Locating and Dispatching Two Types of Ambulances Under Uncertainty and Partial Coverage

Beatriz A. García-Ramos[a], Roger Z. Ríos-Mercado[b], and Yasmin A. Ríos Solís[c]

[a]Graduate Program in Systems Engineering, Universidad Autonoma de Nuevo Leon, Mexico; [b]Graduate Program in Electrical Engineering, Universidad Autonoma de Nuevo Leon, Mexico ; [c]Science and Engineering School, Tecnologico de Monterrey, Mexico

ARTICLE HISTORY

Compiled August 15, 2025

ABSTRACT

This paper addresses an emergency vehicle covering and planning problem that arises from a real-world application. A limited number of two heterogeneous types of ambulances must be located at different city points and dispatched to emergency scenes, considering the uncertainty of the emergency. One of the main challenges is to determine whether an emergency can be fully covered on time, partially covered but with longer response times than ideal, partially covered with delays, or not covered at all. To this end, we use a gradual decay function to represent the partial coverage, within a two-stage integer stochastic program. To find solutions of good quality, we propose a location-allocation methodology that relies on the solution of an auxiliary surrogate model, which is faster to solve. Several aspects were evaluated in our empirical work. First, the benefit of introducing a partial coverage function is assessed, finding 84% fewer uncovered emergencies, which directly translates into saved lives. We also found that the proposed solution methodology produces solutions of very good quality significantly faster than the ones obtained when solving the original model.

KEYWORDS

Emergency medical services; Ambulance location; Stochastic integer programming; Location-allocation method

Corresponding author: Y. A. Ríos Solís. E-mail: *yasmin.riossolis@tec.mx*

1. Introduction

Emergency Medical Service (EMS) systems provide prehospital care to people who suffer a medical incident and transport patients to hospitals [5, 7, 40] for complete care. Given the presence of uncertainties in ambulance demand, often due to accident locations, degree of prehospital care needed, or calls' arrival times, emergency vehicle planning can be very challenging.

The problem addressed in this work is motivated by a real-world application in a developing country. EMS systems in developing countries, as is the case in Mexico, lack around 30-60%[1] the number of ambulances suggested by the World Health Organization (WHO), which should be at least four ambulances per 100,000 people [13]. For the Red Cross, an EMS operating with this small number of ambulances is considered similar to a war situation[1] or after a disaster incident [24]. Thus, the aim is to use the scarce ambulances in the best possible way.

Typically, in an EMS system, there are two types of decisions involved: ambulance location decisions (strategic) and ambulance dispatching decisions (operational). In the real world, these decisions must be made under uncertainty. We propose a two-stage integer stochastic programming model with recourse to maximize a weighted combination of total and partial coverage. In the first stage, ambulance locations must be determined, and in the second stage, when accident occurrence scenarios become known, ambulances are dispatched to the accident points.

Our model incorporates two novel features. First, we consider two different types of vehicles. The two most commonly used types of ambulances in EMS systems: Basic Life Support (BLS) units, typically staffed with two emergency medical technicians, and Advanced Life Support (ALS) units, which may include an emergency medical technician, an advanced emergency medical technician, and one or more paramedics. When a BLS ambulance is dispatched to an emergency that requires an ALS, it can reduce the patient's survival. Thus, we consider that ALS ambulances can be used as BLS units, but the contrary is not allowed [6]. Although most of the literature assumes a single type of ambulance, there are a few works that also consider two types

[1] Anonymous interviews done by the authors.

of ambulances [25, 34, 56].

A second feature incorporated in our model is the consideration of allowing partial covering. This is achieved by considering a gradual coverage decay function that have been used in the context of facility location [9], deterministic ambulance location [18], and fire service facility location [51], but have not been used in stochastic ambulance location problems, to the best of our knowledge. In our particular real-world setting, the number of available ambulances is very scarce, which implies that when ambulances are needed, accidents may not be fully covered. Therefore, it makes sense to consider the benefits of partial coverage to help customers get urgent care. In the ambulance location literature, a total coverage function is typically used. Naturally, dealing with a decay function is more computationally challenging [10, 59]; thus, it is not surprising that most of the work on stochastic ambulance location sorts to full coverage objective functions.

The proposed two-stage integer stochastic program primarily aims to determine ambulance tactical location decisions. We refer to our problem as the *Emergency Vehicle Covering and Planning* (EVCP) problem. As is the case with two-stage stochastic programming models, the second-stage decisions (ambulance deployment in this case) are informative to first-stage decisions because the value of the location decisions can be evaluated through the second-stage objective function values. Each ambulance has an average period-dependent response time to travel from its potential location to the demand point where the patient will be cared for [43, 53]. The location of the ambulance is crucial, as every minute of delay in treatment in a cardiac patient reduces the probability of survival by 24% [4, 38].

Similarly to Yoon et al. [56], to build the parameters for the second stage, we generate call-arrival scenarios by sampling emergency call logs to use in the second stage of our stochastic model. In this way, we address the volume of calls during a short period, such as on Friday nights [28]. Thus, time is not explicitly measured, and it is assumed that a vehicle can be assigned only once during this high-demand period for ambulances [61]. According to Yoon et al. [56], an important reason why sampling call logs is beneficial is the following. Sampling from the call logs does not make assumptions about the probability distributions. Queueing models used to describe EMS-system

performance often need assumptions about the distributions of accident occurrence, such as exponentially distributed interarrival times or service times. However, in many cases, these assumptions can be inconsistent with real-world observations.

To solve the model, we propose a novel location-allocation approach that relies on the solution of an auxiliary surrogate model that is faster to solve. Essentially, a location-allocation method is a two-step process in which location decisions are first determined. Then, the allocation variables are found by solving the problem with the location variables fixed. Location-allocation techniques have been used in other contexts [42, 50], but not in stochastic ambulance location to the best of our knowledge. An interesting feature of our proposed approach is that it can be implemented with relative ease by using off-the-shelf general-purpose solvers, without resorting to sophisticated decomposition algorithms. We name this approach *a surrogate-based feedback method* because the location of the ambulances obtained by this surrogate model is used as input to the original model. With this method, we obtain high-quality solutions in a reasonable time.

Summary of research contributions:

- We present a maximum expected coverage model considering partial coverage. To this end, we use a decay function to handle partial coverage. Although decay functions have been used in facility location problems, to the best of our knowledge, this is the first time they have been used in a stochastic ambulance location problem.
- We develop a surrogate model that is faster to solve, facilitating the solution of the problem.
- We propose a location-allocation method based on intelligent exploitation of the surrogate model solution.
- We present empirical evidence that shows: (i) the benefit of considering gradual coverage, (ii) the value of the stochastic solution, and (iii) the effectiveness of the proposed solution method.

The remainder of this paper is organized as follows. In Section 2, we review and discuss the literature related to our problem. Section 3 formally describes the EVCP

problem, explaining in detail the concept of gradual coverage and discussing the modeling assumptions. This includes introducing a two-stage integer stochastic programming model for the EVCP problem. The surrogate-based feedback method for the EVCP problem is described in Section 4. Experimental results on generated instances based on real-world cases that show the efficiency of our approach are given in Section 5. Final remarks and conclusions are drawn in Section 6.

2. Literature Review

There are excellent surveys and review papers discussing models and algorithms for ambulance location problems, such as the ones by Aringhieri et al. [5], Bélanger et al. [7], Brotcorne et al. [14]. Khelfa and Khennak [31] emphasize the need to develop methodological approaches that improve the deployment of ambulances. This involves optimizing decisions related to their placement, relocation, and dispatch to ensure the continued effectiveness and quality of EMS systems. Although early works dealt more with deterministic models, including backup coverage models [19, 27] and double standard models [17, 20, 22], given the inherent uncertainty of incident occurrence, stochastic approaches have gained more attention over the past 15-20 years. Some of these stochastic approaches are based on queueing systems and/or hypercube models [3, 25, 36, 49, 57], simulation studies [2, 26], sampling approaches [37], or dynamic systems [2, 4, 47, 48, 52].

In our work, we consider a two-stage stochastic programming model with recourse based on scenario generation. These types of models are useful for locating the ambulances in the first stage and then dispatching the vehicles to the accident points in the second stage. Yavari et al. [55] consider an ambulance dispatching and relocation problem taking into account overcrowding of emergency departments. Beraldi and Bruni [8] present a stochastic programming model for determining the optimal location of ambulances in congested emergency systems. Nickel et al. [37] present a stochastic programming model for minimizing the total cost for installing (and maintaining) ambulance location facilities but assuring a minimum coverage level. They develop a sampling approach in which they draw several samples of scenarios and solve the re-

stricted model associated with each of them. Gago-Carro et al. [21] present a two-stage stochastic programming model for ambulance relocation/allocation that balances the response time between densely populated and isolated areas. Sung and Lee [46] propose a scenario-based ambulance location model that explicitly computes the availability of ambulances with stochastic call arrivals under a dispatching policy. The model utilizes two-stage stochastic programming to represent the temporal variations in call arrivals as a set of call arrival sequences. Khosgehbari and Mirzapour Al-e Hashem [32] present a mixed-integer two-stage stochastic programming model that consider uncertainty on parameters such as emergency calls, travel times, and pathways, simultaneously. Other innovations of their model include considering a heterogeneous fleet of ambulances to provide specialized out-of-hospital services and considering different types of patients in terms of the need to be transferred to the hospital. To tackle this problem, they proposed a new decomposition-based heuristic method called the Progressive Estimating Algorithm (PEA). PEA is a modified version of the classic Progressive Hedging Algorithm. PEA attempts to deal with PHA drawbacks such as the possibility of being placed in a loop or prolonging the solution time by changing the method of calculating the first stage variables in each iteration. Therefore, by considering a large number of scenarios, PEA can reach feasible near-optimal solutions more efficiently.

An interesting feature and advantage of our surrogate-based feedback approach is that it can be implemented with relative ease by using off-the-shelf general-purpose solvers, without resorting to sophisticated decomposition algorithms as the ones discussed above. With our method, we obtain high-quality solutions in reasonable computing effort.

There are also studies based on robust optimization models. For instance, Ong et al. [39] present a robust optimization model to minimize the worst-case scenario of the location and dispatching problem. Yuan et al. [58] propose a distributionally robust chance-constrained programming model for an emergency medical system location problem with uncertain demands that minimizes the total expected cost by finding the location of emergency medical stations, the allocation of the ambulances and demand assignments. In the context of post-disaster relief logistics, Sun et al. [44, 45] propose a bi-objective robust optimization model for strategic and operational response to

decide the facility location, emergency resource allocation, and casualty transportation plans. Akıncılar and Akıncılar [1] address an ambulance location model considering uncertainty in vehicle speed under a robust optimization framework.

