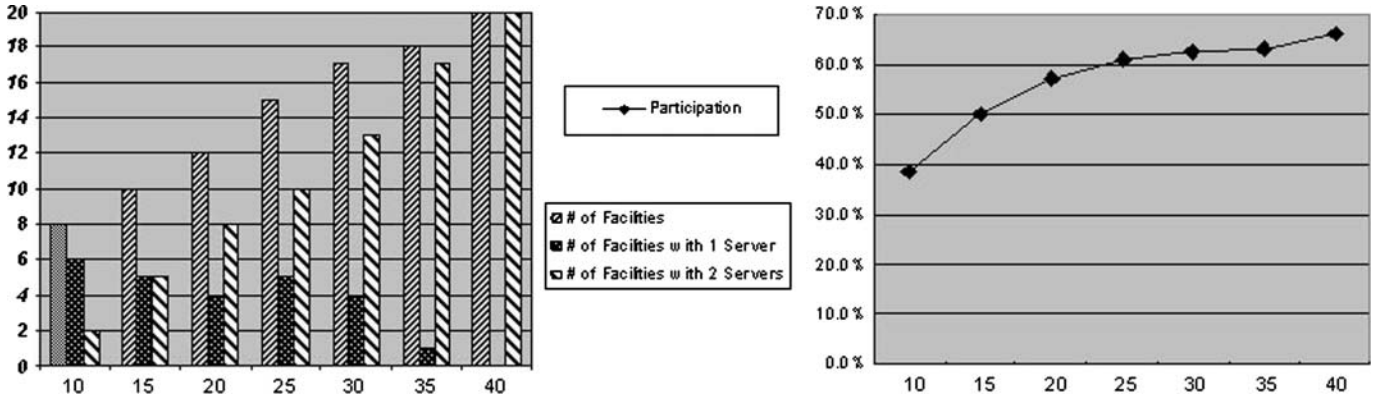

Fig. 2. Parametric analysis on $Q_{\max}$.

number of facilities with one server is decreasing. At the extreme, all the facilities have two servers when $Q_{\max} = 40$. We identify two main causes for this result. As mentioned earlier, capacity pooling may raise participation, by reducing the mean waiting time. Meanwhile, the minimum workload requirement may also favor centralization, since it may not be feasible to accredit many single-server facilities. In particular, we note that the impact of $R_{\min}$ increases with $Q_{\max}$.

Another main result is that with a medium or large number of total available servers, centralizing capacity at the facilities in high-density areas is a better strategy than decentralizing capacity to smaller facilities. In contrast, if there are only limited servers available (e.g., ten), instead of adding one additional server to an existing facility in a high-density area, it is better to locate a new facility in another high-density area so as to increase the spatial coverage. Figure 1 also shows the map of the population density of Montreal in 2001 (Ville de Montreal, 2008), from which we can clearly see that the facilities with two servers are located in the high-density areas and the facilities with one server are typically located in the medium-density areas.

Interestingly, during the computations, we found that it takes a very long time to find a feasible solution when $Q_{\max} = 35$. In fact, the final solution shows that there is one server at facility 3 and all the other facilities have two servers. However, the solutions with $Q_{\max}$ = 25, 30, and 40, all show that there are two servers allocated to facility 3. After investigation, we realized that it is very difficult to open a facility with a single server that can satisfy the minimum workload requirement at a place close to other facilities in a well-developed network. Therefore, in this case, only one server is allocated to facility 3, around which there is no other open facility, and this allows all the other 17 facilities with two servers to satisfy more easily the minimum workload requirement. In other words, if a number of additional servers can be added to a balanced network, it would be a good plan to allocate the servers to existing facilities or to open new facilities with more than one server.

In addition, Fig. 2 demonstrates that total participation is close to a concave function, increasing with $Q_{\max}$. Even with 40 servers, the participation rate is 66.1%, still below the national target of 70%. Therefore, it seems necessary to support capacity expansion projects with initiatives to increase people's awareness about the significance of preventive care (i.e., to reduce their sensitivity to the total time) in order to achieve the target participation levels.

