ATLANTIS PRESS

Advances in Intelligent Systems Research, volume 133

2nd International Conference on Artificial Intelligence and Industrial Engineering (AIIE2016)

A multi-service Healthcare Network Design with Patients' Choice to Exchange between Services

Maryam Radman[1,*] and Kourosh Eshghi[2]

[1]PhD student of Department of Industrial Engineering, Sharif University of Technology, Iran

[2]Professor of Department of Industrial Engineering, Sharif University of Technology, Iran

[*]Corresponding author

***Abstract*—Efficient location of medical services is an issue of paramount importance in healthcare strategic planning. In this research, a mathematical model is developed for the location of multi-service health centers assuming probabilistic demand and service time. Since patients may be shifted to another service after receiving a service by doctors' order, health system is considered a Jackson queue network. The primary factor contributing to patients' choice of one center over another is their proximity to the center. The proposed model seeks to minimize the demand weighted total distance traveled by patients between their residential areas and health centers and also between health centers on the one hand, and the weighted sum of undesired deviations from standard arrival rates at service stations on the other hand. The location of health centers as well as the type of services they offer and the number of servers at each service station are the main determinants of the proposed model. A GA-based heuristic is developed to solve medium and large instances of the proposed model. For evaluation of the suggested heuristic, computational experiments are performed on a number of test problems.**

Keywords-healthcare system; multi-service health centers; location problem; queuing theory; utility theory

I. INTRODUCTION

The design of medical infrastructures directly bears on the community's health and medical costs. In every society, many basic and specialized treatment centers provide services. Among them, polyclinics increase patients' accessibility and speed up service delivery by providing a set of related medical services in a centralized location. These places offer a wide range of services such as a variety of medical and dental specialties, outpatient surgeries, emergency and preventive services as well as paraclinical services.

Because of the multifaceted nature of most diseases, patients should be simultaneously observed by different specialists. As a case in point, a diabetic needs the attention and care of an endocrinologist, a nutritionist and an ophthalmologist all at once. It inevitably follows that by concentrating different related services in one place, not only can we obviate the need for patients' transfer to different parts of a city, but we can also create an opportunity for doctors with diverse specialties to work as a team.

In this research, a model is proposed to determine (1) the location of multi-service health centers, (2) the type of services provided in every center and (3) the number of servers at each service station.

This research introduces a probabilistic model for the location of multi-service health centers assuming the transfer of patients between services and the creation of queue networks in the health system for the first time. In addition, we classify medical services into two groups (normal and emergency) and assume different choice behaviors by patients for these services.

The remainder of the article is organized as follows: the next section reviews the related literature on healthcare network design, customers' choice behavior and congestion modeling in service centers. Section 3 introduces the proposed model. Section 4 explains the proposed solution method for solving the model. In section 5, the performance of the proposed heuristic is tested by several numerical examples. Lastly, the main points of the article are summed up and suggestions for future research are offered.

II. RELATED LITERATURE

Healthcare location models can be classified into three broad areas, namely accessibility, adaptability and availability models. Accessibility models try to provide affordable and right healthcare resources in the right place at the right time. Adaptability models consider different future conditions and attempt to find solutions that perform well across a range of future scenarios. And availability models, which are more applicable for ambulance location problems, attempt to compensate for the short-term unavailability of vehicles [1]. The proposed model straddles both the accessibility and availability models as it tries to offer services in right areas in an attempt to reduce the time traveled by patients and to allocate the right number of servers to service stations in order to reduce the waiting time at the stations.

Before taking any decisions, we should recognize effective factors in the location of health centers. Patients' demand, the current state of health centers, the type and trend of diseases, birth and death rates, geographic setting (urban or rural), transportation infrastructure, patients' income, budget constraint, patients' functional and cultural needs and affiliations to other medical centers like hospitals often feature prominently when it comes to making decisions about health centers [2]. In our model, we have accounted for the impact of patients' demand on each type of service and the number of servers offering each type of service.

The most important goals of healthcare location problems are (1) maximizing patients' coverage, (2) minimizing costs including fixed and variable costs (3), minimizing the sum of weighted distances traveled by patients, and (4) minimizing the sum of travel time, waiting time in the queues and service time [3].

In addition, multi-objective models, which combine two or more of the mentioned goals are used in healthcare location problems. Our proposed model is not multi-objective, but simultaneously seeks to minimize the demand weighted total distance traveled by patients between their residential areas and health centers as well as between health centers and sum of undesired deviations from the standard arrival rates at the medical stations.

The most widely used constraints in these models are (1) system cost constraint, (2) the allowed number of health centers to establish, (3) minimum workload required to establish a center, (4) capacity constraint, (5) critical service distance, and (6) congestion constraints [3]. The most important constraints in our proposed model are congestion constraints, constraints relating to patients' choice behavior including the calculation of patients' utilities with logit utility function, allocating patients to the nearest health centers and transfer of patients between medical services, capacity constraints and minimum workload required to offer services.

In addition, there are four types of uncertainties in these problems, namely (1) uncertainty in demand due to changing birth and death rates, migration and seasonal conditions, (2) uncertainty in travel time due to changing traffic load and unpredictable conditions of transportation routes, (3) uncertainty in service time because of different patients' conditions, and (4) uncertainty in capacity. Multi-period models, queuing theory, stochastic and dynamic programming are the most prevalent methods used in the literature to address the mentioned uncertainties [3]. Using probability distributions and queuing theory, we have addressed uncertainty in demand and service time.

We usually face queues in health centers. The high congestion of these centers adversely affects clients' satisfaction, and in emergency cases, it may lead to irreparable damages. To guarantee service quality, we usually deal with one of these factors: (1) average queue length, (2) average waiting time in queues, or (3) the probability of receiving service in a standard time [4], [5]. These factors can be a part of constrains that are called "constraint-oriented" approaches [6], [7]. If the factors form a part of the objective function, the resulting models are called "objective-oriented" models [8]. In our proposed model, standard values are determined by experts for average waiting time at medical stations and the objective function seeks to minimize the sum of undesired deviations from these values.

