RAIRO-Oper. Res. 56 (2022) 2967–2997
https://doi.org/10.1051/ro/2022140

RAIRO Operations Research
www.rairo-ro.org

AMBULANCE LOCATION UNDER TEMPORAL VARIATION IN DEMAND USING A MIXED CODED MEMETIC ALGORITHM

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

Abstract. Emergency medical services (EMS) are among the most important services in any society due to their role in saving people's lives and reducing morbidities. The location of ambulance stations and the allocation of ambulances to the stations is an important planning problem for any EMS system to ensure adequate coverage while minimising the response time. This study considers a mixed-integer programming model that determines the ambulance locations by considering the time of day variations in demand. The presented model also considers heterogeneous performance measures based on survival function and coverage for different patient types with varying levels of urgency. A memetic algorithm based-approach that applies a mixed chromosome representation for solutions is proposed to solve the problem. Our computational results indicate that neglecting time-dependent variation of demand can underestimate the number of ambulances required by up to 15% during peak demand. We also demonstrate the effectiveness of the proposed solution approach in providing good quality solutions within a reasonable time.

Mathematics Subject Classification. 90B06, 90B80, 90C11, 90C30, 90C90.

Received January 13, 2022. Accepted August 4, 2022.

1. Introduction

Emergency medical services (EMS) are an integral part of modern healthcare systems and are responsible for providing pre-hospital care and transportation services to both emergency and non-emergency patients. Providing a timely response to emergency calls requiring urgent medical care can save lives and reduce morbidity in patients. The response time, *i.e.* the time taken to respond to emergency calls, plays a vital role in patient outcomes, necessitating that the ambulances reach the location within the least possible time. The need to respond to calls in a timely manner leads to the ambulance location problem that requires locating ambulance stations and allocating ambulances to these stations such that all demand zones can be reached within a specified time limit. Determining the optimal location of ambulance stations is a strategic-level planning problem, while the optimal allocation of ambulances to these stations is a tactical planning problem. Despite the difference in planning levels, these problems have often been studied simultaneously in the literature [56]. The two-level

Keywords. Emergency medical service planning, ambulance planning, location-allocation, memetic algorithm, operations research in health services.

[1] Department of Industrial and Systems Engineering, Indian Institute of Technology Kharagpur, Kharagpur 721302, West Bengal, India.

[2] National Rail and Transportation Institute, Vadodara 390004, Gujarat, India.

*Corresponding author: ravi1989.06@gmail.com

nature of the ambulance location is a critical problem since it influences other tactical and operational level problems such as crew scheduling, relocation, and routing of ambulances.

EMS systems often aim to meet a predefined response time threshold while responding to emergency calls [73]. Therefore, response time-based measures are the most commonly used performance measure in ambulance location problems as they are also easy to evaluate [45]. Coverage, which measures the expected number of calls to be served within a pre-specified response time, is used in the majority of ambulance location studies [56]. However, some studies have also considered other performance measures such as survival function [30] and equity [12]. The survival function evaluates EMS system performance based on the probability of patient survival, while equity considers whether demand from all demand zones is equally served. Since an EMS system could serve calls of different urgency levels, location models considering heterogeneous performance measures have been developed [35].

Ambulance location depends on various input factors such as demand arising out of each region, travel time required to reach each location in the region, and response time requirement that needs to be met to achieve the target service performance levels. An important factor that requires consideration while locating ambulances is the probabilistic nature of the problem because ambulances might be busy when a call arrives [22]. Many ambulance location models estimate an area-specific busy probability assuming all ambulances in a given area have equal busy probability [38]. Some researchers have relaxed this assumption to consider server-level busy probability [7]. Another important factor that needs to be considered in determining the location of ambulances is the uncertainty in input parameters [4]. Demand from each region exhibits temporal variation, *i.e.* it usually varies over the day and on various days of the week [11]. Considering an average demand over a period of time may result in underestimating the number of ambulances required to achieve the required service level, especially during peak demand periods. Taking into account these variations in demand can improve the overall allocation of ambulances.

Facility location models, including coverage-based location models, are NP-hard [15,16], due to which many heuristic and metaheuristic approaches have been proposed to solve these models [17,34,51]. Furthermore, many real-life urban EMS systems serve a large region with a large population. For example, the case considered in our test instances is the city of Mumbai, with a population above 12 million and an area of 619 square km (`www.mcgm.gov.in`). For such large urban EMS systems, the demand from the region would need to be divided into a large number of zones for improved accuracy and realism, which will result in large-scale optimisation problems [51]. Solving such large-scale optimisation problems will require significantly high computational power and time. Although the location of ambulance stations is a strategic problem, ambulance allocation is a tactical problem [56]. Therefore, an ambulance location problem is not just required to be solved while initially setting up an EMS system but frequently on a medium-term basis to redistribute ambulances based on the demand forecast. Thus, it becomes necessary to develop effective solution approaches that provide quality solutions within a reasonable time for such realistic problems.

Determining optimal location-allocation of ambulances is thus an important planning problem for the operation of an EMS system due to their direct impact on the survival probability of patients and the overall performance of the EMS system. Incorporating the temporal variation in demand is significant due to the impact on the number of ambulances and ambulance stations required, affecting the busy probability and utilisation of ambulances. As ambulances serve different types of patients, it is also necessary to consider different performance measures to evaluate the performance of the EMS system. Further, ambulance location decisions also affect the operational-level planning of EMS, such as relocation of ambulances and crew scheduling. Based on these observations, we formulate the following research questions to address in this study.

(i) What is the impact of temporal variation in demand on the number of ambulances and ambulance stations located compared to considering average demand?
(ii) How does temporal variation in demand affect the number of ambulances and ambulance stations located compared to considering maximum demand throughout the day?
(iii) How do the busy probability and server utilisation of ambulances vary as demand varies over the day?

(iv) Can a heuristic approach be developed to solve the ambulance location problem with time-dependent variation and server-level busy probability for obtaining better solutions within less time than commercial solvers and the conventional genetic algorithm (GA) for realistic large-scale problem instances?

To address the above research questions, we consider an ambulance location problem that accounts for both temporal variations in demand and server-level busy probability. We also consider heterogeneous performance measures for different patient types, including coverage and survival probability-based measures. A memetic algorithm (MA) based solution approach is developed to solve large-size real-life instances of the problem. The overall contributions of this work can be summarised as follows.

- Present a model for the location of ambulance stations and allocation of ambulances to the stations, which accounts for temporal variation in demand and server-level busy probability.
- Develop and implement a novel mixed coded MA-based solution approach to solve large-scale problems, utilising a combination of integer and real coding for solution representation.
- Analyse the impact of considering variations in demand over the time of day on ambulance allocation decisions.

The organisation of the remainder of this paper is as follows. Section 2 reviews the existing literature in the area of ambulance location. Subsequently, Section 3 discusses the problem background and the model formulation, followed by Section 4 detailing the proposed solution methodology. In Section 5, a detailed summary of the computational results is presented. Finally, Section 6 presents the conclusions and scope for further research.

2. Review of literature

A considerable number of papers that study ambulance planning models can be understood from the numerous review papers published related to the topic [4,9]. Brotcorne *et al.* [9] review ambulance location problems considering static, probabilistic and dynamic models. Aringhieri *et al.* [4] review more recent EMS planning problems across the entire emergency care pathway. Reuter-Oppermann *et al.* [56] classify EMS planning problems by focusing on the interdependencies between different planning levels and problems. Bélanger *et al.* [6] review EMS planning problems at the tactical and operational levels, focusing on the interaction of various decisions related to location, relocation, routing and dispatching of ambulances. These review papers provide a comprehensive overview of the overall research in the area of ambulance location. We present an overview of the papers more closely related to the present study.

The location set covering model (LSCM) and maximum covering location problem (MCLP) constitute the earliest major ambulance location models in the literature. Toregas *et al.* [64] propose LSCM that minimises ambulances required while covering all demand zones. The MCLP, proposed by Church and ReVelle [14], on the other hand, tries to maximise the possible coverage given a fixed number of ambulances. Daskin [22] develops a maximum expected covering location problem (MEXCLP) that addresses the unavailability of ambulances by explicitly considering the busy fraction of ambulances while maximising the expected coverage value. Larson [37] introduces a hypercube queuing model that accounts for the busy probability of ambulances. Davis [23] presents a simulation-based approach that tackles temporal variations in demand and travel time, multiple casualty accidents, and variations in service time at each incident location. The double standard model (DSM) addressing the issue of ambulance unavailability by maximising the demand covered by two ambulances for each location is proposed by Gendreau *et al.* [31]. Various extensions and improvements to these models have been developed and presented in the literature are discussed subsequently.

Hogan and Revelle [33] introduce backup coverage models that maximise the backup coverage for demand zones while enforcing single coverage as a constraint. They analyse the model that considers a trade-off between primary and backup coverages, which they extend to include multiple coverages. Mandell [40] proposes a covering model for a two-tiered ambulance system where advanced life support (ALS) and basic life support (BLS) can serve all calls. A region is considered covered if an ALS is available for service within a pre-specified response

time. Revelle and Hogan [57] introduce a probabilistic model called the maximum availability location problem (MALP) to extend the LSCM to account for the busy probability of ambulances. The MALP-I assumes that the busy probability is the same for all ambulances, which is relaxed to develop the MALP-II by considering area-specific busy probability. Batta *et al.* [5] develop an adjusted MEXCLP model that assumes that ambulances are independent. Marianov and Revelle [42] present an extension to the MALP that relaxes the assumption that the busy probability of ambulances is independent of each other. They propose a queueing MALP (Q-MALP) model that calculates server-level busy probability using queueing theory to make the model more realistic. McLay [44] develops the MEXCLP2 model that extends the MEXCLP to account for two server types.

The demand for emergency care follows a definite pattern and varies continuously over the day, with low demand during the night and peak demand during day time [10, 11, 41]. Disregarding this variation in demand while determining the location of ambulances can result in inaccuracies in estimating the number of ambulances required to meet the demand. Repede and Bernardo [55], while developing a decision support system for the Louisville EMS system, present an extension to the MEXCLP to consider time variation in demand. They solve the model for 35 scenarios of the EMS system with the number of ambulances ranging from 5 to 11. Rajagopalan *et al.* [52] develop a dynamic available coverage location model, a multi-period model applicable for dynamic demand, which accounts for significant changes in demand across different time clusters. Schmid and Doerner [58] develop a mixed-integer linear programming (MILP) model for multi-period planning of ambulance locations while ensuring a pre-specified coverage level throughout the planning horizon. They apply the model for a dataset based on Vienna with potential locations varying from 16 to 163 and 16 available ambulances. Degel *et al.* [24] take a data-driven optimisation and develop a model that determines the flexible location and fleet size based on empirically determined required coverage, which accounts for the time-varying nature of demand. They demonstrate their approach using the case study of Bochum (Germany), with 163 demand zones, 21 potential ambulance stations and 14 ambulances. Van Den Berg and Aardal [67] also develop a probabilistic model to account for the time-dependent nature of demand and travel time that, while maximising expected coverage, minimises the number of stations and relocations required across different periods. They apply their model to a randomly generated instance with 500 demand zones and 50 potential ambulance stations.

Coverage is the most widely used objective function for evaluating EMS location models. However, it is ineffective in discerning the impact of differences in response times [45]. Erkut *et al.* [30] observe that coverage-based models may locate stations at sites just within the response time limit while increasing coverage to additional regions resulting in decreased survival probability of patients. To overcome this limitation, they introduce a maximum survival location problem to maximise the survival function for patients, which measures the impact of response time on the probability of survival. Erkut *et al.* [30] solve problems with 180 demand points, 16 potential stations and up to a maximum of 16 ambulances. Knight *et al.* [35] propose a maximal expected survival location model for heterogeneous patients to account for multiple outcome measures for various patient types and present an approximation methodology that iteratively finds the best solution. They solve the problem based on EMS in Wales with 18 demand nodes, 11 ambulance stations and 36 ambulances. An important objective considered in the ambulance literature is equity in service levels between different zones of a region. Chanta *et al.* [12] introduce a minimum p-envy location problem (MpELP) to maximise service equity among all zones. MpELP models equity using envy, a function of the distance of a demand zone from its nearest and backup stations. Chanta *et al.* [13] improve the p-envy model introduced in Chanta *et al.* [12] by considering envy as a function of the difference between survival probabilities to capture more accurately the difference in patient outcomes among different zones.

Leknes *et al.* [38] formulate a MILP model to consider heterogeneity in demand while accounting for various outcome measures for multiple patient types. They tested the proposed model on instances based on Trondheim and Malvik with 67 demand zones and 44 potential station locations and Sør-Trøndelag with 139 demand zones and 76 potential locations. Yoon and Albert [72] present an EMS system with priority queues for different patient types by employing multiple types of patients to develop a MILP model for ambulance deployment. The developed model is studied using a dataset based on Hanover County, with 270 demand nodes, 16 potential stations and up to 15 ambulances. El Itani *et al.* [29] extend the MEXCLP to consider the problem of utilising

additional private ambulances to improve coverage. Andersson *et al.* [3] extend the model presented by Leknes *et al.* [38] to consider various strategic scenarios, including the closure of emergency rooms, usage of designated vehicles for non-urgent patients, and time-dependent variation in demand. Boutilier and Chan [8] consider the problem of locating and routing ambulances in a lower-middle-income country setting of Dhaka, Bangladesh. They use field data and prediction models to handle uncertainty in the data and apply a simulation-based approach to answer policy-related questions. Nelas and Dias [50] present a location model for an EMS system with multiple types of ambulances that serve multiple types of patients, explicitly considering the substitutability between different types of ambulances for different care services and the assignment of ambulances to specific calls. They applied their model to a dataset based on the Coimbra district (Portugal) for locating 35 ambulances among 34 potential ambulance stations. Naji *et al.* [49] propose a two-server dynamic covering location model to consider two types of patients and two types of servers. Yoon *et al.* [74] present a stochastic model for joint location and dispatching of ambulances under demand uncertainty.