Since the values of A_i and γ are not estimated based on real data or survey, we therefore conducted another parametric analysis on these two parameters. Table 5 shows the corresponding values of the objective function (participation) with different values of A_i and γ when $Q_{\max} = 22$. In particular, participation almost linearly increases with the increase in A_i, while it almost linearly decreases with the increase in γ, when the values of the two parameters are within those ranges. Moreover, by investigating the facility–server distribution, we find that the lower the participation level, the more likely the servers are centralized together. This is mainly due to the minimum workload requirement.

The above discussion is based on the case of $s_j \leq 2$. To study the impact of this restriction, we use the approximation method devised for the allocation problem. Figure 3 displays the solution of our model with $s_j \leq 3$ on the map of the population density of Montreal in 2001. In this solution, 11 facilities are accredited, three of them with three servers and five of them with two servers. The three facilities with three servers are all located in the very-high-density

Table 5. Participation levels (in percentage) as a function A_i and γ

	γ				
A_i	*0.35*	*0.45*	*0.55*	*0.65*	*0.75*
0.95	69.9	64.2	58.8	53.9	46.8
0.85	61.6	56.0	50.2	43.7	39.3
0.75	52.8	45.9	40.2	36.2	31.3

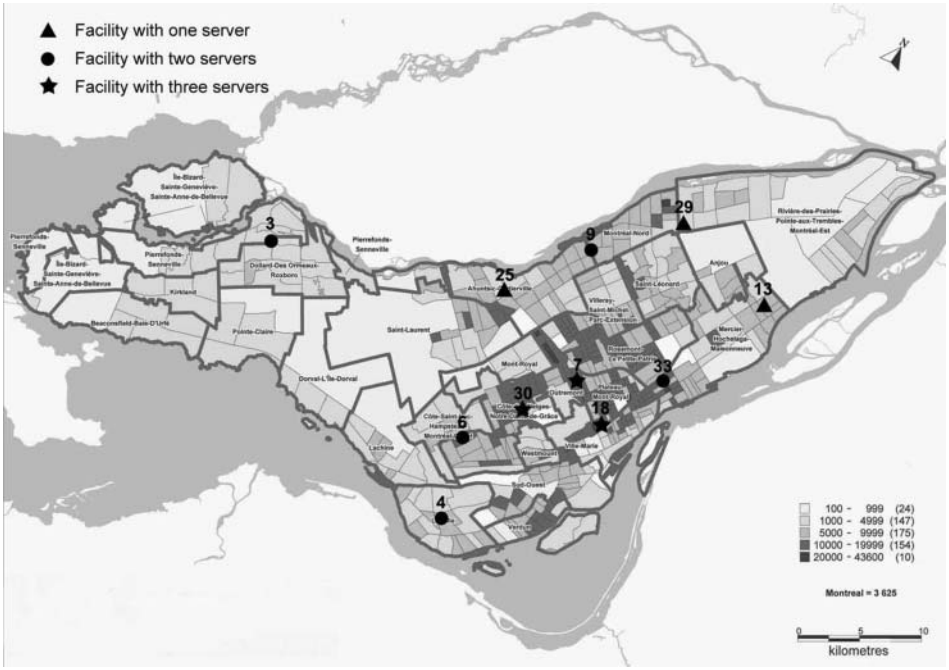

Fig. 3. The solution of the model with $s_j \leq 3$.

areas, and this supports our previous observation about the strategy of capacity pooling. The total participation is 59.5%; i.e., only a 1.2% improvement on the previous solution. This small increase suggests that the current policy of having up to two mammography machines at the facilities is reasonable, and it would not pay off to centralize capacity further.

7. Conclusions and future research

This article presents a model for the problem of preventive healthcare facility network design and provides a solution methodology. We formulate the problem as an MPEC; i.e., a bilevel non-linear optimization model. The lower level problem determines the allocation of clients to facilities and it is formulated as a variational inequality; the upper level is a facility location and capacity allocation problem. Our solution approach is based on the location–allocation framework. The variational inequality is formulated as a convex optimization problem, which can be solved by the gradient projection method; we develop a Tabu search procedure to solve the upper level problem. Our computational experiments indicate that large-sized instances can be solved in reasonable time. Based on the analysis of an illustrative case, the network of mammography centers in Montreal, we derive managerial insights with regards to the impact of capacity pooling on the level of participation and the trade-off between capacity pooling and spatial coverage provided by the facility network.

