Capacitated preventive health infrastructure planning with accessibility-based service equity

Hongzhi Lin, Ph.D.[1]; Min Xu, Ph.D.[2]

Abstract

Hard-to-access health infrastructure is likely to lead to increased morbidity and mortality. The optimal layout of health facilities is undoubtedly of great significant for disease control and prevention. This study aims to propose a method to provide equitable access to capacitated preventive health facilities, which captures the key features of facility congestion in a competitive choice environment. The problem is formulated as a bilevel non-linear integer programming model. The upper level is a bi-objective programming model subject to investment budget constraint, where the primary objective is to minimize the maximum probability of balking (i.e., denied to access service) and the secondary objective is to minimize the maximum queueing time. The lower level is a user equilibrium analogous model resulting from the user choice of facility location. It determines the allocation of users to facilities by a defined generalized cost. An efficient heuristic algorithm is designed according to the bilevel structure where the genetic algorithm (GA) with elite strategy is developed to solve the upper level problem and the method of successive averages (MSA) is adopted to solve the lower level problem. An illustrative case study is employed to validate the performance of the proposed methods, and a number of interesting results and managerial insights are provided with sensitivity analysis.

Keywords: *health services, queueing, facility location, bilevel programming, user equilibrium*

1. Introduction

The health infrastructure planning is an essential part of urban planning. It usually means the planning of hospitals while the planning of preventive health facilities is

[1]Associate Professor, School of Economics and Management, Southeast University, Nanjing 211189, China. ORCID: https://orcid.org/0000-0002-4766-2621. Email:linhz@seu.edu.cn

[2]Assistant Professor, Department of Industrial and Systems Engineering, The Hong Kong Polytechnic University, Hung Hom, Hong Kong (Corresponding author). Email: xumincee@gmail.com

usually neglected. However, preventive health service, such as screening, examination, isolation, and vaccination, is necessary for urban development. It is of utmost importance since it can make massive savings on health expenditure by early detection. This is a painful lesson from the raging COVID-19 pandemic. In fact, before the current COVID-19 pandemic, three historically important epidemics had occurred since 2000: severe acute respiratory syndrome (SARS) in 2003, Middle East respiratory syndrome (MERS) in 2013, and Ebola virus disease in 2014. The arising monkeypox virus has already attracted our attention. The health issue is a real problem disturbing urban development. If the diseases can be detected and controlled earlier, the society would not suffer from huge economic damage and life losses. Therefore, the authorities around the world begin to realize the importance of preventive health facilities. In fact, the users usually face barriers in accessing appropriate, timely, and affordable preventive health service so far. The planning of preventive health infrastructure for disease control and prevention is an urgent problem that has practical implication for urban planning community.

A noticeable disparity in the accessibility to health facilities among different zones, however, is found in theory and practice. This paper tries to propose a method to design a health facility network for disease prevention, with the aim of improving service equity in terms of accessibility. The accessibility usually refers to a measure of the ease of reaching destinations or activities distributed in space. There are various ways to measure the spatial accessibility to facilities. Unlike the conventional definition of accessibility, the implication of accessibility here is straightforward and intuitive which is defined as the accessible demand. In fact, there are two sources of inaccessible demand: one is demand lost due to insufficient coverage and the other is demand lost due to congested facility (Abouee-Mehrizi et al., 2011; Berman et al., 2006). For the first source, demand is elastic with respect to cost and customers are usually assigned to the closest facilities to maximize system total demand (Berman and Drezner, 2006; Davari et al., 2016; Marianov, 2003; Zhang et al., 2010). For the second source, customers could be denied to access the service (i.e., occurrence of balking) upon their arrival, due to capable space. It is seldom explored in past studies because the incorporation of limited capacity is not easy. There are only two closely related references to the best of our knowledge. Marianov et al. (2008) studied the capacitated facility network design problem. They defined travel cost as the travel time and queueing delay but ignored the balking cost. This creates to the paradoxical situation where customers will choose facilities with a greater likelihood of balking, because it will reduce their overall time spent in the system while ignore the inaccessible demand. Motivated by this, Dan and Marcotte (2019) defined user utility considering additional balking cost and formulated a model to maximize the overall accessible demands. However, although system accessibility is maximized, the probability of balking between different facilities could be disparate. This could result in serious service inequity issue. Therefore, this study tries to propose a method to deploy capacitated health facilities to alleviate accessibility-based service inequity arising from balking.

The service equity issue is one of the critical problems concerned by the users, especially for public health services. Tao et al. (2014) and Zhang et al. (2016) proposed to locate health facilities by maximizing equity in accessibility. The disparity in accessibility to health facilities is noticed and optimized. They adopted a general definition of facility accessibility and minimized the variance of accessibility. Mousazadeh et al. (2018) suggested to design an accessible, stable, and equitable health service network where the equity is incorporated by maximizing the minimum service level of each residential zone. It is the well-known John Rawls's social justice approach where the welfare of the worst group is maximized. Filippi et al. (2021) found that the equitable treatment of users was usually neglected. They suggested a way to compromise between efficiency and equity. Pourrezaie-Khaligh et al. (2022) proposed a bi-objective approach for health facility location problem considering both equity and accessibility. The objective is to minimize system costs, maximize accessibility, and minimize inequality among all demand nodes. They employed the accessibility index introduced by Wang and Tang (2013). Different from conventional way of equity measured using the variance of individual accessibility, they defined equity based on a minimum envy criterion. All in all, although accessibility-based service equity has started few attention recently, the congestion effect and user choice behavior have not been incorporated yet.

Service facility location problems have been widely studied because of numerous real-life applications. Most literature is concerned with various versions of the problem where users are simply assigned to closest facilities, while sidesteps the important issue of user choice behavior, as well as the effect of congestion. In fact, the users have freedom to choose facilities. In addition to the travel time, the waiting time of a user at a congested facility also has a significant influence on her/his choice (Marianov et al., 2005; Marianov et al., 2008). It is a congestion game problem. From the perspective of user choice behavior, previous studies could fall into two categories: (i) system optimal models, where users are directed by a central decision-maker to optimize system performances; and (ii) user choice models, where users are free to choose a facility. The congestion at a facility is beginning to be introduced both. Let us give a brief overview on them separately.

The system optimal models accounts for the major part of facility location problems where the users are assigned to the closest facilities. They are also known as all-or-nothing allocation, or winner-takes-all allocation. Verter and Lapierre (2002) investigated the problem of locating preventive health facilities using a system optimal model. The travel time was assumed to be the only determinant of facility choice and users would go to the closest facility without considering the congestion effect. Although the users from the same residential node can be directed to different facilities in theory, the optimization problems will have an optimal solution where all-or-nothing allocation is adopted (Castillo et al., 2009). Zhang et al. (2009) further incorporated congestion effect at a facility where the users are assumed to visit the facility with minimum total cost including travel time and queueing time. The queueing time can also be incorporated as a constraint (Davari et al., 2016). Multi-objective location problems are also proposed recently where multiple performances are evaluated (Dogan et al., 2020; Erdoğan et al., 2019). As the outbreak of COVID-19, Risanger et al. (2021) recently proposed a system optimal model to select pharmacies for COVID-19 testing to ensure accessibility.

