An Intelligent Feedback Approach for Locating and Dispatching Two Types of Ambulances under Uncertainty and Partial Coverage Considerations

Preprint DIE-RR24-01
Graduate Program in Electrical Engineering, Department of Mechanical and Electrical Engineering,
Universidad Autónoma de Nuevo León, San Nicolás de los Garza, NL, Mexico, July 2024

Beatriz A. García-Ramos
Graduate Program in Systems Engineering
Universidad Autónoma de Nuevo León (UANL)
San Nicolás de los Garza, NL 66455, Mexico
beatrizgr95@gmail.com

Roger Z. Ríos-Mercado
Graduate Program in Electrical Engineering
Universidad Autónoma de Nuevo León (UANL)
San Nicolás de los Garza, NL 66455, Mexico
roger.rios@uanl.edu.mx

Yasmín Ríos-Solís
Science and Engineering School
Tecnologico de Monterrey
Monterrey, Mexico
yasmin.riossolis@tec.mx

01 July 2024

Abstract

The Emergency Vehicle Covering and Planning (EVCP) problem locates a limited number of two heterogeneous types of ambulances in different city points and dispatches them to the emergency scenes, considering the uncertainty of the emergency locations, to maximize the emergency total and partial coverage and the response time in which the patients receive medical first aids. We propose a novel two-stage quadratic stochastic program for the EVCP problem that locates the limited number of heterogeneous types of ambulances in the first stage. The second stage deals with dispatching ambulances to accidents. The EVCP stochastic model allows partial coverage of the accidents by the ambulances based on a decay function. Instead of decomposing the stochastic model, we propose a location-allocation methodology that relies on the solution of an auxiliary surrogate model, which is faster to solve. The location of the ambulances obtained by this surrogate model is input to the original model. Experimental results show that we obtain high-quality solutions in a reasonable time.

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

1 Introduction

Emergency Medical Services (EMS) systems provide basic but urgent in-situ medical care for people who suffer a medical incident and then transport patients to hospitals [3, 5, 19]. The first phase of an EMS is the response to an emergency call by an operator that identifies the emergency type: accident, medical, security, fire, etc. The second phase is dispatching one or several ambulances to the emergency scene to provide urgent medical care. Some emergency situations, such as a multiple-car accident, may involve several people; thus, more than one ambulance could be needed.

EMS systems in developing countries, as is the case in Mexico, lack around 30-60%[1] of the number of ambulances suggested by the World Health Organization (WHO), which should be at least four ambulances per 100,000 people [10]. For the Red Cross, an EMS operating with this small number of ambulances is considered similar to a war situation[1]. Thus, one of the main contributions of this work is to deal with the problem of deciding if an emergency will be totally or partially covered. Sadly, some emergencies may remain uncovered by an emergency unit. By using a novel decay function for the total and partial coverage of emergencies, we reduce the average response time of a patient's initial treatment given by a paramedic in an emergency, which is the main objective function in many research works [1, 14, 23]. Indeed, the quickness and the number of ambulances dispatched to the accidents are crucial. Each ambulance has a response time for travel from the potential site where it is located to the demand point where the patient will be cared for. For example, every minute of treatment delay in a cardiac patient reduces the survival probability by 24% [18].

This work considers two different types of ambulances in EMS systems, which implies modeling challenges. When a BLS ambulance is dispatched to an emergency requiring an ALS, it may reduce the patient's survival. Thus, this work considers that ALS ambulances can be used as BLS units, but the contrary is not allowed [4]. A few works deal with different types of ambulances, as we do in this work. McLay [13] determines how to optimally locate and use ambulances to improve patient survivability and coordinate multiple medical units with a hypercube queuing model. Grannan et al. [12] determine how to dispatch multiple types of air assets to prioritized service calls to maintain a high likelihood of survival of the most urgent casualties in a military medical evacuation by a binary linear programming model. In Yoon et al. [27], two types of vehicles are considered, but one of them is a rapid one that cannot offer the first care services of an ambulance. Moreover, neither of these works considers partial covering of the calls.

To summarize, the *Emergency Vehicle Covering and Planning* (EVCP) problem consists of locating the limited number of two heterogeneous types of ambulances in different city locations and dispatching them to the uncertain emergency points to maximize the coverage (even if partially) with short medical first aid response time. Usually, the location and dispatching decisions are made

[1] Anonymous interviews done by the authors.

separately [5, 11, 26]. In the EVCP problem, these two interrelated decisions are simultaneously considered in a novel two-stage stochastic program. The limited number of heterogeneous types of ambulances in the first stage and in the second stage, the dispatching of ambulances to accidents is determined. The EVCP stochastic model allows partial coverage of the accidents by the ambulances based on a decay function [25], which is a main difference with other stochastic approaches Toro-Díaz et al. [23], Ansari et al. [2], Amorim et al. [1].

Similarly to Yoon et al. [27], we generate the call-arrival scenarios by sampling from emergency call logs to use them in the second stage of our stochastic model. In this manner, we address the volume of calls during a short period, such as Friday night hours. Thus, time is not explicitly measured, and it is assumed that a vehicle can be assigned only once during this high ambulance demand period [29]. Boujemaa et al. [9] use a bundle of calls but do not consider a heterogeneous ambulance fleet.

Some works propose stochastic programming models based on call-arrival scenarios as a bundle of calls, the total number of emergency calls in each demand node during a given period. As we do in this work, a two-stage stochastic program deploys the ambulances in the first stage and dispatches them to respond to demand in the second stage. Beraldi and Bruni [6] and Noyan [17] induce a reliability approach by using probabilistic constraints. Nickel et al. [16] minimize the total cost of locating the ambulances while assuring a minimum coverage level. By considering a bundle of calls, they address the volume of calls during a short period, such as the Friday night hours. Bertsimas and Ng [7] implemented stochastic and robust formulations for ambulance deployment and dispatch to minimize the fraction of late arrivals without requiring ambulances to be relocated, sending to demand points the closest available ambulance, and maintaining a call at a queue if there are no ambulances available at the system.

