Scientia Iranica E (2019) 26(2), 909–931

Sharif University of Technology
Scientia Iranica
Transactions E: Industrial Engineering
http://scientiairanica.sharif.edu

Hypercube queuing models in emergency service systems: A state-of-the-art review

M. Ghobadi[a], J. Arkat[a], and R. Tavakkoli-Moghaddam[b,c,*]

a. *Department of Industrial Engineering, University of Kurdistan, Sanandaj, Iran.*
b. *School of Industrial Engineering, College of Engineering, University of Tehran, Tehran, Iran.*
c. *Arts et Métiers ParisTech, LCFC, Metz, France.*

Received 7 April 2016; received in revised form 14 February 2017; accepted 13 January 2018

KEYWORDS
Hypercube queuing model;
Facility location;
Emergency service system.

Abstract. This study provides a review of Hypercube Queuing Models (HQMs) in Emergency Service Systems (ESSs). This survey presents a comprehensive review and taxonomy of models, solutions, and applications related to the HQM after Larson [Brandeau, M.L. and Larson, R.C. "Extending and applying the hypercube queueing model to deploy ambulances in Boston", *Management Science*, 22, pp. 121-153 (1986).] In addition, the structural aspects of HQMs are examined. Important contributions of the existing research are addressed by taking into account multiple factors. In particular, the integration of location decisions with HQMs for designing an ESS is discussed. Finally, a list of issues for future studies is presented.

1. Introduction

Emergency Service Systems (ESSs) provide first care services when incidents occur and ensure public health and safety. In these systems, the customer's situation is usually critical and unstable. This means that a delay in providing services may cause death or serious injuries. Given these conditions and, in general, the uncertainty in these problems, decision-making becomes more complex for managers. The design of ESSs requires strategic and tactical decisions [1]. The strategic decisions determine the number and location of servers. The dispatching policy that illustrates the decision about which a server will respond to a request, and the server's coverage area is specified by tactical decisions.

ESSs can be classified into two main categories, namely customer-to-server and server-to-customer systems. In the first category, servers are immobile, and customers should visit them to receive a service. In the second case, servers are mobile and provide a service at the customer's location. As an example, in the case of fire, fire trucks are dispatched to the scene; in Emergency Medical Systems (EMSs), ambulances travel to the accident location. A system with mobile servers is called emergency response system. In a system with immobile servers, servers are usually considered indistinguishable from each other (e.g., seats of an airplane or beds of a hospital). In these cases, it does not matter exactly which server is busy, and only the number of busy servers is important [1]. On the other hand, mobile servers can be modeled as servers more precisely distinguishable from each other. That is, servers operate independently and can have different features (i.e., different preferences and mean service times), and a server's workload may be changed by the server's location.

The Hypercube Queuing Model (HQM), which was proposed by Larson [2] and used by many re-

[*] *Corresponding author. Tel.: +98 21 82084183; Fax: +98 21 88013102*
E-mail addresses: m.ghobadi@eng.uok.ac.ir (M. Ghobadi); j.arkat@uok.ac.ir (J. Arakat); tavakoli@ut.ac.ir (R. Tavakkoli-Moghaddam)

doi: 10.24200/sci.2018.4515.0

searchers, is a descriptive model suitable for modelling server-to-customer systems. Over the years, the model has been applied to a large number of public and private emergency systems, such as police departments and ambulance services. In this research, development history of the model and its basic ideas are reviewed, and the implementation of this model is discussed.

1.1. Models description

The basic idea of this model is to develop the state-space description of a queuing system to use complex dispatching policies and illustrate each server individually. These models consider the spatial and temporal complexity of the area under study and are suitable for centralized systems. In a centralized system, each customer calls the central unit (i.e., dispatcher), and it dispatches the first idle server to that customer, according to a preference list. This list is prepared based on factors, such as distance and customer requirements. For example, if the list is ordered based on the distances between customers and servers, then once an emergency call is received, the closest server is dispatched. If the closest server is busy, then the second, third or closest available server is dispatched instantly. Therefore, compared with cases with immobile servers, system workload is shared between servers better in the case of mobile servers. In addition, if there are no available servers, the customer enters a waiting line or is transferred to another ESS.

The term hypercube is taken from the space that describes the states of the servers. At any point of time, each server is free (i.e., 0) or busy (i.e., 1). Therefore, there are 2^N states in a system with N servers. A certain state of the system is specified by a list of free and busy servers (an array of 0s and 1s). For example, the state {011} corresponds to a 3-server system, in which server 1 is free and servers 2 and 3 are busy (reading from left to right). For $N = 3$, the state space can be shown by a cube (Figure 1), in which each vertex indicates one state of the system. For $N > 3$, the state space becomes a hypercube.

In an HQM, the performance measures of the system are obtained by calculating limiting probabilities, such as calls per hour, mean travel time, mean response time, mean workload, maximum workload imbalance, fraction of customers answered by primary servers, and fraction of customers answered by backup servers. To determine the limiting probabilities, 2^N balance equations should be solved. This is simply done by using the flow-balance criterion around the states of the system. Accordingly, in a steady state, the rate at which the system enters state i is equal to the rate at which the system leaves that state. By entering or exiting a state, a transition occurs. Actually, a transition arises when a server's state changes from busy to free, or contrariwise. Each transition takes place probabilistically over an edge of the hypercube. When a customer is served by a server, a downward transition happens. Thus, the rate of downward transition is equal to the service rate. An upward transition occurs once a free server is selected to be dispatched to a customer for service. The rate of an upward transition is determined by a set of server assignments and dispatching policies [2].

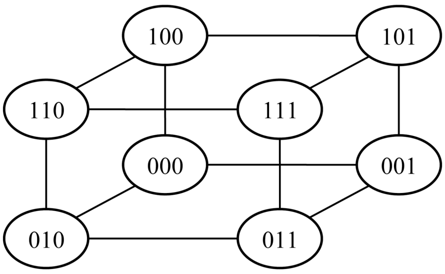

Figure 1. State space of a system with three servers.

To better understand the presented concepts, an example is provided here. Consider a simple network with three atoms connected by a one-way street (Figure 2). The distance matrix between these atoms is presented in Table 1.

It is assumed that the center of each atom is the location of a server, and a fixed-preference dispatch policy is in use. That is, when a call is received, a dispatcher assigns the first available server from a dispatch list ordered from the most preferred to the least preferred for that call. The preference matrix based on the shortest travel distance is shown in Table 2.

To define the hypercube state probabilities, 2^3 balance equations are written. If λ_i represents the arrival rate of customers from atom $i = (1, 2, 3)$ and μ_j indicates the service rate of server $j = (1, 2, 3)$, then

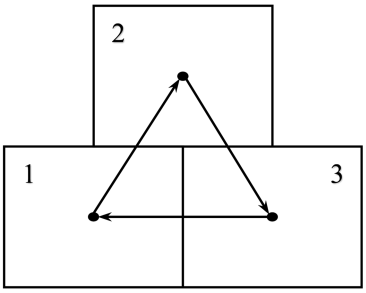

Figure 2. Network with three atoms [3].

Table 1. Travel distance matrix between atoms.

From/to	**1**	**2**	**3**
1	0	1	2
2	2	0	1
3	1	2	0

Table 2. Server dispatch preferences.

Atom	Server preference		
	1st	2nd	3rd
1	1	3	2
2	2	1	3
3	3	2	1

$\lambda = \lambda_1+\lambda_2+\lambda_3$ and $\mu = \mu_1+\mu_2+\mu_3$, where λ and μ are the total arrival and total service rates, respectively.

The following is an explanation of how to build the balance equation for a certain state, like $\{101\}$. The system leaves state $\{101\}$ if a customer arrives or server 1 or 3 completes its service; hence, the transition rate is $(\lambda + \mu_1 + \mu_3)P\{101\}$, where $P\{101\}$ shows the probability that the system is in state $\{101\}$. Moreover, the system enters this state in one of the following three ways:

i) From state $\{001\}$, if a customer arrives from atom 1 or 3 (in accordance with Table 2);

ii) From state $\{100\}$, if a customer arrives from atom 3;

iii) From state $\{111\}$ when the service of server 2 is completed.

The transition rate is $(\lambda_1 + \lambda_3)P\{001\} + \lambda_3P\{100\} + \mu_2P\{111\}$. Since the transition rates of the system out of and into a state are equal in a steady state, the balance equation of state $\{101\}$ is written by:

$$(\lambda + \mu_1 + \mu_3)P\{101\} = (\lambda_1 + \lambda_3)P\{001\} + \lambda_3P\{100\} + \mu_2P\{111\}. \tag{1}$$

The balance equations for other states can be written in a similar manner. To find out more, see Chiyoshi and Morabito [3], who presented a set of hypercube models with different assumptions and indicated steady-state equations and some practical specifications of each model.

In the next section, a brief introduction of the exact and approximate HQM is provided. Section 3 classifies the existing papers in the literature according to their assumptions and highlights the contributions of the model formulations and solution approaches. Section 4 provides the details of studies, which have incorporated the HQM in the location problem. Finally, this paper is ended by discussing future research directions. There are also some papers that cannot be classified in the following categories; however, they are very helpful and prepare basic concepts for the HQM. Potential applications of the HQM, how it works, when it is preferred to other models and the required resource are given in Chaiken [4]. Larson [5] presented a manual for users of the model. Larson [6] prepared a list of computer programs and provided information for users. Sacks [7] and Sacks [8] developed a software, named desktop hypercube and evaluated its performance in a case study. Larson [9] examined the performance of operations research in homeland security and introduced the HQM as an efficient tool in this area. Galvao and Morabito [1] reviewed probabilistic models for the design of emergency service systems. They also surveyed the extensions of these models, embedded into the HQM.

2. Exact and approximate HQMs

As the first HQM, Larson [2] analyzed a multi-server queuing system with distinguishable servers, which they support each other. He also developed a computationally efficient algorithm to evaluate the model analytically and calculate several performance measures. His model was designed for location and districting problems in urban emergency systems; districting is defined as partitioning an area into sub-areas (i.e., districts) according to its features. The following assumptions are considered in the original HQM [10,11]:

1. The district of a server is an area that is handled by the server, of course if it is available. Otherwise, customers in that area will be responded by a server out of the district. If all servers are busy, then the customer enters a queue or is served by another ESS. In addition, there may be more than one server in each district that shares the workload of that area;
2. Each service area is divided into sub-areas, called atoms. This classification can be done based on the census report, urban areas, and so on. Demand points are located at the center of each atom;
3. Demands of each atom are generated independently via a Poisson process with known rate λ_i;
4. There is at least one server in each atom, and exactly one server is dispatched to serve a customer;
5. Server assignment takes place according to a fixed-preference procedure. For each atom, there is an ordered list of preferred servers to dispatch. The dispatcher searches that list in order and sends the first available server. This list is usually obtained by geographical measures (e.g., travel times); however other criteria (e.g., allocating expert personnel) can be considered;
6. Service time follows an exponential distribution with a known rate. Brandeau and Larson [12] showed that service time generally includes travel time, on-scene time, and maybe some follow-up time (Figure 3).