There are only a few works that we know of that deal with two or more different types of ambulances. For instance, McLay [34] address the problem of how to optimally locate and use two types of vehicles to improve patient survivability and coordinate multiple medical units with a hypercube queueing model. In Yoon et al. [56], two types of vehicles are considered, but one of them is a rapid vehicle that cannot offer the first care services of an ambulance. The authors present a two-stage stochastic-programming model that determines how to locate two types of ambulances in the first stage and dispatch them to prioritized emergency patients in the second stage after call-arrival scenarios are disclosed. In their model, they consider probabilistic travel times. They solved their model by means of a Benders-based algorithm. More recently, Nadar et al. [36] consider a joint ambulance location and dispatch problem for a multi-tier ambulance system. The proposed problem addresses three key decisions: the location of ambulance stations, the allocation of ambulances to these stations, and the preference order of stations for dispatching ambulances. Boujemaa et al. [12] present a robust stochastic programming location-allocation model to simultaneously determine the location of ambulance stations, the number and the type of ambulances to be deployed, and the demand areas served by each station in a two-tiered EMS system while accounting for the inherent uncertainty of the demand. In none of these works on multi-tier stochastic ambulance location, gradual partial coverage of the incidents is considered. This is a fundamental difference from our work.

To the best of our knowledge, no previous studies have comprehensively analyzed EMS systems that consider both multiple types of ambulances and partial coverage as we do in our work. This novel approach offers valuable insights for improving emergency medical services and response strategies in real-world scenarios.

3. The Emergency Vehicle Covering and Planning problem

Let us formally describe the EVCP problem. Let I be the set of possible demand points for patients who need medical attention in a city or region. This set can be very large, so we consider all the demand points observed in the historical data. In our case study, $|I|$ can be as large as 1500 demand points. Let L be the set of potential sites or ambulance stations where ambulances could be located, such as hospitals, firehouses, malls, or similar places where ambulances and paramedics can wait for emergency calls. We consider instances with up to 30 potential sites for the experimental results. Let K be the set of types of ambulances available in the system: the BLS (labeled with index $k = 1$) and the ALS ambulances (labeled with index $k = 2$), which are limited by a known parameter η_k for each type $k \in K$. These ambulances must be allocated to a potential site $l \in L$ and dispatched toward a demand point $i \in I$ if there is an emergency situation.

The travel time of any type of ambulance from a potential site $l \in L$ to a demand point $i \in I$ is given by r_{li}. While it is true that these travel times may be affected by many factors such as traffic conditions, within a specific time period where the model is applied, vehicle speed is more or less constant. Thus, we assume that these parameters are known following common practice in the literature. Ideally, ambulances should arrive in less than τ minutes in a life-threatening emergency. Usually, τ is a fixed value in the $[8, 15]$ minute range. This work also considers that the emergency is not covered if an ambulance takes more than a maximum time $\tau_{\max}$ to arrive. In this case, unfortunately, the accident has likely been handled by other means.

3.1. Description of partial coverage

One special feature incorporated in our model is the consideration of allowing partial covering. This is achieved by considering a gradual coverage decay function that have not been used in stochastic ambulance location problems, to the best of our knowledge. In our particular real-world setting, the number of available ambulances is very scarce, which implies that when ambulances are needed, accidents may not be fully covered. Therefore, it makes sense to consider the benefits of partial coverage to help customers

get urgent care. In the ambulance location literature, a total coverage function is typically used.

Since the EVCP problem aims to reduce the response time of the patient's first medical aid, even if it is in a partial or late way, we define a benefit decay function, based on the defined by Berman et al. [9], that only depends on the response time of a location $l \in L$ to any demand point $i \in I$:

$$c_{li} = \begin{cases} 1 & \text{if} \quad r_{li} \leq \tau, \\ 1 - \frac{r_{li}-\tau}{\tau_{\max}-\tau} & \text{if} \quad \tau < r_{li} < \tau_{\max}, \\ 0 & \text{if} \quad r_{li} \geq \tau_{\max}. \end{cases}$$

Thus, c_{li}=1 is the normal coverage definition if the ambulance can arrive in less than τ minutes. However, if the ambulance i in location l can arrive at the emergency in more than τ minutes but in less than τ_{max}, then it takes a decreasing value with respect to the number of minutes. That is, the farther the ambulance, the smaller the value of c_{li}. If an ambulance takes more than $\tau_{\max}$ minutes, then it is too far and is not considered to able to arrive to emergency i from location l.

3.2. Information related to the scenarios

The operational level is represented by a set of scenarios Ω with a bundle list of arriving calls. Each scenario $\omega \in \Omega$ represents a set of emergencies in the demand points. Thus, a scenario is defined by the number and type of ambulances needed at each demand point. Recall that an ALS ambulance can be sent instead of a BLS ambulance, but not vice versa. Thus, each scenario $s \in S$ indicates if there is an accident at a demand point $i \in I$ and provides the value a_{ki} related to the number of ambulances required of type $k \in K$.

For each scenario $\omega \in \Omega$, let $I(\omega) \subseteq I$ contain only the demand points $i \in I$ where ambulances are needed, that is, where $a_{ki}(\omega) \neq 0$ for any $k \in K$. We define five different types of ambulance coverage related to the response times for each demand point $i \in I(\omega)$:

- Total: the $a_{ki}(\omega)$ required ambulances of each type k are dispatched to i, and all

arrive in less than τ time.

- Total-late: the $a_{ki}(\omega)$ required ambulances of each type k are dispatched, but at least one arrives between $(\tau, \tau_{\max})$ time.
- Partial: at least one of the $a_{ki}(\omega)$ required ambulances is not dispatched, for $k \in K$, but all the dispatched ones arrive in less than τ time.
- Partial-late: at least one of the $a_{ki}(\omega)$ required ambulances is not dispatched, for $k \in K$, but at least one of the dispatched arrives between $(\tau, \tau_{\max})$ time.
- Null: none of the $a_{ki}(\omega)$ required ambulances arrives in less than $\tau_{\max}$ time, for $k \in K$.

Example of a scenario: Suppose for instance, that we have five demand points, i.e., $I = \{1, 2, \ldots, 5\}$. A scenario is determined by the realization of the random matrix $a = (a_{ki})$, which indicates whether an accident occurred and, if it did, the number of each type of ambulance required at the demand points. For instance

$$a = \begin{pmatrix} 0 & 1 & 0 & 2 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

indicates that there is no accident in demand points 1, 3, and 5, there is an accident in point 2, requiring 1 type 1 ambulance and none type 2 ambulances, and there is an accident in point 4 requiring 2 type 1 ambulances and 1 type 2 ambulances.

3.3. Illustrative example

Figure 1 illustrates a solution to the EVCP problem considered in this work. Five emergencies (red triangles) occur in the city during a rush hour period. There are three ALS ambulances (blue) and four BLS ambulances (dark blue) located in the city, which are dispatched to emergency situations. Emergency 1 has *total coverage* as it needs one ALS and one BLS that arrive within the ideal response time (green circle). Emergency 2 needs two BLS and one ALS. It has a *total-late coverage* since one ALS and one BLS arrive after the ideal response time (orange circles), while a BLS arrives within the ideal time. Emergency 3 needs one ALS and one BLS. It has a *partial coverage* since only the BLS can respond to the emergency within the

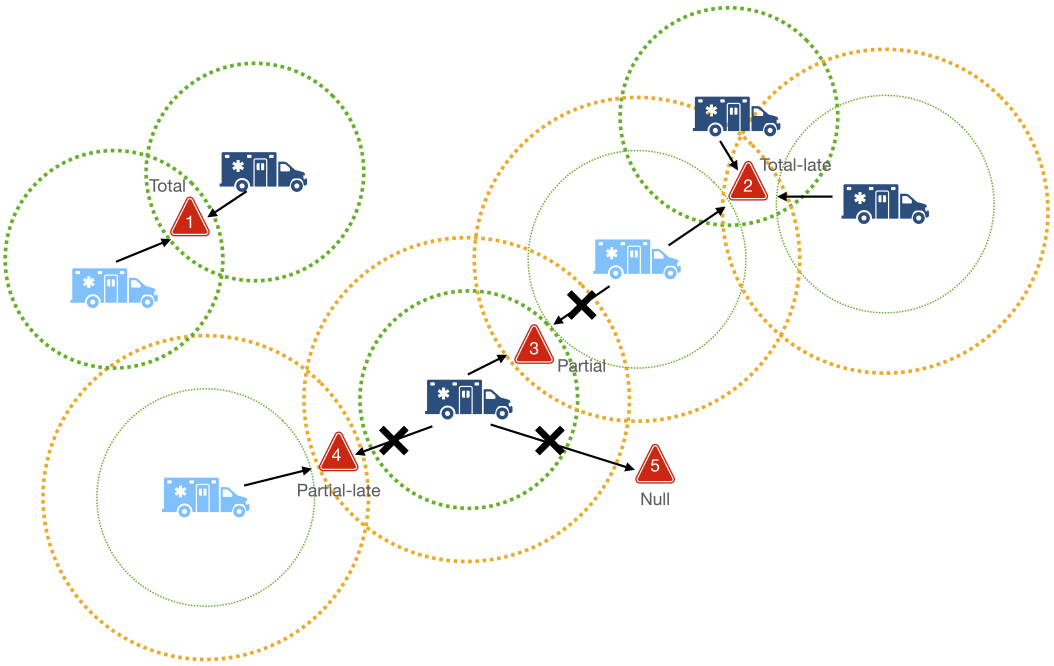


Figure 1. Five emergencies (red triangles), three ALS (blue) ambulances, and four BLS ones (dark blue). Emergency 1 has a *total coverage*; Emergency 2 has a *total-late coverage*; Emergency 3 has a *partial coverage*; Emergency 4 has a *partial-late coverage*; Emergency 5 has a *null coverage*.

ideal response time. Emergency 4 requires two BLS. It has a *partial-late coverage* because only one ALS (replacing a BLS) arrives, but with a longer response time than ideal. Unfortunately, Emergency 5 has a *null coverage* since the BLS ambulance that is required does not arrive within the maximum tolerated response time. Considering the number of ambulances available, their type and requirements, the coverage obtained by solving the EVCP problem is the best. Note that every emergency is treated as a whole event, and the solution tries to cover most of them, if not fully, at least partially, which in reality translates into saving lives.

3.4. Stochastic integer programming formulation for the EVCP problem

In this subsection, we present the Maximum Expected Coverage (MEC) formulation as a two-stage integer stochastic programming model.