In most of the reviewed articles, patients are assigned to health centers by a model called "system choice models". In contrast, in "user choice models", each individual has the right to choose a center rather than being assigned to one by the model. In the user-choice environment, patients' utility to choose health centers is calculated by a probability distribution based on centers' attractiveness. These models are categorized into "optimal choice" and "probabilistic choice" models. In optimal choice models, patients choose the center with the maximum utility. For instance, patients choose to go to the nearest health center [9]. In probabilistic choice models, patients choose every center according to the probability calculated for it [5], [10]. Huff used the probabilistic choice model to calculate the choice probability of shopping centers in 1963 for the first time [11]. Our model is based on user choice behavior and is a combination of optimal and probabilistic choice models. Referral of patients to the nearest health centers and choosing health centers using the multinomial Logit function shows optimal and probabilistic choice behavior respectively. This will be explained in much more detail in the next section.

In general, we can model the probability of choosing the center in area j by people in area i using the equation $p_{ij} = \frac{U_{ij}}{\sum_{k \in X} U_{ik}}$. U_{ij} is the utility gained by people in area i to get the service of the center in area j and X is the set of alternatives to establish centers. We can use different ways to model the utility function (U_{ij}). One of the most famous models is the multinomial Logit function proposed by Feddan in 1974. According to this model, U_{ij} is calculated using the equation $U_{ij} = e^{\sum_{l \in L} \beta_l y_{ijl}}$. In this formula, y_{ijl} is related to the lth attractiveness factor of the center in area j, L is the set of attractiveness factors and β_l is a parameter that shows the effect of its corresponding factor on utility [12]. We use the multinomial Logit function in our model to calculate utility. Travel time is the only attractiveness factor.

Marianov, Rios and Icaza proposed a model to locate multi-server facilities to maximize market capture in 2008. Customers choose the facility to patronize in view of the time needed to travel to the facility and the waiting time at the facility. The Logit function is utilized to model the user-choice environment [13]. Zhang, Berman and Verter in 2012 developed a model to maximize total participation in a preventive care program. The only attractiveness factor used in the Logit function they proposed was the proximity of the service center to its clients [14].

III. Problem Description and Modeling

Before describing the problem, a mention of the two properties of Poisson process, which are used later, should be in order:

1. If $N_1(t)$ and $N_2(t)$ follow Poisson processes with parameters λ_1 and λ_2, then $N_1(t) + N_2(t)$ also follows a Poisson process with parameter $\lambda_1 + \lambda_2$.
2. If customers' arrival follows a Poisson process $N(t)$ with parameter λ and customers belong to two groups with the probabilities p and $1-p$, then the arrivals of group one and two also follow Poisson processes with parameters λp and $\lambda(1-p)$.

A brief explanation of the terms "service", "server", "station" and "center", which are frequently used in this study, should also be in order. By "service", we mean medical services offered by health centers. They fall into to three categories in the proposed model, namely normal, preventive

and emergency. Dental services, mammograms and fire and rescue services are examples of these services respectively. "Server" is the person who provides medical services. Examples of servers can be doctors, nurses and dentists. "Station" is a place where a special service is offered, which has one or more servers on tap. Finally, by "center" or "health center", we mean a place where a combination of different services is provided. So a center can include service stations.

Let $G = (N, E)$ be a network with a set of nodes (N) and a set of links (E). The nodes represent population zones and the links are the main transportation arteries showing travel time between the nodes. The number of patients residing in area i who require service k is denoted by h_{ik} and is Poisson distributed by a rete of λ_{ik} per unit of time (h_{ik} denotes the population of area i, each person generates a demand for service k according to a Poisson process with rate λ_{ik} per unit of time). Therefore, the number of demands in area i for service type k is $h_{ik}\lambda_{ik}$ per unit of time.

Travel time from area i to area j through the shortest path is denoted by t_{ij}. Service time or the time required by a server to provide service k is exponentially distributed at a rate of μ_k per unit of time. We assume that every patient goes to a center for a special service but after receiving that service, he may be referred to another service. So a patient may use several types of services to complete his treatment process. To be more accurate, we also divide services to two types, namely normal and emergency. Emergency cases need immediate visits such as food poisoning, insect stings and other unpleasant events. Other types of services such as dental services, simple and specialized medical services, drug stores fit into the normal group [3]. Other assumptions of the proposed model are as follows (however, it should be noted that these assumptions depend entirely on the conditions of the issue under discussion and can change if necessary):

1. Patients' choice behavior follows this pattern: for emergency services, because of the immediate need for treatment, patients go to the nearest health center which provides the service. For normal services, patients have "probabilistic choice behavior", meaning they choose a center according to the Logit utility function.
2. The attractiveness factor considered by patients to choose a center is travel time.
3. A patient may be referred to another service after receiving a service. In this situation, the person goes to the nearest area that provides the service (for convenience).