For the convenience of the reader, a summary

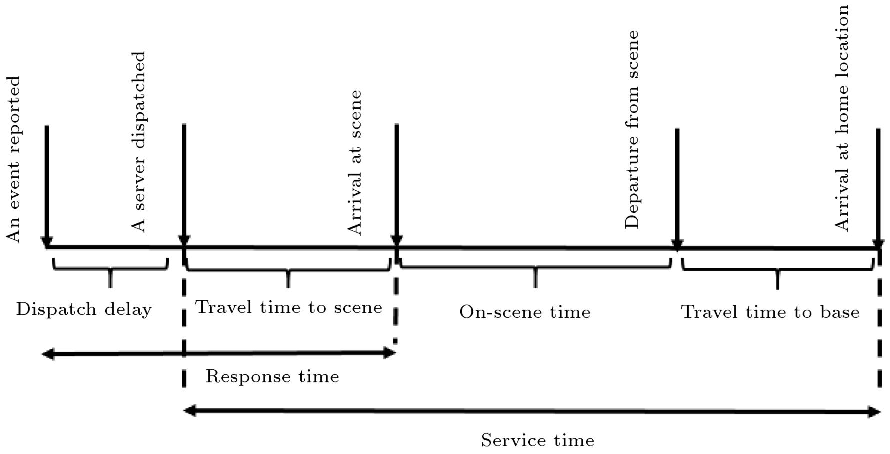

Figure 3. Divisions of service time [12].

Table 3. Notations used in this paper.

N	Number of servers ($j = 1, \cdots, N$)
M	Number of demand points ($m = 1, \cdots, M$)
C	Type of customer ($c = 1, \cdots, C$)
ρ	Average system busy probability
ρ_j	Busy probability of server j
ρ_{jm}	Fraction of dispatches in which server j is sent to atom m
λ	System-wide arrival rate
λ_c	Arrival rate of customers from node c
τ	System-wide mean service time
τ_{jc}	Expected service time for server j and a customer of type c
f_{jc}	Probability that a customer of type c is assigned to server j
h_m	Proportion of demand that is generated at node m
t_{ij}	Expected travel time between customer m and server j
W_m	Set of potential sites covering demand point m;
S_k	A state in which exactly k servers are busy
$P\{S_k\}$ or P_k	Probability that the system is in state S_k
$P\{S_0\}$ or P_0	Probability that all servers are idle
$P\{S_N\}$ or P_N	Probability that all servers are busy
a_{ck}	Index of the kth preferred server for customers of type c
B_j	The event that the jth selected server is busy
$F_j = B_j^c$	The event that the jth selected server is free
$P\{V_k\}$	Steady-state probability of the state corresponding to vertex V_k
E_{jm}	Set of states where server j is the nearest available server to customer m
C_N	Vertices of N-dimensional unit hypercube
d_{im}^-, d_{im}^+	Downward and upward Hamming distances between vertices V_i and V_m
$Q(N, \rho, j)$	Larson's correct factor

definitions of symbols frequently used in this paper is presented in Table 3.

2.1. Approximate HQM

Each system is evaluated by its performance measures. Larson [2] divided these measures according to the evaluated characteristics. As an example, the mean region-wide travel time (considering all types of calls), workload imbalance, and fractions of dispatches that are inter-district dispatches are suitable performance measures from a region-wide perspective; workload (fraction of time that server is busy), mean travel time and fraction of responses of each response unit that are inter-district can measure the performance of a response unit (server). It is possible to focus simultaneously on several performance measures as objectives while the other measures are maintained at an admissible level.

There are three main ways to evaluate the performance of a system [13]:

1. Exact approaches (e.g., HQM proposed by Larson [12]);
2. Discrete-event simulation;
3. Approximate approaches (e.g., Approximate Hypercube (AH) model proposed by Larson [14]).

The advantages of the approximate procedures in comparison with two other ones are that their computation time is low and is not influenced by the features of the system.

The AH model is a simple iterative procedure to estimate the performance measures of a system. In a system with N servers, the approximate model requires only N equations rather than 2^N, as it is necessary in the original model. Although, in practice, it has been approved that the solution of the AH model is good enough, simplified assumptions are used such as no cooperation between servers. Furthermore, in most approximate approaches presented so far, it is assumed that only one server is located at each base. Regarding these two assumptions, the state of each server (i.e., free or busy) is independent of the state of other servers; therefore, assuming non-cooperation can be considered almost equivalent to assuming independent servers. Hence, the HQM is reduced to an $M/M/N$ queuing system. In this system, if state S_k represents that k servers are busy exactly and $P\{S_k\}$ is the probability that the system is in state S_k, the steady-state probabilities are as follows:

$$P\{S_k\} \equiv P_k = \frac{N^k \rho^k P_0}{k!}, \qquad k = 1, 2, \cdots, N-1,$$

$$P\{S_N\} \equiv P_N = \frac{N^N \rho^N P_0}{N!(1-\rho)},$$

$$P\{S_0\} \equiv P_0 = \frac{1}{\sum\limits_{i=0}^{N-1} \frac{N^i \rho^i}{i!} + \frac{N^N \rho^N}{N!(1-\rho)}}, \tag{2}$$

where $\rho = \frac{\lambda}{N\mu} < 1$, where λ and μ indicate the demand and service rates, respectively. These results can be extended to include the $M/M/N/\infty$ queuing system.

In the AH model, the probability that the ith customer is served by the jth server in his/her preference list is equal to the probability of the first $j-1$ servers being busy and the jth server being the first free server. In Larson's study [14], servers are selected randomly until an idle server is found. If B_j indicates the event that the jth selected server is busy and $F_j = B_j^c$ indicates the event that the jth selected server is free, then $P\{B_1 B_2 \cdots B_j F_{j+1}\}$ shows the probability of the $(j+1)$th selected server being the first idle server. Selection of servers is done in a completely random process without replacement. Therefore, each server's probability of being busy is ρ, and the probability that the $(j+1)$th selected server is the first idle server is $\rho^j(1-\rho)$, if servers are independent. Thus, Larson [14] presented a factor, $Q(N, \rho, j)$, to correct the results for the case, in which the servers are dependent (or, in other words, there is cooperation between the servers).

$$P\{B_1 B_2 \cdots B_j F_{j+1}\} = P\{F_{j+1} | B_1 B_2 \cdots B_j S_k\}$$
$$P\{B_j | B_1 B_2 \cdots B_{j-1} S_k\} \cdots P\{B_1 | S_k\}$$
$$= Q(N, \rho, j) \rho^j (1-\rho), \tag{3}$$

$$Q(N, \rho, j) = \frac{P\{F_{j+1} | B_1 B_2 \cdots B_j\}}{(1-\rho)}$$
$$\frac{P\{B_j | B_1 B_2 \cdots B_{j-1}\}}{\rho} \cdots \frac{P\{B_1\}}{\rho}$$
$$= \frac{\sum\limits_{k=j}^{N-1} \frac{(N-j-1)!(N-k)}{(k-j)!} \left(\frac{N^k}{N!}\right) \rho^{k-j}}{(1-\rho) \sum\limits_{i=0}^{N-1} \frac{N^i}{i!} \rho^i + \frac{N^N \rho^N}{N!}}. \tag{4}$$

Jarvis [15] extended the approximation procedure in [14] and presented an algorithm for loss systems (zero-line capacity). In his procedure, service time distribution is dependent on the type of server and customer. Different types of customers have different demand rates, and service rate is different for each pair of server and customer. He also estimated server's workload more accurately than Larson [14]. Furthermore, he could show that although the shape of service time distribution was not completely ineffective in the results of the HQM, its influence was very small [15].

The details of the Jarvis algorithm, which approximates the busy probabilities of each server, are presented below [16].

In the initial step of the algorithm, ρ_j is calculated by assuming that every customer is assigned to its first preferred server. Therefore, there is no cooperation between servers.

$$\rho_j = \sum_{c:\alpha_{c1}=i} \lambda_c \tau_{jc}, \tag{5}$$

$$\tau = \sum_{c=1}^{C} \frac{\lambda_c}{\lambda} \tau_{\alpha_{c1},c}. \tag{6}$$

Step 1: Calculate the correction factor using Eq. (4).

Step 2: Calculate an approximation of the server's workload for $j = 1, \cdots, N$:

$$\rho_j(\text{new}) = \frac{V_i}{V_i + 1}, \tag{7}$$

$$V_i = \sum_{k=1}^{N} \sum_{c:\alpha_{ck}=i} \lambda_c \tau_{ic} Q(N, \rho, k-1) \prod_{l=1}^{k-1} \rho_{\alpha_{cl}}. \tag{8}$$

Step 3: Stop if maximum change in ρ_j is less than convergence criterion.

Step 4: Calculate the following equation:

$$P_N = 1 - \frac{\sum\limits_{i=1}^{N} \rho_j}{N\rho}, \tag{9}$$

$$f_{jc} = Q(N, \rho, k-1)(1 - \rho_j) \prod_{l=1}^{k-1} \rho_{\alpha_{cl}}, \tag{10}$$

$$\tau = \sum_{c=1}^{C} \frac{\lambda_c}{\lambda} \frac{\sum\limits_{i=1}^{N} \tau_{jc} f_{jc}}{1 - P_N}. \tag{11}$$

Step 5: Return to Step 1.

3. Classification of the HQM

The HQM can be classified from different perspectives. From a general point of view, HQMs are categorized based on the dispatching policy, backup strategy, and server type. Dispatching strategies are divided into two general categories, namely single and multiple. In a single dispatching strategy, it is assumed that only one server is needed to respond to the demand of a customer. In a multiple dispatching strategy, customers require two or more servers simultaneously. As an example, in a severe car accident, several low-level ambulances are sent for transportation, along with one or two high-level ambulances to provide more professional medical services; in the case of police patrol, two police cars each with one or two officers are usually sent.

Backup strategies are divided into two general categories, namely total and partial. In a total backup strategy, a customer is lost or enters the queue only if there are no idle servers. That is, as long as there is an idle server, customers will not queue up. In a partial backup strategy, only some servers can respond to each customer. For instance, a server may be able only to serve customers who are located at a certain distance from that server. Therefore, in these models, on the arrival of a customer, if its backup servers are busy, the customer is lost or enters a queue, even if there is a free server.

In this survey, the term 'homogeneity' stands for servers with the same rate. A service rate depends on many factors. For example, the service rate for trained servers is usually more than that for newcomers, or, in some cases, the location of customers and servers influences the travel time, which is a part of service time. In addition, a service rate may be different for various types of activities; as an example, in police patrol systems, two kinds of activities are defined, namely dispatching to Call For Service (CFS) and Patrol Initiated Activity (PIA). Officers in police patrol cars control their patrol areas to improve public safety, and they may check buildings, cars or people; these actions are called PIA. Such activities can also be defined in medical systems. Suppose that a patient visits an ambulance base for receiving service. Although, in this case, the server has not been sent by a dispatcher, it would be busy with considerable time spent. Therefore, these activities have different service rates and must be taken into account in calculating the performance measures.

Based on the aforementioned explanations, studies in the field of HQMs can be divided into eight categories. The point that should be noted is that, in cases where servers are not supposed to support each other, they should be treated as non-homogeneous servers; hence, we just review six distinguishable categories below.

3.1. Single dispatch, total backup, and homogeneous servers

The assumptions considered in this part are similar to those of the original model. After Larson [2], Larson and Franck [10] evaluated the performance measures of an emergency response system where the dispatcher has access to Automatic Vehicle Location (AVL) systems. AVL systems estimate the current location of servers in the service area and help the dispatcher forward the closest available server to each customer. Unlike the studies with a fixed-preference dispatching strategy, in this study, the upward transition rate depends on the real-time locations of idle servers. Therefore, to determine the matrix of upward transition rates, a recursive method is used. According to this method, geographical locations of customers are first fixed, and then the hypercube vertices are visited in a unit-step procedure. This matrix is completed when all vertices of the hypercube are visited once for each geographical area.

