Computers & Operations Research 159 (2023) 106355

Contents lists available at ScienceDirect

Computers and Operations Research

journal homepage: www.elsevier.com/locate/cor

Adaptive variable neighbourhood search approach for time-dependent joint location and dispatching problem in a multi-tier ambulance system

Raviarun A. Nadar [a,*], J.K. Jha [a], Jitesh J. Thakkar [b]

[a] Department of Industrial and Systems Engineering, Indian Institute of Technology Kharagpur, Kharagpur 721 302, West Bengal, India
[b] Gati Shakti Vishwavidyalaya, Vadodara 390 004, Gujarat, India

ARTICLE INFO

Keywords:
Location
Ambulance planning
Variable neighbourhood search
Joint location and dispatching

ABSTRACT

Location-allocation of ambulances is a critical planning problem in the efficient operation of an emergency medical services system. This work considers a joint ambulance location and dispatch problem for a multi-tier ambulance system. The proposed problem addresses three key decisions: the location of ambulance stations, allocation of ambulances to these stations, and the preference order of stations for dispatching ambulances. We consider various other important factors, such as temporal variation in demand and travel time, station-specific ambulance busy probabilities and possible relocation of ambulances over the day. A mixed-integer non-linear programming model is formulated with a survival probability-based objective function to represent the problem. A queueing-based iterative approximation approach is presented to estimate the station-level dispatch probability of ambulances. We propose an adaptive variable neighbourhood search approach to solve the problem that utilises the approximation approach to obtain the objective function. A relaxed station location problem combined with a particle swarm approach for ambulance allocation is solved to obtain an initial solution. The effectiveness of the proposed approach is validated using a dataset generated based on the city of Kolkata in India.

1. Introduction

Emergency Medical Services (EMS) systems form an essential component of the healthcare infrastructure across the world. EMS play a significant role in reducing fatalities and injuries through the provision of timely out-of-hospital medical care and transportation services to patients in need of emergency. Many EMS have target response time requirements within which all calls must be served. The optimal location of ambulance stations and allocation of ambulances to these stations can enable EMS to achieve their response time requirements. Dispatch policies are another important decision that affects the performance of an EMS system (Enayati et al., 2019). Ambulance dispatch refers to the process of assigning an ambulance to a specific emergency call, and a dispatch policy could be formed by combinations of multiple dispatch methods (Toro-Díaz et al., 2013). While most researchers have focused on models to optimise ambulance location planning decisions, only a few works related to dispatching decisions have been presented. Toro-Díaz et al. (2013) combined the location-allocation of ambulances with the dispatching problem by not assuming a fixed dispatch policy, thus allowing the model to determine using a preference order for stations.

Variation in demand and travel time is an important consideration while determining the optimal locations of ambulances. Demand for emergency medical services varies spatially across different regions and temporally across day hours (Cantwell et al., 2015). The travel time required for ambulances to reach the ambulance location also varies over the day (Schmid and Doerner, 2010). The impact of this variation in travel time can be significant on time taken to reach patient locations and further transport them to the nearest hospital. In urban locations, especially in developing countries, high population density and narrow roads significantly impact travel time variation on ambulances as vehicles cannot easily yield to emergency vehicles (Boutilier and Chan, 2020). Therefore, various researchers have developed time-dependent models for ambulance locations that incorporate the temporal variation in demand and travel time (Rajagopalan et al, 2008; Schmid and Doerner, 2010; Van Den Berg and Aardal, 2015). These time-dependent models divide the day into a number of periods associated with separate demand or travel time or both, allowing different allocations of ambulances for each period. Coverage is the most commonly utilised measure in ambulance location models to evaluate the performance of EMS. Coverage considers a location covered if an ambulance can reach that

* Corresponding author.
E-mail address: ravi1989.06@gmail.com (R.A. Nadar).

https://doi.org/10.1016/j.cor.2023.106355
Received 1 March 2022; Received in revised form 15 May 2023; Accepted 12 July 2023
Available online 17 July 2023

location within a pre-specified response time limit. However, due to its binary nature, coverage is an inaccurate measure to account for the impact of variation in response time. To overcome this limitation, an objective function based on a continuous non-linear survival function was proposed by Erkut et al. (2008).

Advanced life service (ALS) and Basic life service (BLS) ambulances are the two most common ambulance types employed in an EMS system (Yoon et al., 2021). ALS vehicles usually have at least one or more paramedics and advanced equipment to respond to life-threatening emergencies. Although ALS can respond to any type of emergency, it is expensive to operate due to the personnel and equipment costs. BLS ambulances usually employ emergency medical technicians and may not possess all the equipment of the ALS. However, BLS ambulances can respond faster to emergencies as they do not need time to prepare the equipment and are also cheaper to operate. Another commonly used type of ambulance is the non-transporting first responder vehicle (FRV), which lacks the ability to transport patients in a supine state or provide them with emergency care inside the vehicle. EMS that operate only ALS ambulances for all purposes are called single-tiered or all-ALS systems. Although many ambulance location models assume that all ambulances are homogeneous and can serve all patients, in reality, EMS are multi-tiered. Multi-tiered EMS operate multiple types of ambulances to serve different patient types. For example, an EMS can use a combination of ALS, BLS and FRV type vehicles. In such multi-tier systems, triage is performed when a call is received to determine the emergency level of the patient and the corresponding level of the ambulance is dispatched. Various researchers have modelled the location of ambulances considering multiple vehicle types and patient types (Boujemaa et al., 2018; McLay, 2009; Naji et al., 2021).

One major factor that needs to be considered while locating ambulances is the probabilistic nature of ambulance availability, as ambulances may be busy when a call arrives. Early ambulance location models assume that ambulances are always available (Bélanger et al., 2019). Researchers have relaxed this assumption and presented various probabilistic models (Daskin, 1983; Ingolfsson et al., 2008). One of the major developments is the hypercube queueing model (HQM) introduced by Larson (1974), which applies the spatial queueing representation to evaluate the performance of EMS. Some researchers have integrated the HQM into ambulance location models to evaluate the performance of EMS and account for ambulance availability (Geroliminis et al., 2009; Iannoni et al., 2008). Another approach is to embed an approximate hypercube queueing model-based approach within the optimisation framework (Saydam and Aytuğ, 2003). Simulation has also been used to compute busy probability and system performance within an ambulance location model (Bélanger et al., 2020; Lee et al., 2012; McCormack and Coates, 2015). In this paper, we apply an iterative queueing-based approximation approach similar to Knight et al. (2012) that considers each ambulance station as an $M/M/s$ server. The proposed approach allows computing arrival rate, service rate and busy probability for each station individually while accounting for temporal variations in arrival rate and service time.

We consider a joint ambulance location and dispatching problem that aims to determine the optimal location of ambulance stations, allocation of ambulances to the selected stations and the preference order of stations for each zone. Thus, we do not assume a dispatch policy but treat it as an output of the optimisation model. The model also accounts for different types of ambulances, namely ALS, BLS and FRV, which serve two types of calls. We also consider the temporal variation in demand and travel time by dividing the day into multiple periods. A non-linear survival function based on travel time is used as the objective function to capture the impact of variation in travel time accurately. The arrival rate and service time of ambulances of each type are explicitly calculated for each ambulance station to obtain the busy probability of ambulances. We also calculate the dispatch probability of ambulances from any given station to a patient location. We propose an approximate approach based on queueing equation to obtain the busy probability of ambulances. The proposed problem is modelled as a mixed-integer non-linear programming (MINLP) model. We also propose a hybrid solution approach based on adaptive variable neighbourhood search (AVNS) that uses an initial solution obtained by solving a relaxed ambulance station location problem. We contribute to the literature by proposing a more realistic ambulance location by combining various issues to model the problem and provide an efficient solution approach to solve the problem.

The organisation of the paper is as follows. Section 2 presents a brief review of the existing literature related to the ambulance location and dispatching problem. Section 3 introduces the problem and presents the mathematical formulation of the proposed problem. An approximation-based approach to calculate the dispatch probability of ambulances is presented in Section 4. The proposed AVNS-based solution approach is described in Section 5, followed by the computational results in Section 6. We then conclude with a brief discussion and conclusions in Section 7.

2. Review of existing literature

The critical nature of ambulance planning in emergency services has resulted in significant attention from researchers on ambulance location and related planning problems. ReVelle et al. (1977) present one of the early reviews of the literature on EMS location models focusing on maximal covering location models. Brotcorne et al. (2003) and Goldberg (2004) are other review articles on EMS planning problems. Some of the most recent reviews of EMS planning problems include Aringhieri et al. (2017), Reuter-Oppermann et al. (2017) and Bélanger et al. (2019). In this paper, we present a review of articles that address ambulance location problems and are closely related to our work.

The location set covering problem introduced by Toregas et al. (1971) is one of the early ambulance location models, along with the maximum covering location problem (MCLP) by Church and ReVelle (1974). These early models are static location models and do not consider the probabilistic nature of servers being busy when an emergency call arrives. Larson (1974) introduces the HQM that applies queueing theory to analyse an EMS based on performance measures such as congestion and coverage. Larson (1975) presents an approximate approach for the HQM, as exact solutions for the HQM were computationally prohibitive. Daskin (1983) extends the MCLP by incorporating the probabilistic nature of ambulance availability to introduce the maximum expected covering location problem (MEXCLP). ReVelle and Hogan (1989) introduce the maximal ambulance location problem (MALP) that extends the MCLP by assigning ambulances to ensure a predetermined level of ambulance availability. They first present the MALP-I, which assumes equal busy probability for all ambulances, which they relax to develop the MALP-II with zonal level busy probability. Gendreau et al. (1997) introduce the double standard model (DSM) that maximises the demand covered by at least two ambulances for each location. The early models discussed above often made various assumptions regarding the underlying system to simplify the model. Therefore, various extensions and improvements of these models have been presented, relaxing these assumptions to consider more realistic scenarios.

Variation in demand and travel time over the day is an important factor that has been considered in the literature by various researchers. Repede and Bernardo (1994) present the maximal expected coverage location model with time variation (TIMEXCLP), which is integrated with a simulation model that allows for reallocating ambulances over time based on the variations in demand pattern. Rajagopalan et al. (2008) introduce the dynamic available coverage location model (DACL) that divides the day into multiple periods while minimising the ambulances needed to meet the required level of ambulance availability. Ingolfsson et al. (2008) consider a model for ambulance location that accounts for the uncertainty in the response time through randomness in the delay of dispatch and the travel time required to reach the emergency location. Schmid and Doerner (2010) develop the multi-period

DSM that divides the day into multiple periods to consider temporal variation in travel time and allows for the relocation of ambulances to ensure a predetermined coverage level throughout the planning horizon. Their results show that ignoring temporal variation in travel time results in an overestimation of overall coverage. Saydam et al. (2013) develop the dynamic redeployment model that considers minimising the number of relocations of ambulances required along with the number of ambulances required, which allows them to allocate a more balanced ambulance allocation throughout the day. Van Den Berg and Aardal (2015) extend the TIMEXCLP to consider temporal variation in demand and travel time while incorporating the cost of opening ambulance stations and relocating ambulances. Through the application of their model, they show that time-dependent variation in demand and travel time results in better solutions. Boutilier and Chan (2020) model the problem for simultaneous location and routing of ambulances in Bangladesh while accounting for the variation in travel time and demand. Based on our review, we observe that among the articles that consider temporal variation, server-level busy probability of ambulances is not considered by any of these studies. Another limitation of these articles is that only coverage- and response time-based objective functions have been considered.

Most of the ambulance location models utilise coverage or response time as the performance measure of the system (McLay and Mayorga, 2010). To overcome the limitations due to the binary nature of coverage, Erkut et al. (2008) introduce the maximal survival location problem (MSLP) that considers an objective function based on the survival probability of patients as a function of travel time. Their results show that survival function-based models provide solutions that can save more lives while being slightly intractable compared to coverage-based models. The maximal expected survival location model for heterogeneous patients (MESLMHP) presented by Knight et al. (2012) extends the MSLP to combine coverage and survival probability objectives for heterogeneous patient types. Their results show that considering heterogeneous performance measures and survival probability-based approach instead of average response time or a single patient class results in better allocation of ambulances. McLay and Mayorga (2010) compare different performance measures in an EMS system for ambulance locations. They conclude that response time-based measures can achieve better survival probability or equity by using low or high response time requirements. Leknes et al. (2017) present a model that considers heterogeneous performance measures for different call types while incorporating station-specific busy probabilities. Andersson et al. (2020) extend the work of Leknes et al. (2017) by considering multiple periods to account for temporal variation in demand. Other than coverage and survival probability, equity is another objective considered in ambulance location literature (Chanta et al., 2011; Chanta et al., 2014). Chanta et al. (2011) present the minimum p-envy location problem (MpELP), which defines equity as a function of distance from the nearest station to the demand zone and its backup station. Chanta et al. (2014) present an improvement to the model that defines envy as a function of the difference between survival probabilities among different zones. Khodaparasti et al. (2016) and Enayati et al. (2019) present approaches that consider a trade-off between efficiency and equity-based measures.

Toro-Díaz et al. (2013) introduce the joint location and dispatching problem that integrates the HQM model with a genetic algorithm-based framework. The joint location and dispatch model does not assume a fixed nearest dispatch policy and instead determines the preference order of ambulances for each demand zone within the optimisation model. Toro-Díaz et al. (2015) incorporated an approximate queueing sub-model within a tabu-search-based heuristic for a joint location and dispatch model that maximises fairness in mean response time and workload of servers. Enayati et al. (2019) present a multi-objective model that evaluates the trade-off between equity and efficiency using a joint location and dispatching model. Bélanger et al. (2020) present a recursive simulation-optimisation approach that combines the location and dispatch problem with a simulation model to calculate the availability of ambulances. Nadar et al. (2021) consider a joint ambulance location and dispatch problem under temporal variation in demand and travel time using a survival probability-based objective function. They show that ignoring temporal variation in travel time overestimates coverage and survival probability while underestimating the number of ambulances required during the periods of peak demand. From our review, we observe that the researchers have not considered different ambulance types in the joint location and dispatching problem.

One of the major assumptions of many ambulance location models is the homogeneity of ambulances and the ability of any ambulance to serve all patients. However, EMS usually employ multiple types of ambulances, and these ambulances types perform different responsibilities and often serve different types of patients. Mandell (1998) presents an ambulance location model that considers ALS and BLS as two types of ambulances available and defines coverage based on the time taken for an ALS to respond to a call from a given location. McLay (2009) introduces MEXCLP2, which extends MEXCLP to consider two types of servers and incorporates interdependency between these vehicles using the HQM. They show that non-transporting quick-response vehicles can be effectively used to improve system performance in a cost-effective manner. Liu et al. (2016) present a double standard model for the location of ALS and BLS ambulances, where ALS can serve all calls while BLS serve only high-priority calls. Chong et al. (2016) present a Markov decision process for determining dispatch decisions and an integer programming model for ambulance location in a multi-tiered EMS system of ALS and BLS. Their analysis shows that multi-tiered systems can offer comparable performance to all-ALS systems or even outperform them in some scenarios. Boujemaa et al. (2018) develop a stochastic optimisation model for ambulance location in a two-tier ambulance system under demand uncertainty. Naji et al. (2021) present a dynamic coverage model for the location of ambulances, considering two types of servers and time-dependent variation in travel time. Nelas and Dias (2020) present an ambulance location model that accounts for the hierarchy of different ambulance types and the substitutability of different ambulances based on the type of care they can provide. We observe that among the literature on multiple ambulance types and patient types, the articles considering server-level busy probability do not take into account the temporal variation, while articles that consider temporal variation do not consider server-level busy probability.

