Contents lists available at GrowingScience

International Journal of Industrial Engineering Computations

homepage: www.GrowingScience.com/ijiec

An online real-time matheuristic algorithm for dispatch and relocation of ambulances

Juan Camilo Paz Roa[a], John Willmer Escobar[a,c*] and Cesar Augusto Marín Moreno[b]

[a]Department of Civil and Industrial Engineering, Pontificia Universidad Javeriana Cali, Cali 76001000, Valle del Cauca, Colombia
[b]Universidad Tecnológica de Pereira., Pereira 660001, Risaralda, Colombia
[c]Department of Accounting and Finance, Universidad del Valle, Cali 76001000, Valle del Cauca, Colombia

C H R O N I C L E

Article history:
Received October 5 2019
Received in Revised Format November 6 2019
Accepted November 6 2019
Available online November 6 2019

Keywords:
Ambulances
Emergency Medical Vehicles
Relocation
Dispatch
Matheuristic Algorithm
Optimization
Discrete Event Simulation

A B S T R A C T

The Medical System of Transportation deals with two online real-time decisions: ambulance dispatching and relocation. Dispatching consists of selecting which ambulance to send to an emergency call; relocation consists of determining how to modify the location of available ambulances in response to changes in the system's state. Although the literature regarding this problem is extensive, only a limited number of online real-time approaches for ambulance management have been proposed, much less one taking into consideration different types of emergencies and vehicles. This paper proposes an online real-time matheuristic algorithm that combines: i) a new preparedness index defined as the availability probability of a multi-server queue model which is used as an optimization objective and as a control variable for relocation strategies, ii) two mathematical models to solve the relocation problem, one oriented to the maximization of coverage and other to the minimization of the maximum relocation time, and iii) two heuristic algorithms oriented to the maximization of the preparedness level, one to solve the dispatch problem and other to solve the location problem of one ambulance. The computational experiments, based on discrete event simulation and historical data of Bogotá, Colombia, have shown their capability to adequately respond to the necessities of real-time operation.

1. Introduction

1.1 Background

The Medical System of Transportation (MST) allows for medical attention and transport of patients in a determined geographical area. Between the policies addressed in the MST, there exist decisions related to the management of ambulances in charge of emergencies. The first one is concerned with the real-time location of the ambulances in a set of possible sites in order to continually ensure the best and most effective response capacity of the system to requests. The second one refers to the dispatch of the necessary vehicles in response to requirements of service. Together, these decisions of location-relocation and dispatch are directly responsible for the system performance. This is because if the system operator does a balanced distribution of the vehicles all over the geographical area of coverage, then the vehicles must be able to reach the site of the emergencies quickly. However, once a vehicle is dispatched,

* Corresponding author
E-mail: jwescobar@javerianacali.edu.co (J. W. Escobar)

2020 Growing Science Ltd.
doi: 10.5267/j.ijiec.2019.11.003

it is no longer available to respond to emergencies in the area of influence where it was initially assigned. If this fact is ignored, and another emergency occurs in the same area, the response time to this emergency could be higher than expected; thus, the survival probability of the patient could be less than the one obtained with better management of the relocation decisions. Additionally, the system works nonstop, not knowing the places where future emergencies will occur, facing the natural uncertainty of the system (demand, displacement times, capacity) and under complex operational restrictions arising from the different types of vehicles available to respond to specific types of emergencies. Given the complexity and the implications of this system in the survivability of patients, the design of these policies has been a challenging and relevant problem which has taken the attention of the scientific community. Such developments can be tracked from the first approaches dealing with location problems (Hakimi, 1964) to the last known advances of this specific area (presented in the literature review of Aringhieri et al., 2017 and Bélager et al., 2019).

The characteristics mentioned above make the MST a complex system which needs a decision support tool in order to preserve the life of patients. For this reason, multiple approaches have been developed such as mathematical models (e.g., Sung and Lee, 2018; Enayati et al., 2018, b), analytic models of queue theory (e.g., Karimi et al., 2018), dynamic programming algorithms (e.g., Nasrollahzadeh et al., 2018), simulation models (e.g., Kergosien et al., 2015; Pinto et al., 2015; McCormack and Coates, 2015), and hybrid methods (e.g., Enayati et al., 2019). However, now that we have significant advances in information technologies, global positioning systems, and real-time data processing systems, the application of these technologies, together with traditional decision support approaches, have received increased attention in the recent literature (see Section 1.2). In fact, this opportunity has encouraged the development of online real-time optimization systems, which is the main focus of this research.

These kinds of techniques have set new challenges to the scientific community. For example, even if there exists plenty of location models in the literature, some of the classical models were designed without taking into account the possibility of real-time tracking of the system condition (e.g., geolocation of ambulances and patients, availability in hospitals, traffic conditions, etc.). Thus, when taking into consideration the availability of real-time information, it must be determined how this information could be incorporated into online decision-support systems. Also, these techniques must run with efficient computational time, which is a challenge given the complexity of the problem. Additionally, they must give a balance between the relocation and service level because an excessive number of relocations is prohibited in practice. In this context, few online real-time approaches for ambulance management have been proposed, and they tackle the problem based on different assumptions and implementing different characteristics. Thus, a relevant research problem is still the development of approaches able to lead the ambulance operation to one with a good level of performance, or at least better than the classical or empirical approaches.

In this work, we aim to design a new algorithm to support the real-time operation of ambulances. This is an algorithm which, in a given time, takes the information of the state of the system as input and returns to the entire fleet the orders necessary to render the service. Specifically, we propose an online real-time matheuristic algorithm that combines i) a new preparedness index defined as the availability probability of a multi-server queue model, ii) two mathematical models to solve the relocation problem (the double standard model of Gedreau et al., 1997, DSM, and a classical transportation model), and iii) two heuristic algorithms oriented to the maximization of the preparedness level, one to solve the dispatch problem taking in consideration different types of emergencies and vehicles, and other to solve the location problem of one ambulance. This is a matheuristic algorithm which, to the best of our knowledge, has not been proposed. The effectiveness and efficiency of the proposed algorithm were validated through a discrete event simulation (DES), which is based on real information gathered in the city of Bogotá, Colombia. The results of these computational tests have shown its capability to respond to the necessities of real-time operation adequately.

1.2. Related works

The literature on the dynamic ambulance operation management could be divided by their orientation into three categories (van Barneveld et al., 2018): periodic redeployment as well as offline and online real-time ambulance management. However, it must be noted that many of these works are based on developments in the classic ambulance location problem. For a further description of this literature, literature review articles by Goldberg (2004), Basar et al. (2012), Aringhieri et al. (2017), and Bélager et al. (2019) are recommended.

The works in the category of periodic redeployment split the planning horizon into discrete time periods and then solve the static ambulance location problem multiple times. The multi-period mathematical models are also part of this category, but they additionally define how to perform the movements between locations (e.g., Bagherinejad and Shoeib, 2018). The related literature to this relocation plan has been widely developed in the past, but it is evident that those solutions do not take into account real-time aspects of the system, such as the exact geographical position of the ambulances, in which state they are, and other information that is available in today's context.

In contrast to periodic redeployment approaches, real-time ambulance management bases decisions on the actual state of the system. In the offline branch of these tools, the solutions are precomputed, stored, and indexed by control variables or scenarios. The so-called Compliance Tables or System Status Management is one of these approaches. In this approach, the ambulances that must be in each location are usually selected using as reference the number of available ambulances as a control variable (e.g., Sudtachat et al., 2016; van Barneveld et al., 2017, b). Other offline approaches compute solutions in terms of scenarios/state variables, which have additional information about ambulances and emergencies. However, the number of scenarios is too large, yielding an intractable solution space. This issue has been tackled with approximate dynamic programming for the computation of ambulance relocation strategies (Schmid, 2012; Maxwell et al., 2010, 2013, 2014; Nasrollahzadeh et al., 2018). These two techniques have no problem with computational times; however, they simplify the state of the system in terms of control variables or scenarios, which ignore some of the real-time information of the system.

Both approaches — periodic redeployment and offline real-time — precompute solutions based on different types of tools, such as static mathematical optimization models with and without stochastic parameters (e.g., van den Berg et al., 2019), analytic models of queue theory (e.g., Karimi et al., 2018), simulation models (e.g., Kergosien et al., 2015), dynamic programming algorithms (e.g., Nasrollahzadeh et al., 2018), and hybrid methods (e.g., Enayati et al., 2019). However, no matter how these solutions are constructed, in practice, the operators must check the relocation plan, the control variable, or the scenario and make decisions based on those precomputed solutions. Even so, these tools could work fine in systems with a low level of changes in the system because operators could manage empirically those details not included in these approaches and try to keep the ambulance configuration close to the one suggested.

However, in systems with a high demand for operator decisions, the real operation of ambulances often changes, that is, due to the continual arrival of requests at which ambulances are dispatched. In this situation, online real-time approaches could be much more operable in the sense that they could adapt and better keep up with the fluctuating conditions of the system. In this paper, we focus on this type of approach. Thus, we present in this section a brief description of what we deem to be the most relevant works related to online real-time decision support tools for the location, dispatch, and relocation of ambulances. Then, based on this description, the contribution and differences with the literature of this research are explained in Section 1.3. The work of Gendreau et al. (1997), which is especially relevant to our relocation strategy, is detailed in Section 2.3.

To the best of our knowledge, Gendreau et al. (2001) proposed the first online system. This was embedded with a modified version of the double standard model (DSM, Gendreau et al., 1997). The main change is the insertion into the objective function of a penalty coefficient associated with the relocation of an ambulance from its current site to another location. These coefficients must be updated each time the model is solved. Based on this, the system tries to compute before an emergency occurs, and for each available ambulance, find a solution for the model. This assumes that the ambulance was dispatched and therefore is not available for a relocation operation. From this, when the emergency occurs and an ambulance is dispatched, the precomputed solution can be used. Furthermore, this strategy needs to run every single time a change in the system occurs. Given the complexity of the strategy, it is solved by a tabu-search algorithm using parallel computing. However, for some emergencies with small intervals of occurrence, the system was unable to compute a solution on time.

