International Journal of Industrial Engineering, 31(4), 685-708, 2024

PERFORMANCE EVALUATION OF EMERGENCY MEDICAL SERVICE SYSTEMS WITH MULTIPLE AMBULANCE TYPES AND PATIENT TYPES

Raviarun Arumugaraj Nadar[1,*], J K Jha[1], and Jitesh J. Thakkar[2]

[1]Department of Industrial and Systems Engineering
Indian Institute of Technology Kharagpur
West Bengal, India
[*]Corresponding author's e-mail: ravi1989.06@gmail.com

[2]School of Management
Gati Shakti Vishwavidyalaya,
Vadodara, India

Emergency medical services (EMS) are an important part of the modern healthcare system that tries to provide timely medical care and transportation to patients to reduce morbidity and mortality. Performance evaluation of such EMS systems to determine measures such as mean service rates, dispatch probabilities, busy probabilities, and on-scene times is necessary to design effective and efficient systems. In this paper, we consider an urban EMS system that employs three types of emergency vehicles, including advanced life support (ALS), basic life support (BLS) and first responder vehicle (FRV). We consider two types of patients: type A requires ALS to be dispatched, while type B patients are expected to be served by BLS ambulances. We also consider co-located servers so ambulances of different types can be co-located at the same station. The presence of different types of servers (ambulances) and the patients with different dispatch policies, along with co-located servers, makes it applicable to a more realistic system. We first discuss a modification of the hypercube queueing model for the proposed system and then present an approximate approach for application in large EMS systems. These approaches are compared against a simulation-based model by computing server utilization, service times and on-scene time of ambulances.

Keywords: EMS Planning; Performance Evaluation; HQM; Simulation; Approximation Algorithm.

(Received on March 23, 2023; Accepted on May 27, 2024)

1. INTRODUCTION

Emergency medical services (EMS) play a key role in modern healthcare systems by providing out-of-hospital services and transportation for patients in need of urgent care. EMS systems aim to respond quickly to ensure that patients receive care at the earliest. Planning of EMS systems requires addressing various strategic, tactical, and operational level planning problems such as the location of ambulance stations and ambulances, dispatching policies of ambulances, crew allocation and scheduling, and routing of ambulances (Aringhieri *et al*., 2017). Both long-term and short-term planning of EMS systems require the prediction of equilibrium behavior and accurate evaluation of system performance. The need to evaluate system performance motivates the development of models for accurate and efficient estimation of busy probability, availability, and dispatch probability of ambulances.

Assessing the effectiveness of public safety systems like fire and EMS systems is crucial to guarantee efficient services and timely care to those who need it. Response time, which is the time taken to respond to an emergency call, is one of the major performance indicators of the performance of EMS systems (Mendonca and Morabito, 2001). However, coverage and survival probability-based measures are most commonly used to evaluate planning decisions in EMS systems literature (McLay and Mayorga, 2010). Planning decisions require evaluating the various possible configurations of the system to find the most optimal system. Therefore, determining the performance measures of an emergency services system is critical for evaluating the different configurations of the system. Some of the key performance indicators of EMS systems include the proportion of calls that can be served in a given horizon, the probability that a call is lost, the mean service time taken to serve different call types, and the mean response time required to reach patient locations (Beojone and de Souza, 2017).

Computing the common performance measures, such as coverage and survival probability, requires an estimation of the probability that a specific ambulance responds to any arriving call. Dispatch probability represents the probability that a specific ambulance station will dispatch an ambulance to serve an emergency call from a given demand zone. Dispatch probability is a function of the arrival rate of emergency calls, the number of busy ambulances, and the preference order of

the ambulance stations. Thus, dispatch probabilities provide a means to calculate various performance measures using patient locations, location of ambulances, and conditional probabilities of ambulance availability. Busy probability represents the probability that an ambulance will be busy when an emergency call arrives at the ambulance station. The availability of ambulances estimates the probability that an ambulance will be available or free when an emergency call arrives. Thus, busy probability and availability of ambulances are complementary and calculating one enables us to determine the other value. Calculating the busy probability and dispatch probability of ambulances enables us to determine other important performance indicators of the ambulance system.

Various approaches have been proposed and applied in the literature to estimate the dispatch probability of ambulances. These approaches include the hypercube queueing model (HQM), simulation and approximation-based approaches. HQM is an analytical approach developed to evaluate performance measures of public safety systems (Larson, 1974; Larson, 1975). Simulation-based approaches try to develop a representation of a system and analyze various what-if scenarios, thus enabling policy decisions. The need for approximate approaches arises due to high computational power and time requirements, even for reasonable problem size in the case of HQM and simulation approaches (Karimi *et al*., 2018). A major assumption in most studies is that all ambulances are identical and can serve all calls, whereas, in actual systems, different types of ambulances are employed and usually serve different patient types. This means the dispatch policies are dependent on the type of ambulances available at a station.

Additionally, the majority of existing models in the field neglect factors such as multiple vehicles stationed at the same station (co-location) and the scenario where multiple vehicles are dispatched to handle the same emergency call. Instead, these models typically assume that vehicles are positioned at separate locations and that only one vehicle is sent to handle each call. This study aims to bridge this gap in the literature by presenting a model that takes these factors into account. The proposed model can assist public emergency services decision-makers in making decisions related to determining vehicle placement and designing emergency districts.

In this paper, we focus on an urban EMS system that employs three types of ambulances, including advanced life support (ALS), basic life support (BLS) and first responder vehicle (FRV). We consider two types of patients: type A requires ALS to be dispatched, while type B patients are expected to be served by BLS ambulances. We also consider co-located servers, i.e., multiple ambulances can be located at a single station, and even ambulances of different types can be co-located. The presence of different types of servers (ambulances) and the patients with different dispatch policies, along with co-located servers, makes it applicable to a more realistic system. First, we discuss a hypercube queueing model for the proposed system that is adapted for the proposed system, considering multiple ambulances located at the same stations and different service times based on the patient type and ambulance type.

The organization of this article is as follows. Section 2 presents an overview of the existing literature related to the proposed problem. Section 3 presents a brief description of the system under consideration, explaining the assumptions and dispatch policies. Section 4 outlines the HQM-based approach. Section 5 details the proposed approximation-based approach for performance evaluation. The computational experiments and the results obtained are presented in Section 6. Finally, we outline the conclusions of the work in Section 7.

2. REVIEW OF LITERATURE

Computing the key performance measures requires estimating the probability that a specific ambulance responds to any arriving call from a demand zone. This parameter is called dispatch probability, which depends on the arrival rate of calls, service rate, number of ambulances busy or available, the preference order of ambulances, and the busy probability of ambulances. Dispatch probabilities provide a means to calculate other performance measures using the location of the received call, the location of ambulances, and the conditional probabilities of ambulance availability. Various approaches have been proposed in the literature to estimate dispatch probabilities, including HQM (Geroliminis *et al*., 2011; Iannoni *et al*., 2008), simulation (Lee *et al*., 2012; McCormack *et al*., 2015), and approximate approaches (Saydam and Aytug, 2003). Larson (1974) introduced HQM to evaluate the performance of mobile units involved in emergency services, including the fire service, police, and ambulances. Larson (1975) developed an approximate HQM that uses an iterative procedure to provide solutions for systems with a large number of ambulances. Chelst and Barlach (1981) extend the HQM to incorporate multiple simultaneous dispatches and provide both exact and approximate approaches to the problem. They evaluate performance measures specific to multiple dispatches of ambulances, such as paired travel times and delays between the arrival of the first unit and the backup unit. Jarvis (1985) proposes an approximation algorithm for a multi-server loss system with individual servers that are independent and multiple patient types while taking into consideration the server-dependent nature of service times. Brandeau and Larson (1986) extend HQM to incorporate varying service times based on server location and better travel time estimation approaches.

Similarly, Goldberg and Szidarovszky (1991) present an approximate heuristic approach for computing vehicle busy probabilities while considering varying service times. Burwell *et al*. (1993) developed and applied a hypercube-based model

for emergency systems with server units co-located at a station, thus causing ties in the dispatch of servers. Saydam *et al*. (1994) compare the performance of various coverage models and suggest that hypercube models should be used along with coverage models for conducting post-optimality analysis.

Mendonça and Morabito (2001) apply a modified HQM to an ambulance deployment scenario on a highway with partial backups. Galvao and Morabito (2008) review various HQM extensions and present an extension to MEXCLP and MALP that embeds HQM into these location models. Takeda *et al*. (2007) investigate the impact of decentralizing the ambulance services and also the impact of additional ambulances on the system by applying the HQM to an EMS in Brazil. Iannoni and Morabito (2007) present a hypercube model that takes into account various factors such as partial backup, distinct servers and call types, multiple dispatches, and different dispatch policies in an EMS located on a highway. Atkinson *et al*. (2006) present two heuristic approaches for HQM that can be used to evaluate realistic emergency systems with customer-dependent service time and priority in requests. Atkinson *et al*. (2008) present a 3n hypercube model that considers the servers responding to a call from their primary and secondary locations, and they present two heuristics for the problem. Iannoni *et al*. (2008) developed an HQM that considers partial backup and allows for dispatching multiple ambulances for an emergency call. They embed the developed HQM model in a GA-based framework for districting in an EMS system. Morabito *et al*. (2008) evaluate the impact of assuming homogenous servers in an HQM and conclude that the assumption of homogeneity leads to significant inaccuracies in predicting operational performance indicators of non-homogeneous systems.

Geroliminis *et al*. (2009) present a hypercube model that considers the spatial and temporal variation in demand and service rates between different servers. Budge *et al*. (2009) explore a performance evaluation model to consider multiple ambulances at a single station (co-location) and allow them to calculate station-specific busy probabilities. Knight *et al*. (2012) proposed an iterative approach that initially assumes a busy probability for all ambulances and then solves the location problem to get a new solution in the first iteration. The busy probability is then revised using the new solution found in the previous iteration. This process is repeated iteratively until convergence is reached. Boyacı and Geroliminis (2015) present a model that considers the Spatiotemporal uncertainty in demand and service time while also considering partial backup of ambulances. Ansari *et al*. (2017) present an approximate hypercube spatial queueing model that considers the co-location of ambulances at stations and accounts for multiple dispatches of ambulances. Beojone and de Souza (2017) present an HQM that considers queue priorities, i.e., the priority of one patient type over another patient based on their condition. Karimi *et al*. (2018) present a performance evaluation model that considers partial backups and variations in queue capacities. They evaluate models with both zero and infinite queue capacities and allow service times to vary based on call priority and both customer and server locations.

Rodrigues *et al*. (2018) present an approximate AHQ-based method for emergency maintenance systems in the agricultural sector in Brazil, where they consider different service rates along with prioritized queues and partial backup. Yoon and Albert (2018) proposed a spatial approximation model based on the Hypercube approach for a system with cut-off priority queues where a set of servers become reserved for calls with higher priority as soon as the number of ambulances available reaches a prespecified cut-off limit. The proposed model estimates the performance measures of the system where the ambulance queues are thus prioritized under congestion. Beojone *et al*. (2021) propose a model that accounts for dedicated servers to serve only patients of a specific type (or criticality) and co-located servers in an HQM. Liu *et al*. (2021) present a cooperative HQM that allows multiple simultaneous dispatches for a single emergency incident.

