Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2013, Article ID 716534, 11 pages
http://dx.doi.org/10.1155/2013/716534

Research Article

Congestion Service Facilities Location Problem with Promise of Response Time

Dandan Hu,[1] Zhi-Wei Liu,[2] and Wenshan Hu[2]

[1] *School of Management, South-Central University for Nationalities, Wuhan 430074, China*
[2] *Department of Automation, Wuhan University, Wuhan 430072, China*

Correspondence should be addressed to Zhi-Wei Liu; liuzw@whu.edu.cn

Received 19 July 2013; Revised 15 September 2013; Accepted 10 October 2013

Academic Editor: Pui-Sze Chow

In many services, promise of specific response time is advertised as a commitment by the service providers for the customer satisfaction. Congestion on service facilities could delay the delivery of the services and hurts the overall satisfaction. In this paper, congestion service facilities location problem with promise of response time is studied, and a mixed integer nonlinear programming model is presented with budget constrained. The facilities are modeled as M/M/c queues. The decision variables of the model are the locations of the service facilities and the number of servers at each facility. The objective function is to maximize the demands served within specific response time promised by the service provider. To solve this problem, we propose an algorithm that combines greedy and genetic algorithms. In order to verify the proposed algorithm, a lot of computational experiments are tested. And the results demonstrate that response time has a significant impact on location decision.

1. Introduction

Facility location is a critical decision for a wide range of public and private firms. For example, in public sectors, location decisions for fire stations, ambulances, and other emergency service centers relate directly to the safety of citizens' lives and properties. In private sectors, industries need to locate warehouses, distribution centers, retail outlets, and so forth. These locating decisions concern both costs and competitiveness. In short, the success or failure of both public and private facilities partly depends on the locations chosen for these facilities. Facility location theory has been studied in various forms for hundreds of years. The study formally started by Weber to position a single warehouse [1]. Then, many traditional location models are proposed, including p-median problem [2], covering problem [3, 4], and p-center problem [2]. Following these early investigations, the studies of location theory become popular in recent years. Yang and Zhang [5] studied chain stores location problem with bounded linear consumption expansion function on paths. Yu et al. [6] expanded capacitated facility location problem with consideration of serve radius and economic benefit. Ma et al. [7] studied facility location problem, combined with the feature of demand. Liu et al. [8] presented a location model that assigns online demands to the capacitated regional warehouses. For a survey on models and methods, please see the book edited by Daskin [9] and the review done by ReVelle et al. [10].

The studies that mentioned above mainly concentrate on travel time, physical distance, or some other related travel cost. They assume that facilities are sufficiently large to meet any demand immediately. However, facilities could be congested frequently. To address this issue, congestion facility location problem begins to be considered since Maximum Expected Covering Location Problem (MEXCLP) by Daskin [11] who introduced this model in connection with the location of ambulances and assumed that the probability of each server being busy is predetermined. Then Marianov and Serra [12] introduced queueing theory to address the congestion facility location problems. They applied a constraint to ensure that there is at least a server available on demand with probability α. Huang et al. [13] studied the connections of network that are modeled as M/G/1 queues. Decision variables include selecting connections, assigning flows to the connections and sizing their capacities. To solve this model, they developed an algorithm based on Lagrangian relaxation.

Hu et al. [14] examined bi-objective model based on flow interception problem. The model is formulated from the view of M/M/c queuing system. Service quantity and quality are simultaneously considered as objectives. T. Drezner and Z. Drezner [15] studied the server distribution problem with the objective to minimize the combined travel time and waiting time at the facility for all customers. In their method, the distribution of demand among the facilities is governed by the gravity rule. Other references on the subject include [16–18].

In the literature described above, the models are developed with the objective of optimizing congestion indicators, such as waiting time [14, 15, 18], response time [17], work load [16], and relative congestion costs [13]. However, in many service sectors, congestion indicators are often specified by the service providers. For example, Marianov and Serra [12] considered that the waiting time-limit and queueing length-limit are predetermined.

In real life, response time is often a key competitive priority and represents the firm's commitment for the customer satisfaction. Therefore, specific response time guarantees are often advertised. For examples, Yonghe King, famous for soya bean milk and youtiao, reduces its promised response time from 30 to 20 minutes to enhance its competitiveness. Some take away restaurants often give a discount of food if it is not served within promised time. Jingdong on-line mall promises that customers are able to receive goods within 24 hours after the orders. Within a few years' competition, Jingdong has already taken a leading position in China. The similar situation is common in reality but has not been investigated thoroughly. To keep the promise, proper planning of the facilities is one of the most important actions for service providers. Ideally, the more the service resources they have, the shorter the response time the customer can get. However, due to budget limitation in real life, reasonable facilities location and servers allocation are needed. Therefore, developing a model that addresses this issue could be a main contribution for the current study.

In view of the above described, we consider that the response time-limit is predetermined, which is an important characteristic of our model that distinguishes from others. In this case the objective of model we propose is to maximize requests served within a constant response time in consideration of budget limitation. We concentrate on two parts which affect response time. (1) Travel time is determined by the distance between the service facility and demand node. (2) Sojourn time in facilities is affected by many issues, such as shortage in supply, service interruption (power cut or water cut), strike, and other accidents, but one of the most common is service congestion.

Many existing works [11–14, 16, 17] appoint the value of servers' number at each location in advance. Ideally, the more servers deployed at one facility location, the more effective it could be, but in reality the increasing number of servers also leads to more costs, which is often an important concern for decision makers. Another contribution of our work is that servers' number of each potential facility location could be determined endogenously by the proposed model.

Most closely related to the work in this paper is the study of Berman and Drezner [19]. They introduced an m servers allocation problem. In contrast to the research by Berman and Drezner, there are two differences between their model and ours. First, there are different objectives. In their model, the objective function is to minimize average response time, simple and explicitly convex in the number of servers. In view of the reality of response time-limit, however, our goal is to maximize requests served within a given response time. It is much more complicated and needs further analysis. Second, more generalized constraints. Berman and Drezner assumed that facility location incurs no cost and the total number of servers is known in advance. This is not the case in our problem. We propose constrained budget on the facility location and servers. When location cost is zero, our constraints are reduced to theirs, that is; our constraints are more generalized.

With the consideration of the limited budget, the difficulty of our problem is to solve two contradicting objectives. On the one hand, it is desirable to locate facilities as many as possible so that total travel time for customers could be reduced. On the other hand, with locating costs increasing and budget constraint, the available budget for servers is decreasing, which leads to lower service efficiency and longer waiting time. Therefore, when the demand is high and the travel distance is relatively short, the solution is expected to include fewer facility locations and more servers at each location. When travel time is relatively long, the solution is expected to include more locations and fewer servers at each location. Thus the key to maximize demands satisfied within promised response time lies in the balance of servers' costs and facilities location costs.

