Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Modular vehicle technology for emergency medical services

Gaby Joe Hannoun*, Mónica Menéndez

Division of Engineering, New York University Abu Dhabi, Saadiyat Island, Abu Dhabi, United Arab Emirates

**Abstract**

While advancements in vehicular and wireless communication technologies are shaping the future of our transportation system, emergency medical services (EMS) are not receiving enough research attention. Their operations are still plagued by response delays that can often be life-threatening. Dispatching and redeployment systems identify the best practices regarding the allocation of the resources to emergencies and stations. Yet, the existing systems are unfortunately insufficient, and there is a growing need to embrace new technological solutions. This research introduces a smart system for EMS by leveraging the modular vehicle technology initially developed for transit systems. The proposed system relies on the design of vehicular modules that can couple and decouple to transfer patients from one module to another during transport. A fleet of medical transport vehicles is deployed to cooperate with the life support vehicles by providing, for example, transport and hospital admission tasks, thus allowing life support vehicles to answer pending emergency calls earlier. This is especially useful when there is a large demand for EMS (e.g. under the COVID-19 pandemic or other disasters such as the recent explosion in Beirut). This paper introduces a mathematical programming model to determine the optimal assignment decisions in a deterministic setting. This work is a proof of concept that demonstrates the applicability of the modular vehicle technology to EMS, evaluating the upper bound EMS performance that can be ultimately reached. A sensitivity analysis is conducted to provide insights and recommendations that are useful when selecting the weighting coefficients for the optimization function, to ensure a more efficient implementation of the modular vehicle technology for EMS. Also, the results of a comparative analysis show that the proposed system can adapt and offer larger benefits, in terms of response times and times to hospital, as demand increases and/or resources become more limited.

1. Background and motivation

Nowadays, our world is experiencing unprecedented events that make us question the efficiency of existing emergency medical services (EMS). As a matter of fact, deteriorated EMS operations became prevalent in many cities around the world during the COVID-19 pandemic. According to a recent Bloomberg article (Mosendz, 2020), huge peaks in emergency requests were observed during the COVID-19 pandemic with the average daily 911 medical calls reaching 6,500 in New York City and leading to delays of up to 4 h per emergency. Another disastrous event that has happened recently and that has shocked the world by its massive impact radius is Beirut’s port explosion on August 4th, 2020. This blast caused at least 190 deaths, 6,500 injuries and left 300,000 people homeless (Ghazi Balkiz and Mackintosh, 2020). It all happened in just a few seconds, hence overwhelming the emergency response system in the country. For instance, the Lebanese Red Cross, which is the major ambulance services provider in Lebanon, asserted that every available ambulance in the whole country was dispatched to Beirut (Peltier, 2020). As hospitals in Beirut were
severely affected, response operations also consisted of discharging injured people to hospitals outside the city, which led to longer travel times, and hence tremendous delays up to a few hours.

Even during normal conditions, EMS operations are still plagued by excessive response times that do not meet the benchmarks (Cui et al., 2019). According to the EMS Act of 1973, 95% of requests should be reached within 10 min in urban areas compared to 30 min for the ones in rural areas (Ball and Lin, 1993). The National Fire Protection Association 1710 (NFPA, 2020) suggests a target of 8 min, to be achieved for at least 90% of the emergencies. Bailey and Sweeney (2003) claimed that response time goals should be locally determined while considering characteristics of the given system. Nevertheless, the impacts of response time on survival rate will always remain the same regardless of all external factors. Faster response is associated with a lower level of mortality. For instance, Sánchez-Mangas et al. (2010) suggested that a 10 min reduction in response time is associated with a one-third decrease in probability of death for road traffic accidents. Despite such compelling numbers, EMS operations are still plagued by unsatisfactory response times. For instance, the likelihood of meeting the NFPA’s 8-min response benchmark is 65% for EMS care of chest pain in the US (Cui et al., 2019). According to NHTSA (2010), contemporary EMS should be created for stabilization and transportation to hospitals where more advanced interventions and investigations are available. Also, Manka et al. (2010) highlights that unnecessary and inefficient utilization of EMS may pose a great risk on waiting emergencies. Hence, a technology that would free an EMS vehicle earlier (i.e. after providing all possible prehospital care) would be beneficial.

Decision making in EMS planning and control can be categorized into three levels: (a) strategical, (b) tactical and (c) operational (Hulshof et al., 2012; Wiesche, 2017). The first level consists of defining facility locations and resource dimensioning (e.g. Brotcorne et al., 2003; Li et al., 2011; Chong et al., 2016) while the second level considers determining vehicle allocation to stations (deployment) and staff shift scheduling (e.g. Krishnan et al., 2016; Rajagopalan et al., 2011). These two levels rely on covering models, simulation and optimization techniques. Regarding the third operational level, it comprises research about redeployment and dispatching. The redeployment refers to the relocation (or redistribution) of available vehicles to other stations (1) to maintain good coverage as the time-of-day changes (multiperiod, e.g. Schmid and Doerner, 2010) or (2) to compensate for vehicles that became busy responding to emergencies (dynamic, e.g. Maxwell et al., 2010). Studies in this area focus on developing look-up tables or exact/approximate approaches to solve real-time optimization problems. Dispatching consists of determining the assignment of vehicles to emergencies. And it is usually common for the assignment of vehicles to emergencies usually follows a rule-based strategy such as dispatching the closest vehicle (Aboueljinane et al., 2013). Each of these levels has already been extensively explored by researchers and the desired EMS performance is yet to be achieved. Hence, we strongly believe that the potential for EMS operations’ improvement relies within unexplored vehicular technology and information and communication systems, which can drastically change the way the overall EMS system works.

In this paper, a new vehicular technology developed by the Next Future Transportation system (Next Future Transportation Inc., 2018a) is explored. It is an advanced robotic technology that allows smart vehicular modules to seamlessly couple and decouple among each other en-route and on regular road surfaces. This modular vehicle technology immediately grasped the attention of researchers focusing on transit and ride sharing systems. For instance, it led to the emergence of a new transport system with a variable capacity, that can be customized to fit the temporally and spatially varying travel demand. When compared to the traditional public transportation systems (i.e. buses, taxis), this novel modular transport system has the potential of reducing passenger delays and improving vehicle occupancy at high and low demand levels, respectively. Subsequently, the emerging studies (Guo et al., 2018; Gechelin and Webb, 2019; Chen et al., 2019, 2020; Zhang et al., 2020; Dakic et al., 2021; Dai et al., 2020; Shi et al., 2020; Caros and Chow, 2021; Chen and Li, 2021; Shi and Li, 2021; Gong et al., 2021; Pei et al., 2021; Saeed et al., 2022) investigating the use of modular vehicle technology mostly focus on designing these new modular transit systems with variable capacity to ultimately achieve optimal operations. Nevertheless, the modular vehicle technology is largely versatile with promising benefits to services-in-motion and logistic applications (such as door-to-flight direct services, last mile parcel delivery, etc. Next Future Transportation Inc., 2018a,b). In this paper, the capabilities of the modular vehicle technology are reinterpreted to highlight the possible benefits that it can offer to EMS operations. The proposed system relies on the design of different modular emergency vehicles that can couple and decouple to transfer patients and/or personnel from a module to another during transport. Some modules can provide life support (basic or advanced treatment) as well as medical transport while others are only capable of fulfilling the latter. The medical transport (MT) type is a new fleet type that will cooperate with the life support (LS) vehicles to take over transport and hospital admission tasks. Hence, life support vehicles can respond to waiting emergencies to achieve an overall faster response time and arrival to hospital. Such transport-only vehicles can also be borrowed from other transport fleets (e.g. transit systems) during extreme emergency scenarios. They may initiate transport and move patients closer to the life support vehicles to achieve earlier response times and arrivals to hospitals. The contributions of this paper are threefold. First, we design and introduce two new response operations for EMS that become possible due to en-route patient transfers. Second, as identifying the best practices to adopt for efficiently using this technology is not a trivial task, the deployment of the modular vehicle technology for EMS will eventually involve the use of optimization. So, we develop a deterministic combined vehicle routing and scheduling optimization problem assuming full information availability to obtain the optimal dispatching of the EMS fleet if the modular vehicle technology is deployed. Third, we present a proof of concept to demonstrate the applicability and potential benefits of the modular vehicle technology to EMS, in terms of response time and arrival time to hospitals.

In the following section, the proposed system is introduced along with the new types of response operations. Then, the problem statement is described in Section 3 while the optimization model, developed to identify the optimal EMS vehicles’ routing and service scheduling decisions if the modular vehicle technology is deployed, is discussed thoroughly in Section 4. The experimental plan is presented in Section 5, and includes (1) a sensitivity analysis of the weighting coefficients in the objective function and (2) a comparative analysis under varying demand and supply conditions, whose results are discussed in Sections 6 and 7 respectively. Finally, the key findings and potential directions for future studies are presented in Section 8.
2. Proposed system

In this section, the general idea behind the use of the modular vehicle technology for EMS is explored. As discussed previously, two types of vehicular modules are considered: life support vehicle and medical transport vehicle. Typically, upon receiving an emergency call, a life support vehicle is dispatched to the emergency site where field care may be provided. Then, the same unit transports the patient to the hospital (see Operation A, in Fig. 1). In addition to this regular type of response operation, two new types (Operations B and C, in Fig. 1) are possible with the modular vehicle technology which allows in-motion patient transfer from/to a new medical transport vehicle type. Fig. 1 illustrates the three possible operations on a sample grid transportation network where all links have the same travel time equal to 1 unit time. Also, it compares the timelines of these operations to highlight the potential time savings. Note that emergencies that only require transportation are not considered as they can be served by passenger cars (i.e. private, taxis, ride-hailing services, etc.).

Operation B, in Fig. 1, consists of dispatching a life support vehicle to the emergency site. This unit will provide field intervention to the patient and initiate the transport to a transfer point where it will couple with a medical transport vehicle. The latter will transfer in the patient, decouple from the life support vehicle, and resume the transport to the hospital where it will carry out hospital admission tasks. This new operation is expected to be beneficial to the overall EMS performance as the life support vehicle is freed at an earlier point in time, thus allowing it to respond to other waiting emergency calls, if any. Otherwise, the life support vehicle can move back to a station where it will be ready to serve future calls. We acknowledge that this type of operation is not possible for all types of injuries. For instance, severely injured patients cannot be transferred to a medical transport vehicle due to limitations such as equipment attachments. Nevertheless, as transfer is performed in-motion and as no additional cost is incurred in terms of time, transfer can be arranged so that personnel is transferred out of the life support vehicle when the presence of the full team is not required in this emergency yet it is beneficial in other waiting or future emergencies. The transfer of personnel is part of future work and the scope of this paper is limited to the transfer of patients.

