MULTI-OBJECTIVE OPTIMIZATION OF AMBULANCE LOCATION IN ANTOFAGASTA, CHILE

Carlos OLIVOS*, Hernan CACERES
Dept of Industrial Engineering, Catholic University of the North, Antofagasta, Chile
Submitted 17 December 2019; resubmitted 16 March 2020; accepted 15 June 2020

Abstract. In this paper, we solved an ambulance location problem with a multi-objective framework considering the case of the study of Emergency Medical Service (EMS) of Antofagasta (Chile). Nowadays, in Antofagasta, the ambulances are located in bases that are not necessarily the optimal location achieving an estimated 67% of coverage under the 8 min not meeting the requirements dictated by the Chilean Ministry of Health. We used a multi-objective model considering mean response time, maximum response time, and the demand not covered. The model is solved using an iterative ε-constraint method to generate a Pareto set of efficient solutions. We considered historical data from the years 2015 and 2016 to generate the demand and emergency nodes with a clustering algorithm. The results show improvements on all criteria of the multi-objective model, where we highlight a potential increment on coverage within 8 min from 67 to 99%. In order to test the new policy in a real setting, a pilot plan is proposed, which reaches 89% of coverage within 8 min.

Keywords: ambulance location, multi-objective optimization, multi-criteria decision-making, emergency medical service, Chile.

Introduction

The Emergency Medical Service (EMS) is one of the most critical services that cities provide to their citizens. Along with firefighters and the police, their primary goal is a consistent and timely response. The EMS must provide the 1st care attention to anyone involved in an accident and must be available 24/7. The time between an emergency call and the corresponding arrival of an ambulance to the scene is the response time. The response time is an essential indicator when assessing the performance of the ambulance service, but the criterion with which it is used in different cities varies. New regulation of the Chilean Ministry of Health requires response times under 8 min in 90% of cases for all urban areas (Subsecretaría de Redes Asistenciales 2018), but in Antofagasta, the 5th largest urban area in Chile, we estimated that in no more than 67% of the cases the response is below 8 min. This gap needs to be reduced, and especially because there are time dependent pathologies among those 33% of cases for which response is above 8 min. In this work, we aimed to improve response time by means of implementing a new policy for locating ambulances throughout the urban area of Antofagasta city using multi-objective optimization. We found that deploying the new policy for one ambulance would result in responding to 89% of all cases under 8 min, and that for a full deployment with the entire fleet it is possible to respond to 99% of all cases under 8 min.

In the case of English ambulances services, the target is to respond to emergencies in less than 8 min in 75% of cases (Heath, Radcliffe 2007, 2010). In the United States, there is no federal or state law that enforces any standard in relation to response time, and the only regulations are contractual agreements between private EMS providers and local governments. Many of these agreements require response times under 8 min in 90% of cases, and others, like in some municipalities in California, have set response time standards varying from 12 to 15 min in 90% of cases for the same service (EMS World 2004). In Ontario (Canada), the Ambulance Act defines response time without setting a unique goal for the province but mandates municipalities to establish their own (Government of Ontario 2022).

Governments make an effort to lower response time by setting performance targets and investing accordingly. However, there is no consensus regarding the effect of response time on survival rate. Many studies have found little or no effect, and others have found a significant effect for some instances. Blackwell and Kaufman (2002) found that the curve of mortality risk versus response time is flat after the 5 min threshold, showing that survival can be improved if response times are reduced under 5 min.

*Corresponding author. E-mail: carlos.olivos@ucn.cl

Copyright © 2022 The Author(s). Published by Vilnius Gediminas Technical University

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.
However, achieving this objective would incur extraordinary cost expenditure. In trauma emergencies, Newgard et al. (2010) indicated that it was not possible to demonstrate a relationship between response time and survival possibilities; Lerner, Moscati (2001) and Pons, Markovchick (2002) show similar results. In contrast, another stream of studies states the opposite: response times have a significant effect on survival risk, especially for certain types of emergencies, such as cardiac arrest or strokes. Renkiewicz et al. (2014) indicate that the shock-able presenting rhythm decreased by 8% for every minute since a cardiac arrest, and O’Keeffe et al. (2011) mentioned that reducing the response time in 1 min improves the odds of survival in 24%, though achieving such reduction on response time is estimated to cost around £54 million.

Many studies adhere to the 8 min rule when measuring performance even though there are studies that find this threshold is too high. Pons et al. (2005) highlights that gains in survival possibilities due solely to response time are significant when this is under 4 min. In research by De Maio et al. (2003), their main conclusion is that EMSs system leaders should lower the 8 min mark for defibrillation response time. Despite the debate about the effect of response time over the survival rate, the public will always demand lower response times for any emergency. It is the role of local governments and EMS administrators to continually review their policies to ensure a timely and effective response. It follows that location of ambulances is critical for citizen wellness, especially for patients who have a cardiac history, where each minute count to keep an individual alive. Although the ambulance location problem has been studied in different countries, in Chile there is only one study made by Singer and Donoso (2008). They assess the performance of the Mobile Coronary Care Unit (in Spanish – Unidad Coronaria Móvil (UCM)) in Santiago (Chile). Their work was based on queuing theory techniques and the reallocation of ambulances in their existing bases. Our work advances in the application of the ambulance location problem in Chile to show local policy makers the benefits of improving ambulance location strategies.