Chelst and Jarvis [17] proposed an extension of the HQM where the probability distribution of travel time is calculated, in addition to its average. Larson and Rich [18] investigated the relationship between travel times and dispatching policy in a police department. They found that travel times would not increase remarkably as the service area of each server increased.

Souza et al. [19] proposed a modified HQM in the context of emergency systems and considered the priority of customers in the queue based on the degree of severity. In their work, high-level customers are those whose lives are at risk and need more advanced equipment and highly specialized medical team, as compared with the low-level customers. They assumed a non-preemptive priority discipline when a server becomes idle; it serves a low-level customer in the queue if there are no high-level customers. In this study, a layering procedure [20,21] is used to take into account different classes of customers. In this procedure, the total service area is divided into different sub-areas, called atoms, and each atom is distributed into sub-atoms each of which is an independent source of customers from one of the priority classes. There is an example of such systems with infinite queue capacity. Consider a queue system with three priority classes, a, b, and c, where a and c represent the highest and lowest priority classes, respectively. The dispatching matrix is shown in Table 4, in which each atom has three layers. If there are two customers in the queue, then all possible states of the queue are: $\{aa\}$, $\{ab\}$, $\{ac\}$, $\{bb\}$, $\{bc\}$ and $\{cc\}$. If each class r ($r = a, b$ and c) of customers arrives according to the Poisson process with rate λ_r, and service time is distributed exponentially with rate μ, then the transitions into and out of the queue state $\{ab\}$ are shown in Figure 4. It is obvious that according to the priority queuing discipline, when a server becomes free, the only transition is $\{ab\} \rightarrow \{b\}$, and transition $\{ab\} \rightarrow \{a\}$ is not allowed. The balance equation around the state $\{ab\}$ is built as usual by setting the transition rate into a state equal to the transition rate out of that state.

Table 4. Example of the dispatching matrix.

Atom	Layer	Dispatch preference	
		1st	2nd
1	a	1	2
	b	2	1
	c	1	2
2	a	2	1
	b	1	2
	c	2	1

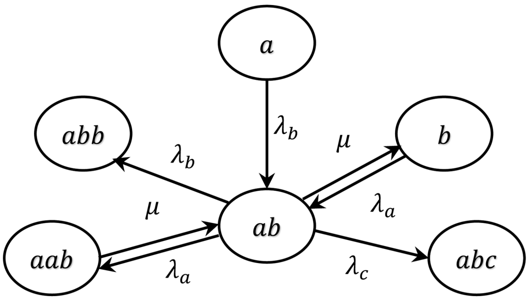

Figure 4. Transition into and out of state $\{ab\}$ [19].

In addition to the studies presented above, some studies have integrated the HQM and location models [22,23]. The performance of emergency systems is associated with the location of servers and their allocation to the customers. Therefore, optimizing these two problems can improve some performance measures (e.g., mean travel time) simultaneously.

Daskin [24] formulated the maximal expected coverage location problem (MEXCLP), where servers are located optimally to maximize the expected coverage of demand in a situation, in which some servers may be unavailable. He recognized the busy probability of each server ρ, which can be calculated by an Erlang loss system equation, $\rho = \lambda/N\mu$. Therefore, if each demand point is covered by n servers, then the probability that a demand is covered by at least one server is equal to $1 - \rho^n$. The formulation of this problem is as follows. In this model, some simplified assumptions (e.g., independent servers and the same busy probability for each server) are considered.

$$\max \quad \sum_{i=1}^{M}\sum_{j=1}^{N}(1-\rho)\rho^{j-1}h_i y_{ij}, \tag{12}$$

$$\text{s.t.} \quad \sum_{j=1}^{N} y_{ij} - \sum_{i=1}^{M} a_{ij}x_j \leq 0, \qquad \forall\, i, \tag{13}$$

$$\sum_{j=1}^{N} x_j \leq N, \tag{14}$$

$$x_j = 0, 1, \cdots, N, \qquad \forall\, j;$$

$$y_{ij} = 0, 1, \qquad \forall\, i, j, \tag{15}$$

where N is the maximum number of facilities, M is the number of demand points, and x_j is the number of servers in facility j. In addition, we have:

$$y_{ij} = \begin{cases} 1 & \text{if node } i \text{ is covered by at least } j \text{ facilities} \\ 0 & \text{otherwise} \end{cases}$$

As mentioned before, the objective function (12) maximizes the expected number of demands that are covered. Constraint (13) calculates how many times demand point j is covered. Constraint (14) limits the maximum number of facilities that can be deployed, and Constraint (15) shows that more than one server can be assigned to each facility.

Batta et al. [25] proposed an adjusted MEXCLP (AMEXCLP) by relaxing three basic assumptions of the MEXCLP, independent servers, the same busy probability for servers, and independence between busy probability of servers and their locations. The formulations of the AMEXCLP and MEXCLP are similar, except for the objective function. The objective function of the AMEXCLP, which relaxes the independence assumption using the Larson's correction factor, is as follows:

$$\max \quad \sum_{j=1}^{N}\sum_{i=1}^{M} Q(N,\rho,j-1)(1-\rho)\rho^{j-1}h_i y_{ij}, \qquad (16)$$

where y_{ij} is one if node i is covered by at least j servers; otherwise, it is zero, and $Q(N,\rho,j-1)$ is computed by Eq. (4). Batta et al. [25] also integrated the HQM with a heuristic optimization procedure to find a set of locations for servers, maximizing the expected coverage. They concluded that there is a conflict between the results of the AMEXCLP and those of a hypercube optimization procedure. Furthermore, Chiyoshi et al. [26] showed that the MEXCLP and its aggregate version (i.e., AMEXCLP) with the HQM were not comparable, because the structures of their objective functions were different and the MEXCLP could not consider queued customers. Furthermore, for queued customers, the waiting time plus travel time may exceed the critical covering time. In an HQM with infinite line capacity, these customers are served and added to the system's workload; however, in the MEXCLP and AMEXCLP, they are not even covered. They investigated these two points in the study carried out by Batta et al. [25]. Galvao et al. [27] compared the MEXCLP with the Maximum Availability Location Problem (MALP). The MALP was proposed by ReVelle and Hogan [28] which aimed at maximizing the number of customers that can be covered at a target response time with reliability of α. Both the MEXCLP and MALP are probabilistic extensions of the Maximum Covering Location Problem (MCLP); actually, they are two different perspectives of the same concept. They also proposed an Extension of the MALP (EMALP), in which each server has a different busy probability. Furthermore, in the EMALP, the Larson's correction factor is used to account for dependency between servers. Finally, Simulated Annealing (SA) solves these two extended models. Chiyoshi et al. [26] developed Tabu Search (TS) [29] for the EMALP and compared the results of this algorithm against those of the SA algorithm developed by Galvao et al. [27]. They showed that the solutions of SA outperformed TS for small networks in terms of quality, while TS performed better than SA for larger networks generated randomly. The hypercube queuing model was used to calculate the server's busy fractions.

Goldberg et al. [30] developed a model to locate emergency facilities. The goal of their model was to maximize the expected number of customers who were responded to within eight minutes (success rate). They used the Jarvis' procedure to estimate server's utilization rate. They also proposed a model toestimate travel time distribution between each pair of server and customer in a case study.

McLay and Mayorga [31] proposed a location model to maximize two performance measures:

1. The expected number of customers who have survived;
2. The expected number of customers who are responded to within the specified time threshold.

These two performance measures are functions of response time, which is affected by the distance between servers and customers. These measures are evaluated only for customers whose lives are threatened, and the Larson's approximation algorithm is used to estimate them. They showed that optimization of patient's survival rate corresponds to how the Response Time Threshold (RTT) is chosen. Toro-Diaz et al. [32] proposed a non-linear mixed-integer optimization model to find the location of ambulances (see Section 3). They found that the use of closest dispatching policy enabled the model to minimize the response time and maximize the coverage.

Usually, in an EMS, servers are located at a fixed base. As population grows, demands for ambulances increase, hence requiring ambulance bases. Demands for ambulances are time dependent and may change weekly, daily, and even hourly; thus, building permanent bases to cover variable demands is costly. In response to demand fluctuation, redeployment strategies are used to change the location of ambulances dynamically [33,34]. There are two types of redeployment strategies:

1. **Multi-period:** In this strategy, the volumes of demands are predicted for different sectors of the service area and for different time periods; then, ambulances are redeployed to face demand fluctuation;
2. **Real-time:** In this strategy, when one or more ambulances are dispatched, the other available ambulances are redeployed to guarantee the desired coverage.

Sudtachat et al. [35] proposed a dynamic relocation strategy by using a nested compliance table. This table indicates where ambulances should be located when there are a certain number of ambulances available. Actually, ambulance stations are specified as a function of the state of the system by a compliance table. A nested compliance table restricts the number of relocations that can occur simultaneously. They extended an integer programming model to determine an optimal nested compliance table strategy and maximize the expected coverage. The relocation and approximation of steady-state probabilities are input parameters of this model. Finally, they compared this dynamic strategy with a non-relocation model (AMEXCLP) proposed by Batta et al. [25], and showed that the expected coverage provided by their model is more suitable.

3.2. Single dispatch, total backup, and non-homogeneous servers

Systems with non-homogeneous servers can be found in many real-world cases. For example, in an EMS, some ambulances only provide basic support; however, some of them can provide advanced support. Thus, when these two types of ambulances share a workload, the service rate becomes the average service rate of basic and advanced ambulances. On the other hand, two systems with similar servers in terms of vehicle, personnel, and equipment may have different service rates based on their locations. For instance, the mean service time may change due to the travel time, which is a function of servers' locations. Halpern [36] investigated the effects of dependency on the service time, customer location, and dispatch units in a simple two-server, two-customer system.

Larson and McKnew [37] proposed the HQM and AH model for police patrol, where officers can be in one of three states, i.e., free, busy on PIA, and busy assigned to a CFS. Therefore, the total number of states in this study is 3^N. They used the Larson's approximation procedure to estimate the performance measures. Further to the previous research, McKnew [38] used a Modified Center of Mass (MCM) dispatching policy at a police department with 3^N states. In this policy, the total service area is divided into several sub-areas, and police cars are located at the center of their mass. Each sub-area is distributed into atoms, and customers are located at their center of them. Upon arrival of a customer, the closest available car is assigned to the respective sub-area even if there is a closer car from other sub-areas. If all cars in a sub-area are busy, then the closest car is assigned to other sub-areas.

A common assumption in the hypercube model is to use a fixed-preference list to dispatch servers. This means that upon receiving a call, the first available server is assigned according to this list. Thus, if there are servers with the same priority for a specific customer, a "tie" occurs. In AH models, it is assumed that there is only one server, which is preferred to dispatch; however, when multiple servers are located at one station, the tie occurs in the dispatch preferences with high probability. Modeling of such a situation by taking an arbitrary fixed-preference order leads to an imbalance in the workload among tied servers, because tied servers, placed in the more preferred positions, receive greater workload than those tied servers that are placed in the less preferred positions. Burwell [39] and Burwell et al. [40] proposed an "Internal Stacking" method to handle this situation. In their study, v_{jm} indicates the number of servers tied with server j, including server j, for the kth preference position of customer m. It is also assumed that these servers are positioned from k to $(k-1+v_{jm})$ in the dispatch list of customer m. The set of these servers is given by H_{jm}:

$$H_{jm}=\left(\bigcup_{l=k}^{k+v_{jm}-1}\{a_{ml}\}\right)-1. \tag{17}$$

Each time, if the first $(k-1)$ servers are busy, then a server is selected randomly from H_{jm}. The workload of server j must be calculated by conditioning the number of tied servers, selected before server j.

