Journal of Industrial and Systems Engineering
Vol. 13, No. 2, pp. 84-104
Spring (April) 2021

Hypercube queuing model for emergency facility location problem considering travel and on-scene service times

Maryam Ghobadi [1], Jamal Arkat [1*], Hiwa Faroughi [1], Reza Tavakkoli-Moghaddam [2]

[1] *Department of Industrial Engineering, University of Kurdistan, Sanandaj, Iran*
[2] *School of Industrial Engineering, College of Engineering, University of Tehran, Tehran, Iran*

mrymghbd@gmail.com, j.arkat@uok.ac.ir, h.farughi@uok.ac.ir, tavakoli@ut.ac.ir

Abstract

The hypercube queuing model is a descriptive model for emergency systems in which servers are mobile and serve customers at their locations. In emergency systems, the service time of each server includes the travel time from the server station to the customer's location, the on-scene time and the travel time from the customer's location to the server station. The on-scene service time depends on factors such as server expertise and the severity of the customer's situation while the travel times depend on factors such as vehicle type, the path, and the traffic volume. Therefore, it is necessary to consider and analyze these two times separately. In the hypercube queuing model presented in this study, the service time is divided into two sections, the travel time and the on-scene service time, both of which follow independent exponential distributions with known rates. A new system state is defined in which the status of servers is classified into idle, serving at the customer's location and traveling. By solving the equilibrium equations with the Gaussian- Elimination method (for small size examples) and simulation (for larger examples), limiting probabilities are obtained, and performance measures (such as the ratio of the on-scene time to the total server busy time) are evaluated. A case study of the road emergency stations of the Red Crescent, which are based in Hamadan province, Iran, is also used to check the model's real-world performance.

Keywords: Emergency systems, Hypercube Queuing Model (HQM), performance measures, discrete event simulation.

1- Introduction

In emergency systems, customers are usually not in a good situation, and any delay in serving them leads to irreparable harm and even death. The goal of emergency systems is to provide service with the highest quality and in the shortest time, and thus, the distance between customers and servers plays an essential role. Usually, in emergency systems, the servers are mobile in which each time a customer contacts a central unit (call center), one or more servers are dispatched to serve the customer. Therefore, the service time begins from the moment a server is assigned to a customer and continues until the server returns to its station. Thus, the service time includes the travel time from the server's station to the customer's location, the on-scene time, and the travel time from the customer's location to the server's station.

*Corresponding author
ISSN: 1735-8272,

Figure 1 shows an overview of service time. As can be seen in this figure, travel time is a significant part of the service time and has a high impact on service quality, so ignoring it will lead to poor analysis.

The Hypercube Queuing Model (HQM), first introduced by Larson (1974), is an efficient descriptive model for systems with mobile servers. This model extends the state space of a queuing system descriptively; thus, each server is considered individually where more complex dispatching policies can be set. The dispatching policy is a priority list that the central unit decides which server to send to serve a new customer. The hypercube term is taken from the state space that describes the status of servers. If there are N servers in a system where each server can be in one of two situations, idle (0) or busy (1) at any time, then 2^N states can be defined, each of them as a vector of zeros and ones. For example, $\{101\}$ is a state in which, first and third servers are busy, and the second server is idle. For $N = 3$, the state space is a cube, and for $N > 3$, it becomes a hypercube.

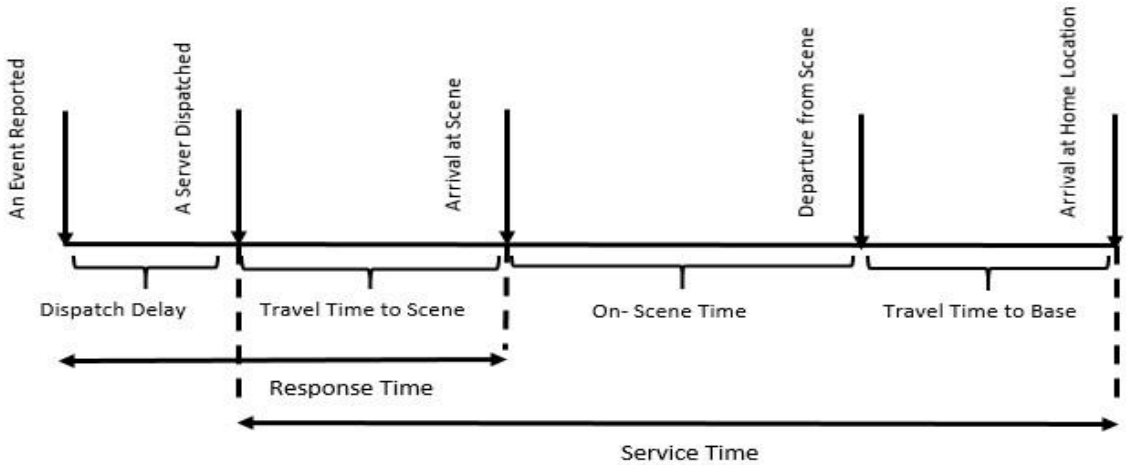

Fig 1. The division of service time (Brandea and Larson, 1986).

In the basic HQM, presented by Larson (1974), the service rate is independent of the locations of servers and customers and also of the type of emergency vehicle. Many other studies in the literature have made the same assumption. In all of these studies, the entire service time follows an exponential distribution with a constant rate. This study is focused mainly on research in which servers have different service rates, known as non-homogenous servers. For the first time, Halpern (1977) evaluated the impact of considering the service rate that was dependent on the locations of customers and servers in a simple system with two servers and two customers. He demonstrated that even in such a small system, the calculation of performance metrics is flawed when the service rate is fixed and similar for all servers.

Larson and Sasanuma (2010) developed a queuing model to measure the impact of seeking on-street parking on traffic congestions. They also developed a model for the case of two types of drivers, patient and impatient. Budge et al. (2010) used an approximation algorithm to find the relationship between travel time and distance. They concluded that a logarithmic transformation makes the travel-time distribution symmetric. Geroliminis et al. (2011) estimated the service rate as the harmonic mean of service rates for all interdistrict and intradistrict requests of demand. Rajagopalan et al. (2011) considered a service time which is varied during the day and developed a two-stage approach to deploying ambulances and schedule crew shifts. They applied this model in a case study to help managers define the preference list for the lengths of personnel shifts. Iannoni et al. (2011) optimized ambulance locations and dispatch policies using several greedy heuristics. In their system, the on-scene and transport times are so large that the impact of travel time has been ignored.

Boyaci and Geroliminis (2012) proposed a mixed-hypercube queuing algorithm (MHQA) in an emergency system in which the total service area is divided into sub-areas. In their presented model, service rate is different for inter- and intra-arrivals and each server thus has three states: free, busy serving an intra-district customer, and busy serving an inter-district customer. They also presented a new definition for the system states. Finally, they merged the sub-areas, and some servers

have therefore been located near the borders between two merged sub-areas and can thus provide service to the two sub-areas at the same rate. As a result, some servers have two states, and others have three states in the final step. Boyaci and Geroliminis (2014) tried to improve the locations of servers in MHQA.

Davoudpour et al. (2014) integrated the hypercube queuing model and the maximal expected coverage location problem (MEXCLP) and presented a probabilistic coverage model for an Emergency Medical System (EMS) center in Tehran with two basic-support and two advanced-support ambulances. In their work, the service rate is dependent on the type of ambulances dispatched according to customer requirements. They calculated the state probabilities by solving steady-state equations for a small-sized problem. Toro-Diaz et al. (2014) eliminated the assumption of exponential service times in the spatial queuing system, and considered the general distribution of service rate, depending upon the server-customer pair. Sudtachat et al. (2014) maximized the patient survival probability in a system with three priorities for customers based on their severity levels. In their research, a combination of ambulances, basic life support (BLS), and advanced life support (ALS) can be sent depending on customer requirements, and service rate is therefore dependent on the type of ambulance.

Boyaci and Geroliminis (2015) proposed a partitioning algorithm to obtain more accurate results in the model presented by Boyaci and Geroliminis (2012). Iannoni et al. (2015) suggested an HQM to increase the probability of serving the higher-level customers immediately. In their study, the situation of a customer defines his/her priority level, and a customer who is not in an emergency is kept waiting until the number of idle servers reaches the threshold number (*i.e.*, cut-off level). In their work, servers have different service rates.