The first-stage integer variables x_{lk} correspond to the number of ambulances of type $k \in K$ located at $l \in L$. The second-stage variables correspond to the ambulance

dispatch decisions at each demand point for each scenario $\omega \in \Omega$ as follows:

$$y_{lki}(\omega) = \begin{cases} 1 & \text{if an ambulance of type } k \in K \text{ in location } l \in L \text{ is dispatched} \\ & \text{to demand point } i \in I(\omega), \text{ for scenario } \omega \in \Omega, \\ 0 & \text{otherwise.} \end{cases}$$

We define the following binary variables related to the *total* and *total-late* coverages related to the response times of the ambulances to the demand point $i \in I(\omega), \omega \in \Omega$:

$$f_i(\omega) = \begin{cases} 1 & \text{if demand point } i \in I(\omega) \text{ has a } \textit{total} \text{ coverage,} \\ 0 & \text{otherwise,} \end{cases}$$

$$g_i(\omega) = \begin{cases} 1 & \text{if demand point } i \in I(\omega) \text{ has a } \textit{total-late} \text{ coverage,} \\ 0 & \text{otherwise.} \end{cases}$$

The following sets of binary variables are for the *partial* and *partial-late* coverages of the ambulances to the emergencies:

$$h_i(\omega) = \begin{cases} 1 & \text{if demand point } i \in I(\omega) \text{ has a } \textit{partial} \text{ coverage,} \\ 0 & \text{otherwise,} \end{cases}$$

$$w_i(\omega) = \begin{cases} 1 & \text{if demand point } i \in I(\omega) \text{ has a } \textit{partial-late} \text{ coverage,} \\ 0 & \text{otherwise.} \end{cases}$$

Finally, to indicate a null coverage of a demand point, we define:

$$z_i(\omega) = \begin{cases} 1 & \text{if active demand point } i \in I(\omega) \text{ has a } \textit{null} \text{ coverage,} \\ 0 & \text{otherwise.} \end{cases}$$

Let $\mathcal{Q}(x, a)$ denote the maximum coverage given decision x and random parameter array a. Given the notation introduced above, $a(\omega) = (a_{ki}(\omega))$, represents a vector of realizations of parameter array $a_{ki}(\omega)$ under scenario $\omega \in \Omega$. Thus, we aim to find x that maximizes the expected coverage. For simplicity, let $Q(q, \omega)$ denote the maximum coverage under the specific realization of scenario ω and let $\pi(\omega)$ the probability of

occurrence of scenario ω. In our work, we assume scenarios are equally likely, so $\pi(\omega) = 1/|\Omega|$ for all $\omega \in \Omega$.

The MEC formulation is as follows.

$$\max_{x} \mathbb{E}[\mathcal{Q}(x,a)] \tag{1}$$

$$\text{where} \quad \mathbb{E}[\mathcal{Q}(x,a)] = \sum_{\omega\in\Omega} \pi(\omega)\mathcal{Q}(x,\omega) \quad \text{and}$$

$$\mathcal{Q}(x,\omega) = \max_{(y,f,g,h,w,z)} \sum_{i\in I(\omega)} (\alpha_1 f_i(\omega) + \alpha_2 g_i(\omega) + \alpha_3 h_i(\omega) + \alpha_4 w_i(\omega) - \phi z_i(\omega))$$

$$\text{s.t.} \sum_{l\in L} x_{lk} \le \eta_k \qquad k \in K \tag{2}$$

$$\sum_{i\in I(\omega)} y_{lki}(\omega) \le x_{lk} \qquad l \in L, k \in K, \omega \in \Omega \tag{3}$$

$$\sum_{l\in L} y_{lki}(\omega) \le a_{ki} \qquad i \in I(w), k \in K, \omega \in \Omega \tag{4}$$

$$f_i(\omega) \sum_{k\in K} a_{ki}(\omega) \le \sum_{l\in L}\sum_{k\in K} c_{li} y_{lki}(\omega) \qquad i \in I(\omega), \omega \in \Omega \tag{5}$$

$$a_{2i}(\omega) f_i(\omega) \le \sum_{l\in L} c_{li} y_{l2i}(\omega) \qquad i \in I(\omega), \omega \in \Omega \tag{6}$$

$$g_i(\omega) \sum_{k\in K} a_{ki}(\omega) \le \sum_{l\in L}\sum_{k\in K} y_{lki}(\omega) \qquad i \in I(\omega), \omega \in \Omega \tag{7}$$

$$a_{2i}(\omega) g_i(\omega) \le \sum_{l\in L} y_{l2i}(\omega) \qquad i \in I(\omega), \omega \in \Omega \tag{8}$$

$$g_i(\omega) \le M \left(\sum_{l\in L} (1 - c_{li}) \sum_{k\in K} y_{lki}(\omega) \right) \qquad i \in I(\omega), \omega \in \Omega \tag{9}$$

$$h_i(\omega) \le \sum_{k\in K} a_{ki}(\omega) - \sum_{l\in L}\sum_{k\in K} y_{lki}(\omega) \qquad i \in I(\omega), \omega \in \Omega \tag{10}$$

$$h_i(\omega) \le a_{2i}(\omega) - \sum_{l\in L} y_{l2i}(\omega) \qquad i \in I(\omega), \omega \in \Omega \tag{11}$$

$$\sum_{l\in L} (h_i(\omega) - c_{li}) \sum_{k\in K} y_{lki}(\omega) \le 0 \qquad i \in I(\omega), \omega \in \Omega \tag{12}$$

$$w_i(\omega) \le \sum_{k\in K} a_{ki}(\omega) - \sum_{l\in L}\sum_{k\in K} y_{lki}(\omega) \qquad i \in I(\omega), \omega \in \Omega \tag{13}$$

$$w_i(\omega) \le a_{2i}(\omega) - \sum_{l\in L} y_{l2i}(\omega) \qquad i \in I(\omega), \omega \in \Omega \tag{14}$$

$$w_i(\omega) \le M \left(\sum_{l\in L} (1 - c_{li}) \sum_{k\in K} y_{lki}(\omega) \right) \qquad i \in I(\omega), \omega \in \Omega \tag{15}$$

$$\sum_{l\in L}\sum_{k\in K} y_{lki}(\omega) + z_i(\omega) \geq 1 \qquad i \in I(\omega), \omega \in \Omega \tag{16}$$

$$f_i(\omega) + g_i(\omega) + h_i(\omega) + w_i(\omega) + z_i(\omega) = 1 \qquad i \in I(\omega), \omega \in \Omega \tag{17}$$

$$x_{lk} \in \mathbb{Z}^+, y_{lki}(\omega) \in \{0,1\} \qquad l \in L, k \in K, i \in I(\omega), \omega \in \Omega \tag{18}$$

$$f_i(\omega), g_i(\omega), h_i(\omega), w_i(\omega), z_i(\omega) \in \{0,1\} \qquad i \in I(\omega), \omega \in \Omega \tag{19}$$

The objective function (1) maximizes the expected value of the weighted coverage of emergencies. The parameters $\alpha_1, \alpha_2, \alpha_3$ and α_4 are normalized weights that ponder the coverage type, and ϕ is a penalty for null coverage. We assume that every scenario is equally probable since each $\omega \in \Omega$ represents a sample of the high-demand period in which we are interested. Constraints (2) establish the number of ambulances available per type. Constraints (3) establish the relationship between the first and second-stage variables, which means that no ambulances can be dispatched from a potential site if no ambulances are located there. Constraints (4) bound the number of dispatched ambulances with the number of required ones for each emergency location.

The *total* coverage of an emergency is defined by constraints (5)–(6). We have in the first constraint that, if $f_i(\omega) = 1$ and is multiplied by the total number of ambulances of the same type needed at a demand point, then c_{li} is equal to 1. However, if at least one ambulance is late, i.e., if $c_{li} < 1$ for one or more ambulances, then $f_i(\omega) = 0$. In the second constraint, we guarantee that ALS ambulances can only cover demand points that need ALS ambulances. If $f_i(\omega) = 1$, then the number of ambulances dispatched of type 2 is not late, that is, $c_{li} = 1$. Otherwise, if the number of dispatched ambulances of this type with $c_{li} = 1$ is not sufficient to cover at least the ALS needed, then $f_i(\omega) = 0$.

The *total-late* coverage is defined by constraints (7)–(9). Constraints (7)–(8) allow the total-late coverage variables $g_i(\omega)$ to be one when the dispatch variables are active. In (7), $g_i(\omega) = 1$ multiplied by the total number of ambulances of both types needed, implying that all must be dispatched, but not necessarily in a time less than τ. Note that c_{li} could be less than 1, allowing for late coverage. If not all needed ambulances are dispatched, then $g_i(\omega) = 0$. Constraints (9) indicate that at least all needed ALS ambulances must be dispatched.

Meanwhile, constraints (9) track the demand points where the response time is between $(\tau, \tau_{\max})$ when the difference on the right-hand side of the equation is positive, that is, when there is a value $c_{li} < 1$ associated to a dispatched ambulance, for $l \in L, i \in I(\omega), \omega \in \Omega$. Note that, when $c_{li} = 1$ for all ambulances dispatched, then $g_i(\omega) = 0$ because that case implies a total coverage. Otherwise, the term multiplying M is a positive fractional number. The value of M introduced has to be large enough to make the constraint redundant in this case. In our testing, we observed that a value of $M = 1000$ suffices.

The *partial* coverage is defined by constraints (10)–(12). Recall that in this case not all needed ambulances are dispatched to the emergency, but the ones dispatched have an ideal response time. Thus, constraints (10)–(11) activate variables $h_i(\omega)$ if the number of ambulances dispatched is less than the required. In the first equation, $h_i(\omega) = 1$ implies that the difference between the needed and the dispatched ambulances is one or more. If the difference is zero, then $h_i(\omega)$ must be zero. For the second equation, we guarantee that the ambulances type 2 needed are covered only for ALS ambulances although not all must be dispatched. Quadratic constraints (12) ensure that ambulances dispatched arrive within the ideal response time, that is, their corresponding value $c_{li} = 1$, for $l \in L, i \in I(\omega), \omega \in \Omega$. If $h_i(\omega) = 1$ and the ambulances arrive at an ideal time, the difference between $h_i(\omega)$ and c_{li} is zero and guarantees that all ambulances must be dispatched at a time less than τ. If at least one ambulance is late, the difference would be positive and the constraint is infeasible because that case is a *partial-late* coverage.

Constraints (13)–(15) define the *partial-late* coverage. Note that the case of the ALS ambulances ($k = 2$) replacing a BLS one is also considered. Constraints (13)–(14) activate the variables $w_i(\omega)$ when the number of ambulances required exceeds the number of ambulances dispatched. Similarly to the total-late coverage, the constraints (15) track ambulances with a response time larger than the ideal and must be multiplied by a M since there could be cases with a sum that is less than 1. Here we also deal with the ALS ambulances that may replace the BLS.

The *null* coverage is activated by constraints (16). All coverage restrictions are related to the constraint (17) which ensures only one type of coverage for each emergency.

Finally, (18) and (19) establish the nature of the decision variables.

The novelty of the MEC model is the stochastic total/partial coverage per emergency by two types of ambulances. However, the related number of variables and constraints is usually large; moreover, the constraints (12) are quadratic. An integer linear stochastic model could easily be formulated with a classical linearization method. Still, previous experiments showed similar times between the linearized and the quadratically constrained models when solved with integer programming solvers, so we kept the quadratic one for the surrogate-based feedback methodology presented in the next section.