Due to the NP-hard nature of the ambulance location problem, solution approaches based on various heuristic and metaheuristic approaches have been proposed in the literature. Gendreau *et al.* [31] develop and apply a solution approach based on tabu-search to a static ambulance location model with double coverage. Doerner *et al.* [27] propose an ant colony optimisation-based solution approach to extend the double coverage model and compare it with the tabu search metaheuristic. Rajagopalan *et al.* [51] compare four metaheuristic approaches, including an evolutionary algorithm, tabu search, simulated annealing, and a hybridised hill-climbing algorithm for solving the MEXCLP. Toro-Díaz *et al.* [65] present a GA-based optimisation framework for the location and dispatching of ambulances simultaneously. Toro-Díaz *et al.* [66] propose a tabu search-based heuristic for a model that considers fairness in large-scale EMS systems. Zhen *et al.* [76] develop a simulation-optimisation framework with GA to deploy and relocate ambulances. Akdoğan *et al.* [2] construct a GA to solve the ambulance location model based on an approximate queueing model that minimises the system's response time. Kaveh and Mesgari [34] present an improved biogeography-based optimisation algorithm for solving the MCLP, which they apply to solve the ambulance location problem for a real dataset in Tehran.

Table 1 compares selected articles in the literature based on some key features of the models. The table shows that only Knight *et al.* [35], Leknes *et al.* [38], and Andersson *et al.* [3] consider both server-level busy probability and heterogeneous performance measures (coverage and survival function). However, Knight *et al.* [35] and Leknes *et al.* [38] do not consider temporal variation in demand, while Andersson *et al.* [3] consider stations to be fixed in their time-dependent model. Among the papers that consider relocation of ambulances based on temporal variation, none consider server-level busy probability or survival-based objective. Table 1 also shows the different solution approaches applied to solve the ambulance location problems. Based on our literature review, we identified a need to analyse the impact of temporal variation in demand on the number of ambulances located while incorporating server-level busy probability and heterogeneous performance measures. Therefore, we propose a time-dependent maximum expected performance location problem for heterogeneous patients that accounts for the time-dependent variation in demand and estimates station-specific busy probabilities. The proposed problem also considers an objective function based on the combination of coverage and survival function for different patient types. A mixed-integer non-linear programming (MINLP) model is presented and converted to a MILP model by linearising the non-linear constraints. A memetic algorithm-based approach is proposed to solve the model as the presented model is difficult to solve using commercial optimisation solvers.

3. Problem description

The problem we address in this work is characterised by a set of demand zones (nodes), where each zone represents a neighbourhood from where calls for ambulances are received. The calls received from demand zones can be of three types (A, B, and C) with varying levels of urgency [35]. Type A calls are time-critical emergency calls, where reaching the patient location within the shortest possible time is important for patient survival. Type B calls are emergency calls that need immediate transportation but are not life-threatening, and type C calls require non-emergency transportation. The arrival rate for each type of call from each zone is assumed to

TABLE 1. Comparison of selected relevant articles with the current work.

Author(s)	Performance measure		Temporal demand variation	Flexible station location and fleet size (Relocation)	Server-specific busy probability	Solution approach
	Coverage based	Survival function-based				
Church and ReVelle [14]	✓					Greedy algorithms
Daskin [22]	✓					Heuristic
Batta *et al.* [5]	✓					Heuristic
Repede and Bernardo [55]	✓		✓	✓		–
Erkut *et al.* [30]		✓				Commercial solver
Rajagopalan *et al.* [52]	✓					Tabu search heuristic
McLay [44]	✓					Commercial solver
Schmid and Doerner [58]	✓		✓	✓		Variable neighbourhood search
Knight *et al.* [35]	✓	✓			✓	Iterative approach
Degel *et al.* [24]	✓		✓	✓		Commercial solver
Van Den Berg and Aardal [67]	✓		✓	✓		Commercial solver
Leknes *et al.* [38]	✓	✓			✓	Commercial solver
Yoon and Albert [72]		✓	✓			Branch and Benders cut
Kaveh and Mesgari [34]	✓					Improved biogeography-based optimisation
Boutilier and Chan [8]	✓		✓	✓		Simulation and heuristic
Nelas and Dias [50]		✓	✓		✓	Commercial solver
Yoon *et al.* [74]	✓		✓			Simulation optimisation
This work	✓	✓	✓	✓	✓	Memetic algorithm

be known. The planning horizon is divided into a set of periods, and each period is associated with different call arrival rates for each zone. A set of potential sites are available where ambulance stations can be located. A fleet of identical ambulances is available, and these ambulances need to be assigned to the selected ambulance stations, where multiple ambulances can be assigned to each station.

Demand for ambulances from a zone can be assigned to two (primary and secondary) nearest ambulance stations. Primary and secondary stations are ranked as one and two, respectively, to indicate the priority of stations in sending an ambulance to the demand zone. If all ambulances at the primary station are busy, the demand will be satisfied from the secondary station. Thus, the proposed problem consists of three simultaneous decisions related to ambulance planning: (i) to select the optimal sites for ambulance stations from the available potential sites, (ii) to assign demand for ambulances from each zone to primary and secondary stations, and

(iii) to allocate ambulances to each station. The objective function of the model is formulated to maximise the sum of the survival function for type A calls and the coverage for type B and type C calls for all ambulance station locations.

3.1. Assumptions

The proposed problem makes the following assumptions to formulate the model.

(i) All ambulances are identical and can serve all call types.
(ii) The complete demand for a zone is assumed to be served by ambulances from either a primary or secondary station since other stations covering the demand zone will be located considerably far away in most cases [38].
(iii) The service time for an ambulance includes the time taken to return to the base station, travel time from the station to the patient location, then to the hospital, and time spent at the patient location (On-scene time).
(iv) To simplify the model, the maximum number of ambulances assigned to a station is assumed to be the same for all stations. However, it can be easily relaxed by defining the input parameter $A_{\max}$ separately for each location j as $A^j_{\max}$.

3.2. Mathematical formulation

In this section, we formulate the proposed problem as an MINLP model to determine the optimal location of stations, assignment of demand zones to stations and ambulance allocation to stations by maximising the weighted sum of survival probability and coverage for all stations. The non-linearity of the model arises due to station-specific service rates and busy probability. The non-linear equations are then linearised in Section 3.2.4 using Special Ordered Set 2 (SOS2) type variables.

3.2.1. Notation

Sets

I	Set of demand zones, $i \in I$
J	Set of potential stations for ambulances, $j \in J$
T	Set of time periods, $t \in T$
R	Set of ranking of stations, $r \in R = \{1, 2\}$, 1 = primary and 2 = secondary
H	Set of different types of calls, $h \in H$

Parameters

D_{iht}	Number of calls associated with call type h from demand zone i during period t
W_{ijh}	Performance weight related to demand zone i if covered by station j for call type h
A_{total}	Total number of ambulances available
$Z_{\max}$	Maximum number of stations that can be located
$A_{\max}$	Maximum number of ambulances that can be assigned to a station
λ_{it}	Number of calls per unit time received from demand zone i during period t
T_{ij}	Travel time for an ambulance at station j to reach demand zone i
S_{ij}	Service time to serve a call from demand zone i from an ambulance at station j

Decision variables

x_{jt}	1 if a station is located at location j during period t, 0 otherwise
y_{jt}	Number of ambulances allocated to station j during period t
d_{ijrt}	Proportion of demand from demand zone i covered by station j with rank r during period t
ρ_{ijt}	1 indicates station j is the primary station for demand zone i during period t, 0 otherwise
θ_{jt}	Number of calls received per unit time at station j during period t

TABLE 2. Performance weight (W_{ijh}) values for different types of calls.

Call type	Performance weight (W_{ijh})
A	$\frac{4}{1+e^{-0.679+0.262T_{ij}}}$
B	1, if $T_{ij} < 12$ 0, otherwise
C	1, if $T_{ij} < 20$ 0, otherwise

μ_{jt} Service rate at station j during period t
r_{jst} Number of ambulances relocated from station j to station s during period t

3.2.2. Objective function

$$\text{Maximise} \sum_{t\in T}\sum_{i\in I}\sum_{j\in J}\sum_{r\in R}\sum_{h\in H} D_{iht}W_{ijh}d_{ijrt}. \tag{3.1}$$

The objective function in (3.1) maximises the total performance measure value of the station locations in all periods. The performance measures are defined as a function of the mean travel time T_{ij} from station j to zone i. For type A calls, a survival function-based performance measure given by the formula $H(t) = \frac{1}{1+e^{-0.679+0.262T_{ij}}}$ is used [25]. However, any similar monotonically decreasing function can be utilised, as the optimal location of stations is insensitive to the parameters of the survival function [30]. For type B and type C calls, a coverage-based binary performance measure is used, *i.e.* $H(t) = 1$ or 0, depending on whether T_{ij} is less than or greater than the response time threshold defined. The value of W_{ijh} in the objective function is calculated by multiplying $H(t)$ by a weightage value of 4 for type A calls and 1 for call types B and C to prioritise type A calls [35]. The performance weight values (W_{ijh}) for all three types of calls are summarised in Table 2.

3.2.3. Constraints

$$\sum_{j\in J} x_{jt} \le Z_{\max} \qquad \forall t \in T \tag{3.2}$$

$$\sum_{j\in J} y_{jt} \le A_{\text{total}} \qquad \forall t \in T \tag{3.3}$$

$$y_{jt} \le A_{\max} x_{jt} \qquad \forall j \in J,\ \forall t \in T \tag{3.4}$$

$$\sum_{j\in J} \rho_{ijt} = 1 \qquad \forall i \in I,\ \forall t \in T \tag{3.5}$$

$$\rho_{ijt} \ge d_{ij1t} \qquad \forall i \in I,\ \forall j \in J,\ \forall t \in T \tag{3.6}$$

$$1 - \rho_{ijt} \ge d_{ij2t} \qquad \forall i \in I,\ \forall j \in J,\ \forall t \in T \tag{3.7}$$

$$\sum_{j\in J} d_{ij1t} \ge \sum_{j\in J} d_{ij2t} \qquad \forall i \in I,\ \forall t \in T \tag{3.8}$$

$$\sum_{j\in J}\sum_{r\in R} d_{ijrt} = 1 \qquad \forall i \in I,\ \forall t \in T \tag{3.9}$$

$$\sum_{i\in I}(\lambda_{it}\rho_{ijt} + \lambda_{it}d_{ij2t}) = \theta_{jt} \qquad \forall j \in J,\ \forall t \in T \tag{3.10}$$

$$\mu_{jt} = \frac{\sum_{i\in I}\sum_{r\in R}\lambda_i d_{ijrt}}{\sum_{i\in I}\sum_{r\in R}\lambda_i S_{ij} d_{ijrt}} \qquad \forall j \in J,\ \forall t \in T \tag{3.11}$$

$$\pi_{jt} = f(\mu_{jt}, \theta_{jt}, y_{jt}) \qquad \forall j \in J,\ \forall t \in T | t < T \tag{3.12}$$

$$d_{ijrt} \leq 1 - \pi_{jt} \qquad \forall i \in I,\ \forall j \in J,\ \forall r \in R,\ \forall t \in T \tag{3.13}$$

$$y_{jt} + \sum_{s\in J} r_{sjt} - \sum_{s\in J} r_{jst} = y_{j(t+1)} \qquad \forall j \in J,\ \forall t \in T | t < T \tag{3.14}$$

$$y_{j|T|} + \sum_{s\in J} r_{js|T|} - \sum_{s\in J} r_{sj|T|} = y_{j1} \qquad \forall j \in J. \tag{3.15}$$

Constraints (3.2) and (3.3) ensure that the number of stations and ambulances allocated in any period is less than the maximum number allowed, respectively. Constraint (3.4) limits the number of ambulances within the maximum limit at each station. Constraints (3.5)–(3.9) are coverage constraints, where constraint (3.5) ensures that one primary station is assigned to every demand zone. Constraint (3.6) ensures that demand can be allocated to a station with rank 1 only if it is the primary station for the corresponding zone during that period. Constraint (3.7) ensures that the primary and secondary stations for any demand zone are different in every period, while constraint (3.8) ensures that the primary station has a higher demand proportion allocated compared to the secondary station. Constraint (3.9) ensures that the total demand from each zone is covered in each period, *i.e.* the total proportion of demand allocated to both primary and secondary stations is equal to 1.

Equation (3.10) represents the aggregate arrival rate associated with station j during period t. In equation (3.10), ρ_{ijt} is used to calculate the arrival rate for a primary station since all calls arrive at the primary station first. In contrast, d_{ij2t} is used for a secondary station since only calls diverted from the primary station arrive at the secondary station. Equation (3.11) represents the aggregate service rate associated with station j during period t. Equation (3.12) represents the busy probability of ambulances at a station as a function of arrival rate, service rate and number of ambulances at that station during each period. The value of π_{jt} is calculated by assuming each station as an $M/M/c$-loss queueing system using the Erlang loss formula. Constraint (3.13) ensures that the probability of availability of ambulances at a station should be greater than the proportion of demand allocated to the station. Constraints (3.14) and (3.15) represent the balance between the number of ambulances being relocated to different stations from each station from one period to the next. A dummy station is considered to account for the unused ambulances from one period to another.

3.2.4. Linearisation of the formulation

The formulation presented above has two non-linear equations, equations (3.11) and (3.12), which are linearised using SOS2 type variables [38, 68]. The notation of parameters and variables used for linearisation is presented in Table 3.

To linearise equation (3.11), the numerator $\sum_{i\in I}\sum_{r\in R}\lambda_i d_{ijrt}$, which represents aggregate demand at station j, is divided into p breakpoint values given by $E_p = \{E_1, E_2, \ldots, E_{|p|}\}$, where E_1 and $E_{|p|}$ are the minimum and the maximum possible values for aggregate demand. Similarly, the denominator $\sum_{i\in I}\sum_{r\in R}\lambda_i S_{ij} d_{ijrt}$, which represents aggregate service time, is divided into q breakpoint values represented by $F_q = \{F_1, F_2, \ldots, F_{|q|}\}$. Then, ν_{pjt} and ω_{qjt} are defined as two SOS2 variables to approximate the numerator and denominator, respectively. Defining the variable ν_{pjt} as SOS2 is implicitly equivalent to defining $\sum_{p\in P}\nu_{pjt} = 1, \forall j \in J, \forall t \in T$, and at most, only two consecutive values of ν_{pjt} can be non-zero for each station j in period t [68]. Suppose for any station j and period t if the value of $\sum_{i\in I}\sum_{r\in R}\lambda_{it} d_{ijrt} = \hat{E}$, then two consecutive values $\nu_{p'jt}$ and $\nu_{(p'+1)jt}$ will be chosen such that $E_{p'}\nu_{p'jt} + E_{(p'+1)}\nu_{(p'+1)jt} \geq \hat{E}$ using constraint (3.16) to approximate the numerator of equation (3.11). Similarly, constraint (3.17) together with constraint (3.23) approximates the denominator of equation (3.11) using ω_{qjt}. Equations (3.18) and (3.19) integrate both the SOS2 variables using the breakpoint variable ζ_{pqjt}. Constraint (3.20) ensures that the breakpoint variable $\zeta_{pqjt} = 1$ if station j is selected and 0, otherwise. Finally, equation (3.21) calculates the value of μ_j using the breakpoint value. Thus, the non-linear

TABLE 3. Summary of linearisation parameters and variables.