The user choice models are emerging ways of facility location problems. Most location models assume that all the demand originating at a particular node is served by the same closest facility. This is not so in competitive situations where the users are free to choose a facility. In this case, the users at each demand node may choose different facilities to patronize. The more attractive the facility for users at a certain demand node, the larger the percentage it captures the demand originating there. The formulation of user choice behavior is the foundation of facility network design. However, the user choice behaviors in facility location problems, are usually sidestepped intentionally or unintentionally. Although the literature concerning facility location is vast, few studies have incorporated user choice behaviors (Dan and Marcotte, 2019). Generally speaking, the user choice models can be classified into two categories: one is proportional allocation and the other is equilibrium allocation. The proportional allocation can be further classified into Huff-based allocation which can revert to a gravity model with pre-specified parameters (Gu et al., 2010; Tao et al., 2016) and logit-based allocation where a multinomial logit function is used to model the probability that users choose facility (Abouee-Mehrizi et al., 2011; Filippi et al., 2021; Kucukyazici et al., 2020). However, the proportional allocation cannot account for congestion effect. It is well-known that as a facility captures more users, it becomes more congested, resulting in longer queueing times. In fact, this effect makes the service facility less attractive and, consequently, user capture is reduced, leading to an eventual user equilibrium state where no user can further reduce his cost by unilaterally changing his behavior. Therefore, the equilibrium allocation was suggested recently which includes deterministic user equilibrium when utility is deterministic and stochastic user aquarium when stochastic utility is assumed (Dan and Marcotte, 2019; Zhang and Atkins, 2019). However, it is few incorporated due to the computation complexity. The facility location problem with equilibrium allocation is still a cutting-edge problem deserved to be explored.

This study makes four main theoretical and practical contributions for urban planning community. (i) We propose a way to improve the accessibility-based service equity for capacitated preventive health infrastructure planning. The equitable accessible flow is achieved by minimizing the maximum probability of balking. (ii) A bilevel decision structure is adopted where the upper level is urban planners and the lower level is facility users. The congestion effect is incorporated in the user utility function including queueing time and probability of balking. (iii) The users competing with each other will lead to user equilibrium state. An equivalent mathematical programming model is proposed to predict facility demand volumes at equilibrium state. (iv) A generic efficient and effective solution algorithm is proposed and validated, which is also applicable for other public service facility planning problems. (v) Several interesting findings and managerial insights for urban planners are provided based on computational experiments.

The remainder of this paper is organized as follows. Section 2 describes the problem and formulates it as a bi-objective bilevel programming model. Section 3 proposes a heuristic algorithm to solve the bilevel problem. Section 4 presents the computational results for the model with managerial insights. Finally, conclusions and future research directions are provide in section 5.

2. Problem modeling

Let $H = (\mathbf{N}, \mathbf{L})$ be a road network with a set of nodes $\mathbf{N}$ and a set of links $\mathbf{L}$.

The nodes represent either demand concentrations, facility locations, or road intersections, and links are main transportation arteries between nodes. We assume that the demand rate requiring preventive health service at population node $i(i \in \mathbf{N})$ follows the Poisson process with an average rate h_i. The set of candidate locations for health facilities is $\mathbf{M}$, and $\mathbf{S}$ is the set of chosen locations where $\mathbf{S} \subset \mathbf{M}$. The shortest path travel time from demand node $i(i \in \mathbf{N})$ to facility location $j(j \in \mathbf{M})$ is denoted by t_{ij}. The government has a limited budget B that can be used to build facilities with associated servers. We assume that servers at all of the facilities are the same, and service time is exponentially distributed providing service to μ clients per unit of time on average. We also assume that clients are homogenous, their arrivals to each facility follow poisson distributions and the queueing discipline is first-come first served (FCFS). These assumptions are reasonable for walk-in facilities, which applies to most routine health services in many countries or regions. Thus, facility $j(j \in \mathbf{M})$ here is assumed to behave as a $M / M / s_j / K_j$ queueing system, where M denotes Markovian (or poisson) arrivals or departures distribution, or equivalently exponential interarrival or service time distribution, s_j denotes the number of servers at facility j, and K_j is the capable number of clients at facility j due to physical constraint. Whenever there are K_j clients at facility j, any arriving client is denied to access and leaves the system as a lost client. The value of K_j is predetermined for each facility location, depending on specific conditions. This assumption is not loss of generality since it could be extended to other queueing system based on estimation from available data.

The problem is to make location and associated capacity decisions, with the aim of equitable probability of balking, subject to the budget constraint B. Three sets of decision variables are defined as follows:

$$y_j = \begin{cases} 1 & \text{if a facility is opened at location } j, \ \forall j \in \mathbf{M}, \\ 0 & \text{otherwise,} \end{cases}$$

s_j = number of servers at facility location j, $\forall j \in \mathbf{M}$,

x_{ij} = number of clients from demand node i to location j, $\forall i \in \mathbf{N}$, $j \in \mathbf{M}$.

Therefore, for a chosen set $\mathbf{S} = \{ j : j \in \mathbf{M}, y_j = 1 \}$, we have

$$\sum_{j \in \mathbf{S}} x_{ij} = h_i, \quad \forall i \in \mathbf{N}. \tag{1}$$

Let λ_j denotes the arrival rate of clients at facility j, $\forall j \in \mathbf{M}$, then we have

$$\lambda_j = \sum_{i \in \mathbf{N}} x_{ij}, \quad \forall\, j \in \mathbf{M}. \tag{2}$$

It defines the demand at each facility as the sum of demands originating from all the demand nodes. Given the arrival rate λ_j and the number of servers s_j at facility j, the probability that there are n clients in the queue is

$$p_{nj}(\lambda_j, s_j) = \begin{cases} \dfrac{\rho_j^n}{n!} p_{0j}, & \text{if } 0 \le n \le s_j, \\ \dfrac{\rho_j^n}{s_j!\, s_j^{n-s_j}} p_{0j}, & \text{if } s_j \le n \le K_j, \end{cases} \tag{3}$$

where $\rho_j = \lambda_j / \mu$ is the intensity of the queueing process and the probability of no client is

$$p_{0j} = [1 + \sum_{n=1}^{s_j} \frac{\rho_j^n}{n!} + \frac{\rho_j^{s_j}}{s_j!} \sum_{n=s_j+1}^{K_j} (\frac{\rho_j}{s_j})^{n-s_j}]^{-1}. \tag{4}$$

Note that the probability p_{nj} at each facility j is a function of λ_j and s_j. The notation p_{Kj} is the probability of balking owing to a limited space. It allows a facility's arrival rate to exceed its service rate, without unbounded grow of queue length. The effective arrival rate, i.e., the number of clients who could access the service, is denoted by $\bar{\lambda}_j$. There is,

$$\bar{\lambda}_j = \lambda_j (1 - p_{Kj}), \quad \forall j \in \mathbf{M}. \tag{5}$$

2.1 The user utility function

Clients are assumed to patronize a facility that maximizes their individual utility, i.e., minimizes their generalized costs. Therefore, it is critical to understand how clients make their choices. Let us now present our user choice modeling, which essentially establishes a utility function depending on the attractiveness of a facility that they are aware of. Let U_{ij} denote the observed utility of users from demand node i receiving the service at facility location j. It mainly comprises four components: (i) u_j, a constant attraction of location j, which might include intrinsic factors such as parking convenience, practitioner reputation, service quality, etc.; (ii) t_{ij}, the shortest path travel time from origin node i to destination facility j; (iii) $w_j(\lambda_j, s_j)$, the average dwell time at location j including queueing time and service time, which is a function of arrival rate λ_j and server number s_j; and (iv) $p_{Kj}(\lambda_j, s_j)$, the probability of unmet service (i.e., balking) due to physical constraint. Note that w_j and p_{Kj} are continuous functions with respect to λ_j and s_j.

As it is an $M/M/s_j/K_j$ queueing system at facility j, for any $s_j \geq 1$, the average dwell time $w_j(\lambda_j, s_j)$ could be given by the following set of equations according to the classical queueing theory:

$$w_j(\lambda_j, s_j) = \frac{L_j}{\bar{\lambda}_j}, \quad \forall j \in \mathbf{S}, \tag{6}$$

$$L_j = \sum_{n=s_j}^{K_j} (n - s_j) p_{nj} + \rho_j (1 - p_{Kj}), \quad \forall j \in \mathbf{S}, \tag{7}$$

where L_j is the average length of the queue in terms of client number, $\bar{\lambda}_j$ is the effective arrival rate according to Eq. (5), p_{nj} is the probability of having n clients at the facility according to Eq. (3), and ρ_j is the intensity of service as defined previously. Eq. (6) is the famous Little's formula in queueing theory.