Operation C, in Fig. 1, consists of initiating the response with the medical transport vehicle that will pick up the patient from the emergency site then move to a transfer point where it will couple with a life support vehicle. The latter will perform in-unit intervention while moving towards the hospital where the patient will be checked in. This type of operation only applies to emergencies that allow the patients to be transported into the medical transport vehicle first. Yet, when feasible, this operation may result in earlier intervention and arrival to the hospital as the medical transport vehicle is bringing the patient closer to the life support vehicle as well as the hospital. When comparing the timelines of Operation A and Operation C in Fig. 1, the start of care and arrival to the hospital are occurring earlier in Operation C (with $\Delta t_1$ and $\Delta t_2$ time savings respectively) and this is due to the cooperation, among the medical transport vehicle and the life support vehicle, made possible through the modular vehicle technology. Note that, in Operations A and B, care starts upon arrival of the life support vehicle at the emergency scene while in Operation C, it starts at the end of the transfer.

As discussed previously, the modular vehicle technology is not limited to the transfer of patients. It also allows personnel’s transfer leading to improved resource management especially during staff shortages as personnel who completed their intervention...
on a given patient can be assigned to another emergency. This is not included in the scope of this paper and will be investigated in future works along with the other capabilities such as (1) the dispatching of multiple vehicular modules to the same emergency where there are several injured persons and this could be useful under disaster scenarios when resources are limited and patients may need different specialized care areas, and (2) the en-route intervention of a life support vehicle during transport due to unexpected complications.

3. Problem statement

First, the problem consists of serving a set of emergencies that arrived during a given time period. Each request has a specific origin, destination (i.e. hospital) and an arrival time. Second, the problem is characterized by a fleet of EMS modular vehicles and as previously discussed, we will consider only two types of EMS modules: LS type and MT type. Third, the vehicles operate, in a transportation network, between stations, emergency request’s origins, and hospitals. With the introduction of the new operations B and C, additional nodes where potential en-route patient transfers can initiate are considered.

The investigated problem is a combined vehicle routing and scheduling problem with (1) time windows as vehicles cannot be dispatched to an emergency before receiving the corresponding call, (2) temporal precedence as some nodes should be visited in a fixed order, (3) temporal synchronization as cooperating vehicles should arrive to a transfer point at the same time, and (4) an objective function that minimizes EMS performance measures such as response times and arrival times to hospitals. The type of problem being solved is also called a pickup and delivery problem (PDP) where each request consists of moving a passenger from an origin to a destination. It is specifically a variant of the Dial-A-Ride Problem (DARP) as it deals with the transportation of people between paired pickup and delivery locations (Parragh et al., 2008). A generalization of the PDP with transfers (Cortés et al., 2010) exists where vehicles can interact at specific locations in the transportation network to exchange passengers. However, these fixed transfer locations actually act as waiting locations where passengers wait a given amount of time for another vehicle to arrive and resume their trip to their respective destinations. In the context of emergency response operations, we mostly focus on the en-route transfer characteristic of the modular vehicular technology and on its capability to bring a patient from one emergency vehicle to another while moving, without needing to stop any vehicle nor incur any passenger waiting time and trip interruption. Hence, this en-route transfer will imply strict temporal and spatial synchronizations of the two vehicles that are transferring a patient. It is the optimization process itself that specifies the location and time of a transfer, if contributing to the improvement of the overall EMS performance. Besides, the execution of only specific EMS response operations (A, B and C as shown in Fig. 1) and the consideration of EMS-related performance measures to explicitly account for the user inconvenience, add to the uniqueness of the investigated problem. Depending on the operation type and the EMS vehicle(s) selected by the optimization process to serve an emergency request, the response time and time to hospital are computed accordingly. It hence motivated us to develop a novel optimization model, inspired by Cortés et al. (2010), that is tailored to our EMS system with modular vehicle technology and strict en-route transfers, in order to demonstrate that this vehicular innovation can indeed bring potential benefits to EMS, making it worth investigating in larger problems replicating the real world.

The mathematical model, described in Section 4, identifies the optimal EMS vehicles’ routing and service scheduling decisions for a given and known set of emergencies and static travel times. Hence, it is a deterministic model that allows us to evaluate and analyze in detail our proof of concept. Since the emergency demand is unknown in advance, it is impossible to rely on an a-priori plan to conduct EMS operations and real-time techniques that can adapt to the arrival of new emergency requests are essential. That being said, this deterministic approach allows us to first assess the benefits that can be offered by the modular vehicle technology to EMS, by isolating the drawbacks that may arise with an improper real-time implementation due to the challenges of the uncertain and dynamic aspects of the real world. Then, if benefits are achievable, it will encourage the development of a more advanced and efficient technique that is suitable for the real-time deployment of the novel technology for EMS. Hence, this deterministic approach is a valuable first building block in the design of new EMS systems with modular vehicle technology, but is not intended to be used in real-time for EMS.

This work assumes that all emergencies are the same, that they all require the same field care and admission durations and that all patients can be transferred from one unit to another. Classifying emergencies into categories with different characteristics and needs will be done in future works. Also note, that this work does not consider mechanical failures linked to the coupling and decoupling of the vehicular modules.

4. Mathematical model

In this section, the mathematical model is thoroughly discussed. Section 4.1 explains how the optimization space represented by a graph is defined. In Section 4.2, the decision variables and parameters in the model are described, while Section 4.3 explains how the travel times are identified in a way to account for the en-route transfers. Finally, Section 4.4 presents the objective and constraints of the mathematical model.
4.1. Graph representation

The optimization problem’s space is defined by transforming the road network into a graph \( G = (N, A) \), shown in Fig. 2, where \( N \) is the set of nodes, each referring to a physical location in the road network, and \( A \) is the set of arcs connecting pairs of nodes.

A main depot node 0, which has no physical location in the road network, is added to the graph. The model will generate a single trip for each vehicle during the planning horizon and this trip should start and end at this node. A station node \( s \in S \) is considered as an intermediate node where a vehicle can wait between response operations during the planning horizon. The number of times each vehicle enters a given station cannot be anticipated as the work assignment of a vehicle is not known in advance and is defined by the optimization process. Due to the scheduling variables which are not vehicle specific for stations (described further in Section 4.2), a given station node cannot be visited by more than one vehicle. Subsequently, duplicate nodes are added in the graph for each physical location where a station is located in the road network, yet not all of these nodes are necessarily visited. Depending on the number of emergencies and number/types of EMS vehicles in the system, the maximum number of EMS vehicle trips is determined, hence identifying the maximum number of duplicate nodes per physical station that could be ultimately visited and that does not affect the optimization process nor reduce the feasible region of the optimization problem. For each emergency \( e \in E \), two nodes are added in the graph, one referring to the physical location at which this emergency originates \((e^- \in E^-)\) and the other referring to the physical location of the hospital \((e^+ \in E^+)\) to which this emergency will be transported. A transfer point is a physical location at which a transfer can start. With the introduction of en-route transfer, it can theoretically be any point in the road network. However, to limit the size of the investigated problem, the road intersections are selected as the candidate transfer locations. We assume that these are the only link entry points and thus the earliest and most beneficial locations to initiate transfers. A given transfer point’s physical location in the road network can be visited by vehicles serving different emergencies. So, to ensure proper coordination of vehicles, each transfer point node \( t \) in the graph should be associated with an emergency \( e \) so that a vehicle assigned to a transfer point node \( t \) knows which emergency \( e \) it is going to serve at \( t \) and which other vehicle it is going to coordinate with. This is why, a set of transfer point nodes \( T^e \) is added in the graph for each emergency \( e \). At each transfer point’s physical location in the road, a transfer point node is added to the graph for each emergency \( e \), resulting in overlapping transfer point nodes in Fig. 2. Their symbols are color coded to indicate their respective emergency. To limit the problem size and for practicality, each set of transfer point nodes \( T^e \) is filtered in a way to exclude the transfer point nodes that are obviously not beneficial to the emergency \( e \). For instance, the transfer point nodes that imply a large deviation from the shortest path from \( e^- \) to \( e^+ \) and the ones that are very close to \( e^- \) or \( e^+ \) in terms of travel time are removed as they are considered disadvantageous and unpractical, respectively.

The set of directed arcs \( A \) represents the possible movements between nodes. Not all pairs of nodes are connected. For example, no movement is defined from a station node \( s \) to a hospital node \( e^- \) or from a hospital node \( e^- \) to another hospital node \( e^- \). Also, a vehicle at an emergency origin node \( e^- \) can only go to its corresponding hospital node \( e^- \) or to a transfer point node \( t \in T^e \). A graph with a set of arcs for the possible movements only results in fewer variables and constraints compared to a complete graph (where an arc is defined for every pair of nodes).

Possible movements are as follows:

- From the depot 0 to:
  - a station \( s \in S \)
- From a station \( s \in S \) to:
  - a station \( s' \in S \) where \( s \neq s' \)
  - an emergency origin \( e^- \in E^- \)
  - a transfer point \( t \in T^e \)
  - the depot 0

![Fig. 2. Graph representation.](image-url)
• From an emergency origin \( e^+ \in E^+ \) to:
  - its corresponding hospital \( e^- \in E^- \)
  - a transfer point \( t \in T^e \)

• From a hospital node \( e^- \in E^- \) to:
  - a station \( s \in S \)
  - an emergency origin \( e^{+e} \in E^+ \), where \( e \neq e' \)
  - a transfer point \( t \in T^e \), where \( e \neq e' \)

• From a transfer point \( t \in T^e \) to:
  - a station \( s \in S \)
  - its corresponding hospital \( e^- \in E^- \)
  - an emergency origin \( e^{+e} \in E^+ \) where \( e \neq e' \)
  - a transfer point \( t \in T^e \) where \( e \neq e' \)

If a movement from \( i \) to \( j \) is allowed, an arc is defined for this pair of nodes and it represents the travel time (\( \tau_{ij} \)) of the shortest path from \( i \) to \( j \) in the original road network. Hence, knowledge of link travel times and the execution of a shortest-path algorithm from all to all nodes is required prior to solving the mathematical formulation. Note that travel times from/to the depot (\( \tau_{0,i} \) and \( \tau_{i,0} \)) are zero.

