Emergency vehicle routing in urban road networks with multi-stakeholder cooperation

Ziling Zeng

Department of Architecture and Civil Engineering, Chalmers University of Technology, 41296 Gothenburg, Sweden

The Hong Kong Polytechnic University Shenzhen Research Institute, Nanshan District, 518057 Shenzhen, China

Email: ziling@chalmers.se

Wen Yi

School of Built Environment, College of Sciences, Massey University, 0632 Auckland, New Zealand

Email: wenyi0906@gmail.com

Shuaian Wang

(Corresponding author)

The Hong Kong Polytechnic University Shenzhen Research Institute, Nanshan District, 518057 Shenzhen, China

Department of Logistics and Maritime Studies, The Hong Kong Polytechnic University, Hong Kong

Email: wangshuaian@gmail.com

Xiaobo Qu

Department of Architecture and Civil Engineering, Chalmers University of Technology, 41296 Gothenburg, Sweden

Email: drxiaoboqu@gmail.com

Abstract

The lack of multi-stakeholder cooperation is one of the main challenges faced by emergency medical services (EMS). Especially in the ambulance routing process, inactive traffic operators fail to provide coordination to prioritize the ambulance, while ignoring the choice of hospitals will lead to inevitable patient transfer between hospitals. To provide efficient decision support for EMS, this paper considers daily ambulance routing problems in a network with high spatial resolution in which two advanced technologies are introduced: pre-hospital screening that provides patient injury diagnosis and lane pre-clearing that ensures the pre-defined driving speed of ambulances. Three different types of ambulances are used to transport and offer first aids to patients based on the screened results. To manage the ambulance fleet properly, a mixed-integer linear programming (MIP) model is proposed to assign vehicles to the injured and plan routes with the shortest travel time. A semi-soft time window constraint is incorporated to reflect the late arrival penalty on-site and at hospitals. Since high-quality EMS responds to the call in seconds, a real-world case in Shenzhen, China, is presented to validate the computational performance by a commercial solver GAMS. In the case study, we further analyzed the effect of different stakeholders' involvement, like the hospitals and traffic operators. This information proves the efficiency of multi-stakeholder participation in ambulance routing.

Keywords

Ambulance routing; ambulance allocation; pre-hospital screening; lane pre-clearing; multi-stakeholder cooperation

1. Introduction

1.1. Motivations

When life-threatening incidents occur, the efficacy of response actions would mean the difference between life or death (Berkoune et al., 2012). Sánchez-Mangas et al. (2010) reported that a 10-min reduction of treatment waiting time reduces mortality by one-third. Motivated by this, the Emergency Medical Service (EMS) responsible for the logistics plan and operations must respond quickly and launch the most effective rescue (Jung and Qin, 2020). A systematic EMS covers phases from pre- to post-disaster. For disaster preparedness concerned with risk-mitigation processes, measures such as infrastructure reinforcement, resource prepositioning, and request prediction are planed (Lee et al., 2013). In the response phase, the post-disaster actions are taken immediately, such as care facilities relocations, resource dispatching (such as ambulances, EMS personnel, and hospital), patients' transportation and treatment (Wang et al., 2012). In the recovery phase, reconstruction activities are taken to return a specific area to 'normality' after being devastated (Salum et al., 2020, Iliopoulou et al., 2020). Among the three stages in the EMS environment, only the disaster response actions are required to be decided within a few seconds or even less (Dimitriou et al., 2018), which raises a strong need for real-time decision support, especially for the ambulance management in the aftermath of accidents.

Several studies tackle ambulance locating (Knight et al., 2012), dispatching (Schmid, 2012), and routing problems (Talarico et al., 2015) as ambulance management. But the suggested methods lack a holistic and integrated methodological framework that coordinates all relevant stakeholders, including emergency medical services providers (EMSP), traffic operators, and hospitals. For ambulance routing problem (ARP), much research emphasizes solving ARP by mathematical models focusing on the collaboration between hospitals and EMSP or by simulation methods with joint consideration of EMSP and traffic operators. However, the absence of any one of the above three decision-makers cannot guarantee the effectiveness of response actions. In the case of ignoring traffic operators, few responses would be provided to prioritize ambulances and ensure the driving speed. When hospitals are inactive, EMSP may select the nearest hospital that lacks the expertise to treat the patients, leading to inevitable transfer (Gao et al., 2020).

The EMSP has to design a rescue plan to cope with the dynamic traffic condition and the scarce resources such as hospitals and ambulances while guaranteeing the shortest transport time and proper resource allocation. Thus, traffic operators and hospitals must be involved in the planning process. It offers an underexplored opportunity to consider both hospitals' expertise and traffic conditions in ARP.

1.2. Literature review

In this section, we present a literature review of earlier work in both disaster preparedness and response phases, focusing on the ambulance (re)locating, dispatching, and routing problems.

1.2.1. Ambulance locating problem

Ambulance locating problem is a branch of resource location design that is of a tactical nature and often based on static information (Li et al., 2019).It aims to find the deployment sites for the ambulance fleet in a certain area to support the EMS. A review of methodologies and algorithms for ambulance locating problem in a coverage notion is summarized by Farahani et al. (2012). The coverage can be defined as response time (Erdemir et al., 2010), total service time (Jagtenberg and Mason, 2020), and the expected survival possibility (Knight et al., 2012). The selected locations ensure that the estimated demand can be satisfied within a given time.