Although coverage is studied considerately in the literature on operations research, in the optimization, this objective ignores the demand located outside of the coverage radius. Therefore, one must consider other objectives such as mean and maximum response time to properly balance all demand. Nevertheless, optimizing only the mean response time may not be the best criteria to choose ambulance locations since it tends to benefit emergencies in areas with higher demand while being detrimental to those demand locations that are dispersed (Calik et al. 2015).

Considering there is not a unique criteria to assess ambulance performance, we analysed the case of study of Antofagasta, with multi-objective optimization scope to study the trade-off among 3 objectives: (1) coverage, (2) mean response time and (3) maximum response time. Before this work, Antofagasta EMS did not have any method to assess the performance of their current policy neither had an analytical method to evaluate any modifications. Our research focused on assessing the performance of ambulances and proposed a new location to improve the performance of the Antofagasta EMS. The results will give Antofagasta EMS a set of efficient solutions where the staff must decide according to their preferences. The model is solved using the ε-constraint method generating an efficient set of solutions. Finally, we proposed a pilot plan to implement and assess the impact of it in the city. This paper is organized as follows: Section 1 shows a brief of ambulance location models. The problem description is presented in Section 2. The mathematical formulation of the multi-objective model is presented in Section 3. The steps to obtain the data used is explained in Section 4. Results and discussion are shown in Section 5 and Section 6, respectively. Conclusions are shown in the last section.

1. Literature review

The ambulance location problem has been studied since the 70s, highlighting 3 measures of performance frequently used: the mean response time, the maximum response time, and the coverage Ahmad-Javid et al. (2017). Among these 3, the most commonly used is the coverage. A demand location is considered “covered” if it can be reached in no more than a certain time or distance. Toregas et al. (1971) propose the Location Set Covering Model (LSCM), where the objective is to minimize the number of ambulances needed to cover all demand locations in a given time or distance radius. However, their model does not consider limited resources number of emergency vehicles or facilities are limited. Church and ReVelle (1974) propose the Maximal Covering Location Problem (MCLP), where they consider a limited number of available facilities. The model objective is to maximize the demand coverage with a limited number of facilities.

Erkut et al. (2008) propose an alternative use of coverage where they maximize the expected number of cardiac arrest survivors using a survival probability function. Then, Knight et al. (2012) extend the previous work with heterogeneous patients using disease dependent survival probability functions. The literature presents surveys of work using coverage as a performance measure on their models, from deterministic to stochastic models, including static and dynamic models (Brotcorne et al. 2003; Li et al. 2011). In another survey, Bélanger et al. (2019) highlighted that there are tools to work with data in real-time to support decision-making, and they should be considered in research of this field. However, it has not been widely used because handling data in real-time is still challenging. Nevertheless, there are works with the application of real-time decisions.

Enayati et al. (2018) proposed a real-time redeployment model where they maximize the coverage and minimize the total travel time with the coverage value as a constraint. Nasrollahzadeh et al. (2018) made real-time...
dispatching and relocation model using stochastic programming solved by approximated dynamic programming. Their model is tested in the EMS system of Mecklenburg Count (North Carolina, United States). Van Buuren et al. (2018) propose the dynamic maximum expected coverage location and the penalty heuristic as a new policy for the EMS of Flevoland (Netherlands). Their new policies were tested for 12 weeks, and because of their success, they were later implemented.

Another formulation for the ambulance location problem is the \( p \)-median (ReVelle, Swain 1970; Hakimi 1964), where the objective is minimizing the average response time. Carson and Batta (1990) use this model to locate ambulances on the Amherst campus of the State University of New York at Buffalo, where they archive 6% reduction of the average response time in field experiments. Serra and Marianov (1998) use the \( p \)-median to locate fire stations in Barcelona (Spain), including uncertainty in travel times and demand. Some researchers have focused on the development of heuristics (Erkut et al. 2008; Alp et al. 2003; Caccetta, Dzator 2015; Dzator, M., Dzator, J. 2013), because the \( p \)-median is considered a Nondeterministic Polynomial (NP) hard problem (Kariv, Hakimi 1979).

Rikalović et al. (2018) propose an approach based on Geographic Information System (GIS) and Strengths, Weaknesses, Opportunities, and Threats (SWOT) analysis to locate a logistic center using as a case of study the Municipality of Apatin (Vojvodina, Serbia). Although they concluded their tool is efficient, they do not show numerical results that support their findings. Valencia-Nuñez et al. (2018) made a Monte Carlo simulation where they concluded that a relocation of the ambulances reduces the arrival time significantly. Moreover, they used GIS to geo-referencing the demand, and to estimate the time matrix of the region of Morona Santiago in Ecuador.

