Health Care Manag Sci (2015) 18:459–474
DOI 10.1007/s10729-014-9274-2

The impact of directed choice on the design of preventive healthcare facility network under congestion

Navneet Vidyarthi · Onur Kuzgunkaya

Received: 19 October 2013 / Accepted: 13 February 2014 / Published online: 31 May 2014
© Springer Science+Business Media New York 2014

Abstract Preventive healthcare (PH) programs and services aim at reducing the likelihood and severity of potentially life-threatening illness by early detection and prevention. The effectiveness of these programs depends on the participation level and the accessibility of the users to the facilities providing the services. Factors that impact the accessibility include the number, type, and location of the facilities as well as the assignment of the clients to these facilities. In this paper, we study the impact of system-optimal (i.e., directed) choice on the design of the preventive healthcare facility network under congestion. We present a model that simultaneously determines the location and the size of the facilities as well as the allocation of clients to these facilities so as to minimize the weighted sum of the total travel time and the congestion associated with waiting and service delay at the facilities. The problem is set up as a network of spatially distributed M/G/1 queues and formulated as a nonlinear mixed integer program. Using simple transformation of the nonlinear objective function and piecewise linear approximation, we reformulate the problem as a linear model. We present a cutting plane algorithm based exact (ϵ-optimal) solution approach. We analyze the tradeoff between travel time and queuing time and its impact on the location and capacity of the facilities as well as the allocation of clients to these facilities under a directed choice policy. We present a case study that deals with locating mammography clinics in Montreal, Canada. The results show that incorporating congestion in the PH facility network design substantially reduces the total time spent by clients. The proposed model allows policy makers to direct clients to facilities in an equitable manner resulting in better accessibility.

Keywords Preventive healthcare · Network design · Location-allocation · Capacity selection · Stochastic demand · Directed choice · Queueing · Congestion · ϵ-optimal · Cutting plane method

N. Vidyarthi (✉)
Department of Supply Chain and Business Technology Management, John Molson School of Business, Concordia University, 1455 de Maisonneuve Blvd. West, Montreal, QC H3G 1M8, Canada
e-mail: navneet.vidyarthi@gmail.com;navneetv@jmsb.concordia.ca

O. Kuzgunkaya
Department of Mechanical and Industrial Engineering, Faculty of Engineering, Concordia University, 1455 de Maisonneuve Blvd. West, Montreal, QC H3G 1M8, Canada
e-mail: onurk@encs.concordia.ca

1 Introduction

Preventive healthcare (PH) programs and services refer to preventive examination, disease-specific screening tests, and procedures based on evidence-based medical information that aim at reducing the likelihood and severity of potentially life-threatening illness by protection and early detection. Examples include mammograms, vaccination, flu shots, blood tests, and anti-smoking advice [1]. PH programs can be categorized into three groups with regard to their objectives: (i) primary, (ii) secondary, and (iii) tertiary. Primary PH programs aim at reducing the likelihood of diseases in people with no symptoms; for example, immunizations of healthy children. Secondary PH prevention aims at identifying and treating people who have risk factors or are at very early stage of diseases; for example, pap smears to detect early forms of cervical cancer. Tertiary prevention aims at treating symptomatic patients in an effort to

decrease complications or severity of disease; for example, sugar control in a diabetic in order to mitigate vision and nerve problems [2].

PH programs have resulted in better quality of life besides substantial savings in the costs of diagnosis and therapy as a result of decrease in the requirement for radical treatments, such as surgery or chemotherapy. For example, Health Canada reports that mammograms taken on a regular basis have the potential to decrease deaths from breast cancer for women between the ages of 50 and 69 by up to 40 %. Furthermore, 36 % of breast cancer patients without a mammogram received the diagnosis of late stage cancer [10]. Studies conducted by Gornick et al. [3] report that among breast cancer patients with a mammogram, of those patients with only one preventive service, 20 % had late stage breast cancer, while 17 % of those patients with two to five preventive services had late stage cancer. Among breast cancer patients without a mammogram, those patients with zero preventive services, 36 % had late stage breast cancer, while 24 % of those with two to five preventive services had late stage cancer. The World Health Organization (WHO) has also emphasized the need for PH programs and services. It has reported that many current healthcare systems are designed to respond to acute problems, urgent needs of patients, and pressing concerns, and hence do not make the best use of their available resources to support this process.

Unlike other healthcare programs and services, most of the PH programs are voluntary and hence it is up to the client to participate and avail themselves of the services. Furthermore, the effectiveness of these programs is impacted by the convenience of access to the facilities [4–6]. The convenience of access is influenced by the waiting time at the facilities, travel time/distance to the facilities, availability of timely and reliable transportation, weather and road conditions, among others. Empirical evidence suggests that the time spent waiting (or, the level of congestion) is a significant factor in a client's choice of facility, especially in preventive healthcare [7]. In Canada, for example, the waiting time for medical imaging diagnostics was reported to be between 30 and 160 days in 2011 [8]. In contrast, the nationwide benchmark was 4 weeks in 2012 [9]. Furthermore, clients do not have accurate information about the waiting times of individual facilities a priori to make the system-optimal choice. As a result, the users' choice of facilities (where clients travel to their closest facility to access the services), is often suboptimal, resulting in congestion and long wait times.

Motivated by the significance of accessibility and congestion at PH facilities, in this paper, we study the impact of a *system-optimal directed choice* model, where a central authority directs the clients to the facilities for the service as opposed to users choosing the facility to access. We show that the configuration of the PH facility network can impact the operational performance of PH services substantially, as measured by the waiting time and travel times of clients. The proposed model captures the relationship between facility location, capacity selection, and client allocation in the design of the PH facility network under congestion. Facilities are modelled as spatially distributed queues with Poisson arrivals and general service time distributions (i.e., M/G/1 queue) to capture congestion due to waiting and service delay at the facilities. More specifically, we present a model to determine the configuration of the network, where the emphasis is on minimizing the total system-wide waiting time and travel times through location of facilities, the acquisition of adequate service capacity and the optimal allocation of clients to these facilities under stochastic client arrival rate. The model is formulated as a nonlinear mixed integer programming (MIP) problem. Using simple transformation and piecewise linear approximation, we reformulate the problem as a linear MIP. We present an exact (ϵ-optimal) solution approach based on cutting plane algorithm. We present an illustrative example to demonstrate the model and a real-life case study on the location of mammography clinics in Montreal, Canada.

The remainder of the article is organized as follows. The next section provides a review of relevant literature. In Section 3, we describe the problem setting and present a nonlinear MIP formulation of the problem. Section 4 describes the piecewise linear approximation and the linear reformulation of the model. The exact solution method is presented in Section 5. The illustrative example and the case study along with computational results are reported in Section 6. Section 7 concludes with some directions for future research.

2 Related literature

Facility location plays an integral role in healthcare planning in many settings. Location-allocation models have been extensively applied in locating hospitals in rural regions [11], healthcare planning with explicit geographical considerations [12], public healthcare planning [13], blood bank location [14], location and service mix for managed health care [15], trauma care with hospitals and ambulances [16], reorganization of liver transplant regions [17], locating specialized healthcare services such as traumatic brain injury treatment centers [18, 19], primary healthcare facility network design [21, 24], location of emergency care services [22], locating medical service facilities for large-scale emergencies, such as natural disasters or terrorist attacks [23], and the design of preventive healthcare facility networks [1, 2, 20, 25, 26]. One of the early reviews of location models in healthcare is by Smith-Daniels et al. [27]. For review of location-allocation models for planning health

services in developing countries, readers are referred to Rahman and Smith [28]. Daskin and Dean [29] provide an extensive review of location set covering, maximal covering and p-median models for addressing the location planning issues in healthcare. Recent developments can be found in Berg [30].

Despite numerous applications of location-allocation models in healthcare planning, the design of PH facility networks that include deciding the location of PH facilities, in particular, remains an interesting area yet to be explored. To the best of our knowledge, one of the earliest studies on PH facility location is by [1]. Recent attempts to consider the distinguishing features of PH services while locating healthcare facilities include [2, 20, 25, 26]. The major drivers that influence the location of PH facilities differs from other healthcare services. The first difference is the fact that clients may select a facility that is not necessarily the closest but the one that is more attractive due to other reasons such as congestion (service and waiting delay), service reputation at the facilities etc.. Another driver is the need to have a minimum number of clients in order to maintain the accreditation which requires accurate estimation of the capacity requirements [20]. In the literature, these drivers are considered in the facility location models by incorporating the congestion effects of the facilities using queueing models, and by defining the choice behavior of the clients through the attractiveness measures. Verter and Lapierre [1] present a solution approach by representing the probability of participation as a function of travelling distance, and minimum workload constraints in a maximal covering location problem (MCLP). Zhang et al. [2] incorporate the congestion effects using M/M/1 queue and consider the travelling and waiting time as a proxy for the accessibility of a facility. Zhang et al. [25] have further advanced the PH facility location models by allowing the clients from the same population zone to choose different facilities, allowing to reach allocation equilibrium.