Brandeau and Larson [12] studied the effects of variable service times in the Larson's AH model. They also considered the mean service time calibration feature in this model and offered an algorithm to estimate the travel time effectively.

Takeda et al. [21] investigated ambulance decentralization in a case study and indicated that decentralization of ambulances can have positive impact on system performance measures. Budge et al. [13] proposed an approximation algorithm based on the Jarvis' algorithm, in which more than one server could be assigned to each station; therefore, they computed a station's (instead of server) busy probability. In addition, in this algorithm, a set of correction factors was formulated based on random sampling of stations. They assumed that N servers were distributed among j stations, with n_j servers at station j, and for all j, $n_j \geq 1$. They also defined P_0 and P_N corresponding to the probability of the system being idle (all servers are available) and the probability of all servers being busy, respectively. This algorithm initiates by calculating:

b_{ki} The kth preferred station for node i,

$n_{ki}=n_{b_{ki}}$ The number of servers at the kth preferred station for node i,

$$z_{ki}=n_{1i}+n_{2i}+\cdots+n_{ki},$$

$$\tau_{ki}=\tau_{b_{ki}i}.$$

In this algorithm, r_j corresponds to the busy fraction of each open station j, ρ is the expected workload per server, and $r=\rho(1-P_N)$ is the expected server utilization.

$$r_j=\frac{1}{n_j}\sum_{i=1}^{I}\lambda_i f_{ij}\tau_{ij}. \tag{18}$$

The station-specific correction factors can be expressed by:

$$Q_i(\{n_{ki}\},\rho,k) = \frac{P_0 \sum\limits_{s=z_{(k-1)j}}^{N-1} \frac{(\rho N)^s}{s!} \left(\prod\limits_{u=0}^{z_{(k-1)i}-1} \frac{s-u}{N-u} - \prod\limits_{u=0}^{z_{(k)i}-1} \frac{s-u}{N-u} \right)}{r^{z(k-1)}(1-r^{n_{ki}})}, \tag{19}$$

where f_{ij} is the multi-server counterpart of Eq. (8) and is calculated by:

$$f_{ij} \approx Q_i(\{n_{ki}\},\rho,k) \prod_{l=1}^{k-1} r_{li}^{n_{li}} \left(1 - r_j^{n_j}\right). \tag{20}$$

In Eq. (20), distribution of servers between stations affects the correction factor, because the probability that a server from station j is dispatched to a customer from node i not only depends on the number of stations that are preferred (by node i) to station j, but also on the number of servers at those stations.

Next, by assuming that the system operates as an $M/M/N/N$ queuing system, the busy fractions and the system's wide average service time are calculated by (superscripts are used as iteration counters):

$$\tau^0 = \frac{1}{\lambda N} \sum_{j=1}^{J} n_j \sum_{i=1}^{I} \lambda_i \tau_{ij}, \tag{21}$$

$$r_j^0 = r^0 = \lambda \tau^0 \frac{1 - P_N^0}{N}, \tag{22}$$

where P_N^0 is calculated using Erlang's loss formula. Set the iteration counter, h, to one and repeat the following steps:

Step 1: Use τ^{h-1}, λ, and N to calculate P_0^h and P_N^h.

Step 2: Calculate V_j^h for all j by using Eqs. (20) and (23):

$$V_j^h = \sum_{i=1}^{I} \lambda_i \tau_{ij} Q_i\left(\{n_{ki}\}, \rho^{h-1}, \alpha_{ij}\right) \prod_{l=1}^{\alpha_{ij}-1} \left(r_{li}^{h-1 \text{ or } h}\right)^{n_{li}}, \tag{23}$$

where $r_{li}^{h-1 \text{ or } h}$ is always the most recently computed station utilization (i.e., r_{li}^h) if it has been computed, and otherwise r_{li}^{h-1}. Then, update the station-specific busy fractions using Eq. (24) if $r^{h-1} \leq 0.5$ and using Eq. (25), otherwise:

$$r_j^h = \frac{V_j^h}{n_j + \left(r_j^{h-1}\right)^{n_j-1} V_j^h}, \tag{24}$$

$$r_j^h = \left(\frac{V_j^h}{V_j^h + n_j / \left(r_j^{h-1}\right)^{n_j-1}} \right)^{1/n_j}. \tag{25}$$

Step 3: Calculate f_{ij}^h and normalize these probabilities using $f_{ij}^h \leftarrow f_{ij}^h(1-P_N^h)/\sum\limits_{j=1}^{J} f_{ij}^h$; then, calculate τ^h, ρ^h, and r^h.

$$f_{ij}^h \approx Q_i\left(\{n_{ki}\},\rho,k\right) \prod_{l=1}^{k-1} r_{li}^{n_{li}} \left(1 - r_j^{n_j}\right), \tag{26}$$

$$\tau^h = \frac{1}{\lambda(1-P_N)} \sum_{i=1}^{I} \lambda_i \sum_{j=1}^{J} f_{ij}^h \tau_{ij}, \tag{27}$$

$$\rho^h = \frac{\lambda \tau^h}{N}, \tag{28}$$

$$r^h = \frac{1}{N} \sum_{j=1}^{J} n_j r_j^h. \tag{29}$$

Step 4: If $|r_j^h - r_j^{h-1}| < \varepsilon$ for all j, stop. Otherwise, set $h = h + 1$.

Budge et al. [41] used this algorithm to find the relationship between travel time and distance. They concluded that a logarithmic transformation made symmetric travel-time distribution. Additionally, Toro-Díaz et al. [42] used this algorithm to extend the work given by Toro-Díaz et al. [32] and presented a multi-objective location model to make a balance between efficiency and fairness, where more than one server can be assigned to each station. In their work, the purpose of fairness is to make the same mean response times and, also, the same server's workload. Ansari et al. [43] used this approximation algorithm to estimate the correction factors, the average server workload, and the individual server workload and treated them as constants in an MILP model. This model maximized the number of high-priority calls that can be covered within a time threshold (i.e., the coverage level) and balanced server workload by determining the location of ambulances and dispatching policy simultaneously. They proposed an iterative algorithm to solve a real-world example and concluded that the server workload maintained equivalence by a small reduction in coverage.

Boyaci and Geroliminis [44] proposed two extensions for the AH model, in which more than one server can be assigned to each atom. In the first extended AH model, it is assumed that the service rate is equal for the intra- and inter-district customers; therefore, each server has two states: free and busy. In addition, in this model, each number in a system state corresponds to the number of busy servers in the corresponding atom. For example, state {302} stands for a state where 3, 0, and 2 servers are busy in the first, second, and third atoms, respectively. Thus, if n_i shows the number of servers assigned to atom i and M indicates the total number of atoms, then there are $(n_1+1)(n_2+1)\cdots(n_M+1)$ states. As usual, the steady-state probabilities in this system are computed by flow-balance equations. As an instance, transition equation for state $\{302\}$ is written as (30), and Figure 5 shows this transition network.

$$P\{302\}(\lambda_1+\lambda_2+\lambda_3+3\mu_1+2\mu_3)=\lambda_3 P\{301\} + \lambda_1 P\{202\}+3\mu_3 P\{303\}+\mu_2 P\{312\} + 4\mu_1 P\{402\}. \tag{30}$$

In the second extended AH model, the service rate is different for inter μ_j and intra μ'_j arrivals; thus, each server has three states: free, busy serving an intra-district customer, and busy serving an inter-district customer. A new definition for the states of the system is proposed, in which the state of each server is shown by two numbers. For example, in a system with two servers in each atom, state $\begin{Bmatrix}1 & 1\\ 1 & 0\end{Bmatrix}$ indicates a situation where two servers in the first atom are busy (according to the first row), and one of them serves an intra-district customer and the other one serves an inter-district customer. Similarly, one server of the second atom is busy and serves an intra-district customer, and the other one is free. Therefore, this system has $\prod_j \binom{n_j+2}{2}$ states. The transition rate of state $\begin{Bmatrix}1 & 1\\ 1 & 0\end{Bmatrix}$ is as Eq. (31), and Figure 6 shows this transition network:

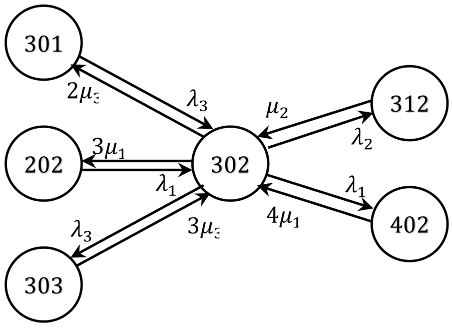

Figure 5. Transition network for state $\{302\}$ with equal inter- and intra-district service rates.

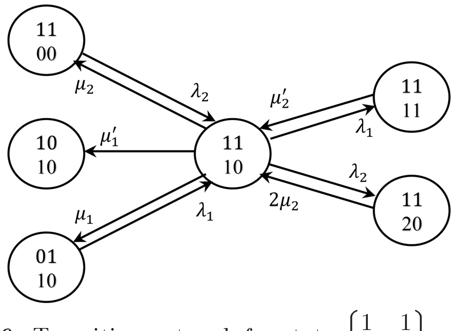

Figure 6. Transition network for state $\begin{Bmatrix}1 & 1\\ 1 & 0\end{Bmatrix}$ with different inter- and intra-district service rates.

$$P_{\begin{Bmatrix}1 & 1\\ 1 & 0\end{Bmatrix}}(\lambda_1+\lambda_2+\mu_1+\mu'_1+\mu_2)=\lambda_1 P_{\begin{Bmatrix}0 & 1\\ 1 & 0\end{Bmatrix}} + \lambda_2 P_{\begin{Bmatrix}1 & 1\\ 0 & 0\end{Bmatrix}} + \mu'_2 P_{\begin{Bmatrix}1 & 1\\ 1 & 1\end{Bmatrix}} + 2\mu_2 P_{\begin{Bmatrix}1 & 1\\ 2 & 0\end{Bmatrix}}. \tag{31}$$

Finally, they showed that Monte Carlo sampling is applicable for the HQM and can represent its features and solve its steady-state probabilities. Boyaci and Geroliminis [45] proposed a Mixed-Hypercube Queuing Algorithm (MHQA) in an emergency system, which can generally be presented in three steps. In the first step, the total service area is divided into the sub-areas iteratively until the size of each sub-problem becomes solvable. In the second step, these sub-problems are solved by considering three states for each server (i.e., the second extended AH model in Boyaci and Geroliminis [44]). In the final step, sub-areas are merged by using an approximate hypercube model, in which some servers have two states and the others have three states, because servers located near the borders between two merged areas can provide service to both areas at the same rate. This algorithm computes the lost rate for the main service area. Boyaci and Geroliminis [46] proposed a partitioning algorithm to obtain highly accurate results in the MHQA.

