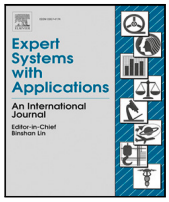
Expert Systems With Applications 249 (2024) 123665

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



A stochastic programming model for ambulance (re)location–allocation under equitable coverage and multi-interval response time

Imanol Gago-Carro [a,b,*], Unai Aldasoro [c], Josu Ceberio [d], María Merino [a,b]

[a] BCAM - Basque Center for Applied Mathematics, Spain
[b] Department of Mathematics, University of the Basque Country UPV/EHU, Spain
[c] Department of Applied Mathematics, University of the Basque Country UPV/EHU, Spain
[d] Department of Computer Science and Artificial Intelligence, University of the Basque Country UPV/EHU, Spain

ARTICLE INFO

Keywords:
Stochastic programming
Location–allocation
OR in health services
Regional equity
Multi-interval response time

ABSTRACT

Emergency Medical Services are essential for health systems as their effective management can improve patient prognosis. Nevertheless, designing an optimized distribution of resources is a difficult task due to the complex nature of these systems. Moreover, locating the resources is particularly challenging in heterogeneous density territories where, in addition to their efficient management, the equity principle in the medical access of inhabitants of rural areas is also desirable.

This paper approaches the ambulance (re)location–allocation problem in the geographical area of the Basque Country. The area has three major cities, which account for a third of the emergencies, while there are few emergencies in rural areas, with a sparse population. To that end, a two-stage stochastic 0-1 integer linear programming model that balances the response time between densely populated and isolated areas is proposed. Specifically, the model incorporates two relevant principles: (1) optimizing emergency attendance through the option of allocating ambulances via a multi-interval response time and (2) equitably responding to emergencies so remote areas are not neglected. Conducted experiments have been validated and indicate that the proposed model can improve the success rate in rural areas by 23 percentage points, while reducing the overall success rate by less than 9 percentage points.

1. Introduction

1.1. Management challenges

Effective management of Emergency Medical Service (EMS) vehicles is crucial for saving people's lives. The corresponding decision-making process frequently includes the location of ambulance stations (strategic phase) and call allocation (operational phase). The main performance measures in the literature rely on the Response Time (RT), which (Aboueljinane et al., 2013) defined as the period between the receipt of a call and the first arrival of a rescue team at the scene.

However, healthcare providers face a trade-off between global behavior and the underlying inequities in heterogeneous regions. For example, concerning the location of ambulances, cities with high population density are usually the preferred sites, to the detriment of rural areas. Additionally, as regards call allocation, call center operators aim to provide the fastest possible response to a call while ensuring that they do not leave areas without assistance. Therefore, a need often arises for managerial frameworks that incorporate multiple intervals.

1.2. Literature review

The literature related to the (re)location–allocation problem is extensive and covers different singularities and points of view: we refer the interested reader to the healthcare resources location–allocation models survey works by Bélanger et al. (2019) and Brotcorne et al. (2003) for extensive reviews. The problem is tackled in many different ways, but the (re)location–allocation models can be divided into two main categories: *static location models* and *dynamic relocation models*.

Static location models focus on strategic decisions such as selecting the appropriate location for the stations or deciding on the suitable number of ambulances in each station. The objective function of these models varies from one to another: Toregas et al. (1971), for instance, present the Location Set Covering Model (LSCM). This model minimizes the number of ambulances required to cover an area. On the contrary, the model presented by Church and ReVelle (1974), the Maximal Covering Location Problem (MCLP), maximizes the covered area for

* Correspondence to: Barrio Sarriena s/n. 48940 Leioa, Basque Country, Spain.
E-mail addresses: igago004@ikasle.ehu.eus, igago@bcamath.org (I. Gago-Carro), unai.aldasoro@ehu.eus (U. Aldasoro), josu.ceberio@ehu.eus (J. Ceberio), maria.merino@ehu.eus (M. Merino).

https://doi.org/10.1016/j.eswa.2024.123665
Received 18 July 2023; Received in revised form 8 February 2024; Accepted 10 March 2024
Available online 14 March 2024

a given number of facilities. Gendreau (1997) presents the Double Standard Model (DSM), which ensures a double coverage using two different time thresholds. All the demand is covered in a particular time r_2, and part of the demand is covered in less time r_1 ($r_1 < r_2$). The main issue with all the above-mentioned deterministic models is that they do not consider the moments when the resources are occupied and consider that they are always available. This issue is avoided in the following stochastic model. Daskin (1983) presents the Maximum Expected Covering Location Problem (MEXCLP), which maximizes the expected coverage by considering the unavailability of resources.

Dynamic relocation models are usually real-time multi-period models that have to be revised in each period to take into account the changes in the system that have occurred up to that point. These models introduce the possibility of the ambulances being relocated every time a decision has to be made, i.e., every time an emergency call has to be allocated. One of the first relocation dynamic works is the one developed by Gendreau et al. (2001), which is tested with actual data from Montreal, Canada. Schmid (2012) uses dynamic programming to follow a relocation strategy. The work is also tested with real data, in that case, with data from Vienna (Austria). Karpova et al. (2023) develop heuristic algorithms to dynamically relocate ambulances and improve the prehospital care services. When it comes to approaching the ambulance (re)location–allocation problem, several research lines are followed: queuing theory models (Larson, 1974; Takeda et al., 2007), simulations (Bell & Allen, 1969 and a survey by Aboueljinane et al., 2013) and mathematical optimization (Aringhieri et al., 2017).

Undoubtedly, EMS inherently contains several sources of uncertainty, such as where and when the following emergency will occur, how long a busy ambulance will be occupied, or the urgency level of the emergency. Therefore, stochastic programming is a useful framework that deals with real-world variability and uncertainty. Stochastic modeling allows us to consider several scenarios, representing the disparity of the situations that can occur. Better decisions can improve the current emergency system by considering all these scenarios and giving them their importance. The stochastic programming model presented by Beraldi et al. (2004) aims to solve dimensioning and location problems using probabilistic constraints. Two-stage models have been used to deal with the stochasticity of the problem: Naoum-Sawaya and Elhedhli (2013) present a stochastic model for relocating ambulances. Noyan (2010) presents a two-stage stochastic model with stochastic demand. Nickel et al. (2016) also deal with stochastic demand and present a two-stage stochastic programming model to optimize the location and number of ambulances and their stations. More recently, Yoon et al. (2021) present a two-stage stochastic model with two types of ambulances, and Wang et al. (2022) formulate a two-stage stochastic programming model incorporating uncertainty in emergency demand and traffic congestion.

Additionally, the characteristics of the area where the problem is addressed greatly influence the optimization approaches. The terrain of the area (Humphreys et al., 2012; Swan et al., 2008), the number of ambulance stations (McLay & Mayorga, 2010) and, particularly, the heterogeneity population density across the country (Leknes et al., 2017) are critical when designing efficient models. Emergency calls are generally concentrated in densely populated areas, while there are far fewer calls in isolated areas per year (Leknes et al., 2017). Accordingly, when optimizing any resource allocation by minimizing the general response time, proposed models tend to ignore the calls in isolated areas and focus their attention on cities and their suburbs (Bélanger et al., 2016). In fact, optimizing ambulance allocation in areas with different population densities has been a recurrent topic in this research field (Erkut et al., 2008; Jagtenberg & Mason, 2020) as ethical issues arise related to the equity of the access of inhabitants in isolated areas to the medical emergency services. In this context, some studies seek to solve the allocation problem by enabling equity decisions (Chanta et al., 2014; Smith et al., 2013). To that end, the term *rurality* is proposed to characterize the areas according to their population density: those that are densely populated, also known as *urban*, and those weakly populated, referred to as *rural*. Intermediate terms such as *suburban* (Clement et al., 2018) or subcategories such as *rural-peripheral* and *rural-accessible* (Jonard et al., 2007) are sometimes used as well. However, there is no agreement in the literature on the definition of rurality. Most research defines it as a characteristic related to geography: Humphreys et al. (2012) and Swan et al. (2008) propose using indicators such as population density or distance to the nearest resources to measure rurality. Conversely, Phillimore and Reading (1992) define rurality as a set of community characteristics, such as being a commuter, industrial, or agricultural village. Rousseau (1995) and, recently Kaneko et al. (2021) propose a wide range of definitions of rurality. As Karsu and Morton (2015) state, there are three main approaches to looking for more equitable solutions: (1) methods based on the maximin principle by Rawls (1971), which optimize the worst-off outcome of a rural area, (2) incorporating inequity indexes directly in the model (for instance, Gutjahr and Fischer (2018) use the Gini coefficient and McLay and Mayorga (2013) use the range between the minimum and maximum outcomes) and (3) using inequity-averse aggregation functions which not only focus on equity but also try to obtain the most efficient solution (Marín et al., 2010). Recently, Xinying Chen and Hooker (2023) provided a study about incorporating inequity measures to mathematical optimization models. In addition, it also includes a compilation of formulations made in previous studies and several references to the literature covering these aspects from an ethical perspective (Lamont & Favor, 2017).

Equity between rural–urban areas is only one of many difficulties when addressing the (re)location–allocation problem. As discussed above, minimizing emergency RT is essential, but it is challenging to model the different responses when allocating ambulances to emergencies. In this sense, the model developed by Schmid (2012) minimizes the average RT of all emergencies. Jagtenberg et al. (2015) and Naoum-Sawaya and Elhedhli (2013), on the contrary, take a response time threshold into account, surpassing which implies a failure to respond to an emergency successfully. Hence, Naoum-Sawaya and Elhedhli (2013) divide the emergency responses into two groups, those under the target time and those over it, which are unacceptable. Nevertheless, this dichotomous division means that attending an emergency in a remote location is impossible when it cannot be reached by any available ambulance. Consequently, different time response thresholds may be added, with greater importance given to lower RT-intervals.

1.3. Contributions and limitations

In this paper, we approach the optimal (re)location–allocation of the ambulances of the Basque Public EMS system. To that end, we present a two-stage stochastic 0–1 integer linear programming model. Locating the ambulances to stations is optimized in the first-stage, and resources are allocated to emergencies in the second. The changes to the ambulance fleet can be carried out in two different ways. The first, which we refer to as *relocation*, is by changing the location of current resources. The second possible change is to add new ambulances to the fleet; we use the *location* term for this case. Unlike the Maximum Covering Model, our model formulation enables tracking the workload carried out by all ambulances: when are where they were allocated. From a management point of view, this is very useful because decision-makers could analyze the occupancy rate of each ambulance and its performance.

We designed the model to deal with the intrinsic characteristics of the Basque Country: three major urban areas which account for a third of the emergencies, along with isolated areas with a sparse population in steep terrain. In particular, the proposed model introduces a new definition of rurality that classifies each municipality of the region according to the number of calls received in its catchment area. This new definition, together with adding a regional equity component to the model, means areas of the Basque Country are not left unattended.

Moreover, we introduce a multi-interval response time to the model, enabling faster response time. We use actual data from the Basque Public EMS system to implement the model. We look for a balance between equity and efficiency using an inequity-averse weighted sum function as the objective function. By varying the weights of this function, we compare several variants of the model, ranging from the most efficient to the most equitable one. A sequence of results is also presented and discussed.

Moreover, as the number of resources that can be added to the fleet is limited due to its high cost, the model settles for several prefixed maximum numbers of changes that can be made to compare the obtained results. As the terrain of the Basque Country is not regular (there are urban, mountainous, and coastal areas), and the habits of its inhabitants vary throughout the year, the distribution of emergencies differs between hours of the day, days of the week, and months of the year. Experiments are conducted for various randomly chosen sets of days and for the day time interval with more activity. The results are validated for the whole year under two different scopes. All the work presented in this paper has been discussed with the decision-makers of the Basque Public EMS system. Together with their previous experiences and knowledge, it can be a valuable tool to improve the current situation of the EMS.

Regarding the limitations, the models proposed in this article suffer from some drawbacks: (a) since the available data only includes the municipality where emergencies occur, we assumed that all emergencies take place at the center of each municipality. While this assumption may have a negligible impact on small municipalities, it could deviate more from reality in larger municipalities where the distances are greater; (b) we have considered deterministic average response times based on data from Google Developers (2020); (c) although the proposed models are designed to be generic, the results obtained are specific to the Basque Country and, therefore, a priori not extrapolable to other regions. In addition, it is worth mentioning the irregular orography of the Basque Country, which has urban and rural areas and mountain and coastal areas.

The rest of the paper is organized as follows: the socio-geographical context of the Basque Country is explained in Section 2. Section 3 provides the problem modeling, the (re)Location–Allocation Baseline (LAB) model and the Equitably Multi-Interval (re)Location–Allocation (EMILA) model. The inputs and calibration of the models, together with the computational experiments for performance assessment and their validation, are considered in Section 4. Finally, the conclusions and future research lines are discussed in Section 5.