P	Set of breakpoints for aggregate call arrival rate, $p \in P$
Q	Set of breakpoints for aggregate call service time, $q \in Q$
U	Set of breakpoints associated with service rate for discretising available probability of ambulances,$u \in U$
V	Set of breakpoints associated with arrival rate for discretising available probability of ambulances, $v \in V$
E_p	Aggregate demand associated with breakpoint p
F_q	Aggregate service time associated with breakpoint q
K_u	Service rate associated with breakpoint u
L_v	Arrival rate associated with breakpoint v
P_{uvk}	Busy probability associated with breakpoints u and v, when k number of ambulances are allocated at a station
ν_{pjt}	SOS2 variable related to breakpoint p for service rate linearisation at station j during period t
ω_{qjt}	SOS2 variable related to breakpoint q for service rate linearisation at station j during period t
ζ_{pqjt}	Breakpoint variable associated with service rate linearisation at station j during period t
β_{vjt}	SOS2 variable related to breakpoint v for arrival rate linearisation at station j during period t
ϕ_{ujt}	SOS2 variable related to breakpoint u for service rate linearisation at station j during period t
α_{uvjt}	Breakpoint variable associated with available probability linearisation during period t
δ_{jkt}	1 if there are more than k ambulances at station j during period t, 0 otherwise

value of μ_j is approximated using four neighbourhood points obtained from constraints (3.16) and (3.17).

$$\sum_{p\in P} E_p \nu_{pjt} \geq \sum_{i\in I}\sum_{r\in R} \lambda_{it} d_{ijrt} \qquad \forall j \in J,\ \forall t \in T \tag{3.16}$$

$$\sum_{q\in Q} F_q \omega_{qjt} \geq \sum_{i\in I}\sum_{r\in R} \lambda_{it} S_{ij} d_{ijrt} \qquad \forall j \in J,\ \forall t \in T \tag{3.17}$$

$$\sum_{q\in Q} \zeta_{pqjt} = \nu_{pjt} \qquad \forall j \in J,\ \forall p \in P,\ \forall t \in T \tag{3.18}$$

$$\sum_{p\in P} \zeta_{pqjt} = \omega_{qjt} \qquad \forall j \in J,\ \forall q \in Q,\ \forall t \in T \tag{3.19}$$

$$\sum_{p\in P}\sum_{q\in Q} \zeta_{pqjt} = x_{jt} \qquad \forall j \in J,\ \forall t \in T \tag{3.20}$$

$$\mu_j = \sum_{p\in P}\sum_{q\in Q} \frac{E_p}{F_q} \zeta_{pqjt} \qquad \forall j \in J,\ \forall t \in T \tag{3.21}$$

$$\{\nu_{1jt}, \nu_{2jt,\ldots,}\nu_{pjt}\} \text{ is SOS2} \qquad \forall j \in J,\ \forall t \in T \tag{3.22}$$

$$\{\omega_{1jt}, \omega_{2jt,\ldots,}\omega_{qjt}\} \text{ is SOS2} \qquad \forall j \in J,\ \forall t \in T \tag{3.23}$$

$$\sum_{v\in V} L_v \beta_{vjt} \geq \theta_{jt} \qquad \forall j \in J,\ \forall t \in T \tag{3.24}$$

$$\sum_{u\in U} K_u \phi_{ujt} \geq \mu_{jt} \qquad \forall j \in J,\ \forall t \in T \tag{3.25}$$

$$\sum_{u\in U} \alpha_{uvjt} = \beta_{vjt} \qquad \forall j \in J,\ \forall v \in V,\ \forall t \in T \tag{3.26}$$

$$\sum_{v\in V} \alpha_{uvjt} = \phi_{ujt} \qquad \forall j \in J,\ \forall u \in U,\ \forall t \in T \tag{3.27}$$

$$\sum_{u\in U}\sum_{v\in V}\alpha_{uvjt} = x_{jt} \qquad \forall j \in J,\ \forall t \in T \tag{3.28}$$

$$\{\beta_{1jt}, \beta_{2jt,\ldots,}\beta_{pjt}\} \text{ is SOS2} \qquad \forall j \in J,\ \forall t \in T \tag{3.29}$$

$$\{\phi_{1jt}, \phi_{2jt,\ldots,}\phi_{qjt}\} \text{ is SOS2} \qquad \forall j \in J,\ t \in T. \tag{3.30}$$

For equation (3.12), we first divide the possible service rate and arrival rate values into u and v breakpoint values given by $K_u = \{K_1, K_2, \ldots, K_{|u|}\}$ and $L_v = \{L_1, L_2, \ldots, L_{|v|}\}$, respectively. Next, we calculate the busy probability P_{uvk} of ambulances for each breakpoint u and v using equation (3.31), assuming $M/M/c$-loss queueing system.

$$P_{uvk} = \frac{\frac{(\rho_{uv})^k}{k!}}{\sum_{l=0}^{k}\frac{(\rho_{uv})^l}{l!}}, \tag{3.31}$$

where $\rho_{uv} = \frac{L_u}{K_u}$ and k is the number of ambulances. Then, constraints (3.25) and (3.26) introduce two SOS2 variables β_{vjt} and ϕ_{ujt} for approximating arrival and service rates, respectively. Equations (3.27)–(3.29) combine the SOS2 variables β_{vjt} and ϕ_{ujt} using α_{uvjt}. The relationship of variable δ_{ijrt} with the number of ambulances allocated is defined in constraint (3.32). Then, constraint (3.13) is expressed in the form of constraint (3.33). The term δ_{jkt} ensures that constraint (3.33) holds only when k ambulances are located at station j during period t.

$$\sum_{k=0}^{A_{\max}}\delta_{jkt} \le y_{jt} \qquad \forall j \in J, t \in T \tag{3.32}$$

$$d_{ijrt} - \delta_{jkt} \le 1 - \sum_{u\in U}\sum_{v\in V} P_{uvk}\alpha_{uvjt} \qquad \forall i \in I, j \in J, r \in R, t \in T, k = 0, 1, \ldots, A_{\max}. \tag{3.33}$$

The objective function (3.1) subject to constraints (3.2)–(3.10), and constraints (3.12)–(3.30), along with constraints (3.32) and (3.33), present the complete mathematical formulation of the proposed ambulance location problem. The presented model incorporates time-dependent variation in demand and allows the relocation of ambulances from one period to another to tackle the variation in demand. Another key feature of the proposed formulation is incorporating the server-level busy probability of ambulances to accurately capture the impact on actual coverage and survival probability of patients. However, this makes the model non-linear, which is linearised with the help of SOS2 type variables and breakpoint values. The linearised model consists of the linear objective function and constraints, making the model solvable by commercial MILP solvers. However, the problem is complex because of a large number of binary variables and integer variables used to represent the selection of ambulance stations, the primary station of demand zones, the number of ambulances allocated and the number of relocations. Additionally, the linearisation of constraints adds many SOS2 type variables, a collection of binary variables with a special structure. Thus, the proposed problem is more realistic and complex.

As both the objective function and constraints are linear, the problem remains feasible for all positive values of demand and travel time. However, the linearisation of busy probability and service rate introduces additional artificial variables and parameters that need to be determined to maintain the feasibility of the model. The breakpoint parameter E_p represents the numerator of the service rate. Therefore, the range of the parameter needs to be chosen such that the problem is feasible, as given in equation (3.34). Similarly, the range of the parameter F_q is given in equation (3.35). The parameter L_v represents the total arrival rate for each station in each period and therefore has the same range of values as E_p, as shown in equation (3.36).

$$E_1 = 0 \text{ and } E_{|p|} = \max_{\forall t}\sum_{i\in I}\lambda_{it} \tag{3.34}$$

$$F_1 = 0 \text{ and } F_{|q|} = \max_{\forall t}\sum_{i\in I}\lambda_{it}S_{ij} \tag{3.35}$$

$$L_1 = 0 \text{ and } L_{|v|} = \max_{\forall t} \sum_{i \in I} \lambda_{it}. \tag{3.36}$$

The parameter K_u represents the service rate of the ambulance stations for each period, given by the range in equation (3.37).

$$K_1 = 0 \text{ and } K_{|u|} = \frac{1}{\max_{\forall i,j} S_{ij}}. \tag{3.37}$$

Within the above ranges of the parameters and limits of the assumptions highlighted in Section 3.1, the proposed model remains feasible and can be applied to any EMS system with a given set of demand zones and potential sites for location of ambulance stations.

4. Solution methodology

Our initial experiments show that the problem is not easily solvable for medium and large-sized instances. In our trial runs, using randomly generated data for instances with more than 30 zones, we could not get feasible solutions even after 12 h of using the commercial solver. This can be explained based on the observation that considering multiple periods in a day increases the solution space exponentially. The basic location-allocation problem of locating m servers in n available sites has a solution space of m^n [51]. Considering r different periods increases the possible solution space to mr^{nr}. This rapid increase in solution space combined with the introduction of individual station-level busy probability and relocation-related constraints increases the complexity of the problem. Especially as the number of demand zones, potential locations and periods increases, the computational time required to obtain even a good quality feasible solution increases rapidly. Therefore, we develop a metaheuristic-based solution approach to solve the problem. Several authors [43, 76] have applied various metaheuristic approaches to solve ambulance location problems. In this work, we adopt a MA-based approach, which improves the GA by embedding a local search routine.

MA is a population-based metaheuristic that combines natural adaptation with individual learning acquired by the members of population [36]. MA expedites the process of finding the global optima and avoids premature convergence by balancing the exploration of the search space using GA while enabling the exploitation of the current neighbourhood using local search. MA has been successfully applied in solving various combinatorial problems such as scheduling [70], vehicle routing [47], assignment problems [46], supply chain network design [69] and location routing [1, 48]. MA has also been used in solving healthcare planning problems such as ambulance routing [28, 75] and home healthcare planning [32]. Memetic algorithms presented in the literature generally use binary or integer coding for solution representation. Real-coded memetic algorithms have also been presented to solve optimisation problems with continuous variables [36, 39]. Like a location routing problem that combines strategic level facility location problem [19, 20] and tactical level routing problem [61], an ambulance location-allocation problem combines strategic and tactical level decisions. The problem presented in this work involves a combination of decisions, including station location, demand allocation and ambulance allocation. Therefore, we adopt a mixed coded solution representation in the proposed algorithm to combine binary coding for station location and real coding for demand allocation decisions. However, mixed coding using a combination of binary and integer representation has been used in MA for location routing problems [26, 71].

The overall solution framework based on the memetic algorithm is presented in Figure 1. Initial solutions are generated using a heuristic approach that randomly allocates demand from different zones to stations, ensuring the diversity of solutions. Individual solutions are evaluated based on the fitness function corresponding to the value of the objective function of the problem. Initial solutions are ranked based on their fitness values, and a selection procedure is applied to choose parent chromosomes that are subject to crossover and mutation to produce offspring. A local search is applied to the offspring solutions to improve the quality of these solutions. A pre-specified set of such offspring is generated during an iteration to generate a pool of offspring by repeatedly selecting a pair of parents from the initial population. These offspring replace the worst-ranked solution among the current generation. The complete procedure is repeated until a fixed number of iterations or convergence is reached. Subsequent sections detail the major phases involved in the application of MA.

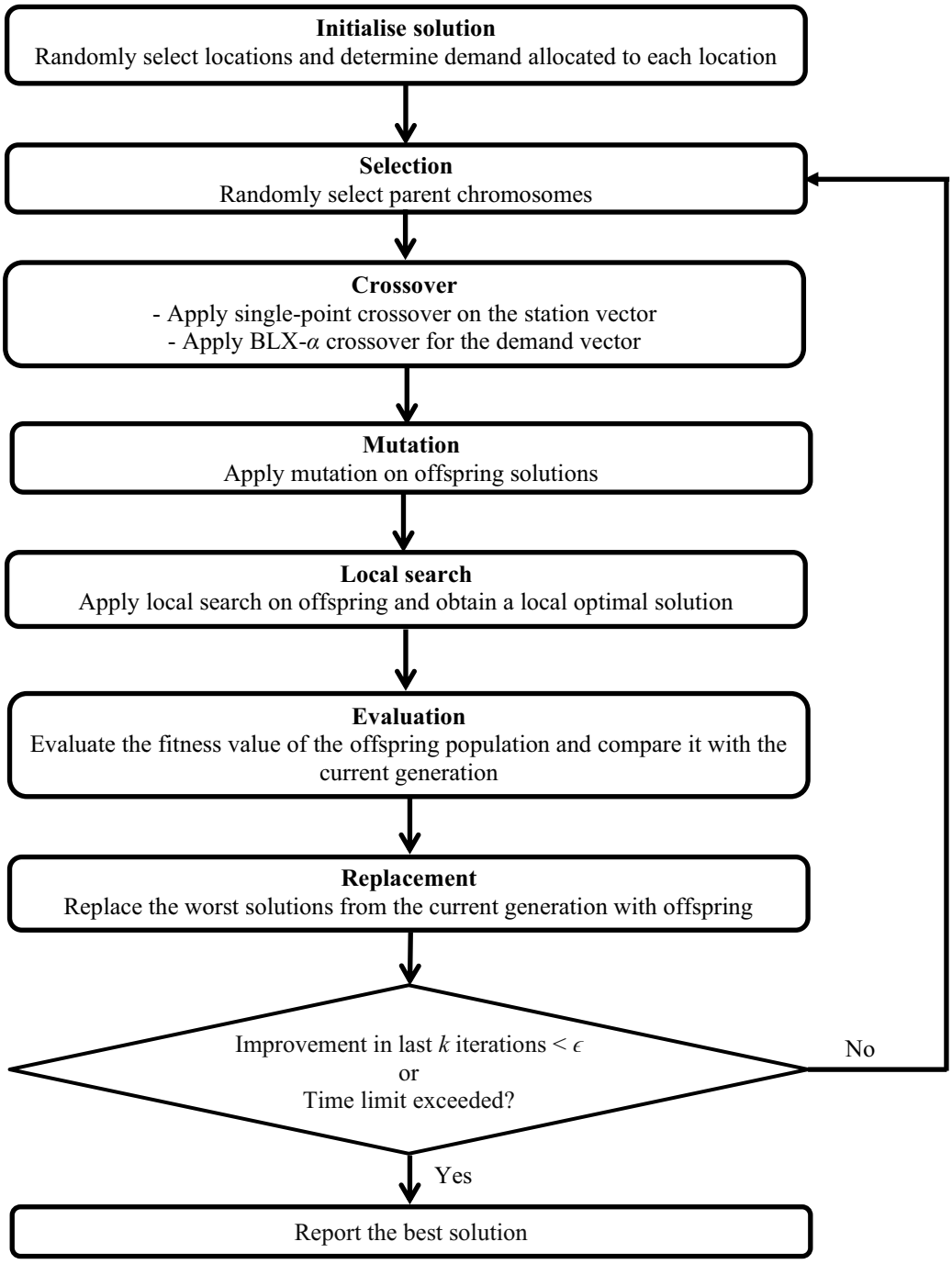

FIGURE 1. Overall solution framework based on memetic algorithm.

