A hybrid optimization method for reallocation of mobile resources

J Catumba-Ruiz¹, L Aguiar², J M Redondo³, R Renteria⁴, and J O Barrera²

¹ Universidad Nacional de Colombia, Manizales, Colombia
² Pontificia Universidad Javeriana, Bogotá, Colombia
³ Universidad Católica de Colombia, Bogotá, Colombia
⁴ Universidad Nacional Abierta y a Distancia, Bogotá, Colombia

E-mail: jcatumbar@unal.edu.co

Abstract. In this paper, we propose a hybrid optimization method to compute a reallocation of ambulances to obtain improved response times. As we want to minimize response times by changing ambulances allocations, we develop a hybrid algorithm based on a genetic algorithm, with randomized ambulances configurations as population individuals. Also, we embed into the genetic algorithm discrete event simulations to model the reporting, assignment, travel, and attendance processes. We later find that the algorithm optimizes the response times for simulated events, even though these times don’t yet compare to response times found in real data. So, we need to evaluate any improvement in real response times. As a study case, we use data from 2014 to 2017 provided by the health authority of Bogotá, Colombia, that contains real values of emergency medical incidents, and the quantity and type of ambulances that attended such incidents.

1. Introduction
Response time is a key performance indicator (KPI) in every emergency service, especially for emergency medical systems (EMS) [1–5], because their mission is to preserve human life. In real applications, it is not cost-effective to put an ambulance in every corner in a city; so, there is considerable research on the resource allocation problem. For example, we can find many attempts across several areas of interest that include disaster recovery planning [6] and operational research [7–10].

However, each resource allocation study case is different from the others. One may focus simply on one of the following goals: availability of resources, area of coverage, demand, of resources, travel time if the resources are mobile, and many more. Also, one may focus on more than one goal, and that changes the approach needed to solve the problem. Therefore, there are many perspectives and methods to address this problem in the case of medical emergency resources or ambulances, including discrete event simulation (DES) [11–13], genetic algorithms (GA) [14,15], reinforcement learning [16], and even social networks [17].

For this research, we design a hybrid method to focus on four goals been the optimization of response time, the effective assignment of resources, the satisfaction of the demand of ambulances, and usage of vehicles.
2. Materials and methods
Bogotá, Colombia, EMS receives emergency medical incidents through the “1-2-3” emergency number. The CAD system records the entire reception-dispatch process data, so it is possible to get the entire log of the attention given by the EMS to each one of the incidents. Given such data, we compute the most frequent incident types, frequencies of incidents per hour, and response times. We describe data structure and computations in subsection 2.1. Using outputs from this data analysis as parameters, we construct a DES that models the assignment, dispatch, and attention process of the EMS and its ambulances. In subsection 2.2, we describe this DES in detail. To optimize ambulance allocations and response times, we enclose several DES into one GA. This GA takes into account randomized ambulance configurations (AC) as population individuals for each GA generation and outputs the best fitting AC after several iterations. We describe the GA in detail in subsection 2.3.

2.1. Structure and analysis of emergency medical system data
Bogotá, Colombia, EMS provided data in two different file templates for each one of the four years: incidents and ambulance statuses. For incidents, we only consider the top 15 types that represent 91.29% of the reported cases from 2014 to 2018. These codes are institutional of the EMS and don’t correspond to any international standard.

With the incident number identifier we match data from both files. Using timestamps we compute assignment, travel and attention times, in seconds, for each incident. Assignment time \( t_{as} \) is computed from incident creation timestamp to log time of ambulance “assigned” status code. Travel time \( t_{tr} \) goes from log time of ambulance “assigned” status code to log time of ambulance “arrived” status code. Finally, attention time \( t_{at} \) goes from log time of ambulance “arrived” status code to incident closure timestamp. We define response time \( t_{re} \) as the sum of assignment time and travel time, so \( t_{re} = t_{as} + t_{tr} \). We focus on \( t_{re} \) optimization because is what users of the EMS perceive when they call “1-2-3”. Moreover, \( t_{at} \) is dependant on a handful of other uncontrolled and unmeasured variables.