Before discussing the model equations, its notations are presented in the following table:

Notation	Description
Model sets	
N	Set of demand and healthcare areas
K	Set of services
K_n	Set of normal services
K_e	Set of emergency services
Model indexes	
i	Index of demand areas
j, l, l'	Index of health center areas
k, k'	Index of services
w	Index of number of servers
Model parameters	
t_{ij}	Travel time between demand area i and health center in area j
t'_{lj}	Travel time between health center in area l and health center in area j
h_{ik}	Number of demands in area i for service k
λ_{ik}	Demand rate of each individual in area i for service k
M_k	Maximum number of servers that can be allocated to service station k
Z_k	Number of available servers for service k that can be assigned to health centers
μ_k	Service rate of service k
R_k	Minimum arrival rate required to provide service k
$w_{std,k}$	Standard values determined by experts for average waiting time at service station k
$\bar{\lambda}_w^k$	Maximum arrival rate of patients at service station k with w servers so that average waiting time in the system equals $w_{std,k}$
$pr_{kk'}$	Probability of referring patients from service k to service k'
π_i	Importance coefficient of section i of the objective function
Model Auxiliary Variables	
p_{ikj}	Probability of choosing service k in area j by patients in area i
U_{ik}	Sum of patients' utility of all centers for service k
Λ_{kj}	Patients' arrival rate at the station type k in area j
w_{kj}	Average waiting time at the station type k in area j
ρ_{ij}^S	Weighted sum of travel time between area i and area j and number of services provided in a center of type S in area j
Model Decision Variables	
x_{ikj}	1 if patients in area i go to the center in area j to receive emergency service type 0 otherwise
q_{lkj}	1 if patients go to the center in area j to receive service type k after receiving services in a center in area l 0 otherwise
s_{kj}^w	1 if at least w servers are assigned to the service station k in area j 0 otherwise

According to the explanations provided so far, we can model the health system as a queue network. Every patient goes to the network to receive a special service, but after receiving that service, he or she may be shifted to another service with a specific probability. This network consists of some alternative systems; each provides a special service with some servers. Each alternative system is called a service station. In every service station, one or more servers work in parallel. In such a network, we cannot consider each station as an independent queuing system, because patients entering a station may be the outputs of other stations [16].

According to the assumptions made above, the health system can be modeled as a Jackson queue network. Our model satisfies four main assumptions of such networks:

1. This assumption states that patients' arrival at each service station from demand areas should follow a Poisson process. According to the Poisson process properties described at the beginning of this section, we can show that patients'

arrival at service station k in area j follows a Poisson process with parameter γ_{kj} and is calculated by the following equations:

$$\gamma_{kj} = \sum_{i\in N} h_{ik}\lambda_{ik}x_{ikj} \qquad \forall_{k\in K_e, j\in N} (1)$$

$$\gamma_{kj} = \sum_{i\in N} h_{ik}\lambda_{ik}p_{ikj} \qquad \forall_{k\in K_n, j\in N} \quad (2)$$

x_{ikj} is a binary variable that takes the value 1 if patients in area i go to the center in area j to receive service type k ($k \in K_e$). p_{ikj} shows the probability of choosing service k in in area j by patients in area i.

The above equations calculate patients' arrival rate from demand areas for emergency and normal services respectively. We showed earlier that patients' demand for every service in each area follows a Poisson process, so according to the first property of this process, the sum of these demands also follows a Poisson process. Consequently, equation (1) follow Poisson process. In addition, because patients' demand for every service in each area follows a Poisson process, p_{ikj} % of them also follows a Poisson process according to the second property of this process. Consequently, equation (2) also follows Poisson process.

2. This assumption states that service time at each station should be exponentially distributed and independent of other stations. In our model, we have assumed that service time at the service station k follows exponential distribution at a rate of μ_k and is independent of other stations.

3. This assumption states that queue capacity should be unlimited in all service stations. This assumption has been accommodated in our model too.

4. According to this assumption, everyone is referred to another station or out of the system after receiving a service with a certain probability. In this model, a patient may be referred to service k' with a probability of $\mathrm{pr}_{kk'}$ After receiving service k, or may leave the system with a probability of pr_{k0}. As we have assumed, these patients go to the nearest health center that provides the service they need.

Let binary variables q_{lkj} be 1 if patients in area l go to the area j to receive service k. Consequently, patients' arrival rate from other stations of the health network to the station k in area j is calculated as follows:

$$\sum_{l\in N}\sum_{k'\in K} \Lambda_{k'l}\,\mathrm{pr}_{k'k}q_{lkj} \qquad \forall_{k\in K, j\in N} \qquad (3)$$

In the above equation, $\Lambda_{k'l}$ is patients' arrival rate at service station k' in area l.

According to the above explanations, in general, patients' arrival rate at service station k in in area j is composed of two parts and is calculated as follows:

$$\Lambda_{kj} = \gamma_{kj} + \sum_{l\in N}\sum_{k'\in K} \Lambda_{k'l}\,\mathrm{pr}_{k'k}q_{lkj} \qquad \forall_{k\in K, j\in N} \qquad (4)$$

The first part is patients' arrival rate from demand areas (γ_{kj}) and the second part is patients' arrival rate from other service stations ($\sum_{l\in N}\sum_{k'\in K}\Lambda_{k'l}\,\mathrm{pr}_{k'k}q_{lkj}$).

If there is no feedback in the health network, it means the patient has not returned to a previously visited service station directly or indirectly, the number of patients entering a station follows a Poisson process and every station is an $M/M/c$ queuing system. If there is feedback, patients' arrival does not necessarily follow a Poisson distribution, but to calculate the average waiting time in the queue/system or the average number of patients in the queue/system, we can still use the equations proved for $M/M/c$ queuing systems [16].

We use the following equation proposed by Kleinrock in 1975 to calculate the average waiting time in a $M/M/c'$ queuing system [14]. $\bar{W}$ shows the average waiting time in a station with c' servers.

$$\bar{w} = \frac{C(c',\mu)}{c'}\frac{1}{\mu(1-\rho)} + \frac{1}{\mu} \qquad u = \frac{\lambda}{\mu}, \rho = \frac{\lambda}{c'\mu}$$
$$C(c',u) = \frac{1-K(u)}{1-\rho K(u)}, \qquad K(u) = \frac{\sum_{l=0}^{c'-1}\frac{u^l}{l!}}{\sum_{l=0}^{c'}\frac{u^l}{l!}} \qquad (5)$$

To guarantee the quality of services provided, constraint (6) is used. w_{kj} is the average waiting time at the service station k in area j and is calculated using equation (5). $w_{std,k}$ is the standard average waiting time at the service station k proposed by experts. So this constraint states that the average waiting time in the service station k should be less than the standard value determined by experts for it.

$$w_{kj} \le w_{std,k} \qquad \forall_{k\in K, j\in N} (6)$$

[Note that if the average waiting time (w_{kj}) is set equal to the standard value determined by experts $(w_{std,k})$, the maximum arrival rate, when there are c' servers in the service station k, can be obtained. This value is denoted by $\bar{\lambda}^k_{c'}$ in the following]

To satisfy the equation (6), it is sufficient that arrival rate at service station k in area j (Λ_{kj}) be less than the maximum arrival rate ($\bar{\lambda}^k_{c'}$). That is $\Lambda_{kj} \le \bar{\lambda}^k_{c'}$.

