Proceedings of the 2018 Winter Simulation Conference
M. Rabe, A. A. Juan, N. Mustafee, A. Skoogh, S. Jain, and B. Johansson, eds.

A SIMULATION AND ONLINE OPTIMIZATION APPROACH FOR THE REAL-TIME MANAGEMENT OF AMBULANCES

Roberto Aringhieri
Simone Bocca
Luigi Casciaro
Davide Duma

Computer Science Department
University of Turin
Corso Svizzera 185
Turin, 10149, ITALY

ABSTRACT

Emergency Medical Service is one of the most important health care services as it plays a vital role in saving people's lives and reducing the rate of mortality and morbidity. A peak of emergency demand can determine overcrowding at the emergency department. In this operative context, a challenge is the definition of proper real-time dispatching, routing and redeployment policies (DRRP) in such a way to maximize the number of emergency requests served within a time threshold, and to minimize the waiting times. The contribution of this paper is twofold. The former is a simulation model capable to deal with the real-time management of the ambulances, and to generate new ad hoc instances. The latter is a set of simple online algorithms to implement several DRRP. An extensive comparison among different DRRP is also provided.

1 INTRODUCTION

Emergency Medical Service (EMS) is one of the most important health care services as it plays a vital role in saving people's lives and reducing the rate of mortality and co-morbidity. The importance and sensitivity of decision making in the EMS field have been recognized by researchers who studied many problems arising in the management of EMS systems since the 1960s, as reported by Aringhieri et al. (2017), Reuter-Oppermann et al. (2017) and Bélanger et al. (2018). Aboueljinane et al. (2013) present a review of the many simulation models that have been developed over the years: most of the available simulation approaches are based on a Discrete Event Simulation (DES) approach.

A peak of emergency demand can determine overcrowding (Paul et al. 2010) at the Emergency Department (ED). Overcrowding is manifested through an excessive number of patients in the ED, long waiting times and patients leaving without being seen. Sometimes patients being treated in hallways and ambulances are diverted (Hwang and Concato 2004). Consequently, the ED overcrowding has a harmful impact on the health care (George and Evridiki 2015).

Simulation is often exploited for the analysis of the overcrowding at the ED and its impact on ambulance diversion. Nafarrate et al. (2010) found that the number of patients in the waiting room is a better trigger for ambulance diversion than inpatient bed availability, as it provides the best balance between accessibility and waiting times. Ramirez-Nafarrate et al. (2012) use a simulation model to determine the effect of several ambulance diversion policies on the patient's waiting time. More generally, van Buuren et al. (2012) uses simulation to evaluate several dynamic dispatching strategies while Maxwell et al. (2009) evaluates redeployment policies computed by approximate dynamic programming using simulation.

A first attempt to consider real-time redeployment policies is due to Ni et al. (2012). In their analysis, the authors use simulation to devise and to evaluate redeployment policies. Jagtenberg et al. (2017) provide

978-1-5386-6572-5/18/$31.00 ©2018 IEEE

a bound existing online solutions in comparison with three different methods to compute the optimal offline dispatching policy for problems with a finite number of incidents. The performance of the offline optimal solution serves as a bound for the performance of an unknown optimal online dispatching policy. Then, they compare such an offline solution to the closest idle vehicle dispatching policy obtaining a bound of 2.7 times on the fraction of late arrivals. Aringhieri et al. (2018) analyzes the interplay between the EMS and the network of EDs operating on a given area at the system level. The analysis of a simple EMS dispatching policy, based on the real-time workload of the EDs, showed that there is room to improve the efficiency of the ED network reducing the patient waiting time. Further, such an improvement is more significant as soon as the percentage of the patients transported by the EMS increases. van Barneveld et al. (2018) evaluate the impact of typical factors influencing the performance of an EMS such as (i) the frequency of redeployment actions, (ii) time bounds on the ambulance relocation, and (iii) the inclusion of busy ambulances in the decision process. The main insights derived by their research are that adding more relocation action is highly beneficial for rural areas and considering ambulances involved in dropping off patients available for newly coming incidents reduces relocation times only slightly. Finally, Nasrollahzadeh et al. (2018) develop a flexible optimization framework for real-time ambulance dispatching and relocation. They formulate the problem as a stochastic dynamic program. Because of the unbounded state space, the authors propose an approximate dynamic programming framework to generate high-quality solutions. Their analysis is performed on an available benchmarks on an EMS system in Mecklenburg County, North Carolina.