Notice that we only use the data of incidents that received an ambulance assignment and were closed as “attended”; we took this decision because many of them are “duplicated” to other real ones. Also, some of them are “false reports” and alike. From given data, we generate a table of the necessary information, and each row has the incident identifier, incident type, incident priority, incident coordinates, creation timestamp, assignment timestamp, attention timestamp, arrival timestamp, and closing timestamp; using incident creation timestamps, we compute the average and standard deviation of incidents’ reporting by type and priority at each hour of the year. For simplicity, we consider only two types of ambulances: medical ambulances (MA) and basic ambulances (BA). An MA differs from a BA because it has a medical doctor among its crew; a MA must attend high-priority incidents, and a BA must attend medium-priority and low-priority incidents.

2.2. Discrete event simulation
The purpose of using a DES is to simulate a simplified version of incident reporting, ambulance assignment, ambulance movement, and attention process. We provide the algorithm as an annex. To achieve incident reporting, we generate a list of incidents with corresponding incident type, timestamp of creation, priority, and coordinates. For the first three, we use computed parameters according to the provided data. Coordinates are randomly selected, for each incident, from historical data by filtering the corresponding incident type. For example, if we have “cardiac arrest” incidents to have a mean of 10 and a standard deviation of 4 at 4:00 pm of February 10th, we randomly generate several cardiac arrest incidents using a normal distribution with those parameters. After that, we use that number to generate the minutes of the hour in which incidents will occur.
The simulation clock runs by a minute and reads which incidents must happen in every simulation moment. We implement a prioritized assignment queue. High-priority incidents are always before all mid-priority ones and, in the same way, mid priority incidents are before all low priority ones. We differentiate this incident queue (IQ) from the DES event queue (EQ). Despite the incident priority, we register the incident at the EQ with a “reported” event. Every record in the EQ has an ambulance id (if assigned), event type, incident id, and a timestamp. The EQ and the simulation clock take over the ambulance dispatch, and control travel and attention times. After we add incidents to the IQ, we check the EQ. First, we filter events to occur at the simulation clock time. From that selection, we see which are “arrival” events and use computed parameters from historic attention time data to generate a random attention time. With this value, we add the “finish” event to the EQ. After that, we select “finish” events and free the associated ambulance for further assignment. When disassociated, ambulance coordinates are not reset.

The assignment process checks first if there are available ambulances. Then, according to the number of available ambulances, we dequeue pending incidents and assign emergency vehicles considering the Euclidean distance, type of incident, and type of ambulance. MAs are preferred to attend high-priority incidents, BAs are preferred to attend mid-priority incidents, and low-priority incidents remain without attention. But, if there is not an MA for a high-priority event, then a BA is assigned and conversely. Using the distance between ambulance location and incident coordinates, we compute the arrival time using the mean speed of vehicles in Bogotá, Colombia, and a low standard deviation. Finally, we register the “arrival” events on the EQ. By the end of the simulation, we have the log of the EQ. We use this to compute the fit of the attention to the desired values from the initial ambulance allocations. For this computation, we consider four criteria scaled from 0 (best) to 10 (worst).

First criterion \( f_{ar} \) is based on response times; we compute this value using Equation (1) where we define \( m_{ar} \) as the average time of response times. For this formula, we consider that a response time of 3600 seconds (1 hour) must be 10, and a response time of 480 seconds (8 minutes) must be 1.

\[
f_{ar} = \frac{3}{1040} m_{ar} - \frac{5}{13}.
\]

The second criterion \( f_{ef} \) is the effective assignment of ambulance type, or how many incidents received the right ambulance type according to its priority. We consider \( n \) as the number of reported incidents and \( w \) the number of incidents that received a not corresponding ambulance; to build Equation (2), we assumed that if all reported incidents received a not corresponding ambulance, the score must be 10, but, if only 10% of incidents received a not corresponding ambulance, the score must be 1.

\[
f_{ef} = \frac{10}{n} w.
\]

The third criterion \( f_{sa} \) is the satisfaction of attention, or how many incidents were left without an ambulance assignment. For Equation (3) we consider \( N \) as the number of high-priority and mid-priority reported incidents, and \( x \) the number of high-priority and mid-priority incidents that weren’t attended; also, we assume that if 20% of incidents didn’t receive an ambulance assignment, the score must be 10, and, if only 1% of incidents didn’t receive an ambulance assignment, the score must be 1.