This paper is organized as follows. We formulate the problem and provide some analysis in Section 2. In Section 3, we present solution algorithms. Computational results and sensitivity analysis are included in Section 4. In the last section, we provide conclusions and suggestions for future research.

2. Formulation of the Model

In this section, to solve facility location problem with response time-limit, a mixed integer model is formulated. A service facility is modeled as an M/M/c queuing system. Customers arrive at the facility according to a Poisson process. Service time is exponentially distributed. Each facility has multiple servers and the number of servers is a decision variable in our model. It is assumed that services at different facilities are homogenous, which means that the efficiency of every server is the same. In real life, customers generally have no idea of congestion at each facility before they arrive at the service facility so they choose the closest facility. Location and servers both incur costs, and servers' costs are assumed to be linear with the number of servers.

Now, we denote index sets by the following:

(i) I: set of demand nodes, denoted by $i \in I$,

(ii) J: set of candidate locations, denoted by $j \in J$.

Next, we denote parameters by the following:

(i) f_i: demand at node i, yielding to random Poisson distribution with parameter λ_i,

(ii) w_j: sojourn time at facility j, including waiting time and service time,

(iii) t_{ij}: response time of facility j to demand node i,

(iv) T: response time that service facilities have promised,

(v) d_{ij}: travel time between facility j and demand node i,

(vi) B: budget limitation,

(vii) a_j: cost of candidate facility location j,

(viii) b: cost of unit server,

(ix) μ: average service rate per unit server.

The decision variables of the problem areThe decision variables of the problem are

(i) $y_j = \begin{cases} 1 & \text{if a facility is located at node } j\,(j \in J) \\ 0 & \text{otherwise} \end{cases}$

(ii) $x_{ij} = \begin{cases} 1 & \text{if facility } j \text{ serves demand node } i \\ 0 & \text{otherwise.} \end{cases}$

(iii) c_j: the number of servers at facility j.

Given the previous definitions, the location model can be formulated as follows:

$$\text{maximize } Z = E\left(\sum_{i\in I}\sum_{j\in J} f_i P\left(t_{ij} \leqslant T\right) x_{ij}\right), \tag{1}$$

subject to

$$x_{ij} \leqslant y_j, \quad \forall i \in I,\ j \in J \tag{2}$$

$$\sum_{j\in J} x_{ij} = 1, \quad \forall i \in I \tag{3}$$

$$\sum_{k\in J \mid d_{ik} \leqslant d_{ij}} x_{ik} \geqslant y_j, \quad \forall i \in I,\ j \in J \tag{4}$$

$$\sum_{j\in J}\left(a_j y_j + b \cdot c_j\right) \leqslant B \tag{5}$$

$$c_j \leqslant M y_j, \quad \forall j \in J \tag{6}$$

$$\sum_{i\in I} \lambda_i x_{ij} < c_j \mu, \quad \forall j \in J \tag{7}$$

$$x_{ij} \in \{0, 1\}, \quad \forall i \in I,\ j \in J \tag{8}$$

$$y_j \in \{0, 1\}, \quad \forall j \in J \tag{9}$$

$$c_j \in \mathbb{N}, \quad \forall j \in J \tag{10}$$

The model has the objective of maximizing expected demand rates that are served within promised time T. Constraints (2) ensure that facility j can serve demand node i only if facility j is open. Constraints (3) guarantee that demand node i is served by one and only one facility. Constraints (4) assure that each demand node is served by the closest open facility. Constraint (5) is a budget limitation, spending on location and servers. Parameter M, which is a very large number in constraints (6), can be chosen as $M = \lfloor B/b \rfloor$ ($\lfloor a \rfloor$ denotes the largest positive integral that is no more than a). Constraints (6) ensure that servers can be deployed at facility j only if facility j is open. Constraints (7) prevent infinite waiting time. Constraints (8) and (9) are binary constraints, and the last constraints preserve the positive integer restrictions on decision variables of the number of servers at each open facility.

Since service response time includes travel time and sojourn time $t_{ij} = d_{ij} + w_j$, the objective function Z can be transformed as follows:

$$\begin{aligned} Z &= E\left(\sum_{i\in I}\sum_{j\in J} f_i P\left(t_{ij} \leqslant T\right) x_{ij}\right) \\ &= E\left(\sum_{i\in I}\sum_{j\in J} f_i P\left(w_j \leqslant T - d_{ij}\right) x_{ij}\right), \end{aligned} \tag{11}$$

Let $T_{ij} = T - d_{ij}$, $Z = \sum_{i\in I}\sum_{j\in J} \lambda_i F(T_{ij}) x_{ij}$.

$F(T_{ij})$ is a probability distribution function, denoting the probability that demand node i's sojourn time at facility j is not larger than T_{ij}. So the response time restriction in the objective function is transformed to the sojourn time restriction.

For further analysis, we introduce some queuing theory relevant to our research.

In an M/M/c queuing system, the following parameters are always known:

μ: average service rate,

λ: average arrival rate,

c: the number of servers.

Let $\rho = \lambda/\mu$ and $\rho_c = \rho/c$. p_0, the probability that no demand sojourns in system, is denoted by

$$\begin{aligned} p_0 &= \left(\sum_{n=0}^{c-1} \frac{\rho^n}{n!} + \frac{\rho^c}{c!\left(1 - \rho_c\right)}\right)^{-1}, \\ p_c &= \frac{\rho^c}{c!} p_0, \end{aligned} \tag{12}$$

expressing the probability that the number of demands sojourning in system is c.