The other critical aspect of PH facility network models is related to the choice of the users. While most works focus on user-optimal, that is, clients selecting the most attractive facility [1, 2, 25], a probabilistic choice allows capturing the client's behavior in the PH facility location problem. Gu et al. [20] use a probabilistic catchment area where both the proximity of clients to the facilities and the number of available facilities within the proximity of clients are considered to determine the attractiveness of facilities and allocation of the clients. Zhang et al. [26] compare the performance of a user-optimal choice with probabilistic choice allocation strategy. Furthermore, the solution methodologies implemented in earlier works are based on heuristics that are modelled on a location-allocation framework due to the highly nonlinear characteristics of the proposed models [1, 2, 25].

Another stream of research related to our work concerns facility location with stochastic demand and congestion (FLSDC). The problem of FLSDC arises in several planning contexts: location of emergency medical clinics, police stations, and fire stations; refuse collection and disposal; location of stores and service centers; telecommunication network design; location of bank branches and automated teller machines; and Internet mirror site location. For an extensive review, the readers are referred to Berman and Krass [31] and Boffey et al. [32]. It is worthwhile mentioning that despite a large body of literature in this area, none of the references discuss the issues and underlying characteristics of PH facility location and network design.

In this paper, we propose a model to simultaneously optimize the location and the capacities of PH facilities in a *system-optimal directed choice* setting. This paper differs from previous works in the following aspects:

- General service time distributions: The congestion at the facilities is modelled using the waiting and service delay in an M/G/1 queue. Most of the earlier works use either M/M/1 or M/M/s models to represent the congestion effects. This implies that the service times follow exponential distribution representing a coefficient of variation (cv) of 1. Unlike emergency clinics, clients arriving at the PH facilities often have a routine set of tasks for clients such as taking x-rays or blood tests. In these cases, the coefficient of variation can be less than 1. In other cases, where clients would like to make use of available multiple services in one visit and the procedure is manual, the coefficient of variation of service times can be substantially greater than 1. In that case, we argue that modelling the facility as an M/M/1 or M/M/s queue is a restrictive assumption. Hence, we model congestion at the facilities using a more general class of queuing systems, that is, an M/G/1 fqueue.
- Non-identical service rates: Unlike previous models that often seek to locate facilities with identical capacities/service rates, the proposed model seeks to locate facilities and select their capacity levels (or service rates) from a list of multiple discrete capacity options. This often results in network configuration with various facilities equipped with non-identical service rates in the network. This is more realistic given the uniform distribution of population density in various client zones.
- Capacity selection and budget constraint: Budget plays an important role in deciding the number and the location of PH facilities. This paper explicitly considers budget constraint, and hence, the optimal number and capacities of the facilities are affected by the overall budget.

– Social-optimal directed choice vs. user-optimal choice: User optimal choice models assume that the clients are fully informed about the optimal facility and they make rational choice to patronize their optimal facility at all times. The optimal location is based on highest attractiveness – which is often the travelling time (or distance) [26]. In this paper, we consider the impact of social-optimal directed choice on the location of the facilities and the design of the network, where the attractiveness of a facility is the weighted sum of the travelling time (or distance) and the congestion and service delay at that facility.

3 Model formulation

The PH facility network design problem determines the location and the capacity of PH facilities that serve various clients (or population zones) with the demand of their PH services. Once the client decides to access the services, a central decision maker/authority directs the clients to a particular facility, such that the total system-wide travel time and waiting time in queue as well as service is minimized. Hence, the central decision-maker has estimates of the average waiting times at the facilities based on their capacities and the population/clients that has been directed to the facilities.

Let I denote the set of clients residing in population zones indexed by i, $i \in I$ and J denote the set of candidate sites for the location of PH facilities indexed by j, $j \in J$. Unlike previous studies that limit the number of capacity levels at the facilities to one, we model situations where there exist multiple capacity levels to choose from. Let k be the index for discrete capacity levels at the facilities, $k \in K(j)$. The service rate of facility j equipped with capacity level k is denoted by μ_{jk} and the fixed cost of installation and operation (amortized over the planning period) is denoted by f_{jk}. We assume that the number of clients from the population zone i that require service is an independent random variable that follows a Poisson process with mean rate of λ_i. Clients arriving at the facilities are served on a first-come first-served (FCFS) basis. We assume that each facility operates as a single flexible-capacity server with an infinite buffer to accommodate clients waiting for service. If x_{ij} is the fraction of clients from zone i allocated to facility j, then the total number of clients served by facility j is also a random variable that follows a Poisson process with mean $\Lambda_j = \sum_{i \in I} \lambda_i x_{ij}$ (due to the superposition of Poisson processes).

Assuming that the service times at each facility follows a general distribution, each facility can be modelled as an M/G/1 queue. Let τ_j represent the mean service time at clinic j ($\tau_j = 1/\mu_j$), cv_j^2 be the squared coefficient of variation of service times ($cv_j^2 = \sigma_j^2/\tau_j^2$), and ρ_j be the utilization of clinic j ($\rho_j = \Lambda_j/\mu_j$). If y_{jk} is a binary variable that equals one if a facility with capacity level k is located at site j and zero otherwise, then the mean service rate of facility j, is given by $\mu_j = \sum_{k \in K(j)} \mu_{jk} y_{jk}$ and the variance in service times is $\sigma_j^2 = \sum_{k \in K(j)} \sigma_{jk}^2 y_{jk}$. This service rate reflects the number of clients a facility can serve in a given time period. Under steady state conditions ($\Lambda_j < \mu_j$) and FCFS queuing discipline, the average waiting time (including the service time) at facility j is given by the Pollaczek-Khintchine (PK) formula:

$$\overline{W}_j = \left(\frac{1 + cv_j^2}{2}\right) \frac{\tau_j \rho_j}{1 - \rho_j} + \tau_j = \left(\frac{1 + cv_j^2}{2}\right) \frac{\Lambda_j}{\mu_j(\mu_j - \Lambda_j)} + \frac{1}{\mu_j} \qquad \forall j$$

The total waiting time of clients at facility j is obtained by multiplying the average waiting time at clinic j by the aggregate arrival rate Λ_j as:

$$\Lambda_j \overline{W}_j = \left(\frac{1 + cv_j^2}{2}\right) \frac{\Lambda_j^2}{\mu_j(\mu_j - \Lambda_j)} + \frac{\Lambda_j}{\mu_j} \qquad \forall j \quad (1)$$

Substituting $\Lambda_j = \lambda_i x_{ij}$ and $\mu_j = \sum_{k \in K(j)} \mu_{jk} y_{jk}$, this can be further expressed as:

$$\lambda_i x_{ij} \overline{W}_j(\mathbf{x}, \mathbf{y}) = \frac{\left(1 + \sum_{k \in K(j)} cv_{jk}^2 y_{jk}\right) \left(\sum_{i \in I} \lambda_i x_{ij}\right)^2}{2 \sum_{k \in K(j)} \mu_{jk} y_{jk} \left(\sum_{k \in K(j)} \mu_{jk} y_{jk} - \sum_{i \in I} \lambda_i x_{ij}\right)} + \frac{\sum_{i \in I} \lambda_i x_{ij}}{\sum_{k \in K(j)} \mu_{jk} y_{jk}} \quad (2)$$

The objective is to minimize the weighted sum of total time that comprises two components: (i) the travel time from zone i to facility j through the shortest path denoted by t_{ij}; and (ii) The average time clients spend at the facility in waiting and receiving service, denoted by $\overline{W}_j$. Let α_t and α_w denote weights associated with the travel time and the waiting time components respectively ($\alpha_t + \alpha_w = 1$). These weights are set by the decision-maker to account for the importance of the components. The objective function can be expressed as follows:

$$T(\mathbf{x}, \mathbf{y}) = \alpha_t \sum_{i \in I} \sum_{j \in J} \lambda_i t_{ij} x_{ij} + \alpha_w \sum_{i \in I} \sum_{j \in J} \lambda_i x_{ij} \overline{W}_j(\mathbf{x}, \mathbf{y}) \quad (3)$$

PH programs often require a minimum number of clients to retain the accreditation. For example, the US Food and Drug Administration (FDA) requires 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 [2]. This requirement can be modelled using the following set of constraints that ensure that a facility cannot be established at node j unless they meet a minimum workload requirement denoted by R_{min}:

$$\sum_{i\in I}\lambda_i x_{ij} \geq R_{min} \qquad \forall j$$

We also include budget constraints on the fixed cost of opening facilities. However, this constraint can be easily modified to include the cost of providing service to clients.