Baptista and Oliveira [47] extended the AH model in which the customer arrival process was not stationary. However, service time, response time, and preference list were independent of time periods. It is obvious that the AH model can be applied only in periods, in which the arrival rate is stationary; therefore, the average unit workloads are computed by:

$$\rho_q=\frac{\sum\limits_{q\in J} t_q\rho_j^q}{\sum\limits_{q\in J} t_q}, \tag{32}$$

where:

T Number of time periods;

r_t Time period, t $(t=1,\cdots,T)$;

q The set of periods, r_t, where the arrival rate is stationary, i.e., $q=r_1,r_2,\cdots,r_{t_q}$ and $q\in J$;

j The set of q different stationary periods;

t_q The number of time periods, r_t, in set q;

ρ_j^q Workload of server j in period q.

They estimated a set of the system performance measures using the presented hypercube model and used a simulation model in a case study to assess the validity of this AH model based on four dispatching rules:

1. *The nearest neighbor rule:* Servers are ranked based on their distances to the customers;
2. *Less occupied preference rule:* Servers are ranked based on the number of times that each server is assigned;
3. *Area preferred server rule:* Only a server with the highest priority can be assigned to serve a customer, and if it is not available, a server is assigned to another ESS;
4. *Area with two most preferred servers:* Only two servers with the highest priority can be assigned to respond to a customer in order, and if neither of them is available, a server is assigned to another ESS.

Iannoni et al. [48] suggested an HQM to analyze the EMS, in which customers had different priorities. In this study, low-level customers were kept waiting until the number of idle servers reached the threshold number (i.e., cut-off level) to increase the probability of serving the higher-level customers immediately, upon arrival. They calculated the performance measures for this cut-off HQM.

As shown in Section 2.1, the studies that proposed location models in a hypercube framework are as follows. Goldberg and Paz [49] modified the model presented in Goldberg et al. [30] by considering the FIFO queue discipline, customer classification, and allocation of multiple servers at each base location. They tested the applicability of the model in more real test problems and proposed a heuristic algorithm. Zhu and McKnew [50] proposed a Workload Balancing Allocation Model (WBAM) to deploy a number of ambulances and balance workload between servers. This model uses a goal programming approach to address this aim and calculates server workload with an AH model developed by Burwell [39]. Lei et al. [51] formulated a four-objective model for a districting-routing problem under dynamic and stochastic conditions. They solved this model by a two-stage stochastic programming approach and an enhanced multi-objective evolutionary algorithm.

Geroliminis et al. [52] proposed a hybrid queuing location model to minimize the mean response time and meet the minimum coverage level. Actually, they extended the MCLP for locating servers and used the HQM for districting and dispatching purposes. In this study, the server's response time depends on the customer's location, and the service rate for intra-district customers located in the server's region is lower than that for customers out of that region (i.e., inter-district). Later, Geroliminis et al. [53] extended the previous model for locating emergency vehicles in urban networks with many servers, subject to hypercube flow-balancing equations. This model will be described in Section 4. Geroliminis et al. [54] used a GA combined with the hypercube model to solve this model in a two-stage approach. In the first stage, the overall service area is districted into subareas, and a number of servers are allocated to each subarea. In the second stage, the optimal location of servers is determined in their subarea. In both stages, the AH model is used to evaluate the fitness function. The results of the model application indicated that this model was suitable as an optimization tool, particularly when many servers must be located.

Erkut et al. [55] extended ten existing covering models for emergency systems by considering the survival function, which maximized the expected number of patients who survived cardiac arrest. They used the Jarvis' algorithm to evaluate this function. Ingolfsson et al. [56] designed a location model to minimize the number of ambulances and to satisfy the minimum threshold of the service level. They determined the service level by the number of customers responded to at a time interval. They also considered random pre-trip delays, in addition to random travel times. They used the approximation procedure proposed by Budge et al. [13] to evaluate the server's busy fraction.

McLay [57] proposed an MEXCLP with two types of servers and multiple types of customers (MEXCLP2) to determine the locations of ambulances optimally. The goal of their model was to maximize the expected number of customers in life-threatening situations covered within a specified time. To calculate server's busy probabilities, they extended an approximation algorithm based on the Jarvis' algorithm for a case with an infinite queue.

Rajagopalan and Saydam [58] formulated a Minimum Expected Response Location Problem (MERLP) to determine the locations of ambulances in order to minimize the expected response distances and meet minimum coverage requirements. They incorporated the concept of coverage in their model by using the Daskin's expected coverage [24], as presented in Section 2-1, and the Marianov and Revelle's available coverage [59], in which only those customers were incorporated in the coverage statistics and covered with predetermined reliability. They also incorporated server busy probabilities, computed by the Jarvis' algorithm, in the expected and available coverage statistics. By applying the MERLP to a case study, they compared this model with the MEXCLP [24], and showed that, in MERLP, the response time was faster; hence, more lives could be saved.

3.3. Single dispatch, partial backup, and non-homogeneous servers

When assuming that there is partial cooperation between servers, some servers cannot respond to some customers for some reasons, such as the location

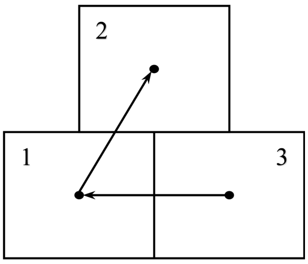

Figure 7. Network with three atoms.

Table 5. Server dispatch preferences.

Atom	**Server preference**		
	1st	**2nd**	**3rd**
1	1	3	—
2	2	1	3
3	3	—	—

of the customer or the type of the customer's demand. In these systems, on arrival of a customer, if the backup servers are busy, the customer is lost or enters a queue even if there are other free servers.

To better understand the presented concepts, the reader may consider the example provided in Section 1. Assume that there is no way between atoms 2 and 3 (Figure 7). Therefore, the preference matrix based on the shortest travel distance is changed, as shown in Table 5.

Now, the balance equation for state {011} is as follows. The system leaves state {011} if a customer arrives from atom 1 or 2, or server 2 or 3 completes its service; hence, the transition rate is $(\lambda_1 + \lambda_2 + \mu_2 + \mu_3)P\{011\}$. Moreover, the system enters this state in one of the following three ways (in accordance with Table 4):

i) From state {001} if a customer arrives from atom 2 (it is noteworthy that, in this state, if a customer arrives from atom 3, it has been lost or served by another system, although servers 1 and 2 are idle);

ii) From state {010} if a customer arrives from atom 3;

iii) From state {111} when the service of server 1 is completed.

The transition rate is $(\lambda_2)P\{001\} + \lambda_3 P\{010\} + \mu_1 P\{111\}$. The balance equation of state {011} is written by:

$$(\lambda_1 + \lambda_2 + \mu_2 + \mu_3)P\{011\} = (\lambda_2)P\{001\} + \lambda_3 P\{010\} + \mu_1 P\{111\}. \tag{33}$$

Mendonca and Morabito [60] investigated ambulance deployment on a highway, connecting the cities of Sao Paulo and Rio de Janeiro in Brazil. In this study, only a part of the highway was analyzed, covered by the Anjos do Asfalto's emergency system. This EMS had six ambulance bases (i.e., 2^6 states) along the highway, and one ambulance was stationed at each base. The central dispatcher was located in Rio de Janeiro and recorded all the movements of ambulances, even for fueling. When an emergency call is received, the nearest ambulance is dispatched to the place of incident, and if it is busy, the second nearest ambulance is dispatched. If this ambulance is busy too, then the customer is lost and transferred to another EMS. They used a hypercube model to evaluate the performance measures of this system, such as mean response time and workload of ambulances. They showed that the ambulance workload became more balanced only by changing the sizes of the atoms.

Atkinson et al. [61] proposed one exact and two heuristic methods to estimate loss probabilities and ambulance utilization rate for the EMS studied by Mendonca and Morabito [62]. They also showed the accuracy of the proposed heuristic methods with a numerical example. Atkinson et al. [63] extended these two heuristic methods for a system with 3^n states: free, busy serving a first-preference customer, and busy serving a second-preference customer. Iannoni et al. [64] presented several greedy heuristic algorithms to optimize ambulance location and dispatch policies for the EMS presented in Mendonca and Morabito [62]. The hypercube approximation algorithm proposed by Atkinson et al. [61] was embedded in each optimization procedure to solve large-scale problems fast with acceptable precision.

Morabito et al. [65] investigated the effects of considering homogeneous against non-homogeneous servers in calculating the performance measures with HQM. They concluded that even when the level of non-homogeneity was not significant, it led to different results to consider it in performance measures. Kim and Lee [66] used the HQM to compute steady-state probabilities in a Probabilistic Location Set Covering Problem (PLSCP) to satisfy the reliability requirements. In the PLSCP presented by ReVelle and Hogan [67], the total number of ambulances was minimized while the location of ambulances was determined such that the number of ambulances covering each node would be higher than minimum requirement. They also suggested two iterative optimization algorithm based on the HQM and simulation, and found that the performance of these two algorithms was almost equivalent.

3.4. Multiple dispatch

As noted above, in the original model, only one server is dispatched to serve a customer; however, in real-world situations (e.g., large fire or severe accidents), more than one server is usually dispatched. This section presents the studies with multiple dispatch assumption in hypercube models, in which some customers require simultaneous dispatching of two or more servers.

Now, assume that, in the example presented in Section 1, customers from each atom, j, can be of two types. Customers of type 1 require services by a server with arrival rate $\lambda_j^{[1]}$, and customers of type 2 require the simultaneous service by two servers with arrival rate $\lambda_j^{[2]}$. The total arrival rate of the system is as follows:

$$\lambda = \lambda_1^{[1]} + \lambda_1^{[2]} + \lambda_2^{[1]} + \lambda_2^{[2]} + \lambda_3^{[1]} + \lambda_3^{[2]}. \tag{34}$$

Suppose that each server can be dispatched to each atom (total backup), and when a type 1 customer asks for service, a server with the highest priority among available servers (Table 2) will be dispatched. In the case of type 2 customer, the first two preferred servers are dispatched simultaneously. If one of them is busy, the third server from the priority list is sent. When only one of the three servers is available, it is assigned as a single dispatch. As a result, in addition to the transitions which occur at the cube edges (Figure 1), some upward transitions can occur on the diagonals of this cube, such as:

$$\{000\} \to \{110\}, \qquad \{100\} \to \{111\},$$
$$\{010\} \to \{111\}.$$

The mean service rate for a customer of type 1 served by server i is μ_i. In the case of a customer of type 2, the model considers two servers that are servicing simultaneously the same customer, similar to two independent servers that service two separated type 1 customers. Thus, when a customer of type 2 is served by servers i and k, the mean service rates are μ_i and μ_k, respectively.

Now, the balance equation for state {110} is as follows. The system leaves this state if a customer arrives, or server 1 or 2 completes its service; therefore, the transition rate is $(\lambda + \mu_1 + \mu_2)P\{110\}$. Moreover, the system enters this state in one of the following four ways (in accordance with Table 4):

i) From state {000} if the customer of type 2 arrives from atom 2;

ii) From state {010} if the customer of type 1 arrives from atom 1 or 2;

iii) From state {100} when the customer of type 1 arrives from atom 2;

iv) From state {111} when a service of server 3 is completed.