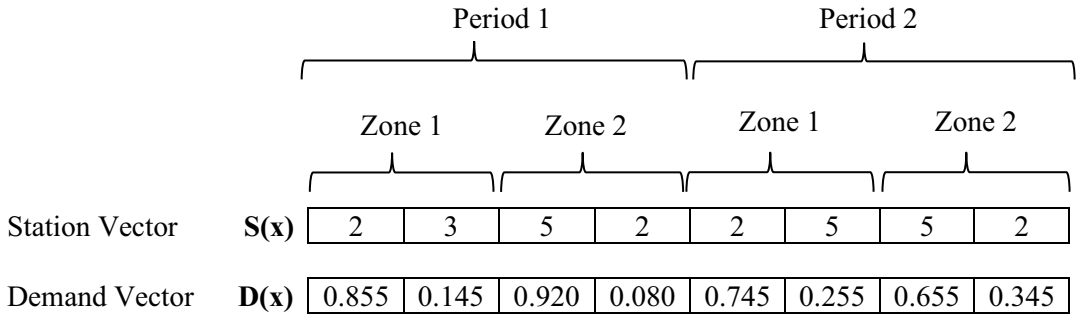

FIGURE 2. Chromosome representation of an individual solution.

4.1. Chromosome representation of a solution

One of the most critical points in the design of MA is the representation of each solution in the form of a chromosome. A mixed coded representation is proposed where each chromosome depicting a solution is represented as a combination of two vectors, namely station vector and demand vector, to accommodate both continuous and discrete variables in the model. The station vector **S(x)** represents the station assigned to each demand zone (which is a discrete variable), and demand vector **D(x)** represents the fraction of demand from a zone that is being served by the corresponding station (which is a continuous variable). The number of ambulances allocated to a station depends on the total demand allocated to that station. The proposed chromosome representation is illustrated in Figure 2. The first element in the station vector represents that station 2 is the primary station and station 3 is the secondary station for zone 1 during period 1. Similarly, station 2 is the primary station and station 5 is the secondary station for zone 1 during period 2. The demand vector represents that 85.5% of calls from zone 1 during the first period are expected to be covered by primary station 2, while 14.5% of calls are covered by secondary station 3.

4.2. Selection, crossover and mutation

Selection is an important phase of the MA process. It is necessary to select good chromosomes from the existing population to generate good quality offspring while maintaining diversity in the solution pool to explore a larger solution space. Various selection procedures exist, such as tournament selection, roulette wheel selection, rank selection, and Boltzmann selection [53]. The procedure employed in this work is adapted from the ranking selection procedure suggested by Reeves [54]. We first order all the individuals from the current population in ascending order of their objective function value, considering their feasibility. The solutions with no constraint violations are ranked higher than those with violations (*i.e.*, infeasible solutions) irrespective of their objective function value. The probability of selection of each individual is then given by the formula $P(k) = \frac{2k}{M(M+1)}$, where k is the position of the chromosome in the ordered list of objective values, and M is the population size. The advantage of this procedure is that while being simple to implement, it gives a higher preference to the solution with better objective function value and lesser constraint violations. The solution with the best fitness value has M times higher probability than the lowest-ranked solution and two times the probability of the median value.

The two individual chromosomes selected during the selection phase are recombined to produce new offspring during the crossover phase. We apply a single point crossover separately on the station vector and the BLX-α crossover for the demand vector. A single point is randomly chosen for the station vector, and the parent chromosomes are swapped along this point to produce station vectors for the child chromosome. The BLX-α crossover for the demand vector part of the chromosome is performed as follows.

Step 1. Consider d_{i1} and d_{i2} to be the two demand proportion values in the ith location of parent 1 and parent 2, respectively.
Step 2. Determine $d_{\min} = \min\{d_{i1}, d_{i2}\}$ and $d_{\max} = \max\{d_{i1}, d_{i2}\}$.
Step 3. Calculate range $I = d_{\max} - d_{\min}$.
Step 4. Generate a random number in the range $d_i = [d_{\min} - I\alpha, d_{\max} + I\alpha]$, where $\alpha \in [0, 1]$.
Step 5. If the random number lies outside the range of the demand proportion value, *i.e.* $[0, 1]$, then select another random number.

The mutation operator in MA is used to prevent the solution from converging to local optima by reducing the convergence rate. In our problem, we apply two separate operators for mutation on the station and demand vectors. For the station vector, we randomly remove one station location from all existing station allocations and replace it with a station chosen from the set of potential locations that are not allocated. This allows the introduction of a new station allocation into the solution, thus diversifying the solution. For the demand vector, we simply replace the existing value of a given chromosome element with a random value in the feasible range for that element.

4.3. Local search

Local search is an optimisation technique that tries to generate a local optimal solution by exploring the neighbourhood of a given solution. A neighbourhood of a solution is explored by performing a few modifications to the initial solution. We apply multiple neighbourhood exploration procedures to explore the neighbourhood more efficiently and avoid rapid convergence to some local optima. Different neighbourhood operators applied in the proposed algorithm are detailed below.

4.3.1. For station vector

Swap stations: we select two random elements representing two zones on the station vector and swap the stations allocated at these locations on the chromosome to generate a new solution.

Replace: in this operator, we select an element on the station vector, and it is replaced by a station that is allocated to at least one of the other zones.

4.3.2. For demand vector

Since the demand vector is a continuous variable, a fixed step random search is applied to determine the local optima in this case. Once the local search is applied to the station vector, we change an element of the current solution by a small step in a random direction to generate a new solution. If the new solution is better than the original one, it is updated, and the next improvement is applied.

5. Computational results

Computational experiments were carried out on the dataset generated based on the urban location of Mumbai in India. The results were used to analyse the performance of the proposed solution approach and the impact of temporal variation in demand on the ambulance location decisions. We first describe the details related to various input parameters in Section 5.1, followed by the summary of results and related discussion in Sections 5.2 and 5.3.

5.1. Generating input data

This section presents the details related to various input parameters such as potential sites for station locations, demand, travel time, and service time for the dataset generated.

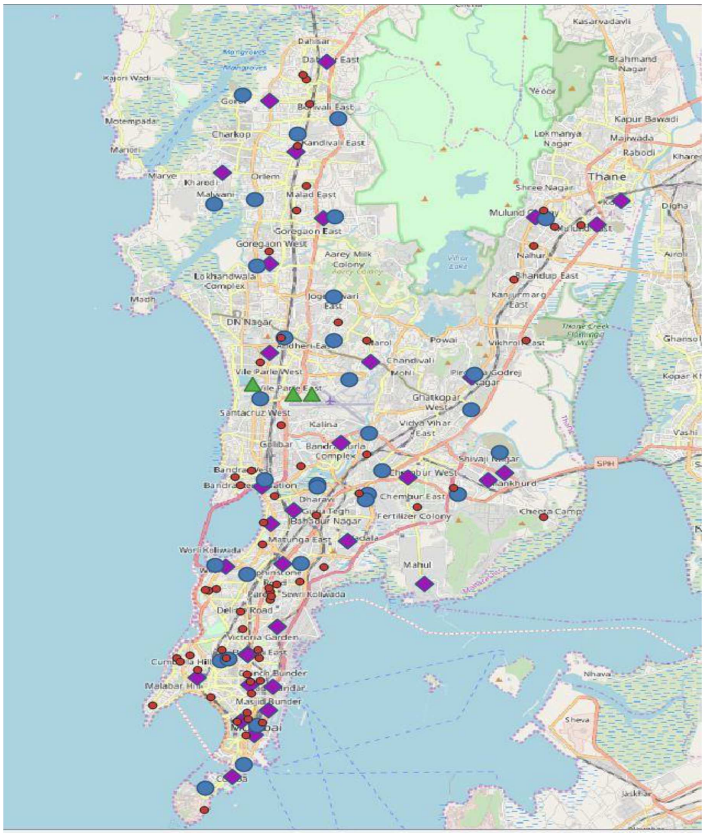

FIGURE 3. Region of interest with potential locations for ambulance stations.

5.1.1. Demand zones and base locations

To make the dataset realistic, the region was divided into smaller demand zones from where the calls for ambulances arise. Potential base locations for ambulances were selected by considering various public places such as railway stations, shopping malls, public hospitals, and schools/colleges. Figure 3 shows all the different selected potential locations for the region of Mumbai under consideration. These locations were considered since the potential locations need to provide some basic features such as parking space for ambulances, security for ambulances and equipment, and electrical supply for recharging or operating equipment. The coordinates of the selected locations were obtained using the QGIS software. A subset of these potential locations and demand zones were considered to generate smaller test instances of the problem.

5.1.2. Call demand

We considered five randomly simulated values for total demand to create five test instances. The total demand is then divided among 144 zones such that demand is approximately proportional to the population of each zone. The smaller problem instances are generated by considering only a subset of the zones. The variation in demand data over a day is considered by breaking the day into four periods: morning (6.00 am to 12.00 pm), afternoon (12.00 pm to 06.00 pm), evening (6.00 pm to 12.00 am) and night (12.00 am to 06.00 am). The total demand is divided into three types, type A, type B, and type C calls, based on their requirements for response time within which they need to be covered. The demand was also separated by the day of the week as weekdays and weekends, as weekends tend to have slightly lower demand and less pronounced peak demand.

5.1.3. Travel and service time

The mean service time needed to serve any demand location from an ambulance station depends on the travel time from the station to the patient location, on-scene time, travel time from the scene to the nearest hospital and then back to the base station. To obtain a better approximation of the travel time, a large number of random points were simulated using the QGIS software for the location of calls. The time taken to reach these locations and then to the nearest hospital is determined based on the distance between locations. The time spent at a patient location is assumed to be a constant value of 15 min [38]. The total service time was calculated by adding the on-scene time to all travel time components from leaving a station until the ambulance returns to the station.

The computational tests were performed on a PC with Intel(R) Core (TM) i5-4570T CPU @2.90 Hz and 8 GB of memory. The MILP model used for comparison was implemented using IBM ILOG CPLEX in concert technology with Java. The proposed MA was implemented using MATLAB 2018b. Parameter settings for the proposed MA and the GA used for comparison were obtained using multiple trial runs. The optimal population size was 12, with a crossover probability of 0.90% and a mutation probability of 0.05%. The selection procedure was repeated to obtain an offspring pool of 8 solutions in each iteration. A maximum time limit of 7200 s is considered for all computational experiments as both MA and GA converged within this time for even large instances. The same time limit is also used for CPLEX, as trial runs longer than 7200 s did not result in any meaningful improvement in the objective function value for medium and large-size instances.

5.2. Performance of the proposed approach compared to CPLEX and GA

To validate the effectiveness of our proposed solution approach, we compared the performance of the proposed approach with the exact solutions obtained using CPLEX and a GA-based approach. Various small-size instances were developed from our initial dataset by varying the number of demand zones from 5 to 15, station locations from 10 to 30, and considering five different total demand values for each case. Table 4 summarises the results obtained for these test instances from CPLEX, GA and MA. From Table 4, we observe that although CPLEX could find an optimal solution for most instances, the time taken to reach the optimal solution increases drastically as the problem size increases. While it only takes less than 41 s to reach optimality for all instances with five demand zones, CPLEX could not find the optimal solution within the specified time limit of 7200 s for any instances with 15 zones. Both GA and MA could find optimal solutions for some of the small-size instances. As problem size increases, both GA and MA converge to a solution within 2% of the CPLEX solution but take significantly less CPU time than CPLEX. For most instances, the difference between GA and MA is less than 1%, with MA slightly outperforming GA in all instances.

Medium-size instances of the problem were generated by varying the number of demand zones from 20 to 60, potential station locations from 40 to 124, and considering five different total demand values for each case. Table 5 summarises the results for the medium-sized instances for comparing the proposed approach with CPLEX and GA. From Table 5, we observe that CPLEX was able to find feasible solutions within the time limit for only three instances out of the total 25 instances, while both MA and GA were able to find feasible solutions for all instances. As there is inherent uncertainty in MA and GA, five trials of each instance were solved. The average solution value and the best solution out of the five trials are reported in Table 5. Similarly, the table also reports the percentage improvement in the best objective value found using MA compared to the exact approach and GA. The proposed MA-based approach found a slightly better solution than GA for almost all instances, with the difference varying in the range 0–3%.

Large-size problem instances were considered with more than 80 demand zones and 150 potential locations. Table 6 summarises the results for the large-size problem instances reporting the best solution found and the average of five trials found using MA and GA. CPLEX is not included in the comparison as it was unable to find a feasible solution for any of the instances within the given time limit. Among both GA and MA, there was a clear difference in the best solution found, as the memetic-based approach consistently converged to a better solution than GA. The difference between both algorithms varied up to 15%, with MA outperforming

TABLE 4. Comparison of the proposed approach with exact approach and GA for small-size instances.