From the above discussion, we can observe that many researchers have used hypercube, approximation, and simulation models to estimate the busy probability of ambulances. Based on the literature review, we also observe that the performance evaluation of systems with multiple vehicle types and patient types with hierarchy and priority between them has not been addressed. Similarly, the possibility of variation of service time based on server location, patient location, and type of server utilized has not been considered in the literature. Since hypercube queueing models for large urban EMS systems are computationally difficult to solve, approximate methods to address these issues need to be developed. Therefore, we propose an iterative approximate approach to estimate the busy probability, assuming each station as M/M/s queueing system. This is similar to Knight *et al*. (2012) as they use the M/M/s approximation for each ambulance station. The M/M/s queuing approximation is used to obtain the estimate of the busy probability for the given arrival rate, service rate and number of ambulances at each iteration. However, Knight *et al*. (2012) get a single estimate of the busy probability for each configuration of the ambulance system. In our algorithm, for a given configuration of the EMS, the approximation approach is solved iteratively to obtain a better estimate for busy probabilities. This is necessary since we consider multiple ambulance types and partial backup, which makes it difficult to obtain a single busy probability estimate directly for each solution. Integrating the proposed approximate method within an ambulance location model can better estimate the server-level busy probability and, thus, the expected performance of a given configuration of ambulance locations.

3. DESCRIPTION OF THE SYSTEM

3.1 Assumptions

We wish to develop a performance evaluation approach for the system described above that estimates the dispatch probability of ambulances from different stations to each zone and the busy probability of ambulances at each station. Some key assumptions are as follows.

(i) The arrival rate of calls from each demand zone is assumed to follow the Poisson distribution and is independent of other zones.
(ii) All calls require transportation. Therefore, even when an FRV (non-transporting) ambulance is sent to a patient location, it has to be followed by a BLS.
(iii) The average travel time for each ambulance type between all pairs of demand zones and ambulance stations is known.
(iv) The ambulances are located at pre-determined (base) stations and return to the same base after serving a call.
(v) The number of ambulances of each type located at each station is known.
(vi) The order of preference for dispatching an ambulance to any zone is fixed and known. This preference order is represented using a rank of ambulances.

3.2 Dispatch policies

The dispatch policies for different types of ambulances considered for the two types of patients are as follows.

3.2.1 For type A patients (life-threatening)

(a). An ALS ambulance from the preferred primary station (with rank 1) is sent if available.
(b). If an ALS ambulance is not available at a higher-ranked station, an ALS from the next preferred station is dispatched.
(c). Only if an ALS ambulance is unavailable at any preferred stations, a BLS ambulance is dispatched in the same order of station preferences as ALS.
(d). If all ALS and BLS ambulances at preferred stations are busy, the call is lost.

3.2.2 For type B patients (non-life-threatening)

(a). If a BLS ambulance is available at a primary station, it is dispatched.
(b). If a BLS ambulance is unavailable at a higher-ranked station, then a BLS from the next-ranked station is dispatched.
(c). If a BLS ambulance is unavailable at any of the preferred stations, an FRV from the nearest station is sent. The preference order of stations for FRV is also the same as for ALS and BLS.
(d). If an FRV is dispatched, a BLS is dispatched to the location as soon as one becomes available. Since FRV cannot provide transport, a BLS has to be dispatched.
(e). If no BLS or FRV is available when a call arrives, the call is lost.

In the subsequent sections, we describe the three approaches for the performance evaluation of EMS systems and discuss how they can be implemented for the proposed system.

4. HYPERCUBE QUEUEING MODEL-BASED APPROACH

The hypercube model introduced by Larson (1974) models the emergency response system as a spatially distributed queueing system. The HQM, in combination with Markovian analysis approximations, has been one of the most effective approaches to describe emergency systems (Larson, 1974, 1975; Larson and Odoni, 1981). Some of the major advantages of the HQM include the ability to incorporate uncertainty related to EMS systems and retain the identity of each server while considering the cooperation within servers. The HQM represents each server individually by expanding the state space of the multi-server systems, which allows to incorporate complex dispatch policies of ambulances. The state space equations enable calculating the probabilities of different states of the system at equilibrium by solving a linear system of O(2N) equations. These steady-

state probabilities can then be used to calculate various important performance measures related to the system, such as mean response times, server utilizations, number of dispatches from an ambulance located in any region, and other similar measures.

The original HQM has been extended by various researchers (Larson, 1975; Halpern, 1977; Chelst and Barlach, 1981; Jarvis, 1985; Burwell *et al*., 1993; Mendonca and Morabito, 2001). These extensions relax various limiting assumptions of the original HQM and improve the overall computational efficiency of the model in evaluating emergency response systems. For example, the extension presented by Chelst and Barlach (1981) allows for simultaneous multiple identical dispatches of patrol vehicles to the same zone. Whereas the extension by Mendonça and Morabito (2001) allows not only multiple dispatches of ambulances but also considers partial backup, where only a subset of all servers can be used as a backup for each zone. Other researchers have combined the hypercube model with optimization techniques, as seen in the works of Batta *et al*. (1989), Saydam and Aytug (2003), Chiyoshi *et al*. (2003), and Galva˜o *et al*. (2005). Applications of the hypercube model in urban emergency medical services in the US can be seen in the works of Larson and Odoni (1981), Chelst and Barlach (1981), Brandeau and Larson (1986), Burwell *et al*. (1993), and Sacks and Grief (1994). The hypercube model has been applied more recently to analyze deployment in the event of terrorist attacks and other major emergencies (Larson, 2004). In Brazil, applications of the hypercube model can be found in urban emergency systems (e.g., Takeda *et al*., 2007) and highways (Mendonca and Morabito, 2001; Iannoni *et al*., 2008).

The HQM assumes that the entire region being analyzed is divided into small units called atoms (zones). Ambulances (servers) are considered to be located at stations and dispatched when a call arrives at a station. Each demand zone is associated with a preference list that assigns priority orders for stations representing the order in which the ambulances are dispatched to that specific demand zone. Generally, the nearest server is given the first preference for an atom, and other ambulances are ranked according to the time taken to reach that atom. The call arrival rate and service time for each atom are assumed to be known. The hypercube model expands the simple M/M/m/m queueing system using the state-space description such that each server is represented individually. In a typical HQM, ambulances are considered to have two states. An ambulance could be either busy, represented by (1), or available (free), represented by (0). For example, {010} indicates that there are three ambulances, and 'ambulance 2' is busy serving a patient while 'ambulance 1' and 'ambulance 3' are available. So, when there are m ambulances in a system, the representation is an m-dimensional hypercube. Each state is associated with a linear equation that balances the entering probability and leaving probability of that state. Therefore, there are 2m linear equations associated with an m-ambulance system, and solving these simultaneous equations gives the probability of each state. These probabilities are then used to calculate server-level workload, mean service time, and mean response time.

An example of the states of a simple HQM for an EMS system with two servers and one demand zone is shown in Figure 1. The state represented by {00} indicates both ambulances are idle and available, while {11} indicates that both ambulances are busy and not available. In the system shown in Figure 1, λ represents the arrival rate of calls, whereas μ_1 and μ_2 represent the service rate of ambulance 1 and ambulance 2, respectively. The state-space equations for the two-server system in Figure 1 are given by equations (1) to (5).

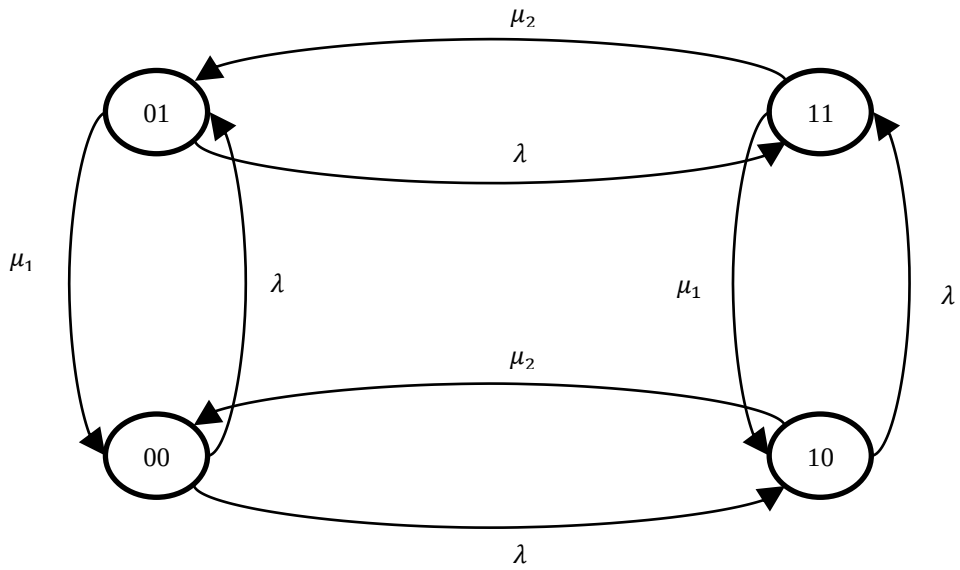

Figure 1. HQM states for two-server system

$$\lambda P\{00\} = \mu_1 P\{01\} + \mu_2 P\{10\} \quad (1)$$

$$(\lambda + \mu_1)P\{01\} = \lambda P\{00\} + \mu_2 P\{11\} \quad (2)$$

$$(\lambda + \mu_2)P\{10\} = \lambda P\{00\} + \mu_1 P\{11\} \quad (3)$$

$$(\mu_1 + \mu_2)P\{11\} = \lambda P\{01\} + \lambda P\{10\} \quad (4)$$

$$P\{00\} + P\{01\} + P\{10\} + P\{11\} = 1 \quad (5)$$

The value of state-space probabilities is obtained by solving the simultaneous equations (1) to (5). The state-space probabilities are used to calculate the busy probability of each ambulance. The value of dispatch probability d_{ij}, the probability that an ambulance *i* is dispatched to attend a call from zone *j*, is calculated using equation (6).

$$d_{ij} = \frac{\lambda_j}{(1 - P_{loss})\lambda} \sum_{B \in E_{ij}} P_B, \quad (6)$$

where λ_j is the arrival rate of the calls from zone *j*, P_B is the probability that a server is busy (busy probability), and P_{loss} is the sum of all probabilities where all preferred servers are busy (called loss probability).

4.1 Dispatch policies

The HQM described above has some limitations when applied to a realistic EMS system. To apply the HQM for the proposed system, we make two key modifications, which are described subsequently.

4.1.1 Layering

To account for multiple ambulance types with ALS as dedicated servers and BLS for type B patients and the different types of patients, we apply the process of layering. Layering refers to representing a demand zone or ambulance station by multiple zones or stations (Beojone *et al*., 2021). For example, a demand zone from where two types of calls (type A and type B) arrive can be considered as two separate demand zones (say zone 1 and zone 2) such that zone 1 is associated with only type A calls and zone 2 is associated with only type B calls. Similarly, a station with three types of ambulances can be replaced by three stations with one type of ambulance at each location. Although layering increases the number of demand zones and stations in the problem, it allows us to consider different arrival rates for different call types and service times for different types of ambulances.

4.1.2 3^n hypercube model

While the general HQM presented above assumes that a server has only two states, i.e., it is either busy or available, there are situations where it is necessary to consider more than two states. For the system under consideration, there are three possible states of BLS ambulances. The possible states for all three ambulance types are explained below.

(i) ALS: The server is either busy serving a type A call or free, indicated by $\{1\}$ or $\{0\}$, respectively.

(ii) A BLS ambulance can be either free, represented by $\{0\}$ or serving a type A or type B call from a zone, indicated by $\{1\}$. Additionally, BLS can be used for the transportation of a patient that has been served by FRV, represented by $\{2\}$. This allows for the consideration of different service times based on the type of patient the ambulance serves.