The analysis of the literature reveals an attention to real-time dispatching policies in contrast to a limited attention to the redeployment of the ambulances in real time. From such an analysis, a list of insights can be derived, that is (i) the number of patients in the waiting room is a better trigger for ambulance diversion (Nafarrate et al. 2010), (ii) the use of the fraction of covered calls as efficiency measures (McLay and Mayorga 2013), (iii) incorporating equity might lead to a lower service or negative outcomes (McLay and Mayorga 2013), (iv) priority dispatching policies can improve the performance for urgent calls at the price of a worsening of the performance for non-urgent calls (Bandara et al. 2014).

The real-time management of the ambulances of an EMS is an online optimization problem in which three main decisions should be addressed to serve an emergency request, that is (1) which ambulance should be dispatched to serve an emergency request, (2) the selection of the ED facility to which the patient will be transported, and (3) where to redeploy the ambulance at the end of its service. A challenge is the definition of proper DRRP in such a way to maximize the number of emergency requests served within a time threshold, and to minimize the waiting times. To the best of our knowledge, a comprehensive analysis of the DRRP is missing. The contribution of this paper is twofold. The former is an ambulance simulation model capable to deal with real-time management and to generate new ad hoc and realistic instances. The latter is a set of simple online algorithms to implement several DRRP. An extensive comparison among different DRRP is also provided.

The paper is organized as follows. The instance generator and the simulation model are described in Sections 2 and 3, respectively. The DRRP are introduced in Section 4. The quantitative analysis is reported in Section 5. Conclusions are discussed in Section 6.

2 INSTANCE GENERATOR

An instance is a planar graph $G = (N,E)$ with n nodes and m arcs. Each node is a centroid representing a small part of the whole area served by the EMS. Each arc models the connection between two nodes. Two labels are associated to each arc: the former represents the length of the arc while the latter is the average speed of a vehicle traveling on it. The number of arcs starting from a node $u \in N$ is equal to a_u. An example is reported in Figure 1.

The length of an arc and, more generally, distances in the graph are Euclidean. Further, a scale factor f_s determines the value of each pixel. The scale factor is useful to generate graph representing different type of areas such as urban or rural: for instance, in our settings for a urban area 1 pixel corresponds to 20 meters.

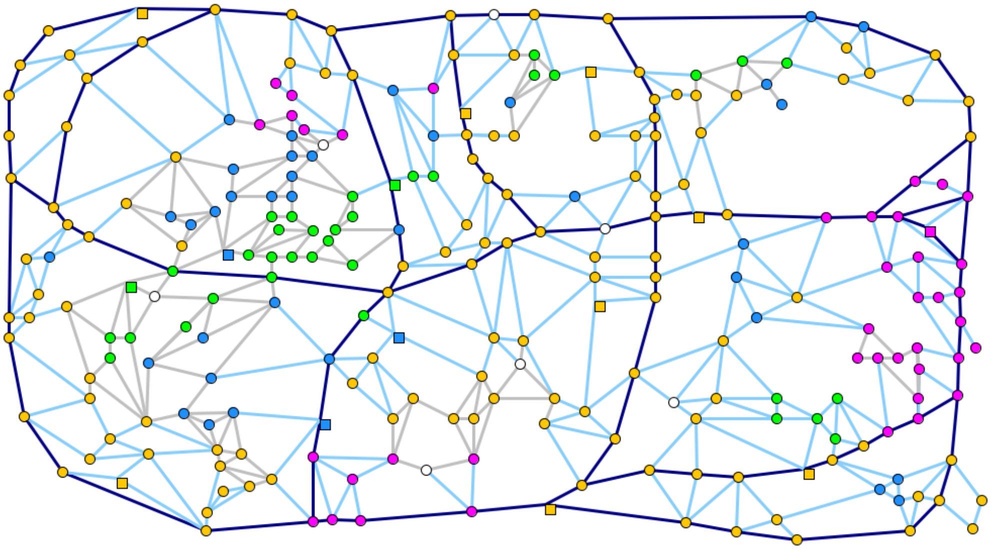

Figure 1: An example of realistic graph in its final version.

As highlighted in Figure 1, there are three type of nodes, that is the emergency demand generator (the colored circle), the ambulance base (the colored square), and the ED facility (the white circle). Note that an emergency request can be generated from an ambulance base node. Globally, we have n_G, n_B and n_{ED} nodes (respectively generators, bases, and ED facilities) such that $n_G + n_B + n_{ED} = n$. Let N_G, N_B, $N_{ED} \subset N$ be respectively the subsets of the $n_G + n_B$ generator nodes, the n_B bases, and the the n_{ED} ED facilities.