Andersson and Värbrand (2007) presented an online system called DYNAROC. This article proposes an integration of a heuristic algorithm for the ambulance dispatch (considering different types of emergencies but not different ambulances) and a nonlinear mathematical model for relocating idle ambulances, which is solved by a tree-search heuristic. The system as a whole operates by seeking to maintain a minimum level of a preparedness index, which is a measure suggested by the authors of the readiness of the system to respond to future emergencies. Also, instead of using a penalization coefficient, the preparedness index is used as the control variable triggering the relocations, allowing savings in computational time.

Haghani and Yang (2007) proposed a deployment system for emergency vehicles that embeds an optimization model to solve the dispatch and relocation decisions. They considered three types of emergency vehicles (police, firefighters, and medical personnel) and allowed dispatched emergency vehicles on route to switch to a new emergency call that is more severe. Although the proposed relocation and dispatch strategies are promising, the computational time required to solve the model is inefficient for an online real-time approach. The authors suggested but did not implement a tabu-search algorithm to overcome this issue.

Jagtenberg et al. (2015) designed a heuristic algorithm based on the online status of the system but for newly idle ambulances only, which we call in this paper a single location strategy. It uses a heuristic strategy in which the marginal contribution to the expected coverage obtained from locating the ambulance in each possible location is computed. Then, the best contribution is selected to send the location instruction.

Bélanger et al. (2016) modeled and analyzed four management strategies related to the location and relocation of an ambulance fleet. Strategies 1 and 2 correspond to cases of periodic redeployment. The first corresponds to a static mathematical model, and the second corresponds to a multi-period mathematical model. Therefore, these are non-real-time strategies. Strategies 3 and 4 correspond to an online real-time approach. In the third strategy, relocations of already located ambulances were not considered, while in the fourth strategy relocations are allowed. As we also do, the proposed strategies of the authors are based on the DSM (Gendreau et al., 1997). All the strategies were evaluated by simulation, and, contrary to the expected, the experiments revealed that multi-period relocation approaches seem to be dominated by the fully static strategy. Besides, it was shown that relocation strategies clearly lead to better service performance than static approaches. However, these strategies also generate significant increases in the total traveled distance and, eventually, the number of relocations, which might be difficult to implement with respect to human resources.

Van Barneveld et al. (2016) proposed a heuristic with optimization features for ambulance relocation that only considers relocation decisions i) when an ambulance is dispatched, and ii) when an ambulance becomes available. A relocation is triggered only if an improvement in an "unpreparedness" index is reached. The last is a measure of the possible cost derived from a future emergency given the current state of the system. If no improvement is expected, depending on the triggering event, the positions of the fleet are maintained, or the newly available ambulance is relocated to the nearest base station. A particular feature is that it is not necessary for an ambulance that needs to be relocated to move from the point of origin to the destination. Instead, the authors allow for several movements; for example, the ambulance that is at the point of origin can be moved to an intermediate location while, at the same time, the ambulance at the intermediate location moves to the destination. These movements are computed by a mathematical model of the Linear Bottleneck Assignment Problem and later are defined as chain relocations (van Barneveld et al., 2018). The tool was validated using real data and different cost functions for the computation of the unpreparedness index. The authors send the nearest server to emergencies, and only one type of emergency and ambulance is considered.

Van Barneveld et al. (2017, a) developed an online tool for the management of ambulances in rural regions with a limited number of ambulances. The problem is formulated as a discrete-time Markov decision process. The computational times of an optimal relocation policy are not efficient; thus, a one-step look-ahead heuristic was developed so that, at each time step, ambulances are relocated in order to minimize the expected response time.

Aringhieri et al. (2018) defined several online ambulance management policies and evaluated their performance using DES. These policies correspond to several combinations based on independent dispatch and relocation strategies. In detail, they combine four dispatch strategies and three relocation strategies. Regarding the dispatch problem, they use i) the classical dispatching policy of the closest server, ii) dispatching from a list of bases capable of reaching the request within the time threshold for the emergency, iii) the cutoff priority queue, and iv) the smart assignment. The last two strategies are possible extensions of the first two; iii) temporarily replaces one emergency request for another more severe, and iv) considers dispatching not only the ambulances available at a base but also those that are in the relocation phase. The first considered relocation strategy is that in which the ambulance always returns to its original base. In the second one, newly available ambulances are redeployed to the closest base. The third policy is almost the same as the second one except that it locates the ambulances to the less-covered base.

Enayati et al. (2018, a) proposed a management scheme for an ambulance fleet based on real-time optimization. The authors' approach used two mathematical models of linear programming in a series. The first model is oriented towards the maximization of coverage, and the second is oriented towards the minimization of relocation time but considers workload restriction. The authors evaluated the need to apply a general relocation each time the state of the system changes, comparing the current coverage against what would be obtained if relocation were made. If the increase in coverage exceeds a minimum percentage increase, the relocation is performed. Computational tests using discrete-event simulation show the applicability of the tool and reveal that the relocation scheme improves the coverage against the static policy scheme. Something noticeable in this work is that it is supposed that, at most, one ambulance is assigned to each location, which is not common in recent literature. Also, the authors suggested that future studies relating to this work consider patient priorities and different types of ambulances, or combine this model with different dispatching strategies.

Van Barneveld et al. (2018) improved the algorithm proposed by Jagtenberg et al. (2015) by using the characteristics of the work of van Barneveld et al. (2016). The effects of incorporating one or several of those characteristics in the new algorithm in both rural and urban areas were studied. The authors found that i) taking the classical 0-1 performance criterion for assessing the fraction of late arrivals differs only slightly from related response-time criteria, ii) it is beneficial for rural areas to consider moments of relocation, both when an ambulance is dispatched and when it goes back into operation, iii) relocation times are not significantly reduced if it is considered that once an ambulance has completed a given service, it is then available for coverage in the immediate area, iv) the use of chain relocations with more than two chain movements for a relocation does not generate significant benefits, v) the proposed tool allows operators to have better control over the number of relocations thanks to the implementation of time restrictions imposed on relocation as well as the number of relocations, and vi) it is of vital importance to use simulation in order to evaluate any relocation policy since the conditions from one Emergency Medical System (EMS) to another can change significantly.

The analysis of this brief literature overview reveals increasing attention on online real-time ambulance management optimization systems. From such an analysis, a list of insights can be derived. First, these approaches have become a prominent line of development, not only for their capacity to integrate robust optimization approaches with information that nowadays could be tracked in real time using GIS and information technology but also for their easy way of treating the time dependence of critical parameters (e.g., demand, traffic condition, etc.). Second, even though all of them face the same problem, the few proposed approaches in the literature have substantial differences in their assumptions and implemented strategies. Thus, the development of new approaches is pertinent, combining existing strategies or even incorporating innovative ones. Third, in a framework of applicable developments, it is of utmost importance that the dispatching and relocation strategies, control of the number of relocations triggered (due to impractical increases of the crew workload), incorporation of strategies dealing with the natural uncertainty of the operation and constraints imposed by evident characteristics of operation, such as a location's capacity and the capabilities of different vehicles to attend different types of emergencies.

1.3. Contribution

In this work, we sought to design an online real-time optimization algorithm for the management of an ambulance fleet. This algorithm is a novel approach that considers the following characteristics:

i) Aleatory Considerations. The algorithm uses the ambulance-available probability derived from a multi-server queue model as the preparedness index of the system.
ii) Flexible Location Policies. The algorithm takes location decisions in which ambulances do not need to be statically assigned to a single station.
iii) Integration of Location-Dispatch-Relocation Decisions. The algorithm could give, based on the system status and every time it changes, decisions that respond to single location necessities (e.g., an ambulance starts its work shift and needs to know where to be located), dispatch necessities (e.g., an emergency needs to be attended), and relocation necessities (e.g., a geographical area has reached a critical level of preparedness for future emergencies). All of this brings a balance between the service level and the workload generated by multiple relocations of ambulances.
iv) Robust Dispatch Criteria. The algorithm uses criteria that not just always assigns the nearest ambulance but looks at the general performance of the system, the priority of the emergencies, and the capabilities of a heterogeneous fleet of ambulances.
v) A Posteriori Approach. The algorithm is able to compute solutions in response to events and based on the actual information of the system without using precomputed solutions.
vi) Heterogeneous Fleet. The algorithm is designed to take into consideration the existence of different types of ambulances and their ability to attend different types of emergencies in location-dispatch and relocation decisions.
vii) Location's Capacity. The algorithm considers that stations could have a capacity equal to or more than one ambulance (some approaches consider locations with space for only one ambulance).
viii) Time-Dependent Parameters. The algorithm is able to change its parameters in function of time.
ix) Optimal results. The algorithm tackles the general location and relocation problems with MILP models that are solved to optimality using commercial optimization software.
x) Real-Time Applicability. The algorithm is able to run in efficient computational time even if processing data of a city with a population of approximately seven million and with responses based on information that today could be tracked in real time using GIS and information technology.

When compared with the mainstream literature described above, our approach provides a number of advantages. First, it approximates the random evolution of the system over time since its preparedness index is based on a queue theory formulation of a multi-server system. Additionally, the decisions made by our algorithm can be computed very quickly as this requires running heuristic algorithms or a simple optimization scheme only when it is required. Finally, and in contrast to all of them, our approach takes into consideration the existence of different types of ambulances and their ability to attend different types of emergencies; some take into consideration dispatch priorities (e.g., Andersson and Värbrand, 2007) but not their relation to different types of vehicles. Due to these advantages, our approach can fully automate the decision-making process and be used in problem instances with realistic dimensions. These are conditions required for any solution to be adopted in practice in systems with a high demand for the operator's decisions.

This paper differs from the mainstream literature in several aspects. Thus, we present here the differences in our work compared with those papers that we deem to be the more associated. For example, our approach does not need to compute a relocation strategy for each change in the system or solve the DSM for each ambulance that could be dispatched, much less use a coefficient of penalization in the objective function to control the relocations (Gendreau et al., 2001). Instead, we use the concept of preparedness introduced by Andersson and Värbrand (2007). Thus, a relocation strategy is triggered only when it is needed. Therefore, even if our approach embeds the same model of Gendreau et al., (2001), it is a fact that the general functionality of our algorithm is completely different.