There are different criteria to assess the performance of ambulance locations. Goldberg (2004) proposes minimizing the overall average time to serve emergency calls, minimizing the maximum travel time to any single call, (1) maximizing the area covered under a certain time, and (2) maximizing the number calls covered under a certain time. The last 2, though similar, are not equivalent since zones do not necessarily have the same call loads. Since more than one of these objectives might be desirable to optimize, authors develop multi-objective formulations of the ambulance location problem. Eaton et al. (1986) work in a case study of Santo Domingo (Dominican Republic). Their model minimizes the numbers of ambulances and maximizes multiple demand coverage. Harewood (2002) also works with a multi-objective problem considering to maximize the coverage and to reduce the cost of such coverage. Talwar (2002) uses the \( p \)-median and \( p \)-center to locate helicopters in South Tyrol (Italy), using heuristics to find approximate solutions to improve the response times. In the field of logistics Milosavljević et al. (2018) made a multi-criteria decision-making analysis to decide the macro location of a railroad container terminal in Serbia. They used Delphi and the entropy method to define the weight of the different regions to be selected, then they use Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS), Elimination and Choice Translating Reality (in French – Élimination Et Choix Traduisant la Réalité – ÉLECTRE) and Multi-Attributive Border Approximation area Comparison (MABAC) as methods to rank the possible macro locations. Karatas and Yakıcı (2018) present an iterative method to solve a multi-objective formulation of \( p \)-median, \( p \)-center, and Maximal Coverage Location Problem (MCLP).

2. Problem statement

In this paper, we focused on assessing the current ambulance locations and propose a new efficient location set to improve the current performance measures in which Antofagasta EMS is interested. The Antofagasta EMS has the responsibility of providing emergency service to the population of Antofagasta, a city located in northern Chile with a population of 361873 (INE 2017). The current ambulance locations were set in place without the use of precise guidelines regarding response time, and they have not been assessed analytically since. Figure 1 shows the distribution of emergencies from years 2015 and 2016, and the current position of the ambulances. We can see that the current locations do not necessarily follow the distribution of the emergencies.

In Antofagasta, ambulances provide not only emergency care but also transportation of critical patients.

![Figure 1. Emergencies heat map in Antofagasta (log scale)](image-url)
among care centers of the city; however, we only considered emergency care in our work. The dispatch occurs when the call center receives a call and verify if there are ambulances available to the dispatch. Otherwise, the patient will have to wait for the availability of other ambulance or go to a health center by its means. The Antofagasta EMS has a fleet of 5 vehicles distributed among 3 health care facilities. During the day shift, 5 ambulances are available to cover all emergencies, but during the night shift there are only 4. Hence, we will solve the problem in section 4 considering 2 shifts as different instances, where each instance will consider their corresponding number of ambulances. Finally, by better positioning the ambulances we aim to decrease the response time and improve survival rates.

3. Mathematical formulation

The following multi-objective model incorporates minimizing 3 objectives: (1) mean and (2) maximum response time and (3) demand not covered. Notice that demand not covered is the complement of coverage; we use the former instead of the latter so that all 3 objectives are minimizations. We adapted the MCLP proposed by Church and ReVelle (1976) objective so that it can be modelled as a \( p \)-median problem. Consider \( N \) as the set of emergency demand nodes. For simplification, we considered that \( N \) also represents the set of potential ambulance locations. Let \( \tau \) be the maximum response time for which we will consider a node to be covered, \( t_{ij} \) is the response time defined as the travel time between \( i \in N \) and \( j \in N \) plus the setup time at \( i \in N \), \( w_i \) is the relative frequency of emergency calls of node \( i \in N \), and \( K \) is the number of available ambulances. The decision variables \( x_{ij} \) is 1 if an ambulance is positioned at node \( j \in N \) and 0 otherwise, and \( x_{ij} \) is 1 if demand node \( i \in N \) is assigned an ambulance located at node \( j \in N \) and 0 otherwise. The formulation of our problem reads as follows:

\[
\begin{align*}
\min f_1 &= \sum_{(i,j) \in N^2} w_i \cdot t_{ij} \cdot x_{ij}; \\
\min f_2 &= \max_{i \in N} \left( \sum_{j \in N} t_{ij} \cdot x_{ij} \right); \\
\min f_3 &= \sum_{(i,j) \in N \setminus \{i,j\} \geq \tau} w_i \cdot x_{ij}; \\
\text{s.t.} & \quad \sum_{j \in N} y_{ij} \leq K; \\
& \quad x_{ij} \leq y_{ij}, \quad \forall i \in N, \ j \in N; \\
& \quad \sum_{j \in N} x_{ij} = 1, \quad \forall i \in N; \\
& \quad y_{ij} \in \{0,1\}, \quad \forall j \in N; \\
& \quad x_{ij} \in \{0,1\}, \quad \forall i \in N, \ j \in N,
\end{align*}
\]

where: objective (1) minimizes the weighted response time; objective (2) minimizes the maximum response time; objective (3) minimizes the uncovered demand fraction within a response time equal to \( \tau \); constraint (4) indicates the number of available ambulances to assign; constraint (5) restricts demand nodes to be assigned ambulance locations that do not have an ambulance; constraints (6) ensure all nodes must be assigned to one and only one ambulance; constraints (7) and (8) are the binary constraints.