A generated graph can be manually adjusted adding or deleting nodes and arcs, and also moving nodes and, by consequence, all connected arcs. For urban area, it is also possible to characterize each node as residential, commercial, public utility, and offices. This classification is useful at running time to model in a proper way the generation of the emergency demand: for instance, a residential node usually generates more demand during the evening or night while a public utility node is likely to generate more demand in the morning. Further, we can change the average speed of each arc by default set to *medium* (30 km/h) to *low* (20 km/h) or *high* (40 km/h).

Figure 1 depicts the final version of an instance in which the yellows are the residential nodes, the greens are the commercial nodes, the light blues are the public utilities, and the purple ones are the areas with offices. Regarding the arcs, the light blue are those with medium speed while the blue and the gray arcs are the faster and the slower ones, respectively. At the end of the process, the graph can be saved on a file.

3 THE SIMULATION MODEL

The simulation model replicates how an EMS serves an emergency request. The EMS receives a phone call from a citizen asking for an emergency care for himself or for a third person. The operators at the EMS's operation center are in charge of answering the calls and assigning a color code to each emergency request, based on the severity of injury, through a phase called *triage*. After the triage phase the operator dispatches an ambulance following a predefined dispatching policy. Ambulance crew rescues the patient and, if necessary, transports him/her to a hospital. Note that usually the ambulance crew is in charge of the patient until he/she is handed to the hospital staff.

The model takes in input seven parameters, that is (i) the duration I of the simulation (expressed in days), (ii) the graph $G = (N, E)$, (iii) the number A_b of ambulances for each base $b \in N_B$, (iv) the maximum number A_b^{max} of ambulances that can be positioned on each base $b \in N_B$, (v) the emergency demand variation table, (vi) the workload of the ambulances, and (vii) the capacity of the ED facilities. While the first four parameters have a straightforward definition, the last three requires a detailed description, which is reported in the following. Let us denote with A the total number of ambulances, given by $A = \sum_{b \in N_B} A_b$.

3.1 Distances and Travel Times

The graph $G = (N, E)$ is an undirected and labeled graph. The labels on the arcs $[u, v]$ are the distance ℓ_d among u and v, and the average speed ℓ_s on that arc. We use such labels to compute both distances and/or travel times among two nodes in G. To this end, we use an ad hoc version of the classic label-setting shortest-path algorithm (e.g., Dijkstra).

3.2 The Emergency Demand Variation Table

As reported by many authors (see, e.g., Channouf et al. (2007), Setzler et al. (2009)), emergency demand is not static, but, rather, fluctuates during the week, according to the day of the week, and hour by hour within a given day. The emergency demand table (Table 1) would model the relative demand fluctuation over the day with respect to different urban areas over the total demand (e.g. office nodes should have a higher relative demand during the business hours of the day). In accordance with the characteristics of the generator node, a negative (*low*) or positive (*high*) variation of the predefined (normal) generation rate is possible. We denote as w_u^i the generic entry of the table with respect to the time interval $i = 1, 2, 3, 4$ (morning, afternoon, evening, night) and the node u.

Table 1: The Emergency Demand Variation Table.

	1: $[7-13]$	**2:** $[13-19]$	**3:** $[19-1]$	**4:** $[1-7]$		**1:** $[7-13]$	**2:** $[13-19]$	**3:** $[19-1]$	**4:** $[1-7]$
residential	normal	normal	high	low	**public utility**	high	high	normal	low
commercial	normal	high	low	low	**offices**	normal	normal	low	low
	low = 0.7				normal = 1.0		high = 1.3		

3.3 The Workload of the Ambulances

The workload of the ambulances W^A is clearly determined by the generation rate of each node in accordance with their characteristics and the fluctuations over the day reported in Table 1. Our model allows to input the number of total emergency requests that should be generated for each time interval along the day: the total number of emergency requests is denoted by D which is equal to $D_1 + D_2 + D_3 + D_4$ corresponding to the number of requests to be generated during the morning, afternoon, evening, and night time intervals, respectively.

During each time interval, the D_i requests are spread to all the nodes belonging in N_G as follows: for each node $u \in N_G$, let d_u^i be the number of nodes that should be generated by u during the time interval i, and defined as

$$d_u^i = \frac{w_u^i D_i}{\sum_{v \in N_G} w_v^i}.$$

Then, the generation rate of the node u is equal to d_u^i divided by the duration of the time interval i.