Also, although we apply the concept of preparedness introduced by Andersson and Värbrand (2007), we use a different definition based on queue theory. Contrary to their proposal, our index i) does not use parameters that need to be calibrated, and ii) does not assume a value proportional to the time required for ambulances to reach the zone (instead, we align it with the binary concept of coverage). Second, our heuristic algorithm for dispatch, while improving the preparedness level of the system, considers different types of vehicles available to respond to specific types of emergencies. Third, our algorithm does not need an independent definition of a minimum preparedness level.

Haghani and Yang (2007) considered multiple types of emergency vehicles (police car, ambulances, etc.) but not different types of ambulances as we do. Our approach solves the whole problem in three computational phases that could solve large instances. In contrast, their optimization model keeps track of each vehicle and solves all the decisions using penalization factors in the objective function. These penalization factors cannot be naturally defined from a practice perspective and easily lead the model to bad solutions. Furthermore, the complexity of the model makes its computational times inefficient for large instances. Thus, our model is simpler and manages the majority of their considerations with results good enough to be applied and even optimal for the location problem.

Our schemes are also completely different from those of Jagtenberg et al. (2015), although we incorporate the same concept of a single-location heuristic algorithm. Instead of the expected coverage as an optimization objective, we use the preparedness index. Furthermore, we consider relocation strategies.

Regarding the work of Bélanger et al. (2016), in their third strategy, each time a vehicle starts its work shift, completes a mission, or resumes its service after a break, they solve the DSM to find the best new location for idle ambulances. For this single location problem, instead, we use a heuristic algorithm, seeking a better computational time. In their fourth strategy, a mathematical model is solved whenever a vehicle is dispatched to a call or when a vehicle completes its mission, but only if the last relocation occurred more than a defined amount of time ago, and the complete coverage within a time threshold cannot be maintained. Instead, for those two cases, we use a single location algorithm and a relocation strategy triggered by the preparedness index. The model used in this relocation strategy is the original DSM, which is easier to solve than the modified version used by Bélanger et al. (2016).

Enayaty et al. (2018, a) also have similarities with our work. However, they built their matheuristic upon the Maximum Covering Location Problem, and they used this mathematical approach for both locations of newly idle ambulances as well as the relocation of idle ambulances. In contrast, we face the first situation with a heuristic algorithm and the second with a similar mathematical strategy while using the DSM as the base. The proposed approach is invoked at each event, causing a system state change. The authors evaluated the need for relocating implicitly each time this occurs. Nevertheless, we use a dispatch algorithm instead of dispatching the nearest server, do not assume that at most, one ambulance could be assigned to each location, and consider ambulances performing a relocation available for dispatch.

Finally, although Kergosien et al. (2014) did not intend to propose an ambulance management policy tool, when they designed their general simulation framework, the DSM was used as the model guiding location and relocation decisions through time. Thus, they implicitly provided an approach to apply mathematical models to online real-time ambulance fleet management. However, their approach differs with ours in the sense that i) they do not use preparedness as a relocation trigger; instead, if in a given time the demand zones could not be reached by an available ambulance within a time threshold, then it is triggered, ii) they treat a constraint's infeasibilities by using penalizations in the objective function while we progressivity increase the radii of service levels until a feasible solution is reached, iii) they consider neither heterogeneous fleet nor different types of emergencies, and iv) they always dispatch the nearest ambulance to emergencies.

The paper is organized in the following manner: Section 2 of the article details the proposed algorithm; Section 3 presents the simulation model; Section 4 details the computational results and discussion; and finally, conclusions are presented Section 5.

2. General Framework of the Proposed Real-Time Optimization Algorithm

The MST never stops working and never should. Hence, is impossible to think in an initial location of the fleet when this problem is looked at in practice. This is because the MST always has ambulances in different positions of its area of operation, even if a central control of the fleet or a decision support system do not exist. Thus, each ambulance in operation must be attending an emergency or providing coverage around the position where it is placed. In summary, from an online real-time perspective, there only exists single-location decisions of ambulances (one at a time as they become available for the system) and relocation decisions (which imply the movement of more than one ambulance). This means that a location decision must be done every time an ambulance report itself as available. For this reason, a criterion must be defined to give the orders of location in these cases. In some of them in which the response capacity of the MST is not the desired one, it could be better to perform a partial or total relocation of the fleet, but in other cases, such movements will be unnecessary, and a single location order for the new ambulance should be enough. However, reports of available ambulances could not only trigger single or generalized relocation orders. When ambulances become unavailable, the same thing could happen. For example, if an ambulance ends its shift or is dispatched to an emergency, then the area of operation of the ambulance could be covered, moving some of the ambulances of the fleet, but it is also possible that no movements would be necessary. Besides, constantly performing relocations of the entire fleet is, in practice, prohibited because it implies over-working the crew, chaos in the operation of the system, and more costs. Hence, an online real-time optimization approach for the management of ambulances must give support to locations, dispatch, and relocation decisions while constantly tracking the need and convenience of a relocation strategy. All of this takes into account the ability of different types of vehicles to handle different types of emergencies.

In order to solve this problem, we describe in this section a proposed matheuristic algorithm. This algorithm evaluates every single change in the system status, and based on this, determines i) how the dispatch decisions of the vehicles to emergencies must be done, taking into account different levels of priority, a heterogeneous fleet, and the future performance of the system, ii) the location decision of ambulances when they start their shift or become newly available (e.g., when they complete a service) and the conditions of the systems do not justify a general or partial relocation of the fleet, and iii) if a total or partial relocation of the fleet must be performed and how. In general, dispatch and single-location decisions, as well as the triggering of relocation, are driven by seeking the improvement of a new definition of the preparedness index of the system. While the relocation decisions rely on a two-step mathematical programming optimization procedure, the first is oriented to the maximization of double coverage (DSM), and the second is oriented to the minimization of the maximum time needed to reach the desired configuration defined in the first step.

In this section, the assumptions of the algorithm, the new preparedness definition, and the optimization models will be described. We will then present the matheuristic algorithm capable of supporting ambulance fleet management decisions in an online context where the ambulance fleet size, the demand, travel times, and other critical aspects can change according to time. In order to verify the functionality of this algorithm, it will be used to undertake the decision process within a discrete event simulation (see Sections 3 and 4). However, we want to clarify that this proposal is not an optimization approach embedded or enhanced by a simulation model. This kind of approach is an interesting research topic but is beyond the scope of this paper.

2.1 Assumptions, parameters and general considerations

This section elaborates on the assumptions behind our proposed modeling approach. The MST considers a heterogeneous fleet to respond to different types of emergencies. This means that restrictions on medical attention are generated for certain types of emergencies as well as for certain types of vehicles. Emergencies are classified into three types according to the "triage" classifications: triage 1 concerns life-threatening emergencies, and triages 2 and 3 concern less severe cases. Consequently, two types of ambulances, each equipped differently, are considered: Basic Care Transport (BCT) and Medical Care Transport (MCT). The difference between these two types of ambulances lies in the fact that while the MCT ambulance has a medical doctor as part of the crew as well as advanced life support equipment, the BCT ambulance has only paramedics as part of the crew. Within the given framework, for a specific time in which an event takes place, the following components and considerations of the system under study are proposed:

- Diverse emergency demands are fulfilled over time in a specific geographical area. Only emergencies due to accidents are considered; no response should be made for requests for transfer services or periodic home care.
- The system representation is simplified by dividing the geographical area into zones and by considering the centroids of these areas as the points of demand. This simplification is widely used in literature.
- Two radii of attention time are taken into consideration; some services must be covered within a time frame which must be less than or equal to r_1; others must be covered within a time frame which is less than or equal to r_2; r_1 and r_2 are given parameters.
- A single ambulance must receive two basic orders over time: i) location decisions, which order the vehicle to move to a given location where the vehicle must be parked (if it was already located, this movement is a relocation), providing coverage to the surrounding geographical area, and ii) dispatch decisions, which establish the type of ambulance to respond to a given emergency.
- For dispatchers: if an emergency call is received and there are no available vehicles, the algorithm generates no response. In this case, the algorithm operator must generate, control, and solve the service queues.
- It is considered that ambulances can be found in the system in any one of three different states at a given time: i) *Located*– when the ambulance is parked in a given location providing coverage to the

system, implying that it is available for location decisions (since it is already stationed, this is a relocation decision) and dispatch; ii) *Not available*–when an ambulance is unavailable for dispatch or location decisions because it is attending an assigned emergency, due for maintenance, or it is outside operating hours, etc., iii) *Available*–the ambulance acquires this temporary status when it enters into operation in a given moment of time; this can occur when the medical unit completes an assigned task or when it first begins its work shift. When the ambulance is in the state of availability, the vehicle is waiting for one of two orders: i) an order to cover a service, or ii) an order to move to another location. The *"Located"* or *"Available"* status allows for the use of the vehicle for dispatch, location, or relocation decisions.

- The displacement time between locations l and demand points d is considered by the matrix t_{ld}. These times are taken by the algorithm from the Google Realtime API. These times are not always the same even if consulted by the same geographic positions because they take into account the road network and traffic congestion based on GPS feedback. These accurate times are used only for the relocation strategy (mathematical models).

2.2 Preparedness Index

In ambulance logistics, the preparedness index is considered to be either a qualitative or a quantitative measure of the MST's response capacity to emergencies in a certain geographical area. Also, it has been recognized that the consideration of preparedness in ambulance dispatching can provide significant benefits in reducing response time (Lee, 2011). Thus, each zone or demand point *d* has a preparedness index P_d, which changes over time in accordance with the conditions of the system.