4.2. Variables and parameters

Fig. 3(a) shows the work assignment of a vehicle \((k, v)\) (i.e. vehicle \( v \) of type \( k \)) that is represented by a sequence of visited nodes \((i, t, j, a, g)\). The routing variable \((x_{ij}^{ka})\) is equal to 1 if vehicle \((k, v)\) moves from node \( i \) to node \( j \) and 0 otherwise. Two scheduling variables are used in the model \((\chi_{ij})\) and \((\psi_{ij}^{+e})\). The former refers to the time at which node \( i \) is reached by a vehicle and it is defined for nodes \( i \in N \setminus T \). This variable is not considered as vehicle specific to limit the number of variables and problem size. For instance, the stations, emergency origins and hospital nodes \((i \in N \setminus T)\) can only be visited by a maximum of one vehicle whose time of arrival at node \( i \) will determine the time at which \( i \) is reached. However, if a transfer is found beneficial at a given transfer point node \( t \), its latter will be visited by two vehicles, and their times of arrival at \( t \) must coincide. Hence, a scheduling variable \((\psi_{ij}^{+e})\) is defined for all \( t \in T \) and it refers to the time of arrival of vehicle \((k, v)\) at the transfer point node \( t \). To ensure a proper en-route transfer, the times of arrival of the two vehicles performing the transfer at \( t \) will have to be equal and this is ensured through a constraint in the model (Eq. (29), discussed in Section 4.4).

The service time \((\zeta_i)\) is the time spent at node \( i \) and depends on the node’s type. For instance, \( \zeta^{+e} \) at an emergency origin node is the time spent on field care, \( \zeta_i \) at a transfer point node \( i \) is the transfer duration \((\tau_i)\), and \( \zeta^- \) at a hospital node is the admission time, and \( \zeta_i \) at a node \( s \) is a minimum duration needed for equipment reload and staff changes (it may be zero).

Vehicles may wait at a specific node \( i \) up to a time limit of \( w_i \) which also varies with the node type. Vehicles are not allowed to wait for a new assignment at emergency origin nodes \( e^+ \) and transfer point nodes \( t \) \((w_{e^+} = w_t = 0)\). A vehicle that picked up a patient \( e \) should immediately go to the hospital node \( e^- \) or to a transfer point node \( t \in T^e \). Similarly, a vehicle that transferred in a patient at a transfer point node \( t \) should immediately go to the hospital node \( e^- \), while a vehicle that transferred out a patient cannot wait at the transfer point as this will require it to stop potentially in the middle of the street (recall that transfers are performed in motion). Vehicles are allowed to wait at hospitals \( e^- \) up to a certain limit \((w_{e^-} = W)\). They are also allowed to wait during the whole planning horizon at stations \((w_s = M \text{ where } M \text{ takes a large value})\).

A segment of the work assignment of a vehicle \((k, v)\), performing an Operation A which consists of (1) moving to \( e^+ \) and performing field care and (2) transporting and discharging patient \( e \) to the hospital at \( e^- \), is represented in Fig. 3(b). The response time \((R_e)\) of emergency \( e \) is the duration elapsed between the call time of emergency \( e \) \((l_e)\) and the start of care. The latter occurs upon arrival of vehicle \((k, v)\) to \( e^+ \), as this vehicle is of type \( k \in K^e \) where \( K^e \) is the life support vehicle type. The time to hospital \((TH_e)\) of emergency \( e \) is the time elapsed between the start of care and the hospital arrival time of patient \( e \) to \( e^- \). In other words, the sum of \( R_e \) and \( TH_e \), also called prehospital time \((PH_e)\), is the total time between the emergency call \((l_e)\) and the hospital arrival time \((\zeta^-)\) which is the admission time at the hospital is not included in \((TH_e)\).

In Fig. 3(c), the work assignment of two vehicles \((k, v)\) and \((k', v')\) is illustrated \((k \in K^e \text{ and } k' \notin K^e)\). These two vehicles are cooperating to service patient \( e \) by following an Operation B. Vehicle \((k, v)\) is picking up patient \( e \) first, hence the response time \((R_e)\) which started at \( l_e \) ends upon arrival of vehicle \((k, v)\) at \( e^+ \). Vehicle \((k, v)\) moves towards a transfer point \( t \) after which it is going to couple then decouple with vehicle \((k', v')\) for a duration of \( \zeta_i \). Then, vehicle \((k', v')\) resumes the transport of patient \( e \) to \( e^- \) and the time to hospital \((TH_e)\) starts upon its arrival at \( e^- \).

A typical Operation C is illustrated in Fig. 3(d) where two vehicles \((k, v)\) and \((k', v')\) cooperate to service patient \( e \), where \( k \notin K^e \text{ and } k' \in K^e \). The response time \((R_e)\) does not end up upon arrival of vehicle \((k, v)\) at \( e^+ \) as this is a medical transport vehicle. Hence, the response time ends when care starts and this is the end of the transfer through which patient \( e \) is transferred to vehicle \((k', v')\) that will provide treatment. The time to hospital \((TH_e)\) starts upon arrival of vehicle \((k', v')\) at the hospital node \( e^- \).

The upper and lower bounds of \( R_e \) and \( TH_e \) for all cases are set in Eqs. (23)–(24) and Eq. (25) respectively.
4.3. Travel times

Knowledge of link travel times and the execution of a shortest-path algorithm from all to all nodes is required prior to solving the mathematical formulation. For instance, the travel time \( \tau_{i,j} \) associated with a link from \( i \) to \( j \) in the graph is deduced from the shortest path's travel time \( S_{P_{i,j}} \) from the physical location of \( i \) to the one of \( j \) in the original road network, as shown in Algorithm 1.

One of the main features of the modular vehicle technology is that it allows en-route transfer of patients so vehicles are not required to stop on the side of the road. A transfer point node \( t \) refers to a physical location in the road network at which the transfer starts. During the transfer, both vehicles will be moving towards the hospital while coupled to each other. In other words, the vehicle \( (k', v') \) in Figs. 3(c) or 3(d) which is transferring in the patient \( e^- \) will be moving towards \( e^- \) while performing the transfer for a duration of \( \zeta_t \). At the end of the transfer, the vehicle will be closer to \( e^- \) and the remaining travel time to \( e^- \) is the travel time from \( t \) to \( e^- \) minus the transfer duration \( \zeta_t \). We could define new nodes in the graph at the end of each transfer but this will result in more variables in the mathematical program. Instead, the travel time from \( t \) to \( e^- \) \( \tau_{t,e^-} \) is simply reduced by \( \zeta_t \) to reflect the true travel time remaining to \( e^- \) at the end of the transfer. The vehicle \( (k, v) \) transferring out the patient in Figs. 3(c) or 3(d) will have to follow vehicle \( (k', v') \)'s movement and can only divert its route to a node \( j \in N \setminus e^- \) when the transfer ends, at a physical location \( (r^*) \) in the road network. To account for the in-motion transfer in the model, the travel time from a transfer point node \( t \)
to \(j\) where \(j \in N \setminus e^c\) should be adjusted in a way to reflect the travel time from the location \(t^*\) to \(j\). In fact, vehicle \((k, j)\) is forced to follow, for a duration of \(\zeta_j\), the shortest path of vehicle \((k', l')\) going from \(t\) to \(e^c\). Note that travel times from/to the depot \((\tau_{0,i}\) and \(\tau_{s,0})\) are zero.

**Algorithm 1: Travel times \(\tau_{i,j}\) identification**

**Input:** Shortest path’s travel times \((SP_{e,j})\), \(\forall i, j \in N\)

**Output:** Travel times \((\tau_{i,j})\), \(\forall i, j \in N\)

**Initialization:** \(\tau_{i,j} \leftarrow SP_{e,j}\), \(\forall i, j \in N\);  

**foreach** \(e \in E\) do  

**foreach** \(i \in T^*\) do  

\[\tau_{i,e^c} \leftarrow (\tau_{i,e^c} - \zeta_i);\]

\(t^*\) \(\leftarrow\) Location reached at the end of the transfer which started at \(t\);  

**foreach** \(j \in N \setminus e^c\) do  

\[\tau_{i,j} \leftarrow SP_{e,j} \]

end

end

To summarize, this preprocessing step defines the travel times from all nodes to all nodes in the graph and accounts for the dynamic aspect of the transfer. Travel times from a transfer point node \(t \in T^*\) to hospital node \(e^c\) are adjusted by just decreasing them by \(\zeta_i\). However, the travel time from \(t\) to \(j\) where \(j \in N \setminus e^c\) is replaced by the travel time from the actual physical location \(t^*\) reached at the end of the transfer to node \(j\). This actual location is along the shortest path from \(t\) to \(e^c\) and may be along a link and not necessarily at a given node.

In the following section, the mathematical formulation is presented along with the objective function and constraints. Table 1 describes the notation of all sets, parameters and variables.