Because in the proposed model, the number of servers at every service station is a decision variable, the mentioned equation ($\Lambda_{kj} \le \bar{\lambda}^k_{c'}$) for is modeled as follows:

$$\Lambda_{kj} \le \sum_{w=1}^{M_k} (\bar{\lambda}^k_w - \bar{\lambda}^k_{w-1}) s^w_{kj} \qquad \forall_{K\in K, j\in N} (7)$$

In the above equation, s_{kj}^{w} is a binary variable that takes the value 1 if at least w servers are assigned to the service station k in area j.

Note that $\bar{\lambda}_0^k$ is zero. To clarify the above equation, assume that it is decided by the model to offer service type k in area j with two servers. So s_{kj}^1 and s_{kj}^2 become 1, s_{kj}^w becomes 0 for $w \geq 3$ and the equation becomes $\Lambda_{kj} \leq \left(\bar{\lambda}_1^k - \bar{\lambda}_0^k\right)(1) + \left(\bar{\lambda}_2^k - \bar{\lambda}_1^k\right)(1) = \bar{\lambda}_2^k$. This is the same as the equation $\Lambda_{kj} \leq \bar{\lambda}_{c'}^k$ as stated earlier. Therefore, according to equation (7), the number of appropriate servers to satisfy standard waiting time in the system is determined.

The number of available servers to allocate to the stations of each type of service is limited and is Z_k .So the limitation in the number of available servers may not allow the model to allocate an enough number of servers to all stations of a special type of service in order to satisfy the average standard waiting time and, consequently, constraint (7) may become infeasible. Therefore, in this model, we change hard constraint (7) to a soft one and then try to minimize the undesired deviations (dev_{kj}^-) in the objective function. The following equation is obtained:

$$\sum_{w=1}^{M_k} \left(\bar{\lambda}_w^k - \bar{\lambda}_{w-1}^k\right) s_{kj}^w - \Lambda_{kj} = dev_{kj}^+ - dev_{kj}^- \quad \forall_{k\epsilon K, j\epsilon N} \quad (8)$$

To model patients' choice behavior for normal services, we have used the probabilistic choice model and Logit utility function. The equations are as follow:

$$\mathrm{p}_{ikj} = \frac{U_{ikj}}{U_{ik}}, U_{ikj} = e^{-t_{ij}} s_{kj}^1 \quad \forall_{i,j\epsilon N, k\epsilon K_n} \quad (9)$$

$$U_{ik} = \sum_{l\epsilon N} U_{ikl} = \sum_{l\epsilon N} e^{-t_{il}} s_{kl}^1 \quad \forall_{i\epsilon N, k\epsilon K_n} \quad (10)$$

p_{ikj} shows the probability of choosing service k in area j by patients in area i. U_{ik} is the utility gained by patients in area i if they choose a center in area j for service k. U_{ik} shows the sum of patients' utilities in area i for service k . p_{ikj} is calculated by dividing U_{ikj} by U_{ik}.
According to the Logit utility function, the significant factor in patients' choice behavior shows itself in the power of the statement $e^{-t_{ij}}$. This factor is the travel time between area i and area j, indicating accessibility.

Against this backdrop of the above explanations, the whole proposed model is as follows:

$$\min Z = \pi_1 \sum_{i,j\epsilon N, k\epsilon k_e} h_{ik}\, \lambda_{ik} x_{ikj} t_{ij} + \pi_2 \sum_{l,j\epsilon N, k\epsilon k_n, k'\epsilon K} \Lambda_{k'l}\, pr_{k'k} q_{lkj} t'_{lj} + \pi_3 \sum_{l,j\epsilon N, k\epsilon k_e, k'\epsilon K} \Lambda_{k'l}\, pr_{k'k} q_{lkj} t'_{lj} + \pi_4 \sum_{j\epsilon N, k\epsilon K} w_{std,k} dev_{kj}^- \quad (11)$$

$$s.t$$

$$\sum_{j\epsilon N} \sum_{w} s_{kj}^w \leq \mathrm{Z}_k \quad \forall_{k\epsilon K} \quad (12)$$

$$\sum_{w} s_{kj}^w \leq \mathrm{M}_k \quad \forall_{k\epsilon K, j\epsilon N} \quad (13)$$

$$\Lambda_{kj} \geq R_k s_{kj}^1 \quad \forall_{k\epsilon K, j\epsilon N} \quad (14)$$

$$x_{ikj} \leq s_{kj}^1 \quad \forall_{k\epsilon K_e, i,j\epsilon N} \quad (15)$$

$$t_{ij} x_{ikj} \leq t_{il} + M(1 - s_{kl}^1) \quad \forall_{i,j,l\epsilon N, k\epsilon K_e} \quad (16)$$

$$\sum_{j\epsilon N} x_{ikj} = 1 \quad \forall_{i\epsilon N, k\epsilon K_n, K_p} \quad (17)$$

$$t'_{lj} q_{lkj} \leq t'_{ll'} + M\left(1 - s_{kl'}^1\right) \quad \forall_{k\epsilon K, l,l', j\epsilon N} \quad (18)$$