Andersson and Värbrand (2007) published one definition of this index, considering C_d, a weight that mirrors the demand for ambulances in the demand point; t_d^a, the travel time to point d for each considered ambulance a; and γ^a, a contribution factor for each vehicle considered. Based on these parameters, the preparedness of a given demand point P_d is expressed in (1). The authors used this index as a control variable for relocations, only triggering such a strategy when a zone reached a preparedness level lower than a minimum desired.

$$P_d = \frac{1}{C_d} \sum_{a=1}^{A_d} \frac{\gamma^a}{t_d^a} \tag{1}$$

In (1), A_d is the number of ambulances contributing to the preparedness of d, and the following conditions are stablished:

$$t_d^1 \leq t_d^2 \leq \cdots \leq t_d^{A_d}$$
$$\gamma^1 > \gamma^2 > \cdots > \gamma^{A_d}$$

As can be seen in (1), the earlier an ambulance can reach a zone, the better. However, the proposed approach uses the DSM as the representation of the relocation problem, which is based on the concept of coverage. Thus, we prefer a P_d definition showing that if d is covered or is not in a time threshold, than a value proportional to travel time. Moreover, the γ^a parameter needs to be tuned into each context in which the index is applied and does not have a natural definition from the operational context of ambulances. For these reasons, we propose here a new definition of the preparedness index solving the mentioned issues. Furthermore, it is based on queue theory, which gives aleatory considerations to the proposed algorithm, and with this, we seek better preparation for future events.

Specifically, we propose to use as the preparedness index the probability that a vehicle is available, which is derived from an M/M/n queue model but adapted to our considerations of different types of emergencies and ambulances. To our knowledge, the first time a related approach was used for ambulance management was in the simple location model proposed by Daskin (1983). The last is known as the Maximum Expected Covering Location Model (MEXCLP), and it incorporates the probability that the vehicle is not available when an emergency occurs, which is known in the literature as the busy fraction. Then, if the arrival rate of calls is λ, the average service rate is μ, and n is the number of ambulances that could reach the zone within a time threshold, then the probability that a vehicle is available is $p = \max\{0,1 - \lambda/n\mu\}$, where $\lambda/n\mu$ correspond to the busy fraction in MEXCLP and related models. The use of the maximum function is necessary because, in contrast with Daskin (1983), we do not assume that ambulances return to fixed bases of operation. Therefore, once an ambulance is dispatched, the queue system no longer has that server; thus, n decreases while λ keeps its value, allowing negative values if the maximum with zero is not introduced. The same happens if more emergencies are expected from a given zone (e.g., for different hours of the day); in this case, λ changes to a higher value than $n\mu$, dropping the value of the preparedness to zero.

As can be seen, our proposal has several advantages: i) implicitly, it has a covered or not-covered approach, which is more aligned with the DSM, ii) instead of using an external parameter to be tuned, it is based on already accepted concepts in the literature, iii) it is easily constructed from information usually available for operators (λ and μ), iv) it does not need to define a minimum preparedness level because the need for triggering a relocation strategy is implicitly detected when zero is reached by our proposed preparedness, and v) if the index is different from zero, it also gives a measure of how well a zone is covered, which is why it is also used in our algorithm (see Section 2.5) as an optimization objective for the single location and dispatch decisions.

Expressions (2) and (3) allow for the formal definition of our preparedness index P_d, based on the busy fraction of queue theory (M/M/n) and adapted for a framework of multiple types of demands:

$$\alpha_{ed} = \frac{\delta_{de}}{\sum_{e'} \delta_{de'}} \ \forall e, d \tag{2}$$

$$P_d \ = \ max\left\{0; 1 - \sum_{e} \alpha_{ed} \left(\frac{\lambda_{de}}{n_e \mu_e}\right)\right\} \ \forall d \tag{3}$$

In this case, n_e refers to the number of ambulances that can cover an emergency type e at a given demand point d within the standard time r established for e ($r1$ for triage 1, and $r2$ for triage 2 and 3); λ_{de} is the average arrival rate of emergency type e at demand point d; μ_e is the average service rate for emergency type e at demand point d; and α_{ed} is the weight of the demand for emergency type e at demand point d. The weight is calculated as the ratio between the demand of the zone d of an emergency type e (δ_{de}) and the total demand of a determined demand point d. An average time of attention by each ambulance for each type of emergency is considered. This time is measured beginning with the dispatch of a service to an ambulance until the required tasks have been completed and the ambulance is available for operation.

2.3 Double Coverage Optimization Model

The relocation optimization approach has two steps. The first is oriented to determine where the ambulance should be located, and the second is oriented to define how to implement the relocation process of the available vehicles. We did not combine both models into one because even using independent objective functions, it is still possible to find the best potential coverage. This is gained without having to weigh the importance of different objectives (coverage vs. relocation time) and at the same time with efficient computational times, which is critical in our online real-time approach. In the first step, the problem is the maximization of the double coverage, understood as the number of demands to be covered by the fleet in times less than r_1 and r_2, in which $r_2 > r_1$. The mathematical model of this problem is known as the Double Standard Model, which was proposed by Gendreau et al. (1997) but is adapted to handle different types of emergencies and ambulances. It was selected as the base of the relocation strategy of this paper because it has already led to many variants and extensions of the ambulance location problem, is inspired by different governmental rules, it is easily understood and can be adapted to many cases. The second step is a standard transportation model. It has, as input from the first optimization step, the excess and lack of ambulances in each location. Therefore, the problem in this step is to determine where to relocate the leftover ambulances to the locations with a lack of ambulances, thus minimizing the maximum travel time of the relocations.

The problem of location of medical emergency vehicles can be defined as an incomplete graph $G = (V \cup W, A)$ where V is the set of nodes which represents the demand points. W is a set of possible locations for ambulances, and $A = \{(i,j) \in V \times W, i \neq j\}$ is a set of arcs. For each arc (i,j), a displacement time t_{ij} is associated. The problem is to determine the quantity, type, and node $j \in W$ in which the ambulances should be located. The point of demand $i \in V$ is covered by the location $j \in W$, if and only if $t_{ij} \leq r$, where r is a standard coverage time. The main objective of this approach is to cover a demand totally or in part.

The mathematical notation of the coverage model is as follows:

Sets:

E Types of emergencies, $e \in E$
K Types of ambulances, $k \in K$
D Demand points, $d \in D$
L Locations for ambulances, $l \in L$

Initial Parameters:

t_{ld} Travel time between location $l \in L$ and the point of demand $d \in D$

Induced subsets:

KE_e Ambulance types $k \in K$, which can fulfill the requirements of a given emergency type $e \in E$
L_d^1 Locations l, which are able to cover to the point of demand d in a given time t_{ld}, which is less than or equal to r_1, $\{l \in L : t_{ld} \leq r_1\}$
L_d^2 Locations l, which are able to cover to the point of demand d in a given time t_{ld}, which is less than or equal to r_2, $\{l \in L : t_{ld} \leq r_2\}$

Parameters:

p_k Refers to the total quantity of operational ambulances type k $\in K$ that are available for relocation. Those ambulances with the status of "located" or "available" are considered to be operational
y_{lk}^p Refers to type k ambulances located (or assigned) to l in the previous state of the system
δ_{de} Demand of type e at point d
c_l Capacity of location l
α Proportion of the total demand that must be covered by an ambulance located within $r1$

Decision variables:

Binary

x_{de}^{1} 1 if demand point $d \in D$ for service e $\in E$ is covered by at least one vehicle in the time radius r_1 of time, otherwise 0.

x_{de}^{2} 1 if demand point $d \in D$ for service e $\in E$ is covered by at least two vehicles in the radius r_1 of time, otherwise 0.

Integers

y_{lk} Number of ambulances needed to locate in $l \in L$ of $k \in K$ type.

w_{lk} Number of ambulances needed to relocate in $l \in L$ of $k \in K$ type from other locations pertinent to L

z_{lk} Number of ambulances located in $l \in L$ of $k \in K$ type that can be relocated in other Locations pertaining to L

Objective Function: Maximize the double demand covered

$$\text{maximize Z} = \sum_{d \in D} \sum_{e \in E} \delta_{de}\, x_{de}^{2} \tag{4}$$

When a demand point is covered by at least two vehicles, x_{de}^{2} takes a value of one; in this way, the numerical value of the demand is added into the objective function. The general purpose is to establish better coverage of the most relevant demand points as allowed by the capacity of the system. Thus, if an ambulance becomes unavailable at some point in time, another ambulance could support a service as alternative option.

Constraints:

Each point of demand $d \in D$ must be covered at least once for every type of emergency $e \in E$ in less time than r_2:

$$\sum_{l \in L_d^2} \sum_{k \in KE_e} y_{lk} \geq 1 \ \forall d \in D, e \in E \tag{5}$$

At least a percentage α out of the total demand must be covered once in less time than r_1:

$$\sum_{d \in D} \sum_{e \in E} \delta_{de} x_{de}^{1} \geq \alpha \sum_{d \in D} \sum_{e \in E} \delta_{de} \tag{6}$$

A logical approach must be guaranteed between the coverage variables (double and single) and the location variables for each demand point $d \in D$ as well as for each type of emergency $e \in E$:

$$\sum_{l \in L_d^1} \sum_{k \in KE_e} y_{lk} \geq x_{de}^{1} + x_{de}^{2} \ \forall d \in D, e \in E \tag{7}$$

$$x_{de}^{1} \geq x_{de}^{2} \ \forall d \in D, e \in E \tag{8}$$

Only the fleet that is in operation can be located, and it must be located in its entirety:

$$\sum_{l \in L} y_{lk} = p_k \ \forall k \in K \tag{9}$$

There cannot be more than c_l ambulances located in each $l \in L$ location point:

$$\sum_{k \in K} y_{lk} \leq c_l \quad \forall l \in L \tag{10}$$

Once this model is run, the number of ambulances to be relocated must be determined in relation to the previous state of the system (y_{lk}^{p}). This is done using (10), which establishes the number of ambulances of type k that are required to be relocated in a given potential site l (w_{lk}) in order to reach the ideal state y_{lk}; at the same time, it establishes the origin of the transfer of those vehicles (z_{lk}).

$$y_{lk} - y_{lk}^{p} = w_{lk} - z_{lk} \quad \forall l \in L, k \in K \tag{11}$$

With this structure, the solutions of the model change according to the input parameters and by the state of the system just before running the model (current locations of the ambulances). Then, the solutions obtained are used as inputs for the mathematical model of transport, which will be explained in the subsequent section.