Then, $F(t)$, the probability that sojourn time in system is not larger than t, is shown as follows [20]:

when $c = 1$, $F(t) = 1 - e^{-(\mu-\lambda)t}$;

when $c > 1$,

$$F(t) = \begin{cases} 1 - \left(1 + \dfrac{p_c \mu t}{1 - \rho_c}\right) e^{-\mu t}, & \rho = c - 1, \\ 1 - \left(1 + \dfrac{p_c}{(c - 1 - \rho)(1 - \rho_c)}\right) e^{-\mu t} + \dfrac{p_c}{(c - 1 - \rho)(1 - \rho_c)} e^{-c\mu(1-\rho_c)t}, & \rho \neq c - 1. \end{cases} \tag{13}$$

As we know, distribution function of waiting time t is

$$W_q(t) = 1 - \frac{p_c}{1 - \rho_c} e^{-\mu(c-\rho)t}, \tag{14}$$

and delay probability, the probability that at least one demand is in system for service, is

$$\mathrm{DT}(c) = 1 - W_q(0) = \frac{p_c}{1 - \rho_c}. \tag{15}$$

Let

$$g(c) = \mathrm{DT}(c)\, e^{-\mu(c-\rho)t}. \tag{16}$$

Lemma 1. *$g(c)$ is a convex function.*

Proof. Second derivative of $g(c)$ is

$$g''(c) = \mathrm{DT}''(c)\, e^{-\mu(c-\rho)t} + \mu^2 \mathrm{DT}(c)\, e^{-\mu(c-\rho)t} - 2\mathrm{DT}'(c)\, e^{-\mu(c-\rho)t}. \tag{17}$$

As $\mathrm{DT}(c)$ is a nonincreasing function in c [21], $\mathrm{DT}'(c) \leqslant 0$ and $\mathrm{DT}''(c) \geqslant 0$. Then, $g''(c) \geqslant 0$ is resulted, and Lemma 1 can be proved. □

Theorem 2. *Sojourn time's distribution function $F(t)$ is a concave function in c.*

Proof. From Lemma 1, $g(c)$ is a convex function. Then, waiting time distribution function $W_q(t) = 1 - g(c)$ is a concave function in c; hence,

$$\frac{\partial^2 W_q(t)}{\partial c^2} \leqslant 0. \tag{18}$$

Sojourn time w is made up of two parts, waiting time w_q and service time s, $w = w_q + s$. Then $F(t)$ can be expressed as follows:

$$\begin{aligned} F(t) &= P\{w \leqslant t\} \\ &= P\left\{w_q + s \leqslant t\right\} \\ &= \int_0^t P\left\{w_q \leqslant t - t_s\right\} dP\left\{s \leqslant t_s\right\}. \end{aligned} \tag{19}$$

Since it is an M/M/c system, service time s yields to a negative exponential distribution, and $F(t)$ can be changed to

$$\begin{aligned} F(t) &= \int_0^t W_q(t - t_s)\, d\left(1 - e^{-\mu t_s}\right) \\ &= \int_0^t W_q(t - t_s)\, \mu e^{-\mu t_s} dt_s, \end{aligned} \tag{20}$$

$$\frac{\partial^2 F(t)}{\partial c^2} = \int_0^t \frac{\partial^2 W_q(t - t_s)}{\partial c^2} \mu e^{-\mu t_s} dt_s. \tag{21}$$

From

$$\frac{\partial^2 W_q(t - t_s)}{\partial c^2} \leqslant 0, \tag{22}$$

$$\frac{\partial^2 F(t)}{\partial c^2} \leqslant 0, \tag{23}$$

and Theorem 2 is resulted. □

According to the proof of Theorem 2, $F(t)$ owns the following two properties:

(1) $F(t)$ is a non-decreasing function in c,

(2) $F(t)$ is a concave function in c, that is to say, $F_{c+1}(t) - F_c(t)$ is a nonincreasing function in c.

The properties are similar to [19], and therefore we can get the optimal solution of servers allocation by greedy algorithm in the condition that location decisions have been made. Now, the problem is transformed to the model only with the location decision.

The model we have built is a combinatorial and nonlinear optimization problem. Without taking account of congestion situation, the waiting time is zero, and then the problem can be reduced to the maximum covering location problem. Thus the maximum covering location problem is a special case of this problem. It is known that the maximum covering location problem is NP-hard [9]. Clearly, this problem is also NP-hard. In order to solve large size problems, heuristics are considered our choice.

3. Algorithms

To solve the traditional location-allocation problem, various algorithms have been provided. Exact solution methods for location problem have been proposed and investigated in previous reports. Holmberg [22] embedded the dual ascent method in a branch-and-bound framework. Sasaki et al.

[23] proposed enumeration-based approach. de Camargo et al. [24] used benders decomposition method for the uncapacitated multiple allocation hub location problem.

Traditional location problems are NP-hard. Increase in the polynomial order of the problem leads to an exponential explosion in the computation time, and exact methods can only solve small instances of the presented problem. Thus, most of the literature applies efficient heuristics to solve the problem, which includes local search, greedy heuristic, simulate annealing, tabu search and Lagrangian relaxation, and so forth. Hybrid methods have also been studied recently [18, 19, 25]. For a more complete review of these algorithms, the reader can be referred to [26].

Algorithms for solving the congestion servers location-allocation problem have been discussed by Berman and Drezner [19] and Aboolian et al. [18]. Berman and Drezner developed three heuristic approaches, namely tabu search, simulated annealing and genetic algorithms. Compared with the precise results of total enumeration, the three heuristics are showed to perform very well, especially the genetic algorithm, which obtains the results with the smallest gap and in the shortest computation time. Aboolian et al. presented descent algorithm and genetic algorithm. Large-scale numerical examples illustrated that genetic algorithm performs more efficiently than descent algorithm. These works have proved that genetic algorithm is more efficient to solve congestion servers location-allocation problem than other heuristics presented in [18, 19].

This problem can be decomposed in two smaller problems: at a higher level, named as master problem (MP), the location decisions are made; while at an inferior level, known as subproblem (SP), the determination of the servers' number at each location is done. The MP is a 0-1 integer programming problem, while the SP is an allocation problem. We consider a hybrid algorithm of combining genetic algorithm with greedy algorithm to solve the model. Genetic algorithm is applied to solve MP, and precise solution of SP can be obtained by greedy algorithm. The results of our algorithm for small size problems are compared with the results obtained from numerical algorithm, in order to verify the performance of the algorithm. Although there are alternative algorithms in the literature [18, 19], our choice is driven by the following reasons. Firstly, genetic algorithm is proved to be more efficient to solve congestion facility location problem than other heuristics [18, 19]. Secondly, due to the properties we proved in Section 2, results of servers allocation are optimal by greedy algorithm.

3.1. Precise Algorithm of Servers Allocation. Greedy algorithms have been comprehensively used in location problems [27–30]. In these works, greedy algorithms are developed to determine location decision and the solutions are approximations. However, different from other literature, greedy algorithm in this paper is used to solve servers allocation problem. Due to sojourn time distribution function's special properties, we can get precise results of servers' number at each facility location provided that location decision has been made. Assuming that location decision is S, now compute the number of servers at each location with the budget constraint. Procedure is as follows.

(1) For each node j ($j \in S$), compute average arrival rate $\overline{\lambda}_j$. $\overline{\lambda}_j = \sum_{i \in L_j(S)} \lambda_i$, where $L_j(S)$ is the set of demand nodes from which facility $j \in S$ is the closest in S.