$$\sum_{j\in J}\sum_{k\in K(j)} f_{jk}y_{jk} \leq B$$

The model formulated below simultaneously determines the location of facilities and their capacity levels, as well as the allocation of clients to these facilities in order to minimize the sum of travel time as well as total waiting time subject to budget constraints on the opening of facilities. The resulting nonlinear MIP formulation is as follows:

$$[P]: \min_{x,y} \; T(\mathbf{x},\mathbf{y}) = \alpha_t \sum_{i\in I}\sum_{j\in J}\lambda_i t_{ij} x_{ij} + \alpha_w \sum_{i\in I}\sum_{j\in J}\lambda_i x_{ij}\overline{W}_j(\mathbf{x},\mathbf{y}) \tag{4}$$

$$\text{s.t.} \quad \sum_{i\in I}\lambda_i x_{ij} \leq \sum_{k\in K(j)}\mu_{jk}y_{jk} \qquad \forall j \tag{5}$$

$$\sum_{k\in K(j)} y_{jk} \leq 1 \qquad \forall j \tag{6}$$

$$\sum_{j\in J} x_{ij} = 1 \qquad \forall i \tag{7}$$

$$\sum_{i\in I}\lambda_i x_{ij} \geq R_{min}\sum_{k\in K(j)} y_{jk} \qquad \forall j \tag{8}$$

$$\sum_{j\in J}\sum_{k\in K(j)} f_{jk}y_{jk} \leq B \tag{9}$$

$$0 \leq x_{ij} \leq 1, \quad y_{jk} \in \{0,1\} \qquad \forall i,j,k \tag{10}$$

The objective function (4) minimizes the weighted sum of total travel time and the total waiting (queuing + service) time in the system. Later, we perform sensitivity analysis to see their impact on the location and allocation variables. Constraints (5) ensure that the total demand is less than the total capacity. Furthermore, these constraints ensure that if site j is not opened (right hand side $= 0$), then clients from zone i would not be allocated to site j (i.e. $x_{ij} = 0$). Alternatively, these constraints ensure that the steady state conditions of a queueing system ($\Lambda_j \leq \mu_j$) are met. Constraints (6) state that at most one capacity level is selected at a facility, whereas constraints (7) ensure that the total population is served. Constraints (8) guarantee minimum workload requirements at each facility. Constraints (9) are budget constraints. Constraints (10) are nonnegativity and binary restrictions. The formulation allows for splitting of the client population from a zone to several facilities in order to achieve a system-optimal allocation policy. Furthermore, it reduces computational complexity associated with solving a model with binary variables. However, a restrictive version of the model can handle no-splitting requirements by imposing binary restrictions on x_{ij}.

The nonlinearity in the model $[P]$ is due to the expression for average waiting time, $\overline{W}_j(\mathbf{x},\mathbf{y})$ in the objective function. In the following section, we deal with the nonlinearity due to the expression of total waiting time of clients in the system using a linearization based on a simple transformation and a piecewise linear approximation.

4 Model linearization

By rearranging the terms in Eq. 1, the expression for the total waiting time of clients at any facility can be written as:

$$\Lambda_j\overline{W}_j = \frac{1}{2}\left\{\left(1+cv_j^2\right)\frac{\Lambda_j}{\mu_j-\Lambda_j} + \left(1-cv_j^2\right)\frac{\Lambda_j}{\mu_j}\right\}$$

which is equivalent to

$$\frac{1}{2}\left(1+\sum_{k\in K(j)} cv_{jk}^2 y_{jk}\right)\frac{\sum_{i\in I}\lambda_i x_{ij}}{\sum_{k\in K(j)}\mu_{jk}y_{jk} - \sum_{i\in I}\lambda_i x_{ij}} + \frac{1}{2}\left(1-\sum_{k\in K(j)} cv_{jk}^2 y_{jk}\right)\frac{\sum_{i\in I}\lambda_i x_{ij}}{\sum_{k\in K(j)}\mu_{jk}y_{jk}} \tag{11}$$

In order to linearize this expression, we define nonnegative auxiliary variables U_j and ρ_j such that

$$U_j = \frac{\Lambda_j}{\mu_j-\Lambda_j} = \frac{\sum_{i\in I}\lambda_i x_{ij}}{\sum_{k\in K(j)}\mu_{jk}y_{jk} - \sum_{i\in I}\lambda_i x_{ij}} \quad \text{and}$$

$$\rho_j = \frac{\Lambda_j}{\mu_j} = \frac{\sum_{i\in I}\lambda_i x_{ij}}{\sum_{k\in K(j)}\mu_{jk}y_{jk}} = \frac{U_j}{1+U_j}$$

This implies

$$\sum_{i\in I}\lambda_i x_{ij} = \frac{U_j}{1+U_j}\sum_{k\in K(j)}\mu_{jk}y_{jk} \tag{12}$$

$$= \rho_j\sum_{k\in K(j)}\mu_{jk}y_{jk} = \sum_{k\in K(j)}\mu_{jk}z_{jk}, \tag{13}$$

$$\text{where} \quad z_{jk} = \begin{cases} 0 & \text{if} \quad y_{jk}=0 \\ \rho_j & \text{if} \quad y_{jk}=1 \end{cases} \qquad \forall j,k$$

Note that $\rho_j = \frac{U_j}{1+U_j}$ is the server (facility) utilization. Since there is at most one k' with $y_{jk'} = 1$ while $y_{jk} = 0$ for all other $k \neq k'$, the expression $z_{jk} = \rho_j y_{jk}$ can be ensured by adding the following constraints:

$$\begin{aligned} z_{jk} &\leq y_{jk} && \forall j,k \\ \sum_{k\in K(j)} z_{jk} &= \rho_j && \forall j \end{aligned}$$

By differentiating the function $\rho_j = \frac{U_j}{1+U_j}$ w.r.t. U_j, we get the first derivative, $\frac{\delta\rho_j}{\delta U_j} = \frac{1}{(1+U_j)^2} > 0$, and the second derivative, $\frac{\delta^2\rho_j}{\delta U_j^2} = \frac{-2}{(1+U_j)^3} < 0$, which proves that the function is concave in $U_j \in [0,\infty)$. Let the domain H of the auxiliary variable U_j be a set of indices of points $\{U_j^h\}_{h\in H}$, at which the concave function $\rho_j(U_j) = U_j/(1+U_j)$ can be approximated arbitrary closely by a set of piecewise linear functions that are tangent to ρ_j at points $\{U_j^h\}_{h\in H}$. This implies that $\rho_j(U_j) = U_j/(1+U_j)$ can be expressed as the finite minimum of linearizations of ρ_j at a given set of point $\{U_j^h\}_{h\in H}$ as follows:

$$\rho_j = \min_{h\in H}\left\{\frac{1}{(1+U_j^h)^2}U_j + \frac{(U_j^h)^2}{(1+U_j^h)^2}\right\}, \quad \forall j$$

$$\text{or} \quad \rho_j \leq \frac{1}{(1+U_j^h)^2}U_j + \frac{(U_j^h)^2}{(1+U_j^h)^2}, \qquad \forall j, h \in H$$

$$\text{or} \quad (1+U_j^h)^2\rho_j - U_j \leq (U_j^h)^2 \qquad \forall j, h \in H$$

The expression for total waiting time (11) reduces to:

$$\begin{aligned} \Lambda_j \overline{W}_j &= \frac{1}{2}\left(1 + \sum_{k\in K(j)} cv_{jk}^2 y_{jk}\right)U_j \\ &\quad + \frac{1}{2}\left(1 - \sum_{k\in K(j)} cv_{jk}^2 y_{jk}\right)\rho_j \\ &= \frac{1}{2}\left\{U_j + \rho_j + \sum_{k\in K(j)} cv_{jk}^2 (w_{jk} - z_{jk})\right\} \\ &\quad \text{where} \quad w_{jk} = \begin{cases} U_j & \text{if} \quad y_{jk} = 1 \\ 0 & \text{otherwise} \end{cases} \qquad \forall j,k \end{aligned}$$

Similarly, because there exists at most one k' with $y_{jk'} = 1$ while $y_{jk} = 0$ for all other $k \neq k'$, the expression $w_{jk} = U_j y_{jk}$ can be ensured by adding the following constraints

$$\begin{aligned} w_{jk} &\leq M y_{jk} && \forall j,k \\ \sum_{k\in K(j)} w_{jk} &= U_j && \forall j \end{aligned}$$

where M is the usual Big-M.

The resulting linear MIP formulation is:

$$[L(H)]: \min \ \alpha_t \sum_{i\in I}\sum_{j\in J} \lambda_i t_{ij} x_{ij} + \alpha_w \frac{1}{2}\sum_{j\in J}\left\{U_j + \rho_j + \sum_{k\in K(j)} cv_{jk}^2 (w_{jk} - z_{jk})\right\} \tag{14}$$

s.t. (5) − (9)

$$\sum_{i\in I}\lambda_i x_{ij} - \sum_{k\in K(j)} \mu_{jk} z_{jk} = 0 \quad \forall j \tag{15}$$

$$z_{jk} - y_{jk} \leq 0 \qquad \forall j,k \tag{16}$$

$$\sum_{k\in K(j)} z_{jk} - \rho_j = 0 \qquad \forall j \tag{17}$$

$$(1+U_j^h)^2\rho_j - U_j \leq (U_j^h)^2 \qquad \forall j, h \in H \tag{18}$$

$$w_{jk} - M y_{jk} \leq 0 \qquad \forall j,k \tag{19}$$

$$\sum_{k\in K(j)} w_{jk} - U_j = 0 \qquad \forall j \tag{20}$$

$$y_{jk} \in \{0,1\} \qquad \forall j,k \tag{21}$$

$$0 \leq x_{ij}, z_{jk} \leq 1; \qquad \forall i,j,k \tag{22}$$

$$\rho_j, U_j, w_{jk} \geq 0; \qquad \forall j,k \tag{23}$$

The steady-state conditions ($\lambda_j < \mu_j$) translate into capacity constraints, and are enforced by the constraints (15) and (16) and forced to "<" by the term U_j in the objective.