The advantage of a stochastic programming solution over deterministic approaches where inherent uncertainty has been long proved [8]. A contribution of this work is the methodology to solve the EVCP stochastic model. Indeed, the proposed model can only be solved for relatively small instances with a restrictive number of scenarios. Thus, instead of decomposing the model with cut methods as it is usually done [28, 21], we propose a location-allocation methodology [20, 24] that relies on the solution in an auxiliary surrogate model, which is faster to solve. We name this method *an intelligent feedback approach* because the location of the ambulances obtained by this surrogate model is used as input to the original model. Thus, we obtain high-quality solutions in a reasonable time with an off-the-shelf solver without complex decomposition techniques.

Some works use metaheuristic methods to solve their stochastic models. Toro-Díaz et al. [22] integrate location and dispatching decisions for EMS vehicles to minimize the mean response time of an emergency call and maximize the expected coverage demand, using a continuous-time Markov process to balance flow equations that control the busy fraction of each ambulance. A genetic algorithm can solve mid-size instances. Some others, such as Amorim et al. [1], use simulation

to decide if ambulances stay at the potential sites established by a mathematical model or must be moved to another potential site to maximize the patient's survival. They work on a complete day period while we focus on high-demand periods of some hours. Moreover, we do not need a metaheuristic due to the high-quality solutions that we obtained with the Intelligent feedback approach.

The remainder of this article is organized as follows. Section 2 describes the EVCP problem, emphasizing the emergency coverage types and the two types of ambulances. This includes introducing a two-stage stochastic quadratic model for the EVCP problem (presented in Section 2.2). The intelligent feedback approach for the EVCP problem is detailed in Section 3. Experimental results on real-world-based generated instances that show the efficiency of our approach are given in Section 4. Final remarks and conclusions are drawn in Section 5.

2 The Emergency Vehicle Covering and Planning problem

Let us formally describe the Emergency Vehicle Covering and Planning problem. Let set I include the possible demand points where patients may 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. Set L provides the potential sites or ambulance stations where ambulances could be located, such as hospitals, firehouses, malls, or similar places where the ambulance and the paramedics can wait for emergency calls. We consider instances with up to 30 potential sites for the experimental results. Set K contains the two 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 traveling time of any ambulance type from a potential site $l \in L$ to a demand point $i \in I$ is given by r_{li}. 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, sadly, the accident has probably been dealt with by other means.

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 that only depends on the response time of a location $l \in L$ to any demand point $i \in I$:

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

2.1 Information related to the scenarios

The operational level is represented by a set of scenarios S with a bundle list of arriving calls. Each scenario $s \in S$ represents a set of emergencies in the demand points. Thus, a scenario is represented 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 on a demand point $i \in I$ and provides the value a^s_{ki} related to the number of required ambulances of type $k \in K$.

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

- Total: the a^s_{ki} required ambulances of each type k are dispatched to i, and all arrive in less than τ time.
- Total-late: the a^s_{ki} 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^s_{ki} 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^s_{ki} 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^s_{ki} required ambulances arrives in less than $\tau_{\max}$ time, for $k \in K$.

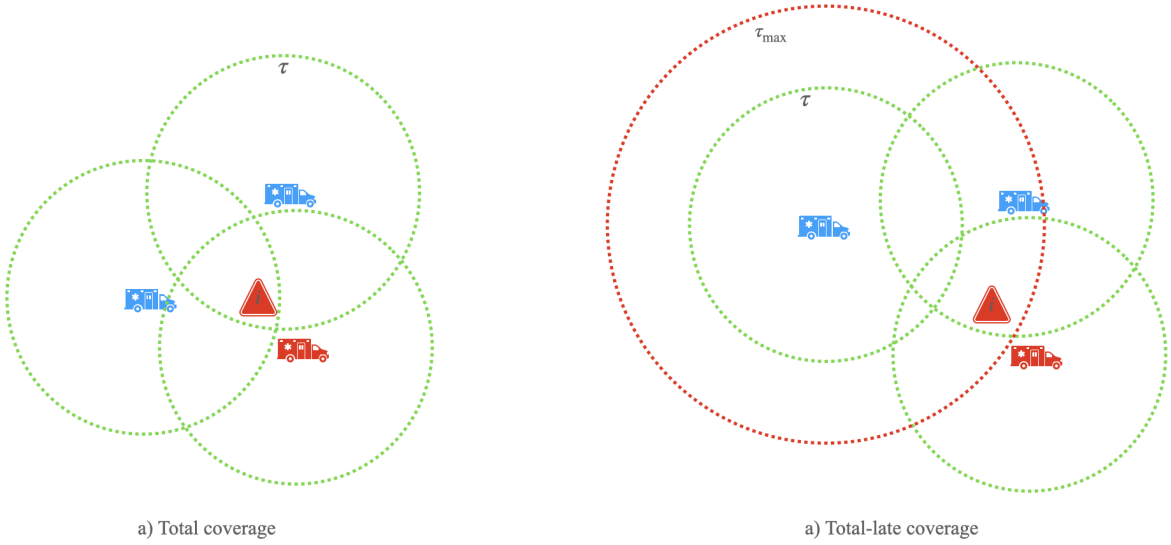


Figure 1: Two different coverage cases for a scenario $s \in S$ where $i \in I^s$ requires $a^s_{1i} = 2$ basic ambulances (blue) and $a^s_{2i} = 1$ advanced ones (red). Total coverage in the left: all ambulances arrive in less than the ideal time τ. Total-late coverage in the right: at least one of the ambulances arrives between $(\tau, \tau_{\max})$.

Figure 1 illustrates two different coverage cases for a scenario $s \in S$ where $i \in I^s$ requires $a^s_{1i} = 2$ basic ambulances (indicated in blue) and $a^s_{2i} = 1$ advanced one (indicated in red). All ambulances arrive in less than the ideal time τ for the Total coverage (left-hand-side figure). In the Total-late coverage line, at least one of the ambulances is late since it arrives between $(\tau, \tau_{\max})$ (right-hand-side figure). In the Partial coverage, the number of required ambulances is not met, but at least they arrive in less than the ideal time τ. In the Partial-late coverage, not only are there not enough ambulances to cover the demand point, but they arrive late, that is, between $(\tau, \tau_{\max})$. In the Null coverage, ambulances may be dispatched to the demand point, but since the arrival times are larger than $\tau_{\max}$, the demand point is considered uncovered.