Alternatively, the workload W^A can be fixed as a target percentage of utilization and determining – by consequence – the corresponding values of D_i, $i = 1, 2, 3, 4$. First, for each node we compute the *minimal mission time* as the minimum time required to serve an emergency request on a given node. Let $t_u^{\min}$ be such a time computed by considering the shortest time needed to follow the path starting at the closest (to node u) ambulance base, passing from u, and ending at closest ED facility (to node u), plus the average service time required at the emergency scene and for releasing the patient at the ED. We note that the assumption under the computation of $t_u^{\min}$ is to have an ambulance always available. After that, we compute the (arithmetic) average $T_u^{\min}$ over all $u \in N_G$. The total number of emergency requests D is then computed as

$$D = A\frac{T(i)}{T_u^{\min}},$$

where $T(i)$ is the duration of the interval $i = 1,2,3,4$. Finally, the value D_i are obtained from D as $D_i = D\frac{r_i}{\sum_i r_i}$ where $r_i = 1, 0.8, 0.5, 0.2$. The basic idea is to spread the daily demand over the time interval in such a way to have a peak in the morning.

Independently of its generation, the urgency code of each request is distributed in accordance with the proportion observed in Aringhieri et al. (2016), that is about the 10% of red codes, 50% of yellow codes, and 40% of green codes. Note that these percentages are due to the so called *over triage*, which is an over estimation of the request urgency, made by the operators answering at the emergency request phone call.

3.4 The Capacity of the ED Facilities

The capacity of each ED facility located in $u \in N_{ED}$ is derived from the total demand D plus the number of patients D' that will arrive at the ED by their own. First we compute the average service time T_S^{ED} from an estimate of the service time of each urgency codes. The minimum necessary hourly capacity of the ED located at the node $u \in N_{ED}$ is given by

$$C_u = \frac{(D+D')T_S^{ED}}{24 n_{ED}}.$$

The main assumption underlying this computation is to have patients evenly distributed among the ED facilities in such a way to have always one patient to exploit a unit of ED capacity as soon as it is released by another patient. Finally, the capacity of the ED facilities is a parameter ranging in $[1,2]$ in such a way that the final capacity varies in $[C_u, 2C_u]$.

In our model we have therefore two sources of patients, that is those transported by the EMS and those arrived by their own. In our setting, $D' = 4D$, that is the number of patients transported by the EMS is about the 20% of the whole emergency demand at the ED facilities. This setting is consistent with the analysis in Aringhieri et al. (2018). The patients arriving by their own follow the same distribution observed in our previous work Duma and Aringhieri (2017), that is 2% of red codes, 15% of yellow codes, and 83% of green codes. To be consistent with such a distribution, the urgency code of the patients transported by the EMS are changed in such a way to obtain the same above distribution decreasing a fraction of the red code to yellow, and the yellow to green. Note that this is consistent with the common practice of an EMS in which over triage determines an overestimates of the emergency demand.

Table 2: Distributions used in the simulation model (Exp=exponential, Tr=triangular).

	Distribution	Parameters	Unit of measure
Ambulance request	$\text{Exp}\left(\frac{1}{\lambda}\right)$	$\lambda = \frac{1}{6D_i}$	patients/h
Autonomous arrivals to ED	$\text{Exp}\left(\frac{1}{\lambda'}\right)$	$\lambda' = \frac{1}{6D'_i}$	patients/h
Ambulance rescue duration	$\text{Tr}(\tau_{min}, \tau_{max}, \tau_{mod})$	$\tau_{min,max,mod} = 10, 20, 15$	min
Urg. patient release at ED	$\text{Tr}\left(\tau_{min}^{ry}, \tau_{max}^{ry}, \tau_{mod}^{ry}\right)$	$\tau_{min,max,mod} = 6, 10, 8$	min
Non-urg. patient release at ED	$\text{Tr}\left(\tau_{min}^{g}, \tau_{max}^{g}, \tau_{mod}^{g}\right)$	$\tau_{min,max,mod} = 6, 20, 13$	min

Probability distributions used to generate the total patient demand, the service times for the ambulance rescue and the release times (of a patient) are reported in Table 2. Further, the ED Length of Stay (EDLOS) is a discrete distribution used to generate the patient treatment duration within the ED is obtained empirically from the real case data-set of the ED studied in Duma and Aringhieri (2017), truncating times exceeding 25 h.

4 REAL-TIME POLICIES

Our main aim is to evaluate real-time policies for the management of the ambulances evaluating their impact in terms of performance of the ambulances and overcrowding of the ED facilities. In this perspective, we recall that the ambulance real-time management is an online optimization problem in which the following three main decisions should be addressed: (1) which ambulance should be dispatched to serve an emergency request, (2) the selection of the ED facility to which the patient will be transported, and (3) where to redeploy the ambulance at the end of its service.