$[L(H)]$ is a minimization problem, hence at least one of the constraints in Eq. 18 will be binding. This implies that

$$\rho_j = \min_{h\in H}\left\{\frac{1}{(1+U_j^h)^2}U_j + \frac{(U_j^h)^2}{(1+U_j^h)^2}\right\} \quad \forall j$$

Note that there is an exponential number of constraints (18) in the linear MIP model $[P_{L(H)}]$ due to the piecewise linear approximation of $\rho_j(U_j)$. However, it is not necessary to generate all. Instead, it suffices to start with a subset of these constraints and generate the rest as needed.

5 Solution method

The proposed exact solution approach relies on obtaining good lower and upper bounds for the linear model $[L(H)]$. The algorithm makes successive improvements to the lower bound and the corresponding upper bound as the iterations progress. Below, we present lower and upper bounds that can be used in the proposed solution approach.

Lower bound: For every given subset of points $\{U_j^h\}_{h\in H^q \subset H}$, the optimal objective function value of the problem $[L(H^q)]$

is a lower bound on the optimal objective of $[L(H)]$ or $[P]$.

Suppose, at an iteration q, we use a subset of points $\{U_j^h\}_{h\in H^q\subset H}$, and solve the corresponding model $[L(H^q)]$, which yields the solution $(\mathbf{x}^q, \mathbf{y}^q, \rho^q, \mathbf{w}^q, \mathbf{z}^q, \mathbf{U}^q)$ and objective function value denoted by $v(L(H^q))$. Since $[L(H^q)]$ is a relaxation of the full problem $[L(H)]$, a lower bound on the optimal objective of $[L(H)]$ or $[P]$ is provided by $v(L(H^q))$, where

$$LB = v(L(H^q)) = \alpha_t \sum_{i\in I}\sum_{j\in J} \lambda_i t_{ij} x_{ij}^q + \frac{\alpha_w}{2}\sum_{j\in J}\left\{U_j^q + \rho_j^q + \sum_{k\in K} c v_{jk}^2 (w_{jk}^q - z_{jk}^q)\right\} \tag{24}$$

Upper bound: For any subset of points $(U_{ij}^h)_{H^q\subset H}$, the objective function of $[P]$ evaluated at the optimal solution $(\mathbf{x}^q, \mathbf{y}^q, \rho^q, \mathbf{w}^q, \mathbf{z}^q, \mathbf{U}^q)$ of $[L(H^q)]$ provides an upper bound to $[L(H)]$ or $[P]$.

Consider iteration q, where we use a subset of tangent points $(U_{ij}^h)_{H^q\subset H}$ and solve the corresponding relaxed problem $[L(H^q)]$. Because the optimal solution $(\mathbf{x}^q, \mathbf{y}^q, \mathbf{U}^q)$ of $[L(H^q)]$ is a feasible solution to $[P]$, it provides an upper bound on the optimal objective of $[P]$, given by

$$UB^q = T(\mathbf{x}^q, \mathbf{y}^q) = \alpha_t \sum_{i\in I}\sum_{j\in J} \lambda_i t_{ij} x_{ij}^q + \frac{\alpha_w}{2}\sum_{j\in J}\left(1 + \sum_{k\in K} c v_{jk}^2 y_{jk}^q\right)\frac{\sum_{i\in I}\lambda_i x_{ij}^q}{\sum_{k\in K}\mu_{jk} y_{jk}^q - \sum_{i\in I}\lambda_i x_{ij}^q} + \frac{\alpha_w}{2}\sum_{j\in J}\left(1 - \sum_{k\in K} c v_{jk}^2 y_{jk}^q\right)\frac{\sum_{i\in I}\lambda_i x_{ij}^q}{\sum_{k\in K}\mu_{jk} y_{jk}^q} \tag{25}$$

5.1 Exact algorithm

The algorithm presented below makes successive improvements to the lower and upper bounds as the iteration progresses. At every iteration, a relaxed version of the linear model $[L(H)]$ is solved to obtain an optimal solution, an upper bound and a lower bound. This solution is used to generate a set of "cuts / constraints" that eliminate the best solution found so far and improve the upper bound on the remaining solutions. The procedure terminates when the gap between the current upper bound and the best lower bound is within the tolerance limits.

The algorithm starts with an initial subset $H^q \subset H$. The resulting model $[L(H^q)]$ is solved and the lower bound (LB^q) and the upper bound (UB^q) are computed using (24) and (25) respectively. If the upper bound (UB^q) equals the best known lower bound (LB^q) within accepted tolerance (ϵ) at any given iteration q, then $(\mathbf{x}^q, \mathbf{y}^q, \mathbf{U}^q)$ is an optimal solution to $[P]$ and the algorithm is terminated. Otherwise, a new set of candidate points $U_{ij}^{h_{new}}$ is generated using the current solution $(\mathbf{x}^q)$ as follows:

$$U_j^{h_{new}} = \frac{\sum_{i\in I}\lambda_i x_{ij}^q}{\sum_{k\in K}\mu_{jk} y_{jk}^q - \sum_{i\in I}\lambda_i x_{ij}^q}$$

This new set of points is appended to $(U_{ij}^h)_{H^q\subset H}$ and the procedure is repeated again, until the stopping criteria is reached. The algorithm is outlined below:

Algorithm 1 Cutting Plane Method

Ensure: $UB \leftarrow \infty$; $LB \leftarrow -\infty$; $q \leftarrow 0$
Require: Choose an initial set of points R^h.

1: **while**$(UB - LB)/LB \geq \epsilon$
2: Solve $L(H^q)$ to obtain $(\mathbf{x}^q, \mathbf{y}^q, \rho^q, \mathbf{w}^q, \mathbf{z}^q, \mathbf{U}^q)$.
3: Update the lower bound: $LB^q \leftarrow v(L(H^q))$.
4: Update the upper bound: $UB^q \leftarrow \min\{UB^{q-1}, Z(x^q, y^q)\}$.
5: Compute new points: $U_j^{h_{new}} = \frac{\sum_{i\in I}\lambda_i x_{ij}^q}{\sum_{k\in K}\mu_{jk} y_{jk}^q - \sum_{i\in I}\lambda_i x_{ij}^q}$
6: Generate new constraints: $(1 + U_j^{h_{new}})^2 \rho_j - U_j \leq (U_j^{h_{new}})^2, \quad \forall j, h \in H$
7: Append new constraints: $H^{q+1} \leftarrow H^q \cup \{h_{new}\}$
8: $q \leftarrow q + 1$
9: **end while**

6 Computational results and insights

In this section, we provide a small illustrative example to demonstrate the tradeoff in the model and understand the interaction among facility location, capacity selection, and allocation of clients to facilities in the design of a PH facility network under congestion. We also present a case study on the location of mammography clinics in the Montreal region. The computational experiments were performed using GAMS IDE 2.0 and the CPLEX 10.2 solver. The ϵ-optimality gap as well as the optimality gap for MIP solved in each iteration is set to 10^{-6}. The proposed algorithm converged to a solution for all the experiments including the Montreal case study. Due to the size of the problem, we limit the CPU time to 240 minutes for each MIP iteration in the Montreal case study.

6.1 Illustrative example

In order to highlight the impact of different parameters on the decision and performance levels, we present an illustrative example with a small data set. The example comprises 16 population zones and 7 potential facility locations to serve the demand for PH services. Table 1 presents the travel time (in minutes) from each population zone to the potential facility location and the maximum demand originating from each zone. Each facility can be equipped with one of the 6 discrete capacity levels. Table 2 presents the service rate, μ_{jk} (in clients/hour) and the corresponding fixed cost, f_{jk} (in units of currency) for each of these capacity levels. For each facility, the minimum workload requirement (R_{min}) is set to 2 clients/hour. All the aforementioned model parameters remain constant throughout the numerical experiments. First, we identify the efficient frontier of the total travel time and the total

Table 1 Travel time (min) between each zone and facility and total demand per zone

Zone	Facility 1	2	3	4	5	6	7	Demand (clients/hr)
1	22	26	33	37	42	55	62	0.61
2	16	28	35	32	45	60	62	1.52
3	15	14	27	28	35	42	48	0.71
4	29	12	24	35	32	37	48	0.66
5	14	26	32	24	39	60	50	0.95
6	38	19	11	31	22	26	39	0.22
7	36	21	16	18	16	28	30	0.29
8	32	30	29	10	21	45	30	1.57
9	45	30	16	30	21	14	35	1.60
10	41	36	24	18	11	27	18	0.70
11	53	38	23	29	15	15	22	1.21
12	50	37	26	25	12	20	19	1.52
13	55	40	27	37	20	12	26	1.21
14	69	58	37	42	26	25	21	0.66
15	65	63	50	40	37	45	19	0.95
16	75	68	47	44	36	40	25	1.62

waiting time. Then, we compare the performance levels of the proposed model and another optimal choice setting where clients are directed to the closest open facility. Finally, we study the impact of changing the available budget and the coefficient of variation of service times on the solutions.