Ansari et al. (2017a) considered a service time for each customer-server pair and used the service rates in an approximation algorithm to estimate the performance measures of the system. They considered these performance measures as constants in a Mixed Integer Linear Programming (MILP) model. Kim and Lee (2016) computed the steady-state probabilities using the HQM and used them in a probabilistic location set covering problem. The objective of their model is to satisfy the reliability requirements. In their work, the service time is not dependent on the distance. Ansari et al. (2017b) used the model by Budge et al. (2009) and provided an approximate HQM in which multiple servers can simultaneously be dispatched to serve a customer. They used a general distribution for the number of servers sent to a customer. They provided an approach for calculating the state probabilities and then estimated the system performance measures. Yoon et al. (2017) used the HQM embedded within mixed-integer linear programming to determine the location of ambulances and dispatching policies. They assumed cutoff priority queue and concluded that the expected coverage would be improved significantly when this paradigm is considered. They also presented an algorithm to define the cutoff value.

Rodrigues et al. (2017) provided an HQM in which customers in the queue are prioritized, and servers support each other partially. They also provided an approximate algorithm to solve more significant size problems. Rodrigues et al. (2018) developed this approximate algorithm for a system with different service rates for servers. They also considered different arrival rates for customers who joined the queue and for those who did not wait in the queue. Karimi et al. (2018) developed the Larson's approximation algorithm for an emergency system in which servers support each other partially. They also considered priorities for customers according to their condition and calculated the performance measures. In their algorithm, the service rate is a weighted average of the dispatch rates and travel times.

In the previously presented research, it is usually assumed that the service time starts when a server is assigned to a customer and continues until the server returns to its base, in which a fixed rate is considered for the total service time. The travel time has either been eliminated or approximated as a part of the service time in most studies while the travel time is a significant part of the service time and has a high impact on the service quality. The travel time from ambulance base to the customer location or when the patient needs to be taken to the hospital should be considered exactly because customers in emergency systems are usually not in a good condition. Also, the service of one server sometimes ends at the customer's location and another customer is in the queue and waiting to receive the service, but the server has to return to its base to relieve manpower fatigue and replenishment of required equipment. Therefore, the travel time and considering its effect on the quality of emergency

services are very important. In this study, the service time similarly starts at the moment a server is assigned to a customer and continues until the server returns to its base; however, this time is divided into two parts (i.e., the on-scene time and the travel time) to study the service time more precisely. Although both on-scene time and travel time follow an exponential distribution; however, the rate of distributions are determined independently and based on the factors which affect them. The travel time (between the station and the customer's location) depends on the factors such as type of vehicle, the distance to be traveled, and the volume of traffic. The on-scene time depends on factors such as the severity of the incident or customer's condition, type of equipment, and the level of expertise and skill of servers. In order to address this issue, a new state definition is presented to demonstrate the statuses of the servers and precisely determine whether they are serving or traveling while they are busy. According to this definition, each server can be in one of the three situations: idle and waiting at its station, traveling from its station to the customer's location or vice versa, and serving at the customer's location. The flow balance equilibrium equations are formed based on the presented state definition, and the performance measures are then calculated. Moreover, new performance measures (e.g., the percentage of time travelled by the servers and percentage of customers served by their closest server) are defined to analyze the impact of the travel time in emergency systems.

The rest of this paper is divided as follows. In section 2, the new state definition for HQM is presented, and then equilibrium equations are formed. In Section 3, some new performance measures are defined, and in section 4, some small numerical examples are presented to describe the states and equations. In section 5, a simulation approach is described and used to solve more abundant examples. The case study is presented in section 6. In section 7, the results and the scopes for future research are presented.

2- Proposed Hypercube Queuing Model

As mentioned earlier, hypercube queueing models are descriptive ones that can analyze the states of emergency systems with mobile servers that support each other. These models calculates the steady-state probabilities under different conditions and dispatching policies. Besides, these models can consider each server individually and evaluate more accurately the server dependent critria and the measures, which relate to the performance of the entire system.Therefore, in this study, an HQM is applied to calculate the performance of an emergency sytem for the cases, in which travel time and service time are independent. There are also other assumptions in this model that are as follows.

- It is assumed that the total area is divided into several sub-areas called atoms.
- A server is located in the center of each atom, and distances between atoms are calculated by their centers.
- The arrival processes of customers are Poisson processes with different rates for each atom.
- The server priority list is fixed and defined based on the distance between the server and customers.
- For each server, a coverage radius is defined, and the server can only serve customers who are at the predefined distance (partial backup).
- A customer might be in the coverage range of several servers, in which case servers will be prioritized in the order of their distance to that customer. When a customer arrives, the closest idle server at the desired radius is sent to serve the customer. If all servers that can serve the customer are busy, the customer will be lost (that is referred to as another emergency system). There might be an idle server in the system, but it cannot serve a new customer because the distance between the server and the customer is greater than the coverage radius.
- The service time in this study is divided into two parts. The on-scene service time follows the exponential distribution. All servers are assumed to be the same in terms of expertise, skill, and associated equipment and therefore have the same rate. The travel time also follows an exponential distribution and the rate of this distribution depends on the distance between the server and the customer.

According to the above assumptions, the status of each server can be idle (0), serving at the customer's location (r_i), or traveling (r_i'). The number r_i is the number of the atom in which the server has been sent to serve. Other symbols are defined as follows.

J: the set of customers, $j = 1, \dots, J$
I: the set of servers, $i = 1, \dots, N$
λ_j: the arrival rate from atom j.
μ: the rate of on-scene service time.
γ_{ij}: the rate of the exponential distribution assumed for the travel time between customer j and server i, which depends on their distance. From now on, this rate referred to as the travel rate. To calculate the service rate, a matrix is used as follows. The matrix is symmetric, and elements in the main diagonal show the rate of the exponential distribution assumed for travel time when the customer and server are from the same atom, which is equal to zero or a small number. Table 1 shows a general frame for the travel rate matrix.

Table 1. The travel rate matrix

	1	**2**	$\cdots$	$\boldsymbol{J}$
1	γ_{11}	γ_{12}	$\cdots$	γ_{1J}
2		γ_{22}	$\cdots$	γ_{2J}
$\vdots$			$\vdots$	$\vdots$
$\boldsymbol{N}$			$\cdots$	γ_{NJ}

k_i: the ordered set of atoms that server i can cover.
$|k_i|$: the total number of atoms covered by server i.
r_i: the status of server i in state R, which is written in two forms r_i and r_i' and can take one of the values $0, 1, 2, \dots, k_i$. If the server status is 0, it means that the server is waiting at its station and is considered to be idle. If the server takes one of the values $1, 2, \dots, k_i$ it means that the server is serving at the atom of that number. Furthermore, if one of the values $1', 2', \dots, k_i'$ is assigned to the server; it means that the server is traveling from its station to the atom of that number to serve a customer or is returning from that atom to its station.
$|r_i|$: the number of the atom that server i dispatched to serve a customer or currently serving a customer in (the numerical value of the status of the server i, regardless of whether it is traveling or serving a customer in its location).
R: the state vector of the system $R = \{r_1, r_2, \dots, r_N\}$.
M_R: the set of servers that are idle in state R.
L_R: the set of servers that are traveling in state R. This set includes servers that are traveling from their station to the customer's location and also includes servers that are returning from the customer's location to their station (i.e., servers with status r_i').
W_R: the set of servers that are serving at the customer's location in the state R (i.e., servers with r_i status).
$P(R)$: the probability of being in state R.

The total number of states of the proposed HQM is equal to:

$$\prod_{i=1}^{N} (2|k_i| + 1) \tag{1}$$

The equilibrium equations for the steady-state of the HQM are written by:

$$P(R)\left(\sum_{j\in\bigcup_{i\in M_R}k_i}\lambda_j+\sum_{i\in W_R}\mu+\sum_{i\in L_R}\frac{1}{2}\left(\gamma_{i,|r_i|}+\gamma_{|r_i|,i}\right)\right)$$
$$=\frac{1}{2}\sum_{i\in M_R}\sum_{k\in K_i}\gamma_{ik}P(R:r_i=0 \to k')+\sum_{i\in L_R}\mu P(R:r_i'\to r_i)+\frac{1}{2}\sum_{i\in W_R}\gamma_{i,|r_i|}P(R:r_i\to r_i')+\sum_{i\in L_R}\lambda_{|r_i|}P(R:r_i'\to 0) \quad (2)$$

$$\sum_R P(R)=1 \quad (3)$$

The left-hand side of equation (2) calculates the exit rate of state R. The first expression computes the total arrival rate for the customers who can receive service in state R. Each customer enters the queue if there is at least one idle server ($i\in M_R$) in its covering radius. The term $\bigcup_{i\in M_R}k_i$ shows the set of atoms supported by idle servers. The term $\sum_{i\in W_R}\mu$ calculates the total service rate for the servers serving a customer in the state R. The term $\sum_{i\in L_R}\frac{1}{2}(\gamma_{i,|r_i|}+\gamma_{|r_i|,i})$ calculates the total rate of the exponential distribution assumed for the travel times in state R. Since the travel status can indicate a server going from its station to the customer's location and can also indicate the server returning from the customer's location to its station, this rate is multiplied by $\frac{1}{2}$. In this HQM, of course, the service time was initially divided into three parts: the travel time from the server station to the customer's location, the on-scene time, and the travel time from the customer's location to the server station, in which case the number of states and, therefore, the number of equilibrium equations would be much higher. Further investigation demonstrated that if the distance traveled from the server location to the customer's location is equal to the distance traveled from the customer's location to the server base; it will lead to similar results to merge the two states as one travel state and add a factor of $\frac{1}{2}$. Further discussion is presented in Section 5.