Instance	Instance details	Total demand	Exact approach		Best objective value		CPU time(s)			% Gap between		
			Best objective value	% Solution gap	GA	MA	Exact	GA	MA	Exact and GA	Exact and MA	GA and MA
1	Zones =	3836	4698	0	4698	4698	41	79	83	0	0	0
2	05	3820	4723	0	4723	4723	26	78	74	0	0	0
3	Locations	7843	9735	0	9735	9735	36	77	85	0	0	0
4	= 10	4332	5349	0	5349	5349	28	81	77	0	0	0
5		8221	9996	0	9996	9996	22	75	71	0	0	0
6	Zones =	6862	8322	0	8322	8322	307	120	140	0	0	0
7	08	6871	8340	0	8340	8340	174	124	191	0	0	0
8	Locations	13 679	16 979	0	16 965	16 979	118	138	188	0.08	0	−0.08
9	= 15	7839	9454	0	9454	9454	149	225	191	0	0	0
10		14 255	17 263	0	17 177	17 263	356	306	197	0.50	0	−0.50
11	Zones =	8805	10 953	0	10 890	10 797	264	419	299	0.57	1.42	0.85
12	10	8853	10 957	0	10 806	10 876	648	340	420	1.38	0.74	−0.65
13	Locations	17 624	21 969	0	21 740	21 680	1607	443	456	1.04	1.32	0.28
14	= 20	9930	12 413	0	12 277	12 391	672	395	432	1.09	0.17	−0.93
15		17 814	22 360	0	22 148	22 360	764	447	445	0.95	0	−0.96
16	Zones =	6862	8480	0	8383	8395	1434	443	572	1.14	1.00	−0.14
17	12	6871	8525	0	8360	8328	2107	453	660	1.94	2.32	0.39
18	Locations	13 679	19 025	0	18 635	18 718	5722	632	727	2.05	1.61	−0.45
19	= 25	7839	9680	0	9490	9495	1823	563	572	1.97	1.92	−0.05
20		14 255	19 719	0	19 220	19 373	3533	352	661	2.53	1.75	−0.79
21	Zones =	12 584	15 474	0.75	14 970	15 392	7232	668	864	3.26	0.53	−2.82
22	15	12 929	15 638	0.22	14 670	15 565	7201	701	763	6.19	0.47	−6.10
23	Locations	25 409	35 352	1.04	35 374	34 952	7239	528	879	−0.06	1.13	1.19
24	= 30	14 204	17 246	3.15	17 151	17 259	7205	569	859	0.55	−0.08	−0.63
25		30 394	36 608	1.16	36 486	36 858	7202	657	862	0.33	−0.68	−1.02

GA in 21 out of 25 instances. This difference is significant compared to the difference observed between both the algorithms in medium and small-size instances. Thus, MA outperforms GA for both medium and large-size instances in the best solution found, and the improvement is also higher as the problem size increases.

Figure 4 presents a typical convergence behaviour of MA compared to GA for one of the instances where both algorithms achieved the optimal solution provided by CPLEX. Figure 4 shows that MA converges to the best solution in significantly fewer iterations than GA. A similar difference in performance was observed in other instances for both algorithms. This difference could be attributed to the local search embedded in MA, which increases the CPU time for each iteration but improves the objective function value. The increase in time required can be observed in Figure 5, which shows the convergence of GA and MA with respect to CPU time for the same problem instance. It is observed that even though MA takes almost 67% fewer iterations to achieve the same objective value but the difference in CPU time for both algorithms is negligible.

5.3. Impact of temporal variation in demand

To illustrate the impact of temporal variation in demand, we compare results obtained from three different models that locate ambulances based on three scenarios. Model 1 takes into account the temporal variation in demand and determines the optimal locations for stations and the allocation of ambulances to each station in each period. In model 2, the call arrival rate is considered equal to the average call arrival rate in a day. In model 3, we assume the same call arrival rate as in the peak period during all periods. Both model 2 and model 3 determine a single set of station locations and ambulance allocations throughout the day. The comparison of

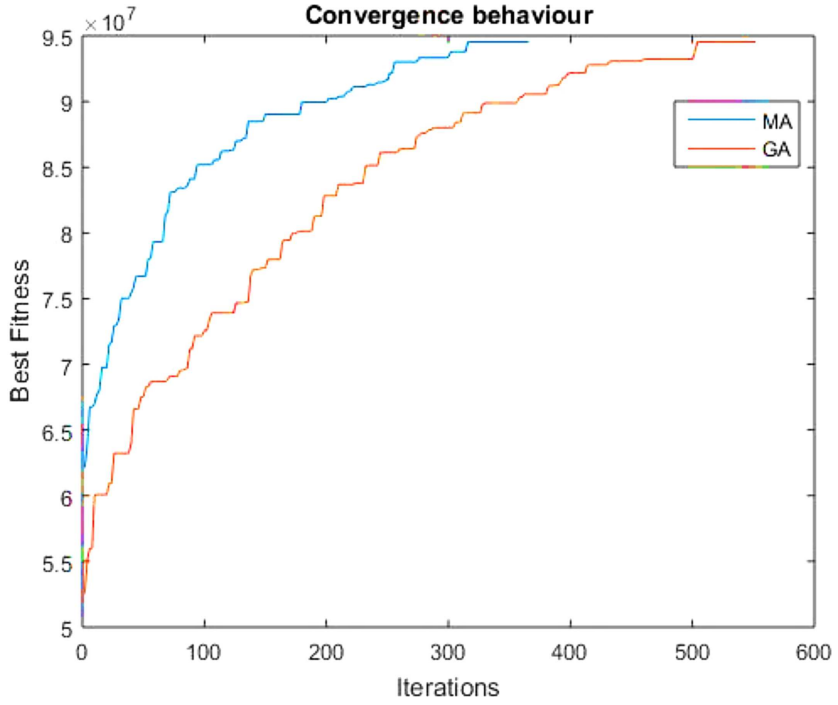

FIGURE 4. Comparison of the convergence behaviour of GA and MA.

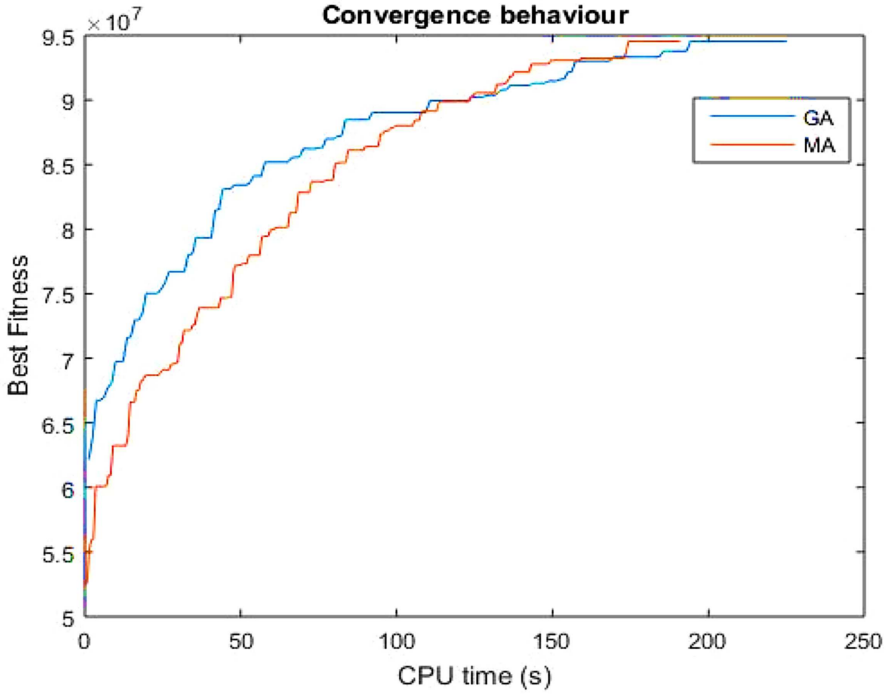

FIGURE 5. Comparison of GA and MA based on running time and best fitness value achieved.

TABLE 5. Comparison of the proposed approach with exact approach and GA for medium-size instances.

Instance	Instance details	Total demand	Exact approach		GA		MA		Improvement in objective (%)	
			Best objective value	% Solution gap	Best objective value	Average objective value	Best objective value	Average objective value	Exact approach	GA
1	Zones =	64 694	94 504	1.1	91 125	88 530	93 715	93 607	−0.84	2.76
2	20	62 874	–	–	92 775	89 826	94 064	93 885		1.37
3	Locations	143 178	–	–	214 285	174 392	214 625	214 204		0.16
4	= 42	71 483	106 840	2.0	101 879	51 789	104 348	99 419	−2.39	2.37
5		149 997	192 660	18.6	215 696	214 136	216 677	213 039	11.08	0.45
6	Zones =	78 089	–	–	115 267	114 004	116 562	58 322		1.11
7	25	81 174	–	–	114 971	114 154	118 168	117 502		2.71
8	Locations	176 673	–	–	263 478	257 665	265 651	263 629		0.82
9	= 60	91 049	–	–	129 652	127 949	133 503	131 081		2.88
10		190 006	–	–	274 332	264 918	277 939	275 952		1.30
11	Zones =	174 392	–	–	246 653	241 950	251 613	245 579		1.97
12	40	161 055	–	–	231 221	230 054	237 451	234 757		2.62
13	Locations	379 531	–	–	539 077	521 883	546 444	545 600		1.35
14	= 80	186 901	–	–	273 622	269 199	277 968	276 481		1.56
15		420 800	–	–	596 849	580 878	606 484	581 486		1.59
16	Zones =	272 854	–	–	378 332	367 176	389 511	381 539		2.87
17	50	254 087	–	–	351 887	342 875	354 253	352 979		0.67
18	Locations	594 067	–	–	837 795	829 052	855 970	855 506		2.12
19	= 110	292 656	–	–	426 530	420 980	427 560	418 427		0.24
20		636 705	–	–	897 666	867 988	916 770	890 755		2.08
21	Zones =	314 947	–	–	466 501	454 246	466 634	455 211		0.03
22	60	318 491	–	–	466 191	462 018	471 637	458 158		1.15
23	Locations	703 324	–	–	1 013 358	1 005 939	1 030 817	1 014 168		1.69
24	= 124	349 541	–	–	509 364	507 221	519 448	517 353		1.94
25		776 626	–	–	1 090 746	1 084 069	1 121 543	1 084 446		2.75

these models is carried out based on the number of stations and ambulances located and the variation in the busy probability of ambulances.

The summary of variation in calls considered across each period for different patient types is presented in Table 7. Maximum demand occurs during the day, and minimum demand is considered during the night on weekdays and weekends. Overall demand is lower during weekends compared to weekdays. Figure 6 plots the percentage difference between the actual call arrival rate and the average call arrival rate in each period. A larger percentage difference indicates that the expected call rate will be higher than the planned demand, which will lead to a decrease in ambulance availability. It can be seen that the maximum difference between the average and actual call arrival rates occurs during weekday afternoons (12.00 pm to 6 pm), as the overall demand is highest during this period. Almost 25% of the calls could be underserved during this peak demand period, while the expected overall under-coverage is about 12% of the total calls across all periods.

Table 8 presents the number of stations and ambulances located using model 1 for the input demand given in Table 7. As expected, the estimated number of ambulances required varies with the variation in demand, and the highest number of ambulances is allocated during the peak period. Also, the number of stations located follows a pattern similar to the variation in the number of ambulances. The maximum number of stations allocated is 51, with the allocation of 80 ambulances to these stations, while at the lowest demand period, 47 stations and 66 total ambulances are allocated. Variations in the number of ambulance stations can be implemented in the EMS system using temporary station locations during peak demand periods [55]. Under this approach, ambulances can initially be located at permanent stations and then relocated to a temporary station to manage

TABLE 6. Comparison of the proposed approach with GA for large-size instances.

Instance	Instance details	Total demand	GA		MA		% Gap between GA and MA
			Best objective value	Average objective in five trials	Best objective value	Average objective in five trials	
1	Zones = 80 Locations = 150	345 709	536 969	516 690	575 708	546 446	6.73
2		362 989	556 551	539 785	584 573	561 102	4.79
3		827 851	1 252 533	1 189 177	1 333 141	1 286 743	6.05
4		415 081	621 505	602 323	663 001	649 572	6.26
5		866 980	1 399 043	1 325 918	1 369 257	1 330 662	−2.18
6	Zones = 100 Locations = 160	441 446	672 864	664 561	712 563	700 750	5.57
7		444 629	694 126	671 141	720 237	697 769	3.63
8		934 001	1 411 343	1 347 745	1 538 981	1 493 137	8.29
9		491 647	685 397	645 430	802 107	779 788	14.55
10		920 042	1 287 920	1 267 128	1 469 789	1 453 901	12.37
11	Zones = 120 Locations = 200	470 669	659 592	659 379	749 388	737 595	11.98
12		512 552	828 374	811 179	783 864	781 086	−5.68
13		1 161 223	1 587 678	1 549 776	1 883 961	1 718 104	15.73
14		584 676	954 293	836 460	917 378	878 309	−4.02
15		1 089 063	1 744 428	1 653 375	1 768 342	1 543 951	1.35
16	Zones = 125 Locations = 240	522 204	758 748	746 089	846 067	789 065	10.32
17		482 946	747 045	734 157	782 902	752 639	4.58
18		1 131 377	1 841 172	1 799 957	1 863 475	1 851 694	1.20
19		527 280	758 086	752 361	871 205	859 561	12.98
20		1 204 515	1 880 338	1 779 590	1 988 320	1 939 944	5.43
21	Zones = 144 Locations = 284	532 109	865 638	847 598	857 886	820 576	−0.90
22		622 232	858 919	828 739	998 650	945 910	13.99
23		1 376 777	2 022 112	2 009 823	2 251 549	2 202 046	10.19
24		612 138	907 186	903 841	1 012 163	954 005	10.37
25		140 9636	2 145 203	2 031 059	2 266 980	2 227 874	5.37

TABLE 7. Distribution of total calls across periods and different call types for a sample instance.