- *2.4 Transportation Optimization Model*

The mathematical notation of the transportation model is as follows:

Sets:

K Types of ambulances, $k \in K$
L Locations for ambulances, $l \in L$

Parameters:

w_{lk} Number of types of ambulances $k \in K$ to be relocated in $l \in L$ from other locations belonging to L (works on demand).
z_{lk} Number of types of ambulances $k \in K$ located in $l \in L$, which can be relocated to other locations belonging to L (functions as an offer).
t_{ij} Travel time from location $i \in L$ to location $j \in L$

Positive integer decision variables:

y_{ijk} The number of types of ambulances $k \in K$ to be sent from location $i \in L$ to location $j \in L$

Positive continuous auxiliary variable:

x Auxiliary variable to establish the objective function of the type minmax

Objective Function: Minimize the maximum transfer time

$$\min Y = x \tag{12}$$

The objective function is to minimize the maximum relocation time because the interest is to bring the system as quickly as possible to a stable level of the preparedness index. In this way, the ambulances to be relocated do not necessarily move to the nearest location l required since the objective is the stability of the system as a whole.

Constraints:

All relocation needs must be met:

$$\sum_{i \in L} y_{ijk} = w_{jk} \;\; \forall j \in L, k \in K \tag{13}$$

The movements of the ambulances should be carried out from the $l \in L$ locations showing an excess number of ambulances according to the coverage model, thus making it impossible to mobilize more than the number of ambulances available for relocation:

$$\sum_{j \in L} y_{ijk} \leq z_{ik} \;\; \forall i \in L, k \in K \tag{14}$$

The minimization of the maximum transport time of all ambulances to be relocated must be guaranteed:

$$t_{ij} * y_{ijk} \leq x \;\; \forall i \in L, j \in L, k \in K \tag{15}$$

2.5 Description of the Proposed Online Matheuristic Algorithm

Matheuristics consist of heuristic algorithms that include interoperation of metaheuristics and mathematic programming techniques (Bélanger et al., 2015). In this section we present a matheuristic algorithm with a decomposition approach in which the problem is divided into smaller and simpler subproblems, and a specific strategy is sequentially applied to each subproblem. This matheuristic consists of three main steps: i) assign idle ambulances to locations, ii) dispatch ambulances to emergencies and iii) trigger relocation strategy if needed. The proposed algorithm is designed in terms of the considerations described in the previous sections and expecting to be activated each time one of three events occurs: emergency arrival, ambulance becomes unavailable (e.g., end of work shift), or ambulance becomes available. Each time one of those events occurs, the whole system information is processed, and the decisions of location and dispatch are returned to the fleet. We exploit the fact that these events occur at discrete times and the computational efficiency of the algorithm; thus, it can be assumed that they never happen at the same time. Furthermore, even if in practice, they could happen almost simultaneously is still possible to process one at a time. Alternatively, if an emergency requires more than one ambulance, they could be processed as two independent emergencies.

Step 1: Assign idle ambulances to locations

If the system has an idle ambulance, then it is located. This is the case of a specific location decision in which only one vehicle moves. The marginal contribution to the preparedness index is used as a criterion under the restrictions of long-distance displacements. The heuristic algorithm doing this is detailed in Procedure 1.

Procedure 1 – Heuristic Procedure for Single-Location Decisions of Idle Ambulances

```
$TotalInitialPreparedness$ ← compute Total Initial Preparedness of locations $l$ as $\sum_l P_l$
for each ambulance $a$ available and without location assigned do:
        for each location $l$ do:
            if $l$ has remaining capacity then
                    choose $l$ as candidate location for $a$
                    $NewTotalPreparedness$ ← compute New Total Preparedness of locations $l$ as $\sum_l P_l$
                    $MarginalContribution_l$ ← $NewTotalPreparedness - TotalInitialPreparedness$
        sort locations $l$ by $MarginalContribution$ in descending order
        for each location $l$ do:
            if assigning $l$ to $a$ does not imply a travel time > $r2$ then
                    assign $l$ to $a$
                    end procedure 1
        assign closer $l$ by travel time to $a$
end procedure 1
```

Step 2: Dispatch ambulances to emergencies

Each time a new emergency is generated, the dispatch problem is solved through a heuristic that takes into account the priority of the emergency as well as the impact on the preparedness of the system for future emergencies. The algorithm assigns vehicles by means of a heuristic based on the one proposed by Andersson et al. (2007). The proposed heuristic considers three different types of emergencies: for the most urgent cases, the nearest ambulance capable of treating the emergency is always assigned; regarding the other two types, the proposed heuristic performs the dispatch in such a way as to guarantee that the level of the preparedness index of the system is least impacted. This is done using an exhaustive computation of marginal impacts in the preparedness of the system. The detailed procedure is explained in Procedure 2.

Procedure 2 – Heuristic Procedure for Ambulances Dispatching Considering Priorities and Heterogeneous Fleet

```
$TotalInitialPreparedness$ ← Compute Total Initial Preparedness of locations $l$ as $\sum_l P_l$
for each emergency $i$ do:
    for each available or located ambulance $a \in A$ do:
        $TravelTime_a$ ← Compute Travel Time from $a$ to $i$
    if $A$ = {} then end procedure 2
    if $i$ is a triage 1 emergency then
        dispatch to $i$ the nearest $a$ by $TravelTime_a$ and of type MCT
        end procedure 2
        if $i$ does not have an $a$ assigned then
            dispatch to $i$ the nearest $a$ by $TravelTime_a$ and of type BCT
            end procedure 2
    if $i$ is a triage 2 or 3 emergency then
        for each available or located ambulance $a \in A$ do:
            choose $a$ as candidate ambulance to be dispatched to $i$
            $NewTotalPreparedness$ ← Compute New Total Preparedness of locations $l$ as $\sum_l P_l$
            $MarginalImpact_a$ ← $TotalInitialPreparedness - NewTotalPreparedness$
        sort ambulances $a$ by $MarginalImpact_a$ in ascending order
        for each ambulance $a$ do:
            if assigning $a$ to $i$ does not imply a travel time > $r2$ then
                assign $a$ to $i$
                end procedure 2
        assign closer $a$ by $TravelTime_a$ to $i$
end procedure 2
```

Step 3: Trigger relocation strategy if needed

The assignment of ambulances to a call may affect the preparedness level in some zones, which is why it is checked if the level of a certain location has dropped to zero. If it has, then a total or partial relocation of the fleet is needed. This subproblem is solved in two stages. In the first stage, a mathematical model oriented towards the maximization of coverage is solved; in the second stage, the maximum time of displacement of the vehicles for relocation is minimized. Both models were previously described, the first as a double coverage model (DSM) and the second as a transportation model (TM). They are solved using exact optimization algorithms (using Cplex 12.5). However, their structure could lead to infeasible solutions if not enough ambulances are in operation; therefore, an incrementing parameter θ is used to progressively increase the time thresholds r_1 and r_2 until a feasible solution is found. The detailed procedure is explained in Procedure 3.

Procedure 3 – Optimization Strategy for the Relocation Problem

```
for each location $l$ do:
    compute the Preparedness $P_l$
    if $P_l$ = 0 then
        do
            optimize DSM
            if DSM does not have a feasible solution then
                $r_1 \leftarrow r_1\theta$
                $r_2 \leftarrow r_2\theta$
        until DSM has a feasible solution
        optimize TM
end procedure 3
```

3. General Framework of the Discrete Event Simulation Model

3.1 General aspects of the system's operation

The operating process that is simulated responds to the structure of events. In the case of the arrival of an emergency to the system, the process illustrated in Figure 1 is followed.

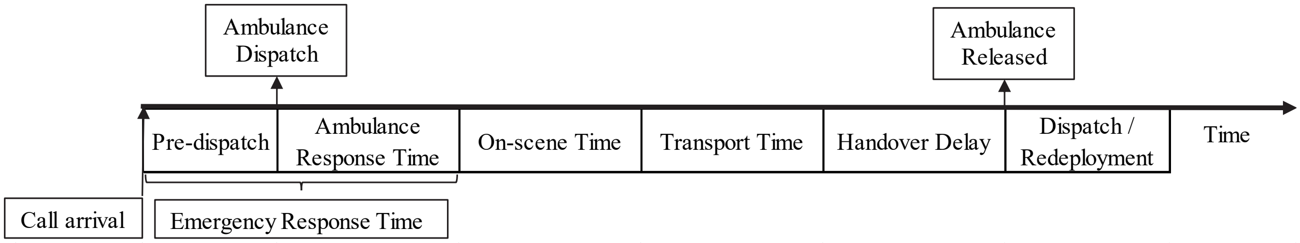


Fig. 1. Operation time of an ambulance

First, the need for an ambulance is reported to the emergency medical system, and the emergency data is verified; then, the ambulance is dispatched. In this first filter, it is typical that a percentage of the alerts received by the system are false or do not merit the dispatch of an ambulance. Several authors have considered this aspect within the simulation models.

When the decision to dispatch is performed, the vehicle operation of the vehicle time begins. The interval of time that the ambulance takes from the moment of dispatch to the place of the event is the response time of the ambulance. The sum of the time from the moment of dispatch to the response time of the ambulance is what the patient perceives as the response time to the emergency. For this research, the pre-clearance process is beyond reach; thus, for the model, a fixed time is assumed.

Once the ambulance reaches the site, an appropriate assessment must be made along with the required medical attention. When doing this, the emergency may not require a transfer to a medical entity, or the patient might not accept the service. In this case, the ambulance is available for dispatch to another emergency, or it can be relocated at some point. If a transfer is required, the entity to which the patient should be sent must be selected. At that time, it should also be taken into consideration that for the context of the application of the algorithm, the nearest service provider entity is not always available.

When the patient is successfully transferred to the appropriate medical entity, the correct procedures and protocol must be followed when the patient is given over to the care of the medical entity (including post-disinfection protocols for the crew and ambulance). The time that takes may vary considerably depending on the type of medical condition of the patient. Once the delivery is completed, the ambulance is available for operation; it can then be sent to attend another emergency or relocated to a base.