The right-hand side of equation (2) calculates the rate of entering state R from other states. The term $\sum_{i\in M_R}\sum_{k\in K}\frac{1}{2}\gamma_{ki}$ calculates the total travel rate for those idle servers in state R that had been returning to their station in the previous state. In these states, all servers have the same status as R except a server that is idle in R but had been traveling in the previous state (the transition of $0\to k'$). The coefficient of 1/2 is also given for the reason similar to that mentioned earlier.

The term $\sum_{i\in L_R}\mu$ calculates the total travel rate for servers that are returning to their station in state R but in the previous state had been serving at a customer's location. Therefore, in these states, the status of all servers is the same as state R except a server that is traveling in state R and had been serving at the customer's location in its previous state (the transition of $r_i'\to r_i$).

The term $\sum_{i\in W_R}\frac{1}{2}\gamma_{i,|r_i|}$ calculates the total service rate for the servers that are serving in state R. These servers had been traveling to reach the customer's location in the previous state. Thus, in these states, the status of all servers is the same as the state R except a server that is serving at the customer's location in state R and had been traveling to reach the customer's location in its previous state (the transition of $r_i\to r_i'$). For the same reasons as before, this expression is multiplied by 1/2.

Equation (3) points out that the sum of all probabilities must be one.

3-Computing system performance measures

System performance measures help decision-makers to focus on different goals simultaneously and gain an overall view of the system. These measures can be classified into two general categories based on customer satisfaction (customer-oriented measures) and system satisfaction (system-oriented measures). For instance, the probability that a customer is covered and the number of times that a customer is served by a server from its atom are two critical measures in terms of customer

satisfaction. The most important criterion in terms of the system is the efficiency or system workload factor; the higher its value, the greater the use of resources. Besides, this criterion is essential because more workload results in a more human error due to reduced ability and concentration (Yazdanparast et al., (2018)).The workload of server i, shown with ρ_i, is equal to the proportion of time that the server i is busy. Therefore, it can easily be calculated by the summation of probabilities of those states in which the server is busy. In this study, the server is busy if it is traveling to or from a customer location or is on-scene. Therefore, for server i, this measure is computed by the sum of the probabilities of states, where $r_i \neq 0$.

$$\rho_i = \sum_{R:r_i \neq 0} P(R) \tag{4}$$

The probability of saturation, *i.e.*, the probability that all the servers are busy, is calculated by $\sum_i \sum_{R:r_i \neq 0} P(R)$. In addition to ρ_i, in this study, two other measures are introduced as ρ'_i and ρ''_i. ρ'_i represents the proportion of time that the server i is busy and traveling and ρ''_i shows a proportion of time that server i is busy and serving a customer at its location. These two measures are calculated by:

$$\rho'_i = \sum_{R:r_i \in L_R} P(R) \tag{5}$$

$$\rho''_i = \sum_{R:r_i \in W_R} P(R) \tag{6}$$

The relationship $\rho_i = \rho'_i + \rho''_i$ always holds. The performance measures that are easily computable in the presented HQM include the following.

Workload factors

T_i: The traveling time proportion of server i dispatched to serve a customer from its atom.
TS_i: The traveling time proportion of server i dispatched to serve a customer outside its atom ($TS_i = 1 - T_i$).
S_i: The service time proportion of server i serving a customer from its atom.
SS_i: The service time proportion of server i serving a customer outside its atom ($SS_i = 1 - S_i$).
TO_i: The traveling time ratio of server i into the total time that the server is busy.
SO_i: The on-scene time ratio of server i into the total time that the server is busy ($SO_i = 1 - TO_i$).
TT: The traveling time ratio of all servers (to go to or come from the customer's location) into the total on-scene time.
TW: The total traveling time ratio of the servers into the total time that the servers are busy.
TR: The ratio of the total on-scene time into the total time that the servers are busy ($TR = 1 - TW$).

Efficiency factors

ZL: The probability that a customer is covered.
RTT: The proportion of time that servers are in travel
NF: The probability that a customer is covered by a server from its closest station.

$$T_i = \frac{\sum_{R:r_i \in L_R \& r_i = i'} P(R)}{\rho'_i} \tag{7}$$

$$TS_i = \frac{\sum_{R:r_i \in L_R \& r_i \neq i'} P(R)}{\rho'_i} \tag{8}$$

$$S_i = \frac{\sum_{R:r_i \in W_R \& r_i = i} P(R)}{\rho''_i} \tag{9}$$

$$SS_i = \frac{\sum_{R:r_i \in W_R \& r_i \neq i} P(R)}{\rho''_i} \tag{10}$$

$$TO_i = \frac{\rho'_i}{\rho_i} \tag{11}$$

$$SO_i = \frac{\rho''_i}{\rho_i} \tag{12}$$

$$TT = \frac{\sum_{i=1}^{I} \rho'_i}{\sum_{i=1}^{I} \rho''_i} \tag{13}$$

$$TW = \frac{\sum_{i=1}^{I} \rho'_i}{\sum_{i=1}^{I} \rho_i} \tag{14}$$

$$TR = \frac{\sum_{i=1}^{I} \rho''_i}{\sum_{i=1}^{I} \rho_i} \tag{15}$$

$$ZL = \sum_{j=1}^{J} \sum_{R \in E_j} \frac{\lambda_j}{\lambda} P(R) \tag{16}$$

$$RTT = \frac{\sum_{i=1}^{N} \sum_{R:r_i \in L_R} P(R)}{\sum_{i=1}^{N} \sum_{R:r_i \neq 0} P(R)} \tag{17}$$

$$NF = \sum_{i=1}^{N} \frac{\lambda_i}{\lambda} \frac{\sum_{R:r_i \in W_R \& r_i = the\ number\ of\ closest\ server} P(R)}{\sum_{R:r_i \in W_R} P(R)} \tag{18}$$

E_j in expression (16) includes all the states, in which there is at least one idle server that can serve customers from atom j. Moreover, in Expressions (16) and (18), $\lambda = \sum_j \lambda_j$.

4- Numerical examples and computational results

For a better understanding of the equilibrium equations and further discussions on the proposed HQM, a small size example (example 1) is solved. In example 1, there is an area with three atoms in which an emergency station is located at the center of one of them. The overall view of this area is shown in figure 2. The coverage radius in this example is 1000, so servers are only able to support atoms located less than 1000 meters away. The priority list of servers based on the distance is shown in table 2.

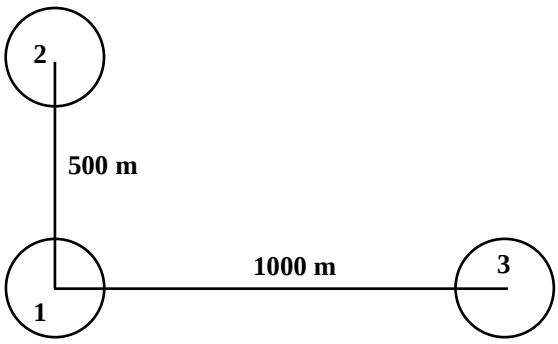

Fig 2. Overview of the area in example 1

Table 2. Priority list for servers in example 1

atom	priority		
	1st	2nd	3rd
1	1	2	3
2	2	1	-
3	3	1	-

As shown in figure 2, the server, which is located in atom 2, cannot serve customers from atom 3, even if it is idle. The server located at atom 3 also cannot support customers from atom 2. Therefore, it is evident that $k_1 = 3, k_2 = 2, k_3 = 2$. As a result, by Equation (1), the total number of states is equal to:

$$\prod_{i=1}^{3}(2|k_i| + 1) = 7 \times 5 \times 5 = 175$$

Figure 3 shows a part of the transition rate diagram for this system. Also, the equilibrium equations for some states are given by:

$$\begin{aligned} P(000)(\lambda_1 + \lambda_2 + \lambda_3) &= \frac{1}{2}\big(\gamma_{11}P(1'00) + \gamma_{22}P(02'0) + \gamma_{33}P(003') + \gamma_{12}P(01'0) \\ &+ \gamma_{21}P(2'00) + \gamma_{13}P(001') + \gamma_{31}P(3'00)\big) \end{aligned} \tag{19}$$

$$\begin{aligned} P(3'00)(\lambda_1 + \lambda_2 + \lambda_3 + \gamma_{13}) &= \frac{1}{2}\gamma_{12}P(3'1'0) + \frac{1}{2}\gamma_{22}P(3'2'0) + \frac{1}{2}\gamma_{33}P(3'03') + \frac{1}{2}\gamma_{13}P(3'01') \\ &+ \mu P(300) \end{aligned} \tag{20}$$

$$\begin{aligned} P(300)(\lambda_1 + \lambda_2 + \lambda_3 + \mu) &= \frac{1}{2}\gamma_{13}P(301') + \frac{1}{2}\gamma_{33}P(303') + \frac{1}{2}\gamma_{12}P(31'0) + \frac{1}{2}\gamma_{22}P(32'0) \\ &+ \frac{1}{2}\gamma_{13}P(3'00) \end{aligned} \tag{21}$$

$$\begin{aligned} P(301')(\lambda_1 + \lambda_2 + \mu + \gamma_{13}) &= \frac{1}{2}\gamma_{12}P(31'1') + \frac{1}{2}\gamma_{22}P(32'1') + \frac{1}{2}\gamma_{13}P(3'01') + \mu P(301) \end{aligned} \tag{22}$$

$$\begin{aligned} P(1'2'3')(\gamma_{11} + \gamma_{22} + \gamma_{33}) &= \mu\big(P(12'3') + P(1'2'3) + P(1'23')\big) + \lambda_1 P(02'3') + \lambda_2 P(1'03') \\ &+ \lambda_3 P(1'2'0) \end{aligned} \tag{23}$$

$$P(321)(3\mu) = \frac{1}{2}\gamma_{13}P(3'21) + \frac{1}{2}\gamma_{31}P(321') + \frac{1}{2}\gamma_{22}P(32'1) \tag{24}$$

The system of equations can be solved by one of the solving methods of linear equations. In this study, the Gaussian-Elimination method is used. This method performs sequential operations on a coefficient matrix to transform this matrix into an upper triangular matrix. The performance measures of Example 1 are calculated using different sets of arrival, travel, and service rates, and the results are shown in table 3. As shown in this table, when the customer's arrival rate decreases (the second row compared to the first row), servers are busy for a shorter time, thus the probability of customer coverage is increased. If the service rate of servers decreases (the third row compared to the first row), servers are busy for longer periods, and the possibility of customer coverage is declined. In the fourth row, the travel rate is less than the first row, so servers spend more time traveling, and the probability of customer coverage is lessened. In this row, TT is also higher than other rows. In the fifth row, the service and travel rates are equal and therefore for each server, T_i and S_i are the same. Besides, in this

row, the proportion of time that each server is traveling is roughly twice the proportion of time that the server spends on the customer's location because each time a server is dispatched for service, it enters the travel status twice.

The comparison of the proposed hypercube model with the basic model, which uses only one rate for the entire service, is impossible in many ways. For example, many of the performance criteria that are accurately quantifiable in the presented model, such as T_i ، TO_i and TW, are not computable in the basic HQM or, at best, can only be approximated. However, the following diagrams help to understand the difference between the performances of these two models. In chart 1, the horizontal axis represents the number of servers, and the vertical axis represents the difference between the coverage probability in the HQM and the basic one. In this chart, N is the number of servers, and other parameters are fixed for each value of N. The value of this measure in the basic HQM is higher than the presented model. This is due to the total service time being considered effective in the basic HQM, whereas, in the presented model, the travel time (which can be considered as the wasted time of each server) is also involved in the calculation of the criterion. Therefore, it can be stated that the basic model reports false coverage. Besides, the magnitude of this difference increases with the problem size, hence using the basic HQM in large-sized problems can lead to tremendous errors in the analysis.

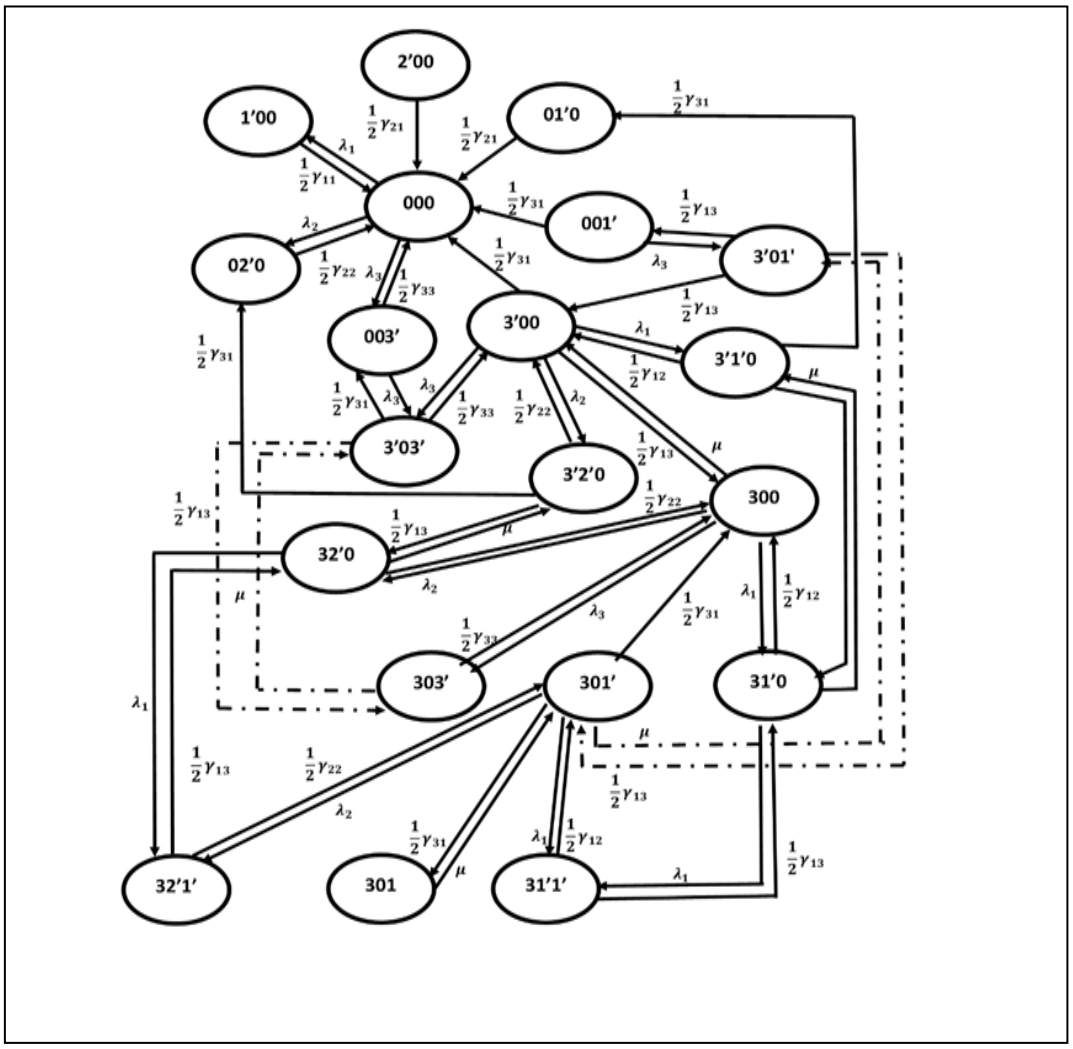

Fig 3. A part of transition rate diagram for example 1

Table 3. Calculation of the performance measures of example 1

	Parameters				Performance Measures							
	λ_j	μ	γ_{ij}		ρ_i	ρ'_i	T_i	S_i	TO_i	TT	TR	ZL
1	(1,2,3)	3	$\begin{pmatrix} 6 & 4 & 2 \\ 4 & 6 & - \\ 2 & - & 6 \end{pmatrix}$	Server1	0.80	0.54	0.12	0.24	0.67	1.46	0.406	0.444
				Server2	0.66	0.35	0.63	0.72	0.53			
				Server3	0.72	0.40	0.67	0.86	0.56			
2	(1,1,1)	3	$\begin{pmatrix} 6 & 4 & 2 \\ 4 & 6 & - \\ 2 & - & 6 \end{pmatrix}$	Server1	0.61	0.37	0.34	0.54	0.61	1.39	0.417	0.689
				Server2	0.52	0.28	0.55	0.65	0.53			
				Server3	0.51	0.3	0.52	0.77	0.59			
3	(1,2,3)	1	$\begin{pmatrix} 6 & 4 & 2 \\ 4 & 6 & - \\ 2 & - & 6 \end{pmatrix}$	Server1	0.88	0.36	0.09	0.21	0.41	0.507	0.66	0.279
				Server2	0.79	0.22	0.61	0.70	0.27			
				Server3	0.84	0.26	0.6	0.82	0.31			
4	(1,2,3)	3	$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & - \\ 1 & - & 1 \end{pmatrix}$	Server1	0.92	0.79	0.19	0.19	0.85	6	0.142	0.192
				Server2	0.87	0.74	0.69	0.69	0.85			
				Server3	0.89	0.76	0.79	0.79	0.85			
5	(1,1,1)	1	$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & - \\ 1 & - & 1 \end{pmatrix}$	Server1	0.88	0.58	0.39	0.39	0.66	2	0.33	0.284
				Server2	0.84	0.56	0.54	0.54	0.66			
				Server3	0.83	0.55	0.59	0.59	0.66			