During daily operations, the ambulance that has been assign to an emergency request can be reallocated to improve the coverage. Lam et al. (2015) proposed a system status management based method to reassign ambulance deployment locations on a daily basis. This problem was formulated as a double coverage model and solved using CPLEX solver. To further consider the dynamic and uncertain behavior of EMS, Acar and Kaya (2019) constructed the integrated location and relocation model for mobile hospitals using a two-stage stochastic programming model. The stochasticity reflects the possibility of the hospital being damaged and the possibility of passenger transfer. The multi-objective function consisted of the total travel time, penalization of capacity shortage, the unused capacity, the hospital transfer time, and the mobile hospital relocating time. The proposed model was solved by GAMS solver. Other models have been introduced to include stochastic travel times (Schmid and Doerner, 2010) and service time (Goldberg and Paz, 1991).

1.2.2. Ambulance dispatching problem

The ambulance dispatching model allocates emergency calls to the vehicle based on its location. Some commonly used dispatching policies are assigning the task to the nearest resource, first-come-first-serve policy and fixed plan (Kuisma et al., 2004). Comparing with these strategies, the benefit of dispatching model is discussed by Jagtenberg et al. (2017).

When making the deployment policy, some studies considered the urgency of the call. McLay and Mayorga (2013) introduced a Markov Decision Process model to dispatch ambulances to prioritized patients considering the classification errors. This model maximized the expected coverage of true high-risk patients. Bandara et al. (2014) developed a simulation model that incorporates the severity of the request to implement the suggested dispatching policies in EMS. They aim to maximize the patients' survival probability. A heuristic algorithm is customized to solve the large-scale problem. Andersson and Värbrand (2007) allocated the ambulance based on the priority of the request and the travel distance. They combined the dispatching and relocating of the ambulance fleet to ensure the coverage of patients.

1.2.3. Ambulance routing problem

The ambulance routing problem seeks to find the shortest path to pick-up and drop-off casualties in a network with high or low spatial resolution.

For high spatial resolution, the physical road network and traffic conditions are formulated in the model. Jotshi et al. (2009) proposed a method to search for the shortest path considering the patient priorities, clustering criteria, distance, road congestion, and hospital availability. Network partition is introduced to reduce the computation complexity.

In contrast, more researches are based on low spatial resolution network. In general, the specific physical path between the hospital and the wounded is simplified into one link. Based on the simplified network, Talarico et al. (2015) consider patient classification. They divided the patients into two categories: red and green. The red should be driven directly from the spot to the assigned hospital, while the green can be taken care of at the accident scene. Accordingly, a MIP model was constructed based on the multi-commodity model to minimize the worst-case patient waiting time. A large neighborhood search metaheuristic was applied to solve the problem. Based on this research, Tikani and Setak (2019) further increased the patient categories to three and classified the ambulance fleet. They added the late arrival penalty cost in the objective function by setting a soft time window. The genetic algorithm is introduced to solve the problem. When an ambulance is allowed to serve a list of patients at different locations, patient clusters are introduced. Similarly, (Zidi et al., 2019) proposed a cluster-first route-second algorithm to tackle the ARP.

These researches formulized the compatible constraints between patients and hospitals as the capacity limitation. Assumptions are made for hospitals' expertise where hospitals can treat all kinds of injuries, which is not realistic.

1.2.4. Joint studies

The interaction among the sub-problems are considered in some studies.

In the field of location-dispatching problems, Toro-Díaz et al. (2013) proposed a joint location and dispatching integer programming model for EMS. The model aims to minimize the response time and maximize the coverage considering queuing elements and congestion phenomena in the dispatching process. A genetic algorithm was introduced to solve the problem. Ibri et al. (2012) developed a decentralized distributed solution approach based on multi-agent systems to jointly locate and dispatch emergency vehicles. They aim to coordinate agents to reach reasonable quality solutions.

For ambulance location-routing problems, Oran et al. (2012) introduced a formulation of emergency facility locating and vehicle routing with time windows that consider the priority of emergency calls. A MIP solver and tabu search algorithm were introduced for problem-solving. Further, Caunhye et al. (2016) presented a two-stage location-routing model with recourse under uncertainty. The objective function is to minimize the total preparedness cost and the worst-case response time with uncertainty consideration. The ambulance location-allocation-routing problem is designed for temporary EMS. Memari et al. (2020) proposed a bi-objective dynamic model to minimize the operational costs and the critical time spent before being treated. Two meta-heuristic algorithms are developed for problem-solving.

1.3. Contribution

The contributions of the present study can be summarized as follows.

(i) This paper focuses on the ambulance routing problem, which strives to involve EMSP, traffic operators, and hospitals in the planning process. In detail, the pre-hospitals screening and lane pre-cleaning are implemented as input to speed up the first aid and avoid inefficient delivery.

(ii) The MIP optimization model is proposed for ambulance dispatching and vehicle routing based on a high-spatial-resolution network to reduce the transport time, the dispatching cost, and the late arrival penalty. Patients with different severity will be allocated to the hospital with the proper expertise, while ambulance allocation to patients depends on travel time after lane clearing.

(iii) The semi-soft time windows constraint is formulated to reflect the urgency of rescue, and a late arrival penalty on-scene and at hospitals is introduced in the objective function.

(iv) A real-world case in Shenzhen, China, is studied to validate the efficiency in rescue time and computational time. The exact optimal solution can be generated within a short computational time by commercial solvers. The comparison with cases with inactive stakeholders is made to verify the resulting efficiency.