Based on our review of the existing literature, we conclude that there is a need for a realistic location model that considers multiple ambulance types and accounts for the temporal variation in demand and travel time. It is also necessary to take into account the station-level arrival rate, service rate and ambulance availability. In this work, we propose a joint ambulance location and dispatch model to address these research gaps by considering multiple patient types, ambulance types, and temporal variations in demand and travel time. We consider a survival function-based objective function that can more effectively capture the impact of travel time variation over the day. We also consider station-specific ambulance busy probability and interdependency between ambulance types.

3. Problem description

We consider a region divided into a set of demand zones with associated demand for ambulances. A set of potential sites for the location of ambulance stations is considered within the selected region. We consider three types of ambulances, ALS, BLS and FRV, that can be located at any station. Patients are classified into two types based on the severity and urgency of calls. Type A call is assumed to be urgent and critical, which requires the dispatch of an ALS type of ambulance. Whereas type B call represents urgent but non-life-threatening, this preferably requires the dispatch of BLS type ambulance. If all nearby ALS ambulances are busy when a type A call arrives, then it can be served by an available BLS ambulance. Similarly, if all BLS type

ambulances are unavailable when a type B call arrives, then FRV is dispatched to serve the demand. The key decisions of the proposed problem are to determine the optimal location of ambulance stations from the available potential sites, allocate the available ambulances to the selected ambulance stations and determine the preference order of ambulance stations for each demand zone. The objective is to maximise the total survival probability of patients.

The preference order of ambulance stations for ambulance dispatch is given by the rank r for each demand zone. The station with rank $r = 1$ has the first preference for dispatching an ambulance to the demand zone, followed by the station with rank $r = 2$, and so on. The day is divided into multiple periods, with the arrival rate for demand and travel time varying over each period. The number of stations selected and the number of ambulances allocated are also allowed to vary over the day based on their availability. Ambulances are allowed to be relocated from one station to another during each period. Fig. 1 summarises the overall output of the ambulance location problem. Demand zones, potential station locations, and the number of ambulances available are provided as the input to the model, along with the travel time between demand zones and potential stations. The ambulance location model selects the optimal location for ambulance stations, allocates ambulances to the selected stations, and determines the preference order of stations for each demand zone indicated by the rank of each station.

3.1. Mathematical model

This section presents the formulation of the proposed ambulance location and dispatching problem.

3.1.1. Notation

Sets	
I	Set of demand zones,$i \in I$
J	Set of potential sites for ambulance stations,$j \in J$
T	Set of periods,$t \in T$
R	Set of rank of ambulance stations,$r \in R$
K	Set of types of ambulances,$k \in K$
L	Set of type of patients representing different types of calls,$l \in L$
Parameters	
D_{it}^{l}	total demand of type l associated with demand zone i during period t
P_{ijt}^{kl}	performance measure (survival probability function) for call type l from zone i if served by an ambulance type k from station j during period t
W_{k}^{l}	performance weight, if patient type l is served by ambulance type k
A_{t}^{k}	total number of ambulances of type k available during period t
Z_{t}	maximum number of ambulance stations that can be located during period t
A_{max}^{k}	maximum number of ambulances of type k that can be assigned to a single station
λ_{it}^{l}	arrival rate of call type l received from demand zone i during period t
R_{ijt}^{kl}	response time to reach demand zone i for an ambulance of type k located at station j during period t to serve call type l
τ_{ijt}^{l}	service time required to serve a call of type l from demand zone i using an ambulance at station j during period t
L	maximum load (total service time per ambulance) allocated to an ambulance in each period
Variables	
x_{jt}	1, if an ambulance station is located at location j during period t 0, otherwise
y_{jt}^{k}	number of ambulances of type k allocated to station j during period t
d_{ijrt}^{kl}	dispatch probability of ambulance type k from station j with rank r to serve call type l from zone i during period t
δ_{ijrt}	1, if station j is assigned rank r for demand zone i during period t, 0, otherwise
r_{jht}^{k}	number of ambulances of type k relocated from station j to station h during period t
π_{jt}^{k}	probability that all ambulances of type k are busy at station j during period t

3.1.2. Model formulation

$$\text{Maximise} \sum_{i\in I}\sum_{j\in J}\sum_{r\in R}\sum_{l\in L}\sum_{k\in K}\sum_{t\in T} P_{ijt}^{kl} W_{k}^{l} D_{it}^{l} d_{ijrt}^{kl} \tag{1}$$

Subject to

$$\sum_{j\in J} x_{jt} \le Z_t \ \forall t \in T \tag{2}$$

$$\sum_{j\in J} y_{jt}^{k} \le A_{t}^{k} \ \forall t \in T, k \in K \tag{3}$$

$$y_{jt}^{k} \le A_{max}^{k} x_{jt} \ \forall j \in J, k \in K, t \in T \tag{4}$$

$$\sum_{j\in J}\sum_{r\in R} \delta_{ijrt} = 1 \forall i \in I, t \in T \tag{5}$$

$$\sum_{r\in R} \delta_{ijrt} \le x_{jt} \forall i \in I, j \in J, t \in T \tag{6}$$

$$d_{ijrt}^{kl} \le \delta_{ijrt} \ \forall i \in I, j \in J, k \in K, l \in L, r \in R, t \in T \tag{7}$$

$$\sum_{j\in J} d_{ij(r-1)t}^{kl} \ge \sum_{j\in J} d_{ijrt}^{kl} \ \forall i \in I, r \in R | r > 1, k \in K, l \in L, t \in T \tag{8}$$

$$\pi_{jt}^{k} = f\left(\lambda_{it}^{l}, \tau_{ijt}^{l}, y_{jt}^{k}, \delta_{ijrt}\right) \ \forall i \in I, j \in J, k \in K, l \in L, r \in R, t \in T \tag{9}$$

$$d_{ijrt}^{kl} = f\left(\pi_{jt}^{k}, \delta_{ijrt}\right) \ \forall i \in I, j \in J, k \in K, l \in L, r \in R, t \in T \tag{10}$$

$$\sum_{i\in I}\sum_{r\in R} \lambda_{it}^{l} \tau_{ijt}^{l} d_{ijrt}^{kl} \le L y_{jt}^{k} \ \forall j \in J, k \in K, l \in L, t \in T \tag{11}$$

$$y_{jt}^{k} + \sum_{h\in J} r_{jht}^{k} - \sum_{h\in J} r_{jh(t+1)}^{k} = y_{j(t+1)}^{k} \ \forall j \in J, k \in K, t \in T | t < |T| \tag{12}$$

$$y_{j|T|}^{k} + \sum_{h\in J} r_{jh|T|}^{k} - \sum_{h\in J} r_{jh1}^{k} = y_{j1}^{k} \ \forall j \in J, k \in K \tag{13}$$

$$d_{ijrt}^{kl}, \pi_{jt}^{k} \in [0,1] \ \forall i \in I, j \in J, k \in K, l \in L, r \in R, t \in T \tag{14}$$

$$x_{jt}, \delta_{ijrt} \in \{0,1\} \ \forall i \in I, \forall j \in J, \forall r \in R, \forall t \in T \tag{15}$$

$$y_{jt}^{k}, r_{jht}^{k} \in \mathbb{Z}^{+} \ \forall j \in J, h \in J, k \in K, t \in T \tag{16}$$

The objective function in (1) maximises the total survival probability of all patients in all demand zones during the entire planning horizon. The survival probability is represented by the performance measure parameter P_{ijt}^{kl} as a function of the response time given below.

$$P_{ijt}^{kl} = \frac{1}{1 + e^{-0.679 + 0.262 R_{ijt}^{kl}}} \ \forall i \in I, j \in J, k \in K, l \in L, t \in T \tag{17}$$

The above survival probability function in (17) was developed by De Maio et al. (2003) for out-of-hospital cardiac arrest patients, which has been applied by Leknes et al. (2017) and Nadar et al. (2021) to ambulance location problems. However, any similar monotonically decreasing function can be utilised, as the optimal station location is insensitive to the parameters of the survival function (Erkut et al., 2008). The parameter W_{k}^{l} is used to capture the impact of patient type l being served by ambulance type k. Together, W_{k}^{l} and P_{ijt}^{kl}, both these parameters capture the impact of ambulance type on the survival probability of patients.

Constraint (2) represents the maximum number of ambulance stations that can be selected during each period. Similarly, constraint (3) presents the limit on the total number of ambulances of each type that can be allocated in a period. Constraint (4) limits the maximum number of ambulances of each type located at any station. Each demand zone is assigned to exactly one station for each rank in each period, as shown in constraint (5). Constraint (6) ensures that only locations selected as a station can be assigned to a demand zone. Constraint (7) ensures that demand from a zone can be served by a station that is assigned to a rank

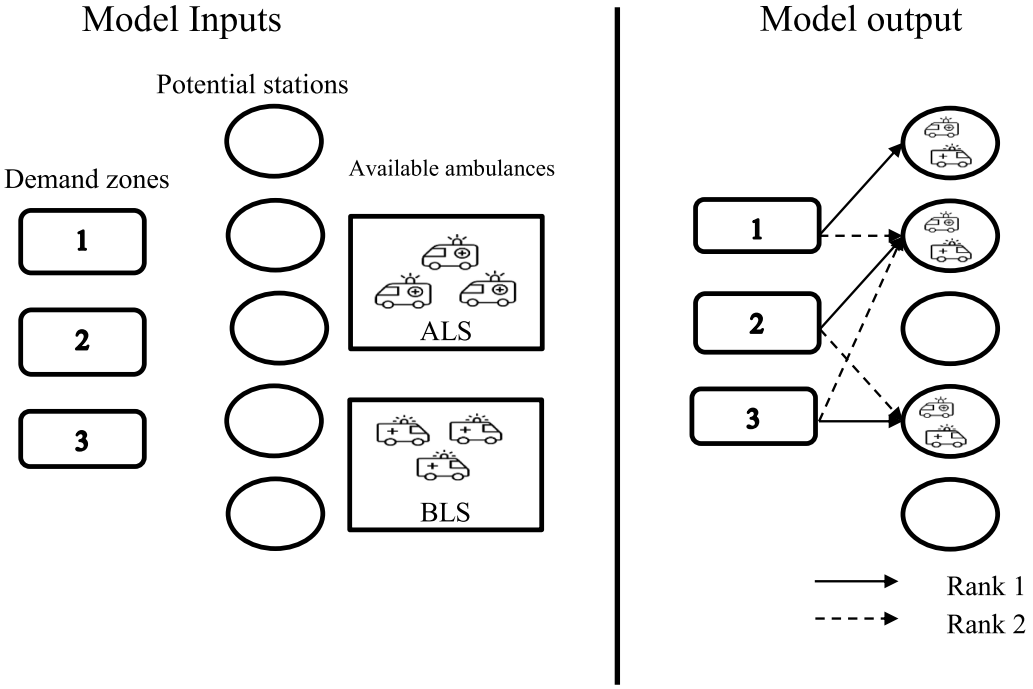

Fig. 1. Description of the ambulance location problem.

for the corresponding zone. Constraint (8) ensures that a station with a higher preference (e.g. rank 1) is assigned more demand than a station with a lower preference (e.g. rank 2). Eq. (9) defines the busy probability of ambulances at a station as a function of arrival rate, service rate, and the number of ambulances located at that station. Eq. (10) defines the dispatch probability of ambulances as a function of the busy probability of ambulances and the dispatch preference order of stations from each demand zone. An approximate approach to estimate the busy probability π_{jt}^{k} and the dispatch probability d_{ijrt}^{kl} is presented in Section 4. Constraint (11) limits the maximum load (i.e. total service time) assigned to an ambulance station based on the number of ambulances allocated to that station. Constraints (12) and (13) are relocation constraints that allow ambulances to be reallocated during each period. The number of relocations at each station is calculated as the difference between the number of ambulances located during period $t+1$ and period t. Since some periods can have more ambulances than others, we use a dummy node to allocate the unallocated ambulances. The nature of different decision variables used in the model is presented in constraints (14)-(16).

4. Approximate approach for estimating busy probability and dispatch probability of ambulances

In the formulation presented in the previous section, we need to estimate the value of dispatch probability d_{ijrt}^{kl} for each station to obtain the objective function value. But from constraint (10), it can be observed that determining the dispatch probability requires the busy probability π_{jt}^{k} of ambulances. The value of busy probability is assumed as an input parameter to the optimisation model in the location models (Knight et al., 2012). However, the value of busy probability depends on the number of ambulances available, the total demand arrival rate, and the required service time for each station. Since the total demand arriving at a station will depend on the preferred dispatch order, the value of δ_{ijrt} is also required to calculate π_{jt}^{k}. Thus, x_{jt}, δ_{ijrt}, and y_{jt}^{k} are required to calculate π_{jt}^{k}, but these are the outputs of the proposed model. This interdependence between π_{jt}^{k} and other variables in the model is depicted in Fig. 2.

Many researchers have used hypercube, approximation, and simulation models to estimate the busy probability of ambulances, and the approach proposed by Knight et al. (2012) is the most similar to ours. Knight et al. (2012) proposed an iterative approach that initially assumes a busy probability for all ambulances and then solves the location problem to get a new solution in the first iteration. The busy probability is then revised using the new solution found in the previous iteration. This process is repeated iteratively until convergence is reached and thus involves solving the ambulance location problem multiple times with different estimates for busy probability. They use M/M/s queuing approximation to obtain the estimate of the busy probability for the given arrival rate, service rate and number of ambulances at each iteration. We also propose an iterative approximate approach to estimate the busy probability assuming each station as M/M/s queueing system. However, we do not solve the ambulance location problem repetitively. In our algorithm, we select a single solution for the location problem at a time and then apply the approximation approach to iteratively obtain a better estimate for busy probabilities for that solution. Thus, we use queuing approximation to obtain an estimate for the busy probability of each solution instance of the algorithm for the same problem. This is necessary since we consider multiple ambulance types and partial backup, which makes it difficult to obtain a single busy probability estimate directly for each solution, like Knight et al. (2012).

In the proposed problem, ALS are dedicated ambulances and preferred only for life-threatening critical type A calls. Thus, the dispatch probability of ALS ambulances is independent of other types of ambulances. However, BLS is considered a general-purpose ambulance and is preferred for type B patients, but it can also be dispatched for type A calls in case of the unavailability of ALS ambulances. Hence, for BLS ambulances, the arrival rate and dispatch probability also depend on the busy probability of ALS ambulances. Similarly, we consider FRV as backup ambulances, which are dispatched if all BLS are busy when type B calls arrive. Therefore, calculating dispatch probability for FRV requires the busy probability of BLS ambulances. Thus, there is a significant interdependence between different ambulance types, and the calculation of exact busy probability will require calculating the conditional probability that an ambulance is serving each patient type along with the accurate estimation of arrival rates, service rate and availability

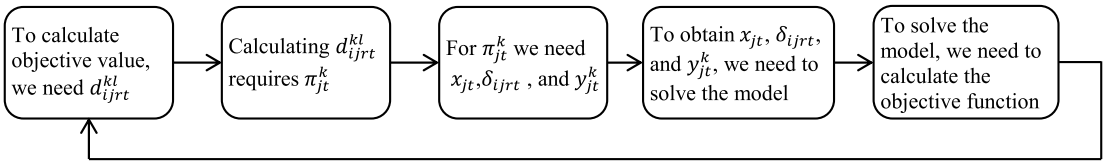

Fig. 2. Interdependence between busy probability π_{jt}^{k} and other decision variables.

of other ambulances at that station. However, calculating conditional probabilities will make the problem computationally intractable as this estimation will only be possible if we apply an exact approach like HQM that considers all possible states or a detailed simulation model. These models are difficult to implement within the ambulance location problem framework, where a large number of alternatives need to be evaluated. Additionally, we only need a good enough approximation of the ambulance availability at each station to evaluate different configurations of the ambulance location decision. Therefore, we focus on the overall long-run busy probability rather than calculating the accurate probability that an ambulance is busy serving a specific type of call. Therefore, we present a simple approximation approach for calculating the long-run busy probabilities of the ambulances and the resulting dispatch probabilities. The long-run dispatch probability approximates the probability that an ambulance from a given station is dispatched for a call from a specific demand zone.