6.1.1 Efficient frontier of travel and waiting time

The objective function of the proposed model consists of two components: the travel time and the waiting time, which may differ in their relative importance. In order to understand the tradeoff between these two components, we conduct an efficient frontier analysis by varying α_t and α_w, where α_t and α_w are respective weights for the travel time and waiting time such that $\alpha_t + \alpha_w = 1$. For a fixed budget level of 35, Fig. 1 shows the tradeoff between the waiting time and travel time as we change their weights. For the same set of instances, Table 3 lists the open facilities with corresponding service rates (capacity levels) as well as the total travel time (TT) and waiting time (WT).

Table 2 Values of service rate and facility cost

	Capacity level $k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$
Service rate (Clients/hr)	3	6	9	12	15	18
Fixed cost (Units)	5	10	15	20	25	30

There are three observations to note from this experiment. First, according to the network configurations in Table 3, the travel time depends on the number of open facilities and their locations, whereas the waiting time depends on the capacity of these facilities. Hence, for a given budget, the model prescribes decentralized facilities if travel times are the priority (e.g., $\alpha_w = 0.01$), whereas capacity pooling through a centralized facility should be the strategy if waiting times are the priority ($\alpha_w > 0.2$).

Secondly, the degree of the centralization depends on the overall resource utilization. The current budget level of 35 corresponds to an aggregate average utilization of 75 %. This implies that even a relatively low value of α_w will result in centralization of facilities that will reduce the waiting time substantially. As we increase the importance of waiting time, the centralization effect will allow reaching a lower bound in waiting time. In this case, the lower bound is reached when α_w is set to 0.2. If the available budget allows a relatively low resource utilization, then the α_w threshold to achieve the lower bound on waiting time will increase as indicated in the Table 4.

The last observation pertains to the average time spent by a client in the system. At the budget level of 35, by increasing α_w from 0.10 to 0.20, the waiting time decreases from 47.39 minutes to 23.33 minutes and the travel time has increased by only 3.49 minutes. On the other hand, by increasing α_w from 0.90 to 0.99, the travel time has increased by 1.73 minutes while the waiting time has decreased by less than 0.1 min. Thus, an appropriate choice of α_t and α_w can be identified in order to minimize the total time spent by clients. As illustrated in Fig. 2, the optimal time in the system can be achieved where the value of α_w is between 0.2 and 0.6, which indicates that the model is quite robust with respect to the weights.

6.1.2 Comparison of directed choice and user choice models

In this section, we conduct experiments to compare the impact of social-optimal directed choice and user choice on the total waiting time and the travel time and analyze its effect on the solutions. Table 5 shows that at any given budget level, targeting the social optimum in the location and allocation decision leads to a better performance for all clients, which will eventually improve their participation in the PH services. As we can see from the table, if the priority is on minimizing the travel time only, the total time spent by clients is very large since waiting times are high. This would be similar to a situation where clients have no information about waiting times at the facilities and patronize the nearest facility. If the objective is to minimize the sum of waiting time and the travel time while the clients patronize the nearest facility, the total time improves substantially as a result of reduction in waiting times. The table shows that the best total time can be achieved in the case of social-optimal directed choice model, where the objective is to minimize the sum of waiting time and travel time while the clients are directed to the facility (not necessarily the nearest) by a central decision-maker. The model also prescribes clients from some population zones to be split to more than one facility. We can see that while clients spend a little more time on the travelling, their waiting times are improved substantially, resulting in the best total time in the system (Table 6).

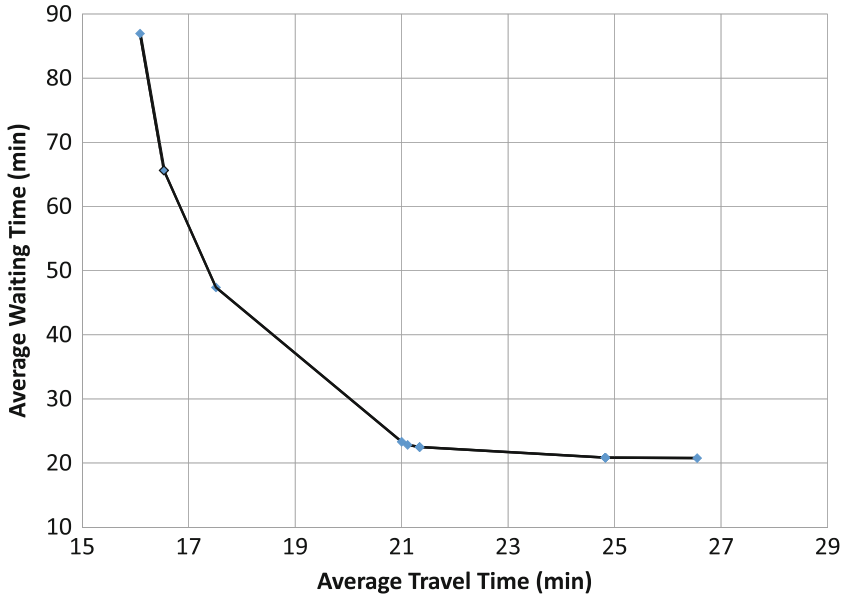

Fig. 1 Trade-off between travel time and waiting time

The advantage of the proposed model can also be observed by comparing our results with a user choice model proposed by [26], where the objective is to maximize total participation under congestion. We use the illustrative example reported in [26]. In order to do this, we modify our model such that the total allocation of the clients is greater than or equal to the participation level achieved by their model. Based on an instance of a problem reported by [26], the following table represents the difference in the optimal solutions. The comparison shows that the proposed model yields lower waiting time and travel time compared to their user choice model for the same level of participation.

6.1.3 Effect of change in budget

One of the most essential considerations in making decisions for PH facility networks is the available budget, since the mammography clinics can be equipped with various technology choices whose costs and capacities are variable. Due to this variability, it may be more appropriate for the policy-makers to allocate a budget and assess its impact on the design of the services. Figure 3 shows the results of this experiment. Note that the coefficient of variation is fixed to 1 for all facilities and capacity levels.

Table 3 Network configuration for different values of α_w

Weight	Service rates of open facilities (clients/hr)							TT	WT
α_w	$j=1$	$j=2$	$j=3$	$j=4$	$j=5$	$j=6$	$j=7$	(min)	(min)
0.01	3	3		3	3	6	3	16.09	86.96
0.05	6			3	3	6	3	16.54	65.63
0.10	6			3		6	6	17.51	47.39
0.20	6				15			21.00	23.33
0.30	6				15			21.11	22.84
0.40	6				15			21.34	22.50
0.50	6				15			21.34	22.50
0.60	6				15			21.34	22.50
0.70	6				15			24.83	20.85
0.80	6				15			24.83	20.85
0.90	6				15			24.83	20.85
0.95	6				15			24.83	20.85
0.99		3			18			26.55	20.78

Table 4 Impact of α_w on travel time and waiting time for various levels of budget

Weight	Travel Time (TT) Budget			Waiting Time (WT) Budget			Total Time (T) Budget			# of Open Facilities Budget		
α_w	35	50	70	35	50	70	35	50	70	35	50	70
0.1	17.51	16.50	15.89	47.39	22.65	14.36	64.90	39.15	30.25	4	5	6
0.2	21.00	17.30	16.38	23.33	17.27	11.53	44.33	34.56	27.91	2	4	5
0.3	21.11	18.81	17.29	22.84	12.07	9.20	43.95	30.88	26.49	2	3	4
0.4	21.34	18.81	17.29	22.50	12.07	9.20	43.84	30.88	26.49	2	3	3
0.5	21.34	21.00	18.60	22.50	8.93	6.81	43.84	29.93	25.41	2	2	3
0.6	21.34	21.21	18.60	22.50	8.76	6.81	43.84	29.96	25.41	2	2	3
0.7	24.83	21.57	18.60	20.85	8.59	6.81	45.68	30.16	25.41	2	2	3
0.8	24.83	22.05	18.60	20.85	8.44	6.81	45.68	30.49	25.41	2	2	3
0.9	24.83	22.07	23.64	20.85	8.44	6.00	45.68	30.51	29.64	2	2	3
0.95	24.83	22.41	23.64	20.85	8.41	6.00	45.68	30.82	29.64	2	2	2
0.99	26.55	22.55	23.81	20.78	8.41	6.00	47.33	30.96	29.81	2	2	2

The left portion of plot in Fig. 3 (budget range: 30 to 50) shows that the average travel time from population zones to open facilities decreases in steps as the budget increases. On the other hand, the right portion of the plot in Fig. 3 (budget range: 70 to 100) illustrates the diminishing rate of improvement in the waiting time as the budget increases. The stepped decrease in the travel time can be explained by the fact that for a given budget range, it is beneficial to increase the capacities of the existing facilities rather than opening new facilities. The capacity increase of an existing facility results in the reduction of the waiting times as seen on the waiting time curve. Once the marginal increase in budget does not improve the waiting times, it is better to open a new facility in order to decrease the average travel time. For example, when the budget increases from 50 to 55, an additional facility is opened, increasing the total number of open facilities from 2 to 3 sites for the given budget. This results in a substantial decrease in travel time albeit with a slight increase in waiting times. The increase in waiting time can be explained by the fact that the budget is now distributed to three sites resulting in facilities with smaller capacities. However, the net effect is a decrease in the total time spent by clients in the system as indicated in Fig. 3.