(iii) FRV ambulances are either free or busy serving a call that could not be served by any of the BLS ambulances.

For example, consider a system with three ambulances, one ambulance for each type of ALS, BLS and FRV. Let $\{0,0,0\}$ represent the idle state for all three ambulances, then $\{0,1,0\}$ represents that BLS is busy serving a type A or type B call. Similarly, $\{0,2,0\}$ represents that BLS is busy serving a type B call where an FRV was already dispatched. In this case, the service time will be less because FRV has already provided emergency care, and there is only a need for transportation by the BLS. Therefore, if there are m ALS type of ambulances, n BLS type of ambulances, and o FRV type of ambulances, then the total number of possible states is given by $2^n\ 3^m\ 2^o$, i.e. $2^{m+o}\ 3^n$. Also, we assume that a BLS will be sent to the location where the FRV is serving as soon as a preferred BLS becomes available, and the FRV will become free immediately. Using the concept of layering and the 3n hypercube model, we can adapt the HQM to be applied to the proposed problem.

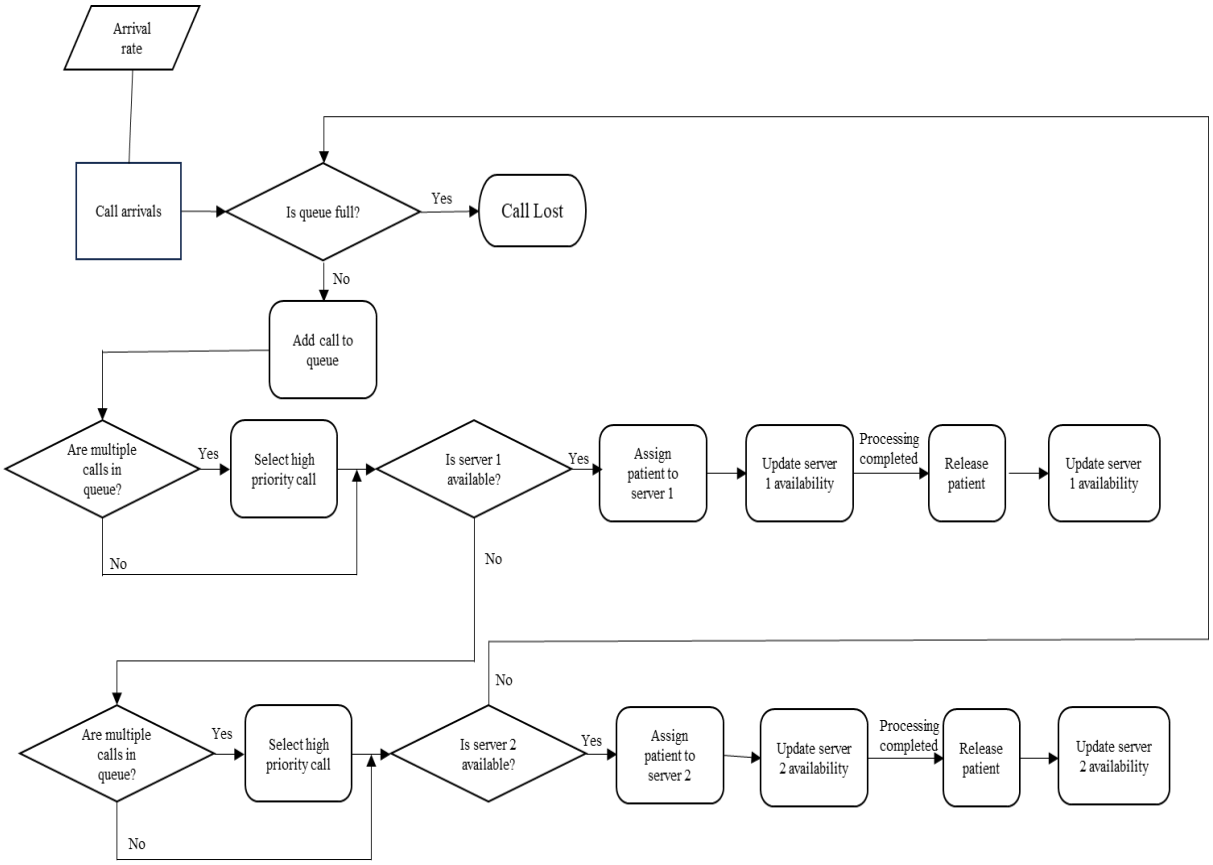

Figure 2. A representative simulation model using Simulink for the system with two ambulances and one demand zone

4.2 Simulation approach

The analytical HQM approach presented above makes various assumptions about the system. A simulation model can be built to represent the system to compare and validate the HQM-based approach (Goldberg *et al*., 1990; de Souza *et al*., 2015). For example, we consider that BLS is assigned to a call served by FRV as soon as it is available, and the FRV is freed. A simulation model can help validate the impact of the assumption on the performance measures. Simulation models are developed for healthcare systems to estimate and optimize the system, especially in emergency departments (Gokalp, 2021; Wang *et al*., 2024; Bang *et al*., 2023)). Therefore, we build a discrete event simulation model to represent the proposed system using 'SimEvents' available in the Simulink package of MATLAB software.

The simulation model was built to reflect the arrival of calls from each zone, and calls of each type were assigned a unique arrival rate. The simulation model also allows calls arriving from different zones belonging to each distinct type to have their unique service time distribution. This is not possible using the HQM, as it assumes each station has a single service time irrespective of call locations. The simulation model can also incorporate individual travel times for each call type. A warm-up period is allowed to account for the transient nature of outputs in the initial period of the simulation run. Similarly, the simulation run length was set as 50000 hours to allow the model to provide more reliable results. An example of the simulation model for a small system with two ambulances (one ALS and one BLS) located at the same station and one demand zone is shown in Figure 2.

5. APPROXIMATE APPROACH

Although the HQM and simulation methodologies discussed in the preceding sections are helpful in evaluating the performance of an EMS system, both of these methods have some serious drawbacks. For real-world systems with a large number of ambulances, the analytical HQM approach necessitates solving a very large number of linear equations. When different types of ambulances and patients are taken into account, the assumption that service time is independent of the type of call used in conventional HQM is severely limiting. The model becomes considerably more challenging to solve when the

dimensions of the HQM are increased from 2^n to 3^n, as shown in Section 4, to account for varied service times. For example, in the case of the 2^n hypercube model, the number of equations required to represent a system with 15 ambulances will be 2^{15}, i.e., 32768. However, if each ambulance is allowed to take three states, the number of equations required to solve for 15 ambulances becomes 3^{15}, i.e. 14348907.

The simulation model allows easily incorporating different service times for calls from different zones and calls belonging to different patient types. However, a separate simulation model is required to be built to evaluate each configuration, which can be a tedious task. Ambulance location problems require evaluating a large number of configurations of different ambulance locations to find an optimal solution, even for medium-sized problems. Another limitation of the HQM and simulation approaches is that the temporal variation in demand and service time cannot be explicitly considered since a separate model has to be developed for each period. The approximate approach can provide an efficient way to obtain quick but reliable estimates for performance measures.

The simulation model allows us to include different service times for calls coming from different zones and calls of different patient types. However, separate simulation models are required to evaluate different configurations of a system. Even for medium-sized problems, evaluating a large number of configurations of various ambulance locations is necessary to determine the optimal locations. The inability to explicitly consider the temporal change in demand and service time due to the need to create a separate model for each period is another drawback of HQM and simulation approaches. The approximate approach can be an effective way to get reliable performance measure estimates quickly. Therefore, an approximate approach can be highly useful for large-scale systems, saving time and computational power with very little loss in the inaccuracy of results. Another advantage of the approximate approach is that it can be easily adapted to account for temporal variations. We describe an approximate approach that uses a queueing-based approximation to estimate the dispatch probability and availability of ambulances in a system.

5.1 Simulation approach

Consider a system with identical ambulances located at different stations and serving a predefined set of demand zones with all calls being of the same patient type. Assume that the arrival rate of emergency calls arriving at each station from each zone is known and that the mean service rate of ambulances at each station is known. Then, each station can be represented by an M/M/m queuing system with a mean arrival rate and service rate where m is the number of ambulances located at the station. But in the proposed problem, we consider a system with heterogeneous ambulance types located at each station that cater to heterogeneous patient types. Therefore, we apply the layering as explained in section 4.1 to first divide each station into multiple stations having only a single ambulance type located. The arrival rate is then determined based on the preference order of ambulance stations assigned for a given demand zone and priority between ambulance types for each call type.

To account for the priority between ambulance types and different stations, we iteratively adjust the estimate using the loss probability of other ambulance stations. We start by assuming the ambulances are independent of each other and obtain an estimate of the dispatch probability of ambulances. This dispatch probability is then used to calculate the probability of lost calls for each station, which is then used to obtain a better estimate of dispatch probability. The loss probability calculated provides a way to obtain the calls arriving at a lower-preference station or an ambulance type with lower priority for a call type. The dispatch probability thus obtained is again used to update the loss probability of calls for each station. The complete process of determining dispatch probabilities, followed by loss probabilities and then updating dispatch probabilities, is repeated until convergence is reached. Although this procedure does not guarantee convergence, it is necessary to limit the number of iterations based on the level of accuracy required. However, our computational experiments show that the dispatch probabilities converge within very few iterations in most cases. The summary of the notation used to represent different sets and key parameters is listed as follows.

I	Set of demand zones, $i \in I$
J	Set of ambulance stations, $j \in J$
T	Set of periods, $t \in T$
R	Set of the rank of ambulance stations, $r \in R$
K	Set of types of ambulances, $k \in K = \{\text{ALS}, \text{BLS}, \text{FRV}\}$
L	Set of types of patients representing different types of calls, $l \in L = \{\text{A}, \text{B}\}$
λ_{it}^{l}	Average arrival rate of call type *l* received from demand zone *i* during period *t*
τ_{ijt}^{l}	Mean service time required to serve a call of type *l* from demand zone *i* using an ambulance from station *j* during period *t*
y_{jt}^{k}	Number of ambulances of type *k* allocated to station *j* during period *t*

d_{ijrt}^{kl}	Dispatch probability of ambulance type *k* from station *j* with rank *r* to serve call type *l* from zone *i* during period *t*
δ_{ijrt}	1, if station *j* is assigned rank *r* for demand zone *i* during period *t*, 0, otherwise
π_{jt}^{k}	Probability that all ambulances of type *k* are busy at station *j* during period *t*
Λ_{jt}^{k}	Average arrival rate of calls associated with ambulance type *k* at station *j* during period *t*
T_{jt}^{k}	Mean service time associated with ambulance type *k* at station *j* during period *t*
M_{jt}^{k}	Average service rate associated with ambulance type *k* at station *j* during period *t*

The general procedure proposed for estimating the dispatch probabilities and busy probabilities of ambulances of any type can be described using the following sequence of steps.