The way of integrating utility could be various. Following the conventional way in the literature, we assume a linear additive functional form of U_{ij} to incorporate the above four components with different weights. It is a standard assumption in the utility theory. In addition, it is also reasonable to assume that U_{ij} is positively associated with benefit u_j but negatively associated with cost t_{ij}, $w_j(\lambda_j, s_j)$, and $p_{Kj}(\lambda_j, s_j)$. In this framework, U_{ij} is given by (Dan and Marcotte, 2019):

$$U_{ij} = u_j - \beta_1 t_{ij} - \beta_2 w_j(\lambda_j, s_j) - \beta_3 p_{Kj}(\lambda_j, s_j), \quad \forall i \in \mathbf{N}, \ j \in \mathbf{S}, \tag{8}$$

where β_1 and β_2 denote the coefficients of the travel time and queueing time respectively, and β_3 is interpreted as the price of service inaccessibility. In practice, parameters β_1, β_2, and β_3 can be estimated empirically using realistic surveys. The different weights on travel time and waiting time could be possible and are allowed, given the different perceptions of clients for them. The definition of real values for these parameters is outside the scope of this paper. Note that besides these specific parts, the utility function can also be extended to incorporate other observable attributes, such as the parking cost and service price, depending on available data.

The users interacts with each other until no one person could increase his utility by unilaterally changing his facility choice, which is known as Nash equilibrium state. Mathematically it is important to note the interdependency between the arrival rate λ_j and the expected waiting time $w_j(\lambda_j, s_j)$ and the probability of balking $p_{Kj}(\lambda_j, s_j)$. According to our modelling framework, λ_j is the sum of x_{ij}, which depends on U_{ij}, which further depends on $w_j(\lambda_j, s_j)$ and $p_{Kj}(\lambda_j, s_j)$. That is, the value of λ_j

depends on itself indirectly. Since we consider a network of competitive facilities, it implies that we need address a Nash equilibrium problem to determine demand allocation x_{ij} given facility locations and associated capacities. Specifically, it is better known as user equilibrium problem.

2.2 The user equilibrium model

It is assumed that clients always choose the facility with the highest observed utility. The clients are assumed to re-evaluate their utilities after several times of visits. They could also learn from others by social network for example. Therefore, they are assumed to know about the queues and capacities of the facilities to make near-optimal decisions. The competition between clients will reach a user equilibrium state finally. Let $\bar{U}_i$ denote the highest utility of clients at demand node i, i.e.,

$$\bar{U}_i = \max_{j \in M} U_{ij}, \quad \forall i \in \mathbf{N}. \tag{9}$$

Given the determined location $\mathbf{S}$ and capacities s_j, $\forall j \in \mathbf{S}$, no client wants to change her/his facility choice at user equilibria. Therefore, the equilibrium condition can be characterized by the following complementarity system

$$U_{ij}^* = u_j - \beta_1 t_{ij} - \beta_2 w_j(\lambda_j^*, s_j) - \beta_3 p_{Kj}(\lambda_j^*, s_j) \begin{cases} = \bar{U}_i^* & \text{if } x_{ij}^* > 0 \\ \le \bar{U}_i^* & \text{if } x_{ij}^* = 0 \end{cases}, \forall i \in \mathbf{N},\ j \in \mathbf{S}, \tag{10}$$

where U_{ij}^* and $\bar{U}_i^*$ denote the utility of clients from demand node i visiting preventive health facility j and the highest utility of clients from demand node i at user equilibrium state, respectively. Moreover, it should be noted that

$$\lambda_j^* = \sum_{i \in \mathbf{N}} x_{ij}^*, \quad \forall j \in \mathbf{S},$$

where λ_j^* denotes the arrival rate of clients at facility j at user equilibrium state, and x_{ij}^* denotes the allocated number of clients from demand node i to facility location j at user equilibrium state.

The equilibrium condition (10) means that if there is a client flow from demand node i to facility location j, then U_{ij}^*, the utility of users from node i to facility j, must be equal to the highest utility $\bar{U}_i^*$; otherwise, it is no more than the highest. It implies that each user patronizes the facility with the highest observed utility. Accordingly, at equilibrium state, users issued from a common origin node will experience identical utilities, thus achieving a well-known Nash equilibrium state. They cannot improve their utility by changing facility choice.

To find λ_j^* and implicit x_{ij}^* in Eq. (10) given determined location $\mathbf{S}$, we can solve the following equivalent nonlinear mathematical programming with symmetric Jacobin matrix of utility function:

$$\max_{\mathbf{x}} \ Z(\mathbf{x}\,|\,\mathbf{S}) = \sum_{i\in\mathbf{N}}\sum_{j\in\mathbf{S}}\int_0^{\lambda_j} U_{ij}(\omega, s_j)d\omega \tag{11}$$

subject to

$$\sum_{j\in\mathbf{S}} x_{ij} = h_i, \quad \forall j\in\mathbf{S}, \tag{12}$$

$$x_{ij} \ge 0, \quad \forall i\in\mathbf{N},\ j\in\mathbf{S}, \tag{13}$$

where

$$\lambda_j = \sum_{i\in\mathbf{N}} x_{ij}, \quad \forall i\in\mathbf{N},\ j\in\mathbf{S}. \tag{14}$$

Theorem 1. *Given the determined location* $\mathbf{S}$*, the mathematical programming (11) -(14) is equivalent to equilibrium condition (10).*

Proof. In order to prove that the mathematical programming is equivalent to Eq. (10), we reformulate the model as a Lagrange function with nonnegative constraints only, i.e.,

$$\begin{aligned} &F = Z(\mathbf{x}\,|\,\mathbf{S}) - \sum_{i\in\mathbf{N}} w_i\Big(\sum_{j\in\mathbf{S}} x_{ij} - h_i\Big) \\ &s.t. \quad x_{ij} \ge 0, \ \ \forall i\in\mathbf{N}, \ \ j\in\mathbf{S}, \end{aligned} \tag{15}$$

where w_i is a Lagrange multiplier of constraint (12).

According to Karush–Kuhn–Tucker (KKT) conditions, the optimal conditions of this Lagrange function are given by

$$x_{ij}\frac{\partial F}{\partial x_{ij}} = 0, \quad \forall i\in\mathbf{N}, j\in\mathbf{S}, \tag{16}$$

$$\frac{\partial F}{\partial x_{ij}} \le 0, \quad \forall i\in\mathbf{N}, j\in\mathbf{S}, \tag{17}$$

$$\frac{\partial F}{\partial w_i} = 0, \quad \forall i\in\mathbf{N}, \tag{18}$$

$$x_{ij} \geq 0, \quad \forall i \in \mathbf{N}, j \in \mathbf{S}. \tag{19}$$

It is straightforward to find that Eq. (18) is equivalent to Eq. (12). Eqs. (16) and (17) imply that

$$\begin{aligned} &\text{if } x_{ij} > 0, \ \frac{\partial F}{\partial x_{ij}} = 0, \quad \forall i \in \mathbf{N}, j \in \mathbf{S}, \\ &\text{if } x_{ij} = 0, \ \frac{\partial F}{\partial x_{ij}} \leq 0, \quad \forall i \in \mathbf{N}, j \in \mathbf{S}. \end{aligned} \tag{20}$$

Note that since we have

$$\begin{aligned} \frac{\partial L}{\partial x_{ij}} &= \frac{\partial}{\partial \lambda_j}[\sum_{i \in \mathbf{N}} \sum_{j \in \mathbf{S}} \int_0^{\lambda_j} U_{ij}(\omega, s_j) d\omega] \frac{\partial \lambda_j}{\partial x_{ij}} - \frac{\partial}{\partial x_{ij}} (\sum_{i \in \mathbf{N}} w_i (\sum_{j \in \mathbf{S}} x_{ij} - h_i)) \\ &= U_{ij} - w_i, \end{aligned} \tag{21}$$