Before introducing the DRRP, we define an estimate of the number of ambulances needed at each base $b \in N_B$. Let $N^c_1, \dots, N^c_{n_B}$ a partition of N in such a way that each node u belongs to N^c_b (with $b = 1, \dots, n_B$) if and only if the j-th base is the closest one to the node u. Let W_b the sum of the morning weights in Table 1 of the nodes in N_b, that is $W_b = \sum_{u \in N_b} w^1_u$. We use morning weights since we supposed to have a peak of demand in the morning. The number A^e_b of estimated ambulance of base $b \in N_B$ is finally given by

$$A^e_b = A \frac{W_b}{\sum_{u \in N_G \cup N_B} W_u}.$$

4.1 Ambulance Dispatching

The most common dispatching policy is that of assigning an ambulance available at the closest base (Cuninghame-Greene and Harries 1988), which has been proven to perform, on average, uniformly better than the other dispatching rules in accordance with (Larsen et al. 2002). In the following, we refer to this policy as **D-Closest**.

Alternatively, the dispatched ambulance can be selected from a list of *enough close* bases, that is those capable to reach the request within the time threshold for the urgency of that request. Let L_B be such a list of enough close bases. The **D-LLCB** policy selects the ambulance to be dispatched from the base

$$\text{argmax}_{b \in L_B : A^a_b > 0} \{A^a_b - A^e_b\}$$

where A^a_b is the number of ambulances available in b at the moment of the decision. The **D-LLCB** policy is similar to those reported in Bandara et al. (2014), Haghani et al. (2004).

The cutoff priority queue (**D-CPQ**) and the smart assignment (**D-SA**) are two possible extensions of the two policies above reported. The **D-CPQ** consists in temporarily stop serving all the emergency requests having green urgency code when the overall number of available ambulance is less than a given threshold. The rationale here is to free up potential resources to deal with the ongoing peak of emergency demand. The **D-SA** consists in considering for dispatching not only the ambulances available in a base but even those who are in the redeployment phase, that is moving from an ED to an ambulance base. In a real setting, this means to have a sort of tracking system that allows us to follow the ambulance in real time.

When **D-SA** is active, the **D-LLCB** should be slightly modified accordingly. First, we consider the list L_R of the destination bases of the redeploying ambulances that are capable to reach the request within the time threshold. This means to assign the value $A^a_b - A^e_b$ to each redeploying ambulance corresponding to the destination base. Then we select a base according to

$$\text{argmax}_{b \in L_B \cup L_R : A^a_b > 0} \{A^a_b - A^e_b\} :$$

if the selected base $b \in L_B$, we dispatch an ambulance from the base B; on the contrary, we dispatch the redeploying ambulance if $b \in L_R$. Finally, if the base $b \in L_R$ belongs also to L_B, we dispatch the closest ambulance between the redeploying ambulance and one of the those available in the base.

Both **D-CPQ** and **D-SA** are introduced and discussed by Aringhieri et al. (2016) while **D-CPQ** is also analyzed by Yoon and Albert (2017). To the best of our knowledge, **D-SA** is surprisingly never cited in the literature: the closest approach we retrieved is that reported in Lee (2014) in which the centrality-based dispatching policy is improved by taking into account both idle and busy ambulances.

4.2 Emergency Department Facility Selection

The **H-Closest** policy selects the closest ED facility. Again, this is a common choice in the real settings. The rationale here is to provide as soon as possible a more accurate medical treatment to the patient. Anyway, the ED managers usually complain about the fact that the workload is not evenly distributed among the ED facilities of a given area. Their claim is that a more fair distribution of the workload could improve the overall efficiency of the ED facility network. Such a claim seems proved by the analysis reported in Aringhieri et al. (2018).

Here, we tested two simple policies addressing the problem of reducing overcrowding at the ED facility in accordance with the remark discussed in Nafarrate et al. (2010), and reported at the beginning of the chapter. The first policy, say **H-SAQ** selects the ED facility having the shortest admission queue counting only those patients that have same or higher urgency code. The second policy, say **H-WLB** tries to balance the workload of the ED facility taking into account the time needed to treat all patients in the admission queue. Note that **H-WLB** implements a policy from the ED point of view while **H-SAQ** implements a policy from the patient point of view. Finally, note that the two policies are applied only to those patients having yellow or green urgency code, while the **H-Closest** is always applied to the red ones.

An estimate of the workload F_u of the ED facility $u \in N_{ED}$ is given by

$$F_u = t^r_w p^r_u + t^y_w p^y_u + t^g_w p^g_u$$

where t^r_w, t^y_w and t^g_w are respectively the average ED length of stay for a red, yellow and green code, while p^r_u, p^y_u and p^g_u are respectively the number of patients inside the ED facility (both waiting for admission and under treatment) for a red, yellow and green code.