2. Case study: Ambulance (re)Location–Allocation problem for the basque public EMS system

The case study conducted to develop the models presented in this paper is set in the Basque Country, a region of 7234 km^2 in the north of Spain. This region is divided into three provinces (Araba, Bizkaia, and Gipuzkoa), which, in turn, are divided into 251 municipalities. The population density in each of these municipalities is heterogeneous (see Fig. 1(a)): while there are some cities where the population density exceeds 5000 inhabitants per km^2, there are also rural municipalities with less than ten inhabitants per km^2. As expected, the number of emergencies depends, to a large extent, on the population of each municipality.

The Basque Public EMS system is responsible for providing the required response to all emergency calls received. To that end, the organization has a call center that manages the responses by assigning the appropriate ambulance to each call. Nowadays, the Basque Public EMS system has a fleet of 88 ambulances distributed in 80 base stations. This fleet can be divided into two groups: 11 Advanced Life Support (ALS) ambulances, with a doctor onboard and which take care of the most serious emergencies, and 77 Basic Life Support (BLS) ambulances, which are usually assigned to medium-low urgency emergencies and, occasionally, are equipped with a nursing technician. Due to the limited number of ALS and their importance, the scope of the research focuses on optimizing the ALS fleet's location and service.

According to the Basque Public EMS Contract Program (Emergentziak Osakidetza, 2019), an emergency is considered to be responded to in time if an ambulance arrives within 15 min. Moreover, the Basque Public EMS system considers that the success rate (the percentage of emergencies reached in time) should be over 75%. Not all emergencies are taken into consideration to measure these success rates. False alarms and special operations (for example, previously scheduled interhospital transfers) are not included in the success rate measures. We conducted a data analysis of the response times for all the emergencies in this region in 2019. This analysis shows that the success rate varies depending on the area of the Basque Country. Table 1 shows that while more dense areas (>5000 inhabitants/km^2) had 3953 calls and 3613 full-service calls (those emergency calls considered to measure the goodness of the system) and a success rate of 90.3%, there were hardly any emergencies in low-density areas (36 emergency calls and 30 full-service calls) and the success rate plummeted to 36.7%. These differences are maintained if we only focus on the time slot between 9:00 and 17:00, the times of the day with the highest emergency per hour rate. Because of such differences, we classify the Basque Country's municipalities according to a call index defined as the number of emergencies in 2019 that an ambulance located in a municipality could have been able to handle in the Threshold Time (TT) (see Fig. 1(b)). In this way, municipalities with a high call index are those where many emergencies occur and which are usually large cities or towns. Conversely, a municipality with a low call index implies that there will not be many emergencies in that municipality or nearby ones. In essence, the lower the call index is, the higher its rurality is, and vice versa.

3. Problem modeling

This section presents two two-stage stochastic 0–1 integer linear programming models to approach the optimal (re)location for the ambulances and the suitable allocations of ambulances to emergencies. We will represent uncertainty by random variables defined on a probability space $(\Omega, \mathcal{F}, P)$, where Ω is a discrete finite set of all possible outcomes (random parameter values related to geographical and time issues of the emergencies), equipped with the σ-algebra $\mathcal{F}$ of all its subsets and $P : \mathcal{F} \to [0, 1]$ a probability measure.

In two-stage problems, two types of decisions are usually represented in the models with variables. The values of first-stage variables are chosen before the uncertainty of the problem is resolved. Once the randomness of the problem is revealed and conditioned by the solution of the first-stage variables, second-stage decisions have to be made. In the problem we face, the first-stage decisions are tactical and consist of (re)locating ambulances to stations. Then, the randomness of the problem, which consists of the location, time, and type of emergency, is revealed, and second-stage decisions have to be made. These operational decisions determine how the emergency calls are responded to, allocating the most appropriate ambulance to each call in each scenario. Both stages are solved simultaneously, and the (re)locations are made considering the emergencies in all the randomly sampled scenarios. Section 3.1 presents a single-objective two-interval variant of the model. In Section 3.2, a regional equity component and the option of dividing the response time in a weighted hierarchy are added to the model, obtaining a bi-objective multi-interval (re)location–allocation model.

3.1. (re)Location–Allocation Baseline (LAB) model

The LAB model presented in this section is inspired by the model presented by Naoum-Sawaya and Elhedhli (2013). Although there are some similarities, such as the definition of variables of the first-stage,

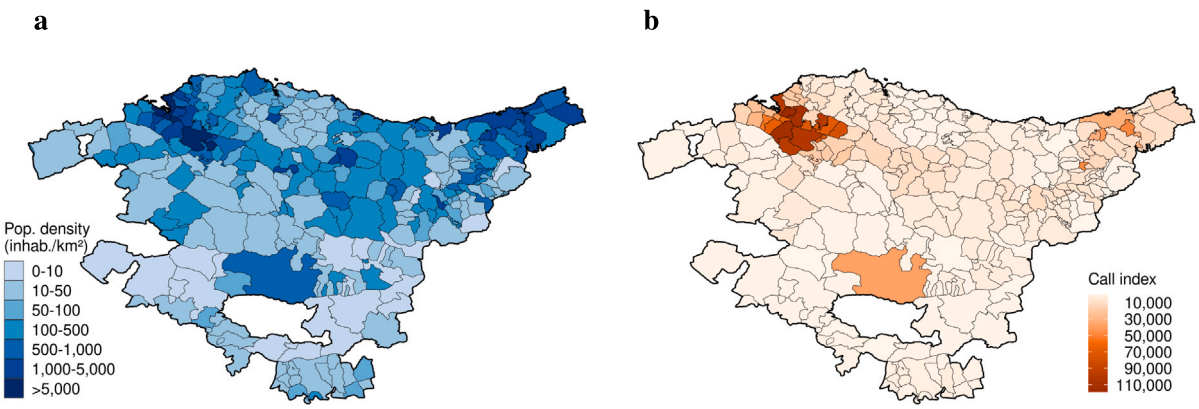


Fig. 1. Population density (a) and call index (b) of the Basque Country by municipality: a low number in the call index (which indicates the number of emergencies that occurred within 15 min of the municipality during 2019) represents a rural area, while most urban municipalities have a high number.

Table 1
Total number of emergencies, full-service emergencies, success rate, and emergencies per hour during 2019 by density area and time slot.

Density group	Number of	0:00–24:00				9:00–17:00			
(inhab./km²)	municipalities	Number of emergencies		Success rate	Emergencies per	Number of emergencies		Success rate	Emergencies per
		Tot.	Full-serv.	(%)	hour	Tot.	Full-serv.	(%)	hour
0–10	13	36	30	53.3	0.00	11	9	66.7	0.00
10–50	79	298	229	56.8	0.03	148	113	58.4	0.05
50–100	41	406	369	59.9	0.05	189	161	57.8	0.06
100–500	67	1,960	1,680	70.9	0.22	879	743	71.6	0.30
500–1000	26	4,158	3,869	84.1	0.47	1822	1676	85.0	0.62
1000–5000	19	4,443	3,991	90.6	0.51	2006	1784	90.6	0.69
>5000	6	3,953	3,613	94.2	0.45	1707	1541	94.6	0.58
Total	251	15,254	13,781	85.8	1.74	6762	6027	86.2	2.32

the model is generalized, adding new relevant contributions: we introduce a new RT-interval to make the model closer to reality as we add the opportunity to respond to emergencies either successfully (faster than the TT) or unsuccessfully (within a second time threshold TT′). Another significant change is the classification of the call types to interact differently with each of them, giving them the priority they deserve. Moreover, a limitation to the number of (re)locations is added to measure the effect of increasing the number of changes allowed. Additionally, the model can also be used to find the optimal location of new ambulances, changing from relocation–allocation to location–allocation by adapting the set of available resources. A final contribution to the model concerns the objective function where, unlike the design of Naoum-Sawaya and Elhedhli (2013), we maximize the number of successfully answered calls in this paper.

Let us define the sets, parameters, and variables of the model:

Sets:

- I, set of resources, $i \in I$, where I_0 is the initial vehicle fleet, and I_1 is the set of resources that can be added to the fleet, $I = I_0 \cup I_1$.
- J, set of potential stations, $j \in J$.
- Ω, set of scenarios (days of the year), $\omega \in \Omega$.
- E_ω, set of emergency calls in scenario ω, $\omega \in \Omega, e \in E_\omega$. Where E^*_ω is the subset of full-service calls, i.e., those calls that are taken into consideration to measure the goodness of the system, $E^*_\omega \subset E_\omega$.

Parameters:

- n', number of emergency calls, where n'_ω is the number of calls in scenario ω, $\sum_{\omega\in\Omega} n'_\omega = n'$.
- n, number of full-service calls, where n_ω is the number of full-service calls in scenario ω, $\sum_{\omega\in\Omega} n_\omega = n$.
- δ_{ij}, 1 if assigning location j to ambulance i involves a (re)location; 0 otherwise, $i \in I, j \in J$.
- c_j, capacity of station j, $j \in J$.
- β, rate of calls that must be attended, $\beta \in (0, 1]$.
- π, rate of full-service calls that must be attended within the TT, $\pi \in (0, 1]$.
- k, number of changes in the system allowed.

Stochastic parameters:

- $u_{e\omega}$, time an ambulance that is attending emergency e of scenario ω in less than TT will be unavailable to attend a new call, $\omega \in \Omega, e \in E_\omega$. Let us define $U_{e\omega}$ as the set of these emergencies.
- $u'_{e\omega}$, time that an ambulance that is attending emergency e of scenario ω between TT and TT′ will be unavailable to attend a new call, $\omega \in \Omega, e \in E_\omega$. Let us define $U'_{e\omega}$ as the set of these emergency calls.
- $a_{je\omega}$, 1 if call e of scenario ω is reachable from station j in less than TT; 0 otherwise, $j \in J, \omega \in \Omega, e \in E_\omega$.
- $a'_{je\omega}$, 1 if call e of scenario ω is reachable from station j in over TT, but less than TT′; 0 otherwise, $j \in J, \omega \in \Omega, e \in E_\omega$.
- $\varepsilon_{ie\omega}$, 1 if ambulance i is operational for attending emergency call e of scenario ω; 0 otherwise, $i \in I, \omega \in \Omega, e \in E_\omega$.
- p_ω, likelihood for each scenario $\omega \in \Omega$, $\sum_{\omega\in\Omega} p_\omega = 1$.

Variables:

- y_{ij}, 1 if ambulance i is assigned to location j; 0 otherwise, $i \in I, j \in J$.
- $x_{ie\omega}$, 1 if ambulance i attends emergency e of scenario ω and it arrives in less than TT; 0 otherwise, $i \in I, \omega \in \Omega, e \in E_\omega$.
- $x'_{ie\omega}$, 1 if ambulance i attends emergency e of scenario ω and it arrives between TT and TT′; 0 otherwise, $i \in I, \omega \in \Omega, e \in E_\omega$. As an example, ambulances that are only operational during the summer months cannot be used during the rest of the year
- z_ω, success rate in scenario ω, defined as the percentage of full-service calls responded to in less than TT, $\omega \in \Omega$.

We consider the following assumptions:

1. Emergency responses are classified into three groups: responses to emergencies in no more than the demanded TT or between TT and TT′ and a third group of non-attended emergencies. We consider that no emergency can be responded to over TT′, so the model does not contemplate assigning an ambulance to an emergency that is too far away.
2. The ambulances are allocated to the calls instantly. Consequently, the model does not allow delayed allocations.