Eq. (20) can be further rewritten as follows:

$$\begin{aligned} &\text{if } x_{ij} > 0, \ U_{ij} - w_i = 0, \quad \forall i \in \mathbf{N}, j \in \mathbf{S}, \\ &\text{if } x_{ij} = 0, \ U_{ij} - w_i \leq 0, \quad \forall i \in \mathbf{N}, j \in \mathbf{S}. \end{aligned} \tag{22}$$

It can be also reformulated in the following complementary form:

$$(U_{ij} - w_i) x_{ij} = 0, \quad \forall i \in \mathbf{N}, j \in \mathbf{S}, \tag{23}$$

$$U_{ij} - w_i \leq 0, \quad \forall i \in \mathbf{N}, j \in \mathbf{S}, \tag{24}$$

$$x_{ij} \geq 0, \quad \forall i \in \mathbf{N}, j \in \mathbf{S}. \tag{25}$$

It can be seen that Eq. (22) means that if there is client flow, i.e., $x_{ij} > 0$, the utility U_{ij} will be equal to w_i, and if there is no client flow, i.e., $x_{ij} = 0$, the utility U_{ij} is no more than w_i. Therefore, the Lagrange multiplier w_i can be interpreted as the highest utility $\bar{U}_i^*$ incurred by clients at demand node i. Hence, Eq. (22) is equivalent to Eq. (10). Therefore, we can conclude that the solution of the mathematical programming (11)-(14) satisfies the equilibrium condition (10). The proof of the theorem is complete.

2.3 The bilevel programming model

The entire problem considered here is a bilevel decision structure where the upper level problem is the determination of facility locations and associated capacities by urban planners, and the lower level problem is the determination of equilibrium flows of users from demand nodes to facility locations given the upper level decisions. Note that the equilibrium flows x_{ij}, as well as the arrival rate λ_j are not decision variables. They are determined endogenously by the lower level model. The decision variables are the location variables y_j and associated capacities s_j in the upper level model. Once the values of these variables are fixed, all of the remaining auxiliary variables and parameters can be computed.

There usually is a limited investment budget to support the establishment and operation of the preventive health facilities in practice. This budget constraint can be used to incorporate the cost differences of establishing and operating facilities at different locations of an urban area. The budget is set to be B. Let c_j^f be the fixed cost of establishing a facility at location $j \in \mathbf{M}$ and c^v be the unit operation cost of adding a server to a facility that is identical for each location. In addition, for cost effectiveness, we assume that facilities cannot be operated unless the number of their clients exceeds a minimum threshold $R_{\min}$. Moreover, the number of servers at facility j cannot exceed an upper bound $\hat{s}_j$ due to physical condition. The value of $\hat{s}_j$ is typically given by the urban planner on the basis of specific conditions and may differ from location to location.

Bi-objective optimization is adopted in the upper level model where the upper level is urban planners and the lower level is facility users. In order to formulate service network design problem considering accessibility-based equity, the maximum probability of balking p_{Kj}, $\forall j \in \mathbf{M}$, is minimized. It is regarded as the primary objective. As there are possible chances that the primary objective is always zero for unsaturated flows, a secondary objective is introduced to minimize the maximum waiting time at a facility in order to reach equitable queueing. Therefore, the upper level model of service network design problem can be formulated as follows:

$$\text{Primary Objective} \quad \min \ E_1(\mathbf{S}) = \max\left\{ p_{Kj}, \forall j \in \mathbf{M} \right\} \tag{26}$$

$$\text{Secondary Objective} \quad \min \ E_2(\mathbf{S}) = \max\left\{ w_j(\lambda_j, s_j), \forall j \in \mathbf{M} \right\} \tag{27}$$

subject to

$$s_j \geq y_j, \quad \forall\, j \in \mathbf{M}, \tag{28}$$

$$s_j \leq \hat{s}_j y_j, \quad j \in \mathbf{M}, \tag{29}$$

$$\sum_{i \in \mathbf{N}} x_{ij} = \lambda_j, \quad \forall j \in \mathbf{M}, \tag{30}$$

$$x_{ij} \leq y_j, \quad \forall i \in \mathbf{N}, \quad j \in \mathbf{M}, \tag{31}$$

$$\bar{\lambda}_j = \lambda_j (1 - p_{Kj}), \quad \forall j \in \mathbf{M}, \tag{32}$$

$$\lambda_j \geq R_{\min} y_j, \quad \forall j \in \mathbf{M}, \tag{33}$$

$$\sum_{j \in \mathbf{M}} c_j^f y_j + c^v \sum_{j \in \mathbf{M}} s_j \leq B, \tag{34}$$

$$y_j \in \{0,1\}, \quad s_j \in \mathbf{Z}^+, \quad \forall\, j \in \mathbf{M}. \tag{35}$$

where x_{ij} is determined by the following lower level model after the location variables y_j and associated capacity variables s_j are determined:

$$\max_{\mathbf{x}} \; Z(\mathbf{x} \,|\, \mathbf{S}) = \sum_{i \in \mathbf{N}} \sum_{j \in \mathbf{S}} \int_0^{\lambda_j} U_{ij}(\omega, s_j) d\omega \tag{36}$$

subject to

$$\sum_{j \in M} x_{ij} = h_i, \quad \forall j \in \mathbf{S} \tag{37}$$

$$x_{ij} \geq 0, \quad \forall i \in \mathbf{N}, \quad j \in \mathbf{S}. \tag{38}$$

The primary objective function (26) is to minimize the maximum probability of balking and the secondary objective function (27) is to minimize the maximum queueing time. They are both Min-Max optimization problems so as to reach service equity, which is robust for any level of demands. Constraints (28) ensure the assignment of at least one server to each open facility. Constraints (29) limit the number of servers not exceeding $\hat{s}_j$. Constraints (30) define the arrival rate λ_j. Constraints (31) ensures that clients can only obtain the service from open facilities. Constraints (32) are the definition of effective arrival rates. Constraints (33) stipulate that the arrival rate at an open facility must satisfy the minimum workload requirement. Constraint (34) is the budget and Constraints (35) define the feasible domain of decision variables y_j and s_j.

3. Solution method

Since the bi-objective bilevel programming model is highly nonlinear and contains integer decision variables, it poses big challenges to solve the model exactly. Therefore, the focus of this study is to adopt efficient and effective heuristic algorithms that have many successful applications for facility network design problem (Auerbach and Kim, 2021; Chambari, et al, 2011; Ershadi and Shemirani, 2021; Zhang and Atkins, 2019). The bilevel framework is carefully followed by our solution method. For the upper level location problem, a meta-heuristic, generic algorithm (GA) with elite strategy, is proposed to find the optimal locations and associated capacities. However, we definitely believe that the more advanced heuristics will improve the computation efficiency. For the lower level allocation problem, we adopt method of successive averages (MSA) to solve the user equilibrium model. This demand allocation algorithm determines the equilibrium flows of users to facilities after the upper level decisions are confirmed. Thus, the demand allocation algorithm serves as an embedded module for the facility location algorithm. For the ease of easier understanding, we describe the demand allocation algorithm first.

3.1 Demand allocation algorithm for the lower level model

Given the upper level facility decisions $\mathbf{S}$ and s_j, $\forall j \in \mathbf{S}$, the lower level problem of the user choice model is to find the equilibrium flows. The adopted algorithm is a kind of iterative method, known as MSA. Let k be the iteration index and K be a maximum iteration number. In addition, let ε be a predetermined error tolerance parameter, and $\theta_k \in (0,1)$, $k = 1,\ldots,K$, be a step-length parameter at iteration k. The specific computation steps are listed below:

Step 0 (Initialization): Set the values of ε and K; initiate $k = 0$; set initial allocation

$$x_{ij}^0 = \frac{h_i}{|\mathbf{S}|}, \forall i \in \mathbf{N}, j \in \mathbf{S}.$$

Step 1 (Calculation of utility): Update $k := k + 1$; calculate λ_j, $\forall j \in \mathbf{S}$, from Eq. (2); calculate the shortest path travel time t_{ij}, $\forall i \in \mathbf{N}$, $j \in \mathbf{S}$, using Dijkstra's algorithm; calculate probability of balking $p_{Kj}(\lambda_j, s_j)$ from Eq. (3), effective arrival rate $\bar{\lambda}_j$ from Eq. (5), waiting time $w_j(\lambda_j, s_j)$ from Eq. (6); calculate U_{ij}, $\forall i \in \mathbf{N}$, $j \in \mathbf{S}$, from Eq. (8); find $\bar{U}_i$, $\forall i \in \mathbf{N}$, from Eq. (9).