Table 1 summarizes the sets and parameters used to describe the EVCP problem.

2.2 Maximum Expected Coverage stochastic formulation for the EVCP problem

The Maximum Expected Coverage (MEC) formulation is a stochastic integer quadratic programming model in which the first stage variables x_{lk} correspond to the number of ambulances of type $k \in K$ located at $l \in L$, and the second-stage variables correspond to the ambulance dispatching

I	set of possible demand points (possible accident places)
L	set of possible ambulance location sites
K	set of ambulance types
η_k	total number of ambulances in the system of type $k \in K$
r_{li}	response time from potential site $l \in L$ to demand point $i \in I$
τ	ideal response time to give the patients the first medical aid in an emergency
τ_{max}	maximum response time to cover an accident
c_{li}	benefit from traveling from potential site $l \in L$ to demand point $i \in I$
S	set of scenarios
a^s_{ki}	number of needed ambulances of type $k \in K$ at demand point $i \in I, s \in S$
I^s	set of demand points for $s \in S$ with at least a value $a^s_{ki} \neq 0$ for $i \in I, k \in K$

Table 1: Sets and parameters to describe the EVCP problem.

decisions at each demand point for each scenario $s \in S$:

$$y^s_{lki} = \begin{cases} 1 & \text{if an ambulance of type } k \in K \text{ in location } l \in L \\ & \text{is dispatched to demand point } i \in I^s, \text{ for scenario } s \in S, \\ 0 & \text{otherwise.} \end{cases}$$

We defined the following binary variables related to the *total* and *total-late* coverages related to the response times of the ambulances to the demand point $i \in I^s, s \in S$:

$$f^s_i = \begin{cases} 1 & \text{if demand point } i \in I^s \text{ has a } \textit{total} \text{ coverage,} \\ 0 & \text{otherwise,} \end{cases}$$

$$g^s_i = \begin{cases} 1 & \text{if demand point } i \in I^s \text{ has a } \textit{total-late} \text{ coverage,} \\ 0 & \text{otherwise.} \end{cases}$$

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

$$h^s_i = \begin{cases} 1 & \text{if demand point } i \in I^s \text{ has a } \textit{partial} \text{ coverage,} \\ 0 & \text{otherwise,} \end{cases}$$

$$w^s_i = \begin{cases} 1 & \text{if demand point } i \in I^s \text{ has a } \textit{partial-late} \text{ coverage,} \\ 0 & \text{otherwise.} \end{cases}$$

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

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

The MEC formulation is as follows.

$$\max_x \mathbb{E}_{s\in S}[\mathcal{Q}^s(x)] \tag{1}$$

where

$$\mathcal{Q}^s(x) = \sum_{i\in I^s}(\alpha_1 f_i^s + \alpha_2 g_i^s + \alpha_3 h_i^s + \alpha_4 w_i^s - \phi z_i^s)$$

subject to

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

$$\sum_{i\in I^s} y_{lki}^s \le x_{lk} \qquad l \in L, k \in K, s \in S \tag{3}$$

$$a_{1i}^s f_i^s \le \sum_{l\in L}\sum_{k\in K} c_{li} y_{lki}^s, \quad a_{2i}^s f_i^s \le \sum_{l\in L} c_{li} y_{l2i}^s \qquad i \in I^s, s \in S \tag{4}$$

$$a_{1i}^s g_i^s \le \sum_{l\in L}\sum_{k\in K} y_{lki}^s, \quad a_{2i}^s g_i^s \le \sum_{l\in L} y_{l2i}^s \qquad i \in I^s, s \in S \tag{5}$$

$$g_i^s \le M\left(\sum_{l\in L}\sum_{k\in K} y_{lki}^s - \sum_{l\in L}\sum_{k\in K} c_{li} y_{lki}^s\right) \qquad i \in I^s, s \in S \tag{6}$$

$$h_i^s \le a_{1i}^s - \sum_{l\in L}\sum_{k\in K} y_{lki}^s, \quad h_i^s \le a_{2i}^s - \sum_{l\in L} y_{l2i}^s \qquad i \in I^s, s \in S \tag{7}$$

$$\sum_{l\in L}\sum_{k\in K} y_{lki}^s h_i^s \le \sum_{l\in L}\sum_{k\in K} c_{li} y_{lki}^s \qquad i \in I^s, s \in S \tag{8}$$

$$w_i^s \le a_{1i}^s - \sum_{l\in L}\sum_{k\in K} y_{lki}^s, \quad w_i^s \le a_{2i}^s - \sum_{l\in L} y_{l2i}^s \qquad i \in I^s, s \in S \tag{9}$$

$$w_i^s \le M\left(\sum_{l\in L}\sum_{k\in K} y_{lki}^s - \sum_{l\in L}\sum_{k\in K} c_{li} y_{lki}^s\right) \qquad i \in I^s, s \in S \tag{10}$$

$$\sum_{l\in L}\sum_{k\in K} y_{lki}^s + z_i^s \ge 1 \qquad i \in I^s, s \in S \tag{11}$$

$$f_i^s + g_i^s + h_i^s + w_i^s + z_i^s = 1 \qquad i \in I^s, s \in S \tag{12}$$

$$x_{lk} \in \mathbb{Z}^+, y_{lki}^s \in \{0,1\} \qquad l \in L, k \in K, i \in I^s, s \in S \tag{13}$$

$$f_i^s, g_i^s, h_i^s, w_i^s, z_i^s \in \{0,1\} \qquad i \in I^s, s \in S. \tag{14}$$

The objective function (1) maximizes the expected value of the weighted coverage of the emer-

gencies. The parameters $\alpha_1 > \alpha_2 > \alpha_3 > \alpha_4$ are normalized weights that ponder the coverage type, and ϕ is the penalty for the null coverage. We assume that every scenario is equally probable since each $s \in S$ represents a sample of the high-demand period we are interested in.