Period		Type A	Type B	Type C	Total	% of Total calls
Weekday	Morning	10 899	22 432	30 764	64 095	18.54
	Afternoon	12 645	26 040	35 712	74 397	21.52
	Evening	12 510	25 748	35 309	73 567	21.28
	Night	5737	11 808	16 195	33 740	9.76
Weekend	Morning	4495	9255	12 696	26 446	7.65
	Afternoon	4479	9221	12 642	26 342	7.62
	Evening	5164	10 623	14 567	30 354	8.78
	Night	2851	5869	8048	16 768	4.85
Total calls		58 780	120 996	165 933	345 709	100

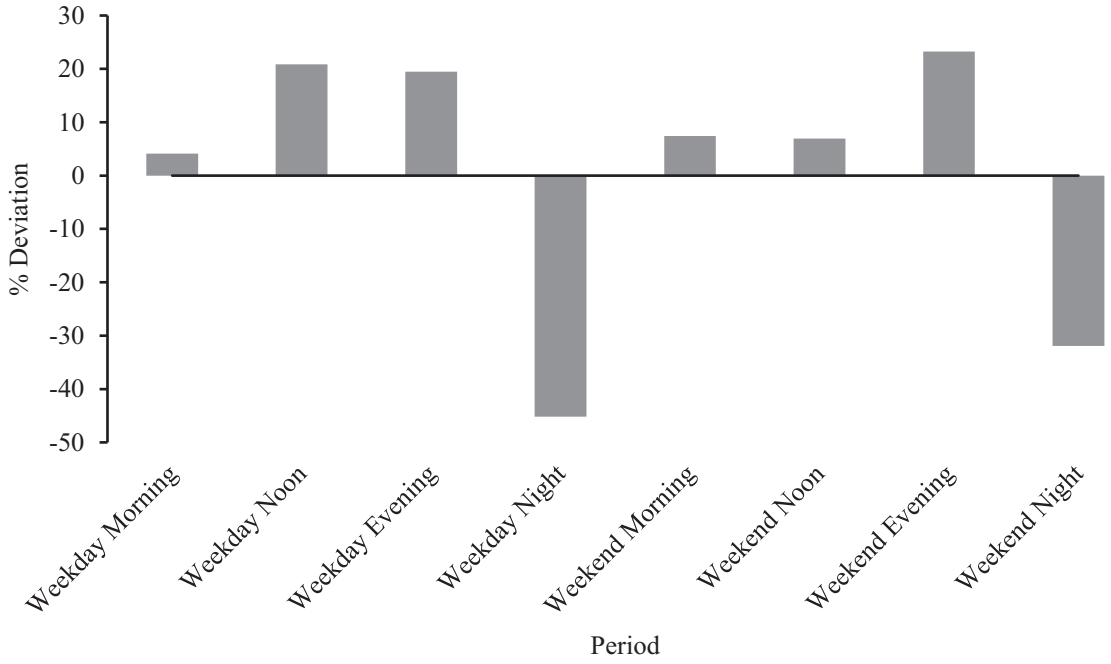

FIGURE 6. Percentage deviation between actual and average call arrival rates in each period.

TABLE 8. Number of stations and ambulances located using model 1.

Period		Total calls	Stations	Ambulances	Busy fraction
Weekday	Morning	64 095	48	73	0.1930
	Afternoon	74 397	51	80	0.2108
	Evening	73 567	50	78	0.2112
	Night	33 740	49	67	0.1170
Weekend	Morning	26 446	48	74	0.1969
	Afternoon	26 342	48	76	0.1932
	Evening	30 354	49	77	0.2186
	Night	16 768	47	66	0.1438

additional demand. Table 8 also shows the mean value of the busy fraction for each period, *i.e.* the probability that no ambulance will be available when a call arrives. It is observed that the busy probability value increases when demand is high and decreases for periods when demand is low. Thus, it indicates that despite a higher number of ambulances located, there is a higher probability that calls will be lost due to the unavailability of ambulances during the peak demand period. It can be inferred that allocating a fixed number of ambulances based on average demand as in model 2 would lead to even higher busy fractions and a higher probability of calls lost.

Figure 7 shows the comparison between the number of stations located in all three models. As model 2 assumes an average call arrival rate for all periods, it results in a significantly lower number of ambulance stations than model 3. The number of stations located by model 3 is almost equal to that of model 1 during the peak demand period. Figure 8 shows a similar comparison for the number of ambulances allocated in each period. It can be

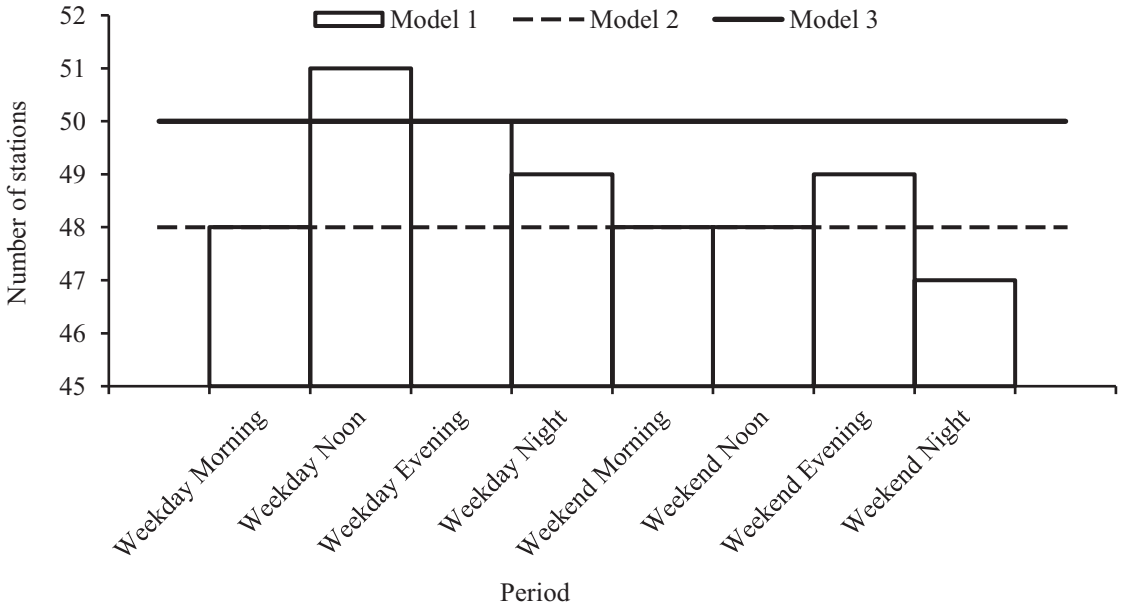

FIGURE 7. Variation in the number of stations located by model 1 compared to models 2 and 3.

observed that model 2 again results in an underestimation of ambulances during the peak period, while model 3 locates almost the same number of ambulances as in model 1 during the peak period. Another important observation is that model 1 allocates significantly fewer ambulances during low-demand periods than the other two models. Figure 8 also shows the variation in the average arrival rate of calls for each period. The number of ambulances allocated follows a pattern similar to the variation in call arrival rate, indicating model 1 assigns ambulances proportional to the demand arriving during each period.

Figure 9 depicts the comparison between the mean busy probability obtained using models 1 and 2, along with the percentage difference between both values. A positive difference in the figure indicates that model 1 has a higher busy fraction, and a negative difference indicates otherwise. The busy fraction was calculated assuming each station to be an $M/M/c$-loss queueing system. It is observed that model 1 results in a lower mean busy probability value during peak periods while slightly higher busy probability during the periods of low demand compared to model 2. Figure 10 shows the comparison between models 1 and 2 based on mean server utilisation across different periods. A pattern similar to that of busy probability is observed with slightly lower utilisation for model 1 during higher demand periods and higher utilisation during lower demand periods. This is expected as model 1 allocates fewer ambulances during the low demand period and more ambulances during other periods. This shows that using average demand throughout the day will result in significantly higher utilisation of ambulances and busy probability during peak periods. As maximum demand occurs during peak periods, this will lead to increased total lost calls.

Figure 11 shows the comparison of busy probabilities for models 1 and 3. As model 3 assumes a maximum call arrival rate for all periods, a higher number of ambulances are allocated, which results in a lower busy fraction during all periods. This low busy fraction comes at the expense of lower resource utilisation, as shown in Figure 12. Figure 12 compares the mean server utilisation for both cases throughout the day, and it is clearly seen that the utilisation is significantly higher in all periods for model 1. Thus, locating ambulances assuming maximum demand as in model 3 definitely provides a good estimate of the maximum number of stations and ambulances required with lower busy probabilities. However, taking into account temporal variation leads to better resource utilisation with a slightly higher busy probability during the lower demand periods, such as night

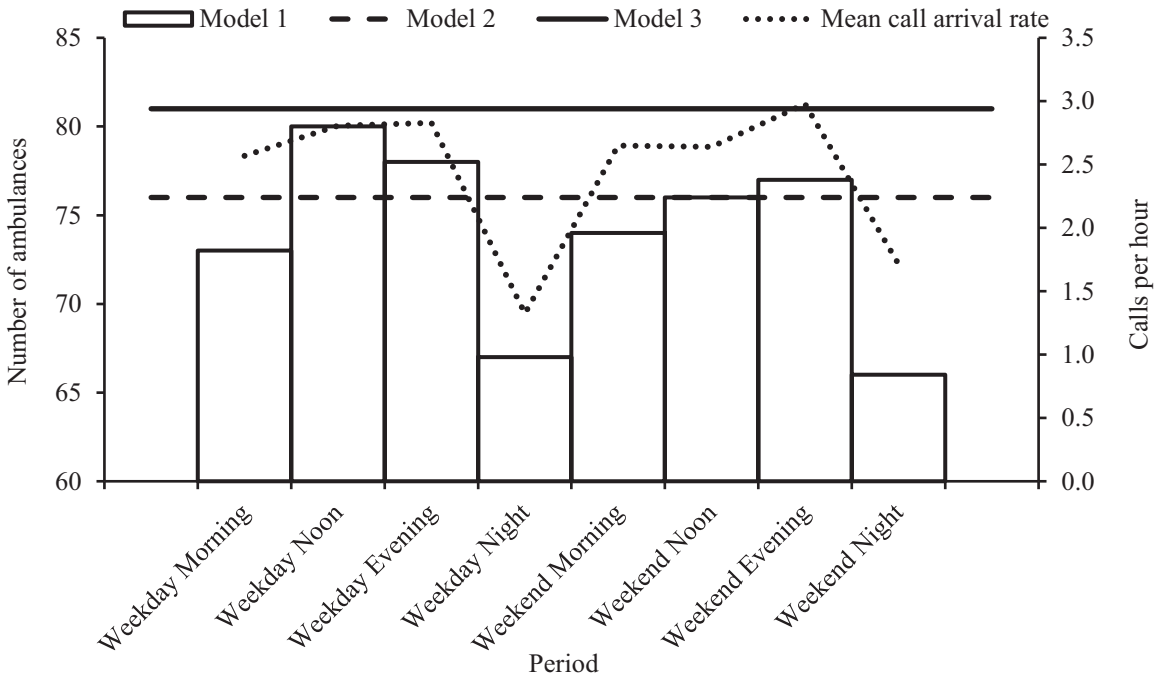

FIGURE 8. Variation in the number of ambulances located by model 1 compared to models 2 and 3.

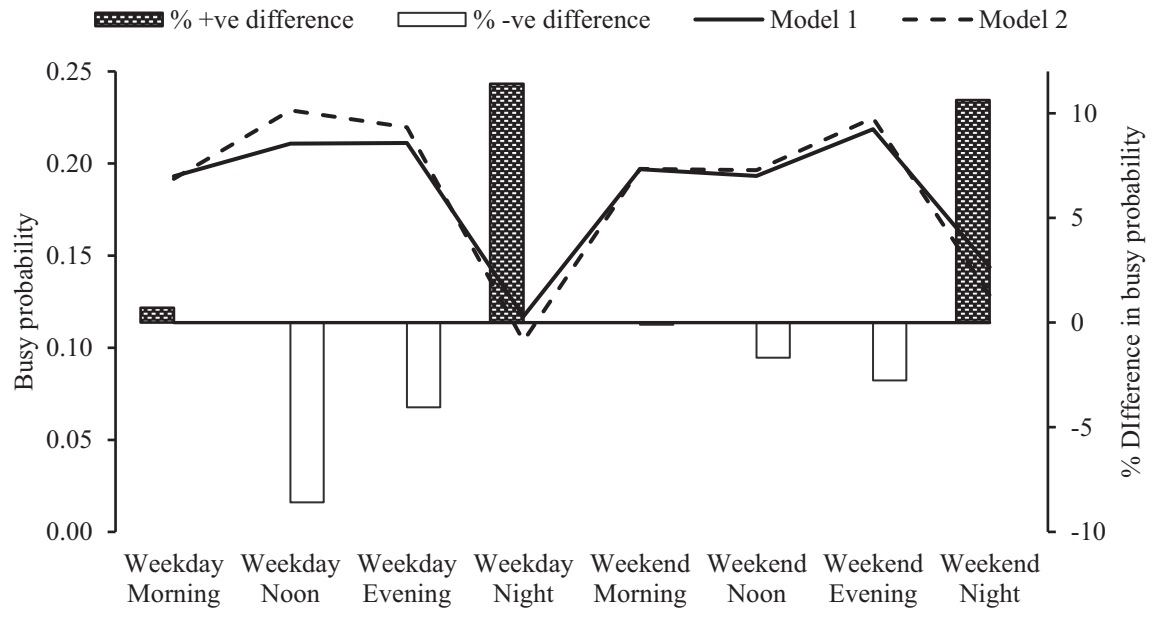

FIGURE 9. Comparison of variation in busy probability for models 1 and 2.

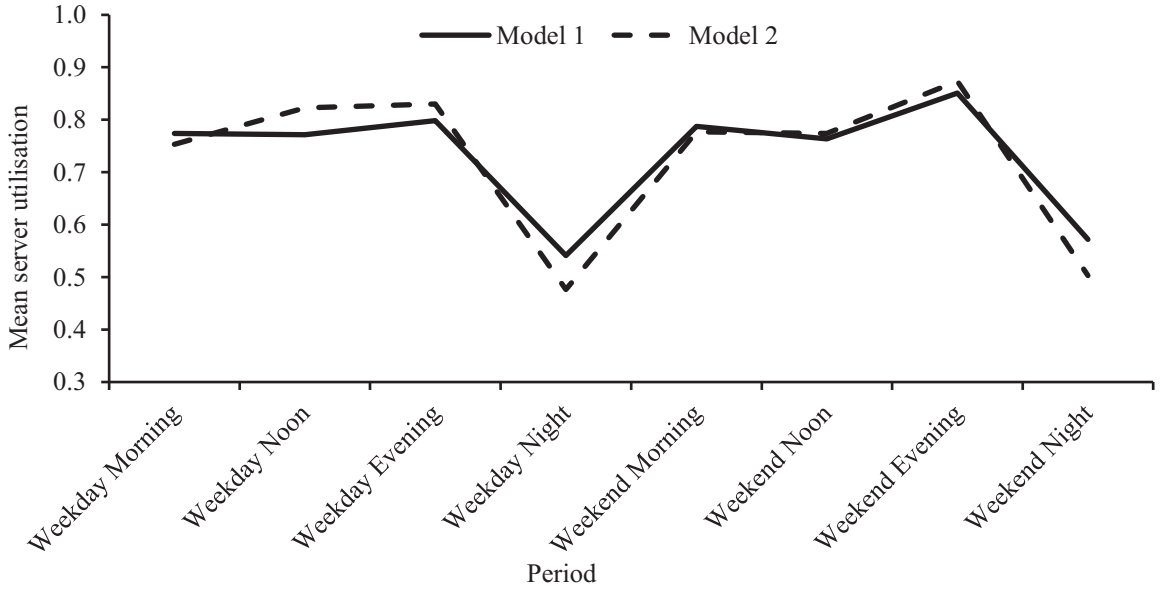

FIGURE 10. Comparison of variation in mean server utilisation for models 1 and 2.