\[
f_{sa} = \frac{900}{19} N + \frac{10}{19}.
\]

The fourth criterion \( f_{us} \) is ambulance usage, understood as a measure that indicates that all ambulances are used equally. In Equation (4), we consider \( M \) as the number of ambulances,
and $u$ as the mean of incidents attended by ambulance; if usage $u$ is $\frac{N}{M}$ the score is 1, and, if $u = 2 \frac{N}{M}$ the score is 10.

$$f_{us} = \frac{9M}{N} u - 8. \tag{4}$$

We compute the overall score for each DES simulation using Equation (5). Notice that we gave response time more importance.

$$f = \sqrt{0.4 * (f_{ar})^2 + 0.2 * (f_{ef})^2 + 0.2 * (f_{sa})^2 + 0.2 * (f_{us})^2}. \tag{5}$$

### 2.3. Genetic algorithm

The goal of the GA is to optimize allocations and the number of ambulances by type given a fixed number of ambulances to attend emergency medical incidents. Parameters for this algorithm are number of ambulances, initial and final timestamps for DES, time step for DES, number of AC by population (number of individuals by population or generation), number of iterations (number of generations), number of AC to use as parents, coordinates and type of historical incidents, and a random set of possible ambulance allocations (universe of allocations). We provide the algorithm as an annex.

We fix the number of ambulances, but we may also let it variate to optimize it. On the other hand, we require initial and final times, and time step, so all DES simulations are time comparable. One of Bogotá’s political divisions is the zonal planning units (UPZ). These are used by the public administration for localized management, because of their similar socio-economic features. So, we randomly generate one hundred points inside each one of the 117 UPZ polygons to be the universe of possible ambulance allocations. Before the GA started, we generated the simulation incidents list and the first generation of ambulances AC. Each AC is composed of a fixed amount of triplets ($X$, $Y$, ambulance type) for each ambulance, where $X$ and $Y$ are the coordinates of its initial position and “ambulance type” is MA or BA. We randomly selected the coordinates from the universe of allocations, and the type of the ambulance.

Once we finish all DES, we compute their scores from the event log and show the mean score. We saved the individual with the best score, and use the top best to create the new generation. This action describes the selection operator. The combination operator chooses, from the previous selection, two items at random and then selects the 45% of triplets from each AC. The mutation operator selects the remaining 10% ambulance locations and types randomly from the universe of allocations to complete the number of AC triplets. The optimization function is the one given at Equation (5). We want the lesser possible value of $f$ among ACs.

### 3. Results and discussion

We execute the algorithm with provided data and simulate the most critical week of the year in terms of the number of EMS incidents, which is the last week of the year because of Christmas and New Year. We get that the median simulation response time was 1140s (19 minutes), which is still far from the recommended time of 8 minutes. If we look at the box-plots of real times and simulated times by incident type, Figure 1 and Figure 2, we can see that generally, the real data response times are less dispersed than simulation ones.

Even though there are improvements for certain incident types, the overall performance of the method doesn’t improve real response times. Yet, we find allocations that optimize those simulated response times. As we generate the points at random, there might be allocations in the middle of a park or a lake, so for real implementation, we must consider the closest viable allocation. On the other hand, as we take into account ambulance types, we find that the given solution has a 3:1 proportion for MAs and BAs. We provide output files for optimal solutions and times as annexes.
Concerning the ambulance usage criterion, we find that the median of incidents attended by ambulance is 24, but the minimum is 1, and the maximum is 37, Figure 3. For the demand criterion, we get that 2215 out of 2767 high and mid priority incidents received attention. And finally, for the effective assignment criterion, we get that 1913 out of 2215 attended incidents received the corresponding ambulance type according to the priority. All of these criteria show that the hybrid method complies with the proposed requirements. We must consider that by the end of the DES there are still incidents left to attend because reporting didn’t stop before.
Also, as the resulting files show, there are 36 busy ambulances. Therefore, criteria measures could be even better if we left the simulation run for more time without queueing new incidents beyond the simulation ending time.

Now, considering the execution time of the hybrid method, it takes five days to run one hundred iterations with twenty individuals by generation, in a dedicated computing cluster. We also must consider the fact that each execution of the GA algorithm runs all DES in parallel. Therefore, if we want to simulate an entire year, it will take a considerable amount of time with the same input parameters. So, this is a decisive factor for this algorithm to be deemed inefficient.