Step 2 (All-or-nothing allocation): Set flow x'_{ij} by all-or-nothing rule as follows, i.e., allocate all clients from the same demand node to the most attractive facility, i.e.,

$$x'_{ij} = \begin{cases} h_i & \text{if } U_{ij} = \bar{U}_i \\ 0 & \text{if } U_{ij} < \bar{U}_i \end{cases}, \quad \forall i \in \mathbf{N}, j \in \mathbf{S}.$$

Step 3 (Generation of search direction): Define $d_{ij} = x'_{ij} - x_{ij}^{k-1}$, $\forall i \in \mathbf{N}$, $j \in \mathbf{S}$, as a search direction.

Step 4 (Flow update): Update client flow $x_{ij}^{k} = x_{ij}^{k-1} + \theta_k d_{ij}$, $\forall i \in \mathbf{N}$, $j \in \mathbf{S}$, where θ_k is the step-length parameter given by,

$$\theta_k = \frac{1}{k+1}.$$

Step 5 (Stopping criteria): If the relative difference between x_{ij}^{k} and x_{ij}^{k-1} is equal or less than ε, or $k \geq K$, set $x_{ij} := x_{ij}^{k}$ and stop; otherwise, go to Step 1. The relative error is defined as,

$$\frac{\| x_{ij}^{k} - x_{ij}^{k-1} \|}{\| x_{ij}^{k-1} \|} \leq \varepsilon, \forall i \in \mathbf{N}, j \in \mathbf{S}.$$

Step 6 (Return results): Return the incumbent solution to the upper level model, including equilibrium flows, balking probabilities, and waiting times.

The suggested method in each iteration identifies a new search direction for x_{ij} in Step 3 and then updates x_{ij} by a step-length in Step 4. The procedure continues until one of the stopping conditions in Step 5 is met. The step-length θ_k in each iteration is determined in advance. There are a variety of ways to set θ_k. To achieve convergence, θ_k should decrease with k and locate between zero and one. Here we set θ_k as the reciprocal of the iteration number $(k+1)$ as usual. It is worth noting that the x_{ij}^{k} updated in Step 4 may result in an arrival rate at a facility exceeds the capable space allowed. At this situation, excess clients will be denied to access health services and become lost demand.

3.2 Facility location algorithm for the upper level model

We develop a genetic algorithm with elite strategy to solve the upper level problem, because it is one of the most popular meta-heuristics for addressing combinatorial optimization problems with many successful applications. It has the ability to explore other parts of the feasible space while avoiding local optima. Although it is time-consuming and the global optima is not guaranteed mathematically, it is still widely used for nonlinear programming problems.

In genetic algorithms, each chromosome represents a solution to the problem, and the quality of a solution is measured by a fitness value. Note that since it is a bi-objective optimization problem, the primary objective works as the first fitness value, and the secondary objective is adopted when there are equal primary objectives. In this study, an integer coding technique is employed to define a chromosome. Each chromosome is made up of several genes that are nonnegative integer numbers. Each gene corresponds to a candidate location in M, and its value represents the number of servers. If there is no server at a location, the facility is not opened at that location. The following is how we implement the genetic algorithm with elite strategy:

Step 0 (Initialization): Set the used parameters, including the population size P, the maximum number of generations G, the crossover probability p_c, the mutation probability p_m, the label of generation $g=1$, and the fraction of elite p_e.

Step 1 (Generation of initial population): Randomly generate P feasible solutions as an initial population of chromosomes, scattering the entire range of possible solutions. If one chromosome is not feasible according to the constraints, generate another one until a feasible solution is found.

Step 2 (Calculation of fitness value): For each chromosome in the population, the value of fitness is generated that is the objective function value. It is used to evaluate the quality of each chromosome in the population. Note that there are two objective functions in the upper level model. One is primary objective and the other is secondary objective. Therefore, there are two fitness values in order.

Step 3 (Generation of new population):

Step 3.1 (Selection): According to the values of fitness evaluated in Step 2, the best fraction p_e is labeled for elites, and the worst fraction p_e is discarded. A stratified sequencing method is used here where the primary objective value is sorted first and the secondary objective value is sorted next.

Step 3.2 (Crossover): The remaining $(1-p_e)P$ chromosomes are used for crossover operation. These chromosomes are matched in pairs randomly. The probability of carrying out the crossover is p_c. If the two parent chromosomes are chosen for crossover, a gene location is randomly identified to across over to generate two off-springs as new chromosomes. If newborn chromosomes are not feasible according to constraints in the upper level model, try another gene location until they are feasible.

Step 3.3 (Mutation): A chromosome is determined for mutation with probability p_m. Randomly choose two genes with at least one positive, and interchange their values. If the new chromosome is not feasible, try another two gene locations until a feasible off-spring is generated.

Step 3.4 (Elitism): Form a new generation. After genetic operations, there are still $(1-p_e)P$ feasible chromosomes. The labeled p_eP elites are added to ensure the population size P. This allows the best chromosomes from the current generation to carry over the next generation unaltered. It guarantees that the solution quality will not decrease from one generation to the next. Update the notation of generation be $g := g+1$.

Step 4 (Stopping criterion): If the maximum number of generations G is achieved, i.e., $g \geq G$, terminate the iteration process and output the results. Otherwise, turn to Step 2.

4. Computational experiments

4.1 An illustrative case

We conduct a computational experiment to assess the performance of proposed model and algorithm with Sioux Falls network. This network has been widely used for validation in the network design problems. It is a medium sized network as depicted in Fig. 1. The network consists of 24 nodes and 76 links. In the computational experiments, it is assumed that there are 8 population nodes and 8 potential locations in the region. Therefore, there are a total number of 64 origin-destination (O-D) pairs. The travel time and length of each link are given in Table 1. The link length can be converted to the link travel time, by assuming a constant link travel speed of 30 miles/hour. Recognizing that clients are only a very small part of road travelers, the facility choice and route choice of clients is assumed not to affect road travel times. That is, the link travel times are constants. This is different with classical road network design problems as the congestions take place in facilities other than roads. The preventive health demand data, i.e., the number of clients per hour (clients/hr), are listed in Table 2. The demand for preventive health services is fixed at origin zones while the trip distribution is not fixed and it is determined by facility planning. The clients have freedom to choose their favorite facilities.

Fig. 1. The Sioux Falls test network

Insert Table 1 here

Insert Table 2 here

Based on the proposed model and solution method, the following parameter values are used in the case study.

Problem parameters

- the service rate of each server $\mu = 6$ clients/hr;
- the constant facility attraction $u_j = 0$;
- the coefficient to travel time $\beta_1 = 1$ and that to waiting time $\beta_2 = 1$;
- the price of service inaccessibility $\beta_3 = 1$;
- the maximum number of servers $\hat{s}_j = 10$;
- the fixed establishment cost $c_j^f = 0$;
- the unit cost of a server $c^v = 1$;
- the budget $B = 40$;
- the minimum workload $R_{\min} = 10$ clients/hr;

Method of successive averages parameters

- the maximum iteration number $K = 100$;
- the error tolerance $\varepsilon = 0.01$;

Genetic algorithm parameters

- the population size $P = 100$;
- the maximum number of generations $G = 20$;
- the crossover probability $p_c = 0.5$;
- the mutation probability $p_m = 0.2$;
- the fraction of elite $p_e = 0.1$.