Once a new facility is opened, the model will keep adding capacity to open facilities in order to reduce the waiting time further. The reduction in waiting time presents a diminishing rate of improvement since larger budgets allow reduced utilization levels in the open facilities. As a result, the increment on budget does not improve the waiting times substantially compared to higher utilization levels.

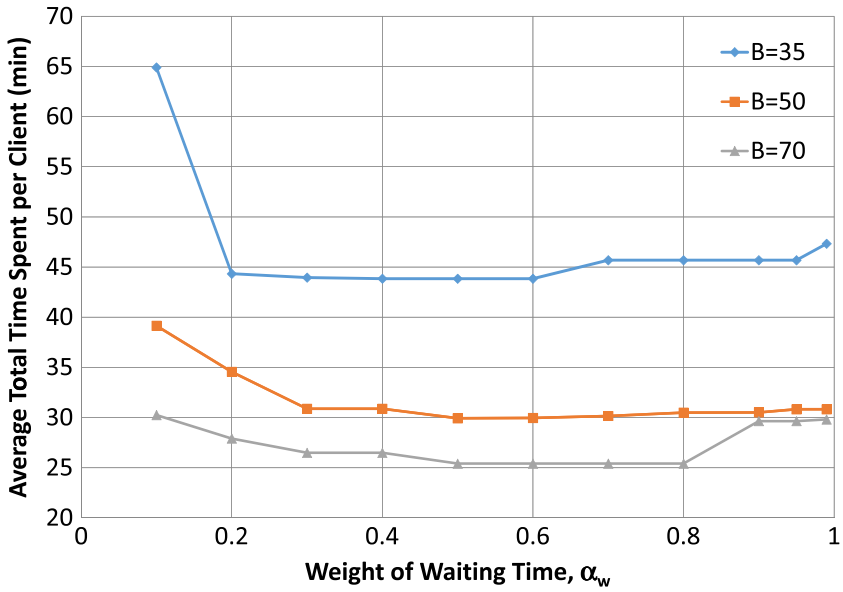

Fig. 2 Effect of changing α_w on the travel time and waiting time

Table 5 Comparison of travel time and waiting time under user choice and directed choice models

	User choice model			User choice model			Directed choice model				
	Min Travel Time; Clients are directed to closest facility			Min Travel + Waiting Time; Clients are directed to closest facility			Min Travel + Waiting Time; Clients allocation are decision variables			Improvement in Total Time	
Budget	TT	WT	T^A	TT	WT	T^B	TT	WT	T^C	T^A-T^C	T^B-T^C
30	20.99	107.02	128.01	27.00	30.00	57.00	27.00	30.00	57.00	71.01	0.00
35	17.27	55.86	73.13	20.99	23.32	44.31	21.34	22.50	43.84	29.29	0.47
40	17.27	55.86	73.13	20.99	16.22	37.21	21.52	13.72	35.24	37.89	1.97
45	17.27	55.86	73.13	20.99	10.38	31.37	20.98	10.38	31.36	41.77	0.01
50	17.27	35.06	52.33	21.00	8.93	29.93	21.00	8.93	29.93	22.40	0.00
55	17.27	55.86	73.13	18.60	10.43	29.03	19.07	9.69	28.76	44.37	0.27
60	17.27	36.84	54.11	18.60	8.98	27.58	18.72	8.56	27.28	26.83	0.30

Table 6 Comparison of directed choice vs. user choice Zhang et al. [26]

	Service rate of open facilities (clients/hr)					WT	TT	T	Allocation
Model	$j=1$	$j=2$	$j=3$	$j=4$	$j=5$	(min)	(min)	(min)	
User choice	3	3		3		48.9	26.1	75.0	52.40 %
Directed choice		6	3			30.5	21.1	51.6	52.40 %

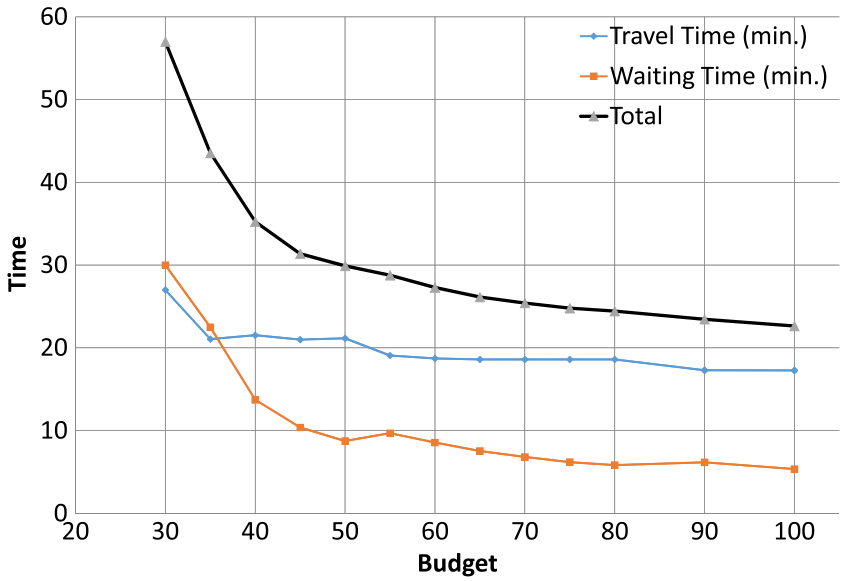

Fig. 3 Effect of changing budget on the travel time and waiting time

Table 7 Coefficient of variation in service times at the facilities

Facility	Service rate (clients/hr)					
	3	6	9	12	15	18
1	2	1.7	1.5	1	0.5	0.10
2	2	1.7	1.5	1	0.5	0.10
3	2	1.7	1.5	1	0.5	0.10
4	2	1.7	1.5	1	0.5	0.10
5	2	1.7	1.5	1	0.5	0.10
6	2	1.7	1.5	1	0.5	0.10
7	2	1.7	1.5	1	0.5	0.10

6.1.4 Coefficient of variation of service times

Most of the healthcare facility network decision models in literature are based on the assumption of exponential service times at the facilities. This implies that the service rate has moderate variability with coefficient of variation of 1 for all the servers, regardless of their capacities. However, the capacity pooling effect should reduce the overall coefficient of variation. In our model, one can account for this change in overall coefficient of variation. The following example shows the impact in terms of capacity allocation and network location decisions when we do consider the effect of cv. In this example, we set the budget level to 85 and R_{min} to 0.5. Then, we identify the optimal solution in two cases: (i) the cv_{jk}^2 is set to 1 for all the capacity levels; and (ii) a variable cv_{jk}^2 as shown in Table 7.

Based on these two instances, Table 8 represents the optimal facility location and performance levels. The impact of reduced coefficient of variation leads to opening facilities with greater capacity with some level of compromise in the travel time. Accordingly, the total time spent by a client improves.

6.2 Case study: location of mammography clinics in Montreal, Canada

In this section, we present a case study on the location of mammography clinics for the greater Montreal region to illustrate the application of our model. This case was previously studied by [2, 25]. The case is based on the initiative of the Quebec Ministry of Health to subsidize mammogram for women between the ages of 50 and 69. According to the case study data, there were 194,475 women in Montreal in this age group in 1996. The region was divided into 497 population zones and a total of 36 potential locations were identified for the mammography clinics. Figure 4 shows these potential sites.

The mammography machines are assigned to these locations in a way that the facility will satisfy a minimum mammography exam load of 4000 per year in order achieve accreditation criteria. This corresponds to R_{min} of 2 clients/hr. The service rate per mammography machine/physician is set to be 5 clients/hr. In order to show the impact of budget on the network configuration, we set it at various levels and observe its effects on the spatial coverage and capacity pooling effects. The composition of open facilities and their capacity levels for different budget levels are shown in Table 9 and Fig. 5. Table 9 shows that the increase in budget results in opening a larger number of facilities with larger capacities. At the highest budget, we observe that out of 24 open facilities, one is capable of handling 20 clients/hr and six facilities are capable of handling 15 clients/hr. Increased budget levels allow for larger size facilities, which represent the capacity pooling effect in terms of waiting time reduction. On the other hand, they also allow opening more facilities in order to reduce the travel time of the clients to the clinics. As the budget becomes more constrained, the number of open facilities decreases along with their capacity levels.

Another observation is the change in allocation of clients to facilities as a result of change in budget. At low budget levels, 11 zones (out of 497 population zones) are split into two facilities in order to reach a social optimum with respect to waiting times compared to the non-splitting scenario. As the budget levels increase, clients in a population zone are directed to a single facility without overloading and creating increased waiting times. Therefore, the optimal solutions provide a balanced load at the facilities as the budget decreases. In fact, the standard deviation of utilization levels range between 5 % and 7 %.