Let us define $Z = \{z_\omega\}_{\omega\in\Omega}$ as a discrete random variable over the set of scenarios Ω, representing the success rate. It takes the value z_ω with likelihood p_ω for each scenario ω, where

$$p_\omega = P(Z = z_\omega) = \frac{n_\omega}{n} \qquad \text{and} \qquad z_\omega = \frac{1}{n_\omega}\sum_{i\in I}\sum_{e\in E_\omega} x_{ie\omega} \tag{1}$$

The LAB model is defined as a two-stage stochastic 0–1 integer linear programming model as follows:

$$\max \quad \mathbb{E}_\Omega[Z] \tag{2a}$$

$$s.t. \quad \sum_{j\in J} y_{ij} \le 1 \qquad \forall i \in I \tag{2b}$$

$$\sum_{i\in I} y_{ij} \le c_j \qquad \forall j \in J \tag{2c}$$

$$\sum_{i\in I}\sum_{j\in J} \delta_{ij}\cdot y_{ij} \le k \tag{2d}$$

$$x_{ie\omega} - \sum_{j\in J} a_{je\omega}\cdot y_{ij} \le 0 \qquad \forall i \in I,\ \omega\in\Omega,\ e\in E_\omega \tag{2e}$$

$$x'_{ie\omega} - \sum_{j\in J} a'_{je\omega}\cdot y_{ij} \le 0 \qquad \forall i \in I,\ \omega\in\Omega,\ e\in E_\omega \tag{2f}$$

$$x_{ie\omega} + x_{ie'\omega} + x'_{ie'\omega} \le 1 \qquad \forall i \in I,\ \omega\in\Omega,\ e\in E_\omega,\ e'\in U_{e\omega} \tag{2g}$$

$$x'_{ie\omega} + x_{ie'\omega} + x'_{ie'\omega} \le 1 \qquad \forall i \in I,\ \omega\in\Omega,\ e\in E_\omega,\ e'\in U'_{e\omega} \tag{2h}$$

$$\sum_{i\in I}(x_{ie\omega} + x'_{ie\omega}) \le 1 \qquad \forall \omega\in\Omega,\ e\in E_\omega \tag{2i}$$

$$\sum_{i\in I}\sum_{\omega\in\Omega}\sum_{e\in E_\omega}(x_{ie\omega} + x'_{ie\omega}) \ge \beta\cdot n' \tag{2j}$$

$$\mathbb{E}_\Omega[Z] \ge \pi \tag{2k}$$

$$x_{ie\omega} + x'_{ie\omega} \le \varepsilon_{ie\omega} \qquad \forall i \in I,\ \omega\in\Omega,\ e\in E_\omega \tag{2l}$$

$$y_{ij},\ x_{ie\omega},\ x'_{ie\omega} \in \{0,1\} \qquad \forall i \in I,\ j\in J,\ \omega\in\Omega,\ e\in E_\omega \tag{2m}$$

The objective function (2a) maximizes the expected global success rate over the set of scenarios. The expression of the objective function is equivalent to (3).

$$\mathbb{E}_\Omega[Z] = \sum_{\omega\in\Omega} p_\omega\cdot z_\omega = \frac{1}{n}\sum_{\omega\in\Omega}\sum_{i\in I}\sum_{e\in E^*_\omega} x_{ie\omega} \tag{3}$$

Constraints (2b) ensure that ambulances always return to their station. Constraints (2c) prevent a station from allocating more ambulances than its capacity. Constraint (2d) restricts the number of changes (either relocations or locations) in the system. Constraints (2e) state that if ambulance i attends emergency e of scenario ω in less than TT, then e is reachable in TT from the station of the ambulance. Similarly, constraints (2f) state that if ambulance i attends emergency e of scenario ω between TT and TT′, then emergency e is reachable in that time from the station of the ambulance. Constraints (2g) and (2h) ensure that if ambulance i is attending emergency e of scenario ω, then that ambulance will be unavailable for the following $u_{e\omega}$ (or $u'_{e\omega}$) seconds, a period of time that depends on the type of emergency. Constraints (2i) hold that at most one ambulance can be allocated to emergency call e of each scenario ω. Constraint (2j) forces that at least $\beta\cdot 100\%$ of calls must be handled. Similarly, constraint (2k) forces that at least $\pi\cdot 100\%$ of full-service calls taken into consideration to measure the goodness of the system must be handled in time. This constraint forces that a minimum level of compliance must be fulfilled. Constraints (2l) prevent a non-operational ambulance from being assigned to an emergency. Finally, the integrality conditions for the binary variables are given in the constraints (2m).

3.2 Equitable Multi-Interval (re)Location–Allocation (EMILA) model

As set out above, the LAB model (2) maximizes the percentage of emergencies attended in time. However, when the response to an emergency exceeds the TT, the model has no preferences in the response time: it can indistinctly allocate a resource that is TT + 1 away or one that is TT′ away to the emergency. For example, in the study case of the Basque Country, the allocation of an ambulance that is 16 min away from the emergency has the same impact on the objective function as one that is 45 min away. We improve the model by defining a new set of RT-intervals to prioritize sooner responded emergencies. Moreover, as previously mentioned in Section 2, the studied region is divided into different population densities and emergency call activity areas. The success rate in those areas varies according to the population density, and remarkable differences exist. We incorporate a regional equity component into the model, which induces conflicting criteria for optimization, to reduce the differences in the success rate. On the one hand, the aim is to respond to the maximum possible number of emergencies in time. On the other hand, equity in terms of regions is sought. Thus, the model is updated by adding the following new sets, parameters, and variables to balance the two criteria.

Sets:

- $\mathcal{L}$, set of RT-intervals bounded by their corresponding thresholds, $\ell \in \mathcal{L}$.
- R, set of regions, $r \in R$.
- $E_{\omega r}$, subset of emergency calls originated in region r, $r \in R, \omega \in \Omega$.

Parameters:

- μ^ℓ, the priority of attending calls in RT-interval $\ell \in \mathcal{L}$, such as $\mu^{\ell_1} \ge \mu^{\ell_2} \ge \cdots \ge \mu^{\ell_{|\mathcal{L}|}} \ge 0$.
- n^r, number of full-service calls in region r, $r \in R$, where $\sum_{r\in R} n^r = n$.
- m, number of full-service calls of the region with the most full-service calls: $m = \max_{r\in R} n^r$.
- α, inequity-aversion parameter, $\alpha \in [0,1]$.

Stochastic parameters:

- $u^\ell_{e\omega}$, time an ambulance attending emergency e of scenario ω in RT-interval ℓ will be occupied and, therefore, unavailable to attend a new call, $\ell \in \mathcal{L}, \omega \in \Omega, e \in E_\omega$.
- $a^\ell_{je\omega}$, 1 the emergency call e of scenario ω is reachable from station j in RT-interval ℓ; 0 otherwise, $j \in J, \omega \in \Omega, e \in E_\omega, \ell \in \mathcal{L}$.
- p^r_ω, likelihood for each scenario $\omega \in \Omega$ when referring to the region r, $\sum_{\omega\in\Omega} p^r_\omega = 1, r \in R$.

Variables:

- $x^\ell_{ie\omega}$, 1 if ambulance i attends emergency call e of scenario ω and it arrives in RT-interval ℓ time; 0 otherwise, $i \in I, \omega \in \Omega, e \in E_\omega, \ell \in \mathcal{L}$.
- z^ℓ_ω, percentage of full-service calls responded to in the RT-interval ℓ, $\omega \in \Omega$, $\ell \in \mathcal{L}$.
- $z^{\ell,r}_\omega$, percentage of full-service calls of the region r responded to in the RT-interval ℓ, $\omega \in \Omega$, $\ell \in \mathcal{L}$, $r \in R$.

Assumptions:

1. Attended emergencies are classified into $|\mathcal{L}|$ intervals. The first interval (ℓ_1) consists of response times not exceeding threshold TT. The LAB model can be recovered when $|\mathcal{L}| = 2$.
2. α is a parameter whose value must be chosen by the decision-makers and which is used to balance the relevance of efficiency and equity components in the model. When $\alpha = 0$, efficiency

is maximized. On the contrary, when $\alpha = 1$, the model gives the same importance to success rates of all regions, and regional equity will be sought. Intermediate values of α balance both objectives.

Now, for each RT-interval $\ell \in \mathcal{L}$, let us define Z^ℓ as a discrete random variable over the set of scenarios Ω that measures the percentage of full-service calls responded to in the RT-interval ℓ. The variable Z^ℓ takes the value z^ℓ_ω with likelihood p_ω, see (4). In addition, for each $\ell \in \mathcal{L}$ and $r \in R$, let us define $Z^{\ell,r}$ as a discrete random variable over the set of scenarios Ω that measures the percentage of full-service calls occurred in region r that are responded to in the RT-interval ℓ. It takes the value $z^{\ell,r}_\omega$ with likelihood p^r_ω, see (5):

$$p_\omega = P(Z^\ell = z^\ell_\omega) = \frac{n_\omega}{n} \qquad \text{where} \quad z^\ell_\omega = \frac{1}{n_\omega} \sum_{i\in I} \sum_{e\in E^*_\omega} x^\ell_{ie\omega} \tag{4}$$

$$p^r_\omega = P(Z^\ell = z^\ell_\omega \mid R = r) = \frac{n^r_\omega}{n^r} \qquad \text{and} \quad z^{\ell,r}_\omega = \frac{1}{n^r_\omega} \sum_{i\in I} \sum_{e\in E^*_{\omega r}} x^\ell_{ie\omega} \tag{5}$$

The EMILA model is defined as a two-stage stochastic 0–1 integer linear programming model as follows:

$$\max \quad (1-\alpha)\cdot EF + \alpha \cdot EQ \tag{6a}$$

$$s.t. \sum_{j\in J} y_{ij} \le 1 \qquad \forall i \in I \tag{6b}$$

$$\sum_{i\in I} y_{ij} \le c_j \qquad \forall j \in J \tag{6c}$$

$$\sum_{i\in I} \sum_{j\in J} \delta_{ij} \cdot y_{ij} \le k \tag{6d}$$

$$x^\ell_{ie\omega} - \sum_{j\in J} a^\ell_{je\omega} \cdot y_{ij} \le 0 \qquad \forall i \in I,\ \ell \in \mathcal{L},\ \omega \in \Omega,\ e \in E_\omega \tag{6e}$$

$$x^\ell_{ie\omega} + \sum_{\ell^* \in \mathcal{L}} x^{\ell^*}_{ie'\omega} \le 1 \qquad \forall i \in I,\ \omega \in \Omega,\ \ell \in \mathcal{L},\ e \in E_\omega,\ e' \in U^\ell_{e\omega} \tag{6f}$$

$$\sum_{i\in I} \sum_{\ell\in\mathcal{L}} x^\ell_{ie\omega} \le 1 \qquad \forall \omega \in \Omega,\ e \in E_\omega \tag{6g}$$

$$\sum_{i\in I} \sum_{\omega\in\Omega} \sum_{e\in E_\omega} \sum_{\ell\in\mathcal{L}} x^\ell_{ie\omega} \ge \beta \cdot n' \tag{6h}$$

$$\mathbb{E}_\Omega \left[Z^{\ell_1} \right] \ge \pi \tag{6i}$$

$$\sum_{\ell\in\mathcal{L}} x^\ell_{ie\omega} \le \varepsilon_{ie\omega} \qquad \forall i \in I,\ \omega \in \Omega,\ e \in E_\omega \tag{6j}$$

$$y_{ij},\ x^\ell_{ie\omega} \in \{0,1\} \qquad \forall i \in I,\ \ell \in \mathcal{L},\ j \in J,\ \omega \in \Omega,\ e \in E_\omega \tag{6k}$$

The objective function (6a) is a linear combination of the two criteria to be maximized (Ehrgott, 2005): the efficiency component (EF) and the equity component (EQ). The efficiency component maximizes the expected value over the scenario set of the weighted success rate (it is detailed in (7a) and equivalently, in (7b)). On the contrary, the equity component maximizes the conditional expected value over the scenario set of the regional weighted success rate (7c). By using the non-decreasing weight parameter μ^ℓ over the set of RT-intervals, both components give more importance to emergencies responded to in less time. By using the $\frac{n^r}{n}$ versus the $\frac{m}{n}$ ratio coefficients, the EF component prioritizes the number of answered calls, while the EQ component gives the same weight to all the regions.

$$EF = \sum_{\ell\in\mathcal{L}} \mu^\ell \cdot \mathbb{E}_\Omega \left[Z^\ell \right] = \sum_{\ell\in\mathcal{L}} \mu^\ell \cdot \left(\sum_{\omega\in\Omega} p_\omega \cdot z^\ell_\omega \right) \tag{7a}$$

$$= \sum_{\ell\in\mathcal{L}} \mu^\ell \sum_{r\in R} \frac{n^r}{n} \cdot \mathbb{E}_\Omega \left[Z^\ell \middle| R = r \right] = \frac{1}{n} \sum_{\omega\in\Omega} \sum_{\ell\in\mathcal{L}} \sum_{r\in R} \sum_{i\in I} \sum_{e\in E^*_{\omega r}} \mu^\ell \cdot x^\ell_{ie\omega} \tag{7b}$$

$$EQ = \sum_{\ell\in\mathcal{L}} \mu^\ell \sum_{r\in R} \frac{m}{n} \cdot \mathbb{E}_\Omega \left[Z^\ell \middle| R = r \right]$$

$$= \frac{1}{n} \sum_{\omega\in\Omega} \sum_{\ell\in\mathcal{L}} \sum_{r\in R} \frac{m}{n^r} \left(\sum_{i\in I} \sum_{e\in E^*_{\omega r}} \mu^\ell \cdot x^\ell_{ie\omega} \right) \tag{7c}$$

Table 2
Summary of the proposed model variants.