---

**Table 1**

Sets, parameters and variables notation.

| Sets       | Description                                                                 |
|------------|-----------------------------------------------------------------------------|
| \(E\)      | Emergency (i.e. patient) indexed by \(e\)                                  |
| \(N\)      | All nodes indexed by \(i\) and \(j\)                                       |
| \(S\)      | Station nodes indexed by \(s\), \(S \subseteq N\)                         |
| \(E^+\)    | Emergency origin nodes \(E^+ = \{e^+: e \in E\}\), \(E^+ \subseteq N\)    |
| \(E^-\)    | Hospital nodes (i.e. emergencies destination nodes) \(E^- = \{e^-: e \in E\}\), \(E^- \subseteq N\) |
| \(T\)      | Transfer point nodes indexed by \(t\), \(T \subseteq N\)                  |
| \(T^*\)    | Transfer point nodes specific to emergency \(e\) indexed by \(t\), \(T^* \subseteq T\) |
| \(K\)      | Vehicle types indexed by \(k\)                                            |
| \(K^l\)    | Life support vehicle type, where \(K^l \subseteq K\)                      |
| \(V_i\)    | Vehicles of type \(k \in K\) indexed by \(v\)                            |

| Parameters | Description                                                                 |
|------------|-----------------------------------------------------------------------------|
| \(\tau_{i,j}\) | Travel time from node \(i\) to node \(j\), where \(i, j \in N\)           |
| \(\zeta_i\) | Service time at node \(i\), where \(i \in N\)                              |
| \(l_e\)    | Call time of emergency \(e\)                                               |
| \(\Gamma\) | Transfer duration                                                           |
| \(\tau_{\text{min}}\) | Minimum travel time from \(i\) to any \(e^c\) (excluding transfer duration) |
| \(\delta_{e,j}\) | Binary parameter equal to 1 if \(\tau_{e,j} \geq \tau_{\text{min}}\) and 0 otherwise |
| \(s_{e,v}^{i,v}\) | Binary variable equal to 1 if vehicle \((k, v)\) goes from \(i\) to \(j\) and 0 otherwise, where \(i \in N, j \in N, k \in K, v \in V_i\) |
| \(y_i\)    | Continuous variable equal to the time at which node \(i \in N \setminus T\) is reached. Node \(i\) can only be visited by a maximum of one vehicle. |
| \(s_{e,v}^{k,v}\) | Continuous variable equal to the time at which vehicle \((k, v)\) arrives to the transfer point node \(t \in T\), where \(k \in K, v \in V_i\) |
| \(R_s\)    | Continuous variable taking the value of the response time for patient \(e\), where \(e \in E\) |
| \(TH_s\)   | Continuous variable indicating the time to hospital (i.e. duration elapsed after the start of care until the arrival at hospital node \(e^c\)) for emergency \(e\), where \(e \in E\) |
| \(PH_s\)   | Continuous variable representing the total time between the emergency call \((l)\) and the hospital arrival time for emergency \(e\) (i.e. sum of \(R_s\) and \(TH_s\)), where \(e \in E\) |

---

[Reference to the table with notation of sets, parameters and variables]
4.4. Formulation

The objective function (1) minimizes the response times and the times to hospital using a weighted method. For instance, different weights ($a_1$ and $a_2$) are assigned to each component in Eq. (1), such that $a_1 + a_2 = 1$. The first component consists of minimizing the response times. Variable $R_e$ takes the value of the response time of emergency $e$ and is defined using Eqs. (23)–(24). The second component minimizes the time to hospital through $TH_e$, which refers to the time it takes for patient $e$ to arrive to its assigned hospital $e^-$, excluding its response time. Lower bounds of $TH_e$ are defined using Eq. (25).

The objective function increases linearly with the increase in response times and times to hospital. Assigning values to the weighting coefficients ($a_1$ and $a_2$) is subjective. Nevertheless, a sensitivity analysis is performed and thoroughly discussed in Section 6 to provide useful insights to decision maker(s).

\[
\text{Minimize } z = a_1 \sum_e R_e + a_2 \sum_e TH_e \tag{1}
\]

The constraints of this mathematical formulation are listed below.

Eq. (2) ensures that a vehicle does not initiate more than one route from the depot while Eq. (3) sets the base station of each vehicle $(k, v)$.

\[
\sum_{j \in N} x_{0k, j}^{k,v} \leq 1; \quad \forall k \in K, v \in V^k \tag{2}
\]

\[
x_{0k, 0}^{k,v} = 1; \quad \forall k \in K, v \in V^k \tag{3}
\]

Eq. (4) indicates that all vehicles return to the depot at the end of the planning horizon.

\[
\sum_{i \in N} x_{i,0}^{k,v} = 1; \quad \forall k \in K, v \in V^k \tag{4}
\]

Eq. (5) ensures that stations can be entered by a maximum of one vehicle to maintain consistency of the scheduling variables ($y_e$). For instance, duplicate nodes referring to the same physical location of a given station are added in the graph, as discussed earlier in Section 4.1.

\[
\sum_{k \in K} \sum_{v \in V^k} x_{i,s}^{k,v} \leq 1; \quad \forall s \in S \tag{5}
\]

Eq. (6) maintains the conservation of flow for each vehicle.

\[
\sum_{i \in N} x_{i,j}^{k,v} - \sum_{i \in N} x_{j,i}^{k,v} = 0; \quad \forall k \in K, v \in V^k, j \in N \tag{6}
\]

Eq. (7) ensures that one and only one vehicle is going to $e^+$ to pickup patient $e$, while Eq. (8) guarantees that exactly one vehicle transports him/her to the assigned hospital node $e^-$. 

\[
\sum_{k \in K, v \in V^k} x_{i,e^+}^{k,v} = 1; \quad \forall e \in E \tag{7}
\]

\[
\sum_{k \in K, v \in V^k} x_{i,e^-}^{k,v} = 1; \quad \forall e \in E \tag{8}
\]

Eq. (9) forces a life support vehicle ($k \in K^L$) to serve a given patient $e$ at one point during the response. In this paper, an operation fully provided by a medical transport vehicle is not allowed.

\[
\sum_{k \in K^L, v \in V^k} \sum_{i \in N} x_{i,e^+}^{k,v} + \sum_{k \in K^L, v \in V^k} \sum_{i \in N} x_{i,e^-}^{k,v} \geq 1; \quad \forall e \in E \tag{9}
\]

Eq. (10) ensures that a patient cannot be transferred to a life support vehicle ($k \in K^L$) if the time left to arrive to $e^-$ is below a minimum travel time ($\tau_{min}$). A parameter $\delta_{e,e^-}$ is equal to 1 if $\tau_{e,e^-} \geq \tau_{min}$ and 0 otherwise.

\[
\sum_{k \in K^L, v \in V^k} x_{e,e^-}^{k,v} \leq \delta_{e,e^-}; \quad \forall e \in E, t \in T^e \tag{10}
\]

Eqs. (11)–(18) set the lower and upper bounds for the scheduling variables (i.e. times of arrival at each node in the graph). If a vehicle $(k, v)$ goes from $i$ to $j$ ($\sum_i x_{i,j}^{k,v} = 1$) where, $i, j \in v \setminus T$, then the time of arrival ($y_e$) at node $j$ should be greater than or equal to the sum of time of arrival ($y_e$) at node $i$, the time spent at $j$ ($z_e$) and the travel time from $i$ to $j$ ($\tau_{i,j}$). On the other hand, the time of arrival at $j$ should not exceed its lower bound by more than a wait time limit $w_i$ that depends on the type of the node $i$. For instance, vehicles are not allowed to wait for a new assignment at transfer points and emergency origins ($w_{e^+} = w_{e^-} = 0$). They can wait up to a certain duration at hospitals ($w_e = W^e$) and indefinitely at stations ($w_s = M$). Note that a vehicle may stop receiving new tasks and may return to the depot as we are running this optimization model for a fixed planning horizon.

When transfer points are involved (i.e. vehicle going from and/or to a transfer point node $t$), separate constraints are defined for each vehicle as these nodes, if visited, have to be reached by two vehicles whose arrival times at $t$ must coincide. In fact, Eq. (29) ensures that the arrival times of these two vehicles are the same for a proper transfer. Eqs. (17)–(18) set an upper bound of zero for the scheduling variables of unvisited nodes. We acknowledge the fact that a more compact formulation would result if all scheduling...
variables were defined as vehicle specific (i.e. \( y^{k,v} \) \( \forall i \in N \)), however it will actually entail the addition of unnecessary scheduling variables at nodes \( (i \in N \setminus T) \) that can only be reached by a maximum of one vehicle, and it will hence lead to more variables and a larger problem size, reducing the computational efficiency.

\[
\sum_{k,l} x_{ij}^{k,v} = 1 \Rightarrow y_i + c_i + r_{ij} \leq y_j; \quad \forall i \in N \setminus T, j \in N \setminus \{T, 0\}
\]

(11)

\[
\sum_{k,l} x_{ij}^{k,v} = 1 \Rightarrow y_i + c_i + u_i + r_{ij} \geq y_j; \quad \forall i \in N \setminus T, j \in N \setminus \{T, 0\}
\]

(12)

\[
x_{ij}^{k,v} = 1 \Rightarrow y_i + c_i + r_{ij} = y_j; \quad \forall k \in K, v \in V^k, i \in N \setminus \{T, 0\}, j \in T
\]

(13)

\[
x_{ij}^{k,v} = 1 \Rightarrow y_i + c_i + r_{ij} \leq y_j; \quad \forall k \in K, v \in V^k, i \in N \setminus T, t \in T
\]

(14)

\[
x_{ij}^{k,v} = 1 \Rightarrow y_i + c_i + u_i + r_{ij} \geq y_j; \quad \forall k \in K, v \in V^k, i \in N \setminus T, t \in T
\]

(15)

\[
x_{ij}^{k,v} = 1 \Rightarrow y_i + c_i + r_{ij} = y_j; \quad \forall k \in K, v \in V^k, i \in N \setminus T, t \in T
\]

(16)

\[
y_j \leq M \sum_{k,l} x_{ij}^{k,v}; \quad \forall k \in K, v \in V^k, e \in E, t \in T^e
\]

(17)

\[
y_j \leq M \sum_{k,l} \sum_{i \in N \setminus T} x_{ij}^{k,v}; \quad \forall j \in N \setminus T
\]

(18)

Eq. (19) indicates that a vehicle \((k, v)\) may be requested to pick up patient \(e\) only after receiving the emergency call. In other words, the time of arrival at \(e^+\) cannot be smaller than the call time \(l_e\) plus the travel time from \(i\) to \(e^+\). Similarly, Eq. (20) ensures that the time of arrival at a transfer point node \(t\) to transfer in patient \(e\) cannot be less than the call time of \(e\) plus the travel time from \(i\) to \(e^+\) to \(t\).

\[
y_e^+ \geq l_e + \sum_{k,v} \sum_{i \in E} (x_{ij}^{k,v} \times r_{ij,e^+}); \quad \forall e \in E
\]

(19)

\[
x_{ij}^{k,v} = 1 \Rightarrow x_{ij}^{k,v} \geq l_e + r_{ij}; \quad \forall k \in K, v \in V^k, i \in N \setminus T, t \in T^e
\]

(20)

Eq. (21) forces a vehicle \((k, v)\) that transferred in the patient \(e\) to go immediately to its assigned hospital node \(e^+\). Note that a vehicle transfers in a patient if it goes directly from a node \(i \in \{N \setminus e^+\}\) to a transfer point node \(t\) (i.e. \(\sum_{i \in N \setminus e^+} x_{ij}^{k,v} = 1\)).