4. Conclusion
The proposed hybrid algorithm manages to optimize the four given criteria, even though it is computationally inefficient. We can work in further optimization to achieve a real-time goal so we can deploy it to a real EMS for allocation decision making. We can use the proposed DES as a standalone algorithm, and we can combine it with other optimization paradigms other than GA. Moreover, in the proposed method, the CPU is working on the DES simulations during most of the computational time. If we design an alternative or modification to avoid the amount of DES that we simulate, we could achieve better computational time. Moreover, we can modify the algorithm to optimize the number of ambulances because the respective optimization criterion is closely related to the other criteria shown in this paper. For that, we must compute this new criterion among different DES to compare behavior in different ACs. On the other hand, if the considered problem has an optimal substructure, we can use a different optimization paradigm.

Furthermore, we selected the universe of allocations at random, but we could use a geostatistical-based criterion for better distribution, and therefore better response times. For example, we didn’t consider incident hot-spots as a preferential selection of allocations. We also didn’t take into account that the statistical tendency might variate by hours or weekdays. Moreover, when we generated the ACs we didn’t consider restricting the selection of points by UPZ, so in the provided solution, there might be UPZs without associated ambulance or with more than one.
References

[1] Knight V A, Harper P R, Smith L 2012 Ambulance allocation for maximal survival with heterogeneous outcome measures *Omega* **40**(6) 918–926

[2] Leknes H, Aartun E S, Andersson H, Christiansen M, Granberg T A 2017 Strategic ambulance location for heterogeneous regions *European Journal of Operational Research* **260**(1) 122–133

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

[4] Erkut E, Ingolfsson A, Erdogan G 2008 Ambulance location for maximum survival *Naval Research Logistics (NRL)* **55** 42–58

[5] Zaffar M A, Rajagopalan H K, Saydam C, Mayorga M, Sharer E 2016 Coverage, survivability or response time: A comparative study of performance statistics used in ambulance location models via simulation–optimization *Operations Research for Health Care* **11** 1–12

[6] Doan X V, Shaw D 2019 Resource allocation when planning for simultaneous disasters *European Journal of Operational Research* **274**(2) 687–709

[7] Bertsimas D, Gupta S, Lulli G 2014 Dynamic resource allocation: A flexible and tractable modeling framework *European Journal of Operational Research* **236**(1) 14–26

[8] Bertsimas D, Griffith J D, Gupta V, Kochenderfer M J, Mišić V V 2017 A comparison of monte carlo tree search and rolling horizon optimization for large-scale dynamic resource allocation problems *European Journal of Operational Research* **263**(2) 664–678

[9] Luscombe R, Kozan E 2016 Dynamic resource allocation to improve emergency department efficiency in real time *European Journal of Operational Research* **255**(2) 593–603

[10] Nesterov Y, Shikhman V 2018 Dual subgradient method with averaging for optimal resource allocation *European Journal of Operational Research* **270**(3) 907–916

[11] Robinson S 2014 *Simulation: The Practice of Model Development and Use* (London: Palgrave Macmillan)

[12] Bohez S, Verbelen T, Simoens P, Dhoedt B 2015 Discrete-event simulation for efficient and stable resource allocation in collaborative mobile cloudlets *Simulation Modelling Practice and Theory* **50** 109–129

[13] Fu X, Presbitero A, Kovalechuk S V, Krzhizhanovskaya V V 2018 Coupling game theory and discrete-event simulation for model-based ambulance dispatching *Proceeds Computer Science* **136** 398–407

[14] Deb K 1999 An introduction to genetic algorithms *Sadhana* **24**(4) 293–315

[15] Feng W H, Lou Z, Kong N, Wan H 2017 A multiobjective stochastic genetic algorithm for the pareto-optimal prioritization scheme design of real-time healthcare resource allocation *Operations Research for Health Care* **15** 32–42

[16] Huang Z, van der Aalst W M P, Lu X, Duan H 2011 Reinforcement learning based resource allocation in business process management *Data & Knowledge Engineering* **70**(1) 127–145

[17] Wang D, Qi C, Wang H 2014 Improving emergency response collaboration and resource allocation by task network mapping and analysis *Safety Science* **70** 9–18