Constraints (2) establish the available number of ambulances per type. Constraints (3) establish the relationship between the first and second-stage variables, meaning no ambulances can be dispatched from a potential site if no ambulances are located there. The *total* coverage of an emergency is defined by constraints (4). Indeed, if the time response of the location of the ambulances to the emergency is less than τ, then all $c_{li} = 1$ and total coverage variables f_i^s can be equal to one, for $l \in L, i \in I^s, s \in S$. The *total-late* coverage is defined by constraints (5) and (6). Constraints (5) allow the total-late coverage variables g_i^s to be one when dispatching variables are active. Meanwhile, constraints (6) track the demand points where the response time is between $(\tau, \tau_{\max})$ when the difference in the right-hand side of the equation is positive, that is, when there is a value $c_{lj} < 1$ associated to a dispatched ambulance, for $l \in L, i \in I^s, s \in S$. Note that this difference may be decimal, so we include a big M value. The *partial* coverage is defined by constraints (7) and (8). Recall that, in this case, not all the needed ambulances are dispatched to the emergencies, but the ones that are dispatched have an ideal response time. Thus, constraints (7) activate variables h_i^s if the number of dispatched ambulances is less than the required ones. Quadratic constraints (8) guarantee that the dispatched ambulances arrive within the ideal response time, that is, their corresponding value $c_{li} = 1$, for $l \in L, i \in I^s, s \in S$. Constraints (9) and (10) define the *partial-late* coverage. Constraints (9) activate the w_i^s variables when the number of required ambulances exceeds the number of dispatched ones. Similarly to the total-late coverage, constraints (10) track the ambulances with a response time larger than the ideal one and must be multiplied by a big M. The *null* coverage is activated by constraints (11). All the coverage constraints are related to constraint (12) that ensures only one type of coverage for each emergency. Finally, (13) and (14) 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. Nevertheless, the related number of variables and constraints is usually large. Moreover, constraints (8) are quadratic. An integer linear stochastic model with a classical linearization method could be easily formulated. Still, previous experiments showed similar times between the linearized and the quadratically constrained models when solved with integer programming solvers, so we keep the quadratic one for the Intelligent Feedback methodology presented in the next section.

3 Intelligent Feedback methodology for the EVCP problem

The EVCP problem is $\mathcal{NP}$-hard since the classical facility location problem [15] could be reduced to it. The MEC model is experimentally challenging to solve, even for medium-size instances, as

shown in Section 4. Thus, we propose the Intelligent Feedback methodology to obtain approximate solutions for the EVCP problem based on an auxiliary disaggregated model, named *Surrogate Ambulance-Based Coverage* (SABC) model.

In addition to the location variables x_{li}, the SABC model requires the following binary ambulance dispatching variables for $k \in K, l \in L, i \in I^s, s \in S$:

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

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

Variables u^s_{lki} 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^s_{lki} 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 variable ζ^s_{ki}, for $k \in K, i \in I^s, s \in S$. The SABC is as follows.

$$\max_x \mathbb{E}_s[\mathcal{G}^s(x)] \tag{15}$$

where

$$\mathcal{G}^s(x) = \left[\sum_{l \in L} \sum_{k \in K} \sum_{i \in I^s} (\beta_1 u^s_{lki} + \beta_2 v^s_{lki}) - \sum_{k \in K} \sum_{i \in I^s} \phi \zeta^s_{ki} \right]$$

subject to

$$\sum_{l\in L} x_{lk} \leq \eta_k \qquad k \in K \tag{16}$$

$$\sum_{i\in I^s} (u^s_{lki} + v^s_{lki}) \leq x_{lk} \qquad l \in L, k \in K, s \in S \tag{17}$$

$$u^s_{lki} \leq c_{li} \qquad l \in L, i \in I^s, k \in K, s \in S \tag{18}$$

$$u^s_{lki} + v^s_{lki} \leq 1 \qquad l \in L, i \in I^s, k \in K, s \in S \tag{19}$$

$$a^s_{1i} = \sum_{l\in L}\sum_{k\in K} (u^s_{lki} + v^s_{lki} + \zeta^s_{ki}) \qquad i \in I^s, s \in S \tag{20}$$

$$a^s_{2i} = \sum_{l\in L} (u^s_{l2i} + v^s_{l2i} + \zeta^s_{2i}) \qquad i \in I^s, s \in S \tag{21}$$

$$\sum_{k\in K} u^s_{lki} \leq a^s_{i1}, \quad u^s_{l2i} \leq a^s_{i2} \qquad i \in I, l \in L, s \in S \tag{22}$$

$$\sum_{k\in K} v^s_{lki} \leq a^s_{i1}, \quad v^s_{l2i} \leq a^s_{i2} \qquad i \in I, l \in L, s \in S \tag{23}$$

$$x_{lk}, \zeta^s_{ki} \in \mathbb{Z}^+, u^s_{lki}, v^s_{lki} \in \{0,1\} \qquad l \in L, k \in K, i \in I^s, s \in S$$

The objective function (15) 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. Weights $\beta_1 > \beta_2$ are normalized parameters prioritizing the dispatched ambulances with a response time less than τ. As in the previous model, no more than the available ambulances can be located in the sites, corresponding to constraints (16). The number of on-time or late dispatched ambulances is less than the number of located ambulances, as indicated by constraints (17). Constraints (18) define the dispatched ambulances 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 (19) activate the late variables for which their response time is between $(\tau, \tau_{\max})$. With constraints (20) and (21), the non-covered emergencies, ζ^s_{ki} variables, are defined, for $i \in I^s, s \in S$. To allow advanced ambulances to be dispatched instead of basic ones, we add restrictions (22) and (23). Finally, the nature of the variables is stated.

The essential characteristic of the SABC model is that the objective function does not rely on emergency coverage as in the MEC model; it only counts the number of on-time, late, or null ambulances sent to the emergency demand points. Moreover, its resolution time is extremely fast since it requires fewer variables and constraints than the MEC one. Nevertheless, disaggregating an emergency situation into the number of ambulances needed does not capture emergency coverage, which is crucial for an EMS system.

We first solve the SABC stochastic model in the Intelligent Feedback methodology for the EVCP problem. From its optimal solution, we obtain the location of the ambulances of the first

stage corresponding to the value of x_{lk} variables, for $l \in L, k \in K$. We use these values as input to the MEC model. Since the first stage variables are fixed, the MEC(SABC) 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. Nevertheless, experimental results show that the quality of the solutions using the MEC(SABC) methodology is extremely high with a single feedback.