\[
x_{ij}^{k,v} = \sum_{i \in N \setminus e^+} x_{ij}^{k,v}; \quad \forall k \in K, v \in V^k, e \in E, t \in T^e
\]

(21)

Eq. (22) prohibits a vehicle \((k, v)\) that transferred out the patient \(e\) from going to its assigned hospital node \(e^+\). Note that a vehicle transfers out a patient if it goes directly from the emergency origin node \(e^+\) to a transfer point node \(t \in T^e\) (i.e. \(x_{e^+,t}^{k,v} = 1\)).

\[
x_{ij}^{k,v} \leq 1 - x_{e^+,t}^{k,v}; \quad \forall k \in K, v \in V^k, e \in E, t \in T^e
\]

(22)

Eqs. (23)–(24) define the response time \((R_e)\) of emergency \(e\). The response time is equal to the difference between the start of care time and the call time \((l_e)\). In Operations A and B, care starts when vehicle \((k, v)\) with \(k \in K^e\) arrives at \(e^+\) (see Figs. 3(b) and 3(c)). In that case, Eq. (23) sets the value of \(R_e\) while Eq. (24) is not binding. However, in Operation C, care starts when a vehicle \((k, v)\) with \(k \in K^e\) finishes transferring in patient \(e\) (see Fig. 3(d)), and only Eq. (24) is binding in this case. Note that upper bounds are set for \(R_e\) so that the value of \(R_e\) does not affect the value of \(TH_e\) defined through Eq. (25) and explained next. Without an upper bound, \(R_e\) may take a higher value to lower the one of \(TH_e\), especially when \(\alpha_2\) is greater than \(\alpha_1\) (i.e. when more weight is assigned to \(\sum_{e \in E} TH_e\) compared to \(\sum_{e \in E} R_e\)). This is discussed in Section 6.

\[
\sum_{k \in K \setminus e} \sum_{i \in e} x_{ij}^{k,v} = 1 \Rightarrow R_e = y_e^+ - l_e; \quad \forall e \in E
\]

(23)

\[
\sum_{k \in K \setminus e} \sum_{i \in e} \sum_{t \in T^e} x_{ij}^{k,v} = 1 \Rightarrow R_e = \sum_{k \in K \setminus e} \sum_{i \in e} \sum_{t \in T^e} y_i^+ + r_{ij} + l_e; \quad \forall e \in E
\]

(24)

Eq. (25) defines the time to the hospital \((TH_e)\) of each emergency \(e\). This is the hospital arrival time minus the response time \((R_e)\) minus the call time \((l_e)\).

\[
TH_e \geq y_e^+ - R_e - l_e; \quad \forall e \in E
\]

(25)

Eq. (26) indicates that a maximum of two vehicles can cooperate at a given transfer point node \(t\).

\[
\sum_{k,v} \sum_{i \in N \setminus T} x_{ij}^{k,v} \leq 2; \quad \forall e \in E, t \in T^e
\]

(26)
Eqs. (27)–(28) ensure that a vehicle is assigned to transfer out a patient \( e \) at transfer point node \( t \) only if another vehicle is assigned to transfer in the same patient \( e \) at the same transfer point node \( t \) where \( t \in T^e \). Two vehicles can cooperate if one is of type \( k \) and the other is of type \( k' \notin K^e \).

\[
\sum_{k \in K^e \setminus v} x_{i,j}^{k,v} = \sum_{k \in K^e \setminus v} \sum_{i \in N^e \setminus t} x_{i,j}^{k,v}; \quad \forall e \in E, t \in T^e \tag{27}
\]

\[
\sum_{k \in K^e \setminus v} x_{i,j}^{k,v} = \sum_{k \in K^e \setminus v} \sum_{i \in N^e \setminus t} x_{i,j}^{k,v}; \quad \forall e \in E, t \in T^e \tag{28}
\]

Eq. (29) guarantees that the two vehicles cooperating to transfer out and in a patient \( e \) at transfer point node \( t \) arrive at the same time to \( t \) (see Figs. 3(c) and 3(d)).

\[
\sum_{k \in K^e \setminus v} \sum_{i \in N^e} y_{i}^{k,v} = \sum_{k \in K^e \setminus v} \sum_{i \in T^e} y_{i}^{k,v}; \quad \forall e \in E \tag{29}
\]

The following Eqs. (30)–(34) maintain the binary restriction for \( x_{i,j}^{k,v} \) and non-negativity restrictions for \( y_{i}, j_{l}^{k,v}, R_{e} \) and \( TH_{e} \) variables.

\[
x_{i,j}^{k,v} \in [0,1]; \quad \forall k \in K, v \in V^{k}, i \in N, j \in N \tag{30}
\]

\[
y_{i} \geq 0; \quad \forall i \in N \setminus T \tag{31}
\]

\[
j_{l}^{k,v} \geq 0; \quad \forall k \in K, v \in V^{k}, t \in T \tag{32}
\]

\[
R_{e} \geq 0; \quad \forall e \in E \tag{33}
\]

\[
TH_{e} \geq 0; \quad \forall e \in E \tag{34}
\]

Note that indicator constraints appear in the formulation for the sake of clarity. Eqs. (11)–(16), (20), (23)–(24) are in fact linearized using the Big-M method.

5. Experimental plan

The mathematical formulation discussed above finds the optimal EMS operations to serve a set of emergency requests occurring during a given planning horizon. To evaluate the benefits behind the use of the modular vehicle technology for EMS operations, several scenarios (i.e. problem instances, each represented by a set of parameters) are solved. In this paper, we test the proposed model on the classical Sioux Falls transportation network (see Fig. 4) which is commonly used for demonstration and testing purposes in transportation studies (e.g. Di et al., 2018; Chow and Regan, 2011; Tilg et al., 2021). The network data are retrieved from the Transportation Networks repository (Transportation Networks for Research Core Team) supported by TRB’s Transportation Network Modeling committee. The proposed system is implemented in java on a computer with 2.4 GHz Intel Core i9 and a 32 GB 2400 MHz DDR4 memory. The sets and parameters are defined and inputted into the model written in AMPL which uses the CPLEX commercial solver to solve the MILP. Due to the synchronization requirements needed for the en-route transfer in the modular vehicle technology, solving the mathematical problem for large problem sizes and long planning horizons to optimality results in unpractical computation times. The development of an exact solution algorithm to allow the evaluation of the model on large problem instances is part of our future works.

In this paper, we develop a proof of concept for which we limit the experimental analysis to relatively small, yet realistic problem sizes. The sets and parameters fixed across all scenarios are shown in Table 2. Regarding the emergency set, the arrival of requests is generated using a Poisson arrival distribution while a uniform distribution of call locations over the network is adopted. An arrival rate of 2 emergencies per hour is tested. Given the size of the network, this is a reasonable demand value. For that rate, 100 different emergency sets with varying spatial and temporal distributions during a 1-hour planning horizon are defined. As discussed
previously in Section 4.1, a transfer point node is only included in \( T^e \) if the travel time from \( e^+ \) to \( t \) and from \( t \) to \( e^- \) is at least 4 min.

Also, we eliminate, from each \( T^e \) set, all the transfer point nodes that fall at travel times exceeding 1.5\( \tau_{e^+, e^-} \) from the emergency \( e \), as we do not expect a transfer to incur a larger deviation beyond the shortest path from \( e^+ \) to \( e^- \).

For a given emergency set, we solve the model for the three fleet sets below:

- Fleet set 1: 1 life support vehicle
- Fleet set 2: 1 life support vehicle and 1 medical transport vehicle
- Fleet set 3: 2 life support vehicles

In Section 6, we assess how the system reacts to different weighting coefficients \( \alpha_1 \) and \( \alpha_2 \) associated with \( R_e \) and \( TH_e \) respectively. The goal is to identify the combinations leading to the most practical and beneficial EMS operations with the modular vehicle technology, hence providing insights and recommendations that are relevant to decision makers. More importantly, we evaluate the benefits of the modular vehicle technology by comparing the EMS performance of fleet set 2 to the ones of fleet sets 1 and 3. For instance, fleet set 2 is the only set that allows the implementation of the newly introduced Operations B and C (see Figs. 3(c) and 3(d)) leveraging the modular vehicle technology. Fleet sets 1 and 3 can only accommodate Operations A and are added to the experimental plan for comparison and to reflect the current practice (i.e. with no modular vehicle technology). In other words, fleet sets 1 and 3 represent a lower bound and upper bound respectively for the performance of our proposed system fulfilled by fleet set 2.

In Section 7, the performance of the proposed model under varying demand and supply conditions is evaluated. In Section 7.1, the emergency arrival rate is increased from 2 to 3 emergencies per hour. Then, in Section 7.2, different hospital admission durations ranging from 10 to 30 min are tested. Finally, in Section 7.3, link travel times in the network are multiplied by a factor of 1.5 and 2 to emulate more sparse networks with longer travel times (e.g. rural areas). These tests will allow us to better understand the benefits of the modular vehicle technology to EMS, its flexibility and which operations are favored under different conditions. Note that the 100 different emergency sets of each arrival rate and fleet set are aggregated to generate Figs. 5–10 and Table 3, which are discussed below.

6. Trade-off analysis

As previously discussed, the optimization model minimizes an objective function equal to the weighted sum of the response times and the times to hospital for all emergencies. Different combinations of weighting coefficients \( \alpha_1 \) and \( \alpha_2 \) are tested, each ranging from 0 to 1 at 0.1 increments such that \( \alpha_1 + \alpha_2 = 1 \). For each emergency set and weight combination, the percent reductions in the objective value for fleet set 2 and 3 were computed relative to the objective value for fleet set 1, which has 1 life support vehicle
only. The reduction in the objective value due to an additional medical transport vehicle and the introduction of Operations B and C is reflected in fleet set 2, while the one due to an additional life support vehicle is shown using fleet set 3. Notice that one additional medical transport vehicle (fleet set 2) is expected to be much cheaper than one additional life support vehicle (fleet set 3). Results are shown in Fig. 5. Similarly, the mean reductions in response times, times to hospital, and prehospital times are computed for fleet sets 2 and 3 in comparison to fleet set 1 for each weight combination. They are shown in Fig. 6. Recall that the prehospital time \((PH_e)\) of a given emergency \(e\) is simply the sum of \(R_e\) and \(TH_e\), and represents the duration elapsed between the call time and the time at which the patient arrives to the hospital. Fig. 7 reveals how the share of emergencies addressed with each type of operations (A, B and C) for fleet set 2 changes with the weight combinations. To illustrate the trade-off between the weights in the objective function, we will first consider a simple example where 1 LS and 1 MT have to serve 2 emergencies in the system. For a given combination of weights where \(\alpha_1 \ll \alpha_2\), the optimization may determine that the EMS operations with the lowest objective value simply consists of executing two Operations of type A by the LS unit, first for emergency 1 and then emergency 2. In that case, en-route transfers (i.e. Operations B or C) may not be beneficial. Nevertheless, if the same scenario is tested with \(\alpha_1 \gg \alpha_2\), the execution of an Operation B for emergency 1 may lead to a better objective value. In fact, by freeing the LS unit after the transfer, a shorter response time may result for emergency 2 while an increase in time to hospital for emergency 1 is observed due to a slight detour. With more weight assigned to response times, switching from Operation A to Operation B for emergency 1 may result in a decrease in \(\alpha_1 R_e\) that is larger than the increase in \(\alpha_2 TH_e\), thus leading to a new solution with a lower objective value. Consequently, the weights attributed to the response time and time to hospital in the objective function largely affect the optimization process.