4. Surrogate-based feedback method for the EVCP problem

The EVCP problem is $\mathcal{NP}$-hard since the classical $\mathcal{NP}$-hard facility location problem [35] could be polynomially reduced to it. The MEC model is experimentally challenging to solve, even for medium-sized instances, as shown in Section 5. We propose a surrogate-based feedback method (SBFM) to obtain approximate solutions to the EVCP problem based on an auxiliary disaggregated model, named *Surrogate Ambulance-Based Coverage* (SABC) model, which is faster to solve. The main motivation for using this approach instead of traditional methods is its ease of implementation. Our approach can be implemented using any off-the-shelf general-purpose solver without the need to code complex decomposition-based techniques. This adds great value from a practical perspective.

In the original model, we are looking at maximizing the partial coverage. In the surrogate model, the objective function maximizes the expected value of the on-time and late dispatched ambulances minus a penalty for the required ambulances that could not be dispatched in less than the maximum response time.

Procedure 1 `SBFM( )`

Input: An instance to the problem
Output: x^*, a solution to the MEC problem

1: $x^*_{\text{SABC}} \leftarrow$ `solve(` SABC `)`
2: $x^* \leftarrow$ `solve(` MEC(x^*_{SABC}) `)`
3: **return** x^*

The surrogate-based feedback method is depicted in Algorithm 1. Under the SBFM, the SABC stochastic model is solved first. From its optimal solution, we obtain the location of the ambulances of the first stage corresponding to the value of x_{lk} variables, for $l \in L, k \in K$. Let the solution vector of these values be called x^*_{SABC}. Then, in the allocation stage, we solve MEC taking $\mathbf{x}^*_{\text{SABC}}$ as input. We call this feedback model MEC(SABC), implying that it is the solution of the MEC model with the location variables fixed with the solution of the surrogate model SABC. Since the first stage variables are fixed, the MEC(SABC) model becomes easier to solve and yields high-quality solutions. We could implement a local search neighborhood based on the location variables x_{lj} to diversify the solution yield by variables x^{SABC}. However, experimental results show that the quality of the SBFM solutions is exceptionally high with a single feedback.

Let us present the SABC surrogate model. In addition to the location variables x_{li}, the SABC model requires the following ambulance dispatching binary variables for $k \in K, l \in L, i \in I(s), \omega \in \Omega$:

$$u_{lki}(\omega) = \begin{cases} 1 & \text{if ambulance of type } k \text{ is dispatched from site } l \text{ to point } i \\ & \text{with response time less than } \tau, \\ 0 & \text{otherwise,} \end{cases}$$

$$v_{lki}(\omega) = \begin{cases} 1 & \text{if ambulance of type } k \text{ is dispatched from site } l \text{ to } i \\ & \text{with response time in } (\tau, \tau_{\max}), \\ 0 & \text{otherwise.} \end{cases}$$

Variables $u_{lki}(\omega)$ indicate the ambulances with an ideal response time dispatched from the location sites corresponding to a decay function value $c_{li} = 1$. While variables $v_{lki}(\omega)$ indicate the ones with a larger than τ response time, which have a value $c_{li} < 1$. The number of required ambulances k in an emergency demand point i that are not dispatched are counted by integer variables $\zeta_{ki}(\omega)$, for $k \in K, i \in I(\omega), \omega \in \Omega$.

Let $\mathcal{G}(x, a)$ denote the maximum expected value of the on-time and late dispatched ambulances minus a penalty for the required ambulances that could not be dispatched on time, given given decision x and random parameter array a. For simplicity, let

$G(x, \omega)$ denote the maximum coverage under the specific realization of scenario ω and let $\pi(\omega)$ the probability of occurrence of scenario ω. In our work, we assume scenarios are equally likely, so $\pi(\omega) = 1/|\Omega|$ for all $\omega \in \Omega$. Then SABC can be expressed as:

$$\max_{x} \mathbb{E}[\mathcal{G}(x, a)] \tag{20}$$

$$\text{where} \quad \mathbb{E}[\mathcal{G}(x, a)] = \sum_{\omega \in \Omega} \pi(\omega) \mathcal{G}(x, \omega) \quad \text{and}$$

$$\mathcal{G}(x, \omega) = \max_{u, v, \zeta} \left[\sum_{l \in L} \sum_{k \in K} \sum_{i \in I(\omega)} (\beta_1 u_{lki}(\omega) + \beta_2 v_{lki}(\omega)) - \sum_{k \in K} \sum_{i \in I(\omega)} \phi \zeta_{ki}(\omega) \right]$$

$$\text{s.t.} \sum_{l \in L} x_{lk} \leq \eta_k \qquad k \in K \tag{21}$$

$$\sum_{i \in I(\omega)} (u_{lki}(\omega) + v_{lki}(\omega)) \leq x_{lk} \qquad l \in L, k \in K, \omega \in \Omega \tag{22}$$

$$u_{lki}(\omega) \leq c_{li} \qquad l \in L, i \in I(s), k \in K, \omega \in \Omega \tag{23}$$

$$u_{lki}(\omega) + v_{lki}(\omega) \leq 1 \qquad l \in L, i \in I(s), k \in K, \omega \in \Omega \tag{24}$$

$$a_{1i}(\omega) + a_{2i}(\omega) = \sum_{l \in L} \sum_{k \in K} (u_{lki}(\omega) + v_{lki}(\omega) + \zeta_{ki}(\omega)) \qquad i \in I(\omega), \omega \in \Omega \tag{25}$$

$$a_{2i}(\omega) \leq \sum_{l \in L} (u_{l2i}(\omega) + v_{l2i}(\omega) + \zeta_{2i}(\omega)) \qquad i \in I(\omega), \omega \in \Omega \tag{26}$$

$$x_{lk}, \zeta_{ki}(\omega) \in \mathbb{Z}^+, u_{lki}(\omega), v_{lki}(\omega) \in \{0, 1\} \qquad l \in L, k \in K, i \in I(\omega), \omega \in \Omega \tag{27}$$

The objective function (20) maximizes the expected value of the on-time and late dispatched ambulances minus a penalty ϕ for the required ambulances that could not be dispatched in less than $\tau_{\max}$ time response. The weights β_1 and β_2 are normalized parameters that prioritize the ambulances dispatched with a response time less than τ. As in the previous model, no more than the available ambulances can be located on the sites, corresponding to constraints (21). The number of ambulances dispatched on time or late is less than the number of ambulances located, as indicated by constraints (22). Constraints (23) define the ambulances dispatched with an ideal response time of less than τ. Thus, if $c_{li} = 1$, then the ambulance will have an ideal response time, while

constraints (24) activate the late variables for which their response time is between $(\tau, \tau_{\max})$. With constraints (25) and (26), the non-covered emergencies, $\zeta_{ki}(\omega)$ variables are defined for $i \in I(\omega), \omega \in \Omega$. Recall that advanced ambulances can be dispatched instead of basic ones. Finally, the nature of the variables is stated.

The SABC model's essential characteristic is that its objective function does not rely on emergency coverage, as in the MEC model; it only counts the number of ambulances sent on time, late, or null to emergency demand points. Moreover, its resolution time is extremely fast since it requires fewer variables and constraints than the MEC model. However, disaggregating an emergency situation into the number of ambulances needed does not capture emergency coverage, which is crucial for an EMS system. Thus, the main idea is to infer high-quality locations for the ambulances with the SABC model, to fix these locations in the MEC model and to compute (allocate) the dispatching variables for the partial coverage by emergency.

5. Experimental Evaluation

This section presents an empirical assessment of models and the solution methodology previously described to solve the EVCP problem. We used Gurobi Optimizer 10.0.2 with Python 3.10 to solve the integer programming models MEC, SABC, and MEC(SABC). The experiments were carried out on an Intel Core i7 at 3.1 GHz with 16 GB of RAM under the macOS Catalina 10.15.7 operating system. Each execution of the integer linear programming solvers had a CPU time limit of 10800 seconds.

5.1. Instance generation

The value ranges of our instance generator are based on real-world data taken from Monterrey, NL, Mexico. In the literature, there are no suitable benchmarks for our problem. The databases for the Monterrey case study showed a larger number of possible demand points, $|I| \in \{168, 270, 500, 900, 1500\}$ compared to the one from the literature with $|I| \leq 270$ [56]. The number of possible locations for ambulances in Monterrey is $|L| \in \{16, 50, 100\}$, which is also larger than the one from the literature ($\leq$30) since not only hospitals and fire stations can be considered. We consider the whole city

of Monterrey, so the number of ambulances $(\eta_1, \eta_2) = (35, 20)$ is also greater than the ones from the literature cases (6 ambulances per type [56]). This approximate information was provided by the Centro Regulador de Urgencias Médicas (CRUM, Medical Emergency Regulatory Center). Recently, Monterrey acquired 30 new ambulances, increasing from 5 ambulances under the control of CRUM to 47 units. [16]. The value c_{li} is determined by a deterministic response time r_{li}, determined by CRUM, between the potential sites and the demand point, as we can see in literature [56]. The number of scenarios is set to be as large as that in the literature $|\Omega| \in \{10, 50, 100, 150, 200\}$. Thus, our benchmark has 15 instances for which five different scenario settings were built.

For each instance, we simulated a two-hour high-demand period. Each scenario $\omega \in \Omega$ consists of a set of demand values per ambulance type and per demand point $\{a_{ki}(\omega)\}_{k\in K, i\in I, \omega\in\Omega}$. Fewer demand points imply a larger city grid and a larger proportion of emergencies per demand point. Therefore, when $|I| = 168$, around 30% of the demand points may have a value different from 0. In contrast, when $|I| = 1500$, only 1% of the demand points will require ambulances. This setting reflects the number of emergencies per hour observed in the case study. Instances are built such that most emergencies require a single ambulance, but as observed in real cases, some of them may require up to three ambulances.

The ideal ambulance response time is $\tau = 10$ minutes, while the maximum response time is $\tau_{\max} = 30$ minutes. These parameters are not based on WHO guidelines, which suggest an ideal emergency response time of 8 minutes [15, 41]. We established those parameters based on CRUM information, which mentioned 9 minutes was their ideal response time before COVID-19, and 15 minutes in this pandemic. Also, the response time that CRUM considers appropriate for a patient who is not in immediate danger is approximately 25 minutes. With this information, we decided to establish those τ parameters. For the MEC formulation, we use the following weights in the objective function (1): $\alpha_1 = 0.65, \alpha_2 = 0.2, \alpha_3 = 0.1$, and $\alpha_4 = 0.05$. In this manner, the total coverage is the most sought-after, while the partial-late coverage has less benefit. The M value of the MEC model is only used to allow decimal values between [0,1] to trigger the activation of a binary variable. Thus, a simple value $M = 1000$ is set.

For the SABC objective function (20) we use $\beta_1 = 0.7$ and $\beta_2 = 0.3$. These values reflect the goal of sending the necessary ambulances primarily with an ideal response time. The penalty for null coverage in the MEC model or when a required ambulance cannot be dispatched to the emergency in less than $\tau_{\max}$ time in the SABC model is set to $\phi = 1/|\Omega| + 0.0005$.

All instances with related scenarios and detailed solutions are available at `https://doi.org/10.6084/m9.figshare.25928401`.