3.2 Simulation Model

In order to validate the different fleet management strategies in the literature, discrete event simulation has been widely used. In fact, flexible and generic models have been developed in the works of Kergosien et al. (2015), Pinto et al. (2015), and McCormack and Coates (2015). The system is represented with a set of data structures that are lists containing the system's elements and their corresponding characteristics. The lists of elements that make up the systems and their characteristics (in parenthesis) are: i) locations (id, geographic position, and capacity), ii) demand points (id, geographic position and for each type of emergency: demand and expected frequency), iii) ambulances (id, geographic position, type, state, and location assigned), iv) entities providing health services (id and geographic positioning), and v) emergencies (id, geographic position, priority, and ambulance assigned). Following the general architectures of the literature, the proposed simulation model was designed with four main components: a travel and service time estimator, a geospatial estimator, an optimization module (proposed algorithm), and a simulation engine. It also contains two databases: an input list of emergencies and a database that includes all the parameters needed to adequately describe the system. The input list could be either a simulated or a historical data file containing information on historical calls or requests. The service time estimator uses pseudo-random numbers to sample distribution probabilities of travel times and specific operation times (on-scene times and handover times). Each of its elements is now briefly presented.

The travel and service time estimator is the first component of the simulation system. This module computes realistic travel times between ambulances and emergencies, locations, and health entities. In order to produce these estimations, our component multiplies three factors. The geographic distances (16) coming from the formula of the large circle between two coordinates (both with latitude, *lat*, and longitude, *lon*) and two correction factors, one used to approximate the influence of the road network (ς) and its congestion and the other for the use of sirens (s). To model the effect of the road network, a conversion factor r between the geodesic distance and the time for displacement is calculated. To do this, the area under study is divided, and a set of centroids corresponding to those divisions is defined. Then, linear regression is performed for the times of displacement and the geodesic distances between those centroids. The times of displacement come from the Google API and correspond to the traffic condition of a determinate day and hour of operation. However, if needed, it an iteration could be performed of the process explained above to calculate the parameter r as a function of time. Besides, when ambulances are directed to a patient and when they transport the patient, they use the siren; thus, the speed of displacement is affected to a lesser extent by traffic conditions. This is modeled with a constant factor s which value is chosen between 0 and 1. However, the travel time of an ambulance to a location cannot be performed while using a siren; thus, in these cases, travel times are calculated by multiplying the geodetic distance between the ambulance and the location as well as the road correction factor. Nevertheless, the displacement times used by the mathematical models are captured directly from the travel times of the Google API between the coordinates of the centroids of the areas in which the geographical space is divided; this approach was not used for every travel time because it would be too expensive from a computational perspective.

Formula of the large circle, in this equation R represents the radius of the earth:

$$distance = R * acos[sen(lat1) * sen(lat2) + \cos(lat1) * \cos(lat2) * \cos(lon2 - lon1)] \quad (16)$$

Also, this component computes the service times of emergencies. For this, it feeds the simulator with pseudo-random data (from probability distributions which can be selected) that mainly simulate the stochastic process for the pre-dispatch time, the on-scene time, and the handover delay. In general, only triage 1 emergencies require patients to be transported to healthcare institutions (we assume that transport is always directed to the nearest institution); thus, triage 2 and 3 emergencies need on-scene attention only. However, two special cases are taken into account. In the first case, it occasionally happens that even though some patients require transfer (triage 1), a percentage of them do not accept it. This is because they prefer to avoid legal problems due to traffic fines, other pending legal issues, or for other reasons. Thus, there is a transfer only if the emergency is a triage 1 and the patient accepts the transfer. The latter is modeled with a Bernoulli variable. In the second case, the patient needs to be transported, but a probability of not being accepted in a medical entity; therefore, we include a Bernoulli variable modeling this aspect. If the closest entity cannot receive the patient, then the ambulance is sent to the nearest healthcare entity until an available one is found. Even if the latter sounds strange for first-world countries, we use in this research data from a country in which these cases happen more often than desired. We included it for a better representation of the system operation.

In summary, we compute two general service times. In the first, if the emergency is a triage 2 or 3, or is a triage 1 and the patient did not accept the transfer, the service time is equal to the pre-dispatch time, the travel time to the emergency (response time), and the on-site attention as defined in (17). In the second, the emergency needs to be transferred to a medical center. Thus, it is the ServiceTime1 plus the transport time to the healthcare institution that accepted the patient and the handover delay as defined in (18).

$$ServiceTime1 = PreDispatchTime + ResponseTime + OnSiteAttention \tag{17}$$

$$\begin{aligned} ServiceTime2 \\ &= PreDispatchTime + ResponseTime + OnSiteAttention \\ &+ TransportTime + HandoverDelay \end{aligned} \tag{18}$$

In the second component, named geospatial estimator, if in one event an ambulance has been ordered to move, and in a following event its geographical position changes, it is imperative that this aspect is taken into consideration since it critically affects the decisions made. For example, if an emergency takes place near an ambulance's position, even when it is in motion, it can be dispatched to attend the emergency. This aspect was modeled by projecting a geodesic line from the departure point of the ambulance to the point of arrival of the displacement, a line that has a total distance to travel calculated in kilometers. When required, the geographical position of the ambulance is recovered for the distance traveled from that line until the time of consultation. This was done using the library for geographic positioning operations GeographicLib 1.48. In order to determine the distance traveled so far, the percentage of the elapsed time is calculated against the expected total displacement, and the total distance of the geodesic line then multiplies that percentage.

In the third component, named Optimization Module, the whole system information is evaluated every time an event occurs, and the decision to be carried out by the operators are defined. In our case, decisions concerning the location, relocation, and dispatch of vehicles are made according to our algorithm.

The last component is the discrete event simulation engine. Basically, the simulator contains a list of events and a simulation clock that represents the simulated time. The event list is initially populated with historical emergency calls. Then, we rank the events and select the earliest one as the next simulation time point. Each time an event is processed, future events are created and added to the event list. We only consider two events other than emergency arrivals: the arrival of an ambulance to a health entity or emergency point (if the attention is given on-site), and the arrival of an ambulance at a given location in order to provide coverage. Thus, whenever the system clock points to a new emergency event, the simulation engine provides the optimization module with the system information. Then, an ambulance (or more) is dispatched, and its status attribute becomes unavailable (i.e., is no longer considered for dispatch) and an event of “arrival of an ambulance to a health entity or emergency point” is added to the event list. The time of occurrence of this event is defined using the travel and service time estimator module, which will be defined as the current time of the simulation plus the corresponding service time (randomly generated). Moreover, that dispatch could trigger a relocation strategy, a case in which the relocated ambulances changes its attributes of location assigned (to the new one) and status (from located to available). Thus, new events of “arrival of an ambulance at a given location” are added to the event list; the last takes into account the travel times computed by the travel and service time estimator module. However, the relocated ambulances, while moving, keep an available status. This means that they, as well as the located ambulances, could be dispatched to new emergencies, if required; in which case, the geospatial estimator updates these ambulance coordinates before the optimization module is run. Finally, if an ambulance ends its service or just enters into its working shift, it changes its status attribute (from unavailable to available), and the optimization module is activated. Therefore, one event “arrival of an ambulance at a given location” is created. Following the explained rules, the simulation iterates through the list of events, updating for each one of these the whole state of the system. This information is stored and used to generate the results described in Section 4.

Additionally, it is also assumed that if at a given time, there are no ambulances available to respond to an emergency, the model does not respond to it and it continues to iterate; this means that a waiting queue is not taken into consideration. The simulation model was implemented in C++. The verification of the simulation model has been performed through simulation test runs and consistency checks. However, we did not succeed in obtaining enough historical data and, therefore, we were not able to validate the simulator against real system behavior. Nonetheless, it serves its purpose in the framework of this paper, which is not to predict the performance of a specific system but to verify the functionality of our proposed algorithm.

4. Computational Results and Discussion

In this work, we sought to design an online real-time optimization algorithm for the management of an ambulance fleet. To the best of our knowledge, this is the first study proposing to solve this problem using a preparedness index defined as the available probability of a multi-server queuing model. Moreover, together with two mathematical models to solve the relocation problem and two heuristic algorithms taking into consideration different types of emergencies and vehicles.

The proposed algorithm is programmed in C++ and executed employing a computer with a GCC 5.3 compiler with a Linux CentOs operating system and Cplex 12.5. The data file format algorithm used is the JSON type (Java Script Object Notation) because it is susceptible to practical application given its wide use of quick information exchange in real-time execution computer systems. The efficiency of the algorithm was validated through the generation of a discrete event simulation model that represents the operational process of ambulances in the city of Bogotá, Colombia. The simulation model has been programmed on C++, and its objective is to respond, first of all, to the complexity of the characteristics that should be modeled, and secondly, to the algorithm's real-time execution structure. This is a characteristic which implies an iterative interaction between the state of the system and the algorithm, which is impossible to develop in commercial simulation software.

The first objective of the presented experiments is to verify the algorithm’s capability by assessing how the simulated performance of the algorithm corresponds to the expected one. The second objective of these experiments is to illustrate the advantages of our real-time approach compared to the sole application of the MILP ambulance location model already used by the algorithm. In particular, this fact shows that the algorithm works efficiently using the characteristics previously described. Moreover, it shows that taking a MILP model and modifying it to work in real time could lead to better results. However, we want to clarify that these experiments are not intended to compare or judge existing methods for the online real-time management of an ambulance fleet.

The experiments are structured as follows. Two configurations were implemented in the simulation model. One was for the proposed algorithm; other uses the DSM of Gendreau et al. (1997). We explore the behavior of the two approaches by simulating the attendance of 290 emergencies whose coordinates, time of occurrence, and type were taken from real data of the city of Bogotá, Colombia in which the time frames $r1$ and $r2$ were defined in 8 and 13 min. Specifically, the simulation was configured as follows:

An 8-hr shift (480 min) was simulated, assuming that the ambulance crew in operation is constant during this time. The number of ambulances in operation were set as 17 advanced ambulances and 111 basic ambulances. These ambulances were defined from the iterative running of simulations until the preparedness index showed a steady state in both configurations. The proportion between the two types of ambulances is approximately the same as the public fleet of ambulances in Bogotá.