3.1. Solution procedure: \( \epsilon \)-constraint method

Let us consider the following multi-objective optimization problem:

\[
\begin{align*}
\min_{x} & \quad f_1(x), \ldots, f_p(x); \\
\text{s.t.} & \quad x \in S,
\end{align*}
\]

where: \( x \) is the vector of decision variables; \( f_k(x) \) are the objective functions with \( k = 1, \ldots, p \); \( S \) is the feasible region.

The \( \epsilon \)-constraint method reformulates the model using the objectives as constraints as in the following model:

\[
\begin{align*}
\min f_1(x); \\
\text{s.t.} & \quad f_k(x) \leq \epsilon_k, \quad \forall k = 2 \ldots p; \\
& \quad x \in S.
\end{align*}
\]

Mavrotas (2009) proposed an iterative methodology using the \( \epsilon \)-constraint method to solve multi-objective problems generalizing only to maximizing problems. This method allows keeping the model structure including in it each one of objectives as a constraint. Furthermore, it is an exact method that integrated an exact mono-objective resolution for each objective. In this problem, we used his augmented method applied to a multi-objective problem of minimization. The 1st step is to calculate the payoff table through lexicographic optimization and from the payoff table obtain a range between the maximum and minimum value of each one of the \( p-1 \) objectives that will use as constraints. Then divide the range of \( k \) th objective function to \( \epsilon_k \) equal intervals using \( \epsilon_k - 1 \) intermediate equidistant grid points. Figure 2 shows the feasible region when \( q = 4 \) for the objectives \( f_2 \) and \( f_3 \).

Then we have in total \( q_k + 1 \) grid point that are used to vary the parameter \( \epsilon_k \). The total number of runs becomes \( (q_2 + 1) \cdot (q_3 + 1) \cdot \ldots \cdot (q_p + 1) \). The multi-objective problem as follow:

\[
\begin{align*}
\min f_1(x) - \epsilon \cdot \sum_{k=2}^{p} s_k; \\
\text{s.t.} & \quad f_k(x) + s_k = \epsilon_k, \quad \forall k = 2 \ldots p; \\
& \quad x \in S; \\
& \quad s_k \geq 0, \quad \forall k = 2 \ldots p.
\end{align*}
\]

Consider \( s_k \) as slack variables, \( \epsilon_k = u \cdot b_k - \frac{i_k \cdot r_k}{\delta_k} \), where: \( r_k \) is the range of objective \( k = 1, 2, \ldots, p \); \( i_k \) is a
counter of the objective \( k = 1, 2, \ldots, p \); \( \varepsilon \) a value from \( 10^{-3} \) to \( 10^{-6} \).

The objective (14) to minimize the sum of the mean response time (1) minus the slack variables \( s_k \); the 2nd term in the objective function does that the others 2 objectives will be reduced to maximum. The constraint (15) brings near the objectives values to \( k = 2 \ldots p \) with distance of \( s_k \) units. The feasible region (4)–(8) is contained in constraint (16) and the constraint (17) indicates that slack variables are non-negative.

The \( \varepsilon \)-constraint method has been applied successfully in different contexts in the field of operations research. Fetzer et al. (2018) solved the multi-objective location problem of road weather information system. They considered the maximization of vehicle miles traveled of the road segments, coverage of the sensors, and safety. In humanitarian logistics, Haghi et al. (2017) developed a multi-objective model to ensure the well distribution of relief goods under a disaster context. In the same context, Fahimnia et al. (2017) used the \( \varepsilon \)-constraint method combined with Lagrangian relaxation for the efficient design of the supply of blood in disasters. In the electrical sector, Chamandoust et al. (2020) developed a tri-objective model to optimize the distribution of energy in a residential smart electrical grid with renewable energy sources. They used the algorithm to build the Pareto frontier, and among the non-dominated solutions, the best is made based on the ideal point.

4. Model application for the Antofagasta EMS

4.1. Data gathering

The model (1)–(8) considers the following information:

- the demand and ambulance location nodes; the city is represented under a node network;
- the emergency calls as demand; Antofagasta EMS provides historical data of the year 2015 and 2016;
- the expected travel time plus setup time between the nodes of the city.

The Antofagasta EMS data has 13394 and 25429 records in the year 2015 and 2016 respectively, including patient transports and emergency calls. Each record contains the date, a unique id, an emergency type, the type of ambulance used, the event address, and the starting time of each stage of the ambulance dispatching process, from when an emergency call arrives until the corresponding ambulance is available again. Figure 3 shows all the stages of the ambulance dispatching process, where each time stamp is defined as follows:

- \( t_1 \) – time at which an emergency call arrives;
- \( t_2 \) – time at which an ambulance is assigned to the emergency;
- \( t_3 \) – time at which the ambulance leaves the base;
- \( t_4 \) – time at which the ambulance arrives at the place of emergency;
- \( t_5 \) – time at which the ambulance leaves the emergency site;
- \( t_6 \) – time at which the ambulance returns to the base;
- \( t_7 \) – time at which the ambulance leaves the patient;
- \( t_8 \) – time at which the ambulance becomes available again.