Model name	Model variant	α	$\lvert\mathcal{L}\rvert$	$\mu = (\mu^{\ell_1}, \mu^{\ell_2}, \ldots, \mu^{\ell_{\lvert\mathcal{L}\rvert}})$	Reference
LAB	Baseline	–	2	$\mu = (1,0)$	(2)
EMILA	Efficiency	$\alpha = 0$	>2	$\mu^{\ell_1} \ge \mu^{\ell_2} \ge \cdots \ge \mu^{\ell_{\lvert\mathcal{L}\rvert}}$	(6)
	Balanced	$\alpha \in (0,1)$	>2	$\mu^{\ell_1} \ge \mu^{\ell_2} \ge \cdots \ge \mu^{\ell_{\lvert\mathcal{L}\rvert}}$	(6)
	Equity	$\alpha = 1$	>2	$\mu^{\ell_1} \ge \mu^{\ell_2} \ge \cdots \ge \mu^{\ell_{\lvert\mathcal{L}\rvert}}$	(6)

where $\mathbb{E}_\Omega \left[Z^\ell \middle| R = r \right] = \sum_{\omega\in\Omega} p^r_\omega \cdot z^{\ell,r}_\omega$

Therefore, the objective function (6a) can be reformulated as follows:

$$\frac{1}{n} \sum_{\omega\in\Omega} \sum_{\ell\in\mathcal{L}} \sum_{r\in R} \sum_{i\in I} \sum_{e\in E^*_{\omega r}} \left(1 + \alpha \left(\frac{m}{n^r} - 1 \right) \right) \cdot \mu^\ell \cdot x^\ell_{ie\omega} \tag{8}$$

Constraints (6b), (6c) and (6d) remain as in model (2). In the rest of constraints (from (6e) to (6k)), the new formulations for parameters $u_{e\omega}$ and $a^\ell_{je\omega}$ and variables $x^\ell_{ie\omega}$, z^ℓ_ω and $z^{\ell,r}_\omega$ are used.

A summary of the proposed models and their variations is shown in Table 2. By varying the weight α from 0 to 1, the relevance of the equity component is increased. This effect can be easily seen in Eq. (8): when $\alpha = 0$, such a component is not considered and is what we call the *efficiency* model. In this model, regardless of the region where they occur, as many emergencies as possible are attended to in time. On the contrary, when $\alpha = 1$, equity in success rates between different regions is sought. We refer to this model as the *equity* model. This regional equity is achieved through the importance given to each of the regions according to the number of calls ($\frac{m}{n^r}$). In addition, it is easy to simplify the EMILA model into the LAB model (2) by giving appropriate values to specific parameters and sets.

4 Experimental study

This section shows the computational experiments with the models described in Section 3. First, we detail the input parameters for reproducibility purposes. Second, we describe the calibration phase. Third, the optimization outcomes are shown. Finally, the validation of the proposed models is presented. We implemented the model in the optimization software IBM ILOG CPLEX Optimization Studio V20.1 (IBM, 2020) in the computational cluster ARINA from SGI/IZO-SGIker (UPV/EHU) (2020). We used nodes with 128 GB of RAM and a solid-state hard drive for these computational experiments. Eight cores were used for each optimization problem, and the memory and the time were limited to 20 GB and 2 h, respectively. The EMILA model codes and an example of small size are available at Gago-Carro et al..

4.1 Input data

Regarding the specific study case of the Basque Country, we consider the following sets and parameters.

Sets:

- I: the eleven ALS ambulances that, according to Emergentziak Osakidetza (2017), are currently operative in the Basque Country are considered. In addition, we consider at most five potential ambulances with no pre-assigned station to solve the location problem ($|I_0| = 11, |I_1| \le 5, |I| \le 16$).
- J: the set of potential stations for ALS ambulances consists of all the current stations in the Basque Country. Adding the BLS and ALS ambulance stations, there are 80 ambulance stations ($|J| = 80$).
- E_ω: the set emergency calls of each scenario $\omega \in \Omega$ consists of all the emergencies that occurred during the 8 h with more activity, i.e., from 9:00 to 17:00 h. We consider three different emergency call types. First, false alarms, which do not belong to the subset of full-service calls. In that case, an ambulance leaves its base

Table 3
SAA lower and upper bounds and optimality GAPs for 95 CIs when 11 ambulances are relocated.

$\|\Omega\|$	SAA upper bound			SAA lower bound		SAA optimality GAP	
	Estimation	CI	Average time	Estimation	CI	Estimation	CI
30	0.8664	(0.8664,0.8665)	04:22	0.8383	(0.8377,0.8391)	0.0273	(0,0.0288)
60	0.8591	(0.8591,0.8591)	19:19	0.8542	(0.8537,0.8549)	0.0042	(0,0.0055)
90	0.8584	(0.8584,0.8584)	46:31	0.8542	(0.8537,0.8549)	0.0035	(0,0.0048)
120	0.8568	(0.8568,0.8568)	2:05:54	0.8550	(0.8543,0.8556)	0.0012	(0,0.0024)

station, but before reaching the emergency, the emergency is canceled, and it returns to its station. Second, the subset of full-service calls comprises emergencies that do not require hospital evacuation and those that do. In any of the three call types, and for simplification purposes, the ambulance is occupied until it returns to its station, i.e., no call is assigned while en route.

Ω: the set of scenarios consists of randomly selected days from the historical database. We conducted a Sample Average Approximation (SAA) (Kleywegt et al., 2002) to determine the size of Omega. We consider a compromise between the computational tractability of the problem and the SAA results. As the tractability of the problem is crucial in this analysis, we run the relocation variant of the LAB model (6) for $k = 11$, which is the LAB model variant with more computational requirements. The procedure to calculate the appropriate number of scenarios is described in Algorithm 1. We conducted the analysis by increasing the size of Ω from $N_1 = 30$ scenarios to $N_4 = 120$. On the one hand, we obtained the upper bounds of the problem by carrying out randomly sampled M = 30 optimization runs of the model for each size of Ω. On the other hand, the lower bounds were obtained by fixing the first-stage variables (y_{ij}) and optimizing the second-stage allocation problem for the whole year ($K = 365$ days).

Algorithm 1: Sample Average Approximation (SAA)

Input: Potential candidates for $|\Omega| \in \{N_1, \dots, N_P\}$. Sample size M.
1: **for each** $N_i = N_1, \dots, N_P$ **do**
2: **for** $m = 1, \dots, M$ **do**
3: Generate a sample of N_i random scenarios.
4: Solve the EMILA model and save objective value (v^m) and CPU time to solve the problem (T_m).
5: **end for**
6: Obtain most repeated First-Stage solution $y^*_{N_i}$.
7: Compute bilateral $(1-\alpha)$ CI for SAA upper bound.
8: Generate a set K independent scenarios, where $K \gg N_i$.
9: **for** $k = 1, \dots, K$ **do**
10: Fix first-stage solution $y^*_{N_i}$ and solve second-stage EMILA problems over scenario k.
11: Save objective value (v_k).
12: **end for**
13: Compute bilateral $(1-\alpha)$ CI for SAA lower bound.
14: Fix first-stage solution $y^*_{N_i}$ and solve EMILA problem with K scenarios.
15: Save objective value (v_K) and compute $g_m := v_K - v_m$.
16: Compute unilateral $(1-\alpha)$ CI for SAA optimality gap
17: **end for**
Output: CPU Time and three CI for each $N_i \in \{N_1, \dots, N_P\}$.

The results of this analysis are summarized in Table 3. Based on the results, we concluded that 60 is an appropriate size of Ω because the SAA optimality GAP is lower than 0.01 and computationally tractable.

$\mathcal{L}$: since TT and TT′ in our study case are 15 and 45 min, respectively, and since we add an intermediate threshold of 30 min, we have the following four groups: emergencies responded to in less than 15 min (ℓ_1), between 15 and 30 min (ℓ_2), between 30 and 45 min (ℓ_3) and non-attended emergencies.

R: we consider three regions according to the call index as follows: if an ambulance located in a municipality could handle under 1000 emergencies in 2019, the region of the municipality is *rural*; if it could handle between 1000 and 10,000 calls, the region is *suburban* and, finally, if it could handle more than 10,000, it is considered *urban*.

Parameters:

$u^{\ell}_{e\omega}$: the number of seconds an ambulance is unavailable because it is attending an emergency e of scenario ω in the RT-interval ℓ is calculated based on the actual data in the database. The median value that an ambulance is occupied (from the time it is assigned to the emergency until it returns to its station) is considered for each combination of emergency type and RT-interval. Then, the value of its emergency type is selected for each (e, ω) combination.

$a^{\ell}_{je\omega}$: the time required to travel from the ambulance stations to the emergencies (needed to calculate parameter $a^{\ell}_{je\omega}$) is calculated with an API by Google Developers (2020). We compared the actual time ambulances took for each emergency and the simulated time obtained through the API. As this API calculates the time for a regular car and ambulances usually go faster, simulated times are more pessimistic than real times (see Fig. 2(a)).
Let us classify the provided data according to its region. The differences between actual times and calculated times are shown in Fig. 2(b): while these mean differences remain low in urban and suburban areas (2 min and 11 s in urban areas and 2 min 5 s in suburban areas), the difference in rural areas on average reaches 6 min and 55 s (see Fig. 2(b)).

$\varepsilon_{ie\omega}$: as the carried out optimization runs take the most intensive time of the day into account, i.e., between 9:00 to 17:00 h. The parameter indicates whether an ambulance is operational ($\varepsilon_{ie\omega} = 1$) or not ($\varepsilon_{ie\omega} = 0$) for each emergency call.

c_j: the stations where the ambulances wait until they are allocated to emergencies are of several types, such as hospitals or parking lots in the street. The station's capacities (c_j) are calculated from the current capacities and two new potential ambulances added to each station.

π: is set to 0.75, the percentage indicated in the Contract Program (Emergentziak Osakidetza, 2019).

β: is set to 0.99: this value avoids infeasibilities, if any, when many emergencies occur in very remote locations.

α: we consider the following values for α parameter: {0, 0.2, 0.4, 0.6, 0.8, 1}. According to those values, a set of solutions is obtained and provided to the decision-makers.

4.2 Calibration of the EMILA model

The objective function of the EMILA model (6) is affected by two types of parameters: first, μ^{ℓ} parameter, which measures the priority with which emergencies are attended in terms of RT-intervals; second, α nonnegative weighted parameter, which balances the priority between efficiency and equity components. Although α value should be chosen according to decision-makers' inequity-aversion, μ^{ℓ} parameters are calibrated as explained below.

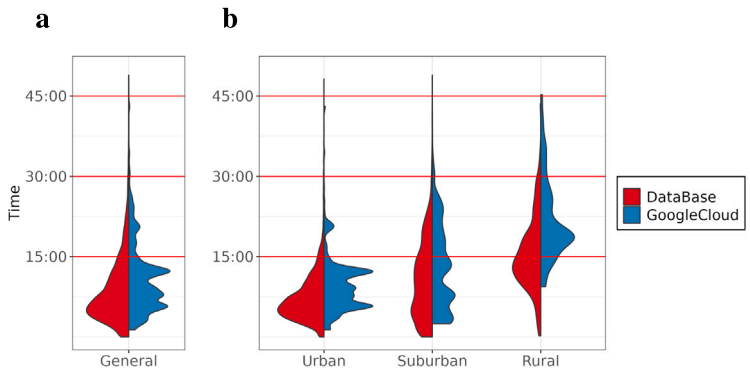


Fig. 2. Comparison of travel times by origin of data and regions. Real travel times acquired from the database are in red, and the simulated times obtained using the Google API are in blue. The general case is shown in (a), while in (b), a partition by region has been done.

In order to make the model as close to reality as possible, let us calibrate the RT-interval priority vector $\mu = (\mu^{\ell_1}, \mu^{\ell_2}, \mu^{\ell_3})$. Based on the fact that it is essential to assign as many calls as possible and since it is desirable to answer them as soon as possible, let us take $\mu^{\ell_1} \geq \mu^{\ell_2} \geq \mu^{\ell_3} \geq 0$. Considering these restrictions, we calibrate the parameter by optimizing a battery of models, with the value of the μ parameter vector components varying from 1 to 6. Additionally, we consider the cases where the number of emergency calls answered in time is maximized $\mu = (1, 0, 0)$ (the LAB model) and the case in which the number of emergencies answered in over 30 min is minimized $\mu = (1, 1, 0)$. For each combination of the parameters, we solve thirty optimization runs of the model's efficiency variant ($\alpha = 0$). The parameter k is set to 0, not allowing any change in the fleet. For this calibration process, we consider 60 scenarios of emergencies in the high activity time slot, i.e., between 9:00 and 17:00. The calibration results are summarized in the ternary diagram of Fig. 3, where each point shows the percentage of emergencies attended to in each of the three RT-intervals when solving each combination of parameters. For instance, the blue point corresponds to the results obtained with the LAB model ($\mu = (1, 0, 0)$): 80.45% in the axis on the right (first RT-interval), 10.38% in the left axis (second RT-interval), and 9.17% in the bottom axis (45′ RT-interval threshold). Together with the results of the calibration, the actual results of the historical database are also shown. A comparison of the results reveals that the parameter combination $\mu = (4, 2, 1)$ is the most suitable one for our model. This combination obtains one of the best possible percentages of emergency calls responded to in time: 79.87%, only 0.56 percentual points below the LAB model. Regarding the worse response to emergencies (these emergencies responded to between 30 and 45 min, and these which were not responded to at all), it is only 0.76 percentage points higher than the one obtained with the combination $\mu = (6, 4, 1)$, which is the one with better results in this interval. Moreover, the results obtained by the $\mu = (4, 2, 1)$ parameter combination are the closest to the actual historical database results (86.46%, 12.64%, 0.90%).