Different weight combinations of \(\alpha_1\) and \(\alpha_2\) result in different levels of EMS performance. As shown in Fig. 5, at low \(\alpha_1\) values between 0.1 and 0.3 (i.e. low weight for \(R_e\) and high weight for \(TH_e\)), higher reductions in the objective values are observed for fleet set 2 compared to fleet set 3. Fig. 6 shows that the reason behind this improved objective value for fleet set 2 at low \(\alpha_1\) is in fact the significant reductions in the times to hospital that can be achieved with the addition of a medical transport vehicle. The time to hospital of a given emergency \(e\) starts when care is provided at \(e^+\) or after the transfer is initiated at \(t \in T^e\). With the addition of a life support vehicle in fleet set 3, no reduction in the \(TH_e\) is possible when compared to fleet set 1 as the patient of emergency \(e\) will receive care at the emergency origin \(e^+\) and the travel time from \(e^+\) to the hospital \(e^-\) will always be the same for both fleet sets 1 and 3. However, with the addition of a medical transport vehicle in fleet set 2, the selection of Operation C will bring the patient closer to the hospital \(e^-\) before providing care, hence reductions in the times to hospital are achieved. Those come at the expense of higher response times. In other words, when more weight is attributed to the times to hospital, (i.e. \(\alpha_1 < \alpha_2\)), the model favors operations of type C, as shown in Fig. 7 with a share of at least 50%. According to Fig. 6, this will only lead to reductions in terms of the times to hospital but will in fact engender increases in the response times, which is not recommended. To sum up, it is not reasonable to claim that the EMS performance achieved with an additional medical transport vehicle is better than the one with an additional life support vehicle based on the observations with weight combinations of \(\alpha_1 < \alpha_2\). As a matter of fact, such combinations should be avoided, because the model will indirectly impose transfers that are in reality not beneficial, hence resulting in inefficient dispatching decisions.

Weight combinations with \(\alpha_1 \geq \alpha_2\) lead to better objective values for fleet set 3 compared to fleet set 2, as shown in Fig. 5. This is expected because the reductions that can be achieved with an additional life support vehicle in terms of response times are higher than the ones that can be achieved with an additional medical transport vehicle. Hence, the difference in objective value

![Fig. 5. Percent reduction in the objective value compared to fleet set 1 for 2 emergencies/hr and different weights.](image-url)
reductions becomes more evident as higher weight is assigned for $R_e$ than for $T_H$. Nevertheless, it is important to note that for fleet set 3, all weight combinations generate the same dispatching decisions, thus the percent reductions due to an additional life support vehicle in terms of response times and times to hospital are constant, see Fig. 6. As discussed previously, an additional life support vehicle can immediately provide care upon arrival at the emergency scene; however, it cannot lead to an earlier arrival to the hospital as the travel time from the emergency to the hospital will remain the same regardless of the fleet size. Hence, the optimal dispatching decisions that lead to the lowest cumulative response times and objective value with fleet set 3 are the same regardless of the weights adopted, as long as $\alpha_1 \neq 0$.

Regarding fleet set 2, the introduction of a medical transport vehicle can lead to reductions in terms of response times but also in terms of times to hospital. Consequently, the weights $\alpha_1$ and $\alpha_2$ affect the solution and may lead to different optimal dispatching decisions with fleet set 2. According to Fig. 6, as $\alpha_1$ increases, the percent reduction in terms of response times increases while the one in terms of times to hospital decreases. This is linked to the lower share of emergencies addressed with Operation C (Fig. 7). Recall that when Operation C is selected, the patient is transported to a transfer point closer to the hospital, hence reducing the remaining times to arrive to the hospital, where advanced treatments and interventions can be provided. Furthermore, Operation C may result in earlier interventions (i.e. shorter response times) as it brings the patient closer to the life support vehicle. However, according to Fig. 7, Operation B becomes more prevalent as more weight is assigned for $R_e$, hence showing that it has more impact on the response times compared to Operation C. As $\alpha_1$ goes from 0.6 to 1, solutions with lower response times are favored at the expense of higher times to hospital, potentially resulting in a slight increase in the prehospital times, as shown in Fig. 6. Thus, it is better to avoid combinations with $\alpha_1 \gg \alpha_2$ (i.e. $\alpha_1 \geq 0.8$), to fully benefit from the capabilities of the modular vehicle technology.

In summary, to achieve the best EMS performance with modular vehicle technology, it is recommended to adopt weighting coefficients such as $\alpha_1 \geq \alpha_2$ but to also avoid combinations such as $\alpha_1 \gg \alpha_2$ (i.e. $\alpha_1 \geq 0.8$), to fully benefit from the capabilities of the modular vehicle technology.

7. Sensitivity to demand and supply

In the following tests, the EMS performance with the modular vehicle technology is evaluated under varying demand and supply conditions. Based on the analysis from Section 6, the recommended weighting coefficients $\alpha_1 = 0.6$ and $\alpha_2 = 0.4$ are selected. First, in Section 7.1, the emergency rate is increased from 2 to 3 emergencies/hr, replicating scenarios with higher demand and when fleet availability and management become more valuable. Then, in Sections 7.2 and 7.3, the proposed system is tested with different admission times at hospitals and travel times in the network, respectively, to assess its performance and flexibility at various supply levels. For instance, increasing these durations indirectly reduces the supply in the system as resources (i.e. vehicles, personnel, and equipment) are blocked and occupied for longer periods of time.

7.1. Emergency demand increase

In this test, the recommended weight combination with $\alpha_1 = 0.6$ and $\alpha_2 = 0.4$ is selected and the emergency rate is increased from 2 to 3 emergencies per hour. Response times and times to hospital are expected to increase as demand becomes higher. Emergencies
may not be served immediately upon the receipt of the emergency call as all vehicles in the system may be occupied and blocked while responding to other emergencies. In Fig. 8(a), the share of fleet set 3 reduction compared to fleet set 1 and secured with fleet set 2 (in terms of the objective value) for each emergency arrival rate is illustrated. The higher the share, the more benefits we get out of introducing Operations B and C with a medical transport vehicle in lieu of a more expensive life support vehicle. Fig. 8(b) provides more details and indicates the mean shares of fleet set 3 reductions (relative to fleet 1) that are secured with fleet set 2 in terms of prehospital times and response times, while Fig. 8(c) shows the reductions that fleet set 2 offers in terms of times to hospital compared to fleet set 1. Recall that an additional life support vehicle in fleet set 3 cannot improve the remaining times to hospital.

According to Fig. 8(b), an additional medical transport vehicle (in fleet set 2) achieves 63.78% of the reduction in prehospital times that an additional and more expensive life support vehicle (in fleet set 3) offers, compared to fleet set 1 for 2 emergencies per
hour. This percent share increases to 64.17% as the emergency rate goes from 2 to 3 emergencies per hour. Similarly, as the rate of emergency demand increases, the value of fleet set 2 with respect to fleet set 3 in terms of response times increases (Fig. 8(b)) and significant reductions in times to hospital compared to fleet set 1 are observed (Fig. 8(c)). This happens because the increased flexibility that is offered by Operations B and C due to the modular vehicle technology allows a better redistribution of resources; which in reality becomes more critical as the emergency demand increases. Therefore, it is highly expected to observe even more considerable benefits with larger network sizes and number of emergencies (see Appendix). Moreover, the increase in arrival rates from 2 to 3 emergencies per hour leads to a decrease in the interquartile range (IQR), as shown in Fig. 8(a). This shows a lower variability in terms of performance level across the 100 sets, indicating that the system becomes more robust.

Recall that an additional medical transport vehicle incurs less costs than an additional life support vehicle, as the latter is fully equipped with a complete EMS personnel team, while the medical transport vehicle is expected to (1) have cheaper fixed and operating costs with reduced human resources and (2) can be potentially borrowed from other modular transportation systems for specific extreme scenarios. Thus, when resources are scarce, EMS providers could then evaluate whether a fleet set 2 is feasible from a financial perspective, knowing that it would still yield significant benefits.

Finally, a thorough inspection of the results indicates that the proposed optimization model often indirectly engenders early transfers resulting in a better performance with fleet set 2 than fleet set 3 due to its full visibility into the future. For instance, in some cases, a patient is being transferred out from a life support vehicle to a medical transport vehicle even when no call exist, so that the life support vehicle is available when a new call arrives. This highlights the value of early transfers which can be potentially encouraged in the future and incorporated into policies as a strategy to anticipate unknown demand, if the modular vehicle technology is deployed and if the budget allows it.

7.2. Admission duration variation

In this test, we evaluate how the proposed system reacts to changes in the admission durations at hospitals to better understand its benefits and which operations are favored under different conditions. The hospital admission duration is the time needed to move the patient from the ambulance to the hospital and to complete the required paperwork by the EMS staff. During this time, the resources are blocked instead of responding to pending calls. With the introduction of Operation B that is possible with the modular vehicle technology, a medical transport vehicle cooperates with a life support vehicle at a transfer point and resumes the transport of the patient to the hospital where it will carry out hospital admission tasks. Thus, the life support vehicle will be able to answer pending emergency calls earlier, leading to shorter response times. In these tests, we select one of the recommended weight combination with $\alpha_1 = 0.6$ and $\alpha_2 = 0.4$ and we test admission times equal to 10, 20 and 30 min.