In chart 2, the difference between the loss probability is shown in the presented HQM and basic HQM. Here, the meaning of loss is that all servers are busy, and if a new customer enters, it will be lost. Other symbols are the same as chart 1. Also, the other parameters are fixed for each value of N. As shown in this chart, the loss probability in the presented HQM is always higher than that in the basic HQM because the travel time in which the servers are busy is taken into consideration in the presented model.

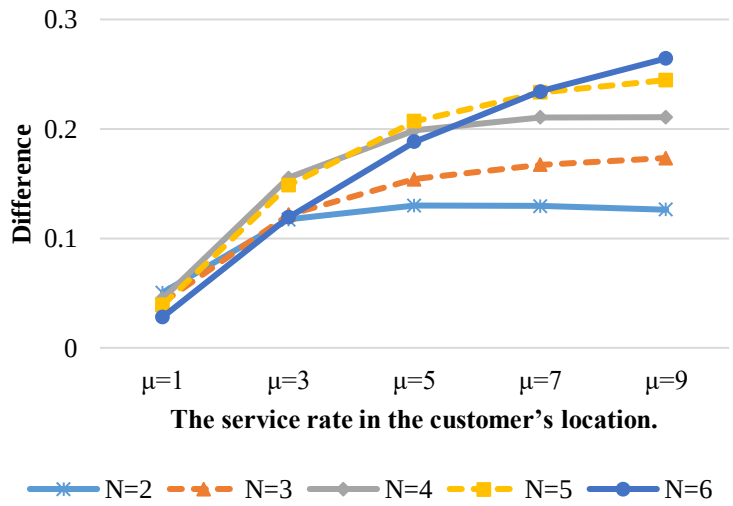

Chart 1. The difference between the covering probability in the presented HQM and the basic HQM.

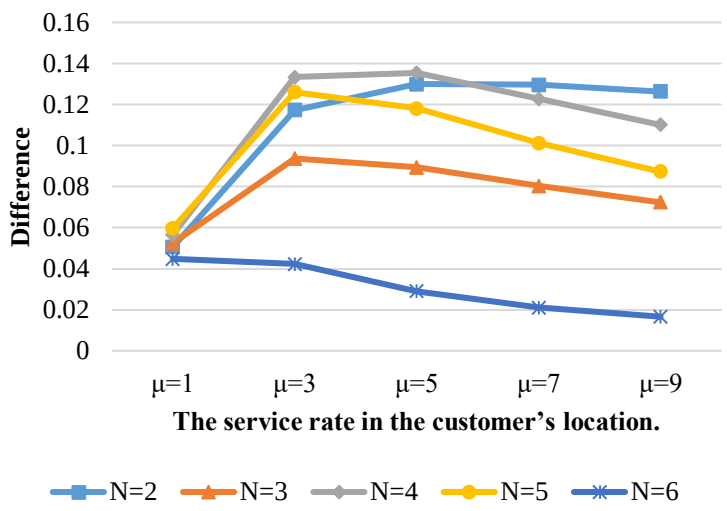

Chart 2. The difference between the loss probability in the presented HQM and the basic HQM.

5- Simulation approach

The existing computer cannot solve examples that have more than 10,000 equilibrium equations. For further analysis, a discrete-event simulation approach is used. Simulation is an appropriate tool for evaluating a queuing model and can analyze larger dimensional examples in a relatively short time. In the simulation approach, the overall behavior of an emergency system is simulated. In this process, some performance measures of the system can be calculated approximately, and there is no need to solve equilibrium equations. Table 4 shows the simulation process used in this study. In this table, t_A is the customer's arrival time, t_s is the server arrival time at scene location or service start time, t_{gf} shows the traveling start time of the server to reach the customer's location and t_{cb} is the traveling start time of the server to return from the customer's location to the station. EL is a N-dimensional vector that shows the current status of the servers. Because of the nature of the simulation process, the service time is divided into three parts, unlike in the presented HQM. Therefore, each EL member can take the values 0 (idle), r_i (service at the customer's location), r_i'' (going to the customer's location) and r_i''' (returning from the customer's location). r_i indicates the atom number in which the server has been sent to serve the customer. W is a set of servers that are serving a customer based on EL. L_1 is the set of servers that are traveling to the customer's location from their station and L_2 shows the set of servers that are traveling from customer's location to their station considering EL. T also shows the total simulation run time, and the value of 100,000 hours is selected. For larger T, only the run time of simulation increased, and the value of performance measures remain unchanged. At all stages, U is a new random number.

To validate the proposed simulation approach, the following examples are designed and solved using both the simulation approach and the proposed HQM. Then, the results of these two methods are compared with each other. Tables 5 and 6 show two categories for the travel rate between two atoms. The blank cells in these tables indicate that the distance between the server and the customer is greater than the coverage radius. Table 7 displays three different categories for the demand rate. The data from these three tables are used to generate examples. For example, when $N = 4$, then only the data related to the first four atoms of each table are considered as an instance; the demand rate based on the first category is $\{6,5,4,3\}$. Furthermore, when $N = 5$, the data related to the sixth atom is not considered, and for $N = 6$, all the data from these tables is valid.

Table 4. Structure of the proposed simulation procedure.

Start $\boldsymbol{t_A = \frac{-1}{\sum_j \lambda_j} \ln U}$; $\boldsymbol{t_S = T}$; $\boldsymbol{t_{cb} = T}$; $\boldsymbol{t_{gf} = T}$ $\quad$ $\boldsymbol{EL = \{0, 0, \ldots, 0\}}$.
First case: if $t_s < t_A$, $t_s \leq t_{gf}$, $t_s \leq t_{cb}$ and $t_s \leq T$ $t = t_S$ From the set of servers that are busy and serving a customer (W), specify the server whose work will be finished sooner. In the state vector, change the status of this server from r_i to r'''_i. $t_S = t - \frac{1}{\sum_{k \in W} \mu_k} \ln U$ $t_{cb} = t - \frac{1}{\sum_{k \in L_2} \gamma_{\lvert r_i \rvert k}} \ln U$
Second case: if $t_{gf} < t_A$, $t_{gf} \leq t_s$, $t_{gf} \leq t_{cb}$ and $t_{gf} \leq T$ $t = t_{gf}$ From the set of servers that are busy and traveling to a customer's location from their station (L_1), specify the server whose work will be finished sooner. In the state vector, change the status of this server from r''_i to r_i. $t_{gf} = t - \frac{1}{\sum_{k \in L_1} \gamma_{k, \lvert r_i \rvert}} \ln U$ $t_S = t - \frac{1}{\sum_{k \in W} \mu_k} \ln U$
Third case: if $t_{cb} < t_A$, $t_{cb} \leq t_s$, $t_{cb} \leq t_{gf}$ and $t_{cb} \leq T$ $t = t_{cb}$ From the set of servers that are busy and traveling to their station from the customer's location (L_2), specify the server whose work will be finished sooner. In the state vector, change the status of this server from r'''_i to 0. $t_{cb} = t - \frac{1}{\sum_{k \in L_2} \gamma_{\lvert r_i \rvert, k}} \ln U$ $t_S = t - \frac{1}{\sum_{k \in W} \mu_k} \ln U$
Fourth case: if $t_A < t_{gf}$, $t_A < t_{cb}$, $t_A \leq t_s$ and $t_A \leq T$ $t = t_A$ Specify the atom where the customer arrives. Add a number to the total number of customers. Identify the closest idle server to that customer. If there is such a server, change its status from 0 to r''_i. $t_A = t - \frac{1}{\sum_j \lambda_j} \ln U$ $t_{gf} = t - \frac{1}{\sum_{k \in L_1} \gamma_{k, \lvert r_i \rvert}} \ln U$ If there is no such server, add a number to the total number of lost customers.
Fifth case: Otherwise, compute the performance measures of the system.