The algorithms are coded using a free open-source language R 3.6.3. All runs are performed at a personal computer with 3.6 gigahertz Intel i7-4790 CPU and 16 gigabytes RAM. The genetic algorithm stopped after 1.42 hours for this case study. The evolutionary process begins to be stable after 19 generations as shown in Fig. 2. It can be concluded that the final results are satisfying solutions. The selected locations to set up preventive health facilities are nodes 3, 9, 16, 19, and 23. Their associated number of servers are 6, 9, 8, 7, and 10 correspondingly. The service quality of each facility is shown in Table 3. It shows that the maximum probability of balking is 0.132 in node 3 and the maximum waiting time is 1.22 hours in node 23. It can be concluded that the service quality among all facilities is quasi-equal in terms of balking probability and waiting time. The accessibility-based service equity is achieved which is the policy goal.

Fig. 2. The evolutionary process of genetic algorithm

Insert Table 3 here

The demand allocation at equilibrium state is presented in Table 4. It shows that clients from the same demand node usually patronize the same facility such as nodes 1, 2, 4, 5, 13, and 14, even if they are free to head for different facilities. However, the clients are possible to be assigned to more than one facility such as nodes 19 and 23 if their utilities are quasi-equal.

Insert Table 4 here

4.2 Sensitivity analysis

It is always beneficial to do a sensitivity analysis which could provide valuable managerial insights. We conduct a sensitivity analysis with varying budget control here, which is also a cost-benefit analysis in economics. The budget is increased from 30 to 60 at step-length 5. The results are shown in Fig. 3 where the horizontal axis is budget and the vertical axis is maximum probability of balking among all of facility locations. At the very beginning, the maximum probability of balking is 37.4% with budget 30. It is a low level of service that is difficult to accept. It is no doubt that the probability of balking decreases with budget. The maximum probability of balking is decreased to 2.2% with budget 45. The maximum waiting time is 1.67 hours at this time. Whether the budget is good enough depends on the policy makers. The probability of balking will continue to decrease until zero. After budget 50, the customers will not be denied to access facilities, which are unsaturated flows. The service network is not that congested. There are enough vacancies for clients. Since then, the secondary objective, minimizing the maximum waiting time, will play an important role as the primary objective will not move forward and keep zero. Therefore, the proposed methods are robust for capacitated facility location problems.

Fig. 3. A sensitivity analysis with varying budget

It is also interesting to do a sensitivity analysis on demand with given infrastructure investment budget. The demand is fixed in the short run, but it can change with time in the long run. In order to investigate the benefit of investment budget, a demand expansion coefficient is adopted varying from 0.7 to 1.3 at step-length 0.1. The budget is set to be 40. The results are shown in Fig. 4 where the horizontal axis is varying demand and the vertical axis is maximum probability of balking. At the very beginning, there is no balking and the clients will not be denied to access health service because demand is insufficient. The maximum waiting time is 0.546 hours when demand expansion coefficient is 0.7. If the demand becomes even less ,the probability of facility idleness will increase, which means there is a waste of investment. It is undoubted that the maximum probability of balking will increase with demand. When demand expansion coefficient is 1.3, the maximum probability of balking will increase to 35.3%. If the probability is unacceptable, more infrastructure investment is needed.

Fig. 4. A sensitivity analysis with varying demand

5. Conclusions

Preventive health services can detect serious diseases at early stage and make a lot of savings on health expenditures. It is critical for urban sustainable development as shown by the current COVID-19 pandemic. The authorities realize to improve the level of preventive health services to avoid expensive social cost. Noticed the disparity in the accessibility to health facilities among different zones, this study proposes a bilevel programming model to improve the accessibility-based service equity for health infrastructure planning problems. The facilities are capacitated where a customer observes the queue on arrival and leaves if there are no vacancies. In the upper level model, a bi-objective programming model is adopted to descript the urban planner where the primary objective is to minimize the maximum probability of balking and the secondary objective is to minimize the maximum queueing time subject to an investment budget. In the lower level model, a deterministic user equilibrium model is adopted to descript facility users, which is formulated as an equivalent mathematical programming problem. The user utility is defined to include travel time, queueing time, and the probability of balking for capacitated health facilities. The solution method is designed to correspond to the bilevel decision framework where a genetic algorithm with elite strategy is adopted for the upper level model and the method of successive averages is used for the lower level model. Note that a stratified sequencing method is used in genetic algorithm where the primary objective value is sorted first and the secondary objective value is sorted next.

To validate the proposed methods, we conduct several computational experiments and derive some interesting managerial insights. We find that the proposed methods are efficient and effective. These methods can reach a satisfactory solution in a reasonable computation time. It is found that although the clients from the same demand node may visit more than one facility, they usually visit the same facility. This is caused by deterministic user equilibrium. The results also indicate that the service quality is quasi-equal in terms of balking probability. The bi-objective method is robust for any level of budget and demand. The sensitivity analysis with varying budget shows that the maximum probability of balking decreases with budget. However, the marginal benefit is decreasing. There is an optimal budget beyond which further increment of investment will not offset its benefits. On the other hand, the sensitivity analysis with varying demand shows that more investment is desired for expanded demand in order to maintain a certain level of service.

This study would be improved in several ways in the near future. First, we would like to find a more realistic case with empirical data to show how our methods can be applied in practice. Second, the user utility function will be extended to include other observable attributes, such as the parking time, the service quality, the service price, etc. The formulation of user choice behavior would be more realistic. Third, there are some other advanced heuristics used for similar problems, such as vibration damping optimization algorithm and cutting plane algorithm. It would be interesting to conduct a comparison of results across different heuristics. Last but not the least, the unobserved utility will be incorporated in terms of a random term. Then the stochastic user equilibrium can be adopted to substitute the deterministic user equilibrium.

Declarations of interest

None

Data Availability Statement

Some or all data, models, or code that support the findings of this study are available from the corresponding author upon reasonable request.

Acknowledgements

This work was supported by the National Social Science Fund of China [grant number 21BGL233]. The second author appreciates the support from the National Natural Science Foundationof China (Grant No. 71901189) and the Hong Kong

Polytechnic University (ZVTK).

References

Abouee-Mehrizi, H., Babri, S., Berman, O., Shavandi, H. (2011) Optimizing capacity, pricing and location decisions on a congested network with balking. *Mathematical Methods of Operations Research* 74, 233-255.

Auerbach, J. D., Kim, H. (2021). Local network connectivity optimization: an evaluation of heuristics applied to complex spatial networks, a transportation case study, and a spatial social network. *PeerJ Computer Science* 7, e605.

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

Berman, O., Krass, D., Wang, J. (2006) Locating service facilities to reduce lost demand. *IIE Transactions* 38, 933-946.

Castillo, I., Ingolfsson, A., Sim, T. (2009) Social optimal location of facilities with fixed servers, stochastic demand, and congestion. *Production and Operations Management* 18, 721-736.

Chambari, A., Rahmaty, S. H., Hajipour, V., Karimi, A. (2011) A bi-objective model for location-allocation problem within queuing framework. *International Journal of Computer and Information Engineering* 5, 555-562.

Dan, T., Marcotte, P. (2019) Competitive facility location with selfish users and queues. *Operations Research* 67, 479-497.

Davari, S., Kilic, K., Naderi, S. (2016) A heuristic approach to solve the preventive health care problem with budget and congestion constraints. *Applied Mathematics and Computation* 276, 442-453.

Dogan, K., Karatas, M., Yakici, E. (2020) A model for locating preventive health care facilities. *Central European Journal of Operations Research* 28, 1091-1121.

Erdoğan, G., Stylianou, N., Vasilakis, C. (2019) An open source decision support system for facility location analysis. *Decision Support System* 125, 113116.

Ershadi, M.M., Shemirani, H.S. (2021) Using mathematical modeling for analysis of the impact of client choice on preventive healthcare facility network design. *International Journal of Healthcare Managemen* 14, 588-602.