To counterbalance the effect of the longer travel times in the case of ED facilities less overcrowded but far from the emergency request, the policies **H-SAQ** and **H-WLB** are applied only taking into account the ED facilities no farther than the *radius* of G, that is half of the longest travel time between a node $u \in N$ and the current ED facility.

4.3 Ambulance Redeployment

In the real practice, a simple policy is that of redeploying the ambulance to its original base. In the following we refer to this policy as **R-Base**. Alternatively, the aim of the EMS management should be to make an ambulance available as soon as possible redeploying it to the closest base. Therefore, the **R-Closest** policy redeploys the ambulance to the closest base at the end of the mission. This is one of the most used policy in the real settings.

A third version of the above two policies is the **R-LCBT** policy: it redeploys the ambulance to the less covered base b within a given time threshold T^R as follows

$$\operatorname{argmin}_{b \in L_R} \{A^a_b - A^e_b\}$$

where L_R is the list of the bases that can be reached from the current ED facility within T^R. Note that T^R is a parameter introduced to counterbalance the effect of the longer travel times as remarked in van Barneveld et al. (2018).

5 QUANTITATIVE ANALYSIS

In this section we provide an analysis of the proposed policies in order to evaluate their impact when are used separately or together.

In order to evaluate the proposed DRRP, we define in Table 3 several performance indices taking into account the ambulance utilization and the most important aspects for the patient safety and satisfaction, which regard the waiting time from the moment of the phone call to the arrival of the ambulance and the waiting time at the ED. Observe that red code patients do not queue at the ED, then their waiting times are omitted.

Table 3: Performance indices.

Index	Definition	Index	Definition
r	Average time to reach a request (min)	u^+	Ambulance utilization considering also redeployment (%)
f_g	Fraction of green code reached within 20 min (%)	u_{ED}	ED utilization (%)
f_{ry}	Fraction of red and yellow code reached within 8 min (%)	w_g	Average waiting time of green code at the ED (min)
u	Ambulance utilization considering only mission time (%)	w_y	Average waiting time of yellow code at the ED (min)

The set of configurations taken into account in our analysis is reported in Table 4 and defined as different combination of the DRRP. All the possible policy combinations have been tested for each scenario, but we report the most significant ones for reasons of synthesis. Given a certain scenario, we start from the baseline configuration (0), in which the basic policies are used, and we change one at a time the policies for ambulance dispatching (1), for ED facility selection (2–3) and ambulance redeployment (4–5). The same policies are defined enabling the D-SA option (0s–5s). Other configurations changing more than one policy w.r.t. the baseline (6 and 7s–9s) have been chosen because they highlight interesting aspects for the analysis. For the same reason, two configuration (6t and 0st) have been selected to study the impact of the option D-CPQ.

Table 4: Configurations of DRRP set for the analysis.

id	Dispatching	ED Selection	Redeployment	id	Dispatching	ED Selection	Redeployment
0	D-Closest	H-Closest	R-Base	2s	D-Closest, D-SA	H-SAQ	R-Base
1	D-LLCB	H-Closest	R-Base	3s	D-Closest, D-SA	H-WLP	R-Base
2	D-Closest	H-SAQ	R-Base	4s	D-Closest, D-SA	H-Closest	R-Closest
3	D-Closest	H-WLP	R-Base	5s	D-Closest, D-SA	H-Closest	R-LCBT ($T^R = 20$ min)
4	D-Closest	H-Closest	R-Closest	7s	D-Closest, D-SA	H-SAQ	R-LCBT ($T^R = 20$ min)
5	D-Closest	H-Closest	R-LCBT ($T^R = 20$ min)	8s	D-LLCB, D-SA	H-WLP	R-Base
6	D-Closest	H-SAQ	R-Closest	9s	D-LLCB, D-SA	H-WLP	R-LCBT ($T^R = 20$ min)
0s	D-Closest, D-SA	H-Closest	R-Base	6t	D-Closest, D-CPQ	H-SAQ	R-Closest
1s	D-LLCB, D-SA	H-Closest	R-Base	0st	D-Closest, D-SA, D-CPQ	H-Closest	R-Base

After fixing a policy configuration and a scenario through the model parameters, we execute 30 simulation runs over a time horizon of $I = 6$ days: the first day is used for the warm-up, while the values of the performance indices are collected over the other five days. The simulation model is implemented using AnyLogic 7.3.7.

All tests are made on the graph illustrated in Figure 1, composed by $n = 248$ nodes and $m = 512$ arcs on a metropolitan area of 880 km^2, among which $n_{ED} = 7$ hospital and $n_B = 14$ bases are distributed approximately in a balanced way. High-speed roads have been drawn through maximum speed arc paths, while traffic areas are located in different points using minimum speed arcs.