In tables 8 to 10, according to the different values of N, μ, λ, and γ, 54 different examples are generated. In each example, the three criteria of RTT, NF, and ZL are calculated by both the simulation approach and the presented HQM.

Table 5. Travel rate (category 1)

	1	**2**	**3**	**4**	**5**	**6**
1	12	-	-	8	-	-
2	-	12	-	10	-	-
3	-	-	12	-	-	-
4	8	10	-	12	-	-
5	-	-	-	-	12	-
6	-	-	-	-	-	12

Table 6. Travel rate (category 2)

	1	**2**	**3**	**4**	**5**	**6**
1	8	-	-	8	-	-
2	-	8	-	8	-	-
3	-	-	8	-	-	-
4	8	8	-	8	-	-
5	-	-	-	-	8	-
6	-	-	-	-	-	8

Table 7. Three categories for the demand rate of customers

Customer's atom	**1**	**2**	**3**	**4**	**5**	**6**
Category1	6	5	4	3	2	1
Category2	2	2	2	2	2	2
Category3	12	10	8	6	4	2

Table 8. Performance measures, RTT, NF, and ZL for demand rates of category 1

μ	**Performance measures**	**Solving method**	**The travel rate**					
			Cat.1			**Cat. 2**		
			$N=4$	$N=5$	$N=6$	$N=4$	$N=5$	$N=6$
5	RTT	Sim	0.479	0.478	0.476	0.540	0.544	0.556
		Hyp	0.480	0.477	0.475	0.555	0.555	0.555
	NF	Sim	0.716	0.753	0.768	0.710	0.747	0.767
		Hyp	0.726	0.753	0.765	0.717	0.745	0.758
	ZL	Sim	0.415	0.438	0.464	0.372	0.400	0.417
		Hyp	0.414	0.431	0.445	0.372	0.387	0.401
7	RTT	Sim	0.550	0.560	0.567	0.628	0.632	0.634
		Hyp	0.563	0.560	0.559	0.636	0.636	0.636
	NF	Sim	0.725	0.759	0.772	0.717	0.755	0.772
		Hyp	0.736	0.762	0.774	0.725	0.753	0.764
	ZL	Sim	0.457	0.480	0.501	0.410	0.436	0.437
		Hyp	0.463	0.478	0.492	0.410	0.425	0.439
9	RTT	Sim	0.627	0.629	0.632	0.699	0.696	0.695
		Hyp	0.623	0.620	0.619	0.692	0.692	0.692
	NF	Sim	0.733	0.763	0.776	0.723	0.760	0.776
		Hyp	0.743	0.769	0.780	0.730	0.757	0.769
	ZL	Sim	0.487	0.507	0.526	0.435	0.461	0.466
		Hyp	0.494	0.509	0.522	0.435	0.442	0.463

- **"sim" simulation approach**
- **"Hyp" the presented hypercube**

Table 9. Performance measures, RTT, NF, and ZL for demand rates of category 2

μ	Performance measures	Solving method	The travel rate					
			Cat.1			Cat. 2		
			$N=4$	$N=5$	$N=6$	$N=4$	$N=5$	$N=6$
5	RTT	Sim	0.461	0.461	0.454	0.535	0.540	0.541
		Hyp	0.476	0.472	0.469	0.555	0.555	0.555
	NF	Sim	0.733	0.800	0.839	0.718	0.790	0.835
		Hyp	0.742	0.793	0.828	0.724	0.779	0.816
	ZL	Sim	0.667	0.647	0.634	0.621	0.602	0.588
		Hyp	0.671	0.652	0.640	0.622	0.603	0.590
7	RTT	Sim	0.545	0.544	0.542	0.627	0.628	0.630
		Hyp	0.558	0.554	0.552	0.636	0.636	0.636
	NF	Sim	0.752	0.810	0.847	0.734	0.801	0.824
		Hyp	0.760	0.808	0.840	0.738	0.790	0.825
	ZL	Sim	0.714	0.692	0.678	0.662	0.640	0.625
		Hyp	0.718	0.698	0.685	0.664	0.643	0.629
9	RTT	Sim	0.615	0.613	0.613	0.700	0.698	0.697
		Hyp	0.618	0.615	0.612	0.692	0.692	0.692
	NF	Sim	0.764	0.819	0.852	0.744	0.808	0.830
		Hyp	0.772	0.817	0.848	0.747	0.798	0.831
	ZL	Sim	0.741	0.720	0.705	0.688	0.664	0.648
		Hyp	0.747	0.726	0.712	0.688	0.667	0.652

The gap between the simulation approach and the presented HQM for calculating the performance measures (a) RTT, (b) NF, and (c) ZL are presented in Chart 3. As shown in this chart, the average gap of the 54 examples for RTT, NF, and ZL is 0.3%, 0, and 0.1%, respectively, and the maximum gap percentage is 2%. Thus, it can be concluded that the simulation has excellent precision and can be used to solve larger examples. The random data are used for generating larger examples. The random numbers generated are between 0 and 50 for the center of each atom, between 1 and 20 for the service rates in customer's location, and between 1 and 15 for the customer's arrival rate. The distance between two atoms is determined by the rectilinear distance between their centers and the coverage radius considered as equal to the average of the distance between the atoms. Chart 4 shows the effect of the increasing travel rate on a) RTT, b) NF and c) ZL. Random numbers between 1 and 12 are generated to set this rate, and each time, two units are added to all travel rates. This procedure is repeated 30 times.

Table 10. Performance measures, RTT, NF, and ZL for demand rates of category 3

μ	Performance measures	Solving method	The travel rate					
			Cat.1			Cat. 2		
			$N=4$	$N=5$	$N=6$	$N=4$	$N=5$	$N=6$
5	RTT	Sim	0.476	0.475	0.475	0.545	0.553	0.555
		Hyp	0.482	0.478	0.476	0.555	0.555	0.555
	NF	Sim	0.686	0.723	0.737	0.682	0.718	0.738
		Hyp	0.692	0.723	0.736	0.687	0.719	0.732
	ZL	Sim	0.239	0.261	0.284	0.213	0.234	0.259
		Hyp	0.241	0.258	0.273	0.213	0.227	0.241
7	RTT	Sim	0.563	0.561	0.560	0.630	0.635	0.639
		Hyp	0.566	0.562	0.559	0.636	0.636	0.636
	NF	Sim	0.691	0.725	0.740	0.686	0.723	0.742
		Hyp	0.698	0.728	0.741	0.692	0.723	0.736
	ZL	Sim	0.271	0.292	0.314	0.239	0.261	0.268
		Hyp	0.276	0.293	0.308	0.239	0.254	0.268
9	RTT	Sim	0.634	0.631	0.626	0.697	0.694	0.691
		Hyp	0.626	0.622	0.620	0.692	0.692	0.692
	NF	Sim	0.695	0.727	0.741	0.690	0.726	0.744
		Hyp	0.702	0.732	0.745	0.695	0.726	0.739
	ZL	Sim	0.292	0.313	0.333	0.256	0.280	0.285
		Hyp	0.299	0.317	0.332	0.256	0.271	0.286