As shown in Figs. 9(a)–9(c), an increase in the admission durations from 10 to 30 min leads to an EMS performance with fleet set 2 that is closer to the upper bound performance (i.e. largest reduction) identified with an additional life support vehicle in fleet set 3. A smaller variability in terms of performance level across the 100 sets is observed as admission times increases (Fig. 9(a)). Moreover, the shares of fleet set 3 reduction (relative to fleet set 1) secured with fleet set 2 in terms of the prehospital times and response times increase (Fig. 9(b)) as the admission time goes from 10 to 30 min. Notice that the increase in the admission time
Table 3
Share of emergencies addressed with each operation for different admission times.

| Share of emergencies addressed with Operation B and C | \( c_e = 10 \) | \( c_e = 20 \) | \( c_e = 30 \) |
|------------------------------------------------------|---------------|---------------|---------------|
| Operation B                                          | 18.5%         | 24%           | 25.5%         |
| Operation C                                          | 21%           | 18%           | 16.5%         |

Fig. 10. (a) Share of fleet set 3 reduction (compared to fleet set 1) secured with fleet set 2, in terms of objective value, for different travel times factors. (b) Share of fleet set 3 reduction (compared to fleet set 1) secured with fleet set 2, in terms of response time and prehospital time, for different travel times factors. (c) Reduction in times to hospital that fleet set 2 offers compared to fleet set 1 for different travel times factors.

at hospital does not affect the patient being served but it actually impacts future emergencies by delaying the time at which they will receive medical intervention. As previously discussed, Operation B frees the life support vehicle earlier so that it can respond to waiting emergencies, hence benefiting future demands rather than the currently addressed emergency. This is why, the system adapts to longer admission times by increasing the share of emergencies addressed with Operation B, from 18.5% to 25.5% as shown in Table 3. As a result, the reductions offered by fleet set 2 in terms of times to hospital compared to fleet set 1 inevitably decrease from 19.25% to 15.85% (as shown in Fig. 9(c)) but this is at the expense of significant improvements in terms of response time (Fig. 9(b)).

To conclude, Operations B are encouraged when expecting long admission durations. The fulfillment of the hospital admission tasks by medical transport vehicles will in fact alleviate the impacts of blocking the life support vehicle and its resources during long admission times while other emergencies are waiting.

7.3. Travel times variation

This test consists of evaluating the benefits of the en-route patient transfers and the new operations in a sparser network with longer travel times, commonly observed in rural areas. Increased travel times indirectly engender limited supply as EMS vehicles and resources are occupied for longer durations while moving from one location to another. Yet, deteriorated EMS performance and the use of more tolerable EMS performance targets compared to urban areas should not be allowed. To replicate these scenarios, the same transportation network is used and the link travel times are multiplied by factors of 1.5 and 2.

As shown in Fig. 10(a), the share of fleet set 3 objective value reduction (compared to fleet set 1) that can be achieved with fleet set 2 increases as the link travel time factor goes from 1 to 2. Similarly, the share in terms of prehospital times considerably increases from 63.78% to 85.97% (Fig. 10(b)). Long transportation times affect response times as well as time to hospitals, which can both be enhanced with Operations B and C. The proposed system does not encourage a specific type of operation to lessen the impact of increased travel times, but it reassigns the fleet and selects the operations that are beneficial to the system without following a clear trend. Note that the very slight decrease in the mean response time as going from a factor of 1 to 1.5 is at the expense of shorter times to hospital which lead to better prehospital times.

Furthermore, with longer travel times, the flexibility of Operations B and C may be reinforced by an increase in transfer point locations for a given emergency. As previously discussed, some transfer points may be excluded if they are too close to the emergency
origin or to the hospital. Hence, with sparser networks, more beneficial transfer points can be identified, leading to an improved EMS performance with fleet set 2 falling closer to the one with fleet set 3.

8. Conclusion

Operations of EMS still require research attention. Advancements in vehicular technologies are continuously emerging but they may be often deemed unnecessary if their capabilities are not fully leveraged. In this paper, we focus on a new technology proposed by the Next Future Transportation system (Next Future Transportation Inc., 2018a) which enables the smooth in-motion transfer of passengers from and to smart vehicular modules that can couple and decouple while moving. This modular vehicle technology primarily targets transit and ride sharing systems. However, we reinterpret and assess its implementation from a new perspective, with the aim of bringing to light its benefits to EMS. New types of response operations are introduced where life support and medical transport vehicles cooperate via patient transfers. We propose a deterministic optimization model which determines the optimal fleet dispatching decisions to identify the ultimate benefits of this technology, if deployed. Tests are executed on the Sioux Falls transportation network with emergency requests generated in a 1-hour planning horizon. A sensitivity analysis for the weighting coefficients associated with the response times and times to hospital is conducted, and the combinations leading to the best EMS performance with the modular vehicle technology are highlighted. When assessing the potential benefits that the modular vehicle technology can offer, the modular vehicle technology results in an improved EMS performance that falls closer to the upper bound performance that can be achieved with a traditional fleet size increase. Hence, it is valuable as the addition of medical transport vehicles requires lower fixed and operational costs compared to the addition of life support vehicles. Moreover, results show that the system with the modular vehicle technology brings larger benefits as the demand increases, and as the supply becomes more limited. It is important to note that the model presented in this paper is not intended to be used as a real-time dispatching method since emergency demand is in reality unknown. This is why the development of a real-time EMS dispatching method, that takes into consideration the high level of uncertainty linked to the demand, service characteristics and travel times, will be included in the scope of future works. In addition, the value of a real-time deterministic (non-anticipatory) approach will also be evaluated to assess the performance of the proposed EMS system with modular vehicle technology in situations where no information about future emergency demand is available or when it is not as expected, since this would be interesting to EMS decision makers. Also, the classification of emergencies into groups to be treated differently could be included in the future. We anticipate the development of additional types of operations that are tailored to specific groups of emergencies, making the system even more flexible and efficient. Finally, future works could include a trade-off analysis in terms of costs to better highlight the fixed and operational costs driven by the addition of the medical transport vehicles in lieu of the life support vehicles.

CRediT authorship contribution statement

Gaby Joe Hannoun: Conceptualization, Methodology, Programming, Validation, Investigation, Results analysis, Writing – original draft. Mónica Menéndez: Conceptualization, Methodology, Results analysis, Writing – review & editing, Supervision.

Acknowledgments

This work was supported by the NYUAD Center for Interacting Urban Networks (CITIES), funded by Tamkeen under the NYUAD Research Institute Award CG001.

Appendix

Solving the proposed MILP to optimality using the CPLEX commercial solver for large problem sizes and long planning horizons results in unpractical computation times. So, we develop here a heuristic to evaluate the proposed EMS system with the modular vehicle technology on larger problem instances to have an idea about how the results scale up to larger problems. For a given set of emergency requests, the heuristic finds a route for each vehicle (LS or MT) in a way that efficiently addresses all the requests. In other words, the heuristic ensures that each request receives treatment and transport to the hospital and that no emergency is left unanswered. As shown in Fig. 11, we start by listing the emergencies first in an ascending order of call time (i.e. FIFO approach). The heuristic will then select the first emergency on the list and will identify the best operation type and corresponding vehicle(s) to fulfill it. Next, it will remove this emergency from the list of unassigned requests and consider the next emergency, if any. So, by sequentially evaluating each emergency in the list, the heuristic will be constructing a route for each vehicle. Since the heuristic considers one emergency at a time and follows a pre-specified order which may not necessarily yield the best EMS performance, an outer loop is added where other potential orders of emergencies (different from FIFO) are tested to identify new plans (i.e. set of EMS vehicle assignments) which may be better. This would be useful especially when two consecutive emergencies happen very closely in time and swapping them (i.e. serving the second emergency first) may be beneficial. After evaluating all potential emergency orders, the one leading to the best overall EMS performance is identified and its corresponding EMS vehicle routes are selected.

When an unassigned emergency e is selected from the list, we assess all possible combinations of operations and vehicle(s) to see which one leads to the lowest total cost \( TC = a_1 \sum_e R_e + a_2 \sum_e T_H_e \) for the previously assigned requests and e. Once the best response operation for a given emergency e is identified, the sequence of the vehicle(s) fulfilling it will be constructed up to the node at which it(they) will become free again. A vehicle may complete a given operation either at a hospital (end of the patient
Fig. 11. Trip-based heuristic.

discharge) or at the end of the transfer. This is why, while constructing the vehicle sequences, a return to a station will not be added (unless \( e \) is the last emergency in the list), as the insertion of the next emergency may require the movement of the vehicle straight from the hospital or transfer point to the next emergency’s origin or to a transfer point to transfer in its patient. Algorithms 1, 2 and 3 identify the lowest total costs \( TCA, TCB, \) and \( TCC \) that can be achieved by implementing a response operation \( A, B \) or \( C \), respectively. The algorithms are not included in this paper but are available to the readers upon request.