5.2. Assessment of benefit of partial coverage

In this experiment, we aim to assess the benefit of the proposed partial coverage model. To this end, we solve our MEC model with partial coverage ($\alpha_1 = 0.65, \alpha_2 = 0.2, \alpha_3 = 0.1$, and $\alpha_4 = 0.05$), and then we solve the Total-MEC model, which corresponds to the MEC model but with $\alpha_2 = \alpha_3 = \alpha_4 = 0$ in the objective function. The Total-MEC is equivalent to eliminating the partial covering terms from the MEC model. Thus, the objective function does not consider partial coverages, only the full and null coverage terms.

Table 1 displays the results. In the first column, the size of the instance is indicated in terms of the number of potential location sites, the number of demand points, and the number of scenarios. The second, third, and fourth columns show the objective function value, the running time (CPU seconds), and the value of the null coverage term in the objective function, for the MEC model (under partial coverage). The remaining columns show the same indicators for the Total-MEC model.

From Table 1, the most interesting result is the comparison of the null coverage term between both models. As can be seen, the Null coverage values obtained by the MEC model are considerably lower than those obtained by the Total-MEC model. This means that when ignoring the partial coverage terms, more people are left without coverage at all, which is, of course, not desirable. Moreover, we can also see that there were even some cases in which the null coverage term was zero under the MEC model, indicating the clear benefit of the partial coverage consideration. Overall, the MEC model improves by 84% on average the null coverage, which means more lives can be saved because at least an ambulance arrives at the emergency. In terms of

the location of the ambulances, when contrasting the MEC and Total-MEC solutions, there was more than 15% difference on average. The solution time was almost the same for both models, with the observation that for three instances, the MEC ran a lot faster. In summary, the most significant result is that the MEC model has a smaller average null coverage compared to the Total-MEC, indicating the substantial benefit of incorporating partial coverage terms into the objective function.

$\|L\|$, $\|I\|$, $\|\Omega\|$	MEC model			Total-MEC		
	Obj. fn. value	Time (CPU sec.)	Null coverage	Obj. fn. value	Time (CPU sec.)	Null coverage
16, 168, 10	4.9	94.8	0.00	7.0	2342.9	13.80
16, 168, 100	4.9	10804.2	0.07	7.6	10801.9	12.04
16, 168, 200	4.8	10808.7	0.14	7.5	10804.0	12.68
16, 500, 10	7.5	64.8	0.00	9.3	919.4	13.80
16, 500, 100	7.4	10809.8	1.69	10.6	10804.0	18.80
16, 500, 200	8.1	10810.9	1.93	11.3	10806.6	19.12
16, 1500, 10	9.9	48.1	5.90	11.6	10801.6	27.8
16, 1500, 100	10.6	10808.2	11.82	14.0	10806.2	26.67
16, 1500, 200	11.0	10838.3	13.56	14.4	10810.2	27.95
100, 168, 10	7.8	10802.6	0.00	7.2	10801.3	10.70
100, 168, 100	1.8	10817.4	0.00	7.5	10809.4	11.63
100, 168, 200	1.9	10843.4	0.24	7.2	10818.3	12.88
100, 500, 10	10.6	10804.2	1.70	9.8	10802.2	18.50
100, 500, 100	2.9	10831.0	1.83	10.7	10815.2	19.60
100, 500, 200	2.7	10905.8	4.36	9.7	10829.8	20.85
100, 1500, 10	13.2	10809.1	4.80	13.1	10803.6	19.60
100, 1500, 100	8.8	10849.6	13.18	14.6	10824.6	28.42
100, 1500, 200	4.2	10910.0	9.70	10.9	10848.9	34.43

Table 1.: Comparison between the MEC and Total-MEC models.

5.3. Sensitivity analysis

In this experiment, we conducted a sensitivity analysis of some model parameters that affect the objective values of the MEC model. We aim to investigate the model's sensitivity to the number of scenarios for various configurations of demand points and potential location sites, in terms of solution quality and solution time. We also want to determine the size of tractable instances when solving the MEC model, without the SBFM methodology.

To this end, we solved the MEC model directly by branch-and-bound based solver for the different configuration sizes discussed above. The results are presented in Figure 2, which consists of six plots. The three plots in the left-hand side column vary the number of demand points (x-axis), comparing each one to the value of the objective function when different scenarios are tested. The three plots on the right-hand side column vary the number of scenarios and show the variation in the solution value for each number of demand points. The upper, middle, and lower plots correspond to $|L| = 16$, $|L| = 50$, and $|L| = 100$, respectively. Straight lines correspond to the best objective values, and dotted lines are the best dual (upper) bounds found.

As can be seen from Figure 2, the difference between the best objective and the best bound (and thus, the relative optimality gaps) is negligible for small instances with 16 potential location sites. However, the relative optimality gaps become larger for instances with 50 and 100 potential sites. The number of demand points where emergencies may occur and the number of scenarios considered make instances harder to solve within the time limit. Thus, solving the MEC model directly with an integer programming solver allows to handle small instances with a few scenarios, demand points, and potential ambulance location sites. Note that the larger the number of scenarios in the plots on the left-hand side, the better the objective function. This implies that a more comprehensive sampling of emergency demand points improves the quality of the solution related to ambulance response times. The plots on the right-hand side show that the larger the number of demand points, the harder it is to solve the instance.

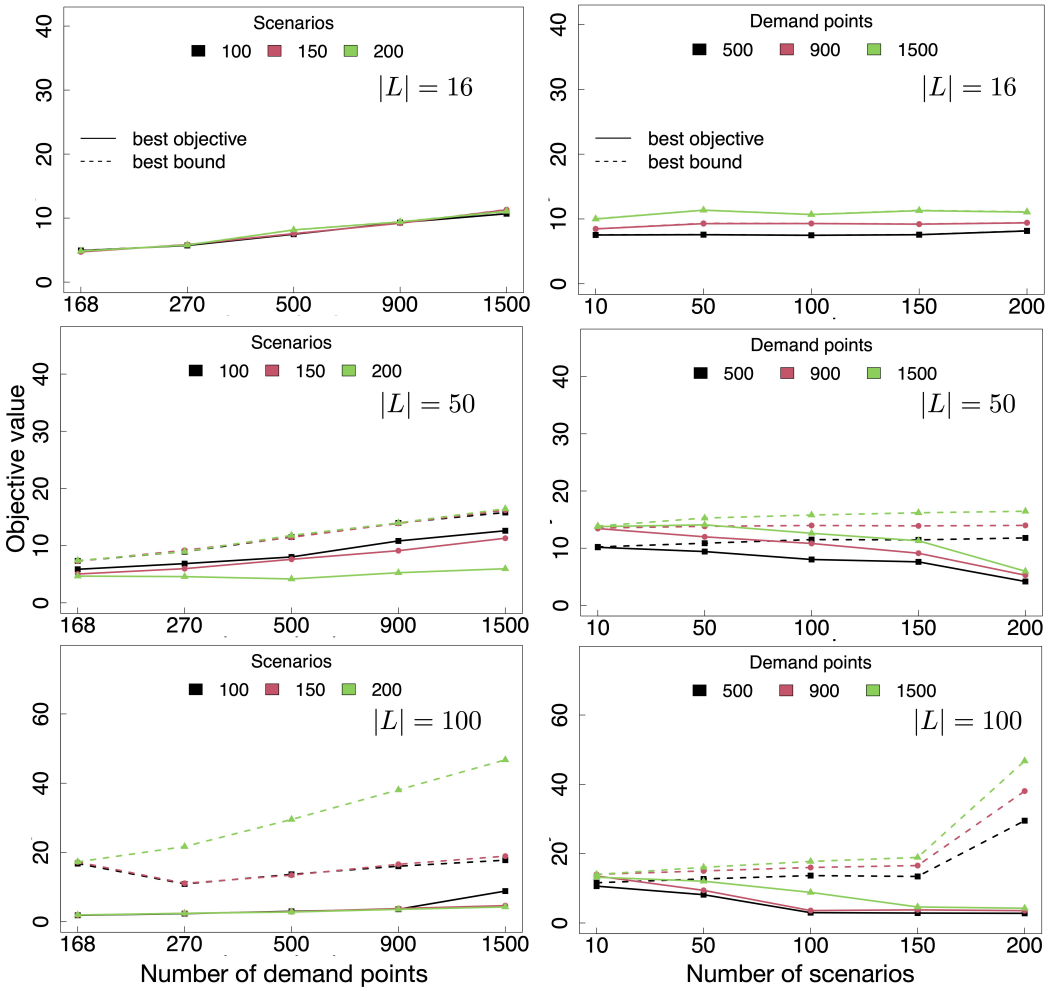


Figure 2. Best objective and the best upper bound of the objective function obtained by the MEC model as a function of $|I|$, $|L|$, and $|\Omega|$.

5.4. Assessment of the surrogate-based feedback method

Naturally, one of the most important aspects to investigate is the value and benefit that the proposed solution method brings to the table. Thus, in this set of experiments, we solved all instances for the different configurations previously discussed under two different methods. We solved the MEC model by directly applying the branch-and-bound method from the solver and compared it with the proposed SBFM.

In Figure 3, the legends indicate the different parameters tested. The only difference from the previous graphs is that the solid and dotted lines indicate the solutions obtained by MEC and SBFM, respectively. As can be seen, while the number of scenarios, demand points, and potential sites slightly affects the performance of SBFM, it