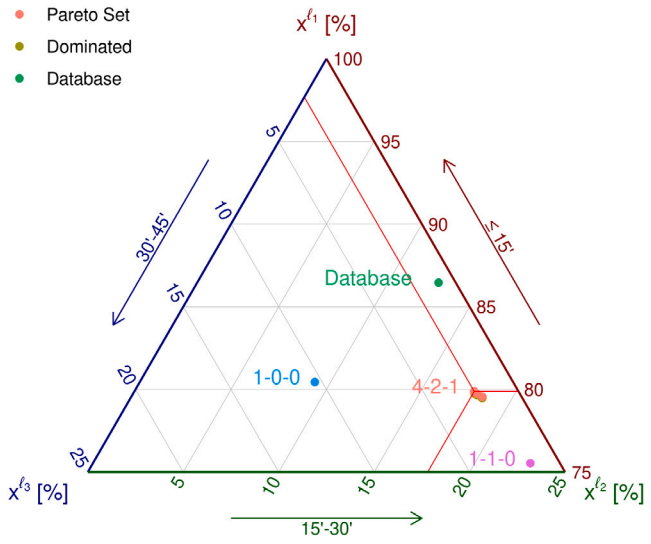


Fig. 3. Ternary diagram with the percentage of full-service emergencies attended to in each RT-interval by optimization run when solving the efficiency variant for $k = 0$.

4.3 Optimization results

Seven optimization variants of the proposed models are solved: the LAB model and the EMILA model for the following values of α parameter: $\{0, 0.2, 0.4, 0.6, 0.8, 1\}$. For each of these variants, 11 instances are solved: control-case (where $k = 0$), five relocation–allocation problems (having $k \in \{1, 2, 3, 4, 5\}$ and $I_1 = \emptyset$), and five location–allocation problems ($k \in \{1, 2, 3, 4, 5\}$ and $|I_1| = k$). For each of the 77 cases, we implement 30 model runs of 60 randomly chosen scenarios (days) for a total of over 2300 instances.

The results for the relocation–allocation models are reported only for the global validation in Section 4.4 since they do not improve the control case significantly. The results for the location–allocation models are reported for the extreme instances ($\alpha = 0$ and $\alpha = 1$) and the balanced one for the intermediate $\alpha = 0.4$ value.

4.3.1 Results of the LAB model

The results obtained for the LAB model (2) for the location–allocation problem are shown in Fig. 4: the boxplot for the success rate of the 30 optimization runs is calculated for the database and the optimization model with $k \in \{0, 1, 2, 3, 4, 5\}$. The gap between the mathematical model and reality can be measured by comparing the percentage of full-service calls attended in time in reality (85.81%) and in the control case (79.75%). This slightly worse result is expected because of the assumptions and the stringency mentioned previously. Regarding the optimization where new ambulances are added to the fleet, the success rate is increased in 3.61%, 6.09%, 8.07%, 9.74% and 11.13% when adding one, two, three, four, and five ambulances, respectively. However, if we focus on the full-service calls answered in over 15 min, the model does not distinguish between the calls answered between 16 and 30 min and the calls answered between 31 and 45 min. Indeed, the model makes more allocations of the last type (see Fig. 4(b)).

4.3.2 Results of the EMILA location–allocation model

The results obtained with the efficiency variant ($\alpha = 0$) of the EMILA model (6) (Fig. 5(a)) are similar to those obtained by the LAB model (Fig. 4(b)), as far as the success rate is concerned: the success rate

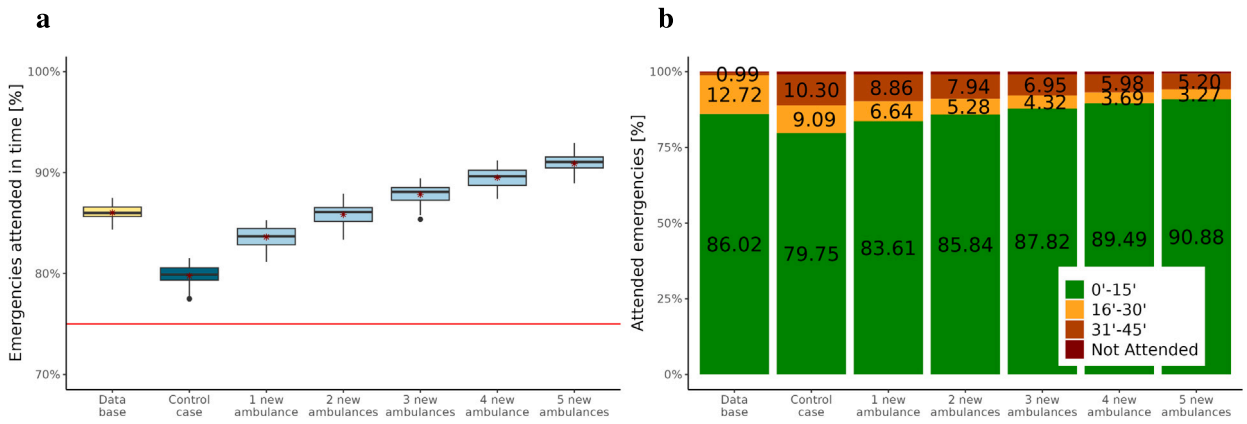


Fig. 4. Results obtained with the LAB model: (a) the success rate of emergencies answered in time; (b) the percentage of emergencies attended in each RT-interval.

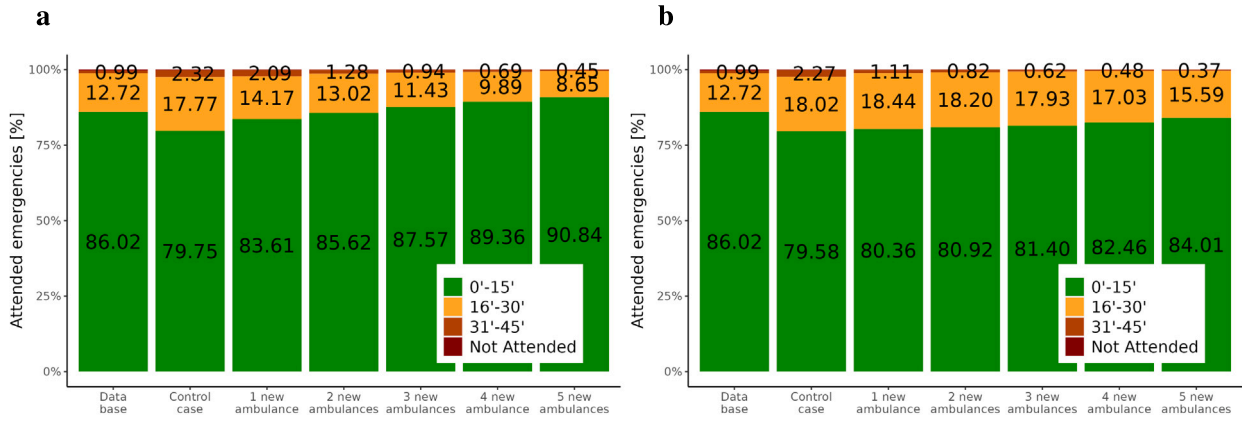


Fig. 5. Percentage of emergencies attended in each RT-interval when optimizing using (a) efficiency variant ($\alpha = 0$) and (b) equity variant ($\alpha = 1$) of the EMILA model.

increases from a 79.75% in the control case to a 90.84% when five ambulances are added to the fleet. However, the major improvement compared to the LAB model lies in the response times of the ambulances when the TT is exceeded: when the model allocates an ambulance that does not arrive in time, the efficiency variant allocates ambulances closer to emergencies: the percentage of emergencies answered between 31 and 45 min, compared with the number of emergencies not responded to in time, is now much closer to reality. However, when we disaggregate the results by region, the results are very different for each region. While urban and suburban areas show improvement as new ambulances are added, the success rate in rural areas remains similar to the control case (see Fig. 6). In addition, the reason for the difference in the success rate between reality and control cases is explained, to some extent, as the success rates in urban areas are similar, but they are not in suburban and rural areas. The significant difference (46.55%) in the success rate in rural areas between reality and control cases is thought to occur because of all the assumptions taken: for example, in reality, some allocations were made when the corresponding ambulance was near the emergency, although it was not at its station. Such allocations mean significant savings in traveling time and increase the success rate.

On the contrary, when applying the equity variant ($\alpha = 1$) of the model, its equity component succeeds in the objective of reducing the differences between regions, as this model gives greater importance to the emergencies in the areas where fewer emergencies occur: rural and suburban areas (see Fig. 5(b)). However, this allocation criterion means that the general success rate does not improve as much as with the efficiency variant when new ambulances are added to the fleet. As mentioned in Section 4.2, intermediate values of α weighted parameter vary call priorities and, consequently, how allocations are made. Fig. 6 presents the results of one of those balanced variants, where the evolution from the efficiency variant to the equity variant is remarkable. Boxplots for reality and control cases solved with the efficiency variant are also shown.

When we optimize one of the two criteria (efficiency or equity), the other can be worsened. An ethical dilemma undoubtedly arises from this conflict of interests: is it better to save more lives and thus locate new ambulances in urban areas at the expense of abandoning rural areas, or, on the contrary, is a slight worsening of the overall performance insignificant in order to reach more municipalities in time? These ethical issues should be discussed thoroughly with the health professionals. It is up to the experts to decide which model best fits their needs and priorities: responding to as many emergencies as possible, providing a more equitable emergency service, or a commitment between both objectives.

4.4 Validation

This section describes the procedure we follow to validate the model. Based on the validation tools explained by Sargent (2013), we develop two own techniques: first, we carry out a so-called *robust validation* to ensure that the assignments proposed in the instances of control case are similar to the real ones, checking that the performance is similar between hours of the day and months of the year. Second, a *global validation* is carried out to guarantee that the results obtained when considering the maximum activity moments of the randomly selected $|\Omega|$ days can be extrapolated to the behavior of the entire year: 24 h a day and 12 months a year.

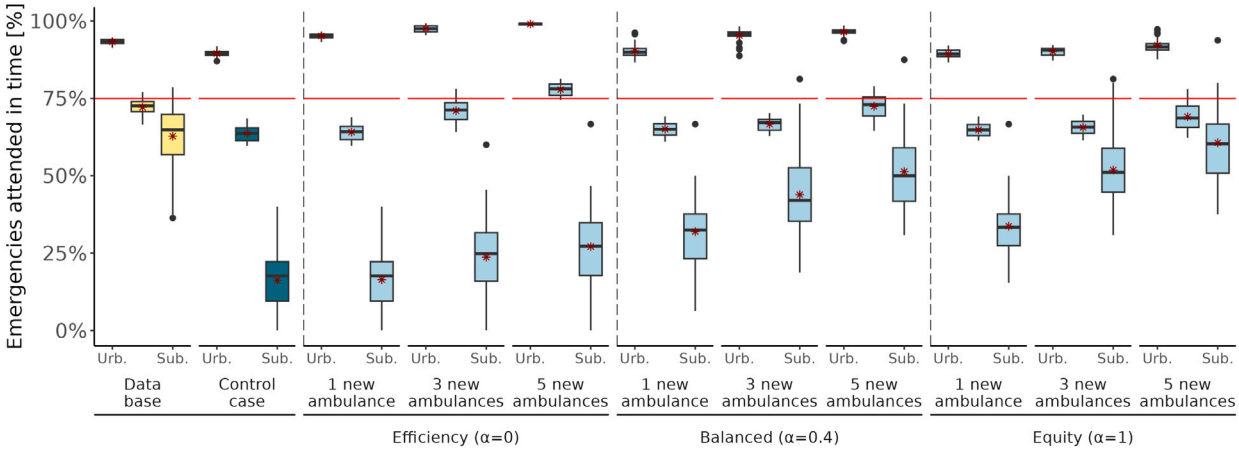


Fig. 6. Boxplots of the success rate obtained with efficiency ($\alpha = 0$), balanced ($\alpha = 0.4$) and equity ($\alpha = 1$) variants of the EMILA model by region, when locating new ambulances to the fleet.