Step 1: Initialise the iteration counter $i = 0$ and $\varepsilon =$ tolerance level required.
Step 2: Set the busy probability of ambulances $\pi_{jt}^{k(0)} = 0, \forall j, t$ and calculate $d_{ijrt}^{k,l(0)}$.
Step 3: Estimate arrival rate $\Lambda_{jt}^{k(0)}$, service rate $T_{jt}^{k(0)}$ and $M_{jt}^{k(0)}$ using the dispatch probability calculated in Step 2.
Step 4: Using the values from Step 3, calculate $\rho_{jt}^{k(0)} = \frac{\Lambda_{jt}^{k(0)}}{M_{jt}^{k(0)}}, \forall j, t.$
Step 5: Update the value of busy probability to obtain $\pi_{jt}^{k(1)}$using $\rho_{jt}^{k(1)}$from Step 4.
Step 6: Set $i = i + 1$.
Step 7: Using the estimated value of π_{jt}^{k} for each station, update $d_{ijrt}^{k,A(i)}$.
Step 8: Update $\rho_{jt}^{k(i)}$, $\Lambda_{jt}^{k(i)}$ and $M_{jt}^{k(i)}$ using $d_{ijrt}^{k,l(i)}$ obtained in Step 7.
Step 9: Calculate $\pi_{jt}^{k(i)}$ using updated values calculated in Step 8.
Step 10: If $\left|\pi_{jt}^{k(i)} - \pi_{jt}^{k(i-1)}\right| < \varepsilon,\ \forall j, t$, stop. Otherwise, go to Step 6.

In the above procedure, the probability that an emergency call arriving at a station is lost, i.e., loss probability due to all the ambulances at the station being busy, needs to be calculated in Steps 5 and 9. This busy probability can be approximated by considering each station as an *M*/*M*/*c* queueing system. Thus, loss probability can be obtained using the Erlang-B formula in equation (7).

$$\pi_{jt}^{k} = \frac{\frac{\rho_{jt}^{y_{jt}^{k}}}{y_{jt}^{k}!}}{\sum_{a=0}^{y_{jt}^{k}} \frac{(\rho_{jt}^{k})^{a}}{a!}} \qquad \forall\, j, k, t, \tag{7}$$

where ρ_{jt}^{k} is the server utilization, expressed as the ratio of arrival rate and service rate for a station, and is given by equation (8).

$$\rho_{jt}^{k} = \frac{\Lambda_{jt}^{k}}{M_{jt}^{k}} \qquad \forall\, j, k, t \tag{8}$$

Steps (6)-(10) in the above procedure can be applied iteratively to update the value of d_{ijrt}^{kl} every time a new value of π_{jt}^{ALS} is estimated.

5.2 Adapting the proposed approach for different ambulance types

The procedure presented in Section 5.1 is a general approach that can be adapted to different ambulance types by modifying the equations used to calculate the busy probability (π_{jt}^{k}), mean arrival rate (Λ_{jt}^{k}), and mean service rate (M_{jt}^{k}). In this section, we discuss how the proposed approximate approach can be adapted for three different types of ambulances discussed in Section 3. As ALS is considered as dedicated ambulance and the first preference to serve life-threatening type A calls, the

dispatch probability of ALS ambulances is independent of other types of ambulances. However, BLS is a general-purpose ambulance that can be used to serve type B calls, preferably while also being dispatched for type A patients in case of unavailability of ALS ambulances. Hence, for BLS ambulances, the arrival rate and dispatch probability also depend on the busy probability of ALS ambulances.

Similarly, FRV is considered a backup ambulance, dispatched only if all BLS ambulances are busy when a type B call arrives. Therefore, calculating dispatch probability for FRV requires the busy probability of BLS ambulances. Due to this hierarchy between different ambulance types, we first estimate the dispatch probability of ALS and then use this to calculate the dispatch probability for BLS, which is then used to calculate the dispatch probability for FRV. For this purpose, we first apply the concept of layering and separate each ambulance type at a station as an individual station.

5.2.1 Estimating busy probability for ALS

For estimating the busy probability of ALS in Step 2, we begin by assuming that the busy probability of all ALS ambulances located at all ambulance stations is zero, i.e. $\pi_{jt}^{ALS} = 0$. This assumption implies that all demand will be served entirely by ambulances from the station with the highest preference (rank 1) since no demand is lost from any station. The proportion of calls served by the station with rank 1 is then equal to 1, and it is 0 for all other stations, as shown in equation (9). ALS is considered a dedicated server and can only serve type A calls, i.e. $d_{ijrt}^{ALS,B} = 0$ for all stations.

$$\left.\begin{array}{l} d_{ij1t}^{ALS,A} = 1, \text{ if } \delta_{ij1t} = 1 \\ d_{ijrt}^{ALS,A} = 0, \text{ otherwise} \end{array}\right\} \qquad \forall i, j, r, t \qquad (9)$$

Using the value for d_{ij1t}^{ALS} from equation (9), the average arrival rate and mean service time for ALS ambulances at each station can be obtained from equations (10) and (11), respectively. Then, the average service rate can be given by equation (12). Equations (10), (11), and (12) are used in Steps 3 and 7 of the procedure to estimate the arrival rate and service rate of ambulances.

$$\Lambda_{jt}^{ALS} = \sum_{i \epsilon I} \sum_{r \epsilon R} d_{ijrt}^{ALS,A} \lambda_{it}^{A} \qquad \forall j, t \qquad (10)$$

$$T_{jt}^{ALS} = \frac{\sum_{i \epsilon I} \sum_{r \epsilon R} d_{ijrt}^{ALS,A} \lambda_{it}^{A} \tau_{ijt}^{A}}{\sum_{i \epsilon I} \sum_{r \epsilon R} d_{ijrt}^{ALS,A} \lambda_{it}^{A}} \qquad \forall j, t \qquad (11)$$

$$M_{jt}^{ALS} = \frac{1}{T_{jt}^{ALS}} \qquad \forall j, t \qquad (12)$$

Using the estimated value of π_{jt}^{ALS}, we can update the dispatch probability $d_{ijrt}^{ALS,A}$ for each pair of zones and stations in Step 9 using equation (13).

$$d_{ijrt}^{ALS,A} = \prod_{q \in R | q < r} \left(\sum_{p \in J | p \neq j} \delta_{ipqt} \pi_{pt}^{ALS} \right) \left(1 - \pi_{jt}^{ALS}\right) \qquad \forall i, j, r, t \qquad (13)$$

5.2.2 Estimating busy probability for BLS

As BLS ambulances are considered general-purpose ambulances, we need to consider the arrival rate for both type A and type B calls to calculate the dispatch probability for BLS ambulances. Similar to the procedure for ALS, we first assume $\pi_{jt}^{BLS} = 0$ and use equations (14) and (15) to calculate an initial estimate for the dispatch probability for BLS ambulances as required in Step 2.

$$\left.\begin{array}{l} d_{ij1t}^{BLS,A} = \prod_{q \in R} \left(\sum_{p \in J} \delta_{ipqt} \pi_{pt}^{ALS} \right), \text{ if } \delta_{ij1t} = 1 \\ d_{ijrt}^{BLS,A} = 0, \text{ otherwise} \end{array}\right\} \qquad \forall i, j, r, t \qquad (14)$$

$$\left.\begin{array}{l} d_{ij1t}^{BLS,B} = 1, \text{ if } \delta_{ij1t} = 1 \\ d_{ijrt}^{BLS,B} = 0, \text{otherwise} \end{array}\right\} \qquad \forall i,j,r,t \qquad (15)$$

Next, the average arrival and service rates for BLS ambulances in Step 3 and Step 8 can be calculated using equations (16) and (18), respectively.

$$\Lambda_{jt}^{BLS} = \sum_{i\in I}\sum_{r\in R} d_{ijrt}^{BLS,A}\lambda_{it}^{A} + \sum_{i\in I}\sum_{r\in R} d_{ijrt}^{BLS,B}\lambda_{it}^{B} \qquad \forall\, j,t \qquad (16)$$

$$T_{jt}^{BLS} = \frac{\sum_{i\in I}\sum_{r\in R} d_{ijrt}^{BLS,A}\lambda_{it}^{A}\tau_{ijt}^{A} + \sum_{i\in I}\sum_{r\in R} d_{ijrt}^{BLS,B}\lambda_{it}^{B}\tau_{ijt}^{B}}{\sum_{i\in I}\sum_{r\in R} d_{ijrt}^{BLS,A}\lambda_{it}^{A} + \sum_{i\in I}\sum_{r\in R} d_{ijrt}^{BLS,B}\lambda_{it}^{B}} \qquad \forall\, j,t \qquad (17)$$

$$M_{jt}^{BLS} = \frac{1}{T_{jt}^{BLS}} \qquad \forall j,t \qquad (18)$$

The value of $\Lambda_{jt}^{BLS}, T_{jt}^{BLS}$ and M_{jt}^{BLS} from equations (16) to (18) are then used to calculate the estimate for π_{jt}^{BLS} in Step 9 using equation (7). The value of π_{jt}^{BLS} is then used to estimate $d_{ijrt}^{BLS,B}$ and $d_{ijrt}^{BLS,A}$ in Step 7 using equations (19) and (20), respectively.

$$d_{ijrt}^{BLS,B} = \prod_{q\in R|q<r}\left(\sum_{p\in J|p\neq j}\delta_{ipqt}\pi_{pt}^{BLS}\right)\left(1-\pi_{jt}^{BLS}\right) \qquad \forall i,j,r,t \qquad (19)$$

$$d_{ijrt}^{BLS,A} = \prod_{q\in R|q<r}\left(\sum_{p\in J|p\neq j}\delta_{ipqt}\pi_{pt}^{BLS}\right)\left(1-\pi_{jt}^{BLS}\right)\prod_{s\in R}\left(\sum_{u\in J}\delta_{iust}\pi_{ut}^{ALS}\right) \qquad \forall i,j,r,t \qquad (20)$$

5.2.3 Estimating busy probability for FRV

FRV-type ambulances are considered backup ambulances for type B calls and are used for cases where BLS ambulances are not available immediately. Therefore, the dispatch probability of ambulances will depend on the busy probability of BLS. Similar to the approaches used for ALS and BLS, we first set $\pi_{jt}^{FRV} = 0$ and calculate dispatch probability using equation (21) in Step 2.

$$\left.\begin{array}{l} d_{ij1t}^{FRV,B} = \prod_{s\in R}\left(\sum_{u\in J}\delta_{iust}\pi_{ut}^{BLS}\right) \quad \text{if } \delta_{ij1t} = 1 \\ d_{ijrt}^{FRV,B} = 0, \text{ otherwise} \end{array}\right\} \qquad \forall i \in I, j \in J, r, t \qquad (21)$$

The arrival and service rates at each station *j* can then be obtained using equations (22) and (24), respectively, in both Steps 3 and 8. The estimate of $d_{ijrt}^{FRV,B}$ in Step 7 can be calculated using equation (25).

$$\Lambda_{jt}^{FRV} = \sum_{i\in I}\sum_{r\in R} d_{ijrt}^{FRV,B}\lambda_{it}^{B} \qquad \forall j,t \qquad (22)$$

$$T_{jt}^{FRV} = \frac{\sum_{i\in I}\sum_{r\in R} d_{ijrt}^{FRV,B}\lambda_{it}^{B}\tau_{ijt}^{B}}{\sum_{i\in I}\sum_{r\in R} d_{ijrt}^{FRV,B}\lambda_{it}^{B}} \qquad \forall\, j,t \qquad (23)$$

$$M_{jt}^{FRV} = \frac{1}{T_{jt}^{FRV}} \qquad \forall\, j,t \qquad (24)$$

$$d_{ijrt}^{FRV,B} = \prod_{q\in R|q<r}\left(\sum_{p\in J|p\neq j}\delta_{ipqt}\pi_{pt}^{FRV}\right)\left(1-\pi_{jt}^{FRV}\right)\prod_{s\in R}\left(\sum_{u\in J}\delta_{iust}\pi_{ut}^{BLS}\right) \qquad \forall i,j,r,t \qquad (25)$$

6. COMPUTATIONAL RESULTS

In this section, we present the results obtained from our computational experiments. The simulation models were implemented using Simulink in MATLAB-2021. The linear equations for the HQM model were solved using MATLAB-2021. We developed small toy instances which can be easily solved using all three approaches. The dispatch probability obtained using all approaches is then also used to obtain various performance measures for comparison.