(2) For each node j ($j \in S$), find the minimum number of servers necessary to serve the customer flows. This minimum number is

$$k_j = \left\lfloor \frac{\overline{\lambda}_j}{\mu} \right\rfloor, \tag{24}$$

in which k_j is the largest integral that is no more than $\overline{\lambda}_j/\mu$.

(3) If

$$\sum_{j \in S} \left(b \cdot k_j + a_j\right) > B, \tag{25}$$

in which B is the given budget, b is unit server's costs and a_j is the location costs of facility j, there is no feasible solution, and the computing should stop.

If

$$B - b < \sum_{j \in S} \left(b \cdot k_j + a_j\right) \leqslant B, \tag{26}$$

the number of servers is k_j, and stop.

If

$$\sum_{j \in S} \left(b \cdot k_j + a_j\right) \leq B - b, \tag{27}$$

go to step 4 to allocate the rest servers.

(4) Compute the objective increase through increasing the number of servers by one at facility j; that is,

$$\Delta_j = \sum_{i \in L_j(S)} \lambda_i \left(F_{k_j+1}(t_i) - F_{k_j}(t_i)\right), \tag{28}$$

in which $L_j(S)$ expresses the set of demand nodes served by facility j. Compute Δ_j for each $j \in S$, choose the maximum Δ_{j^*} ($j^* \in S$), and then let $k_{j^*} = k_{j^*} + 1$. Repeat Procedure (4) until the rest of budget is not enough to pay for another server.

3.2. Genetic Algorithm. Genetic algorithm was first applied to location-allocation problems by Hosage and Goodchild in 1986 [31]. It has been widely applied to solve location problems. Correa et al. [32] proposed a genetic algorithm for solving a capacitated p-median problem. Jaramillo et al. [33] introduced genetic algorithm for solving uncapacitated and capacitated fixed charge problems, the maximum covering problem, and competitive location models. Balakrishnan et al. [34] presented a hybrid approach of combining genetic

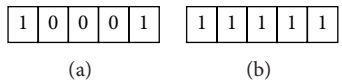

Figure 1: Binary representation of an individual's chromosome.

algorithm with dynamic programming for the dynamic facility location problem. Genetic algorithm is also proved to be efficient to solve congestion facility location problem [18, 19, 35].

In this section, we propose a genetic algorithm to find facility locations. In the proposed method, good chromosomes are identified through crossover, mutation and selection operations. We apply greedy algorithm to decide the number of servers at each location. Crossover and mutation we use are similar as the methods described in the literature mentioned above, which is not the main focus of our work. Only the major operations that we have made are described in this section.

3.2.1. Code. A candidate solution is represented as follows: the representation scheme developed is a $|J|$-bit binary string as the chromosome structure, where $|J|$ is the number of potential facility locations. A value of 1 for the jth bit implies that a facility is located at the jth location. For instance, considering a problem with 5 potential facility locations, the binary representation of an individual's chromosome is illustrated in Figure 1. Figure 1(a) shows the situation when only two facilities are located in potential locations one and five. Figure 1(b) illustrates the situation when facilities are located at all of the potential facility locations.

3.2.2. Crossover, Selection, and Mutation Operations. Randomly select two members from the parental population and merge them to produce offspring. If the offspring is better than the worst parental population member and different from its parent, replace the worst member by the offspring. Repeat above crossover $|J|$ times. Select $|J|$ members from parent and child populations with best objectives. A random mutation pattern is generated for each chromosome. For each gene of the chromosome, a uniform random number between 0 and 1 is generated. If this number is less than a given number p_m, the gene mutates from 0 to 1 or from 1 to 0. If the individual after mutation is not satisfied with budget constraint, the mutation is considered as invalid.

3.2.3. Feasible Operation. We define a candidate solution (individual) for our problem as feasible if the total cost including location and minimum servers costs are no more than the given budget. As initial or crossover individual may produce infeasible solution, for crowds diversity and avoiding local optimum solutions, randomly select one gene, mutating from 1 to 0 until satisfying budget constraint.

The algorithm process is shown in Figure 2.

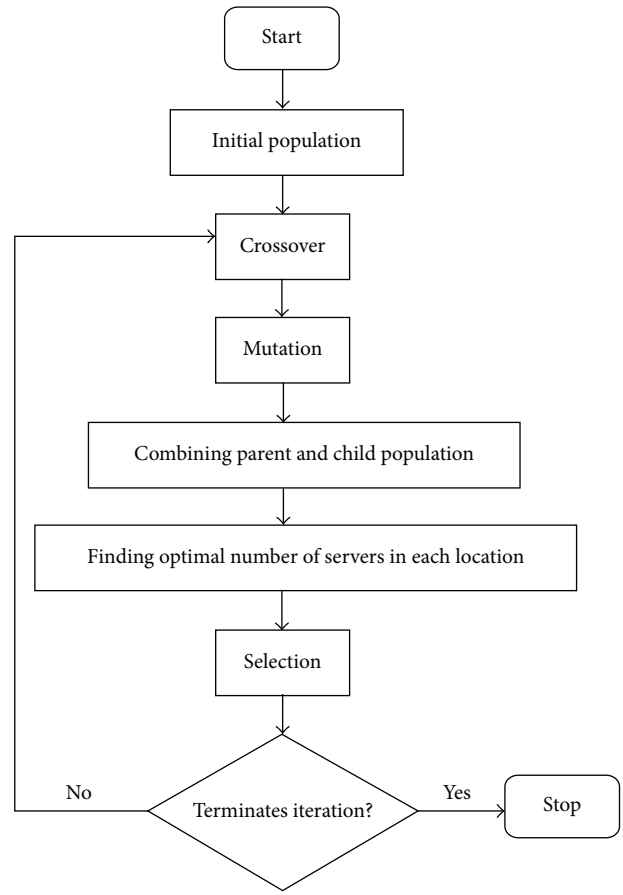

Figure 2: Process flowchart.

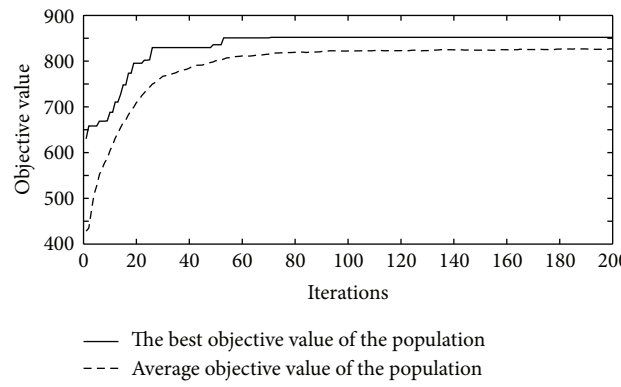

Figure 3: The convergence of genetic algorithm.