$$\sum_{j\epsilon N} q_{lkj} = 1 \quad \forall_{k\epsilon K, l\epsilon N} \quad (19)$$

$$q_{lkj} \leq s_{kj}^1 \quad \forall_{k\epsilon K, l,j\epsilon N} \quad (20)$$

$$\mathrm{p}_{ikj} = \frac{U_{ikj}}{U_{ik}} = \frac{e^{-t_{ij}} s_{kj}^1}{\sum_{l\epsilon N} e^{-t_{il}} s_{kl}^1} \quad \forall_{i,j\epsilon N, k\epsilon K_n} \quad (21)$$

$$\Lambda_{kj} = \gamma_{kj} + \sum_{l\epsilon N} \sum_{k\epsilon K} \Lambda_{k'l}\, \mathrm{pr}_{k'k} q_{lkj} \quad \forall_{k\epsilon K, j\epsilon N} \quad (22)$$

$$\gamma_{kj} = \sum_{i\epsilon N} h_{ik} \lambda_{ik} p_{ikj} \quad \forall_{k\epsilon K_n, j\epsilon N} \quad (23)$$

$$\gamma_{kj} = \sum_{i\epsilon N} h_{ik} \lambda_{ik} x_{ikj} \quad \forall_{k\epsilon K_e, j\epsilon N} \quad (24)$$

$$\sum_{w=1}^{M_k} \left(\bar{\lambda}_w^k - \bar{\lambda}_{w-1}^k\right) s_{kj}^w - \Lambda_{kj} = dev_{kj}^+ - dev_{kj}^- \quad \forall_{k\epsilon K, j\epsilon N} \quad (25)$$

$$s_{kj}^{M_k} \leq \cdots \leq s_{kj}^2 \leq s_{kj}^1 \quad \forall_{k\epsilon K, j\epsilon N} \quad (26)$$

$$dev_{kj}^+ * dev_{kj}^- = 0 \quad \forall_{k\epsilon K, j\epsilon N} \quad (27)$$

$$x_{ikj}, q_{lkj}, s_{kj}^w \in \{0,1\} \quad (28)$$

$$\mathrm{p}_{ikj}, \mathrm{U}_{ik}, \Lambda_{kj}, dev_{kj}^+, dev_{kj}^- \geq 0 \quad (29)$$

The objective function minimizes the demand weighted total distance traveled by patients between their residential areas and health centers for emergency services (the first section) and also between health centers for normal and emergency services (the second and third sections) and the weighted sum of undesired deviations from standard arrival rates at the medical stations (the forth section).

π_i is the importance coefficient of section i of the objective function. $w_{std,k}$ is used to make the forth section of the objective function similar to the first three sections in terms of unit. Note that if dev_{kj}^- takes a non-zero value for a station, then patients' arrival rate (Λ_{kj}) is more than the standard arrival rate of patients ($\bar{\lambda}_w^k$) at that station or, similarly, average waiting time in that service station (w_{kj}) is more than the standard value ($w_{std,k}$) determined for it, therefore, patients should wait at the service station for at least $w_{std,k}$ units of time. That is why this coefficient is used for the forth section of the objective function to make its unit similar to the other sections.

Equation (12) satisfy capacity constraint. Constraints (13) limit the number of allocated servers to service stations. Constraint (14) satisfies minimum patients' arrival rate required to provide a service. Constraint (15) states that if we provide service k in area j, patients can choose to go there.

Constraint (16) states that patients go to the nearest center for emergency services. Note that if variable x_{ikj} takes the value 1, then patients in area i go to the center of in area j to receive service type k. Now if this type of service is offered in area l, then s_{kl}^1 takes the value 1 and the whole equation becomes $t_{ij}(1) \leq t_{il}$ that guarantees t_{ij} is less than all t_{il} $\forall$ l or, similarly, the travel time between area i and area j to receive emergency service type k is less than the travel time between area i and area l where service type k is provided. Consequently, area j is the nearest.

Constraint (17) ensures that patients go to one and only one center for emergency services. Constraint (18) states that after receiving a service, if required, patients go to the nearest center for the next service. Constraints (19) and (20) ensure that the mentioned referral in (18) happens to one and only one center and if the service is provided in that center respectively. Equation (21) calculates ρ_{ij}, which was explained earlier. Constraints (22) to (24) calculate patients' arrival rate to health centers. Description on constraint (25) which calculates the positive and negative deviations of patients' arrival rates from the standard rates was explained earlier. Constraint (26) ensures that w servers are already allocated before allocating the $(w+1)$ *th* server to the service station. Constraint (27) explains that the product of positive and negative deviations must equal zero so that both do not take values simultaneously. Note that because the proposed model is nonlinear, this constraint is not redundant and cannot be omitted.

IV. Solution Methods

The proposed model is a Mixed Integer Non-Linear Programming problem (MINLP). The presence of the term $\Lambda_{k'l}q_{lkj}$ in the second and third part of the objective function and constraint (22), the presence of the term $\mathrm{p}_{ikj}s_{kl}^1$ in constraint (21) and the product of positive and negative deviations in constraint (27) causes the nonlinearity of the model. To linearize the terms $\Lambda_{k'l}q_{lkj}$ and $\mathrm{p}_{ikj}s_{kl}^1$, we use the method in [14]. After linearizing these terms, constraint (27) becomes redundant and can be avoided. Because it is never optimal for both positive and negative deviations to simultaneously assume non-zero values.