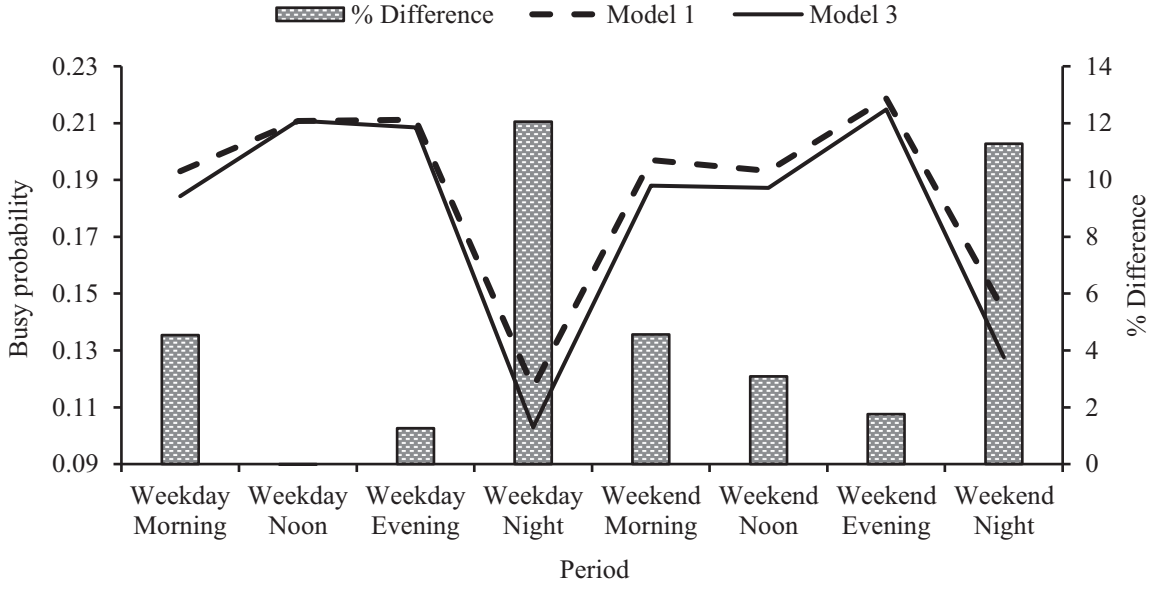

FIGURE 11. Comparison of variation in busy probability for models 1 and 3.

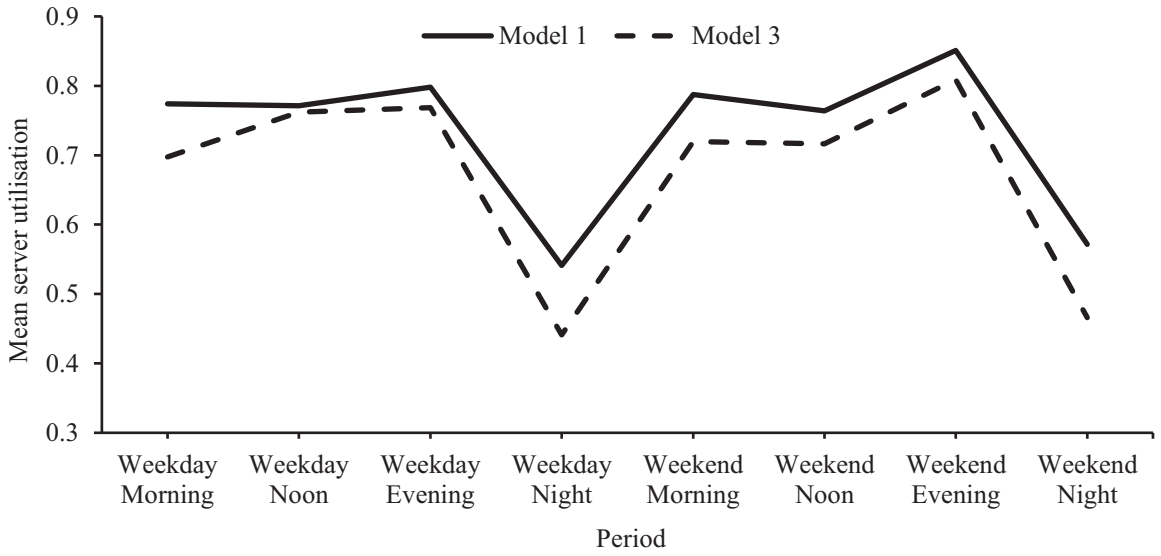

FIGURE 12. Comparison of variation in mean server utilisation for models 1 and 3.

and weekends. This is an important consideration since keeping all ambulances operational during low-demand periods will require additional ambulance crew members, resulting in significantly higher operational costs. From Figure 8, it can be observed that model 1 requires 67 ambulances to be located during weekday nights, while model 3 results in 81 ambulances, which is a significant difference. Thus, model 1 with temporal variation provides a balance between ambulance availability during peak demand periods and resource utilisation during low demand periods.

Table 9 presents the number of stations and ambulances located using model 1 during the peak demand period compared to model 2 (average demand) for large problem instances. It also presents the percentage difference between both cases. The table shows that the number of ambulances and ambulance stations located is significantly higher for model 1 during peak demand. Model 3 is not considered in comparison since the ambulance locations and the number of ambulances obtained using model 3 are similar to values obtained during the peak period using model 1. The number of ambulances required during peak demand is 3.5–15% higher than that determined assuming average demand (model 2), with an average difference of 9.6% between both cases. Table 9 also shows the percentage difference between the maximum demand and average demand values for each instance. It is interesting to note that the difference between the numbers of stations located for both cases was significantly lower and varies in the range 1–5% in most instances, with an average difference of 3.6%. Therefore, we can conclude that temporal variation in demand significantly affects the number of ambulances required, although the impact is lower on the number of stations required.

Based on our results, we can make various key conclusions that can provide insights to decision-makers (DM) of EMS systems. Our model can be applied to EMS systems where multiple ambulances are located at the same station. The model allows DM to determine the optimal sites for the ambulance stations and the number of ambulances allocated to each station. Additionally, DM can also determine the preference order of stations for each zone, *i.e.* the rank of stations. Overall, the proposed model allows DM to determine the best allocation of available ambulances to maximise the coverage and survival probability of patients of different types. Our results show that considering temporal variation in demand provides a better estimate of the number of ambulances required compared to considering average demand. At the same time, considering temporal variation results in increased utilisation of ambulances compared to considering only the peak demand. Therefore, DM should not only focus on average demand or maximum demand over the day but also consider the hourly variation while

TABLE 9. Comparison of stations and ambulances located for large problem instances.

Instance	Problem data	% Difference between peak and average demand	Number of stations located			Number of ambulances allocated		
			Model 1 (peak demand)	Model 2 (average demand)	% Difference	Model 1 (peak demand)	Model 2 (average demand)	% Difference
1	Zones = 80 Locations = 150	18.2	51	48	5.9	80	73	9.2
2		20.8	56	53	5.4	84	81	4.0
3		20.6	46	45	2.2	74	69	7.2
4		19.1	66	63	4.5	106	92	15.5
5		23.0	55	54	1.8	82	79	3.5
6	Zones = 100 Locations = 160	20.5	63	61	3.2	97	94	3.7
7		19.4	76	73	3.9	115	105	9.8
8		16.9	69	67	2.9	112	104	7.4
9		22.2	70	67	4.3	121	105	15.2
10		21.1	76	74	2.6	111	106	5.0
11	Zones = 120 Locations = 200	20.5	95	91	4.2	158	140	13.3
12		25.8	91	87	4.4	149	135	10.8
13		23.9	91	87	4.4	144	136	5.9
14		26.2	80	79	1.3	145	128	13.1
15		25.3	82	81	1.2	145	134	8.4
16	Zones = 125 Locations = 240	19.5	92	89	3.3	148	138	7.6
17		23.3	96	91	5.2	154	138	11.4
18		21.6	101	94	6.9	167	146	14.2
19		21.1	97	93	4.1	166	147	12.9
20		21.6	104	103	1.0	157	149	5.4
21	Zones = 144 Locations = 284	20.6	96	93	3.1	175	161	9.0
22		25.1	138	127	8.0	224	191	17.1
23		25.5	113	110	2.7	190	172	10.3
24		17.2	119	117	1.7	211	188	12.5
25		22.8	96	95	1.0	167	156	7.1
Average					3.6			9.6

estimating the ambulances required. An important observation from our result is that the number of ambulance stations located and the optimal location of the stations do not significantly vary over the day. However, the number of ambulances allocated to stations varies up to 15% with the variation in demand. Thus, DM can fix the optimal location of stations while varying the number of ambulances to balance the total coverage achieved and the number of relocations.

Busy probability of ambulances is affected by the variation in demand and is observed to be high during the afternoon and evening compared to the night and morning. Therefore, DM can utilise the proposed model to solve for the different maximum number of ambulances to determine the optimal number of ambulances that minimise the overall busy probability. Further, the proposed model provides the number of ambulances required during each period, which allows DM to determine the number of EMS crews required to be assigned to each station. This also allows DM to determine the schedule of the crews to operate the required number of ambulances during each period. Ultimately, we can say that our study will enable DM in the decision-making process to determine the optimal allocation of ambulance stations, ambulances, and the preference order of stations that maximise the performance of an EMS system while improving busy probability and server utilisation. Additionally, our model provides a framework for the operational level decisions of relocation of ambulances and scheduling of ambulance crews. We also demonstrate the effectiveness of the MA compared to the exact solver and GA for solving different size instances of the proposed problem. The exact solver is unable

to provide feasible solutions for the problem within a reasonable time, while the memetic algorithm converges to a good quality solution within significantly less computational time compared to both the exact solver and GA.

6. Conclusions

In this paper, we present a location model for ambulances that accounts for the time-dependent nature of EMS demand along with server-level busy probability. An MINLP model is developed that accounts for the spatio-temporal variation in demand while considering heterogeneous performance measures, namely coverage and survival function, for different call types. The presented non-linear model is then linearised using SOS2 variables and breakpoint values to transform into a MILP model. We divided the day into four blocks of six hours with different call arrival rates to account for temporal variation with the minimum call rate during the night (12.00 am to 6.00 am) and maximum call rate during the day (12.00 pm to 6.00 pm). Similarly, the variation in demand during different days of the week is also considered by dividing a week into weekdays and weekends with different call arrival rates for both cases. To test the model, we generated various test instances by dividing an urban region into different demand zones and identifying potential station locations within the region. A memetic algorithm-based solution approach is proposed since commercial MILP solver (CPLEX) could not even find feasible solutions for most medium and large-size problem instances within a reasonable time.

Our results indicate that the solution generated by the proposed memetic algorithm is within 3% of the optimal solution for all the instances for which CPLEX was able to solve to optimality and with significantly less run time than CPLEX. The proposed MA-based approach also outperformed GA in terms of the best objective value for most test instances. We compared the results obtained using the time-dependent model with two models considering average and peak call rates. Computational results show that neglecting time-dependent variation and considering an average call arrival rate can result in an underestimation of ambulances required by up to 15% during peak demand. This can lead to a loss of significant demand due to a higher busy probability during the peak period while there is excess capacity during the low demand periods. Contrarily, considering a peak arrival rate for all periods provides a better estimate of the maximum number of stations and ambulances required and results in lower busy probability but has significantly low resource utilisation for all periods. Thus, considering temporal variation balances ambulance availability during peak periods and resource utilisation for the periods with low demand. Our results illustrate the need for considering the time-varying nature of demand and also demonstrate the effectiveness of the proposed solution approach.

Although this work addresses various research gaps in ambulance location models to develop a model that is applicable for more realistic and complex EMS systems, there are some limitations of our work that can be considered for future research. In this work, we have not considered the impact of variations in demand due to seasonal factors, which could play an important role in locating ambulances. Uncertainty in travel time, especially in urban regions, is another aspect that can be explored. Incorporating uncertainty in data through fuzzy logic or grey numbers and leveraging big data to model stochastic parameters can also be applied to develop decision support systems [18,63]. Equity is another important issue in EMS planning, as heterogeneity in demand can result in an unequal allocation of resources. Incorporating equity-based objective function along with coverage and survival function-based objectives can be an interesting research direction. Developing multi-objective models with different performance measures and solution approaches to obtain Pareto solutions can be another possible direction for further research [21, 59, 60, 62]. Integrating the strategic and tactical level ambulance location problems with operational level problems, such as routing and relocation, is necessary to develop a complete framework for EMS planning decisions. Our model assumes that all ambulances are homogeneous, but the EMS systems generally utilise multiple types of ambulances. Evaluating the impact of various types of EMS vehicles is also an important aspect as it requires different dispatch policies for individual patients. Design of experiments based approach can be applied to fine-tune the parameters of the proposed solution approach. Another possible direction of further research is developing an exact solution approach exploiting the problem structure to improve the solution quality.

References

[1] A. Aghighi, A. Goli, B. Malmir and E.B. Tirkolaee, The stochastic location-routing-inventory problem of perishable products with reneging and balking. *J. Ambient Intell. Humaniz. Comput.* (2021) 1–20.

[2] M.A. Akdoğan, Z.P. Bayındır and C. Iyigun, Locating emergency vehicles with an approximate queuing model and a meta-heuristic solution approach. *Transp. Res. Part C: Emerg. Technol.* **90** (2018) 134–155.