Due to the hierarchy between different ambulance types, we first estimate the dispatch probability of ALS and then use this to calculate the dispatch probability for BLS, which is then used to calculate the dispatch probability of FRV. For this purpose, we first apply the concept of layering (Beojone et al., 2021) and separate each ambulance type at a station as an individual station. For example, if there is a station with three types of ambulances, it can be considered as three distinct stations with one type of ambulance at that station. We first define the set of ambulance types as $K = \{\text{ALS, BLS, FRV}\}$ and two different types of calls as $L = \{\text{A, B}\}$. Then,λ_{it}^{A} and λ_{it}^{B} represent the arrival rate for call types A and B, respectively, from zone i during period t. The arrival rate at station j during period t for ambulance types ALS, BLS, and FRV can be represented by $\Lambda_{jt}^{ALS}, \Lambda_{jt}^{BLS}$, and $\Lambda_{jt}^{\text{FRV}}$, respectively. Similarly, T_{jt}^{ALS}, T_{jt}^{BLS}andT_{jt}^{FRV}, and $M_{jt}^{ALS}, M_{jt}^{BLS}andM_{jt}^{\text{FRV}}$ represent the service time and service rate for the respective ambulance types.

4.1. Overview of the approximate approach

We start by applying layering and separate each ambulance type to be located at different stations to simplify the estimation of busy probability. Then, we apply the approximate approach for the ALS ambulances to obtain the busy probability and dispatch probability of ambulances. Then, the same process is repeated for BLS and FRV. This separation is done because the dispatch probability of BLS ambulances for type A calls depends on the proportion of calls lost by ALS. Similarly, dispatch probability for FRV depends on the proportion of type B calls lost by the BLS ambulances. We also focus only on estimating the long-run busy probabilities rather than the accurate estimation of the dispatch probability of ambulances for each specific demand zone and call type.

The proposed procedure for the approximate approach begins by assuming $\pi_{jt}^{k(iter)} = 0$ for all ambulances of type k. Then, using this assumption, we estimate the dispatch probability $d_{ijrt}^{k,l(iter)}$ of all ambulances for the call type l. This estimated dispatch probability is used to calculate the expected arrival rate, service rate and mean service time of calls for each ambulance station. The estimated arrival and service rates are used to update the busy probability values for all ambulances. Thus, we obtain a newer estimate of the busy probability that considers the number of calls received and the expected service time of calls. This estimate for busy probability is now set as the best estimate for $\pi_{jt}^{k(iter)}$, and the complete process of estimating dispatch probability followed by arrival rate, service rate and service time is repeated. At each step, the newer estimates for arrival rate and service rate are then used to obtain a better estimate for busy probability and dispatch probability. This complete procedure can be summarised as follows.

Step 1: Set iteration number, $iter = 0$ andε = tolerance level required.
Step 2: Set $\pi_{jt}^{k(iter)} = 0, \forall j, t$ and calculate $d_{ijrt}^{k,l(iter)}$.
Step 3: Estimate $\Lambda_{jt}^{k(iter)}$, $\text{M}_{jt}^{k(iter)}$ and $T_{jt}^{k(iter)}$.
Step 4: Calculate.$\rho_{jt}^{k(iter)} = \frac{\Lambda_{jt}^{k(iter)}}{M_{jt}^{k(iter)}}$.
Step 5: Calculate $\pi_{jt}^{k(iter)}$ using Eq. (22).
Step 6: Set $iter = iter + 1$.
Step 7: Using the estimated value of $\pi_{jt}^{k(iter)}$ for each station, update $d_{ijrt}^{kl(iter)}$.
Step 8: Update $\rho_{jt}^{k(iter)}$, $\Lambda_{jt}^{k(iter)}$ and $\text{M}_{jt}^{k(iter)}$ using $d_{ijrt}^{kl(iter)}$ obtained in Step 7.
Step 9: Calculate $\pi_{jt}^{k(iter)}$ using updated values calculated in Step 8.
Step 10: If $\left|\pi_{jt}^{k(iter)} - \pi_{jt}^{k(iter-1)}\right| < \varepsilon, \forall j, t$, stop. Otherwise, go to Step 6.

Although the procedure discussed above is similar for the different ambulance types, the equations required to estimate the dispatch probability, arrival rate and service rate are different for each type of ambulance. The difference in these equations is due to the suitability of different ambulance types for various patient types and the preference of the different ambulance types for each call type based on the dispatch policy considered. This is discussed in the following subsections from Section 4.2 to Section 4.4.

4.2. Estimating busy probability for ALS

We begin by assuming that all ambulances are available at all times, i.e. $\pi_{jt}^{ALS} = 0$, for all stations. This assumption implies that all demand will be served completely by ambulances at the station with the highest preference (rank $r = 1$), as no demand is lost from any station. The proportion of calls served by the station with rank 1 is then equal to 1, and it is 0 for all other stations, as shown in equation (18). Since ALS is considered as a dedicated server and can only serve type A calls, i.e.$d_{ijrt}^{ALS,B} = 0$ for all stations.

$$\left.\begin{array}{l} d_{ij1t}^{ALS,A} = 1, \quad if\ \delta_{ij1t} = 1 \\ d_{ijrt}^{ALS,A} = 0, \quad otherwise \end{array}\right\} \forall i \in I,\ j \in J,\ r \in R, t \in T \tag{18}$$

Using the value for d_{ij1t}^{ALS} from (18), the mean arrival rate and mean service time for ALS ambulances at each station can be obtained from equations (19) and (20), respectively. Then, the mean service rate can be given by equation (21).

$$\Lambda_{jt}^{ALS} = \sum_{i \in I} \sum_{r \in R} d_{ijrt}^{ALS,A} \lambda_{it}^{A} \ \forall j, t \tag{19}$$

$$T_{jt}^{ALS} = \frac{\sum_{i\in I}\sum_{r\in R} d_{ijrt}^{ALS,A}\lambda_{it}^{A}\tau_{ijt}^{A}}{\sum_{i\in I}\sum_{r\in R} d_{ijrt}^{ALS,A}\lambda_{it}^{A}} \ \forall j,t \tag{20}$$

$$M_{jt}^{ALS} = \frac{1}{T_{jt}^{ALS}} \ \forall j,t \tag{21}$$

Now that we have an estimate for the arrival rate and service rate of each station, the probability of calls being lost due to the ambulance being busy can be approximated by considering each station as an $M/M/s$ queueing system, which can be obtained using the Erlang-B formula as given in equation (22)

$$\pi_{jt}^{k} = \frac{\frac{(\rho_{jt}^{k})^{y_{jt}^{k}}}{y_{jt}^{k}!}}{\sum_{a=0}^{y_{jt}^{k}}\frac{(\rho_{jt}^{k})^{a}}{a!}} \ \forall j,k,t \tag{22}$$

where ρ_{jt}^{k} is the server utilisation expressed as the ratio of arrival rate and service rate for a station as in Eq. (23).

$$\rho_{jt}^{k} = \frac{\Lambda_{jt}^{k}}{M_{jt}^{k}} \ \forall j,k,t \tag{23}$$

As we have an estimate of π_{jt}^{ALS}, we can update the dispatch probability $d_{ijrt}^{ALS,A}$ for each pair of zones and stations using Eq. (24).

$$d_{ijrt}^{ALS,A} = \prod_{q\in R|q<r}\left(\sum_{p\in J|p\neq j}\delta_{ipqt}\pi_{pt}^{ALS}\right)\left(1-\pi_{jt}^{ALS}\right) \ \forall i,j,r,t \tag{24}$$

The above estimate for $d_{ijrt}^{ALS,A}$ is then used to obtain a better estimate for the arrival rate and service rate for each station using Eqs. (19)–(21). The above procedure can be applied iteratively to update the value of $d_{ijrt}^{ALS,A}$, every time a new value of π_{jt}^{ALS} is estimated. After a good enough approximation of π_{jt}^{ALS} is obtained, a similar procedure can also be used to obtain the dispatch probability for BLS and FRV type ambulances.

4.3. Estimating busy probability for BLS

To calculate the dispatch probability for BLS ambulances, we need to consider the arrival rate for both type A and type B calls. Similar to the procedure for ALS, we first assume $\pi_{jt}^{BLS} = 0, \forall j,t$ and use Eqs. (25) and (26) to calculate an initial estimate for the dispatch probability of BLS ambulances.

$$\left.\begin{aligned} d_{ij1t}^{BLS,A} &= \prod_{q\in R}\left(\sum_{p\in J}\delta_{ipqt}\pi_{pt}^{ALS}\right), \ \text{if} \ \ \delta_{ij1t}=1 \\ d_{ijrt}^{BLS,A} &= 0\ , \quad \text{otherwise} \end{aligned}\right\} \forall i,\ j,\ r,t \tag{25}$$

$$\left.\begin{aligned} d_{ij1t}^{BLS,\ B} &= 1, \ \ \text{if} \ \ \delta_{ij1t}=1 \\ d_{ijrt}^{BLS,B} &= 0, \ \text{otherwise} \end{aligned}\right\} \forall i,\ j,\ r,t \tag{26}$$

Next, the arrival rate and service rate for BLS ambulances can be calculated using Eqs. (27)–(29).

$$\Lambda_{jt}^{BLS} = \sum_{i\in I}\sum_{r\in R} d_{ijrt}^{BLS,A}\lambda_{it}^{A} + \sum_{i\in I}\sum_{r\in R} d_{ijrt}^{BLS,B}\lambda_{it}^{B} \ \forall j,t \tag{27}$$

$$M_{jt}^{BLS} = \frac{\sum_{i\in I}\sum_{r\in R} d_{ijrt}^{BLS,A}\lambda_{it}^{A}\tau_{ijt}^{A} + \sum_{i\in I}\sum_{r\in R} d_{ijrt}^{BLS,B}\lambda_{it}^{B}\tau_{ijt}^{B}}{\sum_{i\in I}\sum_{r\in R} d_{ijrt}^{BLS,A}\lambda_{it}^{A} + \sum_{i\in I}\sum_{r\in R} d_{ijrt}^{BLS,B}\lambda_{it}^{B}} \ \forall j,t \tag{28}$$

$$T_{jt}^{BLS} = \frac{1}{M_{jt}^{BLS}} \ \forall j,t \tag{29}$$

We use the values of $\Lambda_{jt}^{BLS}, M_{jt}^{BLS}$ and T_{jt}^{BLS} from Eqs. (27)–(29) to calculate the estimate for π_{jt}^{BLS} using Eqs. (22) and (23). The value of π_{jt}^{BLS} is then used to estimate $d_{ijrt}^{BLS,A}$ and $d_{ijrt}^{BLS,B}$ using Eqs. (30) and (31), respectively. The process of estimating $d_{ijrt}^{BLS,A}, d_{ijrt}^{BLS,B}$ and π_{jt}^{BLS} is performed iteratively until a good enough estimate for π_{jt}^{BLS} is obtained.

$$d_{ijrt}^{BLS,B} = \prod_{q\in R|q<r}\left(\sum_{p\in J|p\neq j}\delta_{ipqt}\pi_{pt}^{BLS}\right)\left(1-\pi_{jt}^{BLS}\right) \ \forall i,j,r,t \tag{30}$$

$$d_{ijrt}^{BLS,A} = \prod_{q\in R|q<r}\left(\sum_{p\in J|p\neq j}\delta_{ipqt}\pi_{pt}^{BLS}\right)\left(1-\pi_{jt}^{BLS}\right)\prod_{s\in R}\left(\sum_{u\in J}\delta_{iust}\pi_{ut}^{ALS}\right) \ \forall i,j,r,t \tag{31}$$

4.4. Estimating busy probability for FRV

As FRV type ambulances are backup ambulances for type B calls and are used for cases where BLS ambulances are not available immediately. Therefore, the dispatch probability of ambulances will depend on the busy probability of BLS. Similar to the approach used for ALS and BLS, we first set $\pi_{jt}^{FRV} = 0, \forall j,t$ and calculate dispatch probability using (32).

$$\left.\begin{aligned} d_{ij1t}^{FRV,B} &= \prod_{s\in R}\left(\sum_{u\in J}\delta_{iust}\pi_{ut}^{BLS}\right) \quad \text{if} \ \ \delta_{ij1t}=1 \\ d_{ijrt}^{FRV,B} &= 0\ , \quad \text{otherwise} \end{aligned}\right\} \quad \forall i\in I,\ j\in J,\ r\in R, t\in T \tag{32}$$

The arrival rate and service rate at each station j can then be obtained using Eqs (33)–(35), respectively.

$$\Lambda_{jt}^{FRV} = \sum_{i\in I}\sum_{r\in R} d_{ijrt}^{FRV,B}\lambda_{it}^{B} \ \forall j,t \tag{33}$$

$$M_{jt}^{FRV} = \frac{\sum_{i\in I}\sum_{r\in R} d_{ijrt}^{FRV,B}\lambda_{it}^{B}\tau_{ijt}^{B}}{\sum_{i\in I}\sum_{r\in R} d_{ijrt}^{FRV,B}\lambda_{it}^{B}} \ \forall j,t \tag{34}$$

$$T_{jt}^{FRV} = \frac{1}{M_{jt}^{FRV}} \ \forall j,t \tag{35}$$

$$d_{ijrt}^{FRV,B} = \prod_{q\in R|q<r}\left(\sum_{p\in J|p\neq j}\delta_{ipqt}\pi_{pt}^{FRV}\right)\left(1-\pi_{jt}^{FRV}\right)\prod_{s\in R}\left(\sum_{u\in J}\delta_{iust}\pi_{ut}^{BLS}\right) \ \forall i,j,r,t \tag{36}$$

5. Solution approach

The mathematical formulation presented in Section 3, combined with the busy probability calculations in Section 4, is an MINLP model. The complexity of the problem due to multiple ambulance types and the non-linear nature of equations related to busy probability and dispatch probability in the formulation makes the proposed model difficult to solve using commercial solvers. Therefore, we propose a three-stage

hybrid solution approach that integrates particle swarm optimisation (PSO) and AVNS. We have developed a VNS-based metaheuristic as it requires very few parameters, while it also provides a flexible framework to develop a solution algorithm for any combinatorial optimisation problem (Hansen et al., 2017). The VNS has also been applied to ambulance location problems, and its effectiveness has been shown in the problem with time-dependent variations in demand and travel time (Nadar et al., 2021). The initial solution for the AVNS is obtained by solving a relaxed mixed-integer linear programming (MILP) ambulance station location problem and a PSO-based approach for ambulance allocation. The AVNS is then applied to the initial solution to obtain an improved solution. The application of the relaxed ambulance location problem for the initial solution ensures that the locations of ambulance stations are optimal and comparable to solutions obtained using a simple coverage-based model. The complete solution approach is summarised in Fig. 3.