Our model can be generalized or extended in a number of ways. First, although the model aims at a non-appointment or "walk-in" system, it is possible to calibrate the model to represent the presence of an appointment system at each facility. Under an appointment system, one of the most significant client attraction factors is the waiting time for the appointment rather than the waiting time at the facility. Assuming that clients always take the first available appointment and there are no cancelations or no-shows, an $M/M/C$ queue can be used for approximating an appointment system. To calibrate the travel time and the waiting time for an appointment, we can replace Equation (2) by the following weighted time function,

$$\bar{T}_{ij} = \alpha t_{ij} + \bar{W}(a_j, s_j), \quad i \in N, \quad j \in S, \qquad (27)$$

where α is a constant to balance travel time and wait time and it can be estimated empirically. Also, the slope γ in Equation (5) needs to be calibrated accordingly. The other elements of the model remain the same. During preliminary experiments, however, we found out that the computational requirements of the allocation algorithm drastically increase as the participation function becomes flatter (corresponding to longer appointment waiting times due to high congestion). Thus, the development of an efficient allocation algorithm for dealing with preventive care facilities that use appointment systems remains a challenge for future research. Alternatively, the use of simulation for modeling an appointment system seems to be a promising research direction.

Second, the model presented here can also be applied to other service environments, such as in the context of networks of banks and post offices. The incorporation of congestion in modeling clients' decision making in these service sectors is also crucial. Although the detailed

formulations may be different, the overall structure of the bilevel model and especially the user equilibrium problem remain the same, and our solution methodology can be easily adapted for solving the arising problems.

Third, although we only use the accessibility of a facility (i.e., the total time spent in getting service) as a determinant of client attraction, other factors of attractiveness could be included as well. Fundamentally, these factors can be divided into two categories, static and dynamic. The static factors of attractiveness such as travel time, service time, facility type, and facility reputation typically will not be influenced by the allocation of clients. In contrast, dynamic factors, including the mean waiting time and the average number of clients in the system, depend on the collective outcome of the decisions of all clients, and the resulting user equilibrium needs to be determined. This article models a user equilibrium based on a single dynamic factor in the context of preventive healthcare facility network design. The incorporation of a user equilibrium with multiple dynamic factors needs further work.

Finally, this article studies the issue of capacity optimization from the perspective of allocating a given number of servers. Under the presence of technology alternatives, it may be relevant to optimize the service rate or the throughput at each facility, subject to a budget constraint. This extension constitutes a fruitful avenue for future research.

Acknowledgements

This research was supported in part by a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada (NSERC grant 183631). The authors acknowledge the comments and suggestions of the associate editor and two anonymous referees, which were helpful in improving the article.

References

Abdulaal, M. and LeBlanc, L.J. (1979) Continuous equilibrium network design models. *Transportation Research B*, **13**, 19–32.

Berman, O. and Krass, D. (2002) Facility location problems with stochastic demands and congestion, in *Facility Location: Applications and Theory*, Drezner, Z. and. Hamacher, H.W. (eds.), Springer, New York, NY, pp. 331–373.

Chao, X., Liu, L. and Zheng, S. (2003) Resource allocation in multisite service systems with intersite customer flows. *Management Science*, **49**(12), 1739–1752.

Dafermos, S.C. (1968) Traffic assignment and resource allocation in transportation networks. Ph.D. Dissertation, Johns Hopkins University, Baltimore, MD.

Daskin, M.S. and Dean, L.K. (2004) Location of health care facilities, in *Operations Research and Health Care: A Handbook of Methods and Applications*, Brandeau, M.L., Sainfort, F. and Pierskalla, W.P. (eds.), Kluwer, New York, NY, pp. 43–76.

Facione, N.C. (1999) Breast cancer screening in relation to access to health services. *Oncology Nursing Forum*, **26**, 689–696.