The remainder of the paper is as follows. In Section 2, the problem is defined in detail, and notations are explained. Moreover, two advanced technologies are presented to involve the crucial stakeholders. The high-spatial resolution-based MIP model is illustrated and described in section 3. Section 4 provides a real-world case study in Shenzhen, China, to test the proposed mathematical model and evaluate the performance compared with inactive stakeholder cases. In section 5, the main conclusions and further remarks are summarized.

2. Problem statement

The ambulance routing problem aims to plan routes to pick up patients and drop them off at the hospitals. This problem will be fundamentally different from the traditional vehicle routing problem by taking into account two advanced technologies: pre-hospital screening and lane pre-clearing, as shown in Fig. 1. Typically, the ambulance routing problem can be categorized into two classes: hospital-based and depot-based. For the depot-based system, ambulances belong to hospitals and are initially located at their hospitals. In some cases, ambulance together with other emergency vehicles, will be positioned at a depot, which is defined as a depot-based system. In this research, we focus on the hospital-based one. By customizing the initial location for the ambulance fleet, the proposed method can be implemented in depot-based scenarios. We formalize the problem using the notation shown in Table 1.

Fig.1 Ambulance routing process. (a) Pre-hospital screening. Patients are diagnosed by screen infrastructures and further divided into different injury levels. The patients will be assigned to the nearest qualified hospital to be treated based on the disanose result. Ambulances are classified into three types according to their on-board equipment. They can serve patients at different injury levels. (b) Ambulance routing. The ambulance from the hospital with the shortest pick-up time will be allocated to help the patients and drop them off at the pre-defined hospital. (c) Lane pre-clearing. To ensure the driving speed of the ambulance, the preceding vehicles will switch to another lane to clear one specific lane following the suggestion given by the traffic operator.

2.1. Pre-hospital screening

After receiving an emergency call, remote screening of the injured helps allocate resources accurately, such as ambulance and hospitals. When there are more than patients, it is necessary to differentiate those with severe injuries acutely requiring specialized care. A minor delay in receiving treatment may cause the difference between lifelong disablement and independent life. Thus, the implementation of pre-hospital screening in EMS becomes popular recently. Persson et al. (2014) tested the accuracy of the brain diagnostic devices based on microwave technology for pre-hospital stroke screening. This equipment comprised triangle patch antennas fitted on the head that transmits signals for measurement and analysis. The signals were processed by a supervised learning algorithm based on training data from patients with the known condition. Two clinical tests were conducted to prove the efficiency of the equipment. Based on this research, Fhager et al. (2018) further summarized the promising microwave-based devices for pre-hospital diagnosis, including the diagnostic ability, methodologies, world-wide progress, and challenges. They claimed that, with the help of microwave devices, the clinical evaluations of trauma and stroke could be performed by research nurses and physicians without the need for technical measurements in the hospital.

Fig. 1(a) shows the simplified result of pre-hospital screening, where patients are classified into different injury levels. In practice, patients' severity is diagnosed based on their specific symptoms, such as abdominal pain, allergic reactions, animal bites, violence, burns, cardiac or respiratory arrest. Besides, the figure indicates that the available ambulance fleet is divided into three types:

(i) Type I, these ambulances are designed for patient transport. The on-board equipment is basic ones for first aid and nursing care.

(ii) Type II, this type is for basic life support. A certain number of medical equipment should be provided. Patients require medical transportation, and continuous medical supervision will be assigned to it.

(iii) Type III serves as a mobile Intensive care unit (ICU). The well-trained professionals and stretchers are on-board. The equipment provided is sufficient to stabilize, treat and transport the injured to the target hospital. This ambulance will be allocated to patients who are severely injured and require ongoing care.

The ambulance fleet located in a hospital is composed of vehicles of different types. The injury screening would match the patient with a list of compliant vehicles but not the exact one. We assume that patients at each injury level will be treated by an exact hospital according to its expertise to determine the destination for each patient.

In this research, we assume that the pre-hospital screening is included in ARP. It helps assign proper ambulance and hospital to the patients following the time limitation and the hospital's expertise. We include the result of screening as one of the inputs. To be specific, a set K_l is defined for the vehicles that satisfied the rescue requirement, and a two-dimension parameter $n_l^{h\prime}$ shows the compatibility between patients and hospitals in detail. To understand the specific meaning of the input, let set K denote the fleet of ambulances available to provide aid to patients. Each ambulance has its unique ID and can handle different levels of injuries according to the on-board equipment. We denote by $K_l \in K$, a subsite of the whole ambulances set K that can provide appropriate aids for patients at injury level l. Each element in the set K_l may locate differently to cope with random calls and reduce the response time. The set $n_l^{h'}$ contains the result after diagnosing, which illustrates the exact number of patients detected as injury level l and should be taken care of at hospital h'. This setting determines the injury level and the destination (hospital) of each patient. We allow multiple ambulances to visit the same incident scene at the same time to serve a group of patients. But these patients are supposed to be dropped off at the pre-defined hospital without any transfer.