The geographical space is represented by 112 areas, which correspond to the divisions created by the municipal administration for public management purposes and are named Units for Zone Planning (Figure 2). It is assumed that each of these areas is associated with a demand point as well as a location; the coordinates of the geometrical centroid of these areas are used to determine distances and travel times between them. Lastly, the locations' capacity was set to infinite, and there were identified and included in the simulation 122 medical entities capable of attending the emergencies.

The on-site attention time, as well as the hand-over time of the patient, was modeled based on theoretical probabilistic distributions. These were determined based on goodness-of-fit tests using the data of these operations. The on-site attention time was adjusted to a Weibull distribution (min: 0, alpha: 2.43, beta: 33.9) while the hand-over time was adjusted to a Uniform distribution (min: 0; max: 59), both of them in minutes. The probability that patients needing transport to a medical entity agree with that was set to 80.41% (in Colombia, some patients refuse transport); this percentage was prescribed based on historical data. Due to lack of data, the probability of a patient being accepted in a medical entity was set to 80%.

The transformation factor ς used to compute travel times between coordinates (e.g., ambulance and emergency) was set to 2.48 min/km (geodesic distance). This was defined by linear regression between the travel times from the Google Maps API and the geodesic distance between the 112 areas of Bogotá. The linear regression presented a coefficient of determination of 94.96%; thus, the use of ς as a means of travel time forecasting has a significant level of accuracy.

When ambulances are directed to a patient and transport the patient, they use the siren; thus, the speed of displacement is affected to a lesser extent by traffic conditions. This is modeled with a constant s factor that is chosen for the study at 0.9.

The study did not have the information on call times available; therefore, a constant time of 2 min was used as pre-dispatch time.

The times between emergency arrivals in the historical data were analyzed, finding (as had already been determined in similar cases by Pinto et al., 2015) that the arrivals follow a heterogeneous Poisson process. The frequency of arrivals per location λ used in the preparedness computations were adjusted based on this.

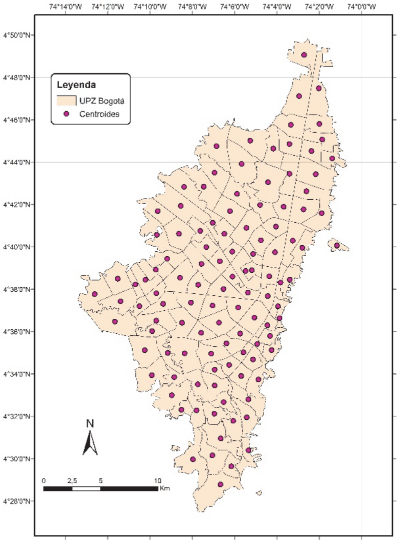


Fig. 2. Zonal Planning Units for the city of Bogotá
Data: Secretaría Distrital de Planeación de Bogotá

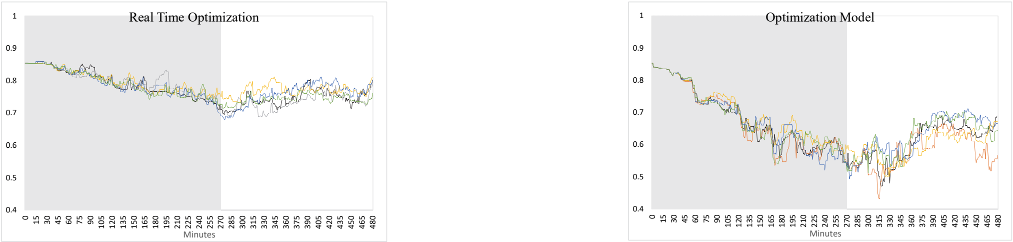
In order to verify the correct and logical behavior of the proposed algorithm and the simulation model, the sequence of events for ambulances and emergencies were rigorously traced on the complete history of all the events and decisions performed during several simulation experiments. In this exercise, a logical performance of the proposed algorithm was noticed. Emergencies were assigned to nearby ambulances; these were responded to in accordance with the proposed considerations, and the operation of the vehicles was coherent. Likewise, the geographical traceability of the ambulances was stable. From this, it was shown that the implementation is correct and the algorithm works accurately.

In order to verify the capabilities of the proposed algorithm, several performance measures were recorded over the simulation runs for the real-time and model-based optimization approaches. We report those measures considering the most relevant with respect to the experiments' objectives. Therefore, the following statistics on the system performance were compiled: i) average preparedness, ii) average response time (min), iii) average $r1$ coverage ratio, iv) average $r2$ coverage ratio, v) average number of relocations, vi) maximum number of ambulances moved per relocation, and vii) maximum algorithm computational time.

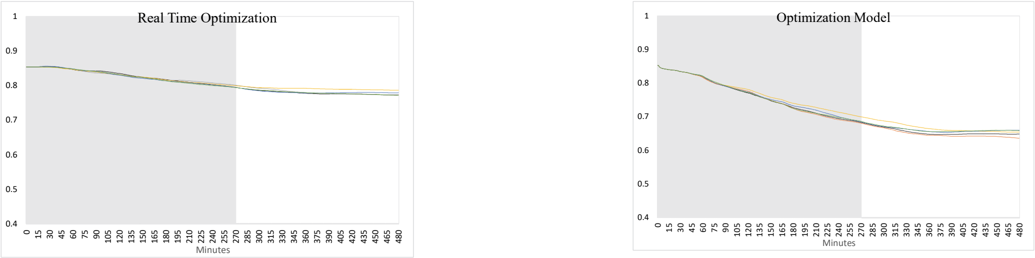
However, in order to perform the simulation runs, it was first necessary to set up an initial state of the system. This was done for both configurations by running the DSM and assuming that the ambulances were already located in the optimal locations indicated by the model. Thus, a transit state was expected; hence, the multiple replications method was used in order to define the time necessary to reach a steady state. As can be seen in Figure 2, the curves showing the preparedness evolution over time of five independent replications are noticeably different between the real-time optimization approach and the optimization model.

From this behavior, two things could be identified. The first one is that, from the perspective of the preparedness index, the real-time approach outperforms the sole application of the optimization model. The second one is that two stages were presented at the time of the simulation. In the first stage (shaded), the preparedness index is in decline. This is due to the fact that during the initialization of the model, most of the events are emergencies which are arriving. However, service termination and relocation events gradually occur, which leads the simulation to a second stage where the system reaches a stable state. Thus, an initialization behavior and one of the full conditions of behavior is clearly recognized. In accordance with this, a window of 270 min was defined as the warm-up period of the simulation. As a result, the performance measures were calculated, ignoring the warm-up period in order to obtain a more stable and representative behavior of the real system (Table 1). The results reported in Table 1 are based on 200 independent replications for each of the evaluated approaches. The computational time for each experiment of the simulation was approximately 6 min.

Let us first analyze the service level performance measures obtained with the algorithm. Notice that the average response time was 7.19 min with only 10.31% of emergencies attended outside the $r1$ and $r2$ time thresholds. The last, together with an average preparedness level of 75.48%, shows a relevant level of service. This comes with the results of the efficiency performance measures. As can be noticed, the algorithm handled in 210 min (simulation length without the warm-up period) an average of 425.5 algorithm calls in which the average number of relocations activated was 33.6. That means that less than 8% of the algorithm calls ended in a global relocation strategy, and even so, a relevant result in the service level was maintained. Although a reduction in the relocations was expected due to the incorporation in the algorithm of the preparedness index as a means to limit the triggering of multiple relocations (which is its purpose), the average of the maximum number of ambulances moved per fleet relocation was only 10.2 ambulances. Considering that the fleet had 128 ambulances in operation, this clearly shows an effective performance of the implemented relocation control strategy in the proposed algorithm: a new preparedness index definition working together with the heuristic algorithms for single location necessities (oriented to the maximization of the same preparedness index) and dispatch necessities.



System Preparedness Index of Five Independent Replications



Cumulative Average of the System Preparedness Index of Five Independent Replications

Fig. 3. Simulated Evolution of the System Preparedness of Five Independent Replications

Table 1
Results of 200 Independent Simulation Replications

Real-Time Optimization Algorithm									
	Preparedness [%]	Response Time [m]	$r1$ Ratio [%]	$r2$ Ratio [%]	$r3$ Ratio [%]	Number of relocations	Maximum ambulances moved per fleet-relocation	Algorithm calls	Maximum computational time [s]
Average	75.48	7.19	72.51	17.18	10.31	33.6	10.2	425.5	0.6
Desvest	1.11	1.67	9.13	3.68	7.32	6.3	3.1	5.9	0.1
Optimization Model									
Average	59.58	12.52	34.64	24.67	40.69	-	-	-	-
Desvest	2.03	2.02	6.91	4.04	8.62	-	-	-	-

As for the optimization model, a general deterioration of the service level compared with the algorithm is noticed. The response time worsened from an average of 7.19 min in the algorithm approach to 12.52 min, the preparedness decreased from an average of 75.48% to 59.58%, and the percentage of emergencies attended in less than $r1$ dramatically declined from 72.51% to 34.64%, which resulted in an average of 40.69% of emergencies attended outside the $r1$ and $r2$ time thresholds. As can be seen, the algorithm's results outperform those of the optimization model. However, we performed a paired-t test to corroborate the statistical difference between them, but more importantly, to define a measure of that difference. This exercise led to the 99% confidence intervals shown in Table 2. Thus, with 99% confidence, we can say that the preparedness, response time, and ratios obtained with the algorithm differs from those obtained from the optimization model (zero is not within the confidence intervals). Furthermore, it appears that using the algorithm is a superior policy, since it leads to a higher preparedness level (between 15.47 and 16.32 percentage points higher), lower average response time (between 4.85 and 5.82 min lower), and higher ratio of emergencies attended in less than $r1$ (between 35.73 and 40 percentage points higher).