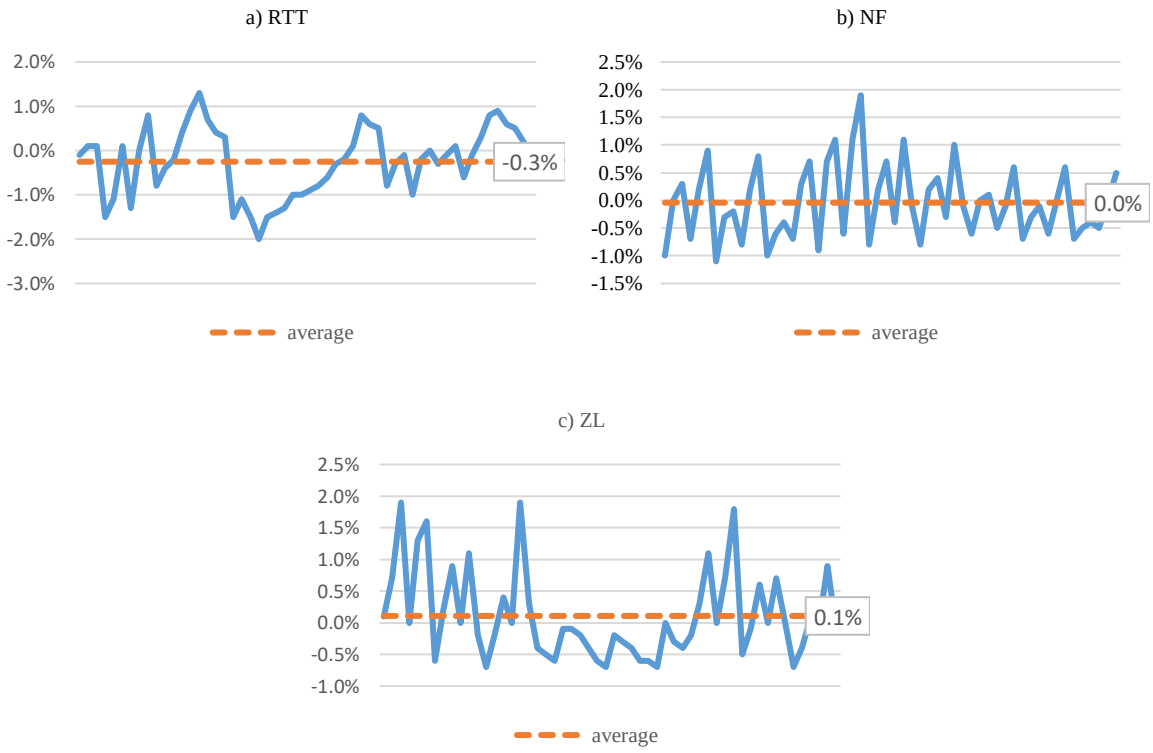

Chart 3. The gap between Simulation & HQM

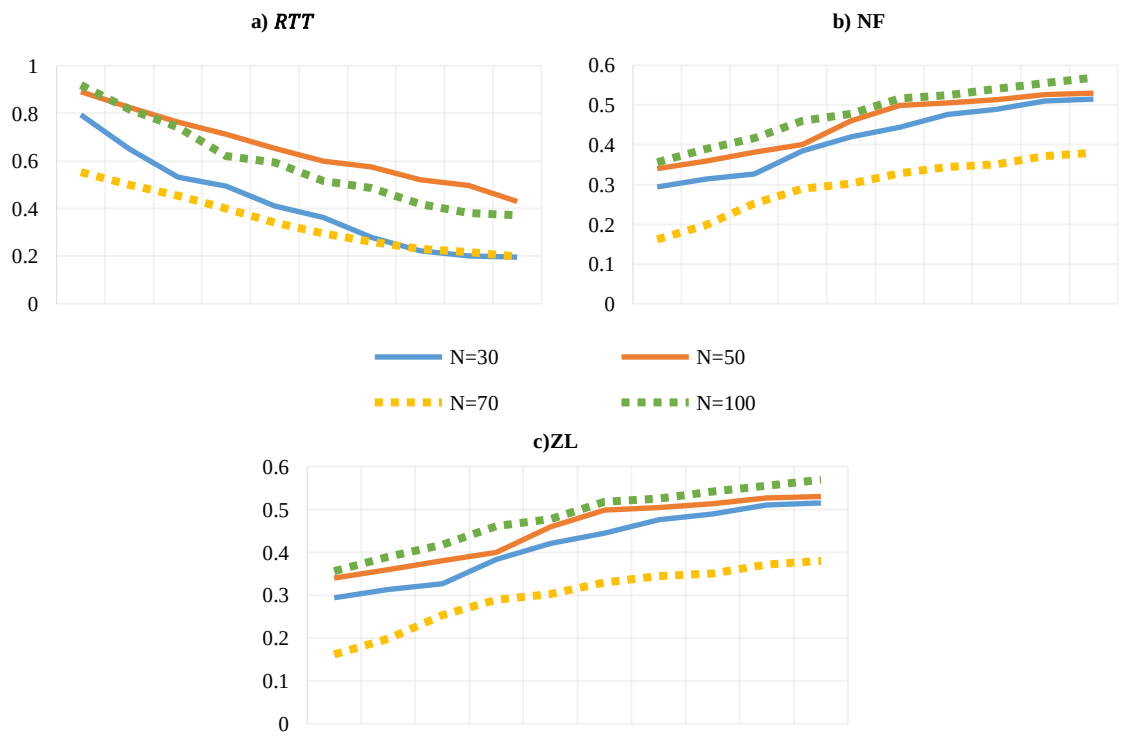

Chart 4. The effect of increasing travel rates on a) RTT, b)NF and c)ZL

6- Case study: Emergency road stations of Red Crescent in Hamedan province

The Red Crescent Society of the Islamic Republic of Iran is an Iranian non-profit organization and member of the International Red Cross and Red Crescent Movement, and it is active in relief and humanitarian activities in the interior of Iran and some cases in other parts of the world. The Iranian Red Crescent has several sub-organizations that cover a wide range of medical, health, educational, and relief services.

In this research, the Red Crescent stations of Hamedan province are studied. Most of the numerical data for the case study is derived from Panahi (2020). The Hamedan province is the 12th province, in terms of the area and the 23rd province, in terms of the population of Iran. The province of Hamadan covers an area of 19493 km^2(1.2% of the total country area), while it holds 2.42% of the total population of Iran. The capital of the Hamedan province is Hamadan city. This province is considered a highway due to its geographical location in the west of the country. Also, the traffic from eight provinces passes through Hamadan province, and this province tolerates 70% of the traffic in the west of the country. In this province, more than 17 road stations of Red Crescent are established and involved in road accidents.

Table 11 shows the travel time between stations. Station 17 is the only mountain station in the area and does not have any road link with other stations. Based on this information, 30 demand points are considered and the demand rates are giving in table 12, which are determined using the number of accidents in each path. The on-scene service time is one hour, and the travel rate from each station to the customer is equivalent to the reverse travel time. According to these data, each route of the network can be analyzed precisely. The coverage radius is 25 minutes, and the travel time from the stations to the demand points that can be covered by them is shown in table 13.

The route between Hamedan to Kangavar is selected to check the validity of the model in the real-world. This route is ranked first in terms of the number of accidents, and 131 accidents in 2016 and 175 accidents in 2017 occurred in this route. In this route, there are five Red Crescent Road stations and eight demand points. The results of the performance criteria of this route are presented in table 14.

Table 11. Travel time between stations

		To														
		2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
From	**1**	26	84	133	120	95	60	106	85	133	57	52	50	23	60	78
	2		84	126	110	88	40	100	78	126	83	45	77	48	104	101
	3			150	135	102	99	125	20	60	54	70	32	62	45	60
	4				45	60	105	102	120	170	168	60	162	153	168	180
	5					63	77	52	134	180	180	65	166	146	170	180
	6						84	76	100	150	160	50	138	116	147	150
	7							67	93	143	120	60	112	81	138	143
	8								120	167	180	55	156	130	165	168
	9									60	73	52	53	80	60	65
	10										97	113	93	122	77	73
	11											106	31	36	21	43
	12												101	75	110	113
	13													27	47	70
	14														56	80
	15															23

Table 12. Demand rates for the case study (numbers should be multiplied by 10^{-5})

Demand point	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
λ_j	1.3	7.1	6.7	1.02	2.3	3.8	2.7	6.7	4.2	4.1
Demand point	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
λ_j	6.8	6.7	0.2	6.8	1.9	1.9	2.6	.38	.69	2.9
Demand point	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**
λ_j	3.1	2.5	1.07	3.9	2.9	1.02	1.3	2.1	0.19	3.9