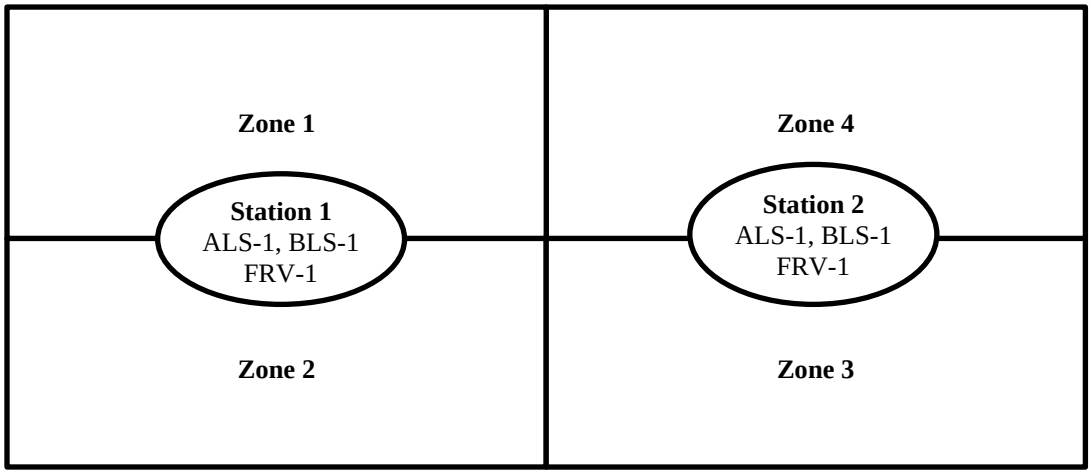

Figure 3. A simple EMS system with four zones and two stations

Table 1. Preference order of station for calls from each zone

Zone	Rank 1	Rank 2
1	Station 1	Station 2
2	Station 1	Station 2
3	Station 2	Station 1
4	Station 2	Station 1

6.1. Illustrative example

Consider a simple system shown in Figure 3, where a region is divided into four demand zones. Each zone has an arrival rate for calls of each type associated with it. Two ambulance stations are located within the region: station 1 and station 2. Each station has one ALS, one BLS, and one FRV ambulance that can serve calls arriving at that station. The preference order for assigning calls from each zone to a station is given in Table 1. Station 1 is the primary station (rank 1) for zones 1 and 2, while station 2 is the primary station for zones 3 and 4. If the preferred ambulance for a call type is not available at the primary station, the call is served using an ambulance from the next ranked station. We assume that each zone has a constant and equal arrival rate for each type of call. The service time for each zone-station pair is assumed to be known and constant. The time-dependent variation in service time and demand is ignored for these instances as we are only interested in comparing the three approaches.

Table 2. Arrival rate for each call type at each station

Problem set	Call type A	Call type B
1	0.25	0.50
2	0.20	0.40
3	0.40	0.60
4	0.40	0.80
5	0.50	1.00

Table 3. Service rate for each station and demand zone pair

Call type	Problem set	Station 1				Station 2			
		Zone 1	Zone 2	Zone 3	Zone 4	Zone 1	Zone 2	Zone 3	Zone 4
Type A	1	1.00	1.00	0.80	0.80	0.80	0.80	1.00	1.00
	2	1.00	1.00	0.80	0.80	0.80	0.80	1.00	1.00
	3	1.25	1.25	1.00	1.00	1.00	1.00	1.25	1.25
	4	1.25	1.25	1.00	1.00	1.00	1.00	1.25	1.25
	5	2.00	2.00	1.50	1.50	1.50	1.50	2.00	2.00
Type B	1	1.20	1.20	1.00	1.00	1.00	1.00	1.20	1.20
	2	1.20	1.20	1.00	1.00	1.00	1.00	1.20	1.20
	3	1.50	1.50	1.20	1.20	1.20	1.20	1.50	1.50
	4	1.50	1.50	1.20	1.20	1.20	1.20	1.50	1.50
	5	2.50	2.50	1.50	1.50	1.50	1.50	2.50	2.50

6.2. Input data

Including the toy example presented in the previous sub-section, we created three small problem instances with the following characteristics.

(a). Two demand zones, One station, Four ambulances
(b). Two demand zones, One station, Six ambulances
(c). Four demand zones, Two stations, Eight ambulances

For each instance, five sets of input data were considered to generate 15 problem instances. These problem instances are used to validate the approaches presented in the previous sections. We apply the three approaches to determine the dispatch probability and other performance measures to compare these approaches. Table 2 shows the arrival data considered for both types of calls. Table 3 presents the service rate for both call types from every zone to each station. These instances are solved using all three approaches, and the results obtained are summarised.

6.3. Key performance measures

We use server utilization, service time, and on-scene time as the key performance measures to compare the results obtained from the HQM, simulation, and the proposed approximate approach.

6.3.1 Server Utilization

Server utilization is the probability that a server is busy during a given period. It can also be described as the proportion of time the server is busy in a given time interval. For the HQM, server utilization can be obtained by adding the probability of all the states in which a given server is busy. For the simulation model, this value is directly obtained as an output of the simulation run. In the case of the approximate approach, it can be obtained using the formula given in equation (26).

$$U_j^k = \rho_j^k(1 - \pi_j^k) \tag{26}$$

6.3.2 Server-level mean service time

The mean server-level service time represents the mean of the time taken by ambulances located at a given station to serve a call. This service time varies from one ambulance type to another as they may serve different types of patients (calls). Also,

server-level service time depends on the travel time required and the arrival rate of each type of call. The server-level service time for different types of ambulances can be calculated using equations (11), (17) and (23) if the dispatch probability is known.

6.3.3 System-level mean service time

The system-level mean service time for a call type is the mean of the service time for all calls arriving in the system. The system-level service time provides a single value for service time for the whole system for each call type to compare different configurations. The system-level service time for call type *l* can be calculated using equation (27).

$$\tau_{sys}^{l} = \frac{\sum_{i\in I}\sum_{j\in J}\sum_{r\in R}\sum_{k\in K} d_{ijr}^{kl}\lambda_i^l\tau_{ij}^l}{\sum_{i\in I}\sum_{j\in J}\sum_{r\in R}\sum_{k\in K} d_{ijr}^{kl}\lambda_i^l} \qquad \forall l \qquad (27)$$

6.3.4 Server-level mean on-scene time

The server-level mean on-scene time represents the average time an ambulance takes to reach a patient location from a given station after a call is received. Although this depends on the delay between the arrival of a call and the dispatch of an ambulance, this delay is considered to be constant in the estimations and omitted. Therefore, only the actual travel time from the station to the patient location is considered in calculating on-scene time. Mean on-scene time can be calculated by replacing the service time τ_{ij}^l with the response time o_{ij}^l in equations (11), (17), and (23).

6.4. Summary of results

Table 4 compares the simulation, HQM and approximate approaches based on the server utilization for a six-ambulance system. Figure 4 presents a summary of the percentage difference between estimates of server utilization using the HQM and the approximate approach compared to the simulation approach. The percentage differences from the simulation approach are less than 2% for the HQM-based approach, and it is within 7% for the approximate approach. Similarly, Table 5 compares the three approaches based on the system-level service time for the six-ambulance system. For the system-level service time, the difference between HQM and simulation is within 3% for all the instances for type A calls, whereas the difference is within 8% for type B calls. A similar range of differences is observed between the approximate approach and simulation for both types of calls. The difference is within 4% for type A calls and within about 8% for type B calls. Figure 5 presents the average percentage difference between estimates of system-level service time using the HQM and approximate approach compared to the simulation approach for both type A and type B ambulances. The average difference is about 2% and 4% for HQM compared to simulation for type A and type B calls, respectively. Whereas the average difference is about 3% and 5% for type A and type B calls, respectively, between approximate approach and simulation. We can clearly observe that both the HQM-based approach and the approximate approach produce results that are close enough to the simulation. Table 6 presents a comparison based on server-level service time for the problem instances.

Similarly, Figure 6 presents the average percentage difference in server-level service time estimates for each server in the six-ambulance system. In the case of server-level service time, the difference between HQM and simulation is within 3%, and the difference is also within 2-3% for the approximate approach compared to the simulation model. Thus, we can observe that, overall, both approaches are able to provide good approximations for the performance measures calculated.

We also solved two more problem sets to compare these approaches, a four-ambulance system and another eight-ambulance system. Figure 7 compares the average difference in estimates for server-level utilization of a four-ambulance system. The difference in server utilization is within 5% between the HQM and simulation for all server types, while the difference between the approximate approach and simulation is within 6% for all server types. Similarly, Figure 8 presents a comparison of the average difference in estimates for server-level utilization for a four-ambulance system using different approaches. The HQM-based and approximate approaches produce a similar result for the service time, with a difference below 3% for all servers. Table 7 presents the percentage difference in the system-level service time for instances based on the four-ambulance system. Similar to the six-ambulance system, the difference is lower for type A calls and slightly higher for type B calls. Figure 9 presents the average percentage difference in system-level service time for the four-ambulance system. The difference is about 3% for type A calls for approximate and HQM approaches compared to the simulation approach. For type B calls, however, HQM results in a slightly lower difference of within 4%, while the approximate approach results in a difference closer to a 6% difference.

Table 4. Comparison of server utilization for a six-ambulance system

Approach	Instance	Server					
		1	**2**	**3**	**4**	**5**	**6**
HQM	1	0.26	0.26	0.43	0.43	0.04	0.05
	2	0.22	0.21	0.35	0.35	0.02	0.02
	3	0.33	0.34	0.42	0.42	0.03	0.03
	4	0.33	0.34	0.55	0.56	0.09	0.09
	5	0.26	0.26	0.45	0.46	0.03	0.03
Simulation	1	0.26	0.26	0.42	0.42	0.04	0.03
	2	0.20	0.21	0.33	0.33	0.01	0.01
	3	0.32	0.32	0.40	0.40	0.03	0.03
	4	0.32	0.32	0.53	0.53	0.07	0.07
	5	0.25	0.25	0.44	0.44	0.03	0.02
Approximate	1	0.29	0.29	0.35	0.35	0.07	0.07
	2	0.25	0.25	0.31	0.31	0.05	0.05
	3	0.34	0.34	0.36	0.36	0.06	0.06
	4	0.34	0.40	0.41	0.40	0.11	0.11
	5	0.29	0.29	0.37	0.37	0.08	0.08
Average % difference	HQM and Simulation	1.04	1.14	1.67	1.69	0.71	1.01
	Approximate and Simulation	2.94	4.04	6.56	6.96	3.90	4.19