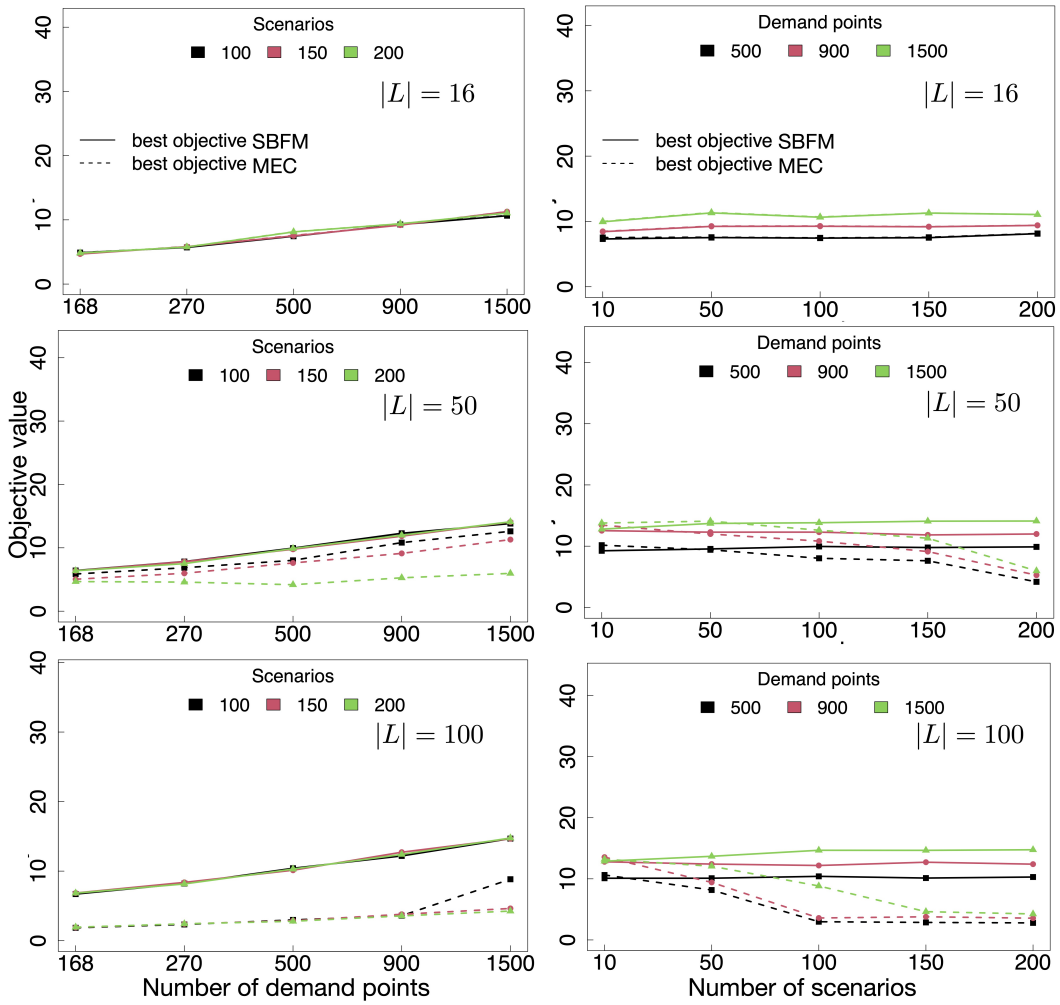
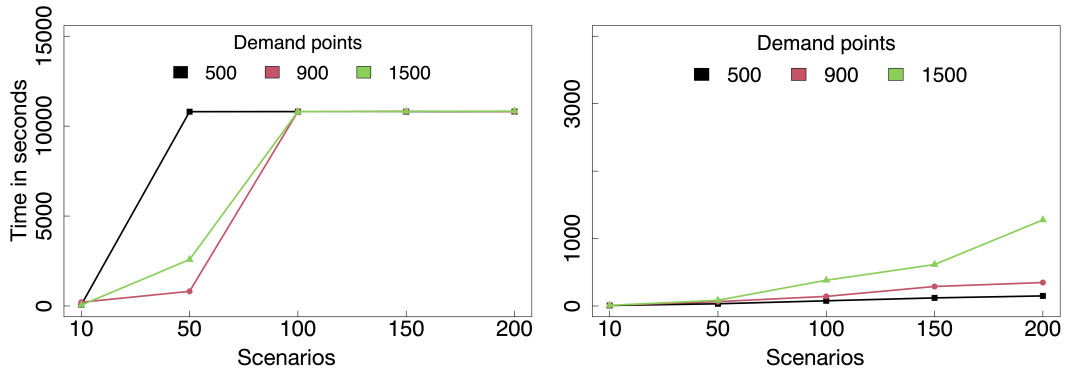


Figure 3. Comparison between MEC and SBFM.

obtains better objective function values than those obtained by the MEC for the larger instances that reported positive gaps. Indeed, the relative optimality gaps found by the SBFM are always equal to 0 within the allotted time limit. In addition, the SBFM tends to be less dependent on the number of scenarios. Thus, although we cannot guarantee optimality with the SBFM, it obtains faster and higher-quality solutions than those obtained by the MEC.

We now compare the running times of the MEC and SBFM. Recall that SBFM attempts to exploit that the surrogate model SABC is very tractable and can be solved relatively quickly. To this end, we used MEC to solve instances with $|L| = 16$ and SBFM to solve instances with $|L| = 16$ (harder instances). Figure 4 shows the



(a) MEC with $|L|$=*16*. (b) SBFM with $|L|$=*100*.

Figure 4. CPU time comparison between MEC and SBFM.

running times between these two methods. The x-axis of the plots corresponds to the number of scenarios and the y-axis corresponds to the running time (in CPU seconds). Recall that when we solved MEC with MEC for $|L| = \{50, 100\}$, it stops by the time limit even for ten scenarios and a few demand points.

As can be seen, the main disadvantage of the MEC is its computational time, which increases significantly with the number of demand points, potential sites, and scenarios, even for small instances with 16 potential location sites for ambulances. The SBFM is extremely fast, even for large instances, and yields an initial solution to the assignment of ambulance location in a short time. The SBFM location-allocation strategy inherits not only its fast computational time from solving the SABC model, but also yields coverage per emergency situation, which is the main objective for the EVCP problem. The SBFM is an approximate method, but it gives solutions that are as good as the MEC model and even better when the MEC instances do not reach optimality and its relative optimality gaps are large. The SBFM solved most of the instances in less than a minute.

An interesting advantage of the SBFM is that only one iteration is needed. In fact, once the location of the ambulances has been retrieved from the SABC model and fed back to the MEC model, we could perturb the ambulance location either randomly or with systematic local search, then re-locate the ambulances, and iterate again. We attempted to improve the solution by generating a neighborhood around a location

solution, to no avail. In other words, it was observed that, under the SBFM approach, local optima were often reached with the first feedback under the neighborhoods tested. An interesting follow-up could be to design more complex or diverse neighborhood structures to try to escape local optima.

We now proceed to compare the methods with respect to another important aspect. One of the most critical objectives of the EVCP problem is to cover the largest number of demand points within a fixed response time. Thus, it becomes relevant to assess the quality of the coverage in terms of its individual components.

Figure 5 displays the proportions of emergency coverage for all instances when solved by both methods. The left-hand (right-hand) side plots correspond to the solutions obtained by the MEC model and the right-hand one to the SBFM. Each plot shows the type of ambulance percentage coverage obtained: T stands for Total coverage (all required ambulances on time), TL stands for Total-late coverage (all required ambulances, but at least one arrives late), P stands for Partial coverage (at least one required ambulance is not dispatched, but the dispatched ones all arrive in time), PL stands for Partial-late coverage (at least one required ambulance is not dispatched, at least one of the dispatched arrives late), and N stands for Null (no ambulances assigned to the demand point). The upper, middle, lower plots are for $|L| = 16$, 50, and 100 potential sites, respectively.

Figure 5a, shows that the MEC tends to leave very few demand points with null coverage, which is the primary concern of the emergency services in our case study. As the number of potential sites $|L|$ increases, the coverage tends to be partial-late for the MEC. This behavior is probably related to the large relative optimality gaps obtained by MEC for large instances, but the number of null coverage is still remarkably low. Column b) shows that the SBFM is robust in terms of the number of scenarios. That is, the demand point coverage is independent of the number of scenarios. In this way, 100 scenarios are sufficient to handle a high-quality coverage solution. Moreover, the SBFM inherits the characteristic of having very few null demand point coverage from the MEC model. Interestingly, partial coverage tends to be larger than partial late coverage for the SBFM, which is mainly desired in real life because it can be translated into first-aid medical care on time, increasing the probability of saving lives.

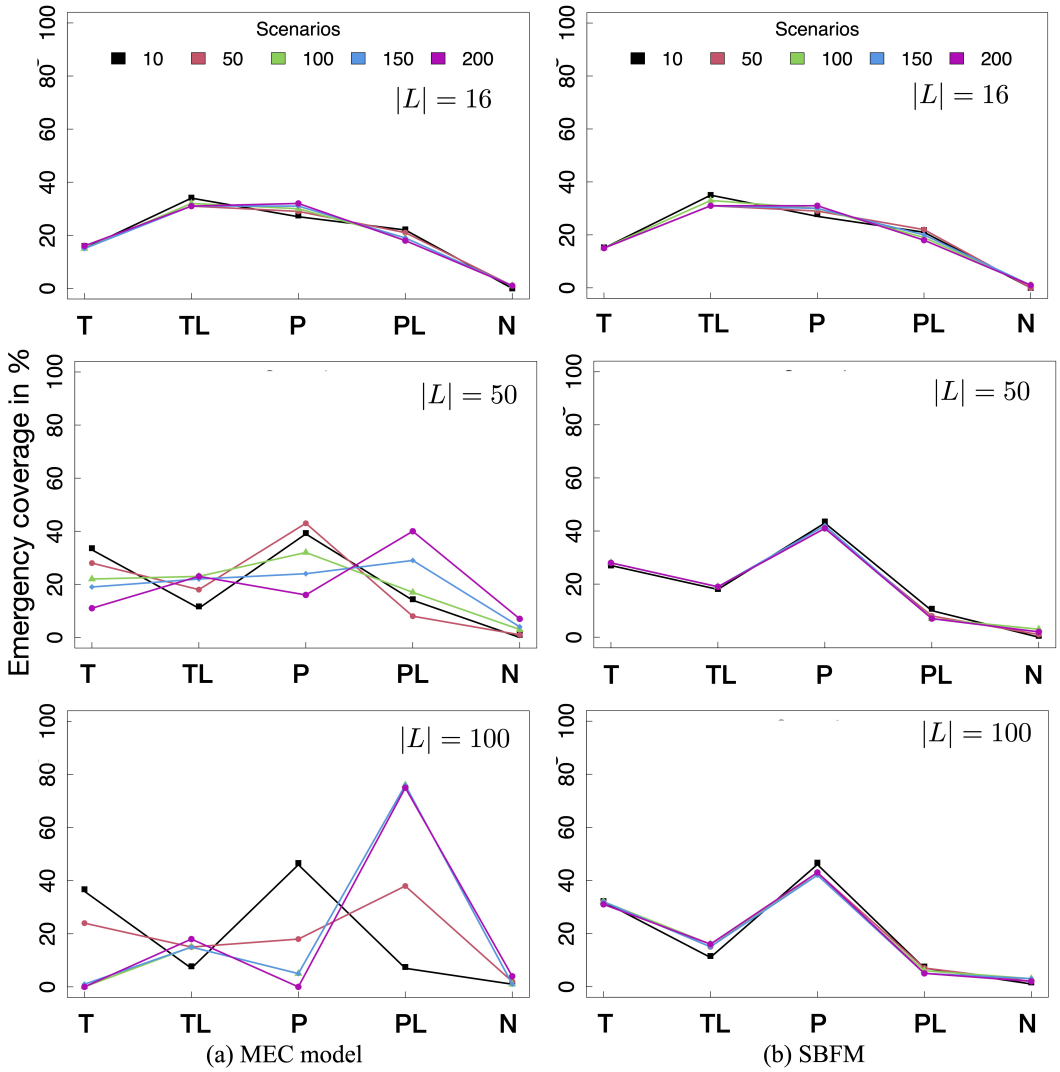
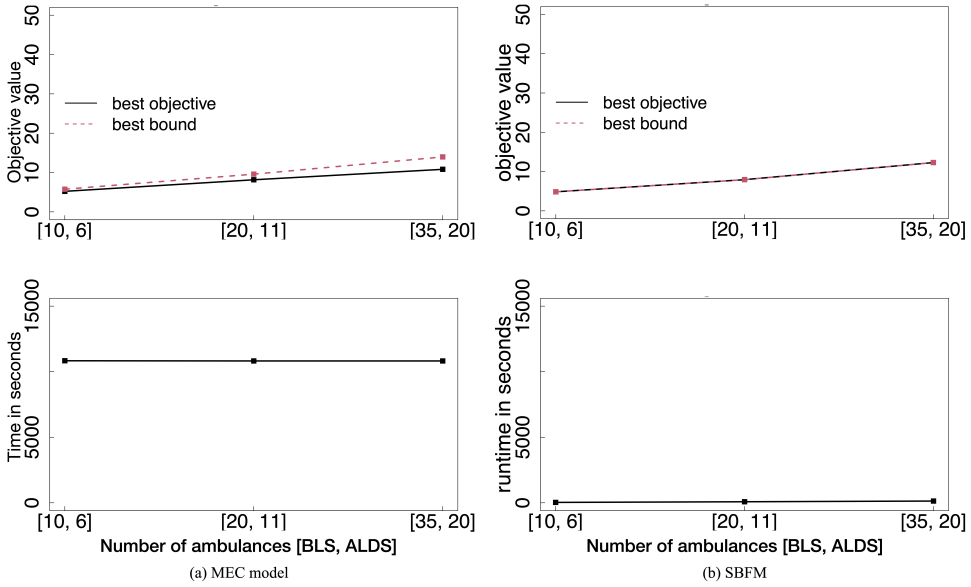


Figure 5. Comparison of MEC and SBFM in terms of disaggregated coverage.