Since q_{lkj} is a binary variable and $\Lambda_{k'l}$ is a continuous variable, the term $\Lambda_{k'l}q_{lkj}$ can be linearized as follows by defining $c_{k'lkj} = \Lambda_{k'l}q_{lkj}$ as an artificial continuous variable:

$$c_{k'lkj} \leq \Lambda_{k'l} \qquad \forall_{k,k'\in K,l,j\in N} \ (30)$$
$$c_{k'lkj} \leq M_1 q_{lkj} \qquad \forall_{k,k'\in K,l,j\in N} \ (31)$$
$$c_{k'lkj} \geq \Lambda_{k'l} - M_2(1-q_{lkj}) \qquad \forall_{k,k'\in K,l,j\in N} \ (32)$$
$$c_{k'lkj} \geq 0 \qquad \forall_{k,k'\in K,l,j\in N} \ (33)$$

Where M_1 and M_2 denote two big numbers. Similarly for constraint (21) we have:

$$\mathrm{p}_{ikj} = \frac{e^{-t_{ij}}s_{kj}^1}{\sum_{l\in N} e^{-t_{il}}s_{kl}^1} \Rightarrow e^{-t_{ij}}s_{kj}^1 = \sum_{l\in N} e^{-t_{il}}s_{kl}^1 \mathrm{p}_{ikj} \qquad \forall_{k\in K_n,i,j\in N} \ (34)$$

$$d_{ikjl} \leq \mathrm{p}_{ikj} \qquad \forall_{k\in K,i,l,j\in N} \ (35)$$
$$d_{ikjl} \leq M_3 s_{kl}^1 \qquad \forall_{k\in K,i,l,j\in N} \ (36)$$
$$d_{ikjl} \geq \mathrm{p}_{ikj} - M_4(1-s_{kl}^1) \qquad \forall_{k\in K,i,l,j\in N} \ (37)$$
$$d_{ikjl} \geq 0 \qquad \forall_{k\in K,i,l,j\in N} \ (38)$$

In the above equations, d_{ikjl} is defined as the product of p_{ikj} and s_{kl}^1 $(d_{ikjl} = \mathrm{p}_{ikj}s_{kl}^1)$.

Consequently by adding constraints (30) to (33) and (35) to (38) and eliminating constraint (27), the proposed model becomes linear.

The model is formulated as an MIP, which can be solved directly by standard MIP solvers, such as CPLEX. However, the proposed model is NP hard, as it contains a p-median problem as a particular case (to see this, assume that, (1) there is just one emergency service (2) the minimum workload required is set at zero (3) the maximum number of servers in a center is set at one (4) w_{std} and μ are considered large enough to make patients' arrival rate at every station less than $\bar{\lambda}_1$). Therefore, heuristic or meta-heuristic methods should be developed to solve the large instances of the model.

The proposed heuristic is based on the breakdown of the model into three subsections, namely (1) determination of the location of service stations, (2) calculation of patients' arrival rates at service stations and (3) allocation of servers to service stations.

A. *Location of Service Stations*

In this subsection, we use Genetic algorithm to determine the location of service stations. The solution structure is shown in Figure I. The rows show the areas and the columns show the services.

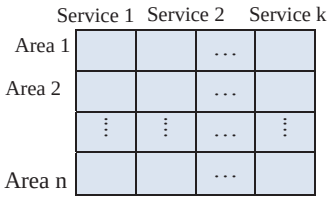

FIGURE I. LUTION STRUCTURE (CHROMOSOME) USED IN HEURISTIC

Step 1 (producing first generation): generate N_{pop} feasible chromosomes randomly. In addition to satisfying the limitation on the number of the servers of each service type, a feasible chromosome must have at least one service station for each normal and emergency service.

Step 2 (calculation of fitness function): after determining the locations of service stations (generating feasible chromosomes), the objective function of the produced solutions is calculated by applying the second (calculation of patients' arrival rates at service stations) and the third (allocation of servers to service stations) subsections. The fitness of each chromosome is inversely proportional to the objective function value.

Step 3 (generating the new generation):

3-1 (parent selection): the roulette wheel selection method is applied to randomly choose two parents from the population.

3-2 (crossover): the one-point crossover operator is applied to all columns of the chromosome.

3-3 (mutation): for each column of the chromosome, a place is randomly chosen. If it is zero, it is changed to one, otherwise, it is changed to zero.

Step 4 (replacement): the first N_{pop} chromosomes, which have better objective function among parents and children, are chosen.

Step 5 (stopping criterion): strong convergence is used as the stopping criterion. According to this criterion, we must calculate the variance of the objective function values of chromosomes in each iteration; the algorithm stops if the difference of the maximum and minimum variance values of the last m iterations is less than ε. m and ε are arbitrary values.

B. Calculation of Patients' Arrival Rate at Service Stations

Step 1: in the first subsection, service stations for establishment are determined. In this step, according to patients' choice behavior explained at the beginning of section 3, the direct arrival rates of patients from demand areas to the service stations are calculated.

Step 2: in the above step, we only calculated the direct arrival rates(γ_{kj}). To calculate the real arrival rates of patients at the service stations (Λ_{kj}) that consider patients' flow between service stations, we form a "system of equations" of variables Λ_{kj}. The number of variables and equations of this system is $(number\ of\ services \times number\ of\ areas)$. You can see the "system of equations" in constraint (22).

Step 3: in this step, the "system of equations" of the previous step is solved by the Gaussian elimination method and Λ_{kj} is calculated**.**

Step 4: in this step, the feasibility of the solution is checked. The calculated arrival rates (Λ_{kj}) for established service stations must be more than minimum arrival rates required(R_k). If the solution is feasible, we go to the next subsection of the algorithm; otherwise, stations with infeasible arrival rates are removed. Then, the resulting solution must be modified if there is not at least one station to provide each normal and emergency service. If this condition is not satisfied, one station is randomly established for that service and then we turn back to the first step.

C. Allocation of Servers to Service Stations

For this subsection, we use the greedy approach.

Step 1: for each service, calculate $\bar{\lambda}_w^k - \Lambda_{kj}$ for all stations (w is the number of the servers of that station). For each station, if the calculated value is positive, the number of servers allocated to that station is enough. Otherwise, it means that patients' arrival rate is more than the standard rate, so more servers should be assigned to it.