5.1. Phase 1 – initial station location (relaxed MILP)

The proposed problem is a joint ambulance location and dispatch problem that determines the optimal location of ambulance stations, the allocation of ambulances to each station and the preference dispatch order of ambulances from stations for each demand zone. For the purpose of solving the problem, we first simplify and solve a station location problem to obtain the best possible location of stations. This relaxed problem determines ambulance stations (x_{jt}) and also assigns order of preference (δ_{ijrt}) for stations from each zone while neglecting the dispatch probability (d^{lk}_{ijrt}) and the number of ambulances (y^k_{jt}) located at each station. The objective function (37) of the relaxed problem is obtained by replacing the dispatch probability d^{lk}_{ijrt} with δ_{ijrt} in the objective function (1). The weightage W_r in (37) assigns higher weight for higher rank.

$$\text{Maximise} \sum_{i\in I}\sum_{j\in J}\sum_{t\in T}\sum_{l\in L}\sum_{r\in R} P^{kl}_{ijt} W_r D^l_{it} \delta_{ijrt} \tag{37}$$

Subject to (2), (5) and (6).

The relaxed problem does not guarantee a feasible solution to the original problem, as constraints related to dispatch could be violated. However, the relaxed problem is a MILP that can be effectively solved using commercial solvers like CPLEX.

5.2. Phase 2 – initial ambulance allocation (Particle swarm optimisation)

Once the initial station locations are obtained from the previous stage, we determine the optimal allocation of ambulances to stations by applying a PSO-based approach. The PSO metaheuristic was introduced by Kennedy and Eberhart (1995). PSO represents the search space of the optimisation problem as positions of particles that move towards the optimal solution iteratively. PSO starts with an initial population of particles randomly generated within the search space, each associated with a position vector $x_i = [x_{i1}, x_{i2}, ..., x_{i\overline{K}}]$ and velocity vector $v_i = [v_{i1}, v_{i2}, ..., v_{i\overline{K}}]$. The number of ambulances is represented using a decimal number between 0 and 1, which makes it easier to apply PSO. Let β^k_{jt} represent the number of ambulances of type k located at station j in period t in the solution representation. Then, the actual number of ambulances y^k_{jt} is obtained using equation (38)

$$y^k_{jt} = \left\lfloor (A^k_{max} + 1)\beta^k_{jt} \right\rfloor \tag{38}$$

where A^k_{max} is the maximum number of ambulances of type k assigned to a station. The position of each particle is updated at each iteration using the global best position and the individual best solution of the particle. If the position and velocity vectors after τ iterations are given by x^τ_i and v^τ_i, respectively, the updated values for $x^{\tau+1}_i$ and $v^{\tau+1}_i$ at iteration $\tau+1$ is obtained using Eqs. (39)–(41).

$$v^{\tau+1}_{ik} = \omega v^\tau_{ik} + c_1 r_1 \left(p^{best}_{ik} - v^\tau_{ik}\right) + c_2 r_2 \left(g^{best}_k - v^\tau_{ik}\right) \tag{39}$$

$$x^{\tau+1}_{ik} = x^\tau_{ik} + v^{\tau+1}_{ik} \tag{40}$$

$$\omega = \omega_{max} - (\omega_{max} - \omega_{min}) \frac{\tau}{\tau_{max}} \tag{41}$$

In the above equations, ω represents the inertial weight, ω_{max} and ω_{min} are the maximum and minimum possible values of ω, respectively, and τ_{max} is the maximum number of iterations. Other parameters in Eq. (39) include c_1 and c_2 are the learning factors, whereas r_1 and r_2 are uniform random numbers,$r_1, r_2 \in (0,1)$. Similarly, p^{best}_{ik} represents the kth element of the local best solution (p^{best}_i) of the ith individual, and g^{best}_k

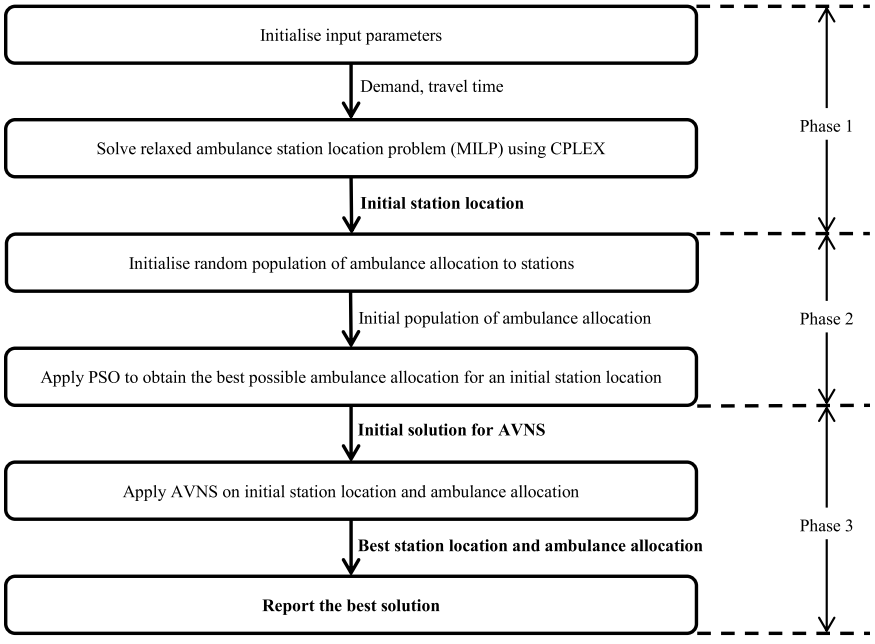

Fig. 3. Overview of the proposed three-stage hybrid solution approach.

represents the kth element of the global best solution (g^{best}). The complete procedure of the PSO algorithm is summarised in Algorithm 1. The initial solutions for the PSO are generated by assigning the available ambulances to the selected ambulance stations from phase I. To ensure the solutions do not become infeasible, we apply penalty for exceeding the total number of ambulances available for each type and also for exceeding the maximum number of ambulances allocated to each station. The algorithm ends with an initial solution for ambulance allocation for the station locations obtained in phase 1 of the solution.

Algorithm 1: PSO for initial ambulance allocation

Set iteration $\tau = 0$ //Initialisation
for $i = 1$: *population_size*
  Initialise position (x_i^τ) and velocity (v_i^τ) of solution i
  Evaluate fitness $f(x_i^\tau)$
  Set local best as current position ($p_i^{best} = x_i^\tau$)
end for
Initialise global best $g^{best} = \max\left(f(x_1^\tau), f(x_2^\tau), .., f\left(x_{popsize}^\tau\right)\right)$
While (stopping criteria is not met) do
  for $i = 1$:*population_size*
    Update $v_{ik}^{\tau+1}$
    Update $x_{ik}^{\tau+1}$
    If $f(x_i^{\tau+1}) \rangle f(p_i^{best})$ **then**
      $p_i^{best} = x_i^{\tau+1}$
    end if
    If $f(x_i^{\tau+1}) \rangle f(g^{best})$ **then**
      $g = x_i^{\tau+1}$
    end if
  end for
end while
Return the current solution as the best solution

5.3. Phase 3 – Adaptive variable neighbourhood search

Combining the solution for station location obtained in phase 1 and ambulance allocation in phase 2, we obtain an initial solution that is improved by applying AVNS. Variable neighbourhood search (VNS) is a simple and effective metaheuristic proposed by Mladenović and Hansen (1997), which combines local search and systematic neighbourhood changes for solving a combinatorial optimisation problem. VNS employs systematic changes of neighbourhoods to escape local optima and thus avoid getting trapped in valleys. We propose and implement several neighbourhood structures to improve the overall performance of the algorithm. Also, we apply an adaptive procedure for the neighbourhood change step to enable an efficient search of the solution space. A detailed description of the proposed algorithm and the neighbourhood structures is presented in the subsequent sections.

5.3.1. Solution representation

An important element for the effective implementation of a metaheuristic is solution representation. The proposed problem has multiple levels of decisions to be represented, including optimal ambulance station locations, the number of ambulances of different types allocated to each station, and the preference order of stations for each demand zone. Therefore, we apply a two-level encoding scheme to represent each solution. The first level, which we refer to as the station array, represents the preference order of stations assigned to each demand zone. The second level, referred to as the ambulance array, represents the number of ambulances of each type allocated to the ambulance stations. The proposed solution representation is shown in Fig. 4 for a hypothetical solution with four demand zones, five potential stations, four periods, and three ranks. In Fig. 4(a), each column of the station array represents a combination of a demand zone and the period, while each row represents the rank r for the corresponding zone. So from the figure, the preferred order for zone 1 in period 1 is station 5, station 1, and station 3. Similarly, in the ambulance array of Fig. 4(b), each column represents the potential station location and the period, while each row represents the type of ambulances. So, from Fig. 4(b), we can observe that in period 1, two ALS, two BLS and one FRV type ambulances are allocated for the first ambulance station.

5.3.2. Neighbourhood structures

We propose various neighbourhood structures to explore the solution space. Along with the station array and ambulance array for solution representation, we also maintain a list of assigned and unassigned stations for each period in each solution to apply some of the neighbourhood structures. The proposed neighbourhood structures are as follows.

(i) NS_1(Swap station): To obtain a solution in the neighbourhood NS_1, we randomly select two demand zones ($j_1, j_2 \in J$). A swap operation is performed by replacing the station assigned to j_1 with that assigned to j_2 and vice-versa. This modifies the stations assigned to both zones.

(ii) NS_2(Swap station rank): In this neighbourhood NS_2, we randomly select a demand zone ($j \in J$) and two ranks ($r_1, r_2 \in R$). The stations allocated to both rank r_1 and rank r_2 are then swapped, which modifies the preference order of the stations assigned to the demand zone.

(iii) NS_3(Replace station): To obtain a solution in this neighbourhood NS_3, we randomly select two demand zones ($j_1, j_2 \in J$). Then, a replace operation is performed by removing the station assigned to j_1 and replacing it with the station assigned to j_2.

(iv) NS_4(Replace stations between periods): Solutions in neighbourhood NS_4 involve assigning the same stations in different periods. We randomly select a demand zone j and two periods (t_1 and t_2). The stations allocated to zone j during period t_1 are replaced with those allocated during period t_2.

(v) NS_5(Add new station): Neighbourhood NS_5 involves adding a new station to the solution. In this procedure, we select a random station from the list of unassigned stations in period t. Then the selected station is assigned to a random zone j in period t with a random rank r.

(vi) NS_6(Remove and replace station): To obtain a solution in neighbourhood NS_6, we randomly select an unassigned station and an assigned station during period t. All occurrences of the currently assigned station are then replaced with the selected unassigned station.

(vii) NS_7(Remove station): Neighbourhood NS_7 produces new solutions by removing an assigned station from the solution. We select a random station j from the list of assigned stations in period t, and each occurrence of station j is replaced with randomly selected stations from the assigned list.

(viii) NS_8(Exchange ambulances): The neighbourhood solution NS_8 is obtained by reassigning ambulances allocated to two stations. This is done by selecting two random stations from the list of assigned stations for period t and swapping the ambulances allocated to both stations.

(ix) NS_9(Add or remove ambulance): For a solution in neighbourhood NS_9, we select an assigned station i, an ambulance type k, and a random number between 0 and 1. If the random number is less than 0.5, we increase the number of ambulances of type k at station i by one; otherwise, we decrease the number of ambulances by one.

5.3.3. Adaptive neighbourhood change procedure

The VNS process involves a systematic neighbourhood change procedure to exploit the proposed neighbourhood structures. The traditional VNS restarts the search procedure from the first neighbourhood when a new best solution is found in any neighbourhood. This leads to a higher probability of exploration in the first neighbourhood. Therefore, we apply two adaptive neighbourhood change procedures that take into account the performance of neighbourhood structures in previous iterations. The first approach is similar to Wang et al. (2021), based on the ranked selection scheme presented by Reeves (1995). The second

approach directly uses the number of times the neighbourhood produces an improvement. A brief description of both these adaptive procedures is presented as follows.

(i) Adaptive neighbourhood change procedure – I

The neighbourhood change procedure uses an improvement counter C_i for each neighbourhood i that is initialised to $C_i = 1$. If in any given iteration, neighbourhood i is chosen and an improvement is obtained in the objective function value, then the value of C_i is increased by one. After every iteration, the desirability value (φ_i) of each neighbourhood is calculated as $\varphi_i = \frac{C_i}{N_i}$, where N_i is the number of iterations of the neighbourhood i applied. The neighbourhoods are then arranged in ascending order of the value of φ_i and are ranked. For example, if there are two neighbourhoods with $\varphi_1 > \varphi_2$, then $r_1 = 2$ and $r_2 = 1$. The probability of selection (P_i) for each neighbourhood is given by

$$P_i = \frac{2r_i}{N(N+1)}, \tag{42}$$

where r_i is the rank of neighbourhood i, and N is the total number of neighbourhoods. This procedure gives an increased value of the probability of selection based on the rank of each neighbourhood. The complete AVNS algorithm with this neighbourhood change approach (referred to as AVNS-I) is summarised in Algorithm 2.

Algorithm 2: AVNS – I

1: Set maximum number of iterations (τ_{max}), maximum time (T_{max})
2: Initialise set of neighbourhood structures, i.e.$NS = \{NS_1, NS_2, .., NS_9\}$ and iteration $\tau = 0$
3: For each NS_k in NS, initialise counter for improvement ($C_i = C_{initial}$)
4: **while** termination criteria not met **do**
5: for each neighbourhood k calculate desirability value (φ_i)
6: rank neighbourhoods based on ascending value of φ_i
7: calculate probability range for each neighbourhood k using equation (42)
8: select a random number r
9: select the neighbourhood k using the probability range between which r lies
10: $S \leftarrow S_{current}$
11: Generate a random solution (S') using S in neighbourhood k: $S' \in NS_k(S)$ *//Shaking*
12: **while** $i <$ *max_iter* **do**
13: Generate new solution S'': $S'' \leftarrow$ *LocalSearch* (S', NS_k)
14: Evaluate the solution to get $f(S'')$
15: **if** $f(S'') \rangle f(S)$
16: Update S: $S \leftarrow S''$
17: **end if**
18: $i \leftarrow i + 1$
19: **end while**
20: If $f(S) \rangle f(S_{current})$
21: Update improvement counter:$C_i \leftarrow C_i + 1$
22: $S_{current} \leftarrow S$
23: **end while**
24: **Return** solution S

(ii) Adaptive neighbourhood change procedure – II

In this approach, we initialise all neighbourhoods with an initial value of improvement counter $C_i$as $C_{initial}$. At the end of each iteration, the counter value is updated as in equation (43)

$$\left.\begin{array}{ll} C_i = C_i + 1 & if \quad S_{new} > S_{old} \\ C_i = \max(1, \quad C_i - 1) & if \quad S_{new} \leq S_{old} \end{array}\right\} \tag{43}$$

The probability of each neighbourhood is then calculated using equation (44)

$$P_i = \frac{C_i}{\sum C_i} \tag{44}$$

Therefore, a neighbourhood being selected depends directly on the number of iterations that have led to an improvement. The complete AVNS-II procedure using this neighbourhood change procedure is presented in Algorithm 3.