Filippi, C., Guastaroba, G., Huerta-Muñoz, D.L., Speranza, M.G. (2021) A kernel search heuristic for a fair facility location problem. *Computers & Operations Research* 132, 105292.

Gu, W., Wang, X., McGregor, S.E. (2010) Optimization of preventive health care facility locations. *International Journal of Health Geographics* 9, 17.

Kucukyazici, B., Zhang, Y., Ardestani-Jaafari, A., Song, L. (2020) Incorporating patient preferences in the design and operation of cancer screening facility networks. *European Journal of Operational Research* 287, 616-632.

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., 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., Ríos, M., 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.

Mousazadeh, M., Torabi, S.A., Pishvaee, M.S., Abolhassani, F. (2018) Accessible, stable, and equitable health service network redesign: A robust mixed possibilistic-flexible approach. *Transportation Research Part E: Logistics and Transportation Review* 111, 113-129.

Pourrezaie-Khaligh, P., Bozorgi-Amiri, A., Yousefi-Babadi, A., Moon, I. (2022) Fix-and-optimize approach for a healthcare facility location/network design problem considering equity and accessibility: A case study. *Applied Mathematical Modelling* 102, 243-267.

Risanger, S., Singh, B., Morton, D., Meyers, L.A. (2021) Selecting pharmacies for COVID-19 testing to ensure access. *Health Care Management Science* 24, 330-338.

Tao, Z., Cheng, Y., Dai, T., Rosenberg, M.W. (2014) Spatial optimization of residential care facility locations in Beijing, China: Maximum equity in accessibility. *International Journal of Health Geographics* 13, 33.

Tao, Z., Cheng, Y., Dai, T., Zheng, Q. (2016) Application and validation of gravity p-median model in facility location research. *System Engineering Theory and Practice* 36, 1600-1608.

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

Wang, F., Tang, Q. (2013) Planning toward equal accessibility to services: a quadratic programming approach. *Environment and Planning B: Planning and Desig* 40, 195-

212.

Zhang, W., Cao, K., Liu, S., Huang, B. (2016) A multi-objective optimization approach for health-care facility location-allocation problems in highly developed cities such as Hong Kong. *Computers, Environment and Urban Systems* 59, 220-230.

Zhang, Y., Atkins, D. (2019) Medical facility network design: User-choice and system-optimal models. *European Journal of Operational Research* 273, 305-319.

Zhang, Y., Berman, O., Marcotte, P., Verter, V. (2010) A bilevel model for preventive healthcare facility network design with congestion. *IIE Transactions* 42, 865-880.

Zhang, Y., Berman, O., Verter, V. (2009) Incorporating congestion in preventive healthcare facility network design. *European Journal of Operational Research* 198, 922-935.

List of Tables

Table 1 Network characteristics for the Sioux Falls network

Link	Length (*mile*)	Travel time (*hr*)	Link	Length (*mile*)	Travel time (*hr*)
1,3	3.6	0.12	33,36	3.6	0.12
2,5	2.4	0.08	34,40	2.4	0.08
4,14	3.0	0.10	37,38	1.8	0.06
6,8	2.4	0.08	39,74	2.4	0.08
7,35	2.4	0.08	41,44	3.0	0.10
9,11	1.2	0.04	42,71	2.4	0.08
10,31	3.6	0.12	45,57	2.4	0.08
12,15	2.4	0.08	46,67	2.4	0.08
13,23	3.0	0.10	49,52	1.2	0.04
16,19	1.2	0.04	50,55	1.8	0.06
17,20	1.8	0.06	53,58	1.2	0.04
18,54	1.2	0.04	56,60	2.4	0.08
21,24	6.0	0.20	59,61	2.4	0.08
22,47	3.0	0.10	62,64	3.6	0.12
25,26	1.8	0.06	63,68	3.0	0.10
27,32	3.0	0.10	65,69	1.2	0.04
28,43	3.6	0.12	66,75	1.8	0.06
29,48	3.0	0.10	70,72	2.4	0.08

30,51	4.8	0.16	73,76	1.2	0.04

Table 2 Demand and facility data for the Sioux Falls network

Population node i	Demand h_i (clients/hr)	Facility location j	Capable space K_j (clients)
1	41	3	50
2	33	7	70
4	23	9	60
5	29	11	60
13	41	16	80
14	35	19	70
15	43	21	50
20	26	23	80

Table 3 The network design scheme and their service level

Facility location	Server number	Balking probability	Effective arrival rate (clients/hr)	Waiting time (hr)
3	6	0.132	36	1.21
9	9	0.096	54	1.12
16	8	0.118	48	1.10
19	7	0.091	42	1.20
23	10	0.128	60	1.22

Table 4 The demand allocation at equilibrium state

Population node	Selected facility location				
	3	9	16	19	23
1	**31.10**	3.77	1.94	1.34	2.55
2	2.07	5.52	**21.77**	2.56	1.08
4	2.48	**16.96**	1.45	1.10	1.10
5	1.79	**21.43**	2.22	1.79	1.37
13	3.16	1.94	1.34	2.55	**31.71**
14	0.63	3.26	1.68	5.36	**24.27**
15	0.13	5.89	**11.01**	**21.90**	3.97
20	0.08	0.87	**13.08**	**9.54**	2.84