2.2. Lane pre-clearing

Lane pre-clearing is a strategy based on cooperative control designed for intelligent connected vehicles (CV). It aims to arrange the trajectories for proceeding vehicles to clear one lane for each ambulance, as shown in Fig. 1(c). The lane clearing request will be sent to the CVs within a communication range through vehicle-to-vehicle or vehicle-to-infrastructure communication. The communication range determines when to conduct the sorting trajectories for the proceeding CVs. In this case, ambulances can drive at the pre-desired speed without congestion and avoid impact on the CVs to some extent.

An ambulance sorting algorithm, developed by Wu et al. (2020), solved the lane pre-clearing problem on normal road segments ensuring the desired speed of ambulances while reducing the disturbances on CVs. They introduced the ambulance speed and real-time locations as decision variables. Based on the A* algorithm, a customized EV sorting algorithm was proposed to decide the optimal communication range and the merging trajectories for CVs. Besides, a linear relationship between the results and road density was calibrated. It provides possibilities for simplifying travel time uncertainties by converting the stochastic travel time into a deterministic equation, which will be discussed in Section 3.2.

3. Emergency vehicle routing problem

3.1. Model formulation

Based on the two technics mentioned above, we propose a mathematical formulation of ambulance routing. Some assumptions are made as follows:

(i) The ambulances are initially located at the hospitals to which they belong. The available ambulances are uniformly dispatched by the EMS. After delivering the injured, ambulances will return to the hospitals from which they depart. But this deadhead trip will not be scheduled in this research.

(ii) The hospital allocation to patients is known after the pre-hospital screening. The nearest hospital that satisfies the treatment requirements will be selected as the patients' destination without any transfer. But the hospital that dispatches ambulance is not pre-determined. The selection of an ambulance will further depend on the pick-up time and distance.

(iii) There is only one accident at a time, and an accident scene may have several patients. We allow more than one ambulance to visit the scene and pick-up the injured. The patients at the same injury level can be served by the same ambulance when the capacity allows.

(iv) We consider three types of ambulances. The capacity of the same ambulance is different when it serves patients at different injury levels. Type III ambulance can provide treatment for all injury levels and accommodate more patients compared with Type I and II.

The ambulance routing problem is modeled as follows:

$$\min \sum_{k\epsilon K}\left(\sum_{i\in N}\sum_{j\in N} t_{ij}x_{ij}^k + \sum_{h\in H}\sum_{j\in N}\tau^k x_{hj}^k + f_k\right) \tag{1}$$

subject to:

$$\sum_{j\in N} x_{hj}^k \le o_h^k \quad \forall h \in H,\ \forall k \in K \tag{2}$$

$$\sum_{j\in N} x_{ij}^k = \sum_{j\in N} x_{ji}^k \quad \forall i \in N\backslash H,\ \forall k \in K \tag{3}$$

$$\sum_{k\epsilon K_l}\sum_{j\epsilon N} c_l^k x_{jp}^k \ge \sum_{h'\in H} n_l^{h'} \quad \forall l \in L,\ \forall p \in H \tag{4}$$

$$\sum_{k\epsilon K_l}\sum_{j\epsilon N} c_l^k x_{jh'}^k \ge n_l^{h'} \quad \forall l \in L,\ \forall h' \in H \tag{5}$$

$$s_i^k + e_i + t_{ij} - M\left(1 - x_{ij}^k\right) \le s_j^k \quad \forall i \in N,\ \forall j \in N,\ \forall k \in K \tag{6}$$

$$s_i^k - \varepsilon_i \ \ge T_i \quad \forall i \in P \cup H,\ \forall k \in K \tag{7}$$

$$x_{ij}^k \in \{0,1\}\ \ \forall i \in N,\ \forall j \in N,\ \forall k \in K \tag{8}$$

The objective function (1) is to minimize the total cost, including the expenses on traveling, ambulance allocation, and delay penalty that will be discussed in Section 3.3. Constraints (2) ensures that each selected ambulance should start from the hospital, where it is initially located. Constraints (3) denotes the flow balance for intermediate nodes. Constraints (4) ensures that all patients detected as injury level l at node p must be picked up by the ambulances that can provide aids to this level. Similarly, Constraints (5) limits that patients at injury level l assigned to hospital h' should be dropped off at the same hospital. Constraint (6) describes the visit time at each node along the route of ambulance k. Constraint (7) describes the delay when arriving at the patient's node or the hospital based on a semi-soft time window. Constraint (8) defines the binary decision variable.

3.2. Travel time calculation

To integrate traditional traffic planning and emergency response methods, innovative technologies such as CVs have enabled new solutions. Transport operators can take actions to prioritize the ambulances, and drivers of other CVs will correspond to give way to ambulances through clear interaction. Thus, ambulances take advantage of real-time traffic information in routing to minimize the delivery time.

In this section, the travel time based on connectivity is generated in Constraint (9) using the method proposed by Wu et al. (2020), which provides a linear relationship between communication range and designed speed.

$$r = (a\sigma + b)\frac{v - m\sigma - n}{v_0 - m\sigma - n} \tag{9}$$

where r is the communication range between the ambulance and the proceeding CVs; a and b are the linear regression coefficients, $ak + b$ denotes the simulated communication range. m and n are the coefficients in the fundamental diagram, $mk + n$ represents the average speed of the proceeding CVs. v_0 is the speed of ambulance used in the simulation, and v is the real-time designed speed. We assume the communication range l is pre-assigned, the travel time of the emergency vehicle in the proceeding node can be calculated in Constraint (10) with only traffic density σ_{ij} which denote the density between nodes i and j:

$$t_{ij} = \frac{D_{ij}(a\sigma_{ij} + b)}{v_0 r + (a\sigma_{ij} + b - r)(m\sigma_{ij} + n)} \tag{10}$$

where D_{ij} represents the distance between two nodes.