The difference between both approaches is that in AVNS-I, the probability of selecting the neighbourhoods always remains constant as it only depends on the rank of each neighbourhood. In AVNS-II, the probability of selection may increase or decrease based on whether there is an improvement in that neighbourhood or not.

Algorithm 3: AVNS – II

1: Set maximum number of iterations (τ_{max}), maximum time (T_{max})
2: Initialise set of neighbourhood structures, i.e.$NS = \{NS_1, NS_2, .., NS_9\}$
3: For each NS_k in NS, initialise improvement counter:$C_i = C_{initial}$
4: **while** termination criteria not met **do**
5: Calculate probability range for each neighbourhood k using equation (44)
6: Select a random number r
7: Select the neighbourhood k using the probability range between which r lies
8: $S \leftarrow S_{initial}$
9: Generate a random solution (S') using S in neighbourhood k: $S' \in NS_k(S)$ *//Shaking*
10: **while** $i <$ *max_iter* **do**
11: Generate new solution S'': $S'' \leftarrow$ *LocalSearch* (S', NS_k)
12: Evaluate the solution to get $f(S')$
13: **if** $f(S'') \rangle f(S)$
14: Update S: $S \leftarrow S''$
15: **end if**
16: $i \leftarrow i + 1$
17: **end while**
18: **If** $f(S) \rangle f(S_{initial})$
19: $C_i \leftarrow C_i + 1$
10: $S_{initial} \leftarrow S$
21: **else**
22: $C_i = max(1, C_i - 1)$
23: **end while**
24: **Return** solution S

6. Computational results

In order to validate the effectiveness of the proposed solution approach for the ambulance location problem, we conducted computational experiments to compare AVNS with the basic variable neighbourhood search (BVNS) approach. We also compare the proposed approximation approach with a simulation model for a few small test instances. The initial solution approach is compared with two approaches based on genetic algorithm (GA) and greedy approach to show that the solution of the relaxed MILP problem gives a better initial solution. The heuristics and algorithms presented in the previous section are developed using Python 3.8. The computational experiments were implemented on a personal computer with Intel Core i5-4570T CPU @ 2.90 GHz CPU and 8 GB RAM. The stopping criteria for all the computational results based on the AVNS-I, AVNS-II and BVNS algorithm was set as 25 continuous iterations without any improvement in the objective value.

6.1. Description of input data for test instances

To validate the proposed model and test the performance of the solution approach, we develop a set of test instances based on Kolkata, an urban centre in India. Kolkata had a huge population of 4.5 million according to the 2011 census, in an area of 185 km^2 with a population density of 24,252 per square kilometre. Kolkata does not have a centralised ambulance system for emergency services like most parts of India. Vehicles and ambulances from private service providers serve the most demand in India. Recently, some Indian states have started deploying a state-owned centralised ambulance system. Even Kolkata plans to introduce 150 ambulances to manage road accidents. Ambulance location planning in a densely populated city like Kolkata poses significant challenges due to the narrow and crowded roads. Lack of dedicated routes for ambulances and high traffic levels can result in a significantly high difference in travel time between peak and non-peak hours. Different wards of the city also have a high difference in

Zone	1				2				3				4			
Period	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
Rank 1	5	5	5	5	3	3	2	3	1	3	3	1	1	5	5	5
Rank 2	1	4	2	3	5	5	5	5	3	4	1	5	5	1	1	1
Rank 3	3	1	1	1	1	1	3	1	5	5	5	3	3	3	3	3

(a) Station array

Station	1				2				3				4				5			
Period	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
ALS	2	1	0	1	0	0	2	0	1	0	2	0	0	1	0	0	1	1	1	1
BLS	2	1	3	3	0	0	2	0	1	2	3	1	0	2	0	0	1	2	1	1
FRV	1	2	1	1	0	0	0	0	1	1	2	0	0	0	0	0	0	0	1	2

(b) Ambulance array

Fig. 4. Solution representation using two arrays for VNS: station array and ambulance array.

population density resulting in significant differences in demand across the region.

6.2. Potential locations, travel time and demand

For generating the test instances, we divide the city into 141 small regions based on the geographical division of wards, where each ward represents a demand zone. We further selected 217 locations in various public places as potential sites for ambulance stations, including railway stations, hospitals, malls, schools/colleges and fire stations. The choice of these sites is based on the requirement of possible space for parking, electrical supply for recharging of equipment, and basic security.

Since there is no available information for ambulance travel time, the travel time data is obtained from Uber-movement (movement.uber.com), which provides data on travel time, speeds and mobility in different cities across the world. The lack of dedicated lanes for ambulances and narrow congested roads with heavy traffic suggest that ambulances will also face similar challenges, and therefore use of this dataset is justified. The Uber-movement dataset provides the mean and standard deviation of hourly travel time between different wards. We obtained travel time from each potential site to each ward and nearest hospital using Google Maps, then adjusted for hourly variation using the data from the Uber-movement dataset. Based on the Uber movement dataset, it is seen that the mean travel time during the peak period is almost 80% higher than the travel time during the lowest period, i.e. there is an 80% difference between the highest and lowest points in mean travel time within a day. Also, the travel time during the peak period was about 30% higher compared to the overall mean travel time during the day.

We assume an average constant service time for all patients of the same type at the location of the patient. For the policy after service completion, we assume a fixed proportion of all calls require travel to hospitals, and for the remaining calls, the ambulances return to the station, similar to Leknes et al. (2017). For each demand zone, we consider the nearest hospitals available for calculating the travel time for reaching that station. We assume a similar proportion to Leknes et al. (2017) for the calculation of service time, i.e. 43% of calls require transportation, and the remaining do not. Additionally, we also assume that the response time is the same for all ambulance types and call types for any pair of ambulance station and demand zone for the computational purpose. The total demand was simulated randomly as no existing data is available for demand in Kolkata. We assume that the spatial distribution of demand for each ward is proportional to the population of the ward. The spatial distribution of demand is an important factor as the difference between the populations of the two wards is significantly high. The population of ward 45 was the lowest, with only 8,394 in 2011, while ward 66 had a population of 98,024. The population density also varies across different wards of the city, with northern wards having higher density compared to the southern part of Kolkata. Wards 23, 39, 44 and 135 have population densities over 100,000 per square km, while wards 106 to 116 in the southern part are some of the wards that have population densities below 25,000 per kilometre.

The time-dependent variation in demand for emergency calls is shown to follow a specific pattern with a peak during noon and then a slight drop to reach a second peak after 7.00 pm again. The temporal variation in demand was obtained by adjusting the total demand for each hour based on the pattern obtained by Cantwell et al. (2013). We consider a day to be divided into six periods of four hours each to account for the temporal variation in demand. Thus, the total demand for each ward is obtained by multiplying the total demand by a factor proportional to the population of a ward and a factor that adjusts for the hour of the day variation. The maximum call arrival rate associated with a demand zone during the peak demand period for type A calls is 3.86 calls per hour, while for type B calls considered is 6.44 calls per hour. The corresponding minimum call arrival rate for a demand zone during the peak period is 0.52 and 0.87 calls per hour for type A and type B calls, respectively. Similarly, during night time, the maximum arrival rate considered is 1.66 calls per hour for type A calls and 2.85 calls per hour for type B calls, while the minimum arrival rate is 0.24 calls per hour and 0.37 calls per hour for type A and type B calls, respectively. The average call arrival rate over the day considered for all demand zones considered is 1.67 calls per hour for type A calls and 2.78 calls per hour for type B calls.

6.3. Evaluation of proposed approximate approach for busy probability

To validate the approximation approach presented in Section 4, we constructed simulation model for various toy models using SimEvents in the Simulink package of MATLAB. The server utilisation obtained using the approximation approach was compared with the simulation model. We consider three small systems with the following characteristics.

(i) 2 demand zones, 1 station, and 4 ambulances with 1 ALS, 2 BLS and 1 FRV
(ii) 4 demand zones, 2 stations, and 6 ambulances with 1 ALS, 1 BLS and 1 FRV at each station
(iii) 4 demand zones, 2 stations, and 8 ambulances with 1 ALS, 2 BLS and 1 FRV at each station

The demand arrival rate was varied from 0.2 to 0.5 calls per hour for type A calls and from 0.4 to 1.0 calls per hour for type B calls to generate five different instances. A comparison between simulation and approximation approaches for estimating server busy probability for all three systems considered above is presented in Tables 1–3. Although there is either an overestimation or underestimation of busy probability for all ambulance types in all three systems, the difference is mostly within 2–7% for all instances. Since we only need an estimate for the busy probability of ambulance stations to calculate the objective function to compare different configurations of ambulance allocation, the approximation approach can be effective. Based on the results, we can conclude that an approximate approach can be used for the estimation of busy and

Table 1
Comparison between simulation and approximation approaches for busy probability (4 ambulance system).

Server		ALS	BLS	FRV
Simulation	Instance 1	0.383	0.430	0.219
	Instance 2	0.318	0.347	0.157
	Instance 3	0.457	0.449	0.220
	Instance 4	0.457	0.569	0.355
	Instance 5	0.386	0.635	0.380
Approximate approach	Instance 1	0.420	0.473	0.253
	Instance 2	0.364	0.390	0.209
	Instance 3	0.487	0.528	0.251
	Instance 4	0.507	0.644	0.422
	Instance 5	0.430	0.652	0.455
Average % difference		4.20	5.17	5.18

dispatch probabilities of ambulances, which provides estimates comparable with the simulation approach.

6.4. Performance comparison of neighbourhood change approaches

We evaluate the performance of the adaptive variable neighbourhood change procedures proposed in Section 5.3.3 by comparing the proposed approaches with the BVNS. We created 20 test instances and solved them using all three approaches (AVNS-I, AVNS-II and BVNS). To account for the uncertainty in algorithms, each instance was solved for ten trials using the same algorithm, and an average of the objective value and the best objective value were reported. The mean CPU time required by each algorithm is also reported for comparison. Further, we consider the average relative percentage difference (ARPD) given by equation (45) in n trials for comparison.

$$S_{ARPD} = \frac{\sum_{i=1}^{n}(S_{best} - S_i)}{n \times S_{best}} \times 100 \tag{45}$$

Table 4 summarises the comparison of the three approaches in terms of the objective function value obtained. The table shows the average objective value obtained in ten trials, the maximum objective value achieved, and the percentage difference between the best solution obtained by the considered method and the overall best solution found. In the table, BVNS indicates the solution obtained using the basic VNS approach, which uses a fixed neighbourhood sequence of change routine. Based on the average objective value, it can be seen that all three approaches produce a very similar range of values. The best solutions were found using adaptive neighbourhood procedures in most cases. AVNS-I and AVNS-II achieved the best solution in nine and eight instances of the total 20 instances, respectively, while the BVNS approach found the best solution in only three instances. Nevertheless, the difference between the best solutions obtained by all three approaches is negligible, with less than a 1% difference for all the instances. Thus, it can be concluded that all three approaches produce solutions of similar quality, although AVNS-I and AVNS-II seem to produce a slightly better solution.

Table 5 shows the comparison between the three approaches in the mean CPU time and the ARPD values. There is a clear difference between the times taken for all three approaches, with the BVNS approach taking significantly longer for almost all the instances. The difference between

Table 2
Comparison between simulation and approximation approaches for busy probability (6 ambulance system).

Station		1			2		
Server type		ALS	BLS	FRV	ALS	BLS	FRV
Simulation	Instance 1	0.256	0.417	0.035	0.258	0.420	0.033
	Instance 2	0.205	0.334	0.014	0.205	0.334	0.013
	Instance 3	0.322	0.401	0.032	0.323	0.403	0.030
	Instance 4	0.323	0.529	0.070	0.324	0.532	0.066
	Instance 5	0.253	0.436	0.026	0.252	0.439	0.021
Approximate approach	Instance 1	0.289	0.349	0.073	0.289	0.349	0.073
	Instance 2	0.250	0.311	0.046	0.250	0.311	0.046
	Instance 3	0.337	0.356	0.064	0.336	0.356	0.064
	Instance 4	0.337	0.408	0.108	0.399	0.399	0.108
	Instance 5	0.292	0.366	0.082	0.291	0.366	0.082
Average % difference		2.93	6.56	3.90	4.04	6.96	4.19

Table 3
Comparison between simulation and approximation approaches for busy probability (8 ambulance system).

Station		1			2		
Server type	ALS (1)	BLS (2)	FRV (1)	ALS (1)	BLS (2)	FRV (1)	
Simulation	Instance 1	0.250	0.755	0.046	0.250	0.756	0.041
	Instance 2	0.203	0.644	0.026	0.203	0.644	0.023
	Instance 3	0.320	0.743	0.038	0.321	0.742	0.033
	Instance 4	0.320	0.867	0.074	0.321	0.876	0.062
	Instance 5	0.258	0.778	0.034	0.258	0.781	0.028
Approximate approach	Instance 1	0.289	0.691	0.073	0.289	0.691	0.073
	Instance 2	0.250	0.689	0.046	0.250	0.689	0.046
	Instance 3	0.337	0.694	0.064	0.336	0.674	0.064
	Instance 4	0.337	0.892	0.108	0.399	0.901	0.108
	Instance 5	0.292	0.735	0.082	0.291	0.735	0.082
Average % difference		3.08	4.51	3.09	4.23	4.98	3.72

Table 4
Comparison of neighbourhood change approaches for VNS based on the objective function value (bold numbers indicate best average objective value/best objective value).

Instance	Number of stations (N_s)	Number of ambulances (N_a)	Average objective value in 10 trials			Best objective value in 10 trials			% difference from overall best		
			BVNS	AVNS-I	AVNS-II	BVNS	AVNS-I	AVNS-II	BVNS	AVNS-I	AVNS-II
1	5	8	**12998.6**	12947.8	12975.7	13043.8	12999.8	**13051.3**	0.06	0.39	0.00
2	5	8	15900.1	**15909.5**	15901.0	**15970.7**	15961.7	15966.3	0.00	0.06	0.03
3	10	15	**20138.1**	20098.2	20086.0	20206.6	**20206.6**	20175.5	0.00	0.00	0.15
4	10	15	23620.3	**23634.5**	23601.9	23730.9	**23737.2**	23692.1	0.03	0.00	0.19
5	15	25	30729.0	**30735.4**	30681.6	30832.0	30844.3	**30896.3**	0.21	0.17	0.00
6	15	25	38532.8	**38565.5**	38543.2	38696.8	**38705.3**	38669.8	0.02	0.00	0.09
7	25	40	55430.0	**55471.8**	55389.8	55751.4	**55881.8**	55793.6	0.23	0.00	0.16
8	25	40	64201.4	**64245.6**	64198.2	64489.6	64496.6	**64568.7**	0.12	0.11	0.00
9	50	78	89815.6	90012.4	**90159.2**	90345.3	90481.5	**90610.1**	0.29	0.14	0.00
10	50	78	**104070.2**	103945.7	104012.2	104396.5	104508.9	**104512.9**	0.11	0.00	0.00
11	62	92	128898.8	129055.0	**129263.1**	129867.8	**130051.7**	129917.9	0.14	0.00	0.10
12	62	92	**148056.5**	148045.7	147832.0	148991.4	149056.6	**149098.1**	0.07	0.03	0.00
13	80	120	158772.3	158872.7	**158969.4**	160269.2	160148.5	**160683.9**	0.26	0.33	0.00
14	80	120	195416.6	195305.8	**195747.0**	**197074.7**	196522.5	196845.8	0.00	0.28	0.12
15	100	152	182514.7	182368.1	**183145.4**	184156.9	183762.6	**184575.2**	0.23	0.44	0.00
16	100	152	224811.3	**225355.1**	225151.1	226169.6	**226871.5**	226192.3	0.31	0.00	0.30
17	120	184	195724.9	195731.2	**196194.6**	197336.3	197478.2	**197775.4**	0.22	0.15	0.00
18	120	184	246593.3	**247206.5**	246932.2	248030.4	248576.5	**249029.8**	0.40	0.18	0.00
19	141	217	212119.7	**212571.7**	212271.7	214576.1	**214989.3**	214826.7	0.19	0.00	0.08
20	141	217	**283961.8**	283482.8	283611.1	285766.0	**286591.3**	286416.0	0.29	0.00	0.06

Table 5
Comparison of neighbourhood change approaches based on CPU time and ARPD (bold numbers indicate lowest CPU time/smallest ARPD).