Six different scenarios are analyzed in our analysis. Scenarios 1–3 are obtained ranging the ambulance workload W^A in $\{30\%, 40\%, 50\%\}$ and setting the total ED capacity equal to 1.5 times the minimum capacity needed to deal with the demand. Then, scenarios 4–6 are defined for the same values of W^A but keeping the same total ED capacity, which is $1.7C_u$, $1.275C_u$ and $1.02C_u$ when $W^A = 30\%$, 40% and 50%, respectively. The initial number of ambulances A_b per base is fixed equal to 2 in all scenarios, while the constraint about the maximum number A_b^{max} is relaxed in such a way to allow a higher flexibility to the redeployment policies, and to analyze their impact in the most favorable situation.

Results of Scenarios 1–3 are reported in Table 5 focusing on indices regarding only performance of the ambulances. As expected, the increasing of the ambulance workload W^A causes a robust lengthening of waiting times, which pass from 6 to 110 min on average for the baseline configuration, and a general worsening of the indices. Such an increasing allows us to appreciate the impact of different configurations, that is for scenarios 2 and 3. Enabling only the policy D-LLCB with respect to the baseline configuration, a slight general worsening can be observed, while H-SAQ and H-WLP provide small variations depending on the considered scenario. More significant is the impact of the redeployment policies and in particular R-Closest, which worsens the fraction f_{ry} of the urgent patients reached within 8 min by an ambulance of

Table 5: Results: ED capacity is $1.5C_u$, that is proportional to demand.

	Scen.1 - $W^A = 30\%$					**Scen.2** - $W^A = 40\%$					**Scen.3** - $W^A = 50\%$				
id	r	f_g	f_{ry}	u	u^+	r	f_g	f_{ry}	u	u^+	r	f_g	f_{ry}	u	u^+
0	6.1	98.2	78.3	27.0	32.4	24.8	59.1	42.8	41.8	52.7	110.0	14.1	14.0	57.9	75.5
1	6.9	98.0	75.8	27.6	33.4	24.3	58.6	41.1	42.1	53.2	111.8	13.7	13.3	58.1	75.7
2	6.0	98.4	78.9	27.0	32.5	22.1	60.8	43.5	41.7	52.6	111.0	14.3	13.9	57.9	75.5
3	6.8	96.1	74.2	28.9	35.6	28.7	51.1	37.0	43.8	56.0	108.0	13.9	14.1	57.6	75.2
4	8.5	96.0	57.9	28.8	31.8	11.2	85.5	46.8	40.2	44.3	32.5	41.0	25.4	54.0	59.1
5	7.0	96.6	70.7	27.8	32.0	11.4	81.4	54.9	39.6	45.3	45.6	31.8	26.3	55.0	62.3
6	8.6	95.2	57.3	29.1	32.2	10.8	87.0	47.5	40.0	44.1	34.9	38.7	24.8	54.6	59.8
0s	5.8	99.0	79.4	25.9	31.6	7.8	94.5	65.4	37.0	42.9	20.7	58.4	37.2	52.3	56.8
1s	6.9	98.6	75.9	27.6	32.4	8.9	93.0	61.3	38.1	44.0	20.3	58.8	36.2	52.8	57.3
2s	5.8	99.1	79.0	27.0	31.8	7.9	94.4	64.6	37.4	43.4	18.7	61.7	38.3	52.0	56.7
3s	6.1	98.8	76.1	28.5	34.1	8.9	91.0	58.9	39.8	46.4	19.4	59.0	35.5	53.6	58.6
4s	8.4	96.4	58.6	28.6	31.5	9.7	91.3	51.1	39.4	42.9	19.8	59.5	32.3	52.9	55.6
5s	6.7	97.2	72.0	27.5	31.4	8.4	92.4	62.7	37.8	42.6	18.6	62.5	40.1	51.7	55.4
7s	6.7	97.1	72.6	27.6	31.5	8.2	92.8	63.8	37.6	42.5	19.1	60.7	39.8	51.9	55.5
8s	7.3	97.6	71.5	29.2	34.7	9.9	89.6	55.8	40.7	47.1	21.2	55.2	33.5	54.4	59.1
9s	7.9	95.7	67.9	29.7	33.7	9.7	92.0	51.0	39.3	42.8	21.2	56.0	35.2	54.0	57.5
	average $u_{ED} = 64.9\%$					average $u_{ED} = 65.1\%$					average $u_{ED} = 65.2\%$				

about 20%, but it halves the waiting times in scenario 2 and reduce them of more than 70% in scenario 3. However, a more relevant impact is observed enabling the D-SA: regardless of which policies are enabled, it always allows better performance than when it is not activated. In particular, considering the best configuration with and without enabling the D-SA, f_{ry} raises from 54.9% to 65.4% in scenario 2 and from 26.3% to 40.1% in scenario 3 confirming the results on the real case study reported in Aringhieri et al. (2016).