The number of open clinics is the same for budget levels of 215, 255 and 275 while their capacities have increased. This is due to the fact that at higher budget levels, the clients have been assigned to their closest facilities, and hence no further improvement in travel time can be realized. Any improvement in total time can be achieved as a result of reduction in the waiting time only. Therefore, the model prescribes increasing the capacity levels of the existing facilities. However, if the budget is tightly constrained, facilities are kept open with the lowest capacity levels in order to avoid degradation in travel times.

Table 10 presents the client's arrival rate (location-allocation decisions), service rate (capacity decisions), the utilization levels, and the waiting time at the various clinics for four different budget levels. From this table, we observe that the increase in the budget levels allow opening of new clinics in high density population zones 12, 27, 34 to reduce the load of the central clinics such as 7, 18, and 30. Once the high-density population zones are covered, clinics in remote locations are opened to further reduce the waiting times.

Table 8 Impact of service rate variability

Coefficient of Variation	Service rate of open facilities (clients/hr)					WT (min)	TT (min)	T
	Facility 1	Facility 4	Facility 5	Facility 6	Facility 7			
Uniform	15	9		12	15	6.75	17.28	24.03
Variable	18		18		15	4.62	18.6	23.22

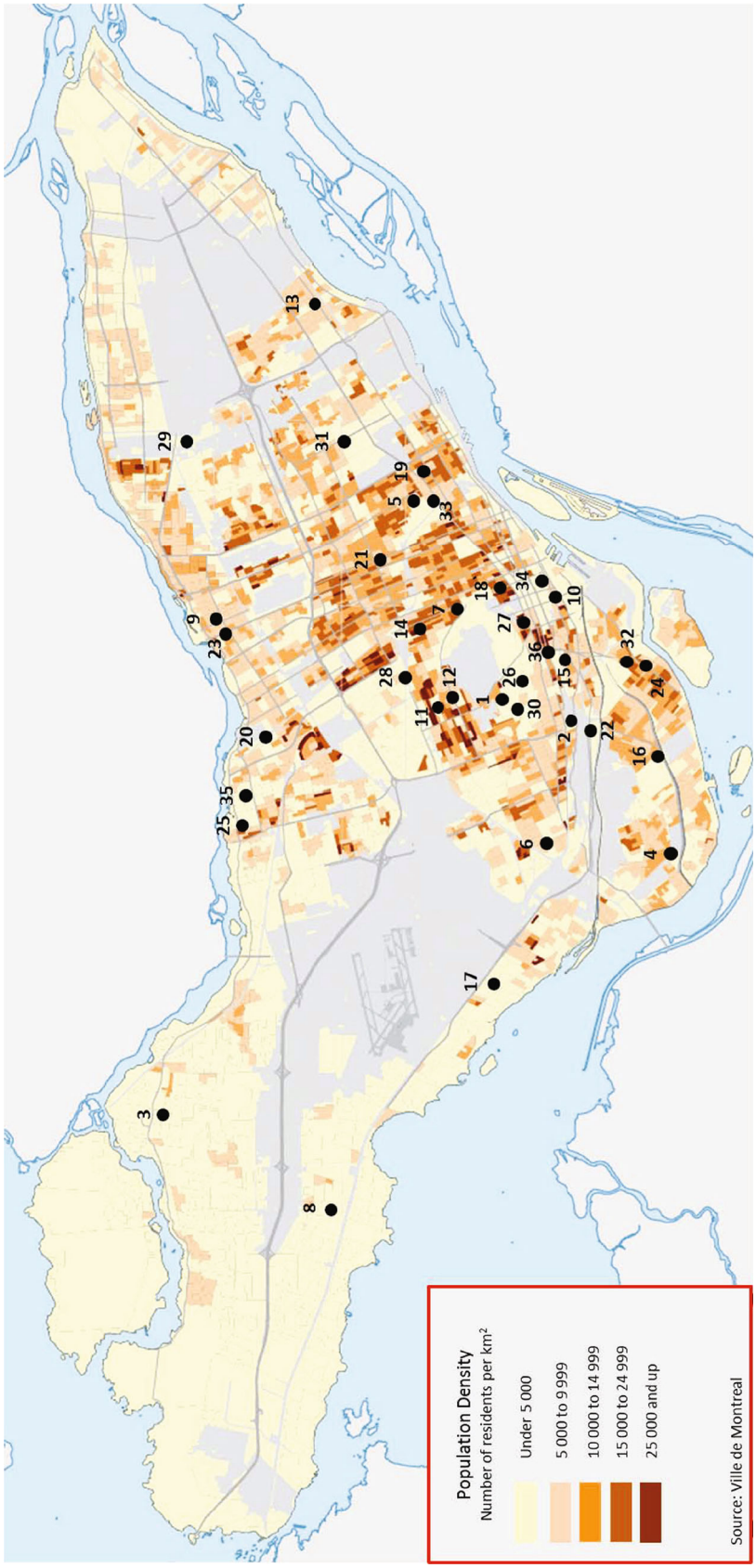

Fig. 4 Potential sites for location of mammography clinics in Montreal

Table 9 Composition of open facilities and their capacity level for different budget

Budget	Number of facilities opened with capacity level					Total Facilities	TT (min)	WT (min)	#Split Zones
	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$				
125	7	4	2	1	0	14	81.51	28.65	11
160	3	13	1	0	0	17	73.50	16.32	6
190	7	14	1	0	0	22	68.88	13.93	10
215	7	15	2	0	0	24	67.46	12.12	9
255	2	17	5	0	0	24	67.14	9.32	6
275	1	16	6	1	0	24	67.09	8.19	2

7 Conclusion and future research

Motivated by the significance of accessibility of the clients to PH facilities, we studied the impact of system-optimal directed choice on the design of a PH facility network. The model captures the tradeoff between the waiting time and travel time while simultaneously determining the location and the capacity of the facilities as well as the allocation of clients to the facilities. Under the assumption that the clients' arrival at the PH facilities follows Poisson process and their service times follow general distribution, the facilities were modelled as a network of spatially distributed M/G/1 queues. We formulated a nonlinear mathematical model and linearized it using simple transformation and piecewise linear approximation. We presented an exact (ϵ-optimal) solution approach based on cutting plane algorithm to solve the model to optimality. Using an illustrative example and a case study of mammography clinics location in Montreal, several observations and managerial insights were presented. The results from the illustrative example demonstrate that considering waiting times in the PH facility network design reduces the total time spent by clients substantially . The results also show that for a given budget, facilities should be decentralized as much as possible if travel times are the main priority (e.g. $\alpha_w = 0.01$), and capacity pooling through centralization should be the strategy if waiting times are important ($\alpha_w > 0.2$). For low budgets, even a relatively low importance of waiting time substantially improves the total time spent by clients. The minimum total time spent can be identified for a larger range of α_w, indicating that the model is robust with respect to the weights. Capacity pooling is captured by the coefficient of variation term in the model.

We have shown that compared to the user choice models, the proposed directed choice model provides better accessibility to clients by minimizing their total time in the system.There are several benefits of the proposed model. First, the solution ensures that all the clients receive their service in a socially equitable manner if they are centrally directed. Secondly, the minimum total time in the system is achieved by finding an appropriate tradeoff between the travel time and waiting time of clients. The model can also accommodate surges in demand as opposed to user choice models where the facility network is designed to serve a fraction of the population. The results of the Montreal case study also shows the benefit of splitting clients from a population zone to several facilities, especially in the case of low budget levels. Furthermore, we observe that considering waiting times in the design of a PH facility network is critical when the total budget is constrained.