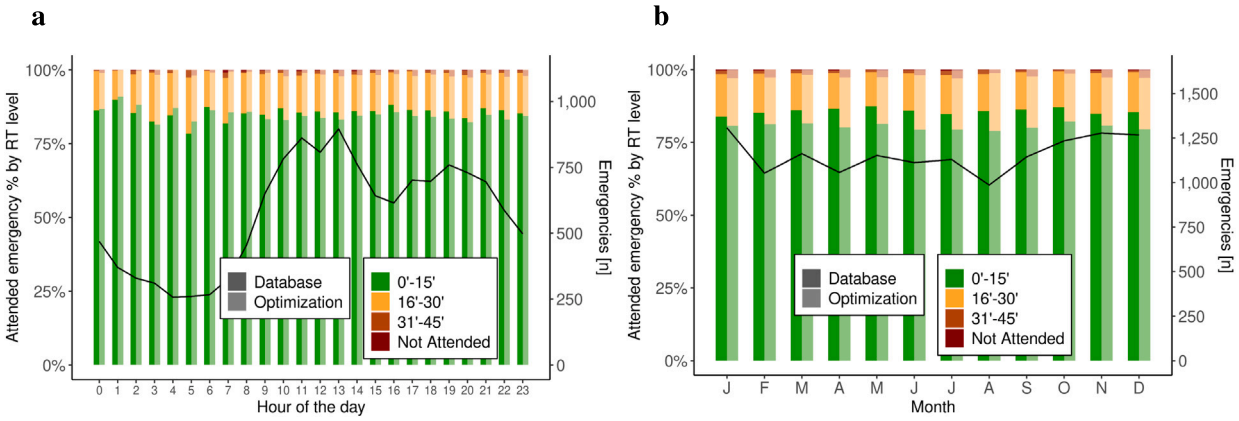


Fig. 7. Robust validation: Comparison of emergency attendance distributions by response time and by (a) hour and (b) month. Real distributions of emergencies (database) are displayed in a dark range of colors, while the distribution of the optimization results is shown in a softer range. The black curve represents the number of emergencies that occurred during each period of time.

Robust validation: We validate the robustness of the efficiency variant ($\alpha = 0$) of the EMILA model (6). We analyze the control case to verify that the assignments made in the optimization do not differ too much from the actual allocations. We fix the parameter k to 0, not allowing relocation or extension. We thus manage to optimize the allocation of ambulances to emergencies, so we can check whether the model's behavior is close to reality. Fig. 7 compares the results of the control case and the actual case by showing the percentage of answered emergencies by RT-interval and by different variables: hour of the day (see Fig. 7(a)) and month (see Fig. 7(b)). In both cases, the general behavior of the results is that the success rate of the optimization instances is slightly lower than the actual success rate. However, the performances are maintained over time, either hour of the day or the month of the year.

Global validation: As mentioned in Section 4.1, each optimization run only considers emergencies between 9:00 and 17:00 h for 60 randomly chosen days. However, it is essential to check whether the changes proposed in those runs are extrapolable to the rest of the day and the whole year. To that end, a validation in two phases is carried out. First, we run 30 tests for each case of parameter k and each of the variants. Each of these optimizations proposes k locations for the new ambulances. In this way, we have a collection of potential combinations of locations of the ambulances added to the fleet. We select the most often repeated combinations of new locations, and in the second phase, we run a total of twelve model runs (one for each month) over the whole year 2019 for each of the selected combinations. Since the results obtained for each location combination are similar, the collection is suitable to be proposed to the experts to decide which best suits their needs.

As the results obtained in this second phase do not differ significantly from one location combination to another, only the results for the combination of new locations most often repeated (the statistical mode) are shown. Table 4 shows the results of the 30 initial tests and the validation of the whole year disaggregated by model variant and value of parameter k. The success rates obtained in the validation phase are slightly higher than those presented in Section 4.3. This slight improvement is because the results of the initial optimization runs are calculated for the time slot of the day with more activity.

Regarding the results of the validation phase, the percentage of emergencies answered in time increases as new ambulances are added to the fleet. For instance, the success rate in the efficiency variant increases from 80.5% when no ambulance is added to 91.2% when adding five ambulances. On the contrary, the percentage of emergencies responded to between 30 and 45 min decreases when k increases: in all three variants of the EMILA model, the number of responded emergencies in this RT-interval decreases to less than a third.

Moreover, in Table 5, each variant behavior is shown: the closer to the equity variant ($\alpha = 1$), the more calls are handled in time in rural areas. Although the success rate in rural areas is similar for the control

Table 4
Global RT validation: percentage of answered calls by RT-interval, the value of α parameter, and number of new ambulances added when solving the LAB and the EMILA models for the chosen samples and the extrapolation to the whole year.

Numb. of new ambulances	RT-interval	Baseline		Efficiency ($\alpha = 0$)		Balanced ($\alpha = 0.4$)		Equity ($\alpha = 1$)	
		Sample	Year	Sample	Year	Sample	Year	Sample	Year
0	[0–15]	79.8	80.5	79.8	80.5	79.6	80.3	79.6	80.3
	(15–30]	9.1	8.7	17.8	17.2	17.9	17.3	18.0	17.4
	(30–45]	10.3	10.1	2.3	2.3	2.3	2.3	2.3	2.2
	NA	0.9	0.8	0.2	0.1	0.1	0.1	0.1	0.1
1	[0–15]	83.6	84.3	83.6	84.3	81.0	82.8	80.4	81.0
	(15–30]	6.6	6.6	14.2	13.6	17.7	15.6	18.4	17.6
	(30–45]	8.9	8.4	2.1	2.0	1.2	1.5	1.1	1.3
	NA	0.9	0.8	0.1	0.1	0.1	0.1	0.1	0.1
3	[0–15]	87.8	88.5	87.6	88.4	85.0	87.3	81.4	81.9
	(15–30]	4.3	3.8	11.4	10.2	14.3	11.8	17.9	17.3
	(30–45]	6.9	6.9	0.9	1.3	0.6	0.8	0.6	0.7
	NA	0.9	0.8	0.1	0.1	0.1	0.0	0.0	0.0
5	[0–15]	90.9	90.9	90.8	91.2	87.8	89.3	84.0	82.3
	(15–30]	3.3	2.8	8.7	8.2	11.8	10.1	15.6	17.1
	(30–45]	5.2	5.5	0.5	0.6	0.4	0.6	0.4	0.5
	NA	0.6	0.7	0.1	0.0	0.0	0.0	0.0	0.0

case of each of the presented variants, when adding five ambulances to the fleet, it varies from 28.8% in the efficient variant to 52.2% in the equity variant (an improvement of over 23 percent points), being 30.5% in one of the balanced variants ($\alpha = 0.4$). The downside of the equity search is the loss of 8.9 percentage points in the overall success rate: from 91.2% when solving the efficiency variant to 82.3% when solving the equity variant.

Regarding the relocation–allocation problem, although the trend is similar, the results are more conservative than those obtained when adding new ambulances to the fleet. The percentage of emergencies responded to in time is lower in the relocation–allocation problem (see Table 6) than in the location–allocation problem: optimizing the location of the fleet by changing up to 5 ambulances only yields an 85.4% success rate while adding five ambulances leads to a success rate of 91.2%. As in the location–allocation problem, the obtained results depend on the objective pursued and the variant of the model used: Table 7 shows that the efficiency variant, seeking a maximization of the overall success rate, improves areas with more activity. On the contrary, the equity variant gives more importance to isolated municipalities and improves rural areas. In this equity variant, the more relocations are allowed, the more the rural success rate improves; however, the overall success rate worsens (from 80.3% to 79.1%). This worsening does not occur in the location–allocation model, where, despite seeking equity between zones, overall success also improves as new ambulances are added. Although the results of both problems follow the same trend, adding more ambulances to the fleet and having them available make a difference in the success rate.

5. Conclusions and future work

In this paper, we formulate a two-stage stochastic model for the (re)location–allocations of medical services in the Basque Public EMS system. Using the presented equity and multi-interval components, a distribution of resources can be found in which the differences in the success rates of the different types of areas are as slight as possible. At the same time, ambulances arrive to emergencies as quickly as possible. In that way, by adding new ambulances to the fleet and locating them strategically, the model can propose appropriate allocations that make an optimal coverage of the demanded attention.

The two-stage stochastic 0–1 integer linear programming models, tested with actual data of the emergencies that occurred in the Basque Country during 2019, propose solutions that meet the objectives pursued: the success rate is increased when new ambulances are added to the fleet: there is an improvement of 10.4, 10.7 and 2.0 percentage points when adding five ambulances to the baseline, efficiency, and equity variants, respectively. Moreover, the developed efficiency variant can propose efficient solutions in which nearer ambulances are allocated to emergencies without worsening the success rates too much: the number of late responses is reduced. Additionally, the equity variant can propose solutions that minimize success rate differences between areas with different population densities. Finally, as the parameters used in the equity variant can be tuned, it is up to the final decision-maker to decide the balanced efficiency-equity level that best suits their needs.

However, even if the objectives are met with the models presented in this paper, there are some challenges to consider in future works. Since the ambulance (re)location–allocation problem is a very complex problem in emergency medical services, improvements in modeling could deal with some of the assumptions made, especially with those that have a significant impact, as seen with the control case, such as the possibility to assign an ambulance that is en route. Incorporating stochasticity in the time considered for moving from station to emergency is interesting for future research. Furthermore, extending the problem to the entire ambulance fleet would be desirable, considering all types of ambulances and the urgency level of emergencies. In addition, one of the problem's main challenges is resolving the conflict of interest between efficiency and equity. While the final decision remains within the purview of professional experts, a future avenue for the research community involves the development of tools that can aid in these decisions. Incorporating multi-objective optimization techniques, such as lexicographic or compromise optimization, is envisioned to play a crucial role in this collective effort. An alternative is a multilevel and hierarchical scheme that optimizes across all the criteria to be considered. Risk aversion functions could also be considered to find solutions protected against worst scenarios. The difficulties that these new challenges may bring, such as time complexity issues when solving the proposed models, make us think it could be interesting to implement decomposition methodologies, such as metaheuristics or matheuristics algorithms, to obtain quasi-optimal results in reasonable computing time.

CRediT authorship contribution statement

Imanol Gago-Carro: Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing. **Unai Aldasoro:** Conceptualization, Investigation, Methodology, Supervision, Writing – review & editing. **Josu Ceberio:** Conceptualization, Funding acquisition, Investigation, Methodology, Supervision, Writing – review & editing.

Table 5
Impact of adding new ambulances over the whole year by region: percentage of answered calls by RT-interval, region, the value of α parameter, and number of new ambulances added when solving the LAB and the EMILA models.

Numb. of new	RT-interval	Baseline			Efficiency ($\alpha = 0$)			Balanced ($\alpha = 0.4$)			Equity ($\alpha = 1$)		
ambulances		Urb.	Sub.	Rur.	Urb.	Sub.	Rur.	Urb.	Sub.	Rur.	Urb.	Sub.	Rur.
0	[0–15]	90.6	63.3	21.2	90.5	63.5	21.2	89.8	64.5	21.7	89.7	64.5	21.7
	(15–30]	4.2	16.6	25.2	7.7	33.9	59.7	8.2	33.1	63.3	8.3	33.3	63.3
	(30–45]	4.5	19.2	46.9	1.9	2.4	15.0	2.0	2.4	11.9	2.0	2.2	11.9
	NA	0.7	0.8	6.6	0.0	0.2	4.0	0.0	0.1	3.1	0.0	0.1	3.1
1	[0–15]	96.1	63.9	21.2	95.9	64.3	21.7	93.3	64.8	27.9	90.3	64.6	37.2
	(15–30]	1.0	16.7	25.7	2.3	33.6	61.9	5.6	33.2	61.1	8.8	33.4	50.0
	(30–45]	2.6	18.2	43.8	1.7	2.1	12.4	1.0	2.0	8.8	0.8	1.9	10.6
	NA	0.3	1.2	9.3	0.0	0.1	4.0	0.0	0.0	2.2	0.0	0.1	2.2
3	[0–15]	96.5	75.7	22.6	97.2	74.0	23.9	96.9	71.1	30.1	91.1	65.2	50.4
	(15–30]	0.8	8.7	26.5	1.8	24.2	63.7	2.7	27.4	60.6	8.7	33.4	39.8
	(30–45]	2.4	14.2	42.9	0.9	1.7	8.8	0.4	1.4	8.4	0.1	1.4	9.7
	NA	0.3	1.4	8.0	0.0	0.1	3.5	0.0	0.0	0.9	0.0	0.0	0.0
5	[0–15]	99.2	77.7	28.8	99.2	78.5	28.8	97.2	76.4	30.5	91.5	65.6	52.2
	(15–30]	0.2	7.0	23.0	0.8	20.2	60.6	2.7	22.4	61.1	8.4	33.4	38.1
	(30–45]	0.5	13.7	38.9	0.0	1.2	8.8	0.1	1.2	8.4	0.1	0.9	9.7
	NA	0.1	1.6	9.3	0.0	0.0	1.8	0.0	0.0	0.0	0.0	0.0	0.0

Table 6
Global RT validation: percentage of answered calls by RT-interval, the value of α parameter, and number of ambulances relocated when solving the LAB and the EMILA models for the chosen samples and the extrapolation to the whole year.