[3] H. Andersson, T.A. Granberg, M. Christiansen, E.S. Aartun and H. Leknes, Using optimization to provide decision support for strategic emergency medical service planning – three case studies. *Int. J. Med. Inf.* **133** (2020) 103975.

[4] R. Aringhieri, M.E. Bruni, S. Khodaparasti and J.T. van Essen, Emergency medical services and beyond: addressing new challenges through a wide literature review. *Comput. Oper. Res.* **78** (2017) 349–368.

[5] R. Batta, J.M. Dolan and N.N. Krishnamurthy, The maximal expected covering location problem: revisited. *Transp. Sci.* **23** (1989) 277–287.

[6] V. Bélanger, A. Ruiz and P. Soriano, Recent optimization models and trends in location, relocation, and dispatching of emergency medical vehicles. *Eur. J. Oper. Res.* **272** (2019) 1–23.

[7] F. Borrás and J.T. Pastor, The ex-post evaluation of the minimum local reliability level: an enhanced probabilistic location set covering model. *Ann. Oper. Res.* **111** (2002) 51–74.

[8] J.J. Boutilier and T.C. Chan, Ambulance emergency response optimization in developing countries. *Oper. Res.* **68** (2020) 1315–1334.

[9] L. Brotcorne, G. Laporte and F. Semet, Ambulance location and relocation models. *Eur. J. Oper. Res.* **147** (2003) 451–463.

[10] K. Cantwell, P. Dietze, A.E. Morgans and K. Smith, Ambulance demand: random events or predicable patterns? *Emerg. Med. J.* **30** (2013) 883–887.

[11] K. Cantwell, A. Morgans, K. Smith, M. Livingston, T. Spelman and P. Dietze, Time of day and day of week trends in EMS demand. *Prehosp. Emerg. Care* **19** (2015) 425–431.

[12] S. Chanta, M.E. Mayorga, M.E. Kurz and L.A. McLay, The minimum p-envy location problem: a new model for equitable distribution of emergency resources. *IIE Trans. Healthc. Syst. Eng.* **1** (2011) 101–115.

[13] S. Chanta, M.E. Mayorga and L.A. McLay, Improving emergency service in rural areas: a bi-objective covering location model for EMS systems. *Ann. Oper. Res.* **221** (2014) 133–159.

[14] R. Church and C. ReVelle, The maximal covering location problem. In: Papers of the Regional Science Association. Vol. 32. Springer-Verlag (1974) 101–118.

[15] J. Current, M. Daskin and D. Schilling, Discrete network location models. *Facility Locat.: App. Theory* **1** (2002) 81–118.

[16] S.K. Das and S.K. Roy, An approximation approach for fixed-charge transportation-p-facility location problem. In: International Conference on Logistics and Supply Chain Management. Springer, Cham (2020).

[17] S.K. Das and S.K. Roy, Effect of variable carbon emission in a multi-objective transportation-p-facility location problem under neutrosophic environment. *Comput. Ind. Eng.* **132** (2019) 311–324.

[18] S.K. Das, S.K. Roy and G.W. Weber, Application of type-2 fuzzy logic to a multiobjective green solid transportation–location problem with dwell time under carbon tax, cap, and offset policy: fuzzy versus nonfuzzy techniques. *IEEE Trans. Fuzzy Syst.* **28** (2020) 2711–2725.

[19] S.K. Das, S.K. Roy and G.W. Weber, An exact and a heuristic approach for the transportation-p-facility location problem. *Comput. Manage. Sci.* **17** (2020) 389–407.

[20] S.K. Das, S.K. Roy and G.W. Weber, Heuristic approaches for solid transportation-p-facility location problem. *Cent. Eur. J. Oper. Res.* **28** (2020) 939–961.

[21] S.K. Das, M. Pervin, S.K. Roy and G.W. Weber, Multi-objective solid transportation-location problem with variable carbon emission in inventory management: a hybrid approach. *Ann. Oper. Res.* (2021) 1–27.

[22] M.S. Daskin, A maximum expected covering location model: formulation, properties and heuristic solution. *Transp. Sci.* **17** (1983) 48–70.

[23] S.G. Davis, Analysis of the deployment of emergency medical services. *Omega* **9** (1981) 655–657.

[24] D. Degel, L. Wiesche, S. Rachuba and B. Werners, Time-dependent ambulance allocation considering data-driven empirically required coverage. *Health Care Manage. Sci.* **18** (2015) 444–458.

[25] V.J. De Maio, I.G. Stiell, G.A. Wells, D.W. Spaite and Ontario Prehospital Advanced Life Support Study Group, Optimal defibrillation response intervals for maximum out-of-hospital cardiac arrest survival rates. *Ann. Emerg. Med.* **42** (2003) 242–250.

[26] H. Derbel, B. Jarboui, S. Hanafi and H. Chabchoub, Genetic algorithm with iterated local search for solving a location-routing problem. *Expert Syst. Appl.* **39** (2012) 2865–2871.

[27] K.F. Doerner, W.J. Gutjahr, R.F. Hartl, M. Karall and M. Reimann, Heuristic solution of an extended double-coverage ambulance location problem for Austria. *Cent. Eur. J. Oper. Res.* **13** (2005) 325–340.

[28] A. El Fallahi and I. Sefrioui, A linear programming model and memetic algorithm for the Emergency Vehicle Routing. In: 2019 4th World Conference on Complex Systems (WCCS). IEEE (2019).

[29] A. El Itani, F. Ben Abdelaziz and H. Masri, A bi-objective covering location problem: case of ambulance location in the Beirut area, Lebanon. *Manage. Decis.* **57** (2019) 432–444.

[30] E. Erkut, A. Ingolfsson and G. Erdoğan, Ambulance location for maximum survival. *Nav. Res. Log.* **55** (2008) 42–58.

[31] M. Gendreau, G. Laporte and F. Semet, Solving an ambulance location model by tabu search. *Locat. Sci.* **5** (1997) 75–88.

[32] G. Hiermann, M. Prandtstetter, A. Rendl, J. Puchinger and G.R. Raidl, Metaheuristics for solving a multimodal home-healthcare scheduling problem. *Cent. Eur. J. Oper. Res.* **23** (2015) 89–113.

[33] K. Hogan and C. ReVelle, Concepts and applications of backup coverage. *Manage. Sci.* **32** (1986) 1434–1444.

[34] M. Kaveh and M.S. Mesgari, Improved biogeography-based optimization using migration process adjustment: an approach for location-allocation of ambulances. *Comput. Ind. Eng.* **135** (2019) 800–813.

[35] V.A. Knight, P.R. Harper and L. Smith, Ambulance allocation for maximal survival with heterogeneous outcome measures. *Omega* **40** (2012) 918–926.

[36] N. Krasnogor and J. Smith, A tutorial for competent memetic algorithms: model, taxonomy, and design issues. *IEEE Trans. Evol. Comput.* **9** (2005) 474–488.

[37] R.C. Larson, A hypercube queuing model for facility location and redistricting in urban emergency services. *Comput. Oper. Res.* **1** (1974) 67–95.

[38] H. Leknes, E.S. Aartun, H. Andersson, M. Christiansen and T.A. Granberg, Strategic ambulance location for heterogeneous regions. *Eur. J. Oper. Res.* **260** (2017) 122–133.

[39] M. Lozano, F. Herrera, N. Krasnogor and D. Molina, Real-coded memetic algorithms with crossover hill-climbing. *Evol. Comput.* **12** (2004) 273–302.

[40] M.B. Mandell, Covering models for two-tiered emergency medical services systems. *Locat. Sci.* **6** (1998) 355–368.

[41] R. Manfredini, O. La Cecilia, B. Boari, J. Steliu, V. Michelini, P. Carli, C. Zanotti, M. Bigoni and M. Gallerani, Circadian pattern of emergency calls: implications for ED organization. *Am. J. Emerg. Med.* **20** (2002) 282–286.

[42] V. Marianov and C. ReVelle, The queueing maximal availability location problem: a model for the siting of emergency vehicles. *Eur. J. Oper. Res.* **93** (1996) 110–120.

[43] R. McCormack and G. Coates, A simulation model to enable the optimization of ambulance fleet allocation and base station location for increased patient survival. *Eur. J. Oper. Res.* **247** (2015) 294–309.

[44] L.A. McLay, A maximum expected covering location model with two types of servers. *IIE Trans.* **41** (2009) 730–741.

[45] L.A. McLay and M. E. Mayorga, Evaluating emergency medical service performance measures. *Health Care Manage. Sci.* **13** (2010) 124–136.

[46] P. Merz and B. Freisleben, Fitness landscape analysis and memetic algorithms for the quadratic assignment problem. *IEEE Trans. Evol. Comput.* **4** (2000) 337–352.

[47] D.M. Miranda, R.S. de Camargo, S.V. Conceição, M.F. Porto and N.T. Nunes, A metaheuristic for the rural school bus routing problem with bell adjustment. *Expert. Syst. Appl.* **180** (2021) 115086.

[48] A. Nadizadeh and A.S. Zadeh, A bi-level model and memetic algorithm for arc interdiction location-routing problem. *Comput. Appl. Math.* **40** (2021) 1–44.

[49] H.Z. Naji, M. AL-Behadili and F. AL-Maliky, Two server dynamic coverage location model under stochastic travel time. *Int. J. Appl. Comput. Math.* **7** (2021) 1–19.

[50] J. Nelas and J. Dias, Optimal emergency vehicles location: an approach considering the hierarchy and substitutability of resources. *Eur. J. Oper. Res.* **287** (2020) 583–599.

[51] H.K. Rajagopalan, F.E. Vergara, C. Saydam and J. Xiao, Developing effective meta-heuristics for a probabilistic location model via experimental design. *Eur. J. Oper. Res.* **177** (2007) 83–101.

[52] H.K. Rajagopalan, C. Saydam and J. Xiao, A multiperiod set covering location model for dynamic redeployment of ambulances. *Comput. Oper. Res.* **35** (2008) 814–826.

[53] R. Reeves, Genetic algorithms Part A: Background. **28** (2003).

[54] R. Reeves, A genetic algorithm for flowshop sequencing. *Comput. Oper. Res.* **22** (1995) 5–13.

[55] J.F. Repede and J.J. Bernardo, Developing and validating a decision support system for locating emergency medical vehicles in Louisville, Kentucky. *Eur. J. Oper. Res.* **75** (1994) 567–581.

[56] M. Reuter-Oppermann, P.L. van den Berg and J.L. Vile, Logistics for emergency medical service systems. *Health Syst.* **6** (2017) 187–208.

[57] C. ReVelle and K. Hogan, The maximum availability location problem. *Transp. Sci.* **23** (1989) 192–200.

[58] V. Schmid and K.F. Doerner, Ambulance location and relocation problems with time-dependent travel times. *Eur. J. Oper. Res.* **207** (2010) 1293–1303.

[59] L. Shaw, S.K. Das and S.K. Roy, Location-allocation problem for resource distribution under uncertainty in disaster relief operations. *Soc.-Econ. Planning Sci.* **82** (2022) 101232.

[60] B. Tirkolaee and N.S. Aydin, Integrated design of sustainable supply chain and transportation network using a fuzzy bi-level decision support system for perishable products. *Expert. Syst. Appl.* **195** (2022) 116628.

[61] E.B. Tirkolaee, A. Goli and A. Mardani, A novel two-echelon hierarchical location-allocation-routing optimization for green energy-efficient logistics systems. *Ann. Oper. Res.* (2021) 1–29.

[62] B. Tirkolaee, A. Goli, S. Gütmen, G.W. Weber and K. Szwedzka, A novel model for sustainable waste collection arc routing problem: pareto-based algorithms. *Ann. Oper. Res.* (2022) 1–26.

[63] E.B. Tirkolaee and A.E. Torkayesh, A cluster-based stratified hybrid decision support model under uncertainty: sustainable healthcare landfill location selection. *Appl. Intell.* (2022) 1–20.

[64] C. Toregas, R. Swain, C. ReVelle and L. Bergman, The location of emergency service facilities. *Oper. Res.* **19** (1971) 1363–1373.

[65] H. Toro-Díaz, M.E. Mayorga, S. Chanta and L.A. McLay, Joint location and dispatching decisions for emergency medical services. *Comput. Ind. Eng.* **64** (2013) 917–928.

[66] H. Toro-Díaz, M.E. Mayorga, L.A. McLay, H.K. Rajagopalan and C. Saydam, Reducing disparities in large-scale emergency medical service systems. *J. Oper. Res. Soc.* **66** (2015) 1169–1181.
[67] P.L. Van Den Berg and K. Aardal, Time-dependent MEXCLP with start-up and relocation cost. *Eur. J. Oper. Res.* **242** (2015) 383–389.
[68] H.P. Williams, Model Building in Mathematical Programming. John Wiley & Sons (2013).
[69] E. Yadegari, A. Alem-Tabriz and M. Zandieh, A memetic algorithm with a novel neighborhood search and modified solution representation for closed-loop supply chain network design. *Comput. Ind. Eng.* **128** (2019) 418–436.
[70] X. Yan and X. Gu, Bi-objective no-wait multiproduct multistage product scheduling problem with flexible due dates based on MOIDE-MA. *Comput. Oper. Res.* **137** (2022) 105543.
[71] X. Yang, N. Bostel and P. Dejax, A MILP model and memetic algorithm for the hub location and routing problem with distinct collection and delivery tours. *Comput. Ind. Eng.* **135** (2019) 105–119.
[72] S. Yoon and L.A. Albert, An expected coverage model with a cutoff priority queue. *Health Care Manage. Sci.* **21** (2018) 517–533.
[73] S. Yoon and L.A. Albert, Dynamic dispatch policies for emergency response with multiple types of vehicles. *Transp. Res. Part E: Logistics Transp. Rev.* **152** (2021) 102405.
[74] S. Yoon, L.A. Albert and V.M. White, A stochastic programming approach for locating and dispatching two types of ambulances. *Transp. Sci.* **55** (2021) 275–296.
[75] Z. Zhang, M. Liu and A. Lim, A memetic algorithm for the patient transportation problem. *Omega* **54** (2015) 60–71.
[76] L. Zhen, K. Wang, H. Hu and D. Chang, A simulation optimization framework for ambulance deployment and relocation problems. *Comput. Ind. Eng.* **72** (2014) 12–23.