3.3. Late arrival penalty

We consider the late arrival at the accident scene and the hospital. The on-site delay affects the efficacy of the first aid while the arrival delay at the hospital might be more fatal for the injured. Based on the semi-soft time window introduced in Constraint (7), the delay calculation can be presented as follow:

(i) To describe the on-site delay when ambulance k picks up the patient(s) at the accident scene $p \in P$, we set the penalty coefficient as ψ_1 to depict the degree of impact. The detailed delay can be calculated as $\sigma_k = \max\left(s_p^k - T_p, 0\right)$. When the arrival time s_p^k is earlier or equals the preferred arrival time T_p, the penalty is 0. Otherwise, the time equals $s_p^k - T_p$.

(ii) Similarly, when the ambulance drop off the patient(s) at injury level l at hospital node $h' \in H$, the penalty coefficient is set as is set as ψ_2, the delay at the hospital is denoted by $\theta_k = \max\left(s_{h'}^k - T_{h'}, 0\right)$.

The late arrival penalty for each ambulance k at two import places can be formulated as:

$$f_k = \psi_1\sigma_k + \psi_2\theta_k \tag{11}$$

The objective function can be rewritten as:

$$min \sum_{k \epsilon K}\left(\sum_{i \in N}\sum_{j \in N} t_{ij}x_{ij}^k + \sum_{a \in H}\sum_{j \in N}\tau_a^k x_{aj}^k + \psi_1\sigma_k + \psi_2\theta_k\right) \tag{12}$$

3.4. Applications for disaster response

In addition to serving a single accident point, the proposed model can also be used in large-scale multi-accident scenarios to support disaster response. In consideration of this, we extend the pre-hospital screening parameter $n_l^{h'}$ to $n_{l,p}^{h'}$ which indicates the number of patients who are diagnosed as injury level l and assigned to hospital h' calling from accident scene p. We modified the patient pick-up and delivery constraints (4-5) as follows:

$$\sum_{k \epsilon K_l}\sum_{j \epsilon N} c_l^k x_{jp}^k \geq \sum_{h' \in H} n_{l,p}^{h'} \quad \forall l \in L,\ \forall p \in H \tag{13}$$

$$\sum_{k \epsilon K_l}\sum_{j \epsilon N} c_l^k x_{jh'}^k \geq \sum_{p \in P} n_{l,p}^{h'} \quad \forall l \in L,\ \forall h' \in H \tag{14}$$

Constraint (13) ensures that the level l patients at scene p will be picked up by the ambulances that can serve patients at level l. Similarly, Constraint (14) indicates that all level l patients that are assigned to hospital h' will be delivered to the hospital without transfer.

After substituting constraints (4-5) with constraints (13-14), the proposed model makes it possible to solve the ambulance routing problem for disaster response.

4. Real-world case study

4.1. Description

This section focuses on a central part of Shenzhen in southern China, as shown in Fig. 2, where 6 hospitals are surrounded, including one emergency medical center with the largest ambulance fleet. The point of interest (POI) is marked in dark grays such as commercials, residential, hospitals, schools, and parking. We further sketch the road network in Fig. 3 to simplify overpasses, small intersections, and roads within residential areas. The travel time after lane pre-cleaning is calculated based on Constraint (10) in minutes and is shown above each link in Fig. 3. In this case study, we made two assumptions:

(i) Patients will be diagnosed and divided into four levels: 1, 2, 3 and 4. Ambulances can serve patients at different levels. For example, a type III ambulance can pick up the wounded at all injury levels; a Type II ambulance is capable for patients at level 1,2 and 3, and a Type I ambulance can only serve patients at injury level 1 and 2.

(ii) The on-site service time for patient picking up is around 1 minute, and is 0 at the intersection nodes where patients remain on-board.

Fig. 2 Physical road network. The red cross represents the hospitals' location, and the dark gray shows the points of interest in this area.

Fig. 3. A sketch network. This sketch abstracts the physical network and ignores internal roads and small intersections within residential areas. The red dot shows the hospitals' location, and the dark orange one represents the emergency medical center where locates the largest number of ambulances. The hollow dot shows the simplified intersection, and the blue one is where the accident happened.

Like other worldwide cities, in Shenzhen, the ambulance fleet located differently belongs to the EMSP and is uniformly dispatched after receiving the request. As mentioned before, the ambulances are classified into three types according to their on-board equipment. The detailed inventory is summarized in Table 2 according to the statistical data provided by Shenzhen EMSP.

One real-world incident is introduced as an example. In a residential area, node 11, a medium-sized traffic accident, sent a request to EMS at 9:00 am. Two drivers were seriously injured when cars collided, and two passengers and two pedestrians were wounded to varying degrees. After the pre-hospital screening, the wounded are classified into different levels and assigned to the hospitals with the proper expertise. The screening results are shown in Table 3.

To balance the weights of the three components of the objective function (1), we convert both the dispatch cost and the delay penalty into time units. We set the ambulance allocation price τ^k as 10,20, and 30 for type I, II, and III, respectively. Besides, the set the coefficient of late arrival as 10 for both on-site ψ_1 and at hospitals ψ_2. To ensure the service level, the preferred on-site pick-up time is 10 mins after the call, and the preferred delivery time is 20 mins after the request.