Numb. of	RT-interval	Baseline		Efficiency ($\alpha = 0$)		Balanced ($\alpha = 0.4$)		Equity ($\alpha = 1$)	
relocations		Sample	Year	Sample	Year	Sample	Year	Sample	Year
0	[0–15]	79.8	80.5	79.8	80.5	79.6	80.3	79.6	80.3
	(15–30]	9.1	8.7	17.8	17.2	17.9	17.3	18.0	17.4
	(30–45]	10.3	10.1	2.3	2.3	2.3	2.3	2.3	2.2
	NA	0.9	0.8	0.2	0.1	0.1	0.1	0.1	0.1
1	[0–15]	82.9	83.6	82.9	83.6	80.5	80.4	78.5	80.4
	(15–30]	7.3	7.0	14.6	14.0	17.1	17.3	19.2	17.5
	(30–45]	9.0	8.5	2.4	2.3	2.3	2.2	2.2	2.1
	NA	0.8	0.8	0.1	0.1	0.1	0.1	0.1	0.1
3	[0–15]	84.9	85.2	84.9	85.2	82.8	83.6	78.1	77.6
	(15–30]	5.8	5.7	12.6	12.1	14.9	14.1	19.3	20.2
	(30–45]	8.5	8.3	2.4	2.5	2.2	2.2	2.5	2.1
	NA	0.8	0.8	0.2	0.1	0.1	0.1	0.1	0.1
5	[0–15]	85.6	85.4	85.5	85.4	84.0	85.2	80.0	79.1
	(15–30]	5.4	6.2	12.1	11.9	13.6	12.3	17.4	18.7
	(30–45]	8.2	7.6	2.3	2.6	2.3	2.4	2.5	2.1
	NA	0.8	0.8	0.1	0.1	0.2	0.1	0.2	0.1

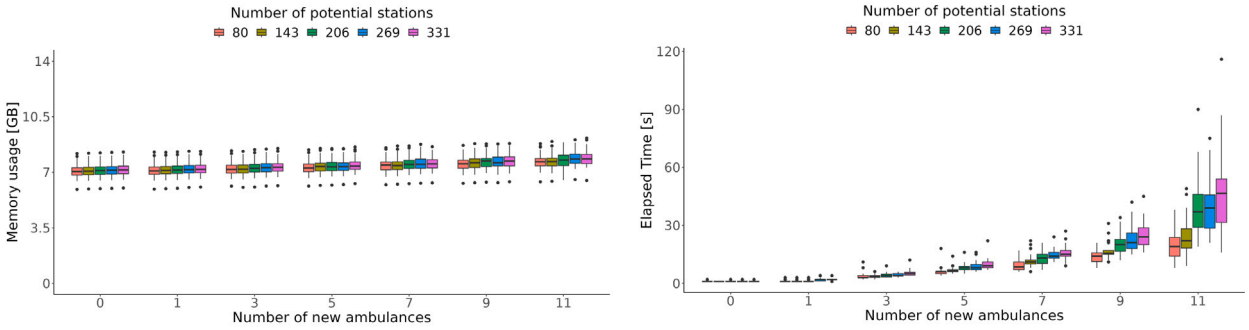
Table 7
Impact of relocations over the whole year by region: percentage of answered calls by RT-interval, region, the value of α coefficient, and number relocations allowed when solving the LAB and the EMILA models.

Number of	RT-interval	Baseline			Efficiency ($\alpha = 0$)			Balanced ($\alpha = 0.4$)			Equity ($\alpha = 1$)		
relocations		Urb.	Sub.	Rur.	Urb.	Sub.	Rur.	Urb.	Sub.	Rur.	Urb.	Sub.	Rur.
0	[0–15]	90.6	63.3	21.2	90.5	63.5	21.2	89.8	64.5	21.7	89.7	64.5	21.7
	(15–30]	4.2	16.6	25.2	7.7	33.9	59.7	8.2	33.1	63.3	8.3	33.3	63.3
	(30–45]	4.5	19.2	46.9	1.9	2.4	15.0	2.0	2.4	11.9	2.0	2.2	11.9
	NA	0.7	0.8	6.6	0.0	0.2	4.0	0.0	0.1	3.1	0.0	0.1	3.1
1	[0–15]	95.4	63.4	21.2	95.3	63.5	21.2	89.8	64.5	28.8	89.7	64.5	28.8
	(15–30]	1.9	16.6	22.1	2.8	33.8	61.1	8.2	33.4	58.8	8.3	33.6	58.8
	(30–45]	2.2	19.1	47.8	1.9	2.5	13.7	2.0	2.0	10.6	2.0	1.8	10.6
	NA	0.6	0.9	8.8	0.0	0.2	4.0	0.0	0.1	1.8	0.0	0.1	1.8
3	[0–15]	96.4	66.1	21.7	96.4	66.2	21.7	94.5	64.3	36.3	85.2	64.2	43.8
	(15–30]	0.8	14.6	23.0	1.8	30.3	61.1	3.4	33.4	52.7	13.1	33.0	43.8
	(30–45]	2.2	18.4	46.5	1.9	3.3	13.3	2.1	2.2	9.3	1.6	2.8	10.6
	NA	0.6	0.9	8.8	0.0	0.2	4.0	0.0	0.1	1.8	0.0	0.0	1.8
5	[0–15]	96.6	66.6	13.7	96.6	66.6	13.7	95.8	66.5	36.3	86.6	66.1	43.8
	(15–30]	0.6	16.1	29.6	1.5	29.8	66.8	2.1	30.5	52.7	12.2	30.4	43.8
	(30–45]	2.4	16.1	46.0	1.8	3.3	15.5	2.0	2.9	9.3	1.1	3.5	10.6
	NA	0.3	1.2	10.6	0.0	0.2	4.0	0.0	0.1	1.8	0.0	0.0	1.8

María Merino: Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Writing – review & editing.

Declaration of competing interest

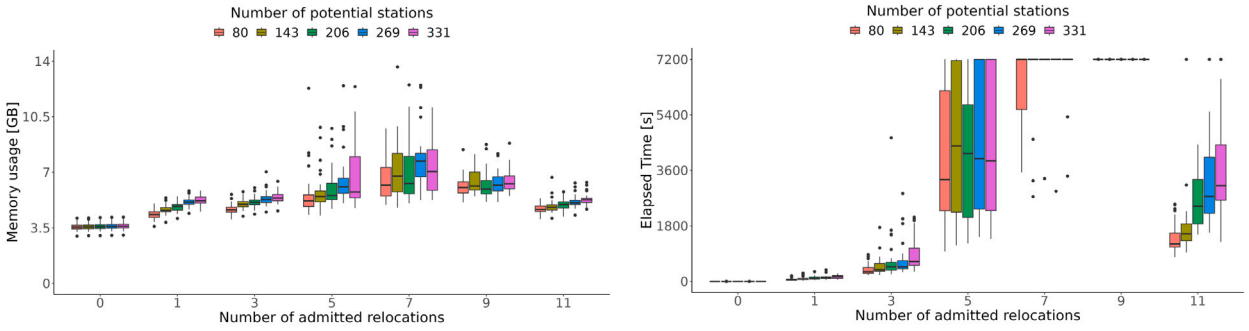
The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:



a Memory usage by number of new ambulances and number of potential stations

b Execution time by number of new ambulances and number of potential stations

Fig. A.8. Memory usage (left) and execution time (right) by number of new ambulances and number of potential stations.



a Memory usage by number of admitted relocations and number of potential stations

b Execution time by number of admitted relocations and number of potential stations

Fig. A.9. Memory usage (left) and execution time (right) by number of admitted relocations and number of potential stations.

Imanol Gago-Carro reports administrative support was provided by Osakidetza Emergencies.

Data availability

The data that has been used is confidential.

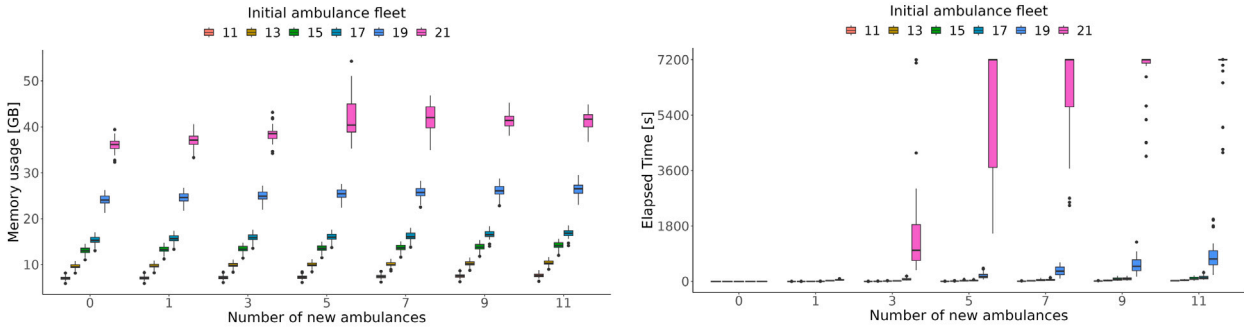
Acknowledgments

The authors are especially grateful to Emergentziak-Emergencias Osakidetza, the organization in charge of emergency health care coordination throughout the Basque Country, for the data and knowledge provided. The authors also thank IZO-SGI SGIker of UPV/EHU for the technical and human support provided.

This research has been partially supported by the Spanish Ministry of Science and Innovation through the project PID2019-104933GB-I00/AEI/10.13039/501100011033 and BCAM Severo Ochoa accreditation CEX2021-001142-S; and by the Basque Government through the program BERC 2022–2025, Elkartek Programs KK-2021/00065 and KK-2022/00106 and the projects IT1504-22 and IT-1494-22. Imanol holds a PRE2020-091984 Severo Ochoa grant from the Spanish Ministry of Science and Innovation. Open Access funding provided by University of Basque Country.

Appendix. Scalability analysis

In this appendix, we provide a scalability analysis of the dimensions of two specific sets of the model: ambulance stations (J) and ambulance initial fleet (I_0). First, we have included the analysis of the scalability and performance of the models depending on the size of the set of ambulance stations. Thus, optimization runs were conducted for five cases, from the initial one with $|J| = 80$ stations up to the last one, where we consider all the municipalities. As the Basque Country has 251 municipalities and in order to increase the size of the potential stations regularly, it is considered that $|J| = 80 + \lceil \frac{251 \cdot p}{4} \rceil$, for $p \in \{0, 1, 2, 3, 4\}$. Therefore, it has been tested $|J| \in \{80, 143, 206, 269, 331\}$ in the analysis. For each iteration, the municipalities have been added randomly. Figs. A.8(a) and A.8(b) show the results of this analysis for the location–allocation problem in terms of memory and elapsed time, respectively. There is no big deal when increasing the number of potential stations when solving the location–allocation problem. The memory usage (in GB) grows linearly when the number of potential stations increases and when the number of ambulances without pre-assigned stations does. Concerning the elapsed time, the growth is non-linear. Nevertheless, both the memory usage and elapsed time present no computational challenges since the former remains between 6 GB and 9 GB, and the latter does not exceed two minutes. Figs. A.9(a) and A.9(b) show the computational results for the relocation–allocation problem. When tackling the relocation–allocation problem, both the time and memory requirements show substantial growth compared to the location–allocation problem. It is worth highlighting the runtime limit of 2 h (7200 s). While the time limitation poses no issues when $k \leq 3$, it affects the relocation–allocation problem when the number of relocations available is equal to or greater than five ($k \geq 5$). This limitation is particularly remarkable when 7 and 9 ambulance relocations are permitted. In the former case, almost every execution run reaches this limit of 2 h. In the latter, not a single case is completed before this limit. Regarding the memory usage, the most demanding cases are $k = 7$ and $k = 9$. The optimization run with the most used resources does not reach 14 GB.



a Memory usage by number of new ambulances and size of the initial ambulance fleet

b Execution time by number of new ambulances and size of the initial ambulance fleet

Fig. A.10. Memory usage (left) and execution time (right) by number of new ambulances and size of the initial ambulance fleet.