All the previous experiments were carried out with the number of ambulances equal to $(\eta_1, \eta_2) = (35, 20)$. A central feature of the EVCP problem is that an ALS ambulance can be sent instead of a BLS one, which provides a more flexible setting but may introduce difficulty when solving the models. Now, we wish to investigate the effect of the number of available ambulances in the objective function value and the running time. To this end, we solved all the instances with 900 emergency demand points, 100 scenarios, and 50 ambulance location sites. For this experiment, we vary the number of ambulances. Figure 6 shows two columns of two plots each. The objective value (upper plots) and the running time (lower plots) are on the y-axis, while the x-axis



(a) MEC model (b) SBFM

Figure 6. Comparison between MEC and SBFM in terms of number of available ambulances.

varies the number of ambulances: $(\eta_1, \eta_2) \in \{(10, 6), (20, 11), (35, 20)\}$. The left-hand side plots correspond to the MEC model, while the ones on the right-hand side are for the model solved by the SBFM.

In Figure 6a, we observe that the difference between the best objective and the best bound for the MEC model (left-hand side plots) increases slightly with the number of ambulances. Thus, the larger the number of ambulances, the harder the instances for the MEC model. Furthermore, the time limit is reached for every tested instance of the MEC model. For the SBFM, the relative optimality gaps are equal to 0 for all instances. In addition, the objective values are comparable to the MEC model for all different ambulance settings, which is a prominent characteristic. Furthermore, in the SBFM, all instances are resolved in less than one minute, and this time is not affected by the number of ambulances.

5.5. Measures of the value of information and modeling

In this experiment, we compute the Expected Value of Perfect Information (EVPI) and the Value of the Stochastic Solution (VSS), which are two concepts used in stochastic programming to assess the accuracy of the model [11]. The EVPI measures the max-

imum amount a decision maker would be ready to pay in return for complete (and accurate) information about the future. For a maximization problem, this is computed as EVPI = WS − RP, where RP is the value of the Recourse Program (our MEC model) and WS is the wait-and-see solution. In our case, we have an infinite number of scenarios for which the MEC model gives optimal solutions; thus, it is difficult to compute an exact expression. However, for instances with 900 demand points, 16 possible location points, and only 10 scenarios, we obtain small but positive EVPI values of 0.03 on average, confirming the value of the MEC model, especially as the number of scenarios will increase.

The Value of the Stochastic Solution (VSS) measures how good or, more frequently, how bad a decision based on solving the deterministic case for the average scenario with respect to the RP solution is. The VSS is computed as RP − EVV, where EVV is the expected result of using the EV solution in the MEC model. Table 2 shows in the first column the size of the instances for which we could obtain optimal solutions with the MEC model (demand points, location points, scenarios). The second column displays the RP value, followed by the EVV and then the VSS.

$\|I\|, \|L\|, \|S\|$	RP value	EVV	VSS
168, 16, 5	5.36	5.27	0.09
270, 16, 5	5.32	5.14	0.18
500, 16, 5	7.00	6.68	0.32
900, 16, 5	8.87	8.37	0.50
1500, 16, 5	9.29	8.44	0.85
168, 16, 10	5.00	4.64	0.36
270, 16, 10	5.58	5.2	0.38
500, 16, 10	7.53	7.13	0.39
900, 16, 10	8.47	7.87	0.59
1500, 16, 10	9.99	8.75	1.24

Table 2.: The value of the stochastic solution for the MEC model.

Table 2 shows that as the size of the instance increases, the VSS also increases. Note that the small number of scenarios for which we could obtain optimal solutions is relatively small. However, the behavior of the VSS shows the benefit of considering a stochastic setting even for instances with few scenarios.

6. Conclusions

EMS systems in developing countries such as Mexico suffer from a shortage of ambulances. Thus, one of the main goals addressed in this work was to investigate and develop tools that allow us to decide whether an emergency can be totally or partially covered.

The EVCP problem consists of locating a limited number of two heterogeneous types of ambulances in different city locations and dispatching them to the emergency points to maximize the coverage with short medical first aid response time. In the EVCP problem, these two interrelated decisions are simultaneously considered in a novel two-stage stochastic program. The EVCP stochastic model allows for partial coverage of the accidents by the ambulances based on a decay function.

We propose a two-stage stochastic program for the EVCP problem that can be solved by branch-and-bound for small instances with a restricted number of scenarios. We also propose a surrogate-based feedback method, which is essentially a location-allocation procedure that relies on the solution of an auxiliary surrogate model. This method is faster to solve and allows us to obtain high-quality solutions significantly faster than the previous model. The SBFM was tested on a broad set of randomly generated instances based on real-world data from a local system. An essential feature of the proposed approach is that it can be implemented by calling any off-the-shelf general-purpose integer solver without employing complex decomposition techniques.

Naturally, several lines of work can be further investigated. For example, we observed that some private EMS services also dispatch vehicles to accident sites. Some of these are not regulated or coordinated by the state. In some cases, this provokes a conflict as too many ambulances arrive at the site, leaving other points unattended. This situation could, of course, benefit from coordinated decision-making tools as those

developed here. For instance, Gernert et al. [23] examine two business models that attempt to address this issue with a game-theoretic approach. However, further research is needed. Another line of work lies in investigating more sophisticated solution techniques. Although our method is indeed relatively easy to implement, there are other techniques, such as decomposition-based algorithms [54, 60], sample average approximation [33, 52], or heuristics/metaheuristics [30], that have been successfully applied to integer stochastic programs, that are worthwhile exploring.

An interesting question arises when two or more periods are considered. Although our single-period methodology can be applied to each consecutive period (several times within a single day, assuming buffer times), the relocation of ambulances for the next period, along with deadheading/setup times could also be studied from a multi-period perspective [29, 47]. In this regard, some of the ideas developed in this work could prove useful.

Due to uncertainty in response times, we could consider using an API to obtain real-time vehicle transfer times or a simulation to see the system's operation. Since we address partial emergency coverage, a crucial aspect is the preference between these coverages, which will directly influence the ambulance response time in the solution. Thus, it would also be interesting to consider also robust optimization modeling approaches [1, 39, 58].

Acknowledgments: The presentation of this work has been improved thanks to the criticism and remarks of four anonymous reviewers, for which we are very grateful. The first author acknowledges a scholarship for doctoral studies from the Mexican Council for Humanities, Science, and Technology (CONAHCYT). The second author was supported by grant CF-2023-I-880 from CONAHCYT. The third author was supported by grant M20M01-315691 from ANUIES-CONAHCYT.

Disclosure statement: The authors report that there are no competing interests to declare.

Data availability statement: All instances with their related scenarios and detailed solutions are available at `https://doi.org/10.6084/m9.figshare.25928401`.

References

[1] A. Akıncılar and E. Akıncılar. A new idea for ambulance location problem in an environment under uncertainty in both path and average speed: Absolutely robust planning. *Computers & Industrial Engineering*, 137:106053, 2019.

[2] M. Amorim, S. Ferreira, and A. Couto. How do traffic and demand daily changes define urban emergency medical service (uEMS) strategic decisions?: A robust survival model. *Journal of Transport & Health*, 12:60–74, 2019.

[3] S. Ansari, L. A. McLay, and M. E Mayorga. A maximum expected covering problem for district design. *Transportation Science*, 51(1):376–390, 2015.

[4] F. Antunes, M. Amorim, F. Pereira, and B. Ribeiro. Active learning metamodelling for survival rate analysis of simulated emergency medical systems. *Transportmetrica A: Transport Science*, 20(1):2046203, 2024.

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

[6] G. Bakalos, M. Mamali, C. Komninos, E Koukou, A. Tsantilas, S. Tzima, and T. Rosenberg. Advanced life support versus basic life support in the pre-hospital setting: A meta-analysis. *Resuscitation*, 82(9):1130–1137, 2011.

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

[8] P. Beraldi and M. E. Bruni. A probabilistic model applied to emergency service vehicle location. *European Journal of Operational Research*, 196(1):323–331, 2009.

[9] O. Berman, D. Krass, and Z. Drezner. The gradual covering decay location problem on a network. *European Journal of Operational Research*, 151(3):474–480, 2003.

[10] D. Bertsimas and Y. Ng. Robust and stochastic formulations for ambulance deployment and dispatch. *European Journal of Operational Research*, 279(2): 557–571, 2019.

[11] J. R. Birge and F. Louveaux. *Introduction to Stochastic Programming*. Springer,

New York, 2nd edition, 2011.

[12] R. Boujemaa, A. Jebali, S. Hammami, A. Ruiz, and H. Bouchriha. A stochastic approach for designing two-tiered emergency medical service systems. *Flexible Services and Manufacturing Journal*, 30:123–152, 2018.

[13] O. Braun, R. McCallion, and J. Fazackerley. Characteristics of midsized urban EMS systems. *Annals of Emergency Medicine*, 19(5):536–546, 1990.

[14] L. Brotcorne, G. Laporte, and F. Semet. Ambulance location and relocation models. *European Journal of Operational Research*, 147(3):451–463, 2003.

[15] E. L. D. S. Cabral, W. R. S. Castro, D. R. M. Florentino, D. A. Viana, J. F. D. Costa Junior, R. P. Souza, A. C. M. Rêgo, I. Araújo-Filho, and A. C. Medeiros. Response time in the emergency services. systematic review. *Acta Cirúrgica Brasileira*, 33:1110–1121, 2018.