4. Example and Computational Experience

The algorithm is coded in Matlab 7.0, and computational experiments are conducted on a PC with 2.2 GHz Intel Pentium Dual E2200 and 2.00 GB RAM. Demand nodes are randomly generated on a $[0, 50][0, 50]$ plane. Candidate facility locations are defined an dconsidered as demand nodes. Each location cost and average demand rate are both random values which vary between 20 and 50. Unit server cost $b = 8$, and average service rate of unit server $\mu = 8$. We set the parameters of genetic algorithm as follows: population

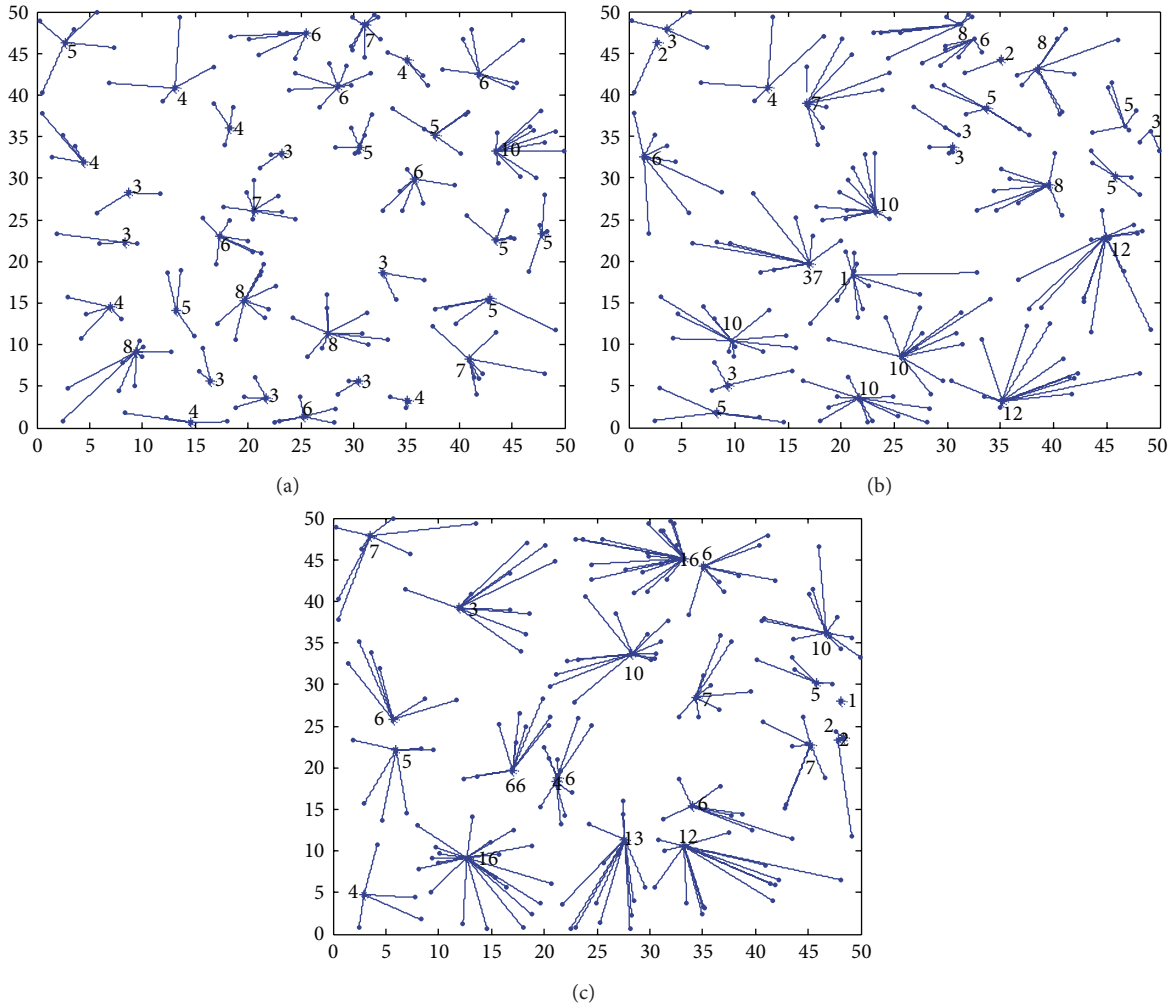

Figure 4: Optimization results with different T. (a) $T = 5$, (b) $T = 15$, and (c) $T = 20$.

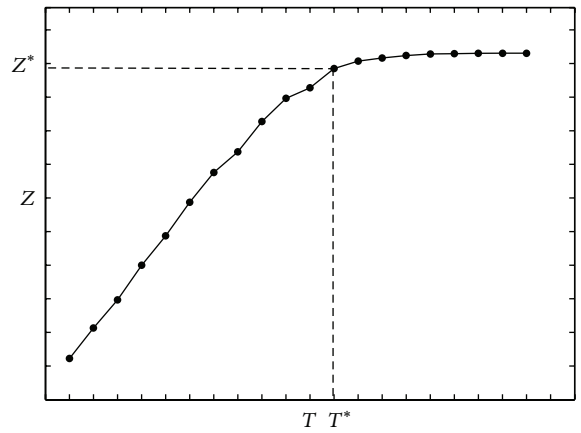

Figure 5: Sensitivity analysis with respect to promised response time T.

size is equal to node, maximum number of generation is equal to 200, and the mutation probability $p_m = 0.1$. Figure 3 shows the convergence of this algorithm (the number of demand nodes $|I| = 50$, $B = 900$, $T = 5$).

Table 1 compares the results of genetic and exact algorithm. Optimal solutions are obtained by numerical algorithm but the computation time increases exponentially with the number of facilities. Considering computation time, the numbers of the facility locations in the test problems are set to be a value no more than 6; that is,

$$\frac{\left(B - b\sum_{i\in I}\lambda_i/u\right)}{20} < 7. \tag{29}$$

seven test problems of varying size are constructed. For each test problem, 50 experiments are randomly generated, leading to a total of 350 experiments. The gap is measured by $(Z^* - Z)/Z^*$, where Z^* is the optimum value and Z is the objective value by our algorithm. From results, the range of gaps is from 0% to 9.65%, and the average gap shifts from 0 to 2.09%. In these experiments, the performance is quiet good and genetic algorithm finds the optimal solution at least 50%.