Step 2: choose the most negative value calculated for the service in the previous step and allocate one server of that type to it.

Step 3: repeat the previous two steps for the service until at least one of the following conditions is satisfied:

- All calculated values in step 1 become non-negative.
- The servers of that type finish.

Step 4: repeat the previous three steps for all services.

V. COMPUTATIONAL RESULTS

To test the computational performance of the GA-based heuristic, ten problems with corresponding demand areas are designed in such a way that, for example, problem number 1 has one demand area, problem number 2 has two demand areas and so forth. For each problem, more cases are produced by changing the number of available servers (Z_k) for each service (a total of 75 cases are produced). There are four services, namely three normal services, and one emergency service, in these problems. The proposed heuristic is coded in C# and all runs are performed on a computer with 2.27 GHz of CPU and 3 GB of RAM.

The input parameters are produced randomly in the following intervals in the ten test problems: the travel time between demand areas and health centers in the interval [0.25 - 1.25] (hour), the travel time between health centers in the interval [0.2- 1] (hour), the demand rates for four services per hour in the intervals [15- 25], [10 - 20], [5- 10] and [3- 10] respectively. The average service rate for four services is 6, 5, 5 and 4 patients per hour, standard waiting time in the system is 25, 30, 35 and 35 minutes and the minimum arrival rate required to provide services is 4, 3, 3 and 1 patient(s) per hour respectively. For determining π_i, we assume that the importance of reaching a service station for emergency services from demand areas to health centers (first section of the objective function) one unit of time earlier is set at 3. We assume this importance is 2 for normal services between health centers, 5 for emergency services between health centers and 0.5 for undesired arrival rates at service stations.

After linearizing the model, we compared the performance of CPLEX (12.2) and the GA-based heuristic for solving the 75 cases explained above.

For all cases of each problem, the average objective function is measured for both CPLEX and the GA-based heuristic. For the cases of the first four problems which have 1, 2, 3 and 4 demand areas respectively, CPLEX reached the optimal solution in at most 1000 seconds. For the cases of other problems, we got the best solution found by CPLEX in the limited time of two hours. The results are presented in Figure II.

As explained, for the cases of the first four problems, CPLEX guaranteed the optimal solution. For these cases, the deviation of the GA-based heuristic from the optimal solutions is only 0.73% on average. For the cases of the problems with 5, 6 and 7 demand areas, CPLEX and GA-based heuristic have almost the same performance. But for the cases of the problems with 8, 9 and 10 demand areas, the GA-based

heuristic produced better objective functions than those of the CPLEX produced in two hours. As can be clearly seen, as the size of the problem increases, so does the gap between the two graphs.

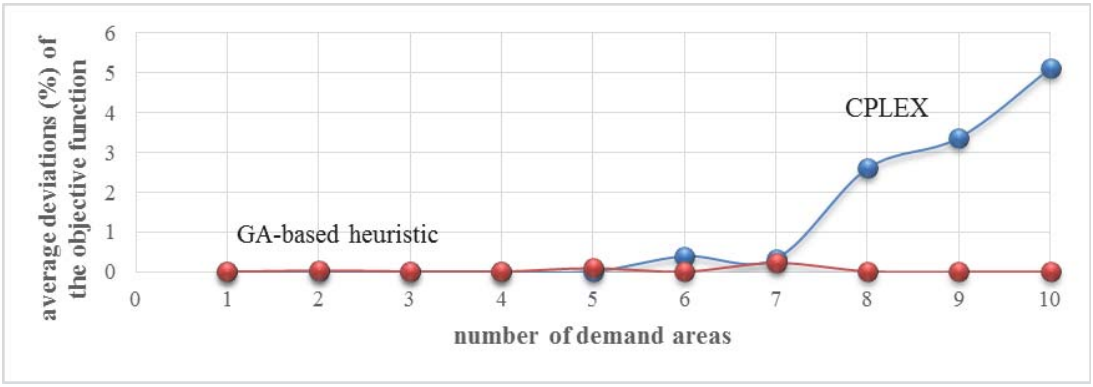

FIGURE II. AGE DEVIATIONS (%) (PERCENTAGE DIFFERENCE OF THE OBJECTIVE FUNCTION VALUES FROM THE BEST VALUE OBTAINED)

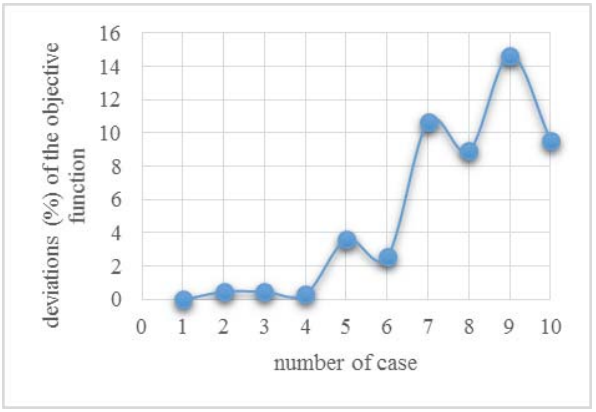

FIGURE III. ATION (%) OF THE OBJECTIVE FUNCTION OF CPLEX FROM GA-BASED HEURISTIC FOR TEN CASES OF THE PROBLEM WITH 10 DEMAND AREAS

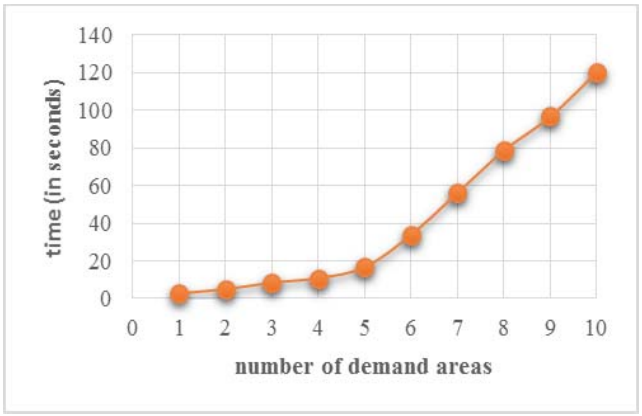

FIGURE IV. 1-AVERAGE SOLUTION TIME (IN SECONDS) FOR THE GA-BASED HEURISTIC FOR TEN PROBLEMS

Generally, in the 75 cases checked, the average deviation is 0.038% for the GA-based heuristic and 1.181% for the CPLEX (what is meant by "deviation" is the percentage difference of the objective function obtained of GA-based heuristic (CPLEX) from the best objective function obtained of both the GA-based heuristic and CPLEX).