The balance equation of state {110} is written by:

$$(\lambda + \mu_1 + \mu_2)P\{110\} = \lambda_2^{[2]}P\{000\}$$
$$+ \left(\lambda_1^{[1]} + \lambda_2^{[1]}\right)P\{010\} + \lambda_2^{[1]}P\{100\} + \mu_3 P\{111\}. \tag{35}$$

Chelst and Barlach [68] studied an example of multiple dispatch HQM in a police patrol system for the first time. They proposed the HQM and AH models for ESSs, in which two servers could be dispatched together to one customer; however, these servers were homogeneous. Then, they estimated the performance measures of this system by the exact and approximate models. In the approximate model, they assumed that:

- U_j is the jth ranked server is unavailable;
- F_j is the jth ranked server is available (free);
- $P\{U_1U_2\cdots U_{j-1}F_j\}$ is the probability that the jth ranked server will be dispatched to the customer of type 1, which requires service by a server;
- $P\{U_1\cdots U_{m-1}F_mU_{m+1}\cdots U_{j-1}F_j\}$ is the probability that the jth and mth ranked servers will be simultaneously dispatched to the customer of type 2, which requires the simultaneous service by two servers.

For a set of identical servers, we have:

$$P\{U_1U_2\cdots U_{j-1}F_j\} = Q(N, j-1, 1)\rho^{j-1}(1-\rho), \tag{36}$$

$$P\{U_1U_2\cdots U_{k-1}F_kU_{k+1}\cdots U_{j-1}F_j\}$$
$$= Q(N, j-1, 2)\rho^{j-2}(1-\rho)^2, \tag{37}$$

where $Q(N, j-1, 1)$ and $Q(N, j-1, 2)$ are the correction factor for customers of types 1 and 2. The first correction factor is calculated like Larson's correction factor (Eq. (4)); however, unlike Larson's models, there is not a simple closed form for $P(S_k)$ (the reader may refer to [68], Appendix II):

$$Q(N, j-1, 1) = \sum_{k=j-1}^{N-1} \frac{\binom{k}{j-1}}{\binom{N}{j-1}} \times \frac{N-k}{N-(j-1)}$$
$$\times \frac{P(S_k)}{\rho^{j-1}(1-\rho)}. \tag{38}$$

To find the second correction factor, suppose that k out of N servers is busy. The probability that only one server out of the first $(j-1)$ servers is free is as follows:

$$\binom{k}{j-2}\binom{N-k}{1} \Big/ \binom{N}{j-1}. \tag{39}$$

The probability that the free server is the mth ranked server is $1/(j-1)$ and is not dependent on m. Therefore, the probability that only the mth server out of the first $(j-1)$ servers is free is as follows:

$$\binom{k}{j-2}\binom{N-k}{1}\Big/\binom{N}{j-1}(j-1)\,. \tag{40}$$

The conditional probability that the jth server is available is as follows:

$$(N-k-1)/N-(j-1). \tag{41}$$

In addition, Eq. (37) becomes as follows:

$$P\{U_1U_2\cdots U_{k-1}F_kU_{k+1}\cdots U_{j-1}F_j\}$$
$$=\sum_{k=j-2}^{N-1}\frac{\binom{k}{j-2}}{\binom{N}{j-1}}\frac{(N-k)}{(j-1)}\frac{N-k-1}{N-(j-1)}P(S_k). \tag{42}$$

The second correction factor is calculated by Eqs. (38) and (42) and is independent of m.

$$Q(N,j-1,2)=\frac{\sum\limits_{k=j-2}^{N-1}\frac{\binom{k}{j-2}}{\binom{N}{j-1}}\frac{(N-k)}{(j-1)}\frac{N-k-1}{N-(j-1)}P(S_k)}{\rho^{j-2}(1-\rho)^2}. \tag{43}$$

Davoudpour et al. [69] introduced a probabilistic coverage model that integrates the MEXCLP with hypercube queuing model. They indicated the applicability of this model by applying it to an EMS center in Tehran with two basic support and two advanced support ambulances. Because of the small size of their problem, they solved steady-state equations to calculate state probabilities. They concluded that the number of servers at the center had large effect on the number of customer, who has been responded to. Then, they showed that the relationship between this performance measure (i.e., system responsiveness) and the parameters of the system was linear. Sudtachat et al. [70] tried to maximize the patient survival probability in a system with two types of ambulances: Basic Life Support (BLS) and Advanced Life Support (ALS). They also considered three priorities for customers based on their severity level, which is determined at first by the dispatcher and can be updated when the server arrives on the scene. The customers of priority 1 need to be served by two servers simultaneously, an ALS and a BLS; the customers of priority 2 are served by the closest available BLS. The customers of priority 3 need one BLS which is selected according to an ordered preference list, because this type of customers is considered non-critical and sending the closest BLS unit to them may make ambulances unavailable for the next life-threatening customers. They developed a simulation model for small problems and proposed a heuristic algorithm based on the AH model for large-scale problems to design dispatching strategies. They concluded that dispatching based on customer priorities improved patient survival probability rather than dispatching based on the closest strategy. They also showed that the number of ALS units and their location are important factors in the efficiency of the heuristic policy.

Iannoni and Morabito [71] formulated an HQM that simultaneously takes into account various assumptions including different types of customers and servers, partial backup, single or multiple dispatch, and a third state for servers. They stated that these extended HQM could be embedded directly into an optimization procedure and might be suitable for evaluating the performance measures. Iannoni et al. [72] embedded HQM into a GA algorithm to optimize the size of each atom for the model presented in Iannoni and Morabito [71]. In this GA/hypercube algorithm, each generated configuration (represented by a chromosome) is evaluated by the hypercube model. They verified that this GA/hypercube algorithm was effective in calculating the performance measures, and showed that these measures could be improved only by modifying the atom sizes of the system, and it was not necessary to relocate the ambulances and additional investments on capacity. Iannoni et al. [73] used this algorithm to determine the locations of ambulances and their coverage areas to minimize the response time and imbalances in the ambulance workload. In this study, ambulance bases can be located anywhere along the highways. Table 6 presents a list of studies in the literature with focus on their assumptions.

4. Location models and solution approaches

In recent years, the increasing costs of emergency service, high volume of emergency calls, and traffic problem have made the location-allocation problem a major issue in designing emergency service systems. On the other hand, the server performance is related to the dispatching and districting policy, which are dependent upon the locations of emergency facilities and allocation of servers. In addition, there are usually a limited number of servers that must be allocated to the facilities to ensure adequate coverage and appropriate response time. Most studies in the literature tried to integrate location models with HQM because this model is able to assess the potential node of facility location from different perspectives. Table 7 represents a summary of these studies. As shown in this table, most studies used approximate hypercube

Table 6. List of studies in the literature with focus on their assumptions.

Category	**Type of customer**	**Queue discipline**		
		No queue	**FIFO**	**Other type**
Single dispatch, total backup, homogeneous servers	Identical customer	Larson [2], Larson and Franck [10], Chelst and Jarvis [17], Berman et al. [22], Larson and Rich [18], Chiyoshi et al. [26], Rajagopalan et al. [33], Galvao et al. [27], Chiyoshi and Morabito [3], Toro-Diaz et al. [32], Saydam et al. [34], and Sudtachat et al. [35]	Larson [14], Batta et al. [25]	—
	Non-identical customer	Goldberg et al. [30]	Berman and Larson [23], McLay and Mayorga [31]	Souza et al. [19]
Single dispatch, total backup, non-homogeneous servers	Identical customer	Brandeau and Larson [12], Zhu and McKnew [50], Geroliminis et al. [52], Erkut et al. [55], Rajagopalan and Saydam [58], Larson and McKnew [37] Geroliminis et al. [53], Budge et al. [41], Geroliminis et al. [54], and Toro-Díaz et al. [42]		—
	Non-identical customer	Halpern [36], McKnew [38], Jarvis [15], Burwell [39], Burwell et al. [40], Ingolfsson et al. [52], Budge et al. [13], Boyaci and Geroliminis [45], Boyaci and Geroliminis [46], and Ansari et al. [43]	Goldberg and Paz [49], Takeda et al. [21], McLay [57], Baptista and Oliveira [47]	Iannoni et al. (2015)
Single dispatch, partial backup, homogeneous servers	Identical customer	—	—	—
	Non-identical customer	—	—	—

Table 6. List of studies in the literature with focus on their assumptions (continued.)

Category	**Type of customer**	**Queue discipline**		
		No queue	**FIFO**	**Other type**
Single dispatch, partial backup, non-homogeneous servers	Identical customer	Kim and Lee [66]		
	Non-identical customer	Mendonca and Morabito [60], Atkinson et al. [61], Atkinson et al. [63], Iannoni et al. [64]	Morabito et al. [65]	—
Multiple dispatch, total backup, homogeneous servers	Identical customer	—	—	—
	Non-identical customer	Chelst and Barlach [68]	—	—
Multiple dispatch, total backup, non-homogeneous servers	Identical customer	—	—	—
	Non-identical customer	Davoudpour et al. [69], Sudtachat et al. [70]	—	—
Multiple dispatch, partial backup, homogeneous servers	Identical customer	—	—	—
	Non-identical customer	—	—	—
Multiple dispatch, partial backup, non-homogeneous servers	Identical customer	—	—	—
	Non-identical customer	Iannoni and Morabito [71], Iannoni et al. [72], Iannoni et al. [73]	—	—

queuing model since this model is easier to solve than the exact model. Batta et al. [25] showed that using the AHQM in location models usually overestimated the coverage and underestimated the number of required servers.

Geroliminis et al. [52-54] and Toro-Diaz et al. [32] developed a model in the framework of hypercube to optimize the performance measures in large urban networks with many servers, without using approximate approaches. The objective of this model is to minimize the mean system response time, subject to hypercube constraints. The general structure of these models is as follows:

min :

$$z=\sum_{j=1}^{N}\sum_{m=1}^{M}\rho_{jm}t_{jm}, \tag{44}$$

s.t. :

$$\sum_{i\in W_m} x_i\geq y_m, \qquad m=1,2,\ldots,M, \tag{45}$$

$$\sum_{i=1}^{I} x_i=N, \tag{46}$$

Table 7. List of studies in the literature with focus on location models.