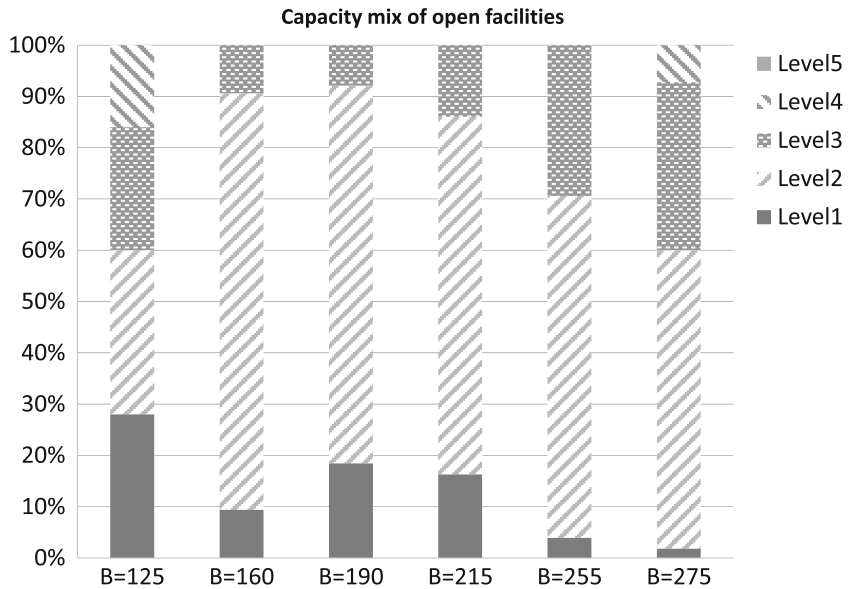

Fig. 5 Capacity mix of open facilities

Table 10 Results of the Montreal case for different levels of budget

	Budget = 125				Budget = 160				Budget = 190				Budget = 275			
Facility (j)	Λ_j	μ_j	ρ_j	W_j	Λ_j	μ_j	ρ_j	W_j	Λ_j	μ_j	ρ_j	W_j	Λ_j	μ_j	ρ_j	W_j
1	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
2	–	–	–	–	5.97	10	60	15	4.83	10	48	12	4.34	10	43	11
3	3.94	5	79	56	5.52	10	55	13	5.51	10	55	13	4.86	15	32	6
4	3.62	5	72	44	5.68	10	57	14	4.9	10	49	12	4.93	15	33	6
5	7.87	10	79	28	6.02	10	60	15	4.71	10	47	11	3.73	10	37	10
6	7.25	10	72	22	6.09	10	61	15	5.78	10	58	14	5.78	15	39	7
7	12.34	15	82	23	6.54	10	65	17	5.63	10	56	14	4.22	10	42	10
8	3.89	5	78	54	2.84	5	57	28	2.67	5	53	26	3.49	10	35	9
9	7.87	10	79	28	6.05	10	60	15	5.56	10	56	13	4.9	15	33	6
10	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
11	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
12	–	–	–	–	–	–	–	–	2.21	5	44	22	3.64	10	36	9
13	7.25	10	72	22	5.55	10	55	13	5.04	10	50	12	5.81	15	39	7
14	–	–	–	–	–	–	–	–	–	–	–	–	2.71	10	27	8
15	–	–	–	–	6.16	10	62	16	–	–	–	–	3.39	10	34	9
16	3.62	5	72	44	–	–	–	–	2	5	40	20	2.38	10	24	8
17	3.27	5	65	35	2.53	5	51	24	2.47	5	49	24	2.78	10	28	8
18	16.83	20	84	19	10.87	15	72	15	9.09	15	61	10	8.24	20	41	5
19	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
20	–	–	–	–	–	–	–	–	–	–	–	–	2	5	40	20
21	–	–	–	–	6.37	10	64	17	5.76	10	58	14	5.41	15	36	6
22	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
23	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
24	–	–	–	–	–	–	–	–	2	5	40	20	–	–	–	–
25	3.62	5	72	44	3.27	5	65	35	4.43	10	44	11	3.49	10	35	9
26	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
27	–	–	–	–	–	–	–	–	4.89	10	49	12	3.18	10	32	9
28	–	–	–	–	6.16	10	62	16	4.89	10	49	12	3.44	10	34	9
29	3.62	5	72	44	5.08	10	51	12	4.89	10	49	12	4.25	10	42	10
30	12.23	15	82	22	6.54	10	65	17	5.33	10	53	13	4.2	10	42	10
31	–	–	–	–	–		–	–	–	–	–	–	–	–	–	–
32	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
33	–	–	–	–	–	–	–	–	2.45	5	49	23	3.4	10	34	9
34	–	–	–	–	–	–	–	–	2.21	5	44	22	2.65	10	26	8
35	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
36	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–

Λ_j: Total Workload; μ_j: Service Rate; ρ_j: Utilization (Percentage); W_j: Average Waiting Time of Client

The proposed model and the solution methodology can be extended to other facility location problems with congestion. While our model can accommodate randomness in service and arrival rates, it does not capture the seasonality of the demand, e.g. surges due to flu seasons and epidemics. Robust optimization models to accommodate variations in demand during various times of the planning period is a potential future research direction.

Acknowledgements This research was supported by the National Science and Engineering Research Council of Canada under grant 386501-2010. Their support is highly acknowledged. The authors would like to acknowledge Mr. Abderrahamane Abbou for writing the code and conducting the experiments. The authors would like to thank Prof. Yue Zhang of the University of Toledo for providing the data for the case study. Thanks are due to the anonymous referees for their constructive comments that led to the improved version of the paper.

References

1. Verter V, Lapierre SD (2002) Location of preventive health care facilities. Ann Oper Res 110:123–132
2. Zhang Y, Berman O, Verter V (2009) Incorporating congestion in preventive healthcare facility network design. Eur J Oper Res 198:922–935
3. Gornick M, Eggers PW, Riley GF (2004) Associations of race, education, and patterns of preventive service use with stage of cancer at time of diagnosis. Health Ser Res 39:1403–1427
4. McNoe B, Richardson A, Elwood JM (1996) Factors affecting participation in mammography screening. N Z Med J 109:359–361
5. Zimmerman S (1997) Factors influencing hispanic participation in prostate cancer screening. Oncology Nursing Forum 24:499–504
6. Facione N (1999) Breast cancer screening in relation to access to health services. Oncology Nursing Forum 26:689–696
7. Gerard K, Shanahan M, Louviere J (2003) Using stated preference discrete choice modelling to inform healthcare decision-making: a pilot study of breast screening participation. Appl Econ 35:1073–1085
8. Canadian Institute for Health Information (2011). www.cihi.ca. Accessed Oct 13 2013
9. Wait Time Alliance (2012) Shedding Light on Canadians' Total Wait for Care - Report Card on Wait Times in Canada. Technical Report. http://www.waittimealliance.ca/media/2012reportcard/WTA2012-reportcarde.pdf. Accessed Oct 13 2013
10. Canadian Breast Cancer Foundation (2013) Technical Report. http://www.cbcf.org/central/AboutBreastCancerMain/AboutBreastCancer/Pages/BreastCancerinCanada.aspx. Accessed 18 Jan 2014
11. Mehrez A, Sinuany-Stern Z, Arad-Geva T, Binyamin S (1996) On the implementation of quantitative facility location models: the case of a hospital in a rural region. J Oper Res Soc 47:612–625
12. Harper P, Shahani A, Gallagher J, Bowie C (2005) Planning health services with explicit geographical considerations: a stochastic location-allocation approach. Omega 33:141–152
13. Lapierre SD, Myrick J, Russell G (1999) The public health care planning problem: a case study using geographic information systems. J Med Sys 23:401–417
14. Price W, Turcotte M (1986) Locating a blood bank. Interfaces 16:17–26
15. Schweikhart S, Smith-Daniels V (1993) Location and service mix decisions for a managed health care network. Socio-Econ Plann Sci 27:289–302
16. Branas C, MacKenzie E, ReVelle C (2000) A truama resource allocation model for ambulances and hospitals. Health Serv Res 35:489–507
17. Stahl JE, Kong N, Shechter SM, Schaefer AJ, Roberts MS (2005) A methodological framework for optimally reorganizing liver transplant regions. Med Decis Making 25:35–46
18. Cote M, Syam S, Vogel W, Cowper DC (2007) A mixed integer programming model to locate traumatic brain injury treatment units in the Department of Veterans Affairs: a case study. Health Care Manage Sci 10:253–267
19. Syam S, Cote M (2010) A location-allocation model for service providers with application to not-for-profit health care organizations. Omega 38:157–166
20. Gu W, Wang X, McGregor S (2010) Optimization of preventive healthcare facility locations. Int J Health Geo 9:1–16
21. Gunes E, Yaman H, Cekyay B, Verter V (2014) Matching patient and physician preferences in designing a primary care facility network. J Oper Res Soc 65:483–496
22. Chata S, Mayorga M, Kurz M, McLay L (2011) The minimum p-envy location problem: a new model for equitable distribution of emergency resources. IIE Tran Healthcare Sys Eng 1:101–115
23. Jia H, Ordonez F, Dessouky M (2007) A modeling framework for facility location of medical services for large-scale emergencies. IIE Tran 39:41–55
24. Mitropoulous P, Mitropoulous I, Giannikos I, Sissouras A (2006) A biobjective model for the location planning of hospitals and health centers. Health Care Manage Sci 9:171–179
25. Zhang Y, Berman O, Macotte P, Verter V (2010) A bilevel model for preventive healthcare facility network design with congestion. IIE Tran 42:865–880
26. Zhang Y, Berman O, Verter V (2012) The impact of client choice on preventive healthcare facility network design. OR Spec 34:349–370
27. Smith-Daniels V, Schweikhart SB, Smith-Daniels DE (1988) Capacity management in health care services: review and future research directions. Dec Sci 19:889–919
28. Rahman SU, Smith DK (2000) Use of location-allocation models in health service development planning in developing nations. Eur J Oper Res 123:437–452
29. Daskin M, Dean LK (2004) Location of health care facilities. In: Sainfort F, Brandeau M, Pierskalla WP (eds) Handbook of OR/MS in health care: a handbook of methods and applications. Kluwer, Boston, pp 43–76
30. Berg B (2013) Location models in healthcare. In: Denton B (ed) Handbook of healthcare operations management: methods. Springer Science and Business Media, New York, pp 387–402
31. Berman O, Krass D (2004) Facility location problems with stochastic demands and congestion. In: Drezner Z, Hamacher H (eds) Facility location: applications and theory. Springer, New York, pp 329–371
32. Boffey B, Galvao R, Espejo L (2007) A review of congestion models in the location of facilities with immobile servers. Eur J Oper Res 178:643–662