Florian, M. and Hearn, D. (1995) Network equilibrium models and algorithms, in *Handbooks in Operations Research and Management Science: Volume 8 Network Routing*, Elsevier, New York, NY, pp. 485–550.

Glover, F. (1986) Future paths for integer programming and links to artificial intelligence. *Computers & Operations Research*, **13**, 533–549.

Gornick, M.E., Eggers, P.W. and Riley, G.F. (2004) Associations of race, education, and patterns of preventive service use with stage of cancer at time of diagnosis. *Health Services Research*, **39**, 1403–1427.

Gunes, E.D., Chick, S.E. and Zeynep, A.O. (2004) Breast cancer screening services: trade-offs in quality, capacity, outreach, and centralization. *Health Care Management Science*, **7**, 291–303.

Health Canada. (2005) Mammography, available at http://www.hc-sc.gc.ca/hl-vs/iyh-vsv/med/mammog-eng.php, accessed August 19, 2010.

Judge, G.G. and Takayama, T. (1973) *Studies in Economic Planning over Space and Time*, North-Holland, Amsterdam, The Netherlands.

Kelley, C.T. (1999) *Iterative Methods for Optimization*, SIAM, Philadelphia, PA.

Kleinrock, L. (1975) *Queueing System I: Theory*, Wiley, New York, NY.

Kontogiorgis, S. and Tibbs, R.W. (2005) An efficient approximation to wait time in M/M/*c* queues with application to staffing planning, in *Proceedings of the 43rd Annual ACM Southeast Regional Conference*, *volume 2*, Association for Computing Machinery (ACM), New York, NY, pp. 98–102.

Magnanti, T.L. and Wong, R.L. (1984) Network design and transportation planning: models and algorithms. *Transportation Science*, **18**, 1–55.

Marcotte, P. (1986) Network design problem with congestion effects: a case of bilevel programming. *Mathematical Programming*, **34**, 142–162.

Marcotte, P., Marquis, G. and Zubieta, L. (1992) A Newton-SOR method for spatial price equilibrium. *Transportation Science*, **26**, 36–47.

Marianov, V. (2003) Location of multiple-server congestible facilities for maximizing expected demand, when services are non-essential. *Annals of Operations Research*, **123**, 125–141.

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

Marianov, V., Rios, M. and Icaza, M.J. (2008) Facility location for market capture when users rank facilities by shorter travel and waiting times. *European Journal of Operational Research*, **191**, 32–44.

Marianov, V. and Serra, D. (2002) Location problems in the public sector, in *Facility Location: Applications and Theory*, Drezner, Z. and Hamacher, H.W. (eds.), Springer, New York, NY, pp. 119–150.

McNoe, B., Richardson, A.K. and Elwood, J.M. (1996) Factors affecting participation in mammography screening. *New Zealand Medical Journal*, **109**, 359–362.

Nagurney, A. (1992) The application of variational inequality theory to the study of spatial equilibrium and disequilibrium, in *Readings in Econometric Theory and Practice: A Volume in Honor of George Judge*, Grinths, W.E., Lutkepohl, H. and Bock, M.E. (eds.), North-Holland, Amsterdam, The Netherlands, pp. 327–355.

Nagurney, A. (1999) *Network Economics: A Variational Inequality Approach*, Kluwer, Norwell, MA.

National Cancer Institute of Canada. (2006) Progress in cancer control: screening. Annual statistics special report, available at http://www.phac-aspc.gc.ca/publicat/prccc-relccc/pdf/F244_HC_Cancer_Rpt_English.pdf, accessed August 19, 2010.

Nolte, S. (2008) The future of the world suger market—a spatial price equilibrium analysis. Working paper, University of Ghent, Belgium.

Public Health Agency of Canada. (2006) Organized breast cancer screening programs in canada—report on program performance in 2001 and 2002, available at http://www.phac-aspc.gc.ca/publicat/obcsp-podcs01/index-ena.php, accessed August 19, 2010..

Samuelson, P.A. (1952) Spatial price equilibrium and linear programming. *American Economic Review*, **42**, 283–303.