Ref.	Hypercube		Location model	Objective function	Solution approach	HQM as an optimization	Case study
	Exact	App[1].					
[22]		*	Q-median	Minimization of the travel time	—		
[23]		*	Q-median	Minimization of the steady state expected travel time	—		
[25]		*	AMEXCLP	Maximization of the expected coverage	Daskin's heuristic procedure followed by post heuristic and post-IP analysis		
[30]		*		Maximization of the expected success rate	Pairwise interchange heuristic		*
[49]		*		Maximization of the expected success rate	Pairwise interchange heuristic		
[52]	*			Minimization of the mean response time	An iterative heuristic		*
[33]		*	DECL	Minimization of the number of ambulances	Incremental search heuristic		
[27]		*	EMALP	Maximization of the number of customers covered with reliability α	Simulated annealing and pure Vertex Substitution (VS) local search heuristic		
[55]		*		Maximization of the number of survivors	Jarvis's location-allocation heuristic	*	
[52]		*		Maximization of the expected coverage	Branch and bound algorithm and an iterative heuristic		*
[57]		*	MEXCLP2	Maximization of the number of high level customer covered within a threshold	Branch and bound algorithm		*
[58]		*	MERLP	Minimization of expected response distances	Reactive Tabu Search (RTS) algorithm and a greedy search heuristic		*
[53]	*			Minimization of mean response time	Discrete-event simulation		*
[31]		*		Maximization of the number of customers covered in a time threshold	—		*
[54]	*			Minimization of mean response time	Embedding heuristic of the GA with hypercube model	*	*
[32]	*			Minimization of the mean response time	Genetic algorithm		*
[34]		*	DRCL	Minimization of the total number of redeployments and ambulances	Reactive Tabu Search (RTS) algorithm and a search heuristic		*
[69]	*		MEXCLP	Maximization of the coverage of emergency region	Branch and bound algorithm		*
[42]		*		Minimization of the mean response time and maximization of the fairness	Reactive Tabu Search (RTS)		*
[43]		*		Maximization of the coverage	An iterative heuristic		*
[66]	*		PLSCP	Minimization of the number of ambulances	Two iterative algorithms based on hypercube and simulation	*	

[1]: Approximate.

$$x_i \in \{0,1\}, \qquad i = 1,2,\cdots,I, \tag{47}$$

$$y_m \in \{0,1\}, \qquad m = 1,2,\cdots,M, \tag{48}$$

$$\rho_{jm} = h_m \frac{\sum\limits_{V_i \in E_{jm}} P\{V_i\}}{1 - P\left\{V_{2^N-1}\right\}},$$

$$j = 1,\cdots,N; \qquad m = 1,\cdots,M, \tag{49}$$

$$P(V_m)\left[\sum_{\substack{i \\ \left\{V_i \in C_N \,:\, d^{+}_{im} = 1\right\}}} \mu_{im} + \sum_{\substack{i \\ \left\{V_i \in C_N \,:\, d^{-}_{im} = 1\right\}}} \lambda_{im}\right]$$

$$= \sum_{\substack{i \\ \left\{V_i \in C_N \,:\, d^{-}_{im} = 1\right\}}} \mu_{im} P\{V_i\} + \sum_{\substack{i \\ V_i \in C_N \,:\, d^{+}_{im} = 1}} \lambda_{im} P\{V_i\},$$

$$m = 1,\cdots,2^N - 1, \tag{50}$$

$$\sum_{i=0}^{2^N-1} P\{V_i\} = 1, \tag{51}$$

there are two decision variables:

$$x_i = \begin{cases} 1 & \text{if afacility is located at potential site } i \\ 0 & \text{otherwise} \end{cases}$$

$$y_m = \begin{cases} 1 & \text{if demand point } m \text{ is covered} \\ 0 & \text{otherwise} \end{cases}$$

where $P\{V_k\}$ is the steady-state probability of the state corresponding to vertex V_k, $k = 0,1,\cdots,2N-1$; d^{-}_{im} and d^{+}_{im} represent the downward and upward Hamming distances between vertices V_i and $V_m (d^{-}_{im} + d^{+}_{im} = d_{im})$; the Hamming distance between two vertices V_i and V_m is the number of digits that are different between two vertices. For example, the Hamming distance between states {0011} and {0111} is equal to one; however, the distance between states {0110} and {1001} is equal to 4. d^{-}_{im} and d^{+}_{im} represent the number of digits changed from 0 to 1 and from 1 to 0, respectively. Given that the system is in state i, λ_{ij} and μ_{ij} are the upward and downward rates of transition from state i to state j corresponding to vertices V_i and V_m.

In this model, Constraint (46) specifies the demand points to be covered. Constraint (47) specifies the number of facilities that must be deployed. Constraints (48) and (49) are integrality constraints for the decision variables. Constraint (51) calculates ρ_{jm}, where its denominator shows the probability that all servers are not busy. Constraint (51) shows the set of flow-balancing equations, which compute the probability of states. Actually, these equations are the modified versions of the original HQM [2] to take into account different service rates for inter-district and intra-district customers. Constraint (52) guarantees that the sum of probabilities is equal to one. This model is in fact a two-step model, in which the service area is distributed into sub-areas in the first step, and, in parallel, the required number of servers is assigned to each sub-area. In the second step, the optimal locations of servers in their atoms are determined.

5. Conclusion and future research

This study reviewed the literature on hypercube queuing model with focus on the research published after Larson [2]. These studies were classified with respect to their assumptions, such as dispatch policy, backup policy, and the homogeneity of servers. The growing attention to the hypercube model in comparison with simulation approaches results from its application in real-world problems. Actually, the HQM can reflect various aspects of an emergency service system and describe the states of these systems, too. However, the existing models are far from real-world situations, and still much work remains to be done. As an example, Souza et al. [19] discussed the provision of a model to subdivide a service area into smaller sub-areas optimally. Davoudpour et al. [69] considered a number of uncertain parameters, such as demand rate and service time. Chiyoshi et al. [3] suggested that different demand rates should be considered during a day or a week. They also offered multiple dispatching, in which servers are not sent simultaneously.

The authors of this paper suggest that eliminating the server's returning time can improve the HQM. The main purpose of an emergency system is to provide service as fast as possible. In real-world examples, the response time is reduced by eliminating the server's returning time to its base. In addition, when a customer enters during the time of returning, the server can go to the customer's location directly. Because of the complexity of modeling such assumptions, the existing studies have supposed that a server returns to its base when a service is completed and will then be dispatched to serve a queued customer. Therefore, taking into account these assumptions helps calculate the performance measures more accurately. On the other hand, in EMS papers, there are no various policies for returning servers, and most articles suppose that servers should come back to their home base station; however, this is not necessary. Sometimes, servers should come back to another station for better coverage. It appears that a new state definition in the HQM models should be formed for this repositioning problem to save the lives of more people.

Furthermore, due to the complex nature of the problems in this field, many studies have used a fixed procedure to dispatch servers. In real-world problems, server dispatch and backup policies are dependent upon the conditions of the system, such as the real-time location of the server, demand fluctuation, and the type of the customer's demand. Therefore, it is recommended that a flexible procedure be defined to reflect these conditions. In view of the aforementioned literature and Table 7, which summarizes the studies that have tried to integrate the HQM with location model, a number of suggestions for future studies are given below:

- As noted before, to design an ESS, decisions are divided into two types: strategic decisions for determining the number and locations of servers and tactical decisions for specifying dispatching policies and server coverage areas. As it turns out, strategic and tactical decisions are related to each other. Thus, the location model and HQM should be integrated to increase their effectiveness. The studies presented so far have not been successful in fully integrating the two models and are, at best, designed in two phases, and there is no relationship between the state probabilities and decision variables of the location model. In fact, the balance equations have not been considered in location models explicitly, and existing models use an approximate hypercube or are designed in two phases. Therefore, it is a promising area for the further study to design a model that can provide a relationship between balance equations and location variables;
- As it can be seen in Table 7, the objective functions of most location models are either travel time minimization or coverage maximization. Furthermore, all of these studies have only a single objective function. Despite the humanitarian nature of these problems, the economic concern is noteworthy, because, in these types of problems, available financial resources are very restricted. Thus, models with multiple objectives considering economic and performance measures are more effective. Additionally, environmental measures can be regarded;
- Due to the significance of travel time as a portion of service time, many studies have tried to analyze this time more accurately [12,18,30,41]; however, travel time is always evaluated as part of the service time. Therefore, it can be very helpful to define a measure to examine the travel time solely. It seems reasonable to consider separate rates for travel time and on-scene time, because travel time is dependent on factors (e.g., distance and type of vehicles) while on-scene time depends on the expertise of the medical team and severity of the incident. It is recommended that distinct distributions with different rates of travel time and on-scene time be used.

References

1. Galvao, R.D. and Morabito, R. "Emergency service systems: The use of the hypercube queueing model in the solution of probabilistic location problems", *International Transactions in Operational Research*, **15**(5), pp. 525-549 (2008).
2. Larson, R.C. "A hypercube queuing model for facility location and redistricting in urban emergency services", *Computers & Operations Research*, **1**(1), pp. 67-95 (1974).
3. Chiyoshi, F.Y. and Morabito, R.A. "Tabu search algorithm for solving the extended maximal availability location problem", *International Transactions in Operational Research*, **18**(6), pp. 663-678 (2011).
4. Chaiken, J.M., *Hypercube Queuing Model: Executive Summery*, Department of Housing and Urban Development, Santa Monica, CA: RAND Corporation (1975).
5. Larson, R.C., *Hypercube Queuing Model: User's Manual*, New York City Rand Institute (1975).
6. Larson, R.C., *Hypercube Queuing Model: Program Description*, New York City Rand Institute (1975)
7. Sacks, S.R. "Optimal spatial deployment of police patrol cars", *Social Science Computer Review*, **18**(1), pp. 40-55 (2000).
8. Sacks, S. "Evaluation of police patrol patterns", *Economics Working Papers*, p. 200317 (2003).
9. Larson, R., *OR Models for Homeland Security*, OR/MS Today, **31**(5), pp. 22-29 (2004).
10. Larson, R.C. and Franck, E.A. "Evaluating dispatching consequences of automatic vehicle location in emergency services", *Computers & Operations Research*, **5**(1), pp. 11-30 (1978).
11. Larson, R.C. and Odoni, A.R., *Urban Operations Research*, Prentice Hall, Englewood Cliffs, N.J. (1981).
12. Brandeau, M.L. and Larson, R.C. "Extending and applying the hypercube queueing model to deploy ambulances in Boston", *Management Science*, **22**, pp. 121-153 (1986).

13. Budge, S., Ingolfsson, A., and Erkut, E. "Approximating vehicle dispatch probabilities for emergency service systems with location-specific service times and multiple units per location", *Operations Research*, **75**(1), pp. 251-255 (2009).

14. Larson, R.C. "Approximating the performance of urban emergency service systems", *Operations Research*, **23**(5), pp. 845-868 (1975).

15. Jarvis, J.P. "Approximating the balance behavior of multi-server loss systems", *Management Science*, **31**(2), pp. 235-239 (1985).

16. Jarvis, J.P. "Optimization in stochastic service systems with distinguishable servers", Ph.D. Dissertation, Massachusetts Institute of Technology (1975).

17. Chelst, K.R. and Jarvis, J.P. "Estimating the probability distribution of travel times for urban emergency service systems", *Operations Research*, **27**(1), pp. 199-204 (1979).

18. Larson, R.C. and Rich, T.F. "Travel-time analysis of New York city police patrol cars", *Interfaces*, **17**(2), pp. 15-20 (1987).

19. Souza, R.M., Morabito, R. and Chiyoshi, F.Y. "Incorporating priorities for waiting customers in the hypercube queuing model with application to an emergency medical service system in Brazil", *European Journal of Operational Research*, **242**(1), pp. 274-285 (2015).

20. Larsen, R.C. and Odoni, A.R., *Urban Operation Research: Logistical and Transportation Planning Methods*, 2nd Ed., Dynamic Ideas (2007).

21. Takeda, R.A., Widmer, J.A., and Morabito, R. "Analysis of ambulance decentralization in an urban emergency medical service using the hypercube queueing model", *Computers and Operations Research*, **34**(3), pp. 727-741 (2007).

22. Berman, O., Larson, R.C., and Odoni, A.R. "Developments in network location with mobile and congested facilities", *European Journal of Operational Research*, **6**(2), pp. 104-116 (1980).

23. Berman, O. and Larson, R.C. "The median problem with congestion", *Computers and Operations Research*, **9**(2), pp. 119-126 (1982).

24. Daskin, M.S. "A maximal expected covering location model: formulation, properties, and heuristic solution", *Transportation Science*, **17**(1), pp. 48-70 (1983).