Additional scenarios with larger problem instances were tested with this heuristic. We increased the emergency rate up to 7 emergencies per hour in the Sioux Falls network and evaluated the EMS performance with the following fleet sets. Note that fleet sets with the number of MT units exceeding the number of LS units are not considered as these sets will be of more interest when the transportation network is larger (i.e. longer transportation times).

| Fleet set   | LS | MT |
|-------------|----|----|
| Fleet set 1 | 1  | 0  |
| Fleet set 2 | 1  | 1  |
| Fleet set 3 | 2  | 0  |
| Fleet set 4 | 2  | 1  |
| Fleet set 5 | 3  | 0  |
| Fleet set 6 | 2  | 2  |
| Fleet set 7 | 3  | 1  |
| Fleet set 8 | 4  | 0  |

For the sake of brevity, we will limit this analysis to the results illustrating the reductions in terms of average prehospital times per emergency for a given tested fleet set and emergency arrival rate. Recall that the prehospital time of an emergency is simply the sum of the response time and time to hospital. For each fleet set (except fleet set 1), the average reductions in terms of prehospital times per emergency are computed with respect to its lower bound, for different emergency arrival rates. These results are shown in Figs. 12(a) to 12(c). In Fig. 12(a), the reductions achieved with an additional MT unit or LS unit with fleet sets 2 and 3 with
Fig. 12. Average prehospital times reductions per emergency for different emergency rates (a) for fleet sets 2 and 3 compared to fleet set 1, (b) for fleet sets 4, 5 and 6 compared to fleet set 3, (c) for fleet sets 7 and 8 compared to fleet set 5.

respect to fleet set 1 are shown, respectively. Similarly, Figs. 12(b) and 12(c) illustrate the reductions with fleet sets 4 and 5 with respect to fleet set 3, and fleet sets 7 and 8 with respect to fleet set 5, respectively. Note that the reduction offered by fleet set 6 is included in Fig. 12(b) and the reason behind this representation is discussed further below. In these figures, the dashed lines represent the traditional fleet sets (i.e. no modular vehicle technology; fleet sets 3, 5 and 8), which in fact serve as upper bounds for the other remaining sets (solid lines), which are equipped with modular vehicle technology and hence capable of executing Operations B and C (fleet sets 2, 4, and 7). According to these figures, the EMS performance of fleet sets 2, 4 and 7 (with 1 MT unit) fall between their corresponding lower bound with fleet sets 1, 3 and 5 respectively (with the same number of LS units) and upper bound with fleet sets 3, 5 and 8 (with the MT unit replaced by an additional LS unit), respectively. However, it is important to note that the performance of these fleet sets leveraging the en-route transfers are closer to their respective upper bounds. Recall that an MT unit is expected to have cheaper operating costs. So, from a financial perspective, and especially at high emergency demand levels where the procurement of larger fleet sizes to satisfy the demand becomes more challenging, the deployment of MT units becomes more valuable. Furthermore, at high emergency demands, staff shortages may often occur which will primarily affect the operations of the LS units. However, the MT units in charge of transportation and hospital discharge tasks may still be deployed as these require less staff, hence providing a faster service than a system case with no MT unit. On the other hand, an MT unit cannot operate independently. Its service has to be complemented by the one of an LS, while the latter can execute a complete operation independently (Operation A). Hence, it is expected to observe lower performance when one LS is replaced by one MT. Nevertheless, we decided to investigate the case where 1 LS unit is replaced by 2 MT units, so we introduced fleet set 6 and illustrated its corresponding performance with respect to fleet set 3 in Fig. 12(b). According to the results, the EMS performance of fleet set 6 (2 LS and 2 MT units) is better than that of fleet set 5 which has 3 LS units. In other words, EMS operations with 2 additional MT units outperform EMS operations with 1 additional LS unit. Moreover, this slight improvement of fleet set 6 over fleet set 5 is in fact significant as fleet set 6 should be in fact compared to fleet set 3 which has the same number of 2 LS units. For instance, the two additional MT units could potentially belong to a fleet that is borrowed from other transport fleets (e.g. transit systems).

To sum up, as the emergency rate increases, the average reduction in terms of prehospital time per emergency with fleet sets 2, 4, 6, and 7 also increase. No plateau is observed which means that MT units and en-route transfers allow EMS operations with modular vehicle technology to adapt to higher emergency demands. Besides, as previously discussed, the modular technology becomes more valuable at higher emergency arrival rates, as the availability of enough supply becomes more challenging, hence borrowing MT units from other transport modes becomes more valuable. In other words, we would recommend decision makers to always procure
as many LS units as the budget allows, since more LS will always achieve better performance. An MT can be procured with the remaining budget which is not enough to purchase an extra LS, or it can be borrowed from other transport modes. Therefore, it is meaningful to compare a fleet set with MT units to its lower bound fleet set only, and based on the results in this Appendix, the reductions are significant and always increasing with the increase emergency rate. In our future works, the EMS with modular technology will be evaluated in a real size network allowing us to test a wider range of scenarios, while using real-time techniques that take into consideration the unknown emergency demand in a dynamic context.

References

Aboueljinaie, L., Sahin, E., Jemai, Z., 2013. A review on simulation models applied to emergency medical service operations. Comput. Ind. Eng. 66 (4), 734-750.

Bailey, E.D., Sweeney, T., 2003. Considerations in establishing emergency medical services response time goals. Prehospital Emerg. Care 7 (3), 397-399.

Ball, M.O., Liu, F.L., 1993. A reliability model applied to emergency service vehicle location. Oper. Res. 41 (1), 18-26.

Brotcorne, L., Laporte, G., Semet, F., 2003. Ambulance location and relocation models. European J. Oper. Res. 147 (3), 451-463.

Caros, N.S., Chow, J.Y., 2021. Day-to-day market evaluation of modular autonomous vehicle fleet operations with en-route transfers. Transp. B: Transp. Dyn. 9 (1), 109-133.

Chen, Z., Li, X., 2021. Designing corridor systems with modular autonomous vehicles enabling station-wise docking: Discrete modeling method. Transp. Res. E: Logist. Transp. Rev. 152, 102388.

Chen, Z., Li, X., Zhou, X., 2019. Operational design for shuttle systems with modular vehicles under oversaturated traffic: Discrete modeling method. Transp. Res. B 122, 1-19.

Chen, Z., Li, X., Zhou, X., 2020. Operational design for shuttle systems with modular vehicles under oversaturated traffic: Continuous modeling method. Transp. Res. B 132, 76-100.

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.

Chow, J.Y., Regan, A.C., 2011. Network-based real option models. Transp. Res. B 45 (4), 682-695.

Cortés, C.E., Matamala, M., Contardo, C., 2010. The pickup and delivery problem with transfers: Formulation and a branch-and-cut solution method. European J. Oper. Res. 200 (3), 711-724.

Cui, E.R., Beja-Glasser, A., Fernandez, A.R., Grover, J.M., Mann, N.C., Patel, M.D., 2019. Emergency medical services time intervals for acute chest pain in the United States, 2015–2016. Prehospital Emerg. Care 1–9.

Dai, Z., Liu, X.C., Chen, X., Ma, X., 2020. Joint optimization of scheduling and capacity for mixed traffic with autonomous and human-driven buses: A dynamic programming approach. Transp. Res. C 114, 598-619.

Dasik, I., Yang, K., Menendez, M., Chow, J.Y., 2021. On the design of an optimal flexible bus dispatching system with modular bus units: Using the three-dimensional macroscopic fundamental diagram. Transp. Res. B 148, 38-59.

Di, X., Ma, R., Liu, H.X., Ban, X.J., 2018. A link-node reformulation of ridesharing user equilibrium with network design. Transp. Res. B 112, 230-255.

Gecelcin, T., Webb, J., 2019. Modular dynamic ride-sharing transport systems. Econ. Anal. Policy 61, 111–117.

Ghazi Balkiz, T.Q., Mackintosh, E., 2020. Beirut port ablaze, weeks after massive blast. CNN, https://edition.cnn.com/2020/09/10/middleeast/beirut-port-fire/index.html, Accessed: 2020-09-10.

Gong, M., Hu, Y., Chen, Z., Li, X., 2021. Transfer-based customized modular bus system design with passenger-route assignment optimization. Transp. Res. E: Logist. Transp. Rev. 153, 102422.

Guo, Q.-W., Chow, J.Y., Schonfeld, P., 2018. Stochastic dynamic switching in fixed and flexible transit services as market entry-exit options. Transp. Res. C 94, 288-306.

Hulshof, P.J., Kortbeek, N., Boucherie, R.J., Hans, E.W., Bakker, P.J., 2012. Taxonomic classification of planning decisions in health care: a structured review of the state of the art in OR/MS. Health Syst. 1 (2), 129–175.

Krishnan, K., Marla, L., Yue, Y., 2016. Robust ambulance allocation using risk-based metrics. In: 2016 8th International Conference on Communication Systems and Networks (COMSNETS). IEEE, pp. 1–6.

Li, X., Zhao, Z., Zhu, X., Wyatt, T., 2011. Covering models and optimization techniques for oversaturated service location and planning: a review. Math. Methods Oper. Res. 74 (3), 281–310.

Manka, M., Moscati, R., Raghavendra, K., Priya, A., 2010. Sonographic scoring for operating room triage in trauma. West. J. Emerg. Med. 11, 138–143.

NHTSA, 2010. Emergency medical services agenda for the future. https://one.nhtsa.gov/people/injury/ems/agenda/emsman.html, Accessed: 2020-01-30.

Next Future Transportation Inc., 2018a. Next future transportation. http://www.next-future-mobility.com, Accessed: 2020-02-14.

Next Future Transportation Inc., 2018b. Next future transportation inc. Introduces smart airport solutions. Aviation Pros, https://www.aviationpros.com/home/press-release/124341855/next-future-transporation-inc-next-future-transportation-inc-introduces-smart-airport-solutions, Accessed: 2020-04-30.

NFPA, 2020. Standard for the organization and deployment of fire suppression operations, emergency medical operations, and special operations to the public by fire care fire departments. NFPA 1710, URL: https://www.nfpa.org/codes-and-standards/all-codes-and-standards/list-of-codes-and-standards/detail?code=1710.

NFPTA, 2010. Emergency medical services agenda for the future. https://one.nhtsa.gov/people/injury/ems/agenda/emsman.html, Accessed: 2020-01-30.

Parragh, S.N., Doerner, K.F., Hartl, R.F., 2008. A survey on pickup and delivery problems. J. FÜR Betriebswirtschaft 58 (2), 81–117.

Pelletier, E., 2020. How to help Lebanon after beirut explosion. The New York Times, https://www.nytimes.com/2020/08/05/world/how-to-help-lebanon-beirut.html, Accessed: 2020-09-12.

Rajagopalan, H.K., Saydam, C., Setzler, H., Sharer, E., et al., 2011. Ambulance deployment and shift scheduling: An integrated approach. J. Serv. Sci. Manag. 4 (1), 66.

Saeed, Z., He, W., Menendez, M., 2022. Application of modular vehicle technology to mitigate bus bunching. In: Transportation Research Board Annual Meeting, 101th, 2022, Washington, DC, USA.

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? Accid. Anal. Prev. 42 (4), 1048-1056.

Schmid, V., Doerner, K.F., 2010. Ambulance location and relocation problems with time-dependent travel times. European J. Oper. Res. 207 (3), 1293–1303.

Shi, X., Chen, Z., Pei, M., Li, X., 2020. Variable-capacity operations with modular transits for shared-use corridors. Transp. Res. Rec. 2674 (9), 230–244.

Tafreshian, A., Masoud, N., 2020. Modular transit: Using autonomy and modularity to improve performance in public transportation. Transp. Res. E: Logist. Transp. Rev. 141, 102033.