Takayama, T. and Judge, G.G. (1964) An intertemporal price equilibrium model. *Journal of Farm Economics*, **46**, 477–484.
Takayama, T. and Judge, G.G. (1971) *Spatial and Temporal Price and Allocation Models*, North-Holland, Amsterdam, The Netherlands.
U.S. Food and Drug Administration. (1999) The Mammography Quality Standards Act Final Regulations, available at http://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm094441.pdf, accessed August 19, 2010.
Verter, V. and Lapierre, S.D. (2002) Location of preventive health care facilities. *Annals of Operations Research*, **110**, 123–132.
Ville de Montreal. 2008. Theme maps: sociodemographic and economic atlas for Montreal, Available at http://ville.montreal.qc.ca/portal/page?_pageid=2077, 2455180&_dad=portal&_Schema=PORTAL, accessed August 19, 2010.
Walker, K. 1977. Current issues in the provision of health care services. *Journal of Consumer Affairs*, **11**, 52–62.
Zhang, Y., Berman, O. and Verter, V. (2009) Incorporating congestion in preventive healthcare facility network design. *European Journal of Operational Research*, **198**, 922–935.
Zimmerman, S. (1997) Factors influencing hispanic participation in prostate cancer screening. *Oncology Nursing Forum*, **24**, 499–504.

Appendices

Appendix 1

Proof of Theorem 1. First, it is shown that if $\mathbf{x}^* \in \mathbf{X}(\mathbf{y})$ satisfies Equation (7) then it also satisfies Equation (9).

Note that given $\mathbf{s}$, for a fixed pair (i, j), $i \in N$, $j \in S$, one must have that:

$$(\bar{W}(a_j^*, s_j) + t_{ij} - T_i(p_i^*)) \times (x_{ij} - x_{ij}^*) \geq 0, \qquad \text{(A1)}$$

for any non-negative x_{ij}. Hence, summing over all pairs, one has that,

$$\sum_{i=1}^{n} \sum_{j \in S} (\bar{W}(a_j^*, s_j) + t_{ij} - T_i(p_i^*)) \times (x_{ij} - x_{ij}^*) \geq 0, \quad \forall x_{ij} \geq 0. \qquad \text{(A2)}$$

Using Equations (1), (4) and some algebra, Equation (A2) yields:

$$\sum_{i=1}^{n} \sum_{j \in S} t_{ij}(x_{ij} - x_{ij}^*) + \sum_{j \in S} \bar{W}(a_j^*, s_j)(a_j - a_j^*) - \sum_{i=1}^{n} T_i(p_i^*)(p_i - p_i^*) \geq 0, \quad \forall x_{ij} \geq 0, \qquad \text{(A3)}$$

which, in vector natation, gives us Equation (9).

Now it is shown that if $\mathbf{x}^* \in \mathbf{X}(\mathbf{y})$ satisfies Equation (9) then it also satisfies Equation (7). For simplicity, utilize Equation (9) expanded as Equation (A2). For any pair (k, l), let $x_{ij} = x_{ij}^*$, $\forall (i, j) \neq (k, l)$, $i \in N$, $j \in S$, and then Equation (A2) simplifies to

$$(\bar{W}(a_l^*, s_k) + t_{kl} - T_k(p_l^*)) \times (x_{kl} - x_{kl}^*) \geq 0, \qquad \text{(A4)}$$

from which Equation (7) follows for this (k, l) and consequently for every pair. ■

Proof of Theorem 2. The Jacobian matrix can be represented as $\nabla \mathbf{F}(\mathbf{z}) = \mathbf{JA} + \mathbf{JB}$, where $\mathbf{JA}$ and $\mathbf{JB}$, two $mn \times mn$ matrices, are defined as

$$\mathbf{JA} = \begin{bmatrix} \mathbf{JA}_1 & 0 & \cdots & 0 \\ 0 & \mathbf{JA}_2 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{JA}_n \end{bmatrix}, \quad \mathbf{JB} = \begin{bmatrix} \mathbf{JB}_1 & \mathbf{JB}_1 & \cdots & \mathbf{JB}_1 \\ \mathbf{JB}_1 & \mathbf{JB}_1 & \cdots & \mathbf{JB}_1 \\ \vdots & & \ddots & \vdots \\ \mathbf{JB}_1 & \mathbf{JB}_1 & \cdots & \mathbf{JB}_1 \end{bmatrix}.$$