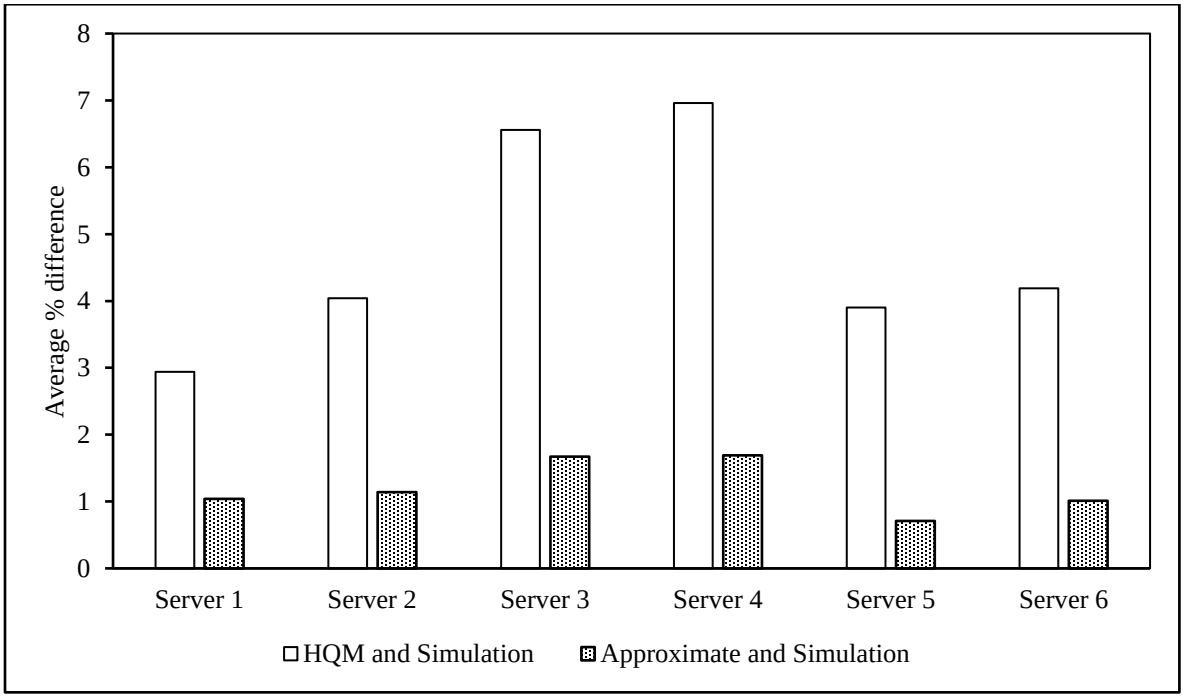

Figure 4. Average percentage difference between server utilization for a six-ambulance system

Table 5. Comparison of system-level service time for a six-ambulance system

Approach		Call type	Instance 1	Instance 2	Instance 3	Instance 4	Instance 5
HQM		Type A	62.72	62.64	50.93	51.01	32.09
		Type B	52.82	54.76	43.29	41.73	26.75
Simulation		Type A	61.74	61.43	49.72	49.71	31.19
		Type B	54.48	53.67	43.86	44.65	28.62
Approximate		Type A	63.31	63.08	50.92	51.56	32.20
		Type B	51.14	57.98	42.26	43.73	26.52
% Difference	HQM and simulation	Type A	1.58	1.98	2.44	2.61	2.88
		Type B	3.05	2.03	1.30	6.52	6.52
	Approximate and simulation	Type A	2.53	2.69	2.42	3.72	3.25
		Type B	6.14	8.05	3.65	2.06	7.32

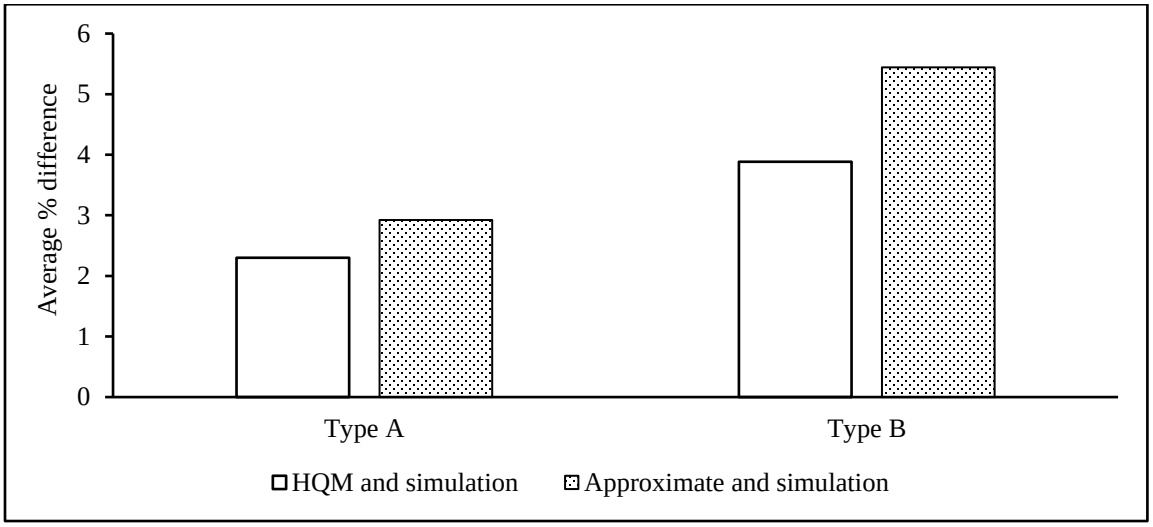

Figure 5. Average percentage difference between system-level service time for a six-ambulance system

Table 6. Comparison of server-level service time for a six-ambulance system

Approach	Instance	Server					
		1	2	3	4	5	6
HQM	1	62.72	62.72	51.75	51.75	35.49	35.49
	2	62.59	62.59	51.42	51.42	35.22	35.22
	3	50.96	50.96	42.15	42.15	25.27	25.27
	4	51.16	51.16	41.89	41.89	25.76	25.76
	5	32.11	32.11	26.68	26.68	9.53	9.53
Simulation	1	61.75	61.75	51.64	52.01	35.58	35.70
	2	61.44	61.42	51.77	51.91	35.51	35.55
	3	49.73	49.73	41.87	42.21	25.53	25.64
	4	49.73	49.73	41.38	42.41	25.54	26.00
	5	31.20	31.19	27.43	28.13	9.64	9.89
Approximate	1	63.36	63.36	51.99	51.99	35.69	35.69
	2	63.00	63.00	51.58	51.48	35.44	35.44
	3	51.03	51.03	42.09	42.09	25.59	25.59
	4	50.81	50.81	43.38	43.38	25.98	25.98
	5	32.26	32.26	26.60	26.60	10.25	10.25
% Difference	HQM and Simulation	2.34	2.35	1.11	1.59	0.82	1.50
	Approximate and Simulation	2.66	2.68	1.89	1.79	1.75	0.85

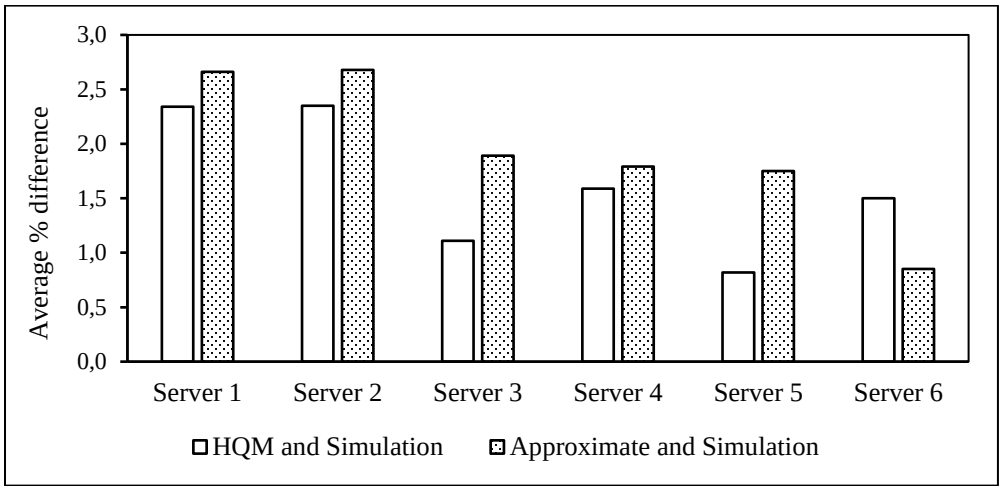

Figure 6. Average percentage difference in server-level service time estimates for a six-ambulance system

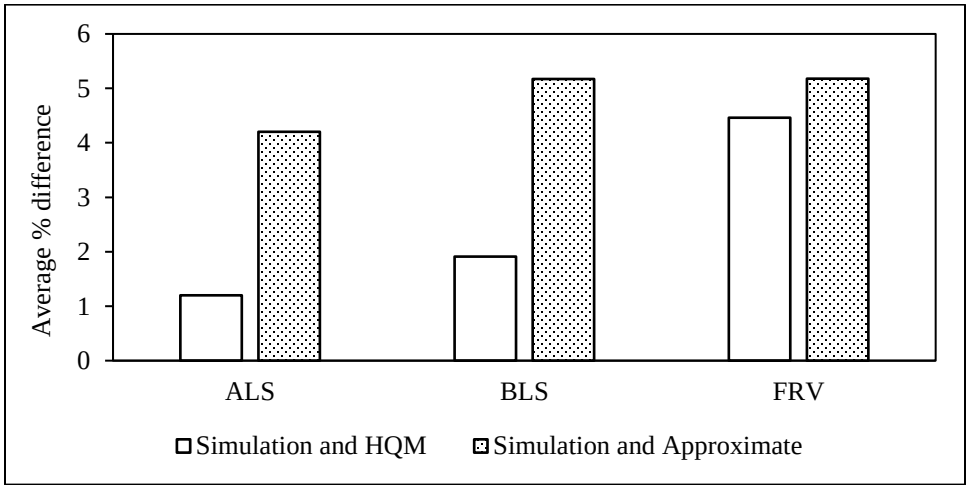

Figure 7. Average percentage difference in server-level utilization estimates for a four-ambulance system

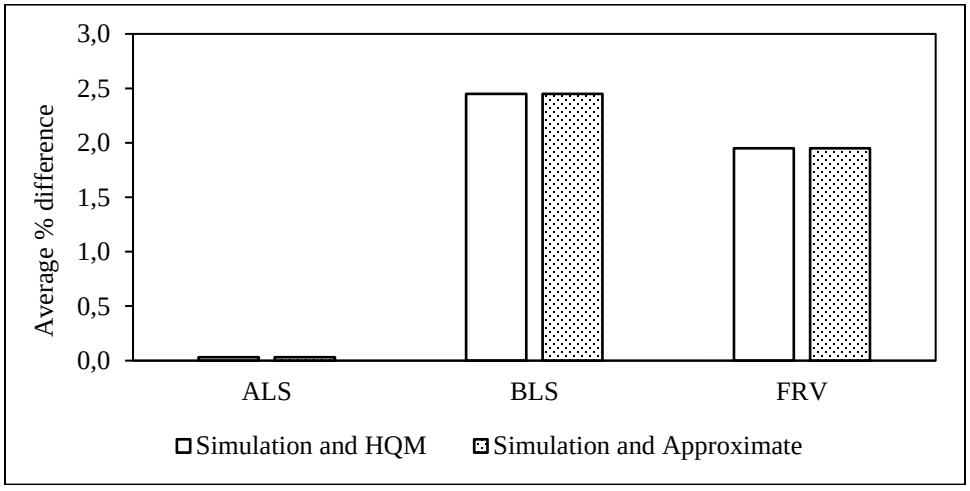

Figure 8. Average percentage difference in server-level service time estimates for a four-ambulance system

Table 7. Percentage difference in system-level service time for a four-ambulance system

	Simulation and HQM (% difference)		Simulation and Approximate (% difference)	
Instance	**Type A**	**Type B**	**Type A**	**Type B**
1	1.58	3.05	2.83	4.04
2	1.98	2.03	2.73	0.91
3	2.44	1.30	2.86	6.33
4	2.61	6.52	3.07	6.39
5	2.88	6.52	3.72	4.06
Average	2.30	3.89	3.04	5.55