From all the records, we excluded 16206 entries corresponding to patient transports. Then, the remaining 22617 records are all the emergency calls for which we proceeded to geocode their addresses using the Google Geocoding Application Programming Interface (API) (Google LLC 2014) to obtain the corresponding geographic coordinates. We retrieved a response with each query or request to the Google Geocoding API that contains, among other information, the coordinates and the location type. The location type indicates the response precision level, and it can be one of the following, sorted in descending precision: rooftop, range interpolated, geometric center and approximate. A request to the Google Geocoding API using a correct address should always yield either rooftop or range interpolated as location type. When it does not, it means that the address was not found, or it could not be approximated. For the most part, addresses are not found when they are incomplete or incorrect in the original data set. Thus, we only considered the emergencies with a correctly identified address, also discarding those emergencies located outside city boundaries. Finally, the resulting data set that we will use hereafter consists of 11372 records, where the number of emergencies and the ambulances available for each of the 2 working shifts are shown in Table 1.

4.2. Application of multi-objective model

The multi-objective model (1)–(8) is applied as follows:

- \( t_1 \)–\( t_2 \)–\( t_3 \)–\( t_4 \)–\( t_5 \)–\( t_6 \)–\( t_7 \)–\( t_8 \)–\( t_9 \), where each time stamp is defined as follows:

Figure 3. Timeline of ambulance dispatch process
Table 1. Number of emergencies in each shift

<table>
<thead>
<tr>
<th>Shift</th>
<th>Hour</th>
<th>Number of emergencies</th>
<th>Ambulances</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>08:00…17:00</td>
<td>5597</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>17:00…08:00</td>
<td>5775</td>
<td>4</td>
</tr>
</tbody>
</table>

4.2. Demand and ambulance location nodes

We generated the set of demand and ambulances location nodes from the records of Antofagasta EMS, a set containing 11372 records. If we consider each record as a node in our model, we obtain $129322384 \times 5597 + 11372 \times 4$ binary variables, a similar number of constraints, and we would also have to estimate the travel time $t_{ij}$ for each $x_{ij}$. Obtaining over 100 million travel times with the shortest path algorithm is very time-consuming. Then, trying to solve a model this size is not feasible for general use computers. Thus, we develop criteria to reduce the number of nodes to avoid computational complexity without tampering too much with the quality of the solution.

We reduced the number of nodes using the algorithm of hierarchical agglomeration of Ward (https://ward.readthedocs.io/en/latest/index.html) in Python 3.6 (https://www.python.org) with the library “scikit.learn” (Pedregosa et al. 2011). Dibene et al. (2017) worked with this methodology to generate 100 demand nodes to simplify the optimization model. However, they did not establish a criterion for choosing the number of clusters. In this paper, we proposed 3 criteria to choose the number of clusters to be used: (1) computational effort, (2) the precision of the objective functions, (3) the distance between an emergency and the center of the cluster to which it belongs.

Because the ward algorithm calculates the Euclidean distance between nodes, we transformed their geographical coordinates to rectangular coordinates using the southwest most point as reference. The algorithm labels each node with their corresponding cluster ID. Then, we set the coordinates for each cluster $k$ with the coordinates of node $\arg\min_{i \in C_k} \left( \max_{j \in C_k} d_{ij} \right)$, where $C_k$ is the set of nodes for cluster $k$.

Recall that we plan to solve the location problem for each of 2 working shifts of Antofagasta EMS as different instances. However, we want the set of locations to be the same for both, with the weight of each location node being different in each case. Then, we merged the data for both shifts to create one set of emergencies that we will use during the procedure to reduce the number of nodes.

4.2.1. Computational effort

We solved the model proposed in section 4 for instances between 100 and 1000 clusters with each objective function representing an independent instance. We implemented the model in Python 3.6 and solved it using Gurobi 8.0.1 (https://www.gurobi.com) and IBM ILOG CPLEX Optimization Studio V 12.8.0 (https://www.ibm.com/support/pages/downloading-ibm-ilog-cplex-optimization-studio-v1280). The computational tests were run on HP Proliant DL360p Gen8, 2 CPU Intel Xeon E5-2650 v2 2.6 GHz with 8 cores

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Variable</th>
<th>$p$-median</th>
<th>$p$-center</th>
<th>MCLP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>CPLEX</td>
<td>Gurobi</td>
<td>CPLEX</td>
</tr>
<tr>
<td>100</td>
<td>10100</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>200</td>
<td>40200</td>
<td>7</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>300</td>
<td>90300</td>
<td>19</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>400</td>
<td>160400</td>
<td>49</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>500</td>
<td>250500</td>
<td>35</td>
<td>0</td>
<td>15</td>
</tr>
<tr>
<td>600</td>
<td>360600</td>
<td>72</td>
<td>0</td>
<td>25</td>
</tr>
<tr>
<td>700</td>
<td>490700</td>
<td>95</td>
<td>0</td>
<td>37</td>
</tr>
<tr>
<td>800</td>
<td>640800</td>
<td>128</td>
<td>0</td>
<td>63</td>
</tr>
<tr>
<td>900</td>
<td>810900</td>
<td>218</td>
<td>0</td>
<td>73</td>
</tr>
<tr>
<td>1000</td>
<td>1001000</td>
<td>259</td>
<td>0</td>
<td>85</td>
</tr>
</tbody>
</table>
each and 64 GB RAM. Table 2 shows a comparison between CPLEX and Gurobi. We decided to use Gurobi because this optimizer reaches better computational time than CPLEX for our case. The computational comparison is done with a time limit of 3600 sec; when an instance reaches this time limit the table also shows the optimality gap.