4.2. Computational result

The incident is implemented in the General Algebraic Modeling System (GAMS) 33.1.0, called by Python 3 installed on a Dell laptop with a 1.9-GHz Intel Core i7 CPU and 8-GB, running on Windows 10. The calculation time for this case is 6.909s.

The ambulance allocation and route plan are generated with detailed departing and arriving time shown in Fig. 4. Hospital 27 dispatches all available ambulances to serve patients at injury levels 2, 3, and 4, respectively. Hospital 3 allocates the type I ambulance to pick-up the patient at injury level 1. It is clear to conclude that the hospital with the shortest travel time are selected to dispatch ambulances. When the ambulance fleet is fully assigned, or the ambulance type is inappropriate, the second nearest hospital will be responsible for serving the patients left that are of relatively low severity and priority.

Fig.4. Ambulance allocation and route plan. (a) Type I ambulance initially located at hospital 3 serves patients at injury level 1 at 9:11 and delivers them to hospital 3 at 9:20. (b) Type I ambulance located at hospital 27 helps patients at injury level 2 at 9:08 and back to hospital 27 at 9:16. (c) Type II ambulance departs from hospital 27 to pick-up the injured at level 3 at 9:08 and delivers them to hospital 7 at 9:24. (d) Type III ambulance from hospital 27 serves one patient at injury level 4 at 9:08 and drop off at hospital 29 after 16 minutes.

Fig. 5 describes the total cost and the detailed cost of each component. The travel time, dispatch price, and delay penalty are shown in blue. The expenses on each passenger list are described in red. It can be observed that the higher injury level leads to a higher rescue cost. This is mainly because the severe casualty requires a better-equipped ambulance, which is more expensive to dispatch. Besides, some patients with fatal injuries or uncommon illnesses are difficult to be treated in the nearest hospital, so a long transport distance will also cause relatively high rescue costs. Conversely, patients at lower injury levels can be treated nearby.

Fig. 5 Rescue cost breakdown

4.3. Method comparison

To validate the efficiency of multi-stakeholder consideration, we take patient list 3 as an example of 2 patients waiting to be treated. We compare the generated result with cases that have inactive hospitals and traffic operators, respectively. In the first case, traffic operators are involved in decision making. Thus, the real-time traffic condition could be included in route choice. For the hospital involvement, each patient's destination is precisely determined by pre-hospital screening equipment, and the nearest qualified hospital will be assigned to the patient. In two cases, we assume that the proper type of ambulance will be assigned to serve the patients. The route plans are illustrated in Fig. 6.

Fig.6. Ambulance routing under different strategies. (a) Route generation with inactive hospitals. (b) Route generation without traffic operators.

When the hospitals are excluded in route planning, EMS will assign patients to the nearest hospital regardless of its expertise, as shown in Fig. 5(a). If the assigned hospital is a comprehensive one with diverse expertise, there will be little difference between active and inactive hospital involvement. But suppose this hospital is not qualified for providing a specific treatment. In that case, it takes several minutes (e.g., 5 mins in this case) to figure out that the patient should be transferred and costs more than 25 minutes for the additional transport.

Fig. 7 compares the breakdown cost in each case. The scenario where hospitals are excluded expenses the most. Followed by the cases with inactive traffic operators, which is three times the baseline cost. Since we assume the proper ambulance allocation, the dispatch fee remains the same for three cases. The largest difference drive from the arrival delay due to the high penalty cost. The on-site delay under the three strategies are 0, 0, and 1 minute respectively, and the hospital arrival delay is 4, 21, and 17 minutes respectively. A few minutes of arrival delay has a tremendous difference in the treatment effect that determines life or death for the wounded. Therefore, the comparative analysis of the three cases fully illustrates the huge advantages of collaboration among EMSP, the hospital, and the traffic operator in planning and dispatching.

Fig. 7 breakdown cost under different strategies

If the traffic operators are excluded in route planning, empirical-based travel time will be adapted for route choice. As pre-hospital screening is considered, the destination for the ambulance is determined. Intuitively, there are two differences from the baseline. First, the ambulance cannot travel at the pre-defined speed for the entire journey because the operators do not coordinate the lane pre-clearing. If some road sections are congested, the travel time will accordingly increase. Second, the exact shortest path may not be assigned to the ambulance based on empirical data, as shown in Fig.5(b). The dynamic traffic conditions will largely influence the travel time.

5. Conclusion

In this paper, the ambulance routing problem with multi-stakeholder cooperation is investigated. Traffic operators respond to provide traffic management strategies to pre-clear the lane and prioritize the incoming ambulance. Hospitals play roles in remote screening to ensure the hospital assigned to the specific patient has sufficient expertise. We have proposed an MIP model to minimize the cost of traveling, ambulance allocation, and delay penalty. The proposed planning method can be used to support EMS decisions regarding the professions of hospitals and ambulance types, and fleet size. The exact solution of the optimization model takes only seconds that enables coordination among stakeholders within a short response time. Despite daily EMS, the scenarios of multiple accidents occurring at the same time can be solved based on the customized model by extending the dimension of the pre-hospital screening parameter. Therefore, the proposed model is sufficient to deal with different emergencies and provide practical rescue plans. Moreover, this method provides possibilities to dispatches the heterogeneous ambulance flee and matches the passengers with the most appropriate one to avoid scarce resource waste.