4 Experimental results

This section presents an empirical assessment of models and 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.

4.1 Instances and parameters

The value ranges of our instance generator are based on real-world data taken from Monterrey, Mexico. In the literature, there are no benchmarks suitable for our problem in terms of size. 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 $|I| \leq 270$ [27]. The number of possible locations of the 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, thus the number of ambulances $(\eta_1, \eta_2) = (35, 20)$ is also larger than the ones from the literature cases (6 ambulances per type [27]). The number of scenarios is set to be as large as the ones from the literature, $|S| \in \{10, 50, 100, 150, 200\}$. 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 $s \in S$ consists of a set of demand values per type of ambulance and per demand point $\{a^s_{ki}\}_{k \in K, i \in I, s \in S}$. Fewer demand points imply a larger grid of the city and a larger proportion of emergencies per demand point thus, when $|I| = 168$, around 30% of the demand points may have a value different from 0. On the contrary, 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. The instances are built such that most emergencies require a single ambulance.

The ideal ambulance response time is $\tau = 10$ minutes while the maximal response time is $\tau_{\max} = 30$ minutes. 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 one, while the partial-late cover is the one with less benefit. For the SABC objective function (3) we use $\beta_1 = 0.7$ and $\beta_2 = 0.3$. These values reflect the aim to send primordially the required ambulances 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 is set to $\phi = 1/|S| + 0.0005$.

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

4.2 Experimental analysis of the MEC and MEC(SABC) stochastic formulations

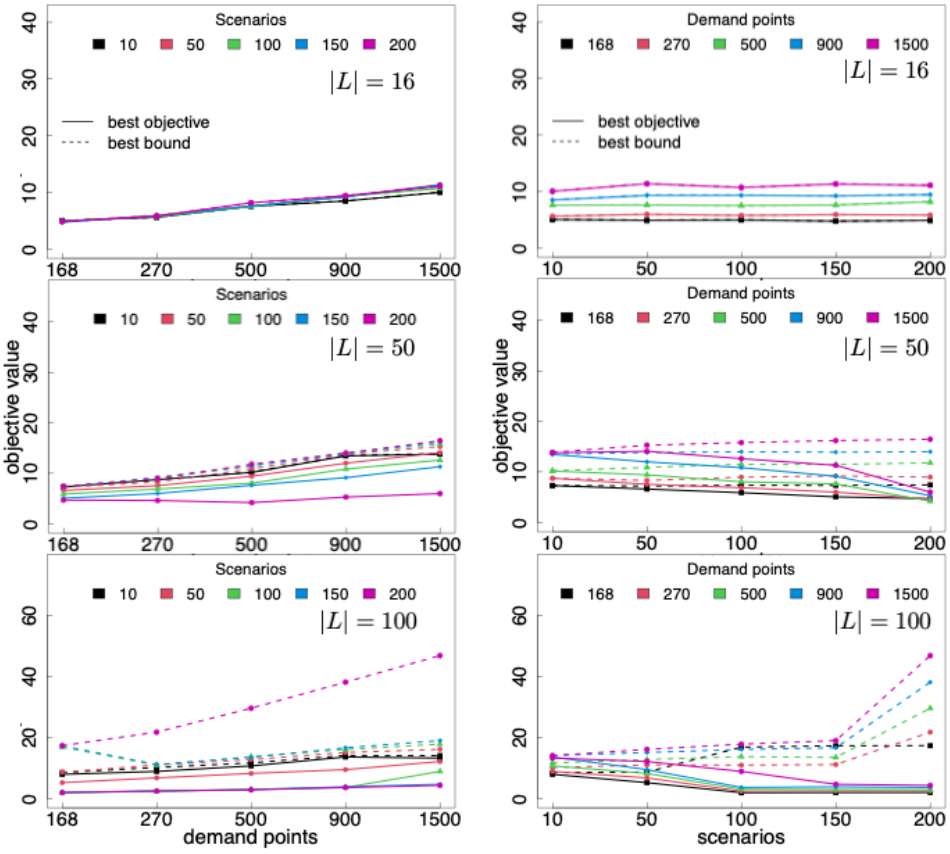


Figure 2: Best objective and the best bound of the objective function obtained by the MEC model with respect to the demand points on the left side and the scenarios on the right side for different sizes of potential sites $|L| = \{16, 50, 100\}$.

In this section, we analyze the parameters of the EVCP problem that impact the performance of the objective values of our stochastic methodologies. Several questions arise. Does the number of scenarios impact the objective function? Does a high number of demand points imply a more

challenging instance? Does the number of possible locations impact the efficiency of the models?

First, we solve all the instances with MEC's deterministic equivalent integer program. In this manner, we can evaluate the size of the instances for which MEC can give optimal solutions and be able to compare the MEC(SABC) intelligent feedback method. Figure 2 consists of six plots. The three plots on the first column vary the number of demand points (x-axis), comparing each one to the objective function value when different scenarios are tested. The three plots in the second column vary the tested number of scenarios and show the variation in the solution value for each number of demand points. The upper plots consider a number of possible locations for the ambulances of $|L| = 16$, the middle plots of $|L| = 50$, and the lower plots of $|L| = 100$. The straight lines are the best objective values, while dotted ones are the best bound found.

As can be seen from the plots, the difference between the best objective and the best bound (and thus, the relative optimality gaps[2]) are negligible for small instances with 16 potential locations sites. Still, the gaps become larger for the instances with 50 and 100 potential sites. The number of demand points where accidents may occur and the considered scenarios make the instances harder to solve optimally within the time limit. Thus, the deterministic equivalent integer program of MEC can only handle small instances with a few scenarios, demand points (emergency points), and potential sites for ambulances. Note that the larger the number of scenarios in the left-hand-side plots, the better the objective function. This implies that a better sampling of the emergency demand points benefits the solution quality related to the ambulance's response time. The right-hand-side plots show that the larger the size of the demand point set, the harder it is to solve the instance.