Then, we set budget $B = 2500$, and the number of demand points $|I| = 200$. Figures 4(a), 4(b) and 4(c) respectively show the facility locations, demand allocation and servers distribution under different promised response time $T = 5$, 15 and 20. The number beside location point means the number

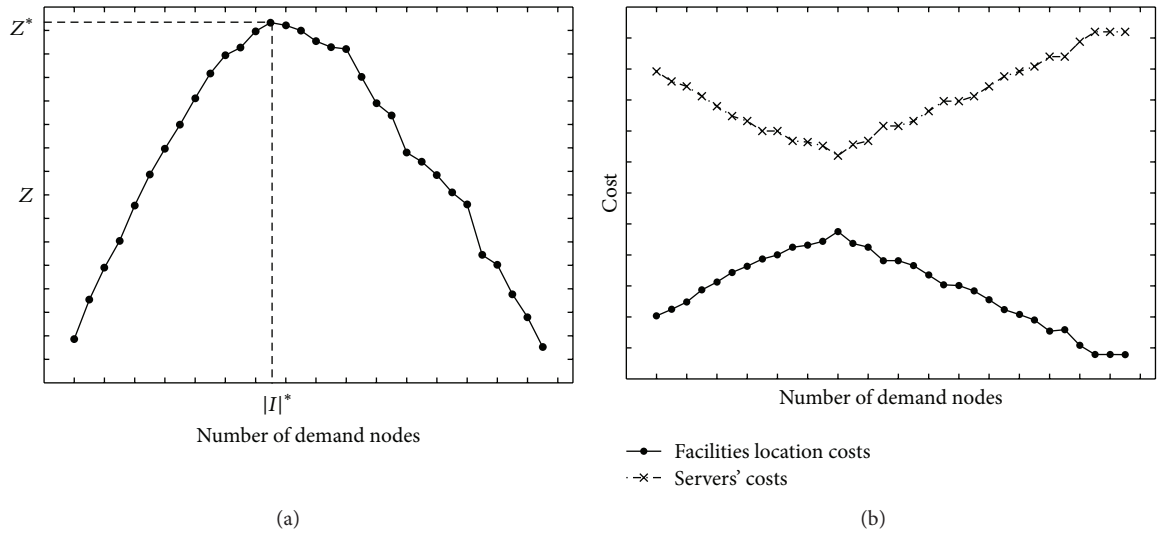

Figure 6: Sensitivity analysis with respect to number of demand nodes.

Table 1: Results of genetic and exact algorithms.

Node	Budget	Genetic algorithm					Numerical algorithm time (S)
		Maximum gap	Minimum gap	Average gap	Optimum solution	Time (S)	
30	350	0	0	0	100%	14.82	45.3
40	400	3.87%	0	1.61%	50%	24.72	78.4
50	450	9.65%	0	2.09%	50%	30.64	523.5
55	480	8.17%	0	1.41%	60%	31.12	335.5
60	500	5.21%	0	0.72%	80%	35.03	487.2
70	550	3.35%	0	1.12%	50%	40.8	1537.6
80	650	6.94%	0	1.23%	70%	55.78	7643.3

of servers at the facility. When $T = 5$, it is desirable to locate more facilities and distribute less servers at each facility. In Figure 4(b), when $T = 15$, we find that the number of facilities is equal to 34, servers at each location are no more than 8, and the decentralized service makes travel time reduced. With the increase of T, the number of facilities is smaller and servers are pooling to reduce service time. Figure 4 illustrates that the promised response time has a strong impact on the location decisions.

In the second set of experiments, we are interested in testing the effect of different parameters. The parameters studied are originally set as follows: promised response time $T = 5$, number of demand nodes $|I| = 30$ and budget $B = 600$. We use three instances for above three parameters sensitivity analysis. For each instance, we vary one parameter under study while keeping two other parameters as original values. The results are presented in Figures 5–7.

Figure 5 presents objective value analysis with respect to promised response time. We find that as promised response time T increases, the objective value, the number of customers served within T, increases as well. However, the objective value Z increases little after the promised response time threshold. Figure 5 shows that when $T > T^*$, the number of customers served within T barely changes. Therefore, through Figure 5 we characterize the reasonable range of promised response time for the service provider, which is no more than T^*. And there is no need to make a longer promise response time decision. The reasons are as follows: (i) longer promised response time ($T > T^*$) has little impact on quantity of customers served within T. (ii) longer promised response time makes service facilities less attractive to customers and weakens their competitiveness.

In Figure 6, we show the sensitivity analysis with respect to the number of demand nodes varying. In Figure 6(a), customers served within promised response time increase initially and then decrease with demand nodes increase. When demand is small, there are enough facilities and servers for service requirements. Although increasing demand leads to reduction of percentage of demand served within T, the impact of percentage reduction is smaller than the increase of demand. Therefore, we can see the result in Figure 6(a) that the objective value is increasing with demand increase before point $|I| = |I|^*$. Z is maximized at point $|I| = |I|^*$. After that point, the percentage of demand served within T drops sharply as $|I|$ increases. Therefore, increasing demand leads to lower demand served within promised time instead. Figure 6(a) illustrates that a service provider may choose service area to provide high-quality service to a limited number