Instance	CPU Time (Seconds)			ARPD		
	BVNS	AVNS-I	AVNS-II	BVNS	AVNS-I	AVNS-II
1	64.4	**44.7**	52.8	**0.40**	0.79	0.58
2	53.9	**39.9**	42.6	0.44	**0.38**	0.44
3	157.3	**141.3**	145.0	**0.34**	0.54	0.60
4	169.5	**110.9**	136.8	0.49	**0.43**	0.57
5	280.6	**226.2**	261.6	0.54	**0.52**	0.69
6	315.5	**179.6**	189.4	0.45	**0.36**	0.42
7	941.9	**756.8**	876.9	0.81	**0.73**	0.88
8	851.8	**599.0**	701.5	0.57	**0.50**	0.57
9	1994.0	**1818.6**	1984.3	0.88	0.66	**0.50**
10	2190.5	**1389.2**	1765.4	**0.42**	0.54	0.48
11	2544.4	2467.8	**2327.7**	0.89	0.77	**0.61**
12	3187.3	**1999.0**	2432.0	**0.70**	0.71	0.85
13	3929.0	**3164.3**	3604.9	1.19	1.13	**1.07**
14	4415.5	**3501.0**	3712.6	0.84	0.90	**0.67**
15	4647.3	**3843.3**	4145.4	1.12	1.20	**0.77**
16	5052.6	**4193.0**	4369.4	0.91	**0.67**	0.76
17	5831.5	4655.2	**4420.7**	1.04	1.03	**0.80**
18	5987.8	5620.3	**5204.2**	0.98	**0.73**	0.84
19	6904.0	5615.5	**5502.6**	1.33	**1.12**	1.26
20	7053.1	6521.8	**6229.7**	**0.92**	1.08	1.04

the other two approaches was relatively small for AVNS-I taking the least time in most instances. The ARPD for all three approaches were within 1% for most instances and within 2% for all instances. There was a slight rise in the ARPD with an increase in the problem size for all three approaches, indicating a possibility of converging to a sub-optimal solution with an increase in problem size. Despite that, in more than 8 out of 10 trials, the solution obtained was within 0.5% of the best-known solution using all three approaches. Our results indicate that an adaptive approach for neighbourhood change performs better than a fixed approach while applying multiple neighbourhoods.

6.5. Comparison among approaches for initial solution

To evaluate the impact of the quality of initial solutions on the overall solution quality of the AVNS algorithm, we compare the initial solutions obtained using three different approaches. We compare initial solutions obtained using an exact solver (CPLEX), a greedy approach and a GA-based approach. CPLEX being an exact solver will provide the optimal solution for the relaxed MILP. The greedy approach usually provides a poor-quality initial solution, while GA is expected to perform better than the greedy approach. Our goal is to observe how the quality of the initial solutions and the time taken to obtain these solutions affect the overall solution quality and time taken for the AVNS. The GA-based approach starts with randomly generated initial solutions that have chromosome representation that is based on the station array shown in Fig. 4(a). The fitness function of the solutions is calculated using the objective function value and the penalties for violation of constraints. We then rank all the solutions and apply a selection based on the probability of selection given by Eq. (42). Then, we apply a single-point crossover and random mutation on the parent chromosomes to obtain the child chromosomes. We then add the newer solutions to the population pool by replacing the lowest-ranked solutions.

The greedy approach directly generates an initial solution to the problem by first ordering all the stations in the ascending order of the P_{ijt} value and assigning the stations until the maximum number of stations is reached. The stopping criteria for the GA to obtain the initial solution were set as a) if there is less than 0.1% improvement in fitness function value in 10 iterations or b) a maximum time limit of 300 s. For the CPLEX solver, we set a maximum time limit of 300 s or a solution gap of 0.1%. The ambulances are then randomly assigned to each allocated station. The test instances used for comparison of all three approaches are the same as the 20 instances described in Table 4, with the number of ambulance stations varying from 5 to 141. To compare all three approaches, we applied the same test instances for ten trials. The same neighbourhood change approach (AVNS-II) is used to obtain the final solution for all three initial solution approaches.

The summary of results obtained using all three approaches is tabulated in Tables 6 and 7. Table 6 shows the average initial and the average final solution values obtained by all three initial solution approaches. The initial solution obtained using CPLEX was significantly better than both GA- and greedy-based approaches, while the GA-based approach produced a solution consistently better than the greedy approach. But the difference in the average final solution for all three approaches was significantly lower. This indicates that although the initial solutions by all three approaches were significantly different, they converged to almost the same objective function value. Table 6 also

Table 6
Comparison of initial solution approaches based on overall objective function value obtained (bold numbers indicate highest average final objective value/highest best objective value).

Instance	Average initial objective value			Average final objective value			Best objective value			% Difference in best objective value		
	Greedy	GA	CPLEX	Greedy	GA	CPLEX	Greedy	GA	CPLEX	Greedy	GA	CPLEX
1	7438.7	9431.2	10666.3	12980.1	**12995.5**	12975.7	12999.8	13038.7	**13051.3**	0.39	0.1	0.00
2	8413.6	11125.2	11646.8	15704.6	15899.7	**15901.0**	15902.6	**15944.2**	15966.3	0.40	0.14	0.00
3	5271.8	7969.3	11167.4	19940.9	**20153.2**	20086.0	20204.7	**20205.3**	20175.5	0.00	0.00	0.15
4	6169.3	11502.1	13316.6	23514.8	23594.5	**23601.9**	23703.2	**23714.4**	23692.1	0.05	0.00	0.09
5	6769.7	12433.3	20758.0	30362.9	30608.9	**30681.6**	30654.5	30776.8	**30896.3**	0.78	0.39	0.00
6	4127.4	15061.6	23428.8	38185.6	38407.8	**38543.2**	38710.8	38667.9	**38669.8**	0.00	0.11	0.11
7	5108.4	17780.5	32723.4	55399.7	**55500.7**	55389.8	54960.9	**55949.1**	55793.6	1.77	0.00	0.28
8	12278.7	27780.4	37933.5	64019.7	**64230.8**	64198.3	64567.7	**64728.7**	64568.7	0.25	0.00	0.25
9	42348.7	53579.9	66462.0	88792.9	89761.5	**90159.2**	90181.5	**90699.8**	90610.05	0.57	0.00	0.10
10	56724.6	58820.1	75282.4	102917.0	103540.5	**104012.2**	104474.9	104489.9	**104512.9**	0.04	0.02	0.00
11	58899.5	73681.7	85119.4	127619.3	128798.8	**129263.1**	128951.7	129917.9	**129917.9**	0.74	0.00	0.00
12	70357.7	84828.9	92285.8	147433.3	**147832.0**	**147832.0**	149062.3	149056.6	**149098.1**	0.02	0.03	0.00
13	90044.5	113167.2	132698.1	157968.0	158706.5	**158969.4**	160148.5	**160833.5**	160683.9	0.43	0.00	0.09
14	99706.7	107914.0	139893.6	194124.8	194782.6	**195747.0**	196573.1	196522.5	**196845.8**	0.14	0.16	0.00
15	97798.3	124857.6	143479.1	181039.9	182667.8	**183145.4**	182762.6	**184749.0**	184575.2	1.08	0.00	0.09
16	118027.6	129747.0	158672.2	223854.0	224957.7	**225151.1**	226121.5	**226664.9**	226192.3	0.24	0.00	0.21
17	119232.4	153183.0	162240.3	192733.4	196105.4	**196194.6**	196378.2	197586.9	**197775.4**	0.71	0.10	0.00
18	120167.4	152924.5	170120.6	245797.5	244836.0	**246932.2**	248029.8	248689.3	**249029.8**	0.40	0.14	0.00
19	126272.2	159172.3	176756.0	210619.9	**212846.5**	212271.70	213995.5	214795.5	**214826.7**	0.39	0.01	0.00
20	128845.8	171465.4	372947.8	282643.0	283588.4	**283611.1**	285655.4	286306.5	**286416.0**	0.27	0.04	0.00

Table 7
Comparison of initial solution approaches based on CPU time and ARPD (bold numbers indicate lowest mean CPU time/ smallest ARPD values).

Instance	Mean CPU Time (initial solution)			Mean CPU Time (Final solution)			ARPD		
	Greedy	GA	CPLEX	Greedy	GA	CPLEX	Greedy	GA	CPLEX
1	0.00	2.22	0.26	53.93	**34.50**	44.70	0.55	**0.43**	0.58
2	0.00	2.40	0.23	26.62	**26.54**	39.90	1.64	0.42	**0.41**
3	0.00	5.41	0.57	119.54	**68.11**	141.30	1.31	**0.26**	0.59
4	0.00	5.15	0.53	176.60	164.62	**110.90**	0.84	0.51	**0.47**
5	0.00	9.98	1.00	315.54	**213.55**	226.20	1.73	0.93	**0.69**
6	0.00	9.40	0.99	232.51	228.90	**179.60**	1.36	0.78	**0.43**
7	0.02	11.91	2.26	1081.77	897.04	**756.80**	0.98	**0.80**	1.00
8	0.02	11.95	2.24	1481.41	1069.55	**899.00**	1.10	**0.77**	0.82
9	0.02	13.62	7.45	2187.30	1962.97	**1818.60**	2.10	1.03	**0.60**
10	0.02	13.63	6.98	1855.46	2007.32	**1389.20**	1.53	0.93	**0.48**
11	0.02	13.70	10.35	3190.48	2642.22	**2467.80**	1.77	0.86	**0.50**
12	0.02	13.89	10.14	2707.61	2358.53	**1999.00**	1.12	0.85	0.85
13	0.02	18.44	16.82	4415.47	3166.44	**3164.30**	1.78	1.32	**1.16**
14	0.02	19.73	16.04	4993.05	**2469.51**	3501.00	1.38	1.05	**0.56**
15	0.02	21.67	24.50	5452.57	**3504.62**	3843.30	2.01	1.13	**0.87**
16	0.02	21.42	24.97	4823.79	4220.17	**4193.00**	1.24	0.75	**0.67**
17	0.02	28.08	35.60	6787.79	5519.39	**4655.20**	2.55	0.84	**0.80**
18	0.02	29.06	34.50	6029.36	5688.99	**5620.30**	1.30	1.68	**0.84**
19	0.02	39.80	48.58	7053.06	5920.09	**5615.50**	1.96	**0.92**	1.19
20	0.02	38.84	47.42	7089.23	6892.98	**6521.80**	1.32	0.99	**0.98**

shows the best solution obtained using all three approaches. It can be seen that in 11 out of 20 instances, the best final solution was obtained when an initial solution was obtained using CPLEX. This indicates that a good initial solution seems to increase the possibility of obtaining a better final solution. The greedy approach led to slightly lower solutions than the best solution in most cases than the other two approaches. The percentage difference between the best solution obtained using each approach when compared with the overall best solution is less than 1% for most cases.

Table 7 shows the mean CPU time for obtaining the initial and final solutions for all three approaches. The time taken to produce the initial solution is the lowest for the greedy approach for all instances. Among CPLEX and GA, for smaller instances, the time taken for obtaining the initial solution is significantly lower for CPLEX, but GA takes lower time as the problem size increases. In the case of time taken to obtain the final solution, CPLEX consistently took lower time than the other two approaches. Similarly, GA took less time than the greedy approach to obtain the best solution. This can be explained by the fact that the initial solution obtained using CPLEX is the best, while GA is better than the greedy approach. Therefore, starting with a better initial solution, both GA and CPLEX converge to a better final solution. This difference can also be seen in ARPD, with both CPLEX and GA resulting in a lower ARPD compared to the greedy approach. The major insight based on the results is that solving a relaxed location-allocation problem to obtain initial solutions for the AVNS algorithm leads to good-quality initial solutions. Even the initial solution generated using a heuristic approach for the relaxed problem improves the performance of the AVNS algorithm than starting with a completely random initial solution.

6.6. Analysis of the impact of variation in resources and temporal variation

Fig. 5 shows the variation in travel time between the two wards based on the Uber-movement dataset. The travel time is typically lowest during the night (12.00——06.00 am), which increases from morning to noon and peaks during noon. The travel time further increases in the

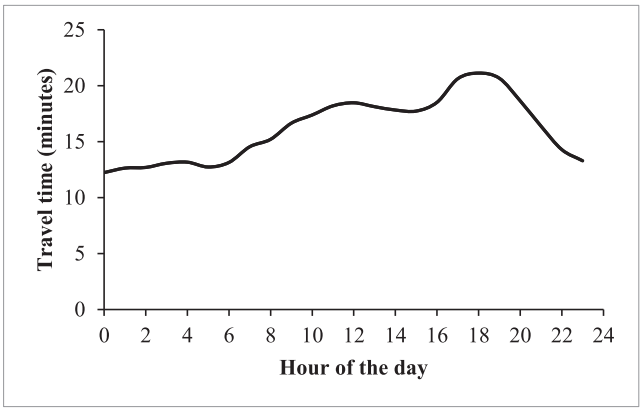

Fig. 5. Variation in travel time by hour of the day between two wards from the Uber-movement dataset.

evening and is highest between 06.00 pm and 08.00 pm, after which it declines. It can be observed that the travel time at the highest is greater than 50% above the minimum time during the night. Similarly, demand is also lowest during the night between 12.00 am to 06.00 am and starts rising after 06.00 am. It reaches a peak during noon, remains high throughout the day until 08.00 pm, and then decreases rapidly. Thus, travel time and demand vary throughout the day and are lowest at night and highest between noon and evening. To capture the impact of this variation on the ambulance location decisions, we solved the instances by varying the amount of resources available throughout the day. This is a practical consideration since the demand is low during the night, and the number of ambulances deployed will also be lower due to crew availability and lower costs.