Now, we compare the solution values of the equivalent integer program of MEC with the ones obtained by MEC(SABC) in Figure 3, which a similar structure than the previous one. As can be seen, while the number of scenarios, demand points, and potential sites slightly affect the MEC(SABC) performance, it obtains better objective function values than those obtained by the MEC model for the larger instances that reported positive gaps with the MEC model. Indeed, the MEC(SABC) model optimality gaps always equal 0 within the time limit we established. Also, the MEC(SABC) model tends to be less dependent on the number of scenarios. Thus, although we cannot guarantee optimality with the MEC(SABC) model, it obtains faster and higher-quality solutions than those obtained by the MEC equivalent model.

The purpose of the following experiment is to compare the running times of the equivalent MEC model with the MEC(SABAC) method. Recall that MEC(SABC) attempts to exploit that the surrogate model SABC is very tractable and solved relatively quickly. To this end, Figure 4 shows two plots of the running time in seconds of the instances with a) $|L| = 16$ potential location sites for the equivalent MEC and the b) $|L| = 100$ potential site instances for the MEC(SABC) methodology. The x-axis of the plots corresponds to the number of scenarios, and we vary the

[2](best objective - best bound)/best objective.

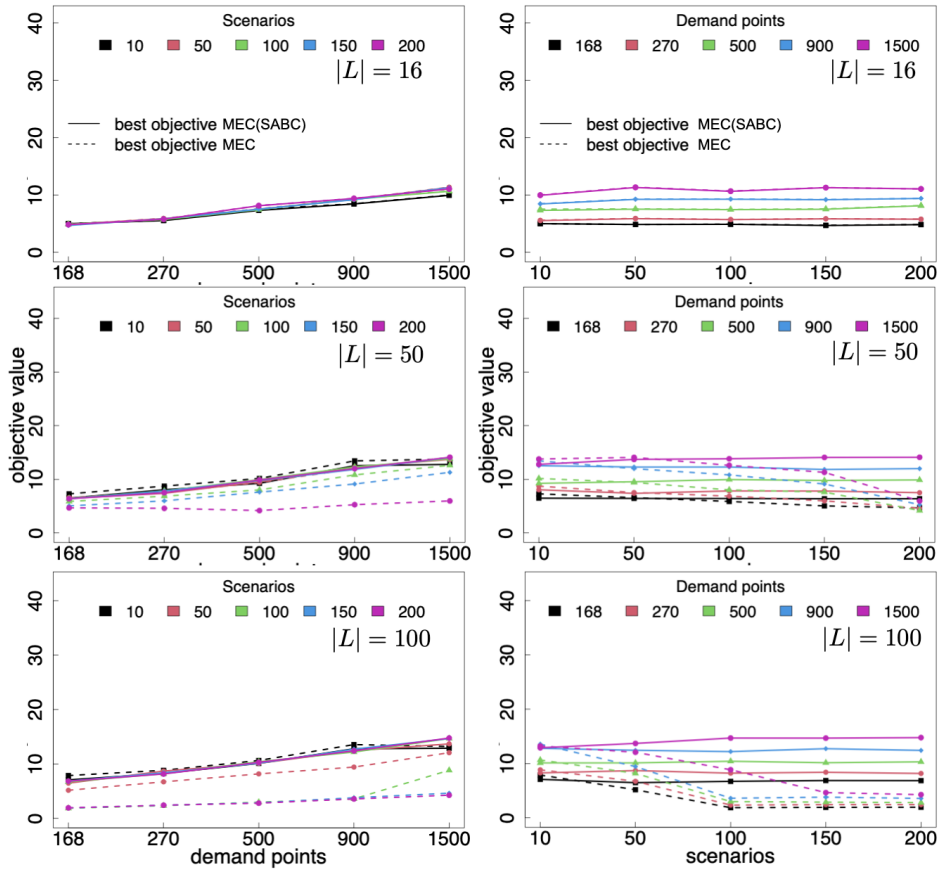
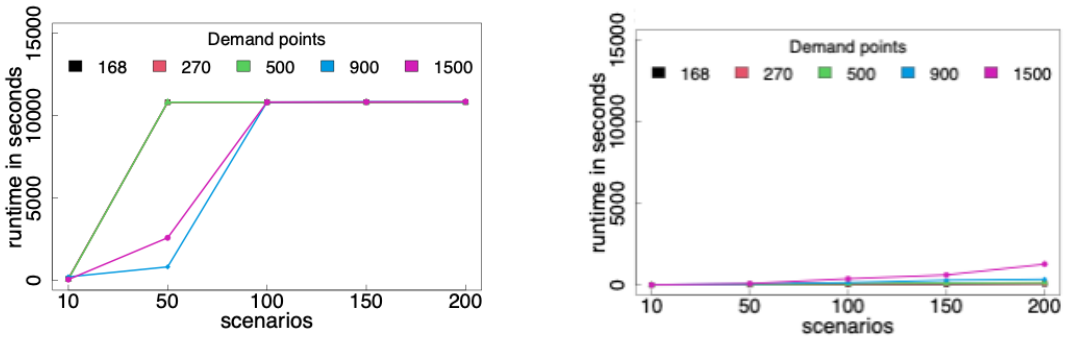


Figure 3: Comparison of the best objective values of MEC and MEC(SABC) concerning the demand points on the left side and the scenarios on the right side for different sizes of potential sites $|L| = \{16, 50, 100\}$.

number of emergency demand points. Recall that the MEC model with $|L| = \{50, 100\}$ reaches the time limit even for ten scenarios and few demand points.

Figure 4 a) shows that the principal disadvantage of the MEC model 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 the ambulances. The SABC model is extremely fast, even for large instances, and yields an initial solution for the ambulance location assignment in a short time to allow the MEC(SABC) model, Figure 4 b), to be solved faster than the MEC model and obtain high-quality solutions. The location-allocation strategy of the MEC(SABC) inherits not only its fast computational time from the SABC but also yields coverage per emergency situation, which is the main objective for the EVCP problem. The MEC(SABC) model is an approximated approach, but it gives solutions that are as good as the MEC and even better when the MEC instances do not reach optimality and its gaps are large. Most instances are solved by the MEC(SABC) intelligent feedback methodology in less than a minute.



(a) MEC with $|L| = 16$.

(b) MEC(SABC) with $|L| = 100$.