[16] Estado de Nuevo León. Fortalece estado cobertura prehospitalaria con 30 nuevas ambulancias. Government news release. URL: `https://www.nl.gob.mx/es/boletines/fortalece-estado-cobertura-prehospitalaria-con-30-nuevas-ambulancias#:~:text=en%20la%20entidad.-,En%20la%20administraci%C3%B3n%20del%20Gobierno%20del%20nuevo%20Nuevo%20Le%C3%B3n%20la,de%205%20a%2047%20unidades.`, 18/12/2024. Accessed: 11-March-2025. In Spanish.

[17] J. C. Dibene, Y. Maldonado, C. Vera, M. de Oliveira, L. Trujillo, and O. Schütze. Optimizing the location of ambulances in Tijuana, Mexico. *Computers in Biology and Medicine*, 80:107–115, 2017.

[18] E. Erkut, A. Ingolfsson, and G. Erdoğan. Ambulance location for maximum survival. *Naval Research Logistics*, 55(1):42–58, 2008.

[19] E. Fadda, D. Manerba, G. Cabodi, P. Camurati, and R. Tadei. Comparative analysis of models and performance indicators for optimal service facility location. *Transportation Research Part E: Logistics and Transportation Review*, 145: 102174, 2021.

[20] E. Fadda, D. Manerba, and R. Tadei. How to locate services optimizing redundancy: A comparative analysis of k-covering facility location models. *Socio-Economic Planning Sciences*, 94:101938, 2024.

[21] I. Gago-Carro, U. Aldasoro, J. Ceberio, and M. Merino. A stochastic programming model for ambulance (re)location-allocation under equitable coverage and multi-interval response time. *Expert Systems with Applications*, 249:123665, 2024.

[22] M. Gendreau, G. Laporte, and F. Semet. Solving an ambulance location model by tabu search. *Location Science*, 5(2):75–88, 1997.

[23] A. K. Gernert, A. P. Calmon, G. Romero, and L. N. Van Wassenhove. Business model innovation for ambulance systems in low-and middle-income countries: "coordination and competition". *Production and Operations Management*. Forthcoming, DOI: 10.1177/10591478231224973.

[24] M. Gharib, S. M. T. Fatemi Ghomi, and F. Jolai. A multi-objective stochastic programming model for post-disaster management. *Transportmetrica A: Transport Science*, 18(3):1103–1126, 2022.

[25] B. C. Grannan, N. D. Bastian, and L. A. McLay. A maximum expected covering problem for locating and dispatching two classes of military medical evacuation air assets. *Optimization Letters*, 9:1511–1531, 2015.

[26] A. Hatami-Marbini, N. Varzgani, S. M. Sajadi, and A. Kamali. An emergency medical services system design using mathematical modeling and simulation-based optimization approaches. *Decision Analytics Journal*, 3:100059, 2022.

[27] K. Hogan and C. ReVelle. Concepts and applications of backup coverage. *Management Science*, 32(11):1434–1444, 1986.

[28] A. Hossain, S. Barua, S. Das, and M. Starewich. Ambulance crash risk dynamics: A baseline (2017–2019) vs. pandemic-era (2020–2022) comparative study using a random parameter logit model. *Transportmetrica A: Transport Science*. Forthcoming, DOI: 10.1080/23249935.2025.2481578.

[29] O. J. Ibarra-Rojas, F. López-Irarragorri, and Y. A. Ríos-Solís. Multiperiod bus timetabling. *Transportation Science*, 50(3):805–822, 2016.

[30] A. A. Juan, P. Keenan, R. Martí', S. McGarragh, J. Panadero, P. Caroll, and D. Oliva. A review of the role of heuristics in stochastic optimisation: From metaheuristics to learnheuristics. *Annals Operations Reseasrch*, 320:831–861, 2023.

[31] C. Khelfa and I. Khennak. A survey on recent optimization strategies in ambulance dispatching and relocation problems. In H. Drias, F. Yalaoui, and A. Hadjali,

editors, *Artificial Intelligence Doctoral Symposium*, volume 1852 of *Communications in Computer and Information Science*, pages 192–203. Springer, Singapore, 2023.

[32] F. Khosgehbari and M. J. Mirzapour Al-e Hashem. Ambulance location routing problem considering all sources of uncertainty: Progressive estimating algorithm. *Computers & Operations Research*, 160:106400, 2023.

[33] A. J. Kleywegt, A. Shapiro, and T. H. de Mello. The sample average approximation method for stochastic discrete optimization. *SIAM Journal on Optimization*, 12(2):479–502, 2001.

[34] L. A. McLay. A maximum expected covering location model with two types of servers. *IIE Transactions*, 41(8):730–741, 2009.

[35] N. Megiddo and A. Tamir. On the complexity of locating linear facilities in the plane. *Operations Research Letters*, 1(5):194–197, 1982.

[36] R. A. Nadar, J. K. Jha, and J. J. Thakkar. Adaptive variable neighbourhood search approach for time-dependent joint location and dispatching problem in a multi-tier ambulance system. *Computers & Operations Research*, 159:106355, 2023.

[37] S. Nickel, M. Reuter-Oppermann, and F. Saldanha-da Gama. Ambulance location under stochastic demand: A sampling approach. *Operations Research for Health Care*, 8:24–32, 2016.

[38] C. O'Keeffe, J. Nicholl, J. Turner, and S. Goodacre. Role of ambulance response times in the survival of patients with out-of-hospital cardiac arrest. *Emergency Medicine Journal*, 28(8):703–706, 2011.

[39] J. Ong, D. Kulpanowski, Y. Xie, E. Nikolova, and N. M. Tran. OpenEMS: An open-source package for two-stage stochastic and robust optimization for ambulance location and routing with applications to Austin-Travis County EMS data. ArXiv preprint arXiv:2201.11208, 2022. URL `https://arxiv.org/abs/2201.11208`.

[40] M. Reuter-Oppermann, P. L. van den Berg, and J. L. Vile. Logistics for emergency medical service systems. *Health Systems*, 6(3):187–208, 2017.

[41] H. Rhodes, B. Rourke, and A. Pepe. Ambulance response in eight minutes or

less: Are comorbidities a factor. *The American Surgeon*, 89(8):3478–3481, 2023.

[42] L. Shaw, S. K. Das, and S. K. Roy. Location-allocation problem for resource distribution under uncertainty in disaster relief operations. *Socio-Economic Planning Sciences*, 82:101232, 2022.

[43] D. Shen, R. Li, and X. Liu. Dynamic-collaborative path planning based on tradable road priority: an interweaving strategy for emergency vehicle. *Transportmetrica A: Transport Science*. Forthcoming, DOI: 10.1080/23249935.2024.2447318.

[44] H. Sun, Y. Wang, and Y. Xue. A bi-objective robust optimization model for disaster response planning under uncertainties. *Computers & Industrial Engineering*, 155:107213, 2021.

[45] H. Sun, Y. Wang, J. Zhang, and W. Cao. A robust optimization model for location-transportation problem of disaster casualties with triage and uncertainty. *Expert Systems with Applications*, 175:114867, 2021.

[46] I. Sung and T. Lee. Scenario-based approach for the ambulance location problem with stochastic call arrivals under a dispatching policy. *Flexible Services and Manufacturing Journal*, 30:153–170, 2016.

[47] Z. Tan, L. Zhen, Z. Yang, L. Liu, and T. Fan. Multi-period emergency vehicle fleet redistribution and dispatching. *Transportmetrica A: Transport Science*, 21 (2):2243344, 2025.

[48] N. Theeuwes, G.-J. van Houtum, and Y. Zhang. Improving ambulance dispatching with machine learning and simulation. In Y. Dong, N. Kourtellis, B. Hammer, and J. A. Lozano, editors, *Machine Learning and Knowledge Discovery in Databases*, volume 12978 of *Lecture Notes in Computer Science*, pages 302–318. Springer, Cham, Switzerland, 2021.

[49] H. Toro-Díaz, M. E. Mayorga, L. A. McLay, H. K. Rajagopalan, and C. Saydam. Reducing disparities in large-scale emergency medical service systems. *Journal of the Operational Research Society*, 66(7):1169–1181, 2015.

[50] M. van Buuren, R. van der Mei, and S. Bhulai. Demand-point constrained EMS vehicle allocation problems for regions with both urban and rural areas. *Operations Research for Health Care*, 18:65–83, 2018.

[51] J. Wang, H. Liu, S. An, and N. Cui. A new partial coverage locating model for

cooperative fire services. *Information Sciences*, 373:527–538, 2016.

[52] W. Wang, S. Wang, L. Zhen, and X. Qu. EMS location-allocation problem under uncertainties. *Transportation Research Part E: Logistics and Transportation Review*, 168:102945, 2022.

[53] J. Wu, Y. Lin, and W. Qi. Timing co-evolutionary path optimisation method for emergency vehicles considering the safe passage. *Transportmetrica A: Transport Science*, 21(2):2253477, 2025.

[54] Y. Wu, S. Wang, L. Zhen, G. Laporte, Z. Tan, and K. Wang. How to operate ship fleets under uncertainty. *Production and Operations Management*, 32(10): 3043–3061, 2023.

[55] M. Yavari, R. Maihami, and M. Esmaeili. Ambulance dispatching and relocation problem considering overcrowding of emergency departments. *IISE Transactions on Healthcare Systems Engineering*, 12(4):263–274, 2022.

[56] S. Yoon, L. A. Albert, and V. M. White. A stochastic programming approach for locating and dispatching two types of ambulances. *Transportation Science*, 55(2): 275–296, 2021.

[57] G. Yu, A. Liu, and H. Sun. Risk-averse flexible policy on ambulance allocation in humanitarian operations under uncertainty. *International Journal of Production Research*, 59:2588–2610, 2021.

[58] Y. Yuan, Q. Song, and B. Zhou. A Wasserstein distributionally robust chance constrained programming approach for emergency medical system planning problem. *International Journal of Systems Science*, 53(10):2136–2148, 2022.

[59] R. Zhang and B. Zeng. Ambulance deployment with relocation through robust optimization. *IEEE Transactions on Automation Science and Engineering*, 16(1): 138–147, 2018.

[60] L. Zhen, X. He, D. Zhuge, and S. Wang. Primal decomposition for berth planning under uncertainty. *Transportation Research Part B: Methodological*, 183:102929, 2024.

[61] Z. Zhou, D. S. Matteson, D. B. Woodard, S. G. Henderson, and A. C. Micheas. A spatio-temporal point process model for ambulance demand. *Journal of the American Statistical Association*, 110(509):6–15, 2015.

Appendix A. Abbreviations

Table A1 contains the abbreviations used throughout the paper.

ALS	Advanced life Support
BLS	Basic life support
CRUM	Medical Emergency Regulatory Center
EMS	Emergency medical service
EVCP	Emergency vehicle covering and planning
EVPI	Expected value of perfect information
MEC	Maximum expected coverage
SABC	Surrogate ambulance-based coverage
SBFM	Surrogate-based feedback method
VSS	Value of the stochastic solution
WHO	World Health Organization

Table A1. List of abbreviations.