Fig. 6 shows the variation in the mean utilisation level of ambulances over the day for the different numbers of ambulances. In Fig. 6, different lines represent the different numbers of ambulances available to allocate. For example, (25, 25, 10) indicates 25 ALS, 25 BLS and 10 FRV available. Server utilisation indicates the proportion of time the ambulances at a station are busy serving an emergency call. This is calculated by dividing the total service time of all ambulances by the total time available. Server utilisation varies over the day, with higher utilisation between 08.00 am to 08.00 pm compared to other periods. Server utilisation decreases with an increase in ambulances as the load gets distributed over more ambulances. Between 08.00 pm and 08.00 am, the utilisation drops below 60%, with 48 ambulances represented by (20, 20, 8) in the figure. However, 60 ambulances are required to achieve an utilisation level of 60% for the remaining period during the day.

Fig. 7 shows the variation in the proportion of calls expected to be served during each period of the day for the different numbers of ambulances. The proportion of calls served represents the percentage of calls where an ambulance is dispatched when a call arrives. The remaining calls can be considered to be lost as the ambulances might be busy when a call arrives. We can observe that the proportion of calls served is higher during the night for the same number of ambulances. This variation can be easily explained by the fact that both demand and travel times are low between 08.00 pm and 08.00 am. For example, with 42 ambulances (represented by (18, 18, 6)) allocated, the proportion of calls served is above 90% between 08.00 pm and 08.00 am. However, it requires more than 50 ambulances to serve more than 90% of the calls during the daytime. This is an important observation since it indicates that a larger number of ambulances are required to be allocated during the day while a lower number of ambulances can achieve similar service levels during the night. Additionally, from Fig. 7, we can observe that as the number of ambulances increases, the percentage increase in the proportion of calls served decreases. This can be very clearly observed during the period 12.00 am to 4.00 am, where an increasing number of ambulances from 25 to 36 increases the proportion of calls served from 50% to 80%. However, increasing from 48 ambulances to 60 resulted in a small improvement during the same period. Also, to serve 90% of calls during the night requires only 42 ambulances, while it requires above 52 ambulances during the day. Accounting for these time-dependent variations is important as the ambulance location decisions serve as input for other planning decisions, such as relocation and crew scheduling.

Further, to study the impact of the parameters Z_t and A^k_{max}, we have reported the results obtained by varying the maximum number of ambulances (A^k_{max}) and ambulance stations (Z_t). Fig. 8 shows the impact of varying the number of ambulances located on the expected proportion of calls served and the mean utilisation of servers. As the number of ambulances is increased, the overall proportion of calls served increases rapidly. However, increasing the total number of ambulances beyond 60 (25 ALS, 25 BLS, 10 FRV) does not result in significant improvement in the service level (calls served) as the proportion of calls served is already closer to 1. Similarly, from the figure we observe that mean server utilisation decreases with the increase in the number of ambulances. This is expected as more ambulances are available to serve the overall demand.

Fig. 9 presents the impact of the number of ambulances on the total number of relocations required throughout the day. The number of relocations increases as the number of ambulances is increased. However,

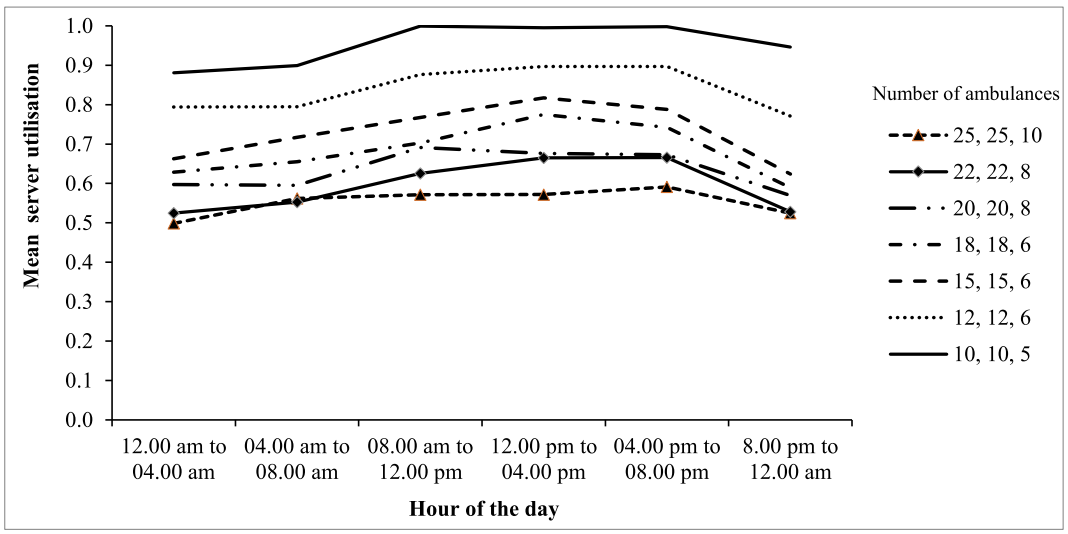

Fig. 6. Impact of temporal variation on mean utilisation level of servers for different numbers of ambulances.

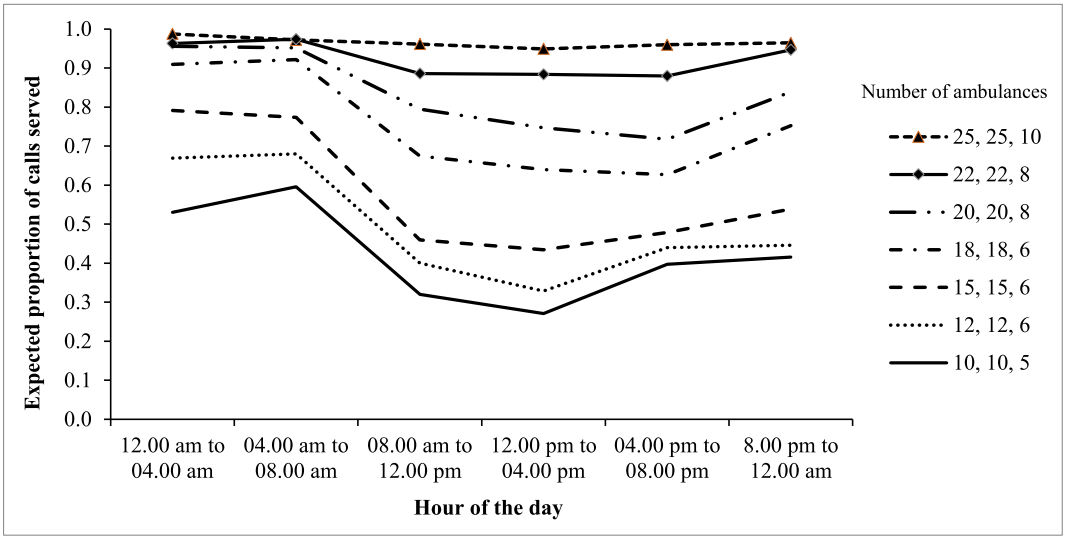

Fig. 7. Impact of temporal variation on the expected proportion of calls for different numbers of ambulances available.

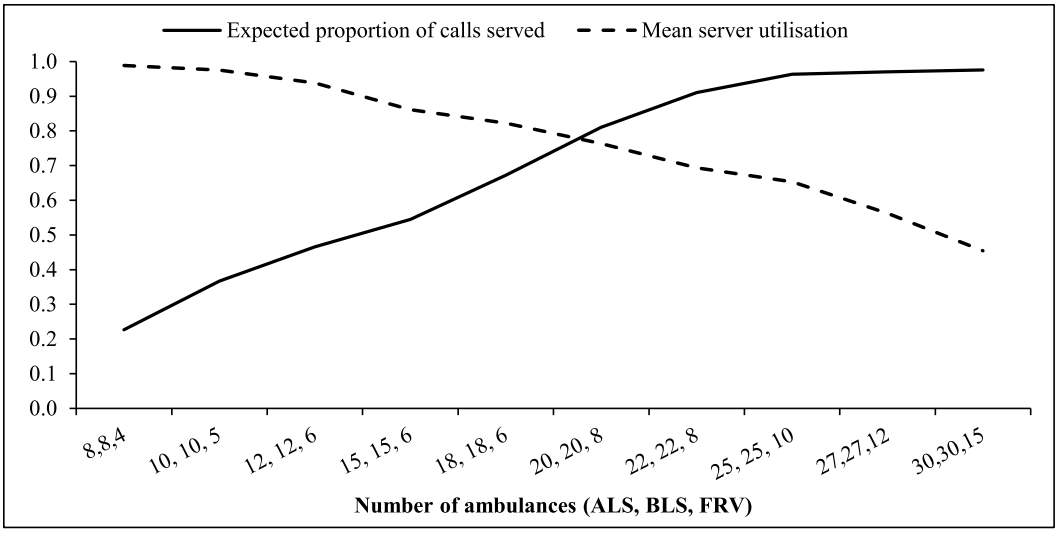

Fig. 8. Impact of the number of ambulances located on the expected proportion of calls served and mean server utilisation.

the proportion of ambulance relocations decreases when the number of ambulances is increased. The proportion of ambulance relocations is calculated as the ratio of the total number of relocations and the total number of ambulances. Another important observation is that as the total number of ambulances available is increased above 60, the number of relocations seems to slightly decrease. This decrease can be explained on the basis that as more ambulances are available, they can be located closer to the demand zones and thus reducing the need for relocations.

Fig. 10 shows the impact of the number of stations on the expected proportion of calls served and the mean server utilisation. The number of stations is increased from 20 to 32, while the number of ambulances is fixed at 60 with 25 ALS, 25 BLS and 10 FRV. The expected proportion of calls served increases slightly as the number of stations is increased. Similarly, mean sever utilisation decreases as the number of stations increases. Increasing the number of ambulance stations results in ambulance stations being located closer to the demand zones which reduces service time for calls and thus improves the expected proportion of calls served. However, the variation in server utilisation and proportion of calls served is lower compared to the variation due to the number of ambulances.

Similarly, Fig. 11 shows the variation in the number of relocations with the increase in number of stations. As can be seen, increasing the number of ambulance stations results in a slight increase in the number of relocations. However, the number of relocations seems to decrease slightly after reaching 20 relocations when the number of ambulance stations is 28.

Based on the above computational results, we can draw some key inferences that can provide key insights to both the managers/decision-makers and researchers in the domain of EMS planning. From the perspective of EMS managers, one important observation is that the number of ambulances required during peak periods during day time is very high compared to the night. EMS managers should, therefore, consider temporal variation in demand and travel time while determining the number of ambulances (i.e. the fleet size decision) required for a region. Another key observation is that increasing the number of ambulances or stations results in an improved expected proportion of calls served with a decrease in the utilisation of the ambulances. Additionally, increasing the number of ambulances and stations above a threshold only results in a marginal improvement in the utilisation and the expected number of calls served. Thus, EMS decision-makers need to find a balance between the utilisation of resources while improving the service level. From the perspective of EMS planning literature, our

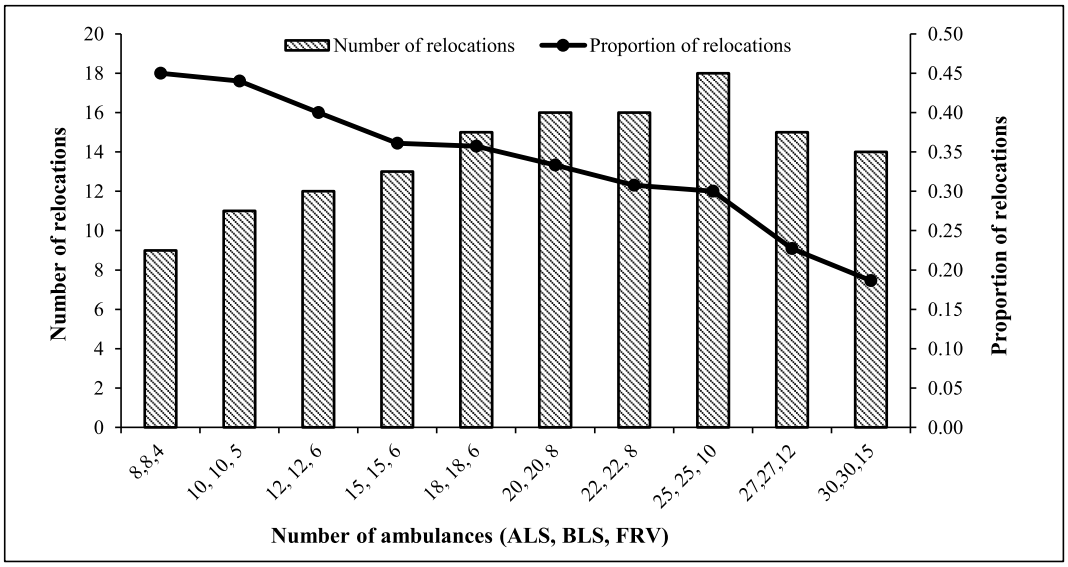

Fig. 9. Impact of number of ambulances located on the number of relocations and the proportion of relocations.

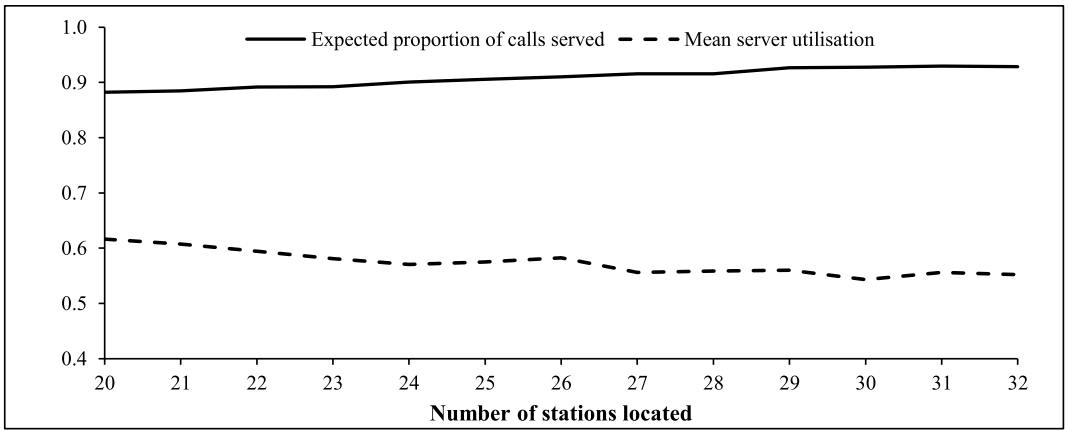

Fig. 10. Impact of number of stations located on the expected proportion of calls served and mean sever utilisation.

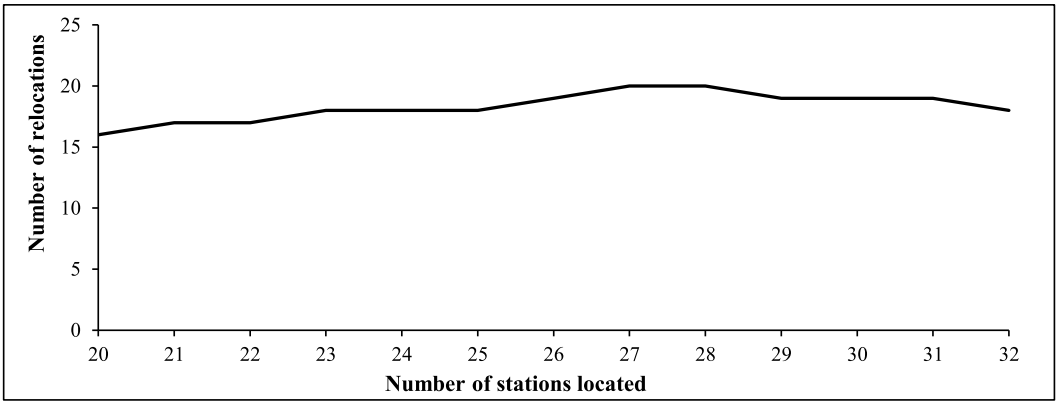

Fig. 11. Impact of number of stations located on the number of relocations.

results indicate the effectiveness of solving a relaxed ambulance location problem to obtain the initial solution for the joint ambulance location and dispatching problem. Similarly, we have demonstrated the application of an approximate approach which can be integrated into the ambulance location problems to obtain a better estimate of busy probabilities for ambulances.