Table 2
Results of 200 Independent Simulation Replications Paired t Test

Paired-t Test, Algorithm vs. Optimization Model				
	Preparedness [%]	Response Time [m]	r1 Ratio [%]	r2 Ratio [%]
Average	15.90	-5.39	37.87	-7.49
UB - 99%	16.32	-4.85	40.00	-6.42
LB - 99%	15.47	-5.82	35.73	-8.56

Although we do not intend to compare our approach in this initial study with existing methods to the real-time management of an ambulance fleet, it is a fact that our algorithm is based on several concepts developed by the works of Gendreau et al. (1997) and Andersson and Varbrand (2007). Regarding the DSM (Gendreau et al., 1997), the results explained above show that even if we use the same model for the location problem, the proposed algorithm displays a better performance. Regarding the results obtained by Andersson and Varbrand (2007), it was already known that a relocation strategy outperforms that of a static location approach, but we did not obtain this result with a large number of relocations as normally happens. A behavior mainly explained by the introduction of other definitions of the preparedness index together with the implementation of our proposed heuristic algorithms for single location necessities and dispatch.

Thus, the proposed algorithm could be implemented without using pre-planned relocations to obtain a lower number of relocations, as proposed by Andersson and Varbrand (2007). Moreover, the algorithm gave effective responses to an average of approximately 425 requests of dispatch, location, and relocation decisions in just 210 min of operation. This shows a frequency of requests not easily managed with precomputed solutions whose assumptions do not fit with the state of the system (such as pre-planned relocations and compliance tables). In this situation, the real-time strategy proposed in this research gains an applicability value, which is even greater nowadays when the algorithm can be connected with current GIS and information technology. The analysis of these results allows us to conclude that the proposed algorithm indeed succeeds in adequately managing in real time an ambulance fleet, handling the data volume representative of a city with a population of 7 million. Therefore, it successfully illustrates the usefulness of the proposed algorithm for real applications.

The results also seem to indicate that different preparedness indexes should be evaluated. Also, a decomposition analysis should be undertaken to determine the level of dominance of the different characteristics implemented in this algorithm in the performance results, and why such results are achieved.

5. Conclusions

The real-time management of ambulance fleets involves challenging decisions. This is mainly due to frequent, complex, and random changes in system conditions. This, with current developments of information technologies and GIS, gives the opportunity to develop online real-time optimization algorithms to help the decision-making process involved in ambulance fleet operations. In line with this, we aimed to design an algorithm capable of evaluating every single change in system status. Based on this, we aimed to first, determine necessities of taking location, dispatch, and relocation decisions, and second, compute their corresponding solutions, ensuring an adequate expected demand coverage while controlling the number of relocations. To our knowledge, this is the first study proposing an online real-time optimization algorithm that combines in one approach i) a new preparedness index defined as the availability probability of a multi-server queue model, ii) two mathematical models to solve the relocation problem (the DSM and a classical transportation model), and iii) two heuristic algorithms oriented to the maximization of the preparedness level, one to solve the dispatch problem (taking into consideration different types of emergencies and vehicles), and another to solve the location problem of one ambulance. The computational experiments have shown its capability of adequately responding to the necessities of a real-time operations. In summary, the proposed algorithm was found to be useful for real applications, and, therefore, provide an opportunity to improve the service level in medical transportation systems.

Acknowledgments

This work has been partially supported by Pontificia Universidad Javeriana Cali, Integra S.A. and by Pontificia Universidad Javeriana Bogotá. This support is gratefully acknowledged.

References

Andersson, T., & Värbrand, P. (2007). Decision support tools for ambulance dispatch and relocation. *Journal of the Operational Research Society*, *58*(2), 195-201.

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

Aringhieri, R., Bocca, S., Casciaro, L., & Duma, D. (2018). A simulation and online optimization approach for the real-time management of ambulances. In Proceedings of the 2018 Winter Simulation Conference (pp. 2554-2565). IEEE Press.

Bagherinejad, J., & Shoeib, M. (2018). Dynamic capacitated maximal covering location problem by considering dynamic capacity. *International Journal of Industrial Engineering Computations*, *9*(2), 249-264.

Başar, A., Çatay, B., & Ünlüyurt, T. (2012). A taxonomy for emergency service station location problem. *Optimization letters*, *6*(6), 1147-1160.

Bélanger, V., Ruiz, A., Soriano, P., & Lanzarone, E. (2015). The ambulance relocation and dispatching problem. CIRRELT.

Bélanger, V., Kergosien, Y., Ruiz, A., & Soriano, P. (2016). An empirical comparison of relocation strategies in real-time ambulance fleet management. *Computers & Industrial Engineering*, *94,* 216-229.

Bélanger, V., Ruiz, A., & Soriano, P. (2019). Recent optimization models and trends in location, relocation, and dispatching of emergency medical vehicles. *European Journal of Operational Research*, *272*(1), 1-23.

Daskin, M. S. (1983). A maximum expected covering location model: formulation, properties and heuristic solution. *Transportation science*, *17*(1), 48-70.

Enayati, S., Mayorga, M. E., Rajagopalan, H. K., & Saydam, C. (2018, a). Real-time ambulance redeployment approach to improve service coverage with fair and restricted workload for EMS providers. *Omega*, *79*, 67-80.

Enayati, S., Özaltın, O. Y., Mayorga, M. E., & Saydam, C. (2018, b). Ambulance redeployment and dispatching under uncertainty with personnel workload limitations. *IISE Transactions*, *50*(9), 777-788.

Enayati, S., Mayorga, M. E., Toro-Díaz, H., & Albert, L. A. (2019). Identifying trade-offs in equity and efficiency for simultaneously optimizing location and multipriority dispatch of ambulances. *International Transactions in Operational Research*, *26*(2), 415-438.

Gendreau, M., Laporte, G., & Semet, F. (1997). Solving an ambulance location model by tabu search. *Location Science*, *5*(2), 75-88.

Gendreau, M., Laporte, G., & Semet, F. (2001). A dynamic model and parallel tabu search heuristic for real-time ambulance relocation. *Parallel computing*, 27(12), 1641-1653.

Goldberg, J. B. (2004). Operations research models for the deployment of emergency services vehicles. *EMS management Journal*, *1*(1), 20-39.

Haghani, A., & Yang, S. (2007). Real-time emergency response fleet deployment: Concepts, systems, simulation & case studies. In Dynamic fleet management (pp. 133-162). Springer, Boston, MA.

Hakimi, S. L. (1964). Optimum locations of switching centers and the absolute centers and medians of a graph. *Operations research*, *12*(3), 450-459.

Jagtenberg, C. J., Bhulai, S., & Van der Mei, R. D. (2015). An efficient heuristic for real-time ambulance redeployment. *Operations Research for Health Care*, *4*, 27-35.

Karimi, A., Gendreau, M., & Verter, V. (2018). Performance Approximation of Emergency Service Systems with Priorities and Partial Backups. *Transportation Science*, *52*(5), 1235-1252.

Kergosien, Y., Bélanger, V., Soriano, P., Gendreau, M., & Ruiz, A. (2015). A generic and flexible simulation-based analysis tool for EMS management. *International Journal of Production Research*, *53*(24), 7299-7316.

Lee, S. (2011). The role of preparedness in ambulance dispatching. *Journal of the Operational Research Society*, *62*(10), 1888-1897.

Maxwell, M. S., Restrepo, M., Henderson, S. G., & Topaloglu, H. (2010). Approximate dynamic programming for ambulance redeployment. *INFORMS Journal on Computing*, *22*(2), 266-281.

Maxwell, M. S., Henderson, S. G., & Topaloglu, H. (2013). Tuning approximate dynamic programming policies for ambulance redeployment via direct search. *Stochastic Systems*, *3*(2), 322-361.

Maxwell, M. S., Ni, E. C., Tong, C., Henderson, S. G., Topaloglu, H., & Hunter, S. R. (2014). A bound on the performance of an optimal ambulance redeployment policy. *Operations Research*, *62*(5), 1014-1027.

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

Nasrollahzadeh, A. A., Khademi, A., & Mayorga, M. E. (2018). Real-time ambulance dispatching and relocation. *Manufacturing & Service Operations Management*, *20*(3), 467-480.

Pinto, L. R., Silva, P. M., et al. (2015). A generic method to develop simulation models for ambulance systems. *Simulation Modelling Practice and Theory*, *51*, 170-183.

Schmid, V. (2012). Solving the dynamic ambulance relocation and dispatching problem using approximate dynamic programming. *European Journal of Operational Research*, *219*(3), 611-621.

Sudtachat, K., Mayorga, M. E., & Mclay, L. A. (2016). A nested-compliance table policy for emergency medical service systems under relocation. *Omega*, *58*, 154-168.

Sung, I., & Lee, T. (2018). Scenario-based approach for the ambulance location problem with stochastic call arrivals under a dispatching policy. *Flexible Services and Manufacturing Journal*, *30*(1-2), 153-170.

van Barneveld, T. C., Bhulai, S., & van der Mei, R. D. (2016). The effect of ambulance relocations on the performance of ambulance service providers. *European Journal of Operational Research*, *252*(1), 257-269.

van Barneveld, T. C., Bhulai, S., & van der Mei, R. D. (2017, a). A dynamic ambulance management model for rural areas. *Health care Management Science*, *20*(2), 165-186.

van Barneveld, T. C., van der Mei, R. D., & Bhulai, S. (2017, b). Compliance tables for an EMS system with two types of medical response units. *Computers & Operations Research*, *80*, 68-81.

van Barneveld, T., Jagtenberg, C., Bhulai, S., & van der Mei, R. (2018). Real-time ambulance relocation: Assessing real-time redeployment strategies for ambulance relocation. *Socio-Economic Planning Sciences*, *62*, 129-142.

van den Berg, P. L., Fiskerstrand, P., Aardal, K., Einerkjær, J., Thoresen, T., & Røislien, J. (2019). Improving ambulance coverage in a mixed urban-rural region in Norway using mathematical modeling. *PloS one*, *14*(4), e0215385.