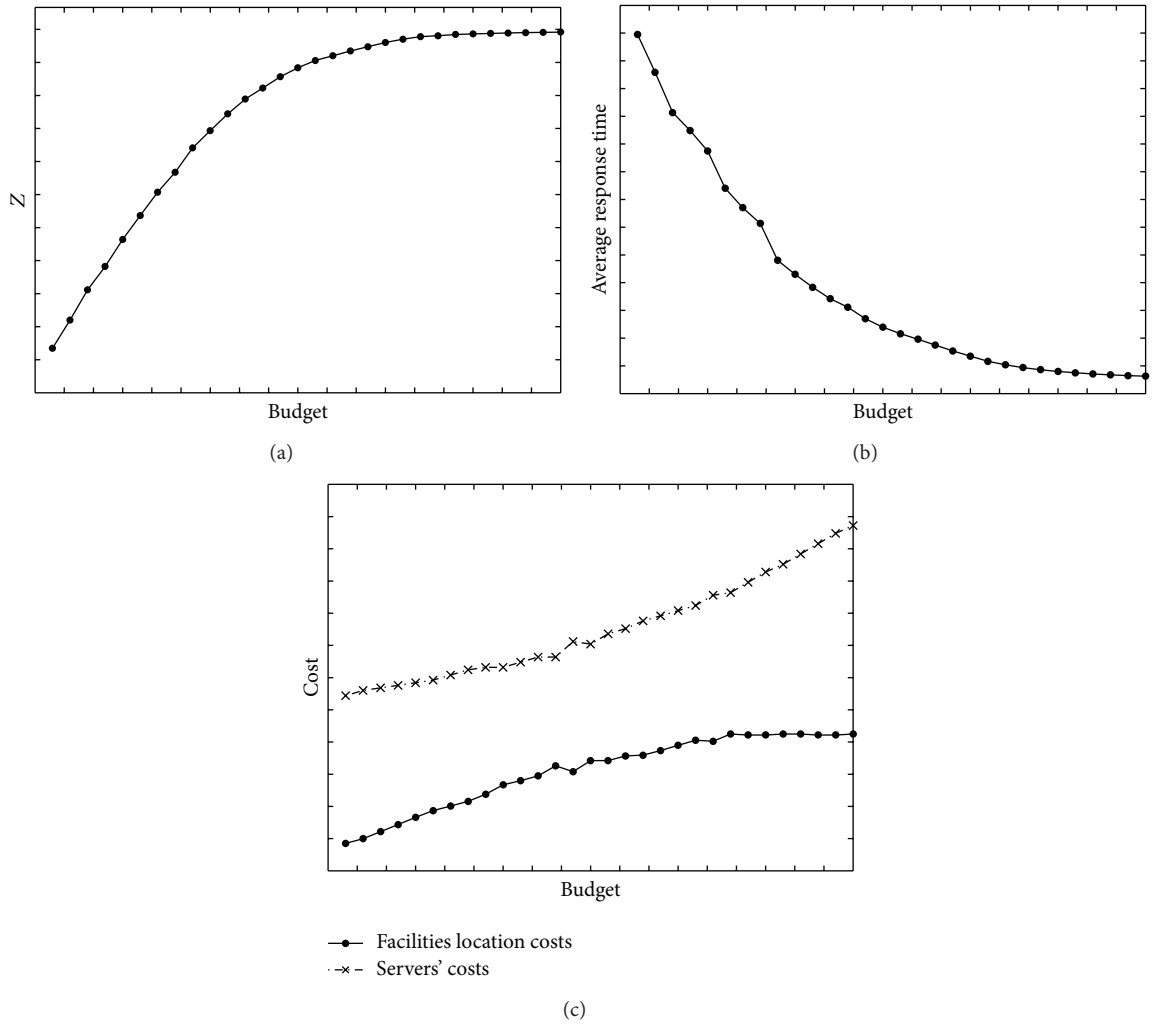

Figure 7: Sensitivity analysis with respect to budget.

of customers, or it may provide lower-quality service to a large number of customers. Therefore, for the service provider, more demand may not be better, which is consistent with the reality. In real life, take away service, such as KFC Delivery, and McDelivery, provides service in designated areas, not all the neighborhoods or streets. The above situation is also our further research on servers location and service areas selection.

Figure 6(b) demonstrates the trade-off between facilities location costs and servers' costs with the increasing demand when the budget is given. From Figure 6(b), as demand expanding, a service provider should increase investment in the construction of new facilities and decrease investment in service efficiency. As demand continues to increase, capital investment in facilities and servers is just converse, from which we can estimate that the demand is too large for the servers to serve and the average waiting time must be longer than the average travel time. Therefore, the action of increasing the number of servers should be taken. The reasons that an increase in the demand results in a decrease in the facility location costs, we think, are as follows: (i) As the budget constraint, increasing servers costs must lead to a decrease in location costs. (ii) Decrease locations will result in the number of servers increase at per facility, and pooling of servers can also decrease waiting time. Consequently, if demand is too small or too large, then we can predict that facilities investment should be small and funds should be focused on hiring more service personnel and purchasing or leasing more equipment for service.

Figure 7 illustrates the variation of satisfied customers, average response time and costs of location and servers with different budget. From Figures 7(a) and 7(b), with the budget increasing, we observe that the increment of Z and decrement of average response time are reduced gradually. We estimate that the objective function and the average response time may have concave-convex quality of the budget. Figure 7(c) demonstrates the investment in facility location and servers with different budgets. Although the investment in locations and servers both increase, their increments are different. The investment in facility location is approximately concave in the budget, whereas the investment in servers is convex. That is, the increment of facility location costs is decreasing, and

the increment of serves costs is increasing with the budget increase. There exists a budget threshold point B^*, on which the increments of facility location and servers are equal. We can estimate that when the budget is small, the increment of location costs is much more than the increase of servers' costs. At this moment, if additional budget is allocated, the management would devote most of attention and budget to increasing distribution of firms. Gradually, with the budget increase, the gap between the increment of location costs and servers' costs becomes smaller and smaller. When budget $> B^*$, increments would be inverse. That means that the investment in service efficiency begins to pay more attention after necessary facilities and service resources construction, if the budget is still abundant.

5. Conclusions and Future Research

In this paper, a model of servers location-allocation problem with promise of response time has been presented. The objective of the model is to maximize the demands satisfied in promised response time, through finding proper locations and according servers. Because of budget constraint, balancing of location costs and servers' costs could decrease response time. System waiting time distribution function is proved to be concave in the number of servers. According to this property, precise algorithm for allocation of servers can be obtained by greedy algorithm. We propose a hybrid algorithm that combines greedy and genetic algorithms, which is proved to be fine compared with exact results by lots of computational experiments.

Finally, the future research could be extended as follows.

(1) More efficient algorithms need to be studied deeply. For example, hybrid methods (also called matheuristics) may be explored as efficient methods for congestion facility location problem.

(2) General service time distribution could be studied instead of exponential distribution in this paper.

(3) The model can be reformulated with incorporation of service response time restrictions, which can be stated as the probability of waiting for no more than x people is greater than α.

(4) All facilities are assumed to be independent in this paper. In future research, cooperative service could be considered.

(5) Customers are assumed to visit the closest open facility in this paper. In future research, the possibility of dynamically assigning a customer to a facility depending on current facility loadings could be studied with an appropriate dispatching rule.

Acknowledgments

This work was supported in part by the Humanity and Social Science Youth foundation of Ministry of Education of China under Grants 11YJC630063, the National Natural Science Foundation of China under Grants 61304152 and 61004030, and the China Postdoctoral Science Foundation funded project 2012M511258 and 2013T60738.

References

[1] A. Weber, *Über den Standort der Industrien*, JCB Mohr, 1909.

[2] S. L. Hakimi, "Optimum locations of switching centers and the absolute centers and medians of a graph," *Operations Research*, vol. 12, no. 3, pp. 450–459, 1964.

[3] C. Toregas, R. Swain, C. ReVelle, and L. Bergman, "The location of emergency service facilities," *Operations Research*, vol. 19, no. 6, pp. 1363–1373, 1971.

[4] R. Church and C. ReVelle, "The maximal covering location problem," *Papers of the Regional Science Association*, vol. 32, no. 1, pp. 101–118, 1974.