7. Conclusions

Location planning of ambulances plays an important role in the ability of emergency service providers to respond to ambulance demand in a timely manner. This work considers a joint ambulance location and dispatching problem in a multi-tier EMS system with three ambulance

types, ALS, BLS and FRV, that respond to multiple patient types. The proposed model considers the determination of the dispatch preference order of stations for each demand zone as a decision variable, along with the optimal location of ambulance stations and allocation of ambulances to these stations. Temporal variation in demand and travel time is incorporated within the model by dividing the day into multiple periods and allowing for possible relocation of ambulances during different shifts in a day. Thus, we contribute to the literature by extending the joint ambulance location and dispatching problem proposed by Toro-Diaz et al. (2013) to include multi-tiered ambulances and temporal variations in demand and travel time. We also incorporate an objective function based on a non-linear survival function to accurately capture the impact of variation travel time. Additionally, we propose and demonstrate the application of a queueing-based approximation approach to estimate the server-level busy probability and dispatch probability of ambulances that allows to evaluate the performance of the ambulance location decision.

We develop a solution approach based on the adaptive VNS meta-heuristic to solve the problem. The initial solution for the algorithm is obtained by solving a relaxed ambulance station location problem to obtain the initial ambulance station locations. The initial solution for the number of ambulances assigned to each station is obtained using a PSO-based approach. The proposed approach is validated using a dataset created based on the city of Kolkata in India. The travel time data is obtained from the Uber-movement dataset for private taxis. We propose two different adaptive approaches for neighbourhood change in the VNS algorithm and compare the proposed approaches with the basic VNS with a fixed sequence of neighbourhood change. Similarly, we also compare three approaches for generating initial solutions for the relaxed ambulance station location problem: GA, CPLEX and greedy. Our results indicate that both the adaptive approaches outperform the basic approach for neighbourhood change in terms of the time required to obtain the best solution. Similarly, our results also show that solving the relaxed problem for the initial solution results in a better final solution in relatively less CPU time than an initial solution obtained using the naïve greedy approach.

We also compare our results by varying the number of ambulances available to consider the impact of temporal variation on server utilisation and the proportion of demand served. Server utilisation decreases during the night when demand is low, and the proportion of demand served increases during the night. This indicates that a relatively lower number of ambulances is required to achieve the same level of service during the night compared to the daytime. One major implication of our results is that considering an average travel time or demand over the day could lead to an underestimation of ambulances required during peak demand periods. Similarly, it could also lead to the underutilisation of ambulances during periods with lower demand. Thus, our contributions to the literature are threefold. One, we have presented a realistic ambulance location and dispatching model that considers multi-tiered EMS and incorporates time-dependent demand and travel time, an approximate performance evaluation approach for estimating dispatch probabilities, and a survival probability-based objective function. Two, we have presented and demonstrated an adaptive VNS to solve the problem and also show that the adaptive approach outperforms the basic VNS. Finally, through our results, we demonstrate the effectiveness of solving a relaxed problem for the initial solution of the algorithm and also the impact of considering a time-dependent variation on ambulance location decisions.

Although our model considers a more realistic ambulance location model, there are some limitations that can be addressed in further research. One major limitation is the calculation of long-run probability using an approximate approach instead of more accurate conditional probabilities for busy and dispatch probabilities. Developing more exact methods for estimating these probabilities is a possible scope for future research. We consider only a single objective while locating ambulances with other objectives such as equity or fairness while allocating ambulances within the context of multi-tier EMS with temporal variation can be evaluated. We consider the variation in demand and travel time over the day, but we do not consider the variation due to other factors, such as weekends and heavy rains, that can affect both demand and travel time. We have also not considered the non-urgent ambulance demand in our model. We have considered the day to be divided into equal lengths of four hours. However, considering unequal lengths and a different number of periods can be considered in future scope. Additionally, the impact of the number of ambulance stations and the number of ambulances located at each station can also be considered for further research. Developing discrete event simulation models that can both validate the results from the approximate approach for large-scale systems and gain better insights into the performance of the system is a possible direction for further research. A simulation model can also be used to test the output configuration from the ambulance location model to determine the quality of the solution obtained. Another possible direction for future research is developing exact methods by relaxing the non-linear constraints or implementing an approximation approach within an exact approach-based framework for solving the problem.

CRediT authorship contribution statement

Raviarun A. Nadar: Conceptualization, Data curation, Software, Formal analysis, Methodology, Validation, Visualization, Writing – original draft. **J.K. Jha:** Conceptualization, Formal analysis, Investigation, Supervision, Validation, Writing – review & editing. **Jitesh J. Thakkar:** Conceptualization, Supervision, Validation, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

Andersson, H., Granberg, T.A., Christiansen, M., Aartun, E.S., Leknes, H., 2020. Using optimisation to provide decision support for strategic emergency medical service planning – Three case studies. Int. J. Med. Inf. 133, 103975.
Aringhieri, R., Bruni, M.E., Khodaparasti, S., van Essen, J.T., 2017. Emergency medical services and beyond: Addressing new challenges through a wide literature review. Comput. Oper. Res. 78, 349–368.
Bélanger, V., Ruiz, A., Soriano, P., 2019. Recent optimisation models and trends in location, relocation, and dispatching of emergency medical vehicles. Eur. J. Oper. Res. 272 (1), 1–23.
Bélanger, V., Lanzarone, E., Nicoletta, V., Ruiz, A., Soriano, P., 2020. A recursive simulation-optimization framework for the ambulance location and dispatching problem. Eur. J. Oper. Res. 286 (2), 713–725.
Boujemaa, R., Jebali, A., Hammami, S., Ruiz, A., Bouchriha, H., 2018. A stochastic approach for designing two-tiered emergency medical service systems. Flex. Serv. Manuf. J. 30 (1), 123–152.
Boutilier, J.J., Chan, T.C., 2020. Ambulance emergency response optimisation in developing countries. Oper. Res. 68 (5), 1315–1334.
Brotcorne, L., Laporte, G., Semet, F., 2003. Ambulance location and relocation models. Eur. J. Oper. Res. 147 (3), 451–463.
Cantwell, K., Dietze, P., Morgans, A.E., Smith, K., 2013. Ambulance demand: Random events or predicable patterns? Emerg. Med. J. 30 (11), 883–887.
Cantwell, K., Morgans, A., Smith, K., Livingston, M., Spelman, T., Dietze, P., 2015. Time of day and day of week trends in EMS demand. Prehosp. Emerg. Care 19 (3), 425–431.
Chanta, S., Mayorga, M.E., Kurz, M.E., McLay, L.A., 2011. The minimum p-envy location problem: a new model for equitable distribution of emergency resources. IIE Trans. Healthcare Syst. Eng. 1 (2), 101–115.
Chanta, S., Mayorga, M.E., McLay, L.A., 2014. Improving emergency service in rural areas: a bi-objective covering location model for EMS systems. Ann. Oper. Res. 221 (1), 133–159.
Chong, K.C., Henderson, S.G., Lewis, M.E., 2016. The vehicle mix decision in emergency medical service systems. Manuf. Serv. Oper. Manag. 18 (3), 347–360.

Church, R., ReVelle, C., 1974. The maximal covering location problem. Papers Regional Sci. Assoc. 32 (1), 101–118.
Daskin, M.S., 1983. Maximum expected covering location model: formulation, properties and heuristic solution. Transp. Sci. 17 (1), 48–70.
De Maio, V.J., Stiell, I.G., Wells, G.A., Spaite, D.W., Ontario Prehospital Advanced Life Support Study Group, 2003. Optimal defibrillation response intervals for maximum out-of-hospital cardiac arrest survival rates. Ann. Emerg. Med. 42(2), 242–250.
Enayati, S., Mayorga, M.E., Toro-Díaz, H., Albert, L.A., 2019. Identifying trade-offs in equity and efficiency for simultaneously optimizing location and multipriority dispatch of ambulances. Int. Trans. Oper. Res. 26 (2), 415–438.
Erkut, E., Ingolfsson, A., Erdoğan, G., 2008. Ambulance location for maximum survival. Naval Res. Logist. (NRL) 55 (1), 42–58.
Gendreau, M., Laporte, G., Semet, F., 1997. Solving an ambulance location model by tabu search. Locat. Sci. 5 (2), 75–88.
Geroliminis, N., Karlaftis, M.G., Skabardonis, A., 2009. A spatial queuing model for the emergency vehicle districting and location problem. Transp. Res. B Methodol. 43 (7), 798–811.
Goldberg, J.B., 2004. Operations research models for the deployment of emergency services vehicles. EMS Manage. J. 1 (1), 20–39.
Hansen, P., Mladenović, N., Todosijević, R., Hanafi, S., 2017. Variable neighborhood search: basics and variants. EURO J. Comput. Optim. 5 (3), 423–454.
Iannoni, A.P., Morabito, R., Saydam, C., 2008. A hypercube queueing model embedded into a genetic algorithm for ambulance deployment on highways. Ann. Oper. Res. 157 (1), 207–224.
Ingolfsson, A., Budge, S., Erkut, E., 2008. Optimal ambulance location with random delays and travel times. Health Care Manag. Sci. 11 (3), 262–274.
Kennedy, J., Eberhart, R., 1995. Particle swarm optimization. In Proceedings of ICNN'95-International Conference on Neural Networks (Vol. 4, pp. 1942-1948). IEEE.
Khodaparasti, S., Maleki, H.R., Bruni, M.E., Jahedi, S., Beraldi, P., Conforti, D., 2016. Balancing efficiency and equity in location-allocation models with an application to strategic EMS design. Optim. Lett. 10 (5), 1053–1070.
Knight, V.A., Harper, P.R., Smith, L., 2012. Ambulance allocation for maximal survival with heterogeneous outcome measures. Omega 40 (6), 918–926.
Larson, R.C., 1974. A hypercube queuing model for facility location and redistricting in urban emergency services. Comput. Oper. Res. 1 (1), 67–95.
Larson, R.C., 1975. Approximating the performance of urban emergency service systems. Oper. Res. 23 (5), 845–868.
Lee, T., Cho, S.H., Jang, H., Turner, J.G., 2012. A simulation-based iterative method for a trauma center—Air ambulance location problem. In Proceedings of the 2012 Winter Simulation Conference (WSC) (pp. 1-12). IEEE.
Leknes, H., Aartun, E.S., Andersson, H., Christiansen, M., Granberg, T.A., 2017. Strategic ambulance location for heterogeneous regions. Eur. J. Oper. Res. 260 (1), 122–133.
Liu, Y., Li, Z., Liu, J., Patel, H., 2016. A double standard model for allocating limited emergency medical service vehicle resources ensuring service reliability. Transp. Res. Part C 69, 120–133.
Mandell, M.B., 1998. Covering models for two-tiered emergency medical services systems. Locat. Sci. 6 (1–4), 355–368.
McCormack, R., Coates, G., 2015. A simulation model to enable the optimization of ambulance fleet allocation and base station location for increased patient survival. Eur. J. Oper. Res. 247 (1), 294–309.
McLay, L.A., 2009. A maximum expected covering location model with two types of servers. IIE Trans. 41 (8), 730–741.
McLay, L.A., Mayorga, M.E., 2010. Evaluating emergency medical service performance measures. Health Care Manag. Sci. 13 (2), 124–136.
Mladenović, N., Hansen, P., 1997. Variable neighborhood search. Comput. Oper. Res. 24 (11), 1097–1100.
Nadar, R.A., Jha, J.K., Thakkar, J.J., 2021. Strategic location of ambulances under temporal variation in demand and travel time using variable neighbourhood search based approach. Comput. Ind. Eng. 162.
Naji, H.Z., Al-Behadili, M., Al-Maliky, F., 2021. Two server dynamic coverage location model under stochastic travel time. Int. J. Appl. Comput. Math. 7 (1), 1–19.
Nelas, J., Dias, J., 2020. Optimal emergency vehicles location: an approach considering the hierarchy and substitutability of resources. Eur. J. Oper. Res. 287 (2), 583–599.
Rajagopalan, H.K., Saydam, C., Xiao, J., 2008. A multiperiod set covering location model for dynamic redeployment of ambulances. Comput. Oper. Res. 35 (3), 814–826.
Reeves, C.R., 1995. A genetic algorithm for flowshop sequencing. Comput. Oper. Res. 22 (1), 5–13.
Repede, J.F., Bernardo, J.J., 1994. Developing and validating a decision support system for locating emergency medical vehicles in Louisville, Kentucky. *Eur. J. Oper. Res.* 75 (3), 567–581.
Reuter-Oppermann, M., van den Berg, P.L., Vile, J.L., 2017. Logistics for emergency medical service systems. Health Syst. 6 (3), 187–208.
ReVelle, C., Bigman, D., Schilling, D., Cohon, J., Church, R., 1977. Facility location: a review of context-free and EMS models. Health Serv. Res. 12 (2), 129.
ReVelle, C., Hogan, K., 1989. The maximum availability location problem. Transp. Sci. 23 (3), 192–200.
Saydam, C., Aytuğ, H., 2003. Accurate estimation of expected coverage: revisited. Socioecon. Plann. Sci. 37 (1), 69–80.
Saydam, C., Rajagopalan, H.K., Sharer, E., Lawrimore-Belanger, K., 2013. The dynamic redeployment coverage location model. Health Syst. 2 (2), 103–119.
Schmid, V., Doerner, K.F., 2010. Ambulance location and relocation problems with time-dependent travel times. Eur. J. Oper. Res. 207 (3), 1293–1303.
Toregas, C., Swain, R., ReVelle, C., Bergman, L., 1971. The location of emergency service facilities. Oper. Res. 19 (6), 1363–1373.
Toro-Díaz, H., Mayorga, M.E., Chanta, S., McLay, L.A., 2013. Joint location and dispatching decisions for emergency medical services. Comput. Ind. Eng. 64 (4), 917–928.
Toro-Díaz, H., Mayorga, M.E., McLay, L.A., Rajagopalan, H.K., Saydam, C., 2015. Reducing disparities in large-scale emergency medical service systems. J. Oper. Res. Soc. 66 (7), 1169–1181.
Uber Technologies, Inc. Data retrieved from Uber Movement, (c) 2021. https://movement.uber.com.
Van Den Berg, P.L., Aardal, K., 2015. Time-dependent MEXCLP with start-up and relocation cost. Eur. J. Oper. Res. 242 (2), 383–389.
Wang, K., Qin, H., Huang, Y., Luo, M., Zhou, L., 2021. Surgery scheduling in outpatient procedure centre with re-entrant patient flow and fuzzy service times. Omega 102.
Yoon, S., Albert, L.A., White, V.M., 2021. A stochastic programming approach for locating and dispatching two types of ambulances. Transp. Sci. 55 (2), 275–296.