Figure 4: Computational time for the a) MEC and b) MEC(SABC) model with respect to the number of scenarios varying the number of emergency demand points.

An interesting advantage of the MEC(SABC) intelligent feedback method is that only one iteration is needed. Indeed, once the location of the ambulances has been retrieved from the SABC model and fed back to the MEC model, we could perturbate either randomly or with a local search, the allocation of the ambulances and iterate again. Nevertheless, we could not systematically generate a neighborhood around a location solution that yields better solutions with the MEC(SABC). This implies that local maximums are often reached with this first feed back and that complex or more diverse neighborhoods should be built to allow escaping from these solutions. Probably it would be interesting to enable local search movements that do not yield immediate benefits.

The objective values and execution times are crucial to evaluate the performance of the models. Nonetheless, the most important objective of the EVCP problem is to cover the largest number of demand points involved in the system within a fixed response time. Thus, a central question arises: is the emergency coverage quality of the MEC(SABC) as good as the one yielded by the MEC model?

The percentage emergency coverage of all instances is presented with the equivalent MEC model and the MEC(SABC) methodology in Figure 5. Two columns with three plots each, varying the number of scenarios and the location sites. Each plot shows the type of ambulance percentage coverage obtained by the a) MEC and b) MEC(SABC) methodology: T is for Total coverage (all required ambulances on time), TL is for Total-late coverage (all required ambulances, but at least one arrives late), P is for Partial coverage (at least one required ambulance is not dispatched, but the dispatched ones all arrive in time), PL is for Partial-late coverage (at least one required ambulances is not dispatched, at least one of the dispatched arrives late), and N for Null (no ambulances assigned to the demand point). The upper plots are for $|L| = \{16\}$ potential sites, the

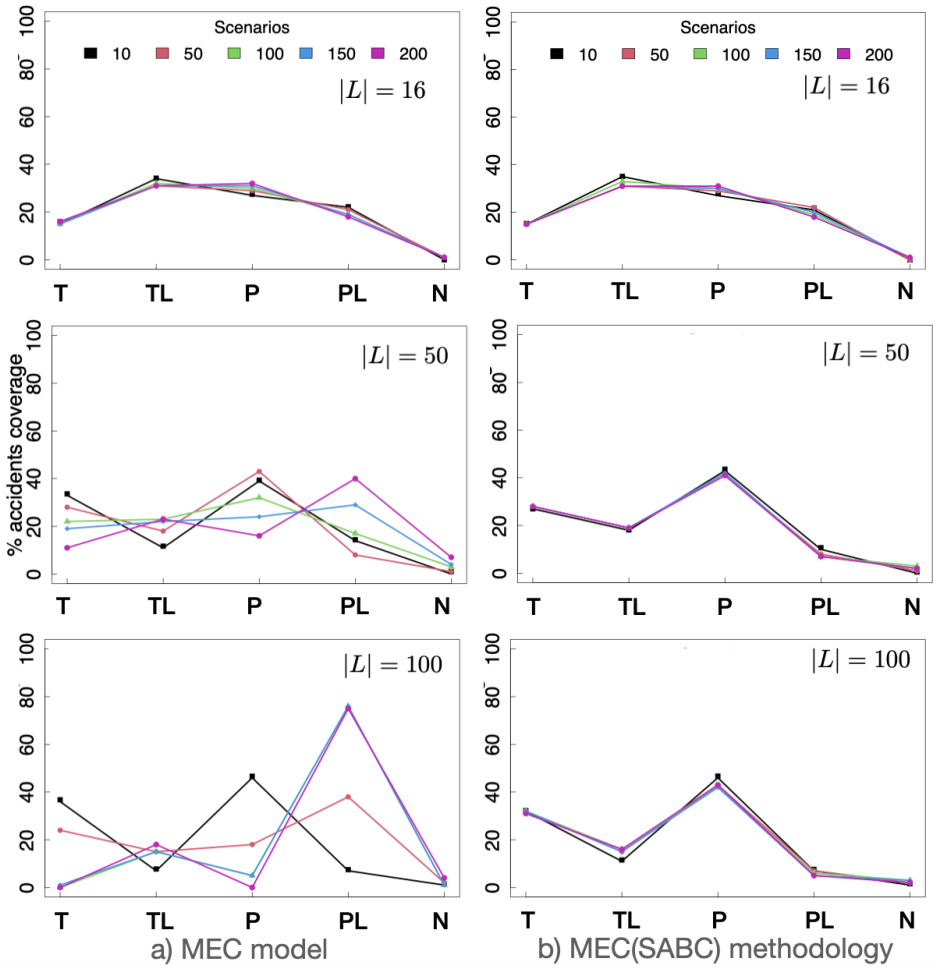


Figure 5: Coverage percentage per type obtained by the a) MEC model, b) SABC model, and the c) MEC(SABC) methodology for potential sites $|L| = \{16, 50, 100\}$.

middle ones for $|L| = 50$, and the lower ones for $|L| = 100$.

Figure 5 column a), shows that the MEC model 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 model. This behavior is probably linked to the large gaps obtained by the MEC model for large instances, but the number of null coverage is still remarkably low. Column b) shows that the MEC(SABC) methodology is robust in terms of the number of scenarios. That is, the demand point coverage is independent of the scenario number. In this manner, 100 scenarios are sufficient for handling a high-quality coverage solution. Moreover, the MEC(SABC) model 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, which is mainly desired in real life since it can be translated into first-aid medical care on time, increasing the probability of saving lives.