Figure 4 shows computational times using Gurobi and with no time limit for each objective in each cluster instance. We run the optimization with each one of the 3 objectives independently and varying the number of clusters from 100 to 1000. When solving the model with 800 or more clusters, the running time increased considerably for the 2nd objective. Therefore, we were not able to run instances over 1000 clusters to optimality with the 2nd objective of how it shows in Figure 4. According to the above mentioned, we can not consider more than 1000 clusters in our solution procedure because the 2nd objective will not reach the optimality result.

4.2.2. Accuracy of objectives values
Simplifying the model by clustering the demand nodes could decrease the accuracy of objectives. Figures 5–7 show how changes each objective function for each cluster. However, the objective functions maintain steady by solving the problem with 300 clusters for the mean response time, 400 clusters for the maximum response time and the coverage are steady on each group. The above implies that we can solve the model with a small number of clusters and the values of the objective functions will not be affected. Furthermore, the difference that exists between each of the instances with different clusters is not relevant, since for example, in the case of Figure 6 the maximum difference between the instance of 100 clusters and that of 600 is 1.1 min. Concerning the uncovered demand, the percentage variation is maximum of 1% because this objective is sensitive to the coverage time $\tau$.

4.2.3. Distance at center of each cluster
The last criterion considered was the distance between an emergency and the center of the cluster to which it belongs. Figure 8 shows a comparison between the distance distribution and the number of clusters. As the number of clusters increases, clusters decrease in size and agglomerate fewer emergencies. When reaching the 600 clusters, the 75% of the emergencies are at a distance less than 100 m from the center of the cluster. This distance is considered negligible for an emergency vehicle. Thus, we chose to set the number of demand nodes to 600, obtaining the distribution of emergency demand nodes in the city presented in Figure 9.
4.3. Expected time
We built the expected travel time matrix using the Open Source Routing Machine (OSRM) developed by Luxen and Vetter (2011). Even though it does not consider the vehicle traffic during the day, in future works we hope to obtain precise information on ambulance travel time including traffic. We obtained the setup time from the historical data, considering only data with a setup time between 1 and 10 min.

5. Results
As mentioned above, we solved the model on Gurobi 8.0.1 coded in Python 3.6. We worked with a parameter of coverage $\tau = 8$ min for objective (3) according to the guidelines of the Department of Health of Chile (SdRA 2018). Furthermore, we considered the parameter $K$ in constraint (4) as 5 and 4 ambulances in shift 1 and 2, respectively.

In our $\varepsilon$-constraint implementation we used $\varepsilon = 10^{-3}$ and $q_2 = q_3 = 4$ resulting in 25 optimization runs. Table 3 shows the resulting efficient frontier, where the 3rd column contains the values obtained in objective (14). The last 3 columns show the objective functions (1), (2) and (3) respectively without considering the slack variables $s_k$ of the constraint (15). Some rows have the same result because the constraint (15) requires equality with the variables $e_k$ (columns 4 and 5). However, there is no combination that allows to fulfil the equality in the constraint (15) using the variables of the functions (2) and (3) only, so they must be activated to satisfy the constraint (15). From all solutions, we obtained 6 and 5 efficient distinct solutions in shift 1 and 2, respectively.

6. Discussion
According to our results, there is a significant difference between the current situation and our set of efficient solutions. Figures 10 and 11 show the current situation in a single red bubble and our efficient solutions in the ideal situation with blue bubbles. Hence, we need to devise a strategy that would allow the city to improve. Implementing any of the efficient solutions we propose would mean taking all 5 ambulances out of the bases and positioning them throughout the city, and it would also mean changing the way paramedics do things. A more conservative, and plausible to be implemented, approach is to have only one ambulance deployed in the city and the rest in their current bases; we call this the pilot plan. This pilot plan consists in relocating one ambulance for each shift keeping the others in their current locations. To find efficient solutions for the pilot plan, we run our $\varepsilon$-constraint implementation with the same parameters but enforcing the current locations to be chosen by the model and allowing only one ambulance to choose any other location. Figures 10 and 11 also show efficient solutions for the pilot plan in orange bubbles. Considering the solutions with the best coverage performance our pilot plan improves coverage in 22 and 25% for shift 1 and 2, respectively, that is, 1231 and 1443 additional emergencies can be attended in less than 8 min. In the ideal case where we can relocate all ambulances, it is possible to achieve 1791 and 1778 additional emergencies regarding the current situation, considering the solution with the best performance for shift 1 and 2.

Figures 12 and 13 show the trade-off by means of ambulance locations where each objective reaches the best performance. We can see the same trade-off saw in Figures 10 and 11. When the mean response time decreases, the ambulances are located near the places where emergencies are more frequent (blue diamonds) but the maximum response time increases. However, when the maximum response time decreases, ambulances are located further from each other (black diamonds), improving the response time of distant emergencies and increasing the mean response time. Moreover, when the coverage is maximum, the ambulances are located at a midpoint between the mean and maximum response time locations (orange diamonds).