Future research may aim to incorporate further aspects such as stochastic service time and demand uncertainty (Legato and Mazza, 2020). This model could be further extended for large scale disaster response with considerations of patient priority and ambulance rerouting for picking up the injured with higher severity.

Acknowledgement

This study is supported by the Sino-Sweden bi-lateral project via National Key R&D Program of China (Project No. 2018YFE0102700) and Vinnova/FFI.

Data Availability Statement

The datasets analysed during the current study are available in the Gitbub repository, https://github.com/ZengZiling/Emergency-vehicle-routing-in-urban-road-networks-with-multi-stakeholder-cooperation.

Reference

Acar, M. & Kaya, O. 2019. A healthcare network design model with mobile hospitals for disaster preparedness: A case study for Istanbul earthquake. *Transportation Research Part E: Logistics and Transportation Review,* 130**,** 273-292.

Andersson, T. & Värbrand, P. 2007. Decision support tools for ambulance dispatch and relocation. *Journal of the Operational Research Society,* 58**,** 195-201.

Bandara, D., Mayorga, M.E. & McLay, L.A. 2014. Priority dispatching strategies for EMS systems. *Journal of the Operational Research Society,* 65**,** 572-587.

Berkoune, D., Renaud, J., Rekik, M. & Ruiz, A. 2012. Transportation in disaster response operations. *Socio-economic Planning Sciences,* 46.

Caunhye, A.M., Zhang, Y., Li, M. & Nie, X. 2016. A location-routing model for prepositioning and distributing emergency supplies. *Transportation Research Part E: Logistics and Transportation Review,* 90, 161-176.

Dimitriou, L., Efthymiou, D. & Antoniou, C. 2018. Saving Lives through Faster Emergency Unit Response Times: Role of Accessibility and Environmental Factors. *Journal of Transportation Engineering, Part A: Systems,* 144, 04018053.

Erdemir, E.T., Batta, R., Rogerson, P.A., Blatt, A. & Flanigan, M. 2010. Joint ground and air emergency medical services coverage models: A greedy heuristic solution approach. *European Journal of Operational Research,* 207, 736-749.

Farahani, R.Z., Asgari, N., Heidari, N., Hosseininia, M. & Goh, M. 2012. Covering problems in facility location: A review. *Computers & Industrial Engineering,* 62, 368-407.

Fhager, A., Candefjord, S., Elam, M. & Persson, M. 2018. Microwave Diagnostics Ahead: Saving Time and the Lives of Trauma and Stroke Patients. *IEEE Microwave Magazine,* 19, 78-90.

Gao, K., Yang, Y., Sun, L. & Qu, X. 2020. Revealing psychological inertia in mode shift behaviorand its quantitative influences on commuting trips. *Transportation Research Part F: Traffic Psychology and Behaviour,* 71, 272-287.

Goldberg, J. & Paz, L. 1991. Locating emergency vehicle bases when service time depends on call location. *Transportation Science,* 25, 264-280.

Ibri, S., Nourelfath, M. & Drias, H. 2012. A multi-agent approach for integrated emergency vehicle dispatching and covering problem. *Engineering Applications of Artificial Intelligence,* 25, 554-565.

Iliopoulou, C., Konstantinidou, M.A., Kepaptsoglou, K.L. & Stathopoulos, A. 2020. ITS Technologies for Decision Making during Evacuation Operations: A Review. *Journal of Transportation Engineering Part A: Systems,* 146.

Jagtenberg, C.J. & Mason, A.J. 2020. Improving fairness in ambulance planning by time sharing. *European Journal of Operational Research,* 280, 1095-1107.

Jagtenberg, C.J., van den Berg, P.L. & van der Mei, R.D. 2017. Benchmarking online dispatch algorithms for Emergency Medical Services. *European Journal of Operational Research,* 258, 715-725.

Jotshi, A., Gong, Q. & Batta, R. 2009. Dispatching and routing of emergency vehicles in disaster mitigation using data fusion. *Socio-Economic Planning Sciences,* 43, 1-24.

Jung, S. & Qin, X. 2020. Connecting motor vehicle crashes with emergency medical services performance: Spatial assessment for the korean freeway system. *Journal of Transportation Engineering Part A: Systems,* 146.

Knight, V.A., Harper, P.R. & Smith, L. 2012. Ambulance allocation for maximal survival with heterogeneous outcome measures. *Omega,* 40, 918-926.

Kuisma, M., Holmström, P., Repo, J., Määttä, T., Nousila-Wiik, M. & Boyd, J. 2004. Prehospital mortality in an EMS system using medical priority dispatching: a community based cohort study. *Resuscitation,* 61, 297-302.

Legato, P., & Mazza, R. M., 2020. Queueing analysis for operations modeling in port logistics. *Maritime Business Review*, 5(1), 67-83.

Lam, S.S.W., Zhang, J., Zhang, Z.C., Oh, H.C., Overton, J., Ng, Y.Y. & Ong, M.E.H. 2015. Dynamic ambulance reallocation for the reduction of ambulance response times using system status management. *The American Journal of Emergency Medicine,* 33, 159-166.

Lee, E.K., Pietz, F., Benecke, B., Mason, J. & Burel, G. 2013. Advancing public health and medical preparedness with operations research. *Interfaces,* 43, 79-98.

Li, X., Medal, H. & Qu, X. 2019. Connected infrastructure location design under additive service utilities. *Transportation Research Part B: Methodological,* 120, 99-124.