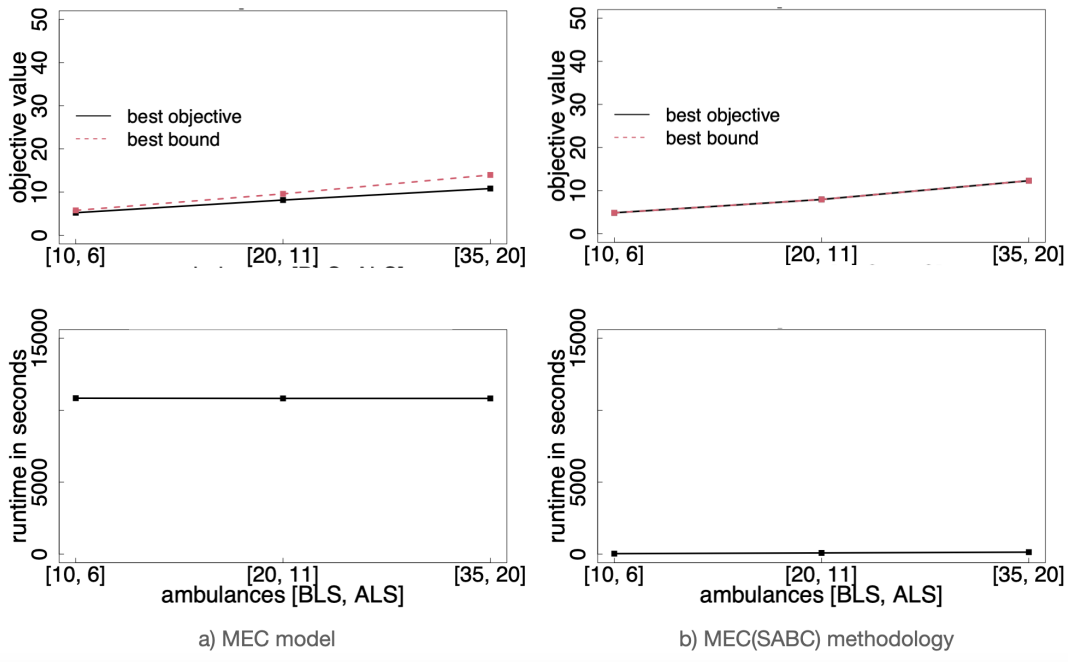


Figure 6: Objective value and execution time versus the number of ambulances for a) MEC model and b) MEC(SABC) methodology.

All the previous experiments were executed 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 the BLS one, which gives a more flexible setting but may induce difficulty when solving the models. Thus, what is the effect of the number of available ambulances in the EVCP problem on the objective function value and the running time?

We execute all the instances with emergency demand points fixed to 900, 100 scenarios, and 50 ambulance location sites. For this experiment, we vary the number of ambulances. In Figure 6, we show two columns of two plots each. The objective value (upper plots) and the execution time (lower plots) are on the y-axis, while the x-axis varies the number of ambulances: (η_1, η_2)=(10,6), (20,11), and (η_1, η_2)=m,(35,20). The left plots correspond to the MEC stochastic model, while the right ones are for the MEC(SABC) methodology.

From Figure 6 a), we observe that the difference between the best objective and the best bound for the MEC model (left plots) slightly increases with the number of ambulances. Thus, the larger the number of ambulances, the harder the instances for the MEC model. Meanwhile, the time limit is reached for every tested instance in the MEC model. For the MEC(SABC) methodology, the gaps are equal to 0 for all instances. Moreover, the objective values are comparable to the MEC model for all different settings of ambulances, which is a main characteristic. Moreover, the MEC(SABC) methodology solves the instances in less than one minute, and this computational time is unaffected by the number of ambulances.

5 Conclusions

EMS systems in developing countries, as in Mexico, lack many ambulance vehicles. Thus, one of the main contributions of this work is to deal with the problem of deciding whether an emergency will be totally, partially, or uncovered.

The *Emergency Vehicle Covering and Planning* (EVCP) problem is about locating the limited number of two heterogeneous types of ambulances in different city locations and dispatching them to the uncertain emergency points so as 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 partial coverage of the accidents by the ambulances based on a decay function.

We propose a novel two-stage stochastic program for the EVCP problem that can be solved by branch-and-bound for small instances with a restrictive number of scenarios. We also propose an Intelligent Feedback methodology, which is essentially a location-allocation procedure that relies on the solution of an auxiliary surrogate model, which is faster to solve. This method allows us to obtain high-quality solutions significantly faster than the previous approach. The Intelligent Feedback method was tested over a wide set of randomly generated instances based on real-world data from Monterrey. The proposed approach is significant because it can be implemented by calling any off-the-shelf integer solver without complex decomposition techniques.

Future research includes incorporating bi-level programming since several EMS operate in the city. Also, there is congestion in the hospitals, and not all emergencies can be dealt with. Moreover, patients are in the ambulance vehicle until they are admitted to the hospital emergency area. With respect to methodology, specialized neighborhoods need to be defined to avoid local maximums in the intelligent feedback method.

Acknowledgements

Beatriz Alejandra García-Ramos wishes to acknowledge a graduate scholarship from the National Council of Humanities, Science and Technology (CONAHCYT) of Mexico, grant 860555. Yasmín Ríos Solís acknowledges ANUIES-CONAHCYT for the grant M20M01-315691.

References

[1] 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.

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

[3] 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.

[4] 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.

[5] 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.

[6] 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.

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

[8] John R Birge and Francois Louveaux. *Introduction to stochastic programming*. Springer Science & Business Media, 2011.

[9] 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.

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

[11] 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.

[12] 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.

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

[14] L. A. McLay and H. Moore. Hanover county improves its response to emergency medical 911 patients. *Interfaces*, 42(4):380–394, 2012.

[15] Nimrod Megiddo and Arie Tamir. On the complexity of locating linear facilities in the plane. *Operations research letters*, 1(5):194–197, 1982.

[16] 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.

[17] N. Noyan. Alternate risk measures for emergency medical service system design. *Annals of Operations Research*, 181:559–589, 2010.

[18] 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.

[19] 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.

[20] 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.

[21] 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, 2018.

[22] H. Toro-Díaz, M. E. Mayorga, S. Chanta, and L. A. Mclay. Joint location and dispatching decisions for emergency medical services. *Computers & Industrial Engineering*, 64(4):917–928, 2013.

[23] 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.

[24] 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.

[25] 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.

[26] Y. Wang, K. L. Luangkesorn, and L. Shuman. Modeling emergency medical response to a mass casualty incident using agent based simulation. *Socio-Economic Planning Sciences*, 46 (4):281–290, 2012.

[27] 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.

[28] Y. Zhang, Z. Li, and Y. Zhao. Multi-mitigation strategies in medical supplies for epidemic outbreaks. *Socio-Economic Planning Sciences*, 87:101516, 2023.

[29] 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.