Table 13. Travel time from the stations to the demand points

Demand point	Station 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	-	-	-	-	-	-	-	-	-	-	-	-	-	15	-	-	-
2	23	-	-	-	-	-	-	-	-	-	-	-	-	23	-	-	-
3	13	13	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
4	-	20	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
5	-	22	-	-	-	-	22	-	-	-	-	-	-	-	-	-	-
6	-	25	-	-	-	-	-	-	-	-	-	25	-	-	-	-	-
7	-	-	-	-	-	-	-	-	-	-	-	12	-	-	-	-	-
8	-	-	-	-	-	25	-	-	-	-	-	25	-	-	-	-	-
9	-	-	-	-	-	-	-	-	15	-	-	-	-	-	-	-	-
10	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	10	-
11	-	-	10	-	-	-	-	-	10	-	-	-	-	-	-	-	-
12	-	-	22	-	-	-	-	-	-	-	-	-	-	-	-	22	-
13	-	-	-	-	-	-	-	-	-	-	-	-	-	-	10	-	-
14	-	-	-	-	-	-	-	-	-	15	-	-	-	-	-	-	-
15	-	-	-	-	-	-	-	-	-	-	15	-	15	-	-	-	-
16	-	-	-	-	-	-	-	-	-	-	18	-	-	-	-	-	-
17	-	-	-	-	-	-	15	-	-	-	-	-	-	-	-	-	-
18	-	-	-	-	-	-	-	15	-	-	-	-	-	-	-	-	-
19	-	-	-	-	20	-	-	-	-	-	-	-	-	-	-	-	-
20	-	-	-	20	-	-	-	-	-	-	-	-	-	-	-	-	-
21	-	-	-	22	22	-	-	-	-	-	-	-	-	-	-	-	-
22	-	-	11	-	-	-	-	-	-	-	-	-	10	-	-	-	-
23	-	-	-	-	25	-	-	25	-	-	-	-	-	-	-	-	-
24	-	-	-	-	-	20	-	-	-	-	-	-	-	-	-	-	-
25	-	-	-	-	-	-	-	-	-	-	15	-	-	-	-	-	-
26	-	18	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
27	-	-	-	-	-	-	-	-	-	-	-	-	13	13	-	-	-
28	-	-	-	-	-	-	-	-	-	-	10	-	-	-	10	-	-
29	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	5
30	-	-	-	-	-	15	-	-	-	-	-	-	-	-	-	-	-

Table 14. Performance measures of the emergency system based on Hamedan-Kangavar route

TT	TR	ZL	RTT	NF
0.5950	0.6269	0.9984	0.3730	0.9884

Number of station	1	2	7	12	14
ρ_i	0.0059	0.0046	0.001	0.0065	0.0089
ρ'_i	0.0017	0.0018	0.0075	0.0024	0.0036
TO_i	0.2954	0.3969	0.4239	0.3772	0.4096

According to table 14, the value of *ZL* and *NF*, the probability of lost customers is very low and most of them are served by their closest servers. The servers are in travel status about 37% of the time based on the value of *RTT*. Thus, the travel time cannot be easily estimated or ignored same as basic HQM. The similar results are obtained by comparing ρ_i and ρ'_i in the table. Therefore, finding a solution to reduce travel time helps to increase the quality of service provided by this emergency system.

In addition to Hamedan-Kangavar route, the entire emergency system of Hamadan province is simulated and the performance criteria $RTT = 0.6955, Zl = 0.9927, NF = 0.9898$ are obtained using simulation. Therefore, almost all customers are served by their closest server; however, the servers are in travel staus about 70% of the time. Thus, a solution to reduce travel time can be very helpful.

7. Conclusion and future research.

Customers are in critical conditions in emergency systems, and it is crucial to provide high-quality service for them. The allocated budget to these systems is usually not enough to cover all customers in the best possible way, and there should always be a balance between the minimum available facilities and the highest possible quality. Therefore, it is necessary to know the dispatching policy and the status of servers at any given moment. HQM is an efficient tool to analyze the different states of a system.

In the studies using the HQM, the service time begins when a server is assigned to a customer and continues until this server returns to its station. This means that the service time includes the travel time to reach a customer, the on-scene service time, and the travel time to return from the customer's location to the server's station. Also, these studies use a service rate for total service time. The on-scene service time is affected by factors such as the severity of the injury, type of equipment, and skill of servers, but travel time is affected by factors such as vehicle type and traffic flow. Therefore, considering one rate for the entire service time would not be appropriate. In this study, different rates are considered for travel time and on-scene time, and as a result, a new state definition is presented for servers. According to this definition, the status of each server can be idle at its station, traveling from its station to the customer or vice versa and serving the customer at its location. By the proposed state definition, new performance measures are defined and calculated precisely using the Gaussian-Elimination method for smaller size examples and approximately using a simulation process for bigger size examples. Of course, the function of the simulation approach was examined, which confirmed that this approach has excellent performance compared to the exact approaches. Finally, by using the real data of Hamedan Red Crescent, the performance of the model was evaluated. The results indicate that the model is working well on real-world issues. In this case study, it is concluded that most customers are served by their closest server, therefore, the value of coverage radius can be considered smaller without affecting the performance of the model. However, one of the main challenges pertains to the emergency system is the large ratio of travel time into the total service time. Therefore, it is not necessary to establish a new emergency station, and it seems that travel time can be reduced by adding some equipped ambulances along the road outside the stations.

Some of the performance measures that can be calculated directly through this new state definition are presented in this study. Providing more performance measures affected by travel time and the distance between customers and servers is recommended for future research.

In this paper, the performance of the presented HQM and the basic model were compared with each other. Although it is not possible to compare these two models in many ways, it is necessary for further studies.

In this study and most of the earlier studies, after completion of service to the customer, the server's work is finished, and it is assumed that the server must first return to its station and then sent it to serve another customer. As is clear from the results of this study, the return time is a significant portion of the total service time. Thus, eliminating this time can help provide better service. Of course, it is noteworthy that the complete removal of the return time is not possible because the server sometimes needs to return to the station to re-stock on supplies and staff to take rest. Thus, providing a strategy that determines whether the server should return to the station and then dispatched to service or dispatched directly to another customer is proposed as a suggestion for future research.

To design an emergency system, the decisions are made at strategic and operational levels. At the strategic level, the number and location of the servers are determined, and at the operational level, the dispatch policies, the coverage area for each server, and the queue length are defined. These two levels depend on each other. For example, the travel time of each server is dependent on its location. In the literature review section, many studies combined the hypercube queuing model with location models. The combination of the proposed hypercube model with location models can be very effective in decision making since this model can evaluate the performance measures more accurately than the basic hypercube model.

As a result of this study, the hypercube model is sensitive to the number of servers, and even by increasing one unit to the number of servers, the number of equilibrium equations might increase significantly. Therefore, solving the system of equations precisely is very time consuming, and providing an approximate or heuristic approach could be greatly useful in solving this model.

References

Ansari, S., McLay, L. A., & Mayorga, M. E. (2017). A maximum expected covering problem for district design. *Transportation Science*, *51*(1), 376-390.

Ansari, S., Yoon, S., and Albert, L. A. (2017). An approximate hypercube model for public service systems with co-located servers and multiple response. *Transportation Research Part E: Logistics and Transportation Review*, 103, 143-157.

Boyaci, B., and Geroliminis, N. (2012). Facility location problem for emergency and on-demand transportation systems. 91th annual meeting of the transportation research board, Washington D.C.

Boyaci, B., and Geroliminis, N. (2014). Hypercube queueing models for emergency response systems. 14th Swiss Transport Research Conference.

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

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

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

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

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

Ghobadi, M., Arkat, J., and Tavakkoli-Moghaddam, R. (2019). Hypercube queuing models in emergency service systems: A state-of-the-art review. *Scientia Iranica*, 26(2), 909-931.

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

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

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

Karimi, A., Gendreau, M., and Verter, V. (2018). Performance approximation of emergency service systems with priorities and partial backups. *Transportation Science*, 52(5), 1235-1252.

Kim, S. H., and Lee, Y. H. (2016). Iterative optimization algorithm with parameter estimation for the ambulance location problem. *Health care management science*, 1-21.

Larson, R. C. (1974). A hypercube queuing model for facility location and redistricting in urban emergency services. *Computers and Operations Research*, 1, 67-95.

Larson, R.C. and Sasanuma, K., (2010). Congestion pricing: A parking queue model. *Journal of industrial and systems engineering*, 4(1), 1-17.

Panahi, P. (2020). Emergency facility location problem considering permanent and temporary stations (case study: Hamadan province), *Unpublished master thesis, University of Kurdistan*

Rajagopalan, H. K., Saydam, C., Setzler, H., and Sharer, E. (2011). Ambulance deployment and shift scheduling: An integrated approach. *Journal of Service Science and Management*, 4(01), 66.

Rodrigues, L. F., Morabito, R., Chiyoshi, F. Y., Iannoni, A. P., and Saydam, C. (2017). Towards hypercube queuing models for dispatch policies with priority in queue and partial backup. *Computers & Operations Research*, 84, 92-105.

Rodrigues, L. F., Morabito, R., Chiyoshi, F. Y., Iannoni, A. P., & Saydam, C. (2018). Analyzing an emergency maintenance system in the agriculture stage of a Brazilian sugarcane mill using an approximate hypercube method. *Computers and Electronics in Agriculture*, 151, 441-452.

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

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

Yazdanparast, R., Hamid, M., Azadeh, A., and Keramati, A. (2018). an intelligent algorithm for optimization of resource allocation problem by considering human error in an emergency department. *Journal of industrial and systems engineering*, 11 (1), 287-309.

Yoon, S., and Albert, L. A. (2017). An expected coverage model with a cutoff priority queue. *Health Care Management Science*, 21(4), 517-533.