$\mathbf{JA}_j$ is a $n \times n$ matrix in which all elements are equal to

$$\frac{1}{\left(\mu - \sum_{i=1}^{n} z_{ij}\right)^2},$$

and $\mathbf{JB}_1$ is a $n \times n$ diagonal matrix as

$$\mathbf{JB}_1 = \begin{bmatrix} \frac{1}{\lambda h_1 \gamma} & 0 & \cdots & 0 \\ 0 & \frac{1}{\lambda h_2 \gamma} & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\lambda h_m \gamma} \end{bmatrix}.$$

We can also obtain that:

$$\mathbf{v}^T \mathbf{JA} \mathbf{v} = \sum_{j=1}^{m} \left[\frac{1}{\left(\mu - \sum_{i=1}^{n} z_{ij}\right)^2} \left(\sum_{k=(j-1)n+1}^{jn} v_k \right)^2 \right] \geq 0 \quad \forall \mathbf{v} \in R^{mn},$$

$$\mathbf{v}^T \mathbf{JB} \mathbf{v} = \sum_{i=1}^{n} \left[\frac{1}{\lambda h_i \gamma} \left(\sum_{k=1}^{m} v_{(k-1)n+i} \right)^2 \right] \geq 0 \quad \forall \mathbf{v} \in R^{mn}.$$

Therefore, $\nabla \mathbf{F}(\mathbf{z})$ is symmetric and positive semi-definite, since $\mathbf{JA}$ and $\mathbf{JB}$ are both positive semi-definite. ■

Appendix 2
The performance of the approximation

Table A1. The performance of the approximation

	$\mu = 5, C = 4$ $a = 0.998, b = 3.159, d = 0.709$			$\mu = 2, C = 10$ $a = 0.983, b = 9.524, d = 0.393$			$\mu = 1, C = 20$ $a = 0.957, b = 19.854, d = 0.234$		
τ	*Exact*	*Approximation*	*Error (%)*	*Exact*	*Approximation*	*Error (%)*	*Exact*	*Approximation*	*Error (%)*
2.0	0.200	0.203	1.44	0.500	0.513	2.57	1.000	1.023	2.31
2.5	0.200	0.202	1.00	0.500	0.510	2.02	1.000	1.019	1.91
3.0	0.200	0.201	0.62	0.500	0.508	1.51	1.000	1.015	1.53
3.5	0.200	0.201	0.31	0.500	0.505	1.03	1.000	1.012	1.16
4.0	0.201	0.201	0.05	0.500	0.503	0.60	1.000	1.008	0.81
4.5	0.201	0.201	0.17	0.500	0.501	0.19	1.000	1.005	0.47
5.0	0.201	0.201	0.34	0.500	0.499	0.17	1.000	1.002	0.16
5.5	0.202	0.201	0.47	0.500	0.498	0.49	1.000	0.999	0.14
6.0	0.203	0.202	0.56	0.500	0.496	0.77	1.000	0.996	0.41
6.5	0.204	0.202	0.62	0.500	0.495	1.00	1.000	0.993	0.66
7.0	0.205	0.203	0.66	0.500	0.494	1.19	1.000	0.991	0.89
7.5	0.206	0.205	0.67	0.500	0.494	1.34	1.000	0.989	1.09
8.0	0.208	0.206	0.66	0.501	0.493	1.45	1.000	0.987	1.26
8.5	0.209	0.208	0.63	0.501	0.494	1.51	1.000	0.986	1.40
9.0	0.212	0.210	0.58	0.502	0.494	1.52	1.000	0.985	1.50
9.5	0.214	0.213	0.53	0.503	0.495	1.50	1.000	0.984	1.57
10.0	0.217	0.216	0.46	0.504	0.496	1.44	1.000	0.984	1.60
10.5	0.221	0.220	0.39	0.505	0.498	1.34	1.001	0.985	1.59
11.0	0.225	0.224	0.31	0.507	0.501	1.20	1.001	0.986	1.54
11.5	0.230	0.230	0.23	0.509	0.504	1.04	1.002	0.987	1.44
12.0	0.236	0.236	0.15	0.513	0.508	0.85	1.003	0.990	1.30
12.5	0.243	0.242	0.07	0.517	0.513	0.64	1.005	0.994	1.11
13.0	0.251	0.251	0.01	0.522	0.520	0.41	1.007	0.998	0.88
13.5	0.260	0.260	0.08	0.529	0.528	0.16	1.011	1.004	0.62
14.0	0.271	0.272	0.15	0.537	0.537	0.09	1.016	1.012	0.32
14.5	0.285	0.286	0.21	0.548	0.549	0.33	1.022	1.023	0.01
15.0	0.302	0.303	0.25	0.561	0.565	0.57	1.032	1.036	0.36
15.5	0.323	0.324	0.29	0.579	0.584	0.79	1.045	1.053	0.71
16.0	0.349	0.350	0.31	0.602	0.608	0.98	1.064	1.075	1.06
16.5	0.383	0.385	0.32	0.633	0.641	1.13	1.090	1.105	1.37
17.0	0.430	0.431	0.31	0.677	0.685	1.21	1.128	1.147	1.64
17.5	0.495	0.497	0.28	0.739	0.748	1.21	1.185	1.207	1.80
18.0	0.594	0.595	0.23	0.834	0.844	1.10	1.275	1.298	1.79
18.5	0.759	0.760	0.15	0.997	1.005	0.83	1.432	1.454	1.51
19.0	1.091	1.092	0.05	1.326	1.330	0.35	1.755	1.769	0.76
19.5	2.090	2.089	0.08	2.321	2.311	0.44	2.745	2.721	0.87