[5] J. Yang and M. Zhang, "Chain stores location problem with bounded linear consumption expansion function on paths," in *Proceedings of the 4th International Conference on Wireless Communications, Networking and Mobile Computing (WiCOM '08)*, October 2008.

[6] H. Yu, L. Gao, and Y. Lei, "Model and solution for capacitated facility location problem," in *Proceedings of the 24th Chinese Control and Decision Conference*, pp. 1773–1776, 2012.

[7] Y. Ma, L. Li, and J. Yang, "A gravitational facility location problem based on prize-collecting traveling salesman problem," in *Proceedings of the IEEE International Conference on Automation and Logistics*, pp. 511–516, 2012.

[8] K. Liu, Y. Zhou, and Z. Zhang, "Capacitated location model with online demand pooling in a multi-channel supply chain," *European Journal of Operational Research*, vol. 207, no. 1, pp. 218–231, 2010.

[9] M. S. Daskin, *Network and Discrete Location: Models, Algorithms, and Applications*, John Wiley & Sons, New York, NY, USA, 2005.

[10] C. S. ReVelle, H. A. Eiselt, and M. S. Daskin, "A bibliography for some fundamental problem categories in discrete location science," *European Journal of Operational Research*, vol. 184, no. 3, pp. 817–848, 2008.

[11] M. S. Daskin, "A maximum expected covering location model: formulation, properties and heuristic solution," *Transportation Science*, vol. 17, no. 1, pp. 48–70, 1983.

[12] V. Marianov and D. Serra, "Probabilistic, maximal covering location-allocation models from congested systems," *Journal of Regional Science*, vol. 38, no. 3, pp. 401–424, 1998.

[13] S. Huang, R. Batta, and R. Nagi, "Distribution network design: selection and sizing of congested connections," *Naval Research Logistics*, vol. 52, no. 8, pp. 701–712, 2005.

[14] D. Hu, C. Yang, and J. Yang, "Budget constrained flow interception location model for congested systems," *Journal of Systems Engineering and Electronics*, vol. 20, no. 6, pp. 1255–1262, 2009.

[15] T. Drezner and Z. Drezner, "The gravity multiple server location problem," *Computers & Operations Research*, vol. 38, no. 3, pp. 694–701, 2011.

[16] R. D. Galvão, L. G. A. Espejo, B. Boffey, and D. Yates, "Load balancing and capacity constraints in a hierarchical location model," *European Journal of Operational Research*, vol. 172, no. 2, pp. 631–646, 2006.

[17] H. K. Rajagopalan and C. Saydam, "A minimum expected response model: formulation, heuristic solution, and application," *Socio-Economic Planning Sciences*, vol. 43, no. 4, pp. 253–262, 2009.

[18] R. Aboolian, O. Berman, and Z. Drezner, "The multiple server center location problem," *Annals of Operations Research*, vol. 167, pp. 337–352, 2009.

[19] O. Berman and Z. Drezner, "The multiple server location problem," *Journal of the Operational Research Society*, vol. 58, no. 1, pp. 91–99, 2007.

[20] Z. K. Guo, L. S. Chang, J. Yin, and J. F. Liu, "Distribution of waiting time in a class of queuing systems," *Journal of Inner Mongolia University*, vol. 39, no. 4, pp. 375–379, 2008.

[21] A. Jagers and E. Van Doorn, "Convexity of functions which are generalizations of the erlang loss function and the erlang delay function," *SIAM Review*, vol. 33, no. 2, pp. 281–282, 1991.

[22] K. Holmberg, "Exact solution methods for uncapacitated location problems with convex transportation costs," *European Journal of Operational Research*, vol. 114, no. 1, pp. 127–140, 1999.

[23] M. Sasaki, T. Furuta, and A. Suzuki, "Exact optimal solutions of the minisum facility and transfer points location problems on a network," *International Transactions in Operational Research*, vol. 15, no. 3, pp. 295–306, 2008.

[24] R. S. de Camargo, G. Miranda Jr., and H. P. Luna, "Benders decomposition for the uncapacitated multiple allocation hub location problem," *Computers and Operations Research*, vol. 35, no. 4, pp. 1047–1064, 2008.

[25] Y. Wang and Z. Cai, "A dynamic hybrid framework for constrained evolutionary optimization," *IEEE Transactions on Systems, Man, and Cybernetics B*, vol. 42, no. 1, pp. 203–217, 2012.

[26] V. Maniezzo, T. Stèutzle, and S. Voss, *Matheuristics: Hybridizing Metaheuristics and Mathematical Programming*, vol. 10, Springer, 2009.

[27] R. A. Whitaker, "A fast algorithm for the greedy interchange for large-scale clustering and median location problems," *INFOR Journal*, vol. 21, no. 2, pp. 95–108, 1983.

[28] S. Guha and S. Khuller, "Greedy strikes back: improved facility location algorithms," in *Proceedings of the 9th Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 649–657, Society for Industrial and Applied Mathematics, New York, NY, USA, 1998.

[29] K. Jain, M. Mahdian, and A. Saberi, "A new greedy approach for facility location problems," in *Proceedings of the 34th Annual ACM Symposium on Theory of Computing*, pp. 731–740, ACM, New York, NY, USA, 2002.

[30] G. Cornuejols and J. M. Thizy, "Location problems, set covering problems and the greedy algorithm," Working Paper 25-80-81, Graduate School of Industrial Administration, Carnegie-Mellon University, 1981, Available as Working Paper 02-51, School of Management, University of Ottawa, 2002.

[31] C. M. Hosage and M. F. Goodchild, "Discrete space location-allocation solutions from genetic algorithms," *Annals of Operations Research*, vol. 6, no. 2, pp. 35–46, 1986.

[32] E. S. Correa, M. T. A. Steiner, A. A. Freitas, and C. Carnieri, "A genetic algorithm for solving a capacitated p-median problem," *Numerical Algorithms*, vol. 35, no. 2–4, pp. 373–388, 2004.

[33] J. H. Jaramillo, J. Bhadury, and R. Batta, "On the use of genetic algorithms to solve location problems," *Computers & Operations Research*, vol. 29, no. 6, pp. 761–779, 2002.

[34] J. Balakrishnan, C. H. Cheng, D. G. Conway, and C. M. Lau, "A hybrid genetic algorithm for the dynamic plant layout problem," *International Journal of Production Economics*, vol. 86, no. 2, pp. 107–120, 2003.

[35] H. Shavandi and H. Mahlooji, "A fuzzy queuing location model with a genetic algorithm for congested systems," *Applied Mathematics and Computation*, vol. 181, no. 1, pp. 440–456, 2006.