The $\varepsilon$-constraint method is an exact method that guarantees to get Pareto frontier points, which is one of the biggest differences regarding metaheuristics where the use of them does not ensure getting the global optimum. If metaheuristics methods are used, dominated solutions can be reached (Jaszkiewicz, Branke 2008). Moreover, the use of metaheuristics is not required in this problem because since it is a planning problem, it is not necessary to be solved in real-time or at the beginning of each day.
In this paper, we present a multi-objective ambulance location problem considering Antofagasta (Chile) as the case study. Previous researches have mostly focused on assessing ambulance performance considering only coverage criteria. However, coverage does not consider the emergencies outside the radio, excepting models that include double or triple coverage. Thus, we formulated this problem with a multi-objective model adding mean response time and maximum response time to the coverage as optimization functions. This approach produces an efficient set of solutions from which the decision-maker can choose based on their trade-off and his or her preference.

We worked with the emergency addresses and geo-coded them using Google Geocoding API. Given the high

<table>
<thead>
<tr>
<th>Shift</th>
<th>Solution</th>
<th>( f_1(x) - \epsilon \cdot \sum_{k=2}^{p} \frac{s_k}{s_k} )</th>
<th>( e_2 )</th>
<th>( e_3 )</th>
<th>( f_1(x) )</th>
<th>( f_2(x) )</th>
<th>( f_3(x) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>-2.457</td>
<td>17.967</td>
<td>0.275</td>
<td>6.081</td>
<td>17.8</td>
<td>0.026</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>-2.454</td>
<td>17.967</td>
<td>0.209</td>
<td>6.081</td>
<td>17.8</td>
<td>0.026</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>-2.452</td>
<td>17.967</td>
<td>0.143</td>
<td>6.081</td>
<td>17.8</td>
<td>0.026</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>-2.449</td>
<td>17.967</td>
<td>0.077</td>
<td>6.081</td>
<td>17.8</td>
<td>0.026</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>-2.133</td>
<td>17.967</td>
<td>0.011</td>
<td>6.308</td>
<td>15.783</td>
<td>0.011</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>-0.863</td>
<td>15.921</td>
<td>0.275</td>
<td>6.161</td>
<td>15.783</td>
<td>0.022</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>-0.86</td>
<td>15.921</td>
<td>0.209</td>
<td>6.161</td>
<td>15.783</td>
<td>0.022</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>-0.858</td>
<td>15.921</td>
<td>0.143</td>
<td>6.161</td>
<td>15.783</td>
<td>0.022</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>-0.855</td>
<td>15.921</td>
<td>0.077</td>
<td>6.161</td>
<td>15.783</td>
<td>0.022</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>-0.633</td>
<td>15.921</td>
<td>0.011</td>
<td>6.308</td>
<td>15.783</td>
<td>0.011</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>0.711</td>
<td>13.875</td>
<td>0.275</td>
<td>6.199</td>
<td>13.833</td>
<td>0.026</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>0.713</td>
<td>13.875</td>
<td>0.209</td>
<td>6.199</td>
<td>13.833</td>
<td>0.026</td>
</tr>
<tr>
<td></td>
<td>13</td>
<td>0.716</td>
<td>13.875</td>
<td>0.143</td>
<td>6.199</td>
<td>13.833</td>
<td>0.026</td>
</tr>
<tr>
<td></td>
<td>14</td>
<td>0.718</td>
<td>13.875</td>
<td>0.077</td>
<td>6.199</td>
<td>13.833</td>
<td>0.026</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>2.419</td>
<td>11.829</td>
<td>0.275</td>
<td>6.309</td>
<td>11.117</td>
<td>0.069</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>2.421</td>
<td>11.829</td>
<td>0.209</td>
<td>6.309</td>
<td>11.117</td>
<td>0.069</td>
</tr>
<tr>
<td></td>
<td>17</td>
<td>2.424</td>
<td>11.829</td>
<td>0.143</td>
<td>6.309</td>
<td>11.117</td>
<td>0.069</td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>2.426</td>
<td>11.829</td>
<td>0.077</td>
<td>6.309</td>
<td>11.117</td>
<td>0.069</td>
</tr>
<tr>
<td></td>
<td>19</td>
<td>5.183</td>
<td>9.783</td>
<td>0.275</td>
<td>7.144</td>
<td>9.783</td>
<td>0.275</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Shift</th>
<th>Solution</th>
<th>( f_1(x) - \epsilon \cdot \sum_{k=2}^{p} \frac{s_k}{s_k} )</th>
<th>( e_2 )</th>
<th>( e_3 )</th>
<th>( f_1(x) )</th>
<th>( f_2(x) )</th>
<th>( f_3(x) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1</td>
<td>-2.105</td>
<td>19.000</td>
<td>0.285</td>
<td>6.319</td>
<td>17.8</td>
<td>0.062</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>-2.102</td>
<td>19.000</td>
<td>0.223</td>
<td>6.319</td>
<td>17.8</td>
<td>0.062</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>-2.100</td>
<td>19.000</td>
<td>0.162</td>
<td>6.319</td>
<td>17.8</td>
<td>0.062</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>-2.097</td>
<td>19.000</td>
<td>0.101</td>
<td>6.319</td>
<td>17.8</td>
<td>0.062</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>-1.502</td>
<td>19.000</td>
<td>0.039</td>
<td>6.69</td>
<td>16.833</td>
<td>0.039</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>-0.387</td>
<td>16.783</td>
<td>0.285</td>
<td>6.462</td>
<td>16.05</td>
<td>0.057</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>-0.385</td>
<td>16.783</td>
<td>0.223</td>
<td>6.462</td>
<td>16.05</td>
<td>0.057</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>-0.382</td>
<td>16.783</td>
<td>0.162</td>
<td>6.462</td>
<td>16.05</td>
<td>0.057</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>-0.380</td>
<td>16.783</td>
<td>0.101</td>
<td>6.462</td>
<td>16.05</td>
<td>0.057</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>1.413</td>
<td>14.567</td>
<td>0.285</td>
<td>6.65</td>
<td>14.55</td>
<td>0.075</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>1.415</td>
<td>14.567</td>
<td>0.223</td>
<td>6.65</td>
<td>14.55</td>
<td>0.075</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>1.418</td>
<td>14.567</td>
<td>0.162</td>
<td>6.65</td>
<td>14.55</td>
<td>0.075</td>
</tr>
<tr>
<td></td>
<td>13</td>
<td>1.420</td>
<td>14.567</td>
<td>0.101</td>
<td>6.65</td>
<td>14.55</td>
<td>0.075</td>
</tr>
<tr>
<td></td>
<td>14</td>
<td>3.485</td>
<td>12.35</td>
<td>0.285</td>
<td>7.017</td>
<td>12.033</td>
<td>0.222</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>3.487</td>
<td>12.35</td>
<td>0.223</td>
<td>7.017</td>
<td>12.033</td>
<td>0.222</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>5.199</td>
<td>10.133</td>
<td>0.285</td>
<td>7.146</td>
<td>10.133</td>
<td>0.285</td>
</tr>
</tbody>
</table>
Figure 10. Pareto frontier shift 1