Biographies

Yue Zhang is an Assistant Professor at the College of Business Administration, the University of Toledo. He holds a Ph.D. degree in Operations Management from Desautels Faculty of Management, McGill University. He also holds a Master's degree in Management Science and Engineering and a Bachelor's degree in Management Information Systems from the School of Economics and Management, Tsinghua University in China. Prior to joining the University of Toledo, he was a Post-Doctoral Fellow at the Sauder School of Business, the University of British Columbia. His research interests include healthcare operations management, supply chain design, logistics, optimization and simulation.

Oded Berman is the endowed Sydney Cooper Chair in Business and Technology and a former Associate Dean of Programs at the Joseph L. Rotman School of Management, the University of Toronto. He received his Ph.D. (1978) in Operations Research from the Massachusetts Institute of Technology. He had been with the Electronic Systems Lab at MIT, the University of Calgary, and the University of Massachusetts at Boston, where he was also the Chairman of the Department of Management Sciences. He has published over 200 refereed articles and has contributed to several books in his field. His main research interests include operations management in the service industry, location theory, network models, and stochastic inventory control. He served as an Area Editor for *Operations Research* and as an Associate Editor for *Management Science*. He is

currently an Associate Editor for *Transportation Science* and a member of the Editorial Board for *Computers & Operations Research*.

Patrice Marcotte is currently chairman of the Computer Science and Operations Research Department of the University of Montreal. Author of more than 70 papers in international journals, his research is mainly concerned with the algorithmic aspects of convex and nonconvex programming, including variational inequalities and equilibrium programming, with a special focus on transportation planning.

Vedat Verter is a Professor of Operations Management at Desautels Faculty of Management, McGill University. He is also an Adjunct Professor at the University of Ottawa M.Sc. in Health Systems Program and MIT-Zaragoza International Logistics Program. His research focuses on supply chain design, hazardous materials logistics, sustainable supply chains, and healthcare operations management. His work in these four areas is well recognized through top-tier journal publications as well as invited presentations around the globe. He is the Founding President of the POMS College of Healthcare Operations Management. Recently, he has been appointed as director of a six-university Ph.D. program on healthcare that is funded by NSERC. He is the incoming Editor-in-Chief of *Socio-Economic Planning Sciences*.