25. Batta, R., Dolan, J.M., and Krishnamurthy, N.N. "The maximal expected covering location problem: Revisited", *Transportation Science*, **23**(4), pp. 277-287 (1989).

26. Chiyoshi, F.Y., Galvao, R.D., and Morabito, R. "A note on solutions to the maximal expected covering location problem", *Computers and Operations Research*, **30**(1), pp. 87-96 (2002).

27. Galvao, R.D., Chiyoshi, F.Y., and Morabito, R. "Towards unified formulations and extensions of two classical probabilistic location models", *Computers and Operations Research*, **32**(1), pp. 15-33 (2005).

28. ReVelle, C.S. and Hogan, K. "The maximum availability location problem", *Transportation Science*, **23**(3), pp. 192-200 (1989).

29. Glover, F. and Laguna, M., *Tabu Search*, Kluwer Academic Publishers (1997).

30. Goldberg, J., Dietrich, R., Chen, J.M., and Mitwasi, M.G. "Validating and applying a model for locating emergency medical vehicles in Tucson, Az", *European Journal of Operational Research*, **49**(3), pp. 308-324 (1990).

31. McLay, L.A. and Mayorga, M.E. "Evaluating emergency medical service performance measures", *Health Care Management Science*, **13**(2), pp. 124-136 (2010).

32. Toro-Diaz, H., Mayorga, M.E., Chanta, S., and McLay, L.A. "joint location and dispatching decisions for emergency medical services", *Computers and Industrial Engineering*, **64**(4), pp. 917-928 (2013).

33. Rajagopalan, H.K., Saydam, C., and Xiao, J. "A multiperiod expected covering location model for dynamic redeployment of ambulances", Euro Working Group on Transportation (2005).

34. Saydam, C., Rajagopalan, H.K., Sharer, E., and Lawrimore-Belanger, K. "The dynamic redeployment coverage location model", *Health Systems*, **2**(2), pp. 1-17 (2013).

35. Sudtachat, K., Mayorga, M.E., and Mclay, L.A. "A nested-compliance table policy for emergency medical service systems under relocation", *Omega*, **58**, pp. 154-168 (2016).

36. Halpern, J. "The accuracy of estimates for the performance criteria in certain emergency service queuing systems", *Transportation Science*, **11**(3), pp. 223-242 (1977).

37. Larson, R.C. and McKnew, M.A. "Police patrol-initiated activities within a systems queueing model", *Management Science*, **28**(7), pp. 759-774 (1982).

38. McKnew, M.A. "An approximation to the hypercube model with patrol-initiated activities: an application to police", *Decision Sciences*, **14**(3), pp. 408-418 (1983).

39. Burwell, T.H. "A spatially distributed queuing model for ambulance systems", Ph.D. Dissertation, Clemson University, Clemson, S.C. (1986).

40. Burwell, T.H., Jarvis, J.P., and McKnew, M.A. "Modeling co-located servers and dispatch ties in the hypercube model", *Computers and Operations Research*, **20**(2), pp. 113-119 (1993).

41. Budge, S., Ingolfsson, A., and Zerom, D. "Empirical analysis of ambulance travel times: The case of Calgary emergency medical services", *Management Science*, **56**(4), pp. 716-723 (2010).

42. Toro-Díaz, H., Mayorga, M.E., McLay, L.A., Rajagopalan, H.A., and Saydam, C. "Reducing disparities in large-scale emergency medical service systems", *Journal of the Operational Research Society*, **66**(7), pp. 1-13 (2014).

43. Ansari, S., McLay, L.A., and Mayorga, M.E. "A maximum expected covering problem for district design", *Transportation Science*, **51**(1), pp. 376-390 (2015).

44. Boyaci, B. and Geroliminis, N. "Extended hypercube models for large scale spatial queuing systems", *90th Annual Meeting of the Transportation Research Board*, Washington D.C. (2011).

45. Boyaci, B. and Geroliminis, N. "Facility location problem for emergency and on-demand transportation systems", *91th Annual Meeting of the Transportation Research Board*, Washington D.C. (2012).

46. Boyaci, B. and Geroliminis, N. "Approximation methods for large-scale spatial queueing systems", *Transportation Research Part B*, **74**, pp. 151-181 (2015).

47. Baptista, S. and Oliveira, R.C. "A case study on the application of an approximated hypercube model to emergency medical systems management", *Central European Journal of Operations Research*, **20**(4), pp. 559-581 (2012).

48. Iannoni, A.P., Chiyoshi, F.Y., and Morabito, R. "A spatially distributed queuing model considering dispatching policies with server reservation", *Transportation Research Part E*, **75**, pp. 49-66 (2015).

49. Goldberg, J. and Paz, L. "Locating emergency vehicle bases when service time depends on call location", *Transportation Science*, **25**(4), pp. 264-280 (1991).

50. Zhu, Z. and McKnew, A. "A goal programming workload balancing optimization model for ambulance allocation: An application to Shanghai, P.R. China", *Socio-Economic Planning Sciences*, **27**(2), pp. 137-148 (1993).

51. Lei, H., Wang, R., and Laporte, G. "Solving a multi-objective dynamic stochastic districting and routing problem with a co-evolutionary algorithm", *Computers & Operations Research*, **67**, pp. 12-24 (2016).

52. Geroliminis, N., Karlaftis, M.G., Stathopoulos, A., and Kepaptsoglou, K. "A districting and location model using spatial queues", *83rd Transportation Research Board Annual Meeting*, at Washington DC, USA (2004).

53. Geroliminis, N., Karlaftis, M.G. and Skabardonis, A. "A spatial queuing model for the emergency vehicle districting and location problem", *Transportation Research Part B*, **43**(7), pp. 798-811 (2009).

54. Geroliminis, N., Kepaptsoglou, K., and Karlaftis, M.G. "A hybrid hypercube-genetic algorithm approach for deploying many emergency response mobile units in an urban network", *European Journal of Operational Research*, **210**(2), pp. 287-300 (2011).

55. Erkut, E., Ingolfsson, A., and Erdogan, G. "Ambulance location for maximum survival", *Naval Research Logistics (NRL)*, **55**(1), pp. 42-58 (2007).

56. Ingolfsson, A., Budge, S., and Erkut, E. "Optimal ambulance location with random delays and travel times", *Health Care Management Science*, **11**(3), pp. 262-274 (2008).

57. McLay, L.A. "A maximum expected covering location model with two types of servers", *IIE Transactions*, **41**(8), pp. 730-741 (2009).

58. Rajagopalan, H.K. and Saydam, C. "A minimum expected response model: formulation, heuristic solution and application", *Socio-Economic Planning Sciences*, **43**(4), pp. 253-262 (2009).

59. Marianov, V. and ReVelle, C. "The queueing maximal availability location problem: a model for siting of emergency vehicles", *European Journal of Operational Research*, **93**(1), pp. 110-120 (1996).

60. Mendonca, F.C. and Morabito, R. "Analysing emergency medical service ambulance deployment on a Brazilian highway using the hypercube model", *Operational Research Society*, **52**, pp. 261-270 (2001).

61. Atkinson, J.B., Kovalenko, I.N., Kuznetsov, N.Y., and Mikhalevich, K.V. "Heuristic methods for the analysis of a queuing system describing emergency medical service deployed along a highway", *Cybernetics and Systems Analysis*, **42**(3), pp. 379-391 (2006).

62. Mendonca, F.C. and Morabito, R. "Analysing emergency medical service ambulance deployment on a brazilian highway using the hypercube model", *Operational Research Society*, **52**, pp. 261-270 (2001).

63. Atkinson, J.B., Kovalenko, I.N., Kuznetsov, N.Y., and Mikhalevich, K.V. "A hypercube queueing loss model with customer-dependent service rates", *European Journal of Operational Research*, **191**(1), pp. 223-239 (2008).

64. Iannoni, A.P., Morabito, R., and Saydam, C. "Optimizing large-scale emergency medical system operations on highways using the hypercube queuing model", *Socio-Economic Planning Sciences*, **45**(3), pp. 105-117 (2011).

65. Morabito, R., Chiyoshi, F.Y., and Galvao, R.D. "Non-homogeneous servers in emergency medical systems: Practical applications using the hypercube queueing model", *Socio-Economic Planning Sciences*, **42**(4), pp. 255-270 (2008).

66. Kim, S.H. and Lee, Y.H. "Iterative optimization algorithm with parameter estimation for the ambulance location problem", *Health Care Management Science*, **19**(4), pp. 362-382 (2016).

67. ReVelle, C. and Hogan, K. "A reliability-constrained siting model with local estimates of busy fractions", *Environment and Planning B: Planning and Design*, **15**(2), pp. 143-152 (1988).

68. Chelst, K.R. and Barlach, Z. "Multiple unit dispatches in emergency services: Models to estimate system performance", *Management Science*, **27**(12), pp. 1390-1409 (1981).

69. Davoudpour, H., Mortaz, E., and Hosseinijou, S.A. "A new probabilistic coverage model for ambulances deployment with hypercube queuing approach", *International Journal of Advanced Manufacturing Technology*, **70**, pp. 1157-1168 (2014).

70. Sudtachat, K., Mayorga, M.E., and McLay, L.A. "Recommendations for dispatching emergency vehicles under multi-tiered response via simulation", *International Transactions in Operational Research*, **21**(4), pp. 581-617 (2014).

71. Iannoni, A.P. and Morabito, R. "A multiple dispatch and partial backup hypercube queuing model to analyze emergency medical systems on highways", *Transportation Research Part E*, **43**(6), pp. 755-771 (2007).

72. Iannoni, A.P., Morabito, R., and Saydam, C. "A hypercube queueing model embedded into a genetic algorithm for ambulance deployment on highways", *Annals of Operations Research*, **157**(1), pp. 207-224 (2008).

73. Iannoni, A.P., Morabito, R., and Saydam, C. "An optimization approach for ambulance location and the districting of the response segments on highways", *European Journal of Operational Research*, **195**(2), pp. 528-542 (2009).

Biographies

Maryam Ghobadi is a PhD Candidate from 2014 in Kurdistan University in Iran at the Faculty of Engineering. She received her MSc degree in Industrial Engineering from Alzahra University in 2013. Her current research interests include location problem and queuing system. Ms. Ghobadi has published a book on Location Problem.

Jamal Arkat received BSc, MSc, and PhD degrees in Industrial Engineering from Iran University of Science and Technology (IUST) in Iran. He is currently an Associate Professor at the Department of Industrial Engineering at University of Kurdistan, Sanandaj, Iran. His research interests include operations research applications in health care, crisis management, and stochastic processes. He is now the Dean of Faculty of Engineering at University of Kurdistan.

Reza Tavakkoli-Moghaddam is a Professor of Industrial Engineering at University of Tehran in Iran. He obtained his PhD in Industrial Engineering from the Swinburne University of Technology in Melbourne (1998). He is an Associate Member at Academy of Sciences in Iran and serves as Editorial Board of the International Journal of Engineering and Iranian Journal of Operations Research. He was the recipient of the 2009 and 2011 Distinguished Researcher Award and the 2010 Distinguished Applied Research Award at the University of Tehran, Iran. He was selected as National Iranian Distinguished Researcher for 2008 and 2010 in Iran. Professor Tavakkoli-Moghaddam has published four books, 22 book chapters, and more than 1000 papers in reputable academic journals and conferences.