In addition, the scalability of the problem has also been tested based on the size of the ambulance fleet. In the optimization runs made for the previous version of the manuscript, the current ALS fleet (11 ambulances) was used. Now, the experiments have been extended to 5 additional tests, increasing the initial ambulance fleet by adding 2 BLS ambulances in each iteration: $|I_0| = 11 + 2 \cdot p$ where $p \in \{0, \dots, 5\}$, hence $|I_0| \in \{11, 13, \dots, 21\}$. For each p, 30 optimization runs have been carried out. Concerning the emergency calls, all the emergencies attended by the initial ambulance fleet during the moment of the day with more activity (from 9 a.m. to 5 p.m.) are considered. This way, the average number of emergencies by optimization run are 1109, 1239, 1387, 1452, 1762 and 2090 for $|I_0| = 11, 13, 15, 17, 19$ and 21, respectively. Figs. A.10(a) and A.10(b) show the memory and time needed for these experiments. Notably, memory usage escalated rapidly: for instance, when considering only ALS ambulances ($|I_0| = 11$), the average memory usage is 7.36 GB. However, for the cases where $|I_0| = 19$ and $|I_0| = 21$, the average memory is 25.34 GB and 39.74 GB, respectively. This computational requirement is also reflected in the execution time: while for $|I_0| \leq 17$, the most demanding case is solved in 34 min, for $|I_0| = 21$ and $k > 3$, a 70% of the cases reached to the 2 h time limit. The memory usage and execution time of the experiments carried out in the scalability analysis are available at Gago-Carro et al..

References

Abouelijinane, L., Sahin, E., & Jemai, Z. (2013). A review on simulation models applied to emergency medical service operations. *Computers & Industrial Engineering*, *66*(4), 734–750. http://dx.doi.org/10.1016/j.cie.2013.09.017.

Aringhieri, R., Bruni, M., Khodaparasti, S., & van Essen, J. (2017). Emergency medical services and beyond: Addressing new challenges through a wide literature review. *Computers & Operations Research*, *78*(2015), 349–368. http://dx.doi.org/10.1016/j.cor.2016.09.016.

Bélanger, V., Kergosien, Y., Ruiz, A., & Soriano, P. (2016). An empirical comparison of relocation strategies in real-time ambulance fleet management. *Computers & Industrial Engineering*, *94*, 216–229. http://dx.doi.org/10.1016/j.cie.2016.01.023.

Bélanger, V., Ruiz, A., & Soriano, P. (2019). Recent optimization models and trends in location. *European Journal of Operational Research*, *272*(1), 1–23. http://dx.doi.org/10.1016/j.ejor.2018.02.055.

Bell, C., & Allen, D. (1969). Optimal planning of an emergency ambulance service. *Socio-Economic Planning Sciences*, *3*(2), 95–101. http://dx.doi.org/10.1016/0038-0121(69)90001-9.

Beraldi, P., Bruni, M., & Conforti, D. (2004). Designing robust emergency medical service via stochastic programming. *European Journal of Operational Research*, *158*(1), 183–193. http://dx.doi.org/10.1016/S0377-2217(03)00351-5.

Brotcorne, L., Laporte, G., & Semet, F. (2003). Ambulance location and relocation models. *European Journal of Operational Research*, *147*(3), 451–463. http://dx.doi.org/10.1016/S0377-2217(02)00364-8.

Chanta, S., Mayorga, M., & McLay, L. (2014). Improving emergency service in rural areas: a bi-objective covering location model for EMS systems. *Annals of Operations Research*, *221*(1), 133–159. http://dx.doi.org/10.1007/s10479-011-0972-6.

Church, R., & ReVelle, C. (1974). The maximal covering location problem. *Papers of the Regional Science Association*, *32*(1), 101–118. http://dx.doi.org/10.1007/BF01942293.

Clement, J., Khushalani, J., & Baernholdt, M. (2018). Urban-rural differences in skilled nursing facility rehospitalization rates. *Journal of the American Medical Directors Association*, *19*(10), 902–906. http://dx.doi.org/10.1016/j.jamda.2018.03.001.

Daskin, M. (1983). A maximum expected covering location model: Formulation, properties and heuristic solution. *Transportation Science*, *17*(1), 48–70. http://dx.doi.org/10.1287/trsc.17.1.48.

Ehrgott, M. (2005). *Multicriteria optimization* (2nd ed.). Springer.

Emergentziak Osakidetza (2017). Cartera de servicios de Emergentziak. https://www.osakidetza.euskadi.eus/contenidos/informacion/emer_cartera_servicios/es_emer/adjuntos/CARTERA_DE_%20SERVICIOS_DE_EMERGENTZIAK_ES.pdf [Online; accessed 15-October-2019].

Emergentziak Osakidetza (2019). Contrato programa emergencias de osakidetza 2019. https://www.legegunea.euskadi.eus/documentacion-relevancia-juridica/contrato-programa-emergencias-de-osakidetza-ano-2019/x59-contfich/es/ [Online; accessed 4-December-2019].

Erkut, E., Ingolfsson, A., & Erdogan, G. (2008). Ambulance location for maximum survival. *Naval Research Logistics*, *55*(1), 42–58. http://dx.doi.org/10.1002/nav.20267.

Gago-Carro, I., Aldasoro, U., Merino, M., & Ceberio, J. Equitable Multi-Interval (re)Location-Allocation (EMILA) models: codes and scalabilty analysis, mar 2024, Zenodo, http://dx.doi.org/10.5281/zenodo.10427658.

Gendreau, M. (1997). Solving an ambulance location model by tabu search. *Location Science*, *5*(2), 75–88. http://dx.doi.org/10.1016/S0966-8349(97)00015-6.

Gendreau, M., Laporte, G., & Semet, F. (2001). A dynamic model and parallel tabu search heuristic for real-time ambulance relocation. *Parallel Computing*, *27*(12), 1641–1653. http://dx.doi.org/10.1016/S0167-8191(01)00103-X.

Google Developers (2020). *Google maps distance matrix API*. Developer's Guide. Google, https://developers.google.com/maps/documentation/distance-matrix/overview [Online; accessed 7-June-2020].

Gutjahr, W., & Fischer, S. (2018). Equity and deprivation costs in humanitarian logistics. *European Journal of Operational Research*, *270*(1), 185–197. http://dx.doi.org/10.1016/j.ejor.2018.03.019.

Humphreys, J., McGrail, M., Joyce, C., Scott, A., & Kalb, G. (2012). Who should receive recruitment and retention incentives? Improved targeting of rural doctors using medical workforce data. *Australian Journal of Rural Health*, *20*(1), 3–10. http://dx.doi.org/10.1111/j.1440-1584.2011.01252.x.

IBM (2020). *Ilog cplex optimizer 20.1*. International Business Machines Corporation.

Jagtenberg, C., Bhulai, S., & van der Mei, R. (2015). An efficient heuristic for real-time ambulance redeployment. *Operations Research for Health Care*, *4*, 27–35. http://dx.doi.org/10.1016/j.orhc.2015.01.001.

Jagtenberg, C., & Mason, A. (2020). Fairness in the ambulance location problem: Maximizing the Bernoulli-Nash social welfare. *SSRN Electronic Journal*, 1–28. http://dx.doi.org/10.2139/ssrn.3536707.

Jonard, F., Lambotte, M., Bamps, C., Dusart, J., & Terres, J.-M. (2007). Review and improvements of existing delimitations of rural areas in Europe. *Jrc*, *EUR 22921*, 1–74.

Kaneko, M., Ohta, R., Vingilis, E., Mathews, M., & Freeman, T. (2021). Systematic scoping review of factors and measures of rurality: toward the development of a rurality index for health care research in Japan. *BMC Health Services Research*, *21*(1), 1–11. http://dx.doi.org/10.1186/s12913-020-06003-w.

Karpova, Y., Villa, F., Vallada, E., & Vecina, M. (2023). Heuristic algorithms based on the isochron analysis for dynamic relocation of medical emergency vehicles. *Expert Systems with Applications*, *212*(2022), Article 118773. http://dx.doi.org/10.1016/j.eswa.2022.118773.

Karsu, O., & Morton, A. (2015). Inequity averse optimization in operational research. *European Journal of Operational Research*, *245*(2), 343–359. http://dx.doi.org/10.1016/j.ejor.2015.02.035.

Kleywegt, A., Shapiro, A., & Homem-de Mello, T. (2002). The sample average approximation method for stochastic discrete optimization. *SIAM Journal on Optimization*, *12*(2), 479–502. http://dx.doi.org/10.1137/S1052623499363220.

Lamont, J., & Favor, C. (2017). Distributive justice. In E. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Winter 2017 ed.). Metaphysics Research Lab, Stanford University, https://plato.stanford.edu/archives/win2017/entries/justice-distributive/.

Larson, R. (1974). A hypercube queuing model for facility location and redistricting in urban emergency services. *Computers & Operations Research*, *1*(1), 67–95. http://dx.doi.org/10.1016/0305-0548(74)90076-8.

Leknes, H., Aartun, E., Andersson, H., Christiansen, M., & Granberg, T. (2017). Strategic ambulance location for heterogeneous regions. *European Journal of Operational Research*, *260*(1), 122–133. http://dx.doi.org/10.1016/j.ejor.2016.12.020.

Marín, A., Nickel, S., & Velten, S. (2010). An extended covering model for flexible discrete and equity location problems. *Mathematical Methods of Operations Research*, *71*(1), 125–163. http://dx.doi.org/10.1007/s00186-009-0288-3.

McLay, L., & Mayorga, M. (2010). Evaluating emergency medical service performance measures. *Health Care Management Science*, *13*(2), 124–136. http://dx.doi.org/10.1007/s10729-009-9115-x.

McLay, L., & Mayorga, M. (2013). A dispatching model for server-to-customer systems that balances efficiency and equity. *Manufacturing and Service Operations Management*, *15*(2), 205–220. http://dx.doi.org/10.1287/msom.1120.0411.

Naoum-Sawaya, J., & Elhedhli, S. (2013). A stochastic optimization model for real-time ambulance redeployment. *Computers & Operations Research*, *40*(8), 1972–1978. http://dx.doi.org/10.1016/j.cor.2013.02.006.

Nickel, S., Reuter-Oppermann, M., & Saldanha-da Gama, F. (2016). Ambulance location under stochastic demand: A sampling approach. *Operations Research for Health Care*, *8*, 24–32. http://dx.doi.org/10.1016/j.orhc.2015.06.006.

Noyan, N. (2010). Alternate risk measures for emergency medical service system design. *Annals of Operations Research*, *181*(1), 559–589. http://dx.doi.org/10.1007/s10479-010-0787-x.

Phillimore, P., & Reading, R. (1992). A rural advantage? urban-rural health differences in Northern England. *Journal of Public Health*, *14*(3), 290–299. http://dx.doi.org/10.1093/oxfordjournals.pubmed.a042745.

Rawls, J. (1971). *A theory of justice*. Harvard University Press.

Rousseau, N. (1995). *What is rurality?: Occasional paper (Royal College of General Practitioners) (71)*, (pp. 1–4).

Sargent, R. (2013). Verification and validation of simulation models. *Journal of Simulation*, *7*(1), 12–24. http://dx.doi.org/10.1057/jos.2012.20.

Schmid, V. (2012). Solving the dynamic ambulance relocation and dispatching problem using approximate dynamic programming. *European Journal of Operational Research*, *219*(3), 611–621. http://dx.doi.org/10.1016/j.ejor.2011.10.043.

SGIker (UPV/EHU) (2020). ARINA: Computational cluster from IZO-SGI. https://www.ehu.eus/sgi/recursos/cluster-arina/.

Smith, H., Harper, P., & Potts, C. (2013). Bicriteria efficiency/equity hierarchical location models for public service application. *Journal of the Operational Research Society*, *64*(4), 500–512. http://dx.doi.org/10.1057/jors.2012.68.

Swan, G., Selvaraj, S., & Godden, D. (2008). Clinical peripherality: Development of a peripherality index for rural health services. *BMC Health Services Research*, *8*, 1–10. http://dx.doi.org/10.1186/1472-6963-8-23.

Takeda, R., Widmer, J., & Morabito, R. (2007). Analysis of ambulance decentralization in an urban emergency medical service using the hypercube queueing model. *Computers & Operations Research*, *34*(3), 727–741. http://dx.doi.org/10.1016/j.cor.2005.03.022.

Toregas, C., Swain, R., ReVelle, C., & Bergman, L. (1971). The location of emergency service facilities. *Operations Research*, *19*(6), 1363–1373. http://dx.doi.org/10.1287/opre.19.6.1363.

Wang, W., Wang, S., Zhen, L., & Qu, X. (2022). EMS location–allocation problem under uncertainties. *Transportation Research Part E: Logistics and Transportation Review*, *168*(January), Article 102945. http://dx.doi.org/10.1016/j.tre.2022.102945.

Xinying Chen, V., & Hooker, J. (2023). A guide to formulating fairness in an optimization model. *Annals of Operations Research*, *326*(1), 581–619. http://dx.doi.org/10.1007/s10479-023-05264-y.

Yoon, S., Albert, L., & White, V. (2021). A stochastic programming approach for locating and dispatching two types of ambulances. *Transportation Science*, *55*(2), 275–296. http://dx.doi.org/10.1287/TRSC.2020.1023.