McLay, L.A. & Mayorga, M.E. 2013. A model for optimally dispatching ambulances to emergency calls with classification errors in patient priorities. *IIE Transactions,* 45, 1-24.

Memari, P., Tavakkoli-Moghaddam, R., Navazi, F. & Jolai, F. 2020. Air and ground ambulance location-allocation-routing problem for designing a temporary emergency management system after a disaster. *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine,* 234, 812-828.

Oran, A., Tan, K.C., Ooi, B.H., Sim, M. & Jaillet, P. Location and Routing Models for Emergency Response Plans with Priorities. *In:* ASCHENBRUCK, N., MARTINI, P., MEIER, M. & TöLLE, J., eds. Future Security, 2012// 2012 Berlin, Heidelberg. Springer Berlin Heidelberg, 129-140.

Persson, M., Fhager, A., Trefná, H.D., Yu, Y., McKelvey, T., Pegenius, G., Karlsson, J. & Elam, M. 2014. Microwave-Based Stroke Diagnosis Making Global Prehospital Thrombolytic Treatment Possible. *IEEE Transactions on Biomedical Engineering,* 61, 2806-2817.

Salum, J.H., Sando, T., Alluri, P. & Kitali, A. 2020. Impact of Freeway Service Patrols on Incident Clearance Duration: Case Study of Florida's Road Rangers. *Journal of Transportation Engineering Part A: Systems,* 146.

Sánchez-Mangas, R., García-Ferrrer, A., de Juan, A. & Arroyo, A.M. 2010. The probability of death in road traffic accidents. How important is a quick medical response? *Accident Analysis & Prevention,* 42, 1048-1056.

Schmid, V. 2012. Solving the dynamic ambulance relocation and dispatching problem using approximate dynamic programming. *European Journal of Operational Research,* 219, 611-621.

Schmid, V. & Doerner, K.F. 2010. Ambulance location and relocation problems with time-dependent travel times. *European Journal of Operational Research,* 207, 1293-1303.

Talarico, L., Meisel, F. & Sörensen, K. 2015. Ambulance routing for disaster response with patient groups. *Computers & Operations Research,* 56, 120-133.

Tikani, H. & Setak, M. 2019. Ambulance routing in disaster response scenario considering different types of ambulances and semi soft time windows. *Journal of Industrial and Systems Engineering,* 12, 95-128.

Toro-Díaz, H., Mayorga, M.E., Chanta, S. & McLay, L.A. 2013. Joint location and dispatching decisions for Emergency Medical Services. *Computers & Industrial Engineering,* 64, 917-928.

Wang, Y., Luangkesorn, K.L. & Shuman, L. 2012. Modeling emergency medical response to a mass casualty incident using agent based simulation. *Socio-Economic Planning Sciences,* 46, 281-290.

Wu, J., Kulcsár, B., Ahn, S. & Qu, X. 2020. Emergency vehicle lane pre-clearing: From microscopic cooperation to routing decision making. *Transportation Research Part B: Methodological,* 141, 223-239.

Zidi, I., Al-Omani, M. & Aldhafeeri, K. 2019. A New Approach Based On the Hybridization of Simulated Annealing Algorithm and Tabu Search to Solve the Static Ambulance Routing Problem. *Procedia Computer Science,* 159, 1216-1228.

Tables

Table 1. Notations used for the ambulance routing representations

Symbol	Notation
Set	
N	Set of all nodes
H	Set of hospital nodes $H \subseteq N$
P	Set of patient nodes $P \subseteq N$
K	Set of the ambulances
L	Set of injury levels
Index	
h, h'	Indices of the hospitals that dispatch ambulances or give treatment
i, j	Indices of nodes
p	Index of patient nodes
k	Index of ambulances
l	Index of injury level
Parameter	
τ^k	Ambulance allocation cost
o_h^k	Binary parameter indicates whether vehicle k is located at hospital h
$n_l^{h'}$	The number of patients at injury level l should be treated by hospital h'
c_l^k	The capacity of ambulance k when serving patients at injury level l
e_i	Service time at node i
t_{ij}	The simulated travel time between nodes i and j
T_i	The preferred arrival time at node i
r	Communication range between ambulance and the proceeding CVs
D_{ij}	Distance between nodes i and j
σ, σ_{ij}	Traffic density on a road segment, or between two nodes i and j
v_0	Ambulance velocity used in simulation planned for real-time cases
v	Real-time ambulance velocity
a, b, m, n	Linear regression parameters for communication range or traffic density
ψ_1, ψ_2	Weights of late arrival penalty on-site or at hospitals.
Variable	
f_k	The penalty cost of late arrival
ε_i	The buffer time at node i, its value can be negative or positive
s_i^k	Arrival time at node i of emergency vehicle k
σ_k, θ_k	Delay of vehicle k on-site or at hospital
Decision Variable	
x_{ij}^k	= 1, if ambulance k traveling from node i to j; = 0, otherwise

Table 2. Available ambulance inventory

Node ID	Ambulance fleet		
	Type I	Type II	Type III
2	0	0	1
3	1	1	1
7	0	1	1
27	1	1	1
28	0	1	1
29	3	4	3

Table 3. Pre-hospital screening

Patient list	Patient number	Accident scene	Level	Hospital
1	2	11	1	3
2	1	11	2	27
3	2	11	3	7
4	1	11	4	29