List of Figures

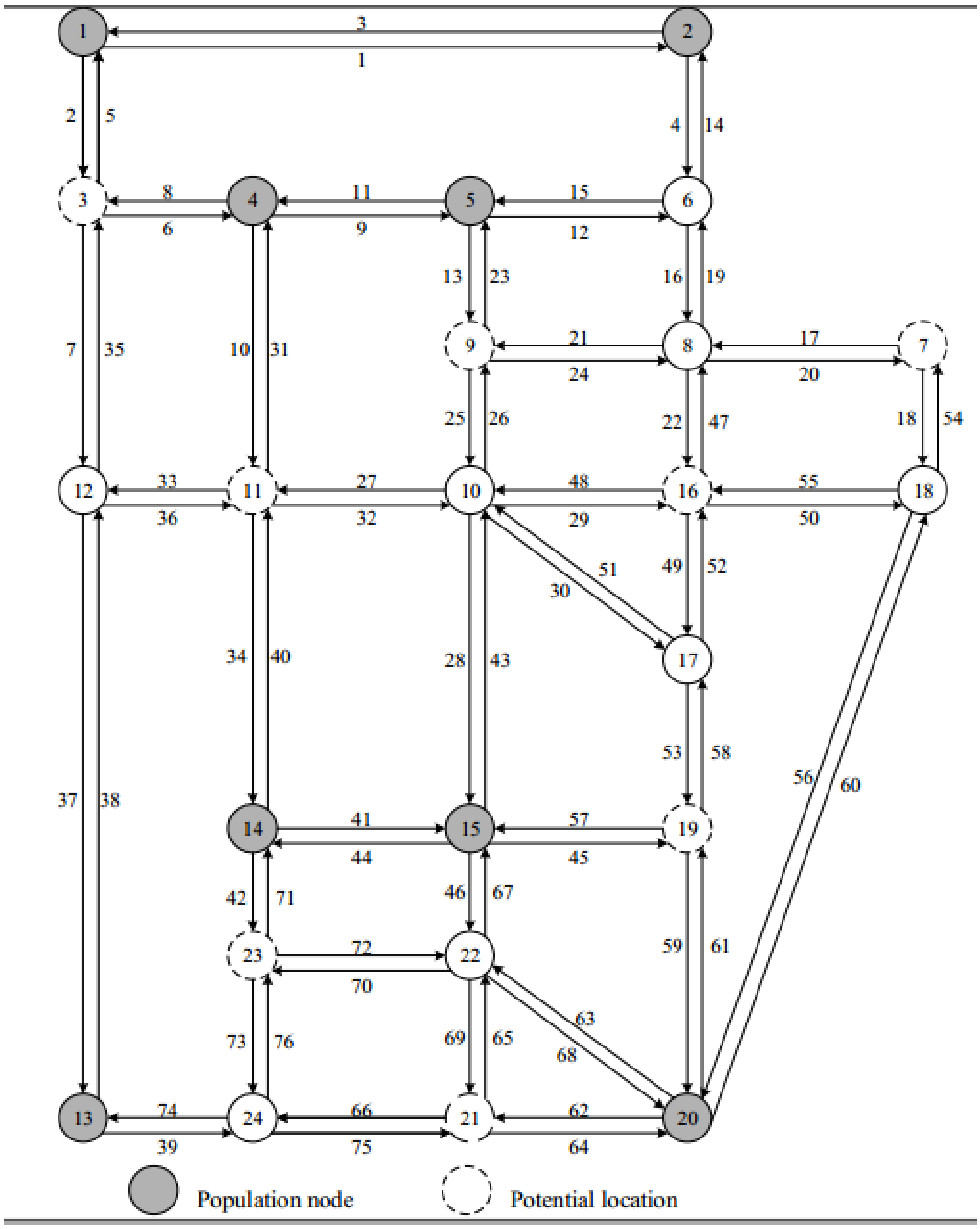

Fig. 1. The Sioux Falls test network

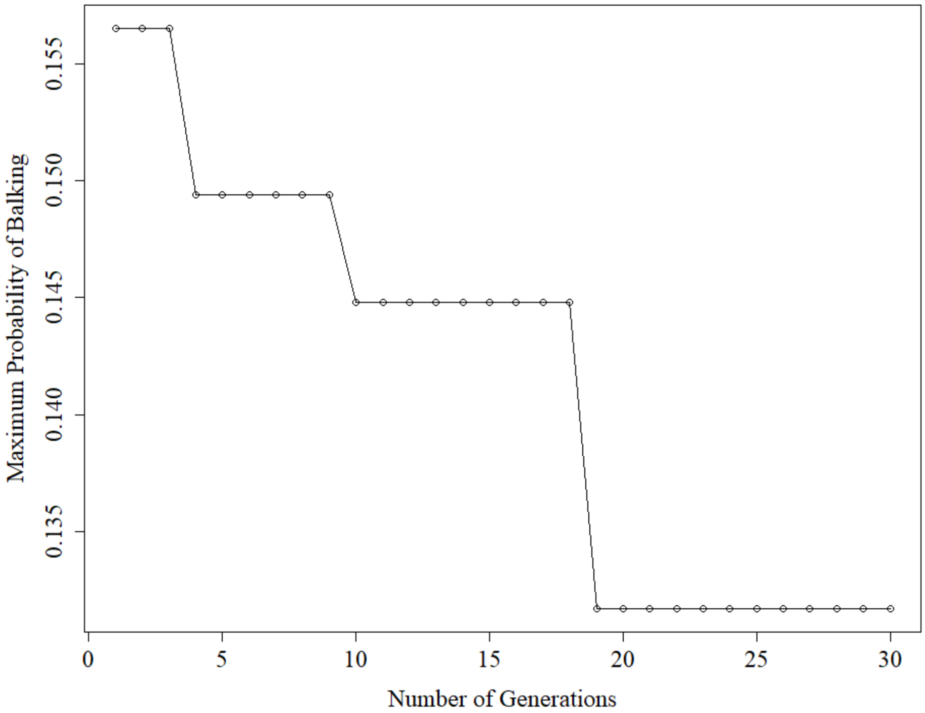

Fig. 2. The evolutionary process of genetic algorithm

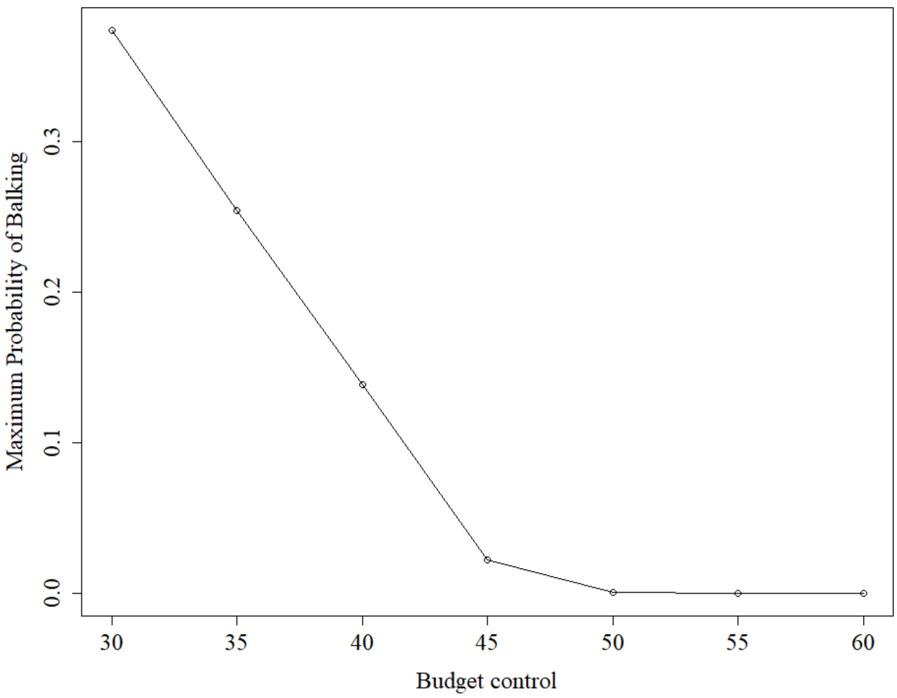

Fig. 3. A sensitivity analysis with varying budget

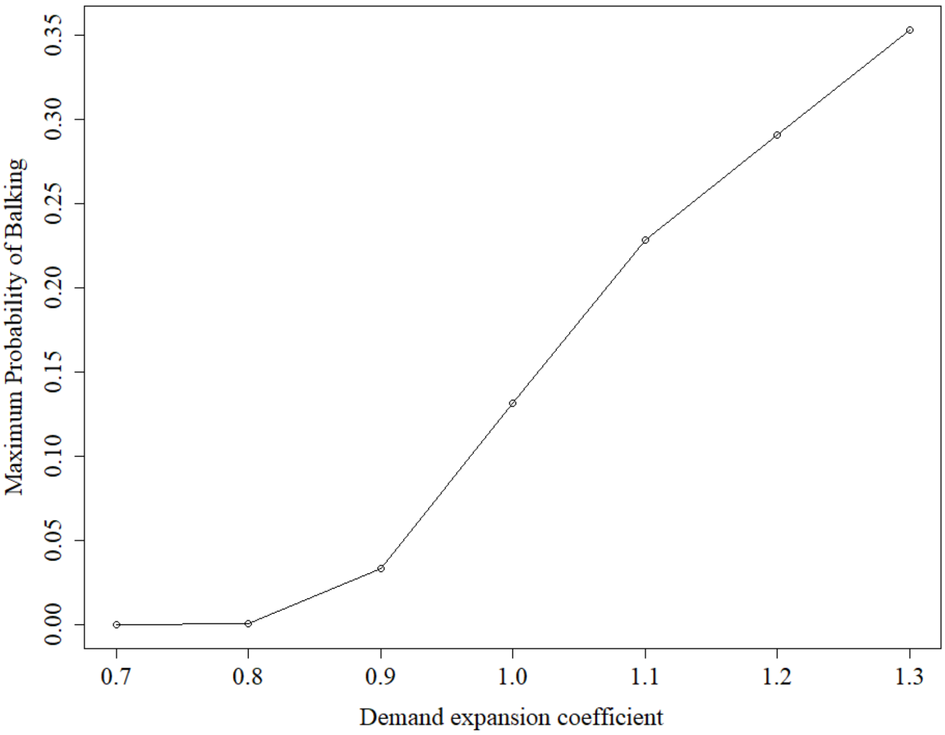

Fig. 4. A sensitivity analysis with varying demand