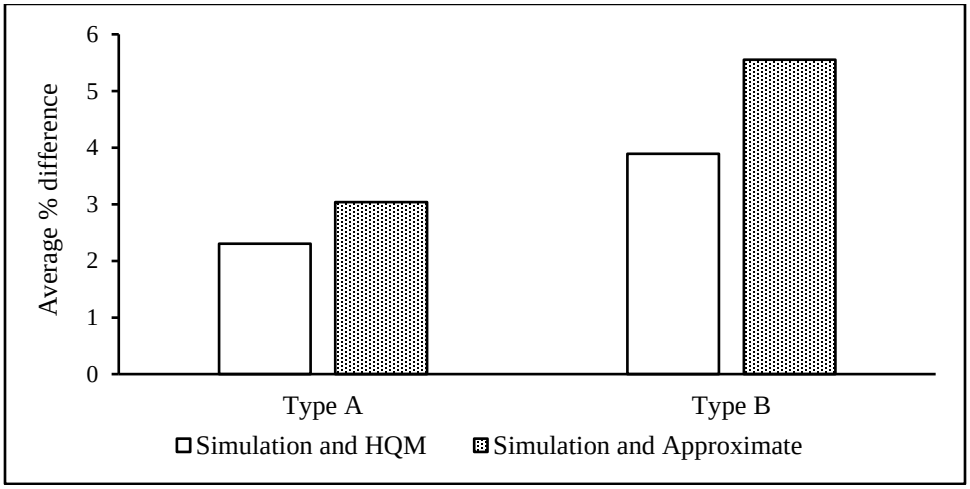

Figure 9. Average percentage difference in system-level service time for a four-ambulance system

Figure 10 shows the average percentage difference in the server-level service time estimates for an eight-ambulance system. The HQM produces service time estimates within 0.5% of the simulation approach for all servers. The mean service time obtained using the approximate approach is within 2% of the simulation approach for all server types. Figure 11 shows the average percentage difference in the server-level utilization estimates between different approaches for an eight-ambulance system. The mean utilization estimates of the approximate approach are within 2% of the simulation approach for all the servers. In the case of the approximate approach, there is a slightly higher difference of about 7% for BLS, while it is lower than 5% for ALS and FRV. The likely reason for the difference is that we have not accounted for the calls that are served by BLS after FRV is sent first. Table 8 presents the percentage difference between the three approaches in the system-level service time for the eight-ambulance system. The difference between the simulation and HQM approach is within 1% for both types of calls in all instances. For the approximate approach, the percentage difference is within 1% for type A calls and within 3% for type B calls. Figure 12 summarises the average percentage difference in system-level service time for an eight-ambulance system. The average percentage difference is within 0.5% for type A calls using both HQM and approximate approaches. The average difference is about 1.5% for type B calls using the approximate approach compared to 0.5% using the HQM approach.

Figure 13 compares the on-scene time estimates for all ambulance types for the four-ambulance system. The average difference is within 4% for the HQM compared to the simulation for all servers of the four-ambulance system. The average percentage difference for the estimates obtained using the approximate approach is within 7% for all servers, with a maximum difference for BLS servers. Figure 14 presents a comparison of the average percentage difference between estimates of on-scene time using different approaches for a six-ambulance system. The comparison between simulation and HQM shows a difference within 5% for all servers, with very low differences in the case of FRVs. Whereas the comparison between simulation and approximate approach shows a difference within 8% for all servers, with the highest difference being in the case of BLS ambulances.

Figure 15 presents a comparison of the average percentage difference between estimates of on-scene time using different approaches for an eight-ambulance system. Both HQM and approximate approach produce results within 2% of the

simulation approach. The comparison between simulation and HQM shows a difference within 0.5% for all servers, whereas the comparison between simulation and approximate approach shows a difference of about 1.2% for all servers.

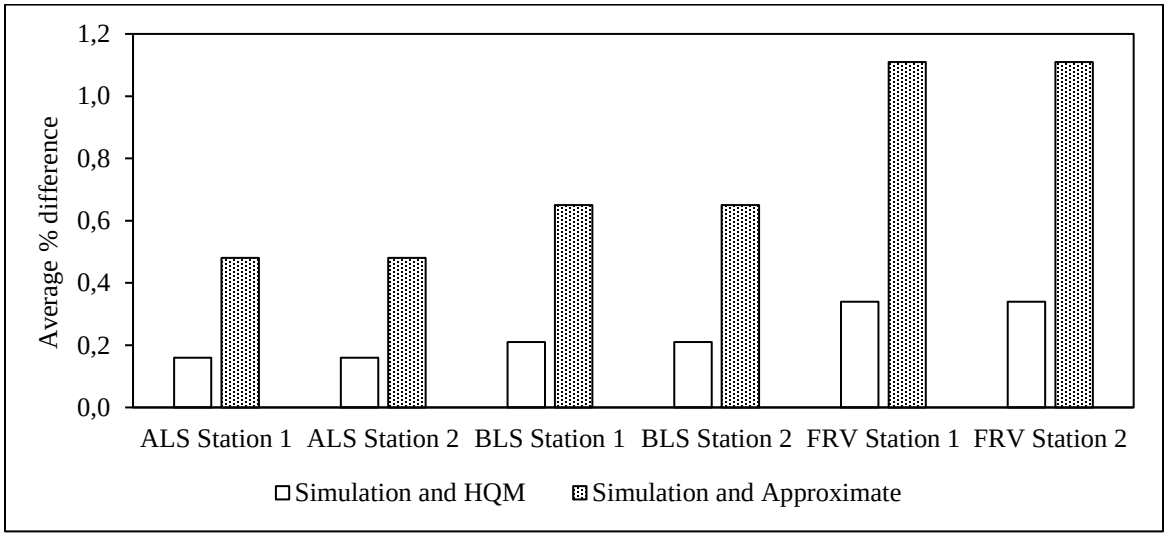

Figure 10. Average percentage difference in server-level service time estimates for an eight-ambulance system

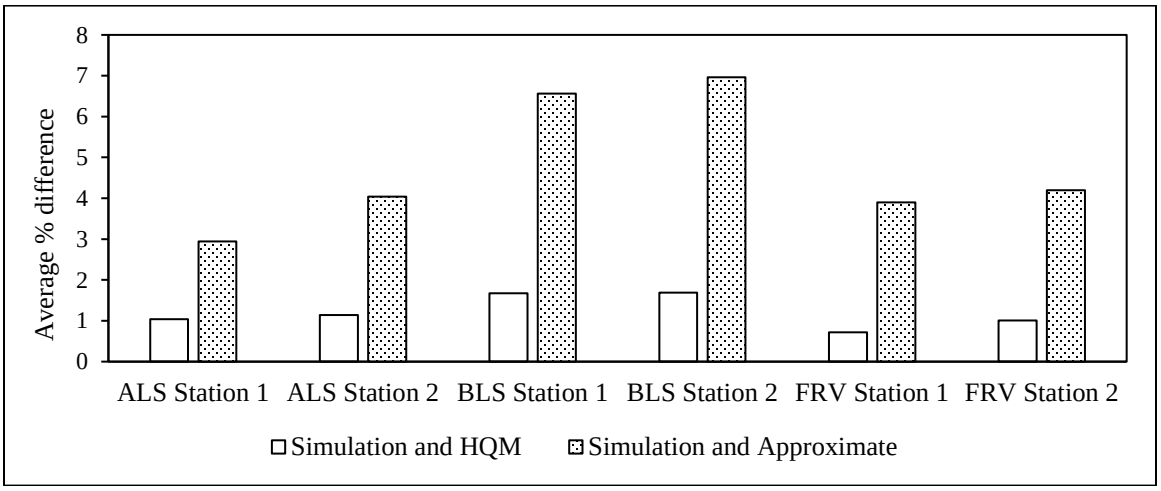

Figure 11. Average percentage difference in server-level utilization estimates for an eight-ambulance system

Table 8. Percentage difference in system-level service time for an eight-ambulance system

	Simulation and HQM		**Simulation and Approximate**	
Instance	**Type A**	**Type B**	**Type A**	**Type B**
1	0.05	0.13	0.51	0.26
2	0.26	0.39	0.76	2.15
3	0.14	0.58	0.25	1.3
4	0.15	1.02	0.15	2.71
5	0.23	0.27	0.78	0.95
Average % difference	0.17	0.48	0.49	1.47

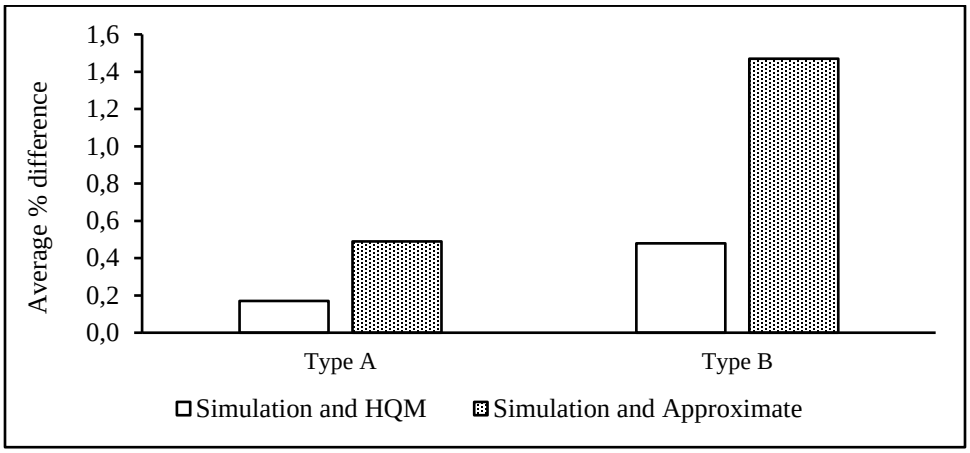

Figure 12. Average percentage difference in system-level service time for an eight-ambulance system

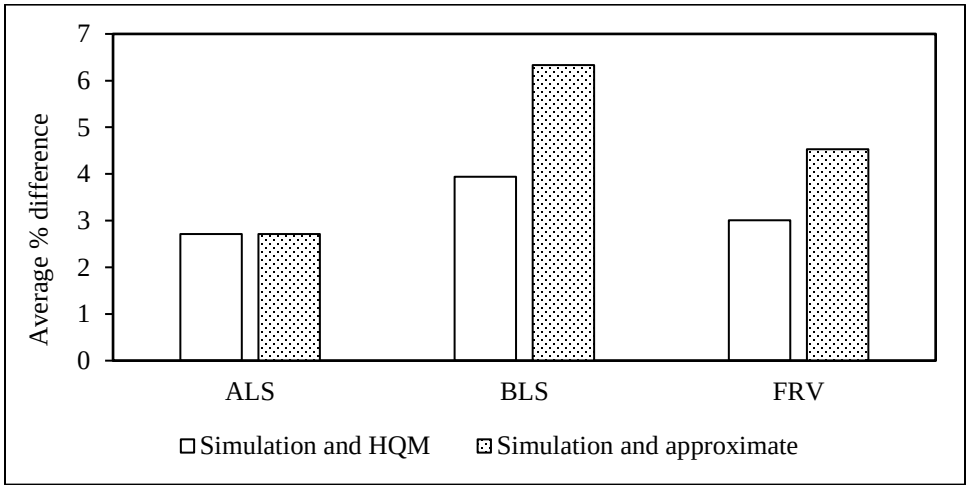

Figure 13. Comparison of the average percentage difference between estimates of on-scene time using different approaches for a four-ambulance system