For the last problem that has 10 demand areas, ten cases are designed (as explained earlier, in these cases, the number of available servers for four services are gradually increasing). For all cases, GA-based heuristic could reach better objective function than CPLEX. The percentage difference of the objective function of CPLEX from those of GA-based heuristic is depicted in Figure III. As can be seen, as the number of available servers increases in these cases, CPLEX produce less qualified objective function.

Figure IV. shows the average time (in seconds) for the GA-based heuristic for ten problems.

VI. CONCLUSION AND FUTURE RESEARCH

In this article, a mathematical model is developed to simulate patients' choice behavior in choosing health centers. In the proposed model, services provided in the health centers are broken down into normal and emergency ones. Since patients may be shifted to another service after receiving a service by doctors' order, patients' flow between services is considered in the proposed model. The model seeks to minimize the demand weighted total distance traveled by patients between their residential areas and health centers and also between health centers and the weighted sum of undesired deviations from standard arrival rates at service stations. The location of health centers as well as the type of services they offer and the number of servers at each service station are the main determinants of the proposed model

We linearize the proposed mixed integer nonlinear programming model and then solve the small instances using optimization software CPLEX. In addition, to solve the medium and large instances, we have broken down the model to three subsections, namely the location of service stations, calculation of patients' arrival rates at service stations and the allocation of servers to service stations, and have developed a GA-based heuristic to solve the model.

To evaluate the performance of the proposed heuristic, some test problems are generated. In the 75 cases checked, the average deviation is 0.038% for the GA-based heuristic and 1.181% for the CPLEX. In general, GA-based heuristic has an acceptable performance to solve the proposed model, since for

small instances (problems with 1 to 7 demand areas) it performs nearly the same as the CPLEX, but for large instances (problems with 8 to 10 demand areas) it produces better results in much less time than CPLEX.

The proposed model can be extended in different ways. In addition to the assumptions for modeling patients' choice behavior, other factors such as average waiting time and number of services provided in the system may be suitable. In the proposed model, patients go to the centers without an appointment; therefore, it can be changed to an appointment system. Adding some constraints like budget constraints, considering another index for controlling the congestion of the system and considering general service time instead of exponential time for services are other directions for extending the model.

REFERENCES

[1] Daskin, M. and Dean, L. , location of healthcare facilities", A handbook of methods and applications, pp. 43-76,2004.

[2] Panwar,M. and Rathi,K. , "Social Sustainability: Contextual Facility Location Planning Model for Multi-facility Hierarchical heathcare system in India," International Journal of Applied Engineering Research, pp. 275-284, 2014.

[3] Afshari,H. and Peng,Q. , "Challenges and solutions for location of healthcare facilities," Industrial Engineering and Management, 2014.

[4] Vidyarthi, N. and Jayaswal, S. , "Efficient Solution of a Class of Location-allocation problems with stochastic demand and congestion," Computers & Operations Research, 2014.

[5] Boffey, B. Galvão, R. and Espejo, L. , "A review of congestion models in the location of facilities with immobile servers," European Journal of Operational Research 178, pp. 643-662, 2007.

[6] Marianov, V. and Serra, D. , "Probabilistic maximal covering location-allocation for congested system," Journal of Regional Science 38, pp. 401-424, 1998.

[7] Marianov, V. and Serra, D. , "Location–allocation of multiple-server service centers with constrained queues or waiting times," Annals of Operations Research 111, p. Annals of Operations Research 111, 2002.

[8] Pasandideh, H. and Akhavan Niaki, T. , "Genetic application in a facility location problem with random demand within queuing framework," J Intell Manuf (2012) 23, pp. 651-659, 2012.

[9] Aboolian, R. Berman, O. and Drezner, Z. , "Location and allocation of service units on a congested network," IIE Transactions 40 422–433, p. 422–433, 2008.

[10] Batty, M. , "Reilly's challenge: new laws of retail gravitation which define systems of central places," Environ. planning A10, pp. 185-219, 1978.

[11] Huff, D. , "A Probabilist Analysis of Shopping Center Trade Areas," Land Economics,Vol. 39, No. 1, pp. 81-90, 1963.

[12] McFadden, D. , "Conditional logit analysis of quantitative choice behavior. In: Zarembkar P (ed) Frontiers in economics," Academic Press, New York, 1974.

[13] Marianov, V. Rıos, M. Icaza, M. , "Facility location for market capture when users rank facilities by shorter travel and waiting times," European Journal of Operational Research 191, pp. 32-44, 2008.

[14] Zhang, Y. Berman, D. and Verter, V. , " the impact of client choice on preventive healthcare facility network design", " OR Spectrum34, pp. 349-370, 2012.

[15] Zhang, Y. Berman, O. and Verter, V. , "Incorporating congestion in preventive healthcare facility network design," European Journal of Operational Research 198, pp. 922-935, 2009.

[16] Gross M, Harris C. Fundamentals of queueing theory, chapter 4: networks, series and cyclic queues. Wiley, New York. 2nd edition 1985; 229-230.