Figure 11. Pareto frontier shift 2

Figure 12. Efficient ambulance location in shift 1

Figure 13. Efficient ambulance location in shift 2
number of records, we generated emergency clusters to reduce computational complexity. The clustering method helps to reduce the computational time, as it is mentioned by Kariv and Hakimi (1979), this type of problem is NP hard, so if it increases the size of the problem will be more complicated to solve it and time-consuming to build the time matrix. The criteria used to decide the number of clusters were the solution time on optimization, the accuracy of objective values for each objective when cluster numbers increase and, the maximum distance between the center of each cluster and its emergencies.

We solved the multi-objective problem using the ε-constraint method, obtaining efficient solutions that improve the current state. In our implementation, we solved the problem exactly, without the use of any heuristic, meaning all points in the efficient frontier are provably Pareto optimal solutions. The results show that the current ambulance locations are not optimal and that changing them can significantly improve EMS key performance indicators, and thus, survival outcome. The Pareto frontier shows the relationship between the objectives, showing that mean response time and coverage follow the same direction. On the contrary, the maximum time increases when others decrease. We suggested a starting point to Antofagasta EMS with a pilot plan that would ease the implementation of a new location policy and would also allow assessing the effect of such relocation to ultimately motivate a full deployment of the proposed policy.

A limitation in our work comes from the fact that our model assumes the potential bases set to be equal that the set of emergencies. Thus, our model could produce solutions with locations where EMS may not position an ambulance for lengthy periods. Before implementing the results of this work in a real setting, the real potential bases have to be defined. EMS workers need to be involved in this definition since it significantly changes their work conditions. Moreover, a communication strategy should be made to announce the new policy in order to keep the population informed.

Future research must focus on better defining potential ambulance locations and improving the estimation of ambulance travel time with the use of GPS data. Moreover, significant effort must be allocated to ensure that emergency location is recorded accurately for every case. Record keeping is human dependent, and in many cases, as it was our case as well, errors in logs force analysts to discard significant numbers of records in detriment of the correctness of any policy decision.

Acknowledgements

We want to thank Antofagasta EMS because they provided us the data to work in this study and its willingness to cooperate us in the research.

Powered by NLHPC: this research was partially supported by the supercomputing infrastructure of the National Laboratory for High Performance Computing (NLHPC) – ECM-02.

This work was partially funded by the Chilean National Agency for Research and Development (Agencia Nacional de Investigación y Desarrollo – ANID) FONDEF ID19110257.

Author contributions

Carlos Olivos and Hernán Caceres worked in the study, and both were responsible for cleaning the data provided by Antofagasta EMSs, make the codes for the optimization process, analyse the results of the experiments and write this paper.

Disclosure statement

Authors declare to have no competing financial, professional or personal interests from other parties.

References


Pons, P. T.; Markovchick, V. J. 2002. Eight minutes or less: does the ambulance response time guideline impact trauma pa-