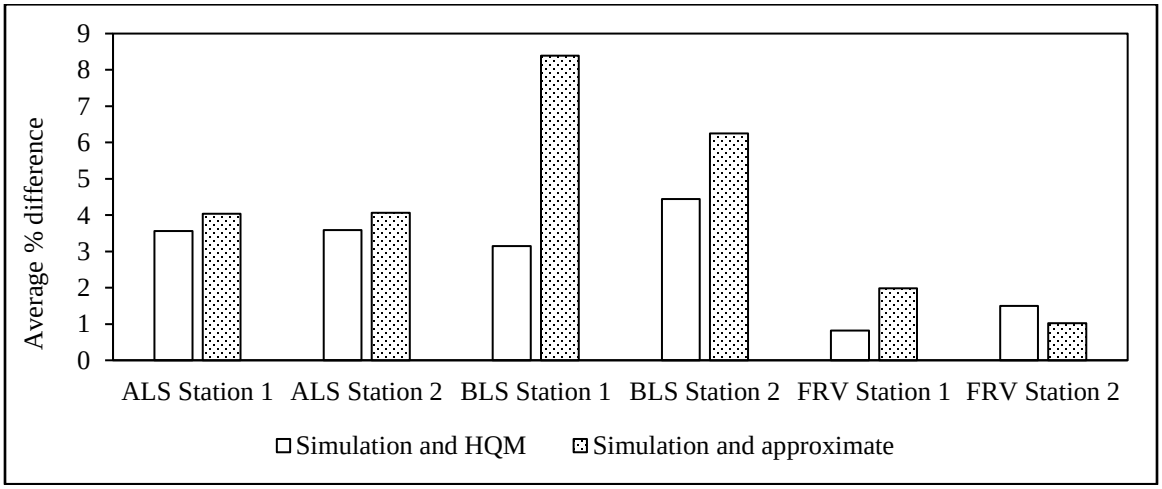

Figure 14. Comparison of the average percentage difference between estimates of on-scene time using different approaches for a six-ambulance system

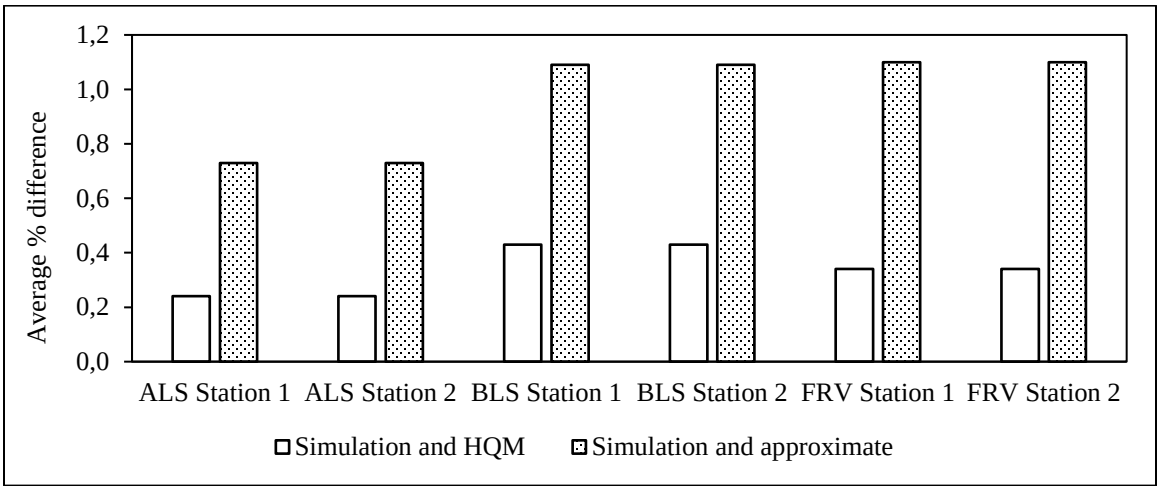

Figure 15. Comparison of the average percentage difference between estimates of on-scene time using different approaches for an eight-ambulance system

Based on the above results, we observe that the approximate approach is able to provide good estimates of the various performance measures. Overall, the HQM performed better than the approximate approach, with very close estimates compared to the simulation approach. Although the difference in some instances was close to 5-8% for the approximate approach, considering it is very easy to implement, it is worthwhile to evaluate large EMS systems. The approximate approach can be easily incorporated within an ambulance allocation system to get performance estimates for the system at each iteration.

7. CONCLUSIONS

Performance evaluations of such realistic public emergency systems are necessary for planning decisions related to the design of the EMS systems. In this work, we consider a complex, realistic EMS system that operates three different types of ambulances, namely ALS, BLS and FRV, that cater to both high-priority and low-priority patients. Also, we consider ALS as dedicated ambulances only sent for high-priority calls, while BLS can serve both high and low-priority calls. FRV is a non-transport ambulance that provides only immediate emergency medical service at a patient location but is not used to transport a patient. We consider a system where these ambulances are used as a backup for BLS ambulances. Performance evaluation of such a realistic system is complicated because different ambulances may serve different types of patients and may, therefore, have different priority rankings. Thus, a preferred station with an ambulance available might not be able to serve a call, as not all ambulances can serve all calls. Additionally, each demand zone can have different arrival rates for different patient types and also different service times based on the type of call and the type of ambulance dispatched. Accounting for the priority between different ambulance types, as well as the preference between different stations combined with varying arrival rates and service rates, makes the proposed system more realistic.

To determine the performance measures, we present an HQM-based approach that uses a 3n queueing system to allow ambulances to serve different types of calls. Additionally, we propose an iterative approximate approach to estimate the performance measures. The approximate approach starts by assuming all calls are served by primary stations alone and is used to estimate call arrival and service rates. These estimates are used further to calculate the probability of calls lost and then iteratively find a better estimate for the arrival and service rates. The proposed approaches are tested on three different simple examples. These example problems showed that the error for HQM is within 3-4% of the simulation estimates. Similarly, the approximate approach consistently provides results within 8% of the simulation results for all the instances considered. As the approximate approach takes significantly less time and is easy to implement for even large systems, it is a more useful approach to planning such EMS systems.

Although the presented problem considers various real-life issues related to the performance evaluation of EMS systems, it has some limitations. We have not accounted for multiple simultaneous dispatches of ambulances to the same zone. The location of ambulances is assumed to be static and known. Thus, the dynamic relocation of ambulances is not considered within the model, which can be considered in future research. Also, the preference order of stations is assumed to be fixed and known. However, the ranking of ambulance stations can change based on the availability of ambulances and expected demand, which needs to be incorporated. The stochastic nature of the arrival rates, service time and response time is also not

considered. Another key factor that could be considered in further research is the possibility of delays due to ambulance diversion and blocking in the emergency department.

REFERENCES

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.

Aringhieri, R., Bruni, M. E., Khodaparasti, S., and van Essen, J. T. (2017). Emergency medical services and beyond: Addressing new challenges through a wide literature review. *Computers and Operations Research*, 78, 349–368.

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

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

Bang, I., Kim, M. I., and Kim, Y. (2023). Performance Impact of Dispatching and Routing in An Automated Guided Vehicle System. *International Journal of Industrial Engineering: Theory, Applications and Practice*, 30(2).

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

Beojone, C. V., and Souza, R. M. D. (2017). Application of the hypercube model with queue priorities and more than one preferential server: a case study on a SAMU. *Gestão & Produção*, 24, 814-828.

Beojone, C. V., Máximo de Souza, R., and Iannoni, A. P. (2021). An Efficient Exact Hypercube Model with Fully Dedicated Servers. *Transportation Science*, 55(1), 222-237.

Brandeau, M., and Larson, R. C. (1986). Extending and applying the hypercube queueing model to deploy ambulances in Boston. A. Swersey, E. Ignall, eds. Delivery of Urban Services, Invited Chapter.

Boyacı, B., and Geroliminis, N. (2015). Approximation methods for large-scale spatial queueing systems. *Transportation Research Part B: Methodological*, 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*, 57(1), 251-255.

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

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

Chiyoshi, F. Y., Galvão, R. D., and Morabito, R. (2003). A note on solutions to the maximal expected covering location problem. *Computers & Operations Research*, 30(1), 87-96.

de Souza, R. M., Morabito, R., Chiyoshi, F. Y., and Iannoni, A. P. (2015). 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), 274-285.

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

Galvao, R. D., and Morabito, R. (2008). Emergency service systems: The use of the hypercube queueing model in the solution of probabilistic location problems. *International Transactions in Operational Research*, 15(5), 525-549.

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

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(2), 287-300.

Goldberg, J., Dietrich, R., Chen, J. M., Mitwasi, M., Valenzuela, T., and Criss, E. (1990). A simulation model for evaluating a set of emergency vehicle base locations: Development, validation, and usage. *Socio-Economic Planning Sciences*, 24(2), 125-141.

Goldberg, J., and Szidarovszky, F. (1991). Methods for solving nonlinear equations used in evaluating emergency vehicle busy probabilities. *Operations Research*, 39(6), 903-916.

Gokalp, E. (2021). Dynamic and flexible staff deployment in accident and emergency departments using simulation-based optimization. *International Journal of Industrial Engineering: Theory, Applications and Practice*, 28(1).

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

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

Iannoni, A. P., Morabito, R., and Saydam, C. (2008). A hypercube queueing model embedded into a genetic algorithm for ambulance deployment on highways. *Annals of Operations Research*, 157(1), 207-224.
Jarvis, J. P. (1985). Approximating the equilibrium behavior of multi-server loss systems. *Management Science*, 31(2), 235-239.

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.

Lee, T., Cho, S. H., Jang, H., and Turner, J. G. (2012, December). A simulation-based iterative method for a trauma center—Air ambulance location problem. In *Proceedings of the 2012 Winter Simulation Conference (WSC)* (pp. 1-12). IEEE.

Liu, H., Yin, H., Zhou, Y., and Li, M. (2021). Cooperative hypercube queuing model for emergency service systems. *Journal of Advanced Transportation*, 2021.

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

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

Larson, R. C., and Odoni, A. R. (1981). *Urban operations research.* Prentice Hall.

Larson, R. C. (2004). OR models for homeland security. *OR/MS Today*, 31(5), 22-29.

McCormack, R., and Coates, G. (2015). A simulation model to enable the optimization of ambulance fle*et al*location and base station location for increased patient survival. *European Journal of Operational Research*, 247(1), 294-309.

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

Mendonça, F. C., and Morabito, R. (2001). Analysing emergency medical service ambulance deployment on a Brazilian highway using the hypercube model. *Journal of the Operational Research Society*, 52(3), 261-270.

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

Rodrigues, L. F., Morabito, R., Chiyoshi, F. Y., Iannoni, A. P., and 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.

Sacks, S.R., Grief, S. (1994). Orlando Police Department uses OR/MS methodology, new software to design patrol districts. *OR/MS Today*, 30–32 (Baltimore)

Saydam, C., and Aytuğ, H. (2003). Accurate estimation of expected coverage: revisited. *Socio-Economic Planning Sciences*, 37(1), 69-80.

Saydam, C., Repede, J., and Burwell, T. (1994). Accurate estimation of expected coverage: a comparative study. *Socio-Economic Planning Sciences*, 28(2), 113-120.

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

Wang, J., Xu, L., Zhang, G., Xu, B., Peng, Y., and Zheng, L. (2023). Modelling of Crowd Dynamics Considering Emergency Signs and Emotions. *International Journal of Industrial Engineering: Theory, Applications and Practice*, 30(3).

Yoon, S., and Albert, L. A. (2018). An expected coverage model with a cutoff priority queue. *Health Care Management Science*, 21, 517-533.