As a counter intuitive result, we observe that the configuration with the highest average waiting time r in scenario 2 (0) becomes the better just enabling the D-SA (0s). Similarly, in scenario 3 the value of r passes from 110 to 21 min (-81%) with only the contribution of the D-SA. Finally, we observe that ambulance utilizations u and u^+ are consistent with the value of the parameter W^A and little variations cause significant differences in performance. The same observation worths for the ED utilization, which have not significant variations among configurations and whose average value is reported in the last row of Table 5.

Table 6: Results: ED capacity is fixed, that is $1.7C_u$, $1.275C_u$ and $1.02C_u$ in scenarios 1, 2 and 3, respectively.

	Scen.4 - $W^A = 30\%$					**Scen.5** - $W^A = 40\%$					**Scen.6** - $W^A = 50\%$				
id	r	f_g	f_{ry}	w_g	w_y	r	f_g	f_{ry}	w_g	w_y	r	f_g	f_{ry}	w_g	w_y
0	6.2	97.5	77.6	1.8	0.6	21.5	61.5	44.9	41.7	4.6	110.0	13.7	13.8	230.0	5.5
2	6.0	98.2	78.7	1.4	0.7	23.2	58.5	41.1	35.0	4.8	114.4	12.8	12.9	204.2	5.6
4	6.7	96.5	74.1	0.3	0.1	26.7	50.8	37.1	36.4	5.1	69.6	27.1	18.5	199.3	5.6
6	8.5	95.8	57.4	1.9	0.8	11.9	83.2	43.3	28.6	4.8	37.1	34.5	20.5	203.7	5.6
0s	5.8	98.9	79.6	2.3	0.7	7.9	94.6	64.9	40.6	4.7	26.1	52.5	33.6	245.9	5.6
2s	5.9	99.0	79.0	1.4	0.6	8.4	93.0	61.3	26.7	4.9	21.6	52.5	33.6	188.9	5.6
4s	6.2	98.6	75.7	0.3	0.1	9.4	89.6	56.6	28.4	4.9	20.3	56.9	35.1	177.2	5.8
9s	9.4	92.6	52.1	0.2	0.1	11.3	85.4	43.3	29.5	4.9	22.7	51.9	31.9	178.4	5.8
	average $u_{ED} = 57.1\%$					average $u_{ED} = 76.2\%$					average $u_{ED} = 93.2\%$				

In Table 6 we summarize the most significant configuration to analyze the impact of ED facility selection on the waiting time for the ambulance arrival and the admission at the ED. Scenarios 4–6 are obtained ranging the value of W^A as well as for scenarios 1–3 but fixing the total ED capacity, which is equal to $1.7C_u$, $1.275C_u$ and $1.02C_u$, respectively. Such scenarios allow us to analyze the trade-off between the indices regarding the ambulance performance and those about the ED performance. The impact on the ED waiting times is evident: policy H-SAQ give a significant decreasing, but H-WLP is the best policy, up to -13% in scenarios 2 and 3. Both policies perform better enabling the D-SA, although this is not its goal. However, these improvement are not always concurrently with the best solution for the fast arrival of ambulance. For instance, in scenario 2 configuration 0s have better values of r, f_g and f_{ry} than configuration 2s, but green code patients have a waiting time 52% higher at the ED. In some cases, a good compromise can be found combining policies, as happens in scenario 2 for configuration 6 with respect to the other configurations without D-SA. In the last row of Table 6, we can observe the model consistency about the fixed ED workload parameters. The trade-off between the average time to reach an emergency request and the average waiting time at the ED confirms the fact that incorporating equity might lead to a lower service or negative outcomes as reported in McLay and Mayorga (2013).

The impact of the cut-off on the ambulance dispatching is studied in Table 7, where results of 2 configurations (with and without D-SA) with a good trade-off between ambulance and ED indices are reported. The threshold parameter of the policy D-CPQ ranges between 5% and 50% of the total number of ambulances. We observe that waiting times of urgent and non-urgent patients are sensitive to the D-CPQ: at the increasing of the threshold, the formers have an improving at the expense of the latters. However, in this scenario the use of the D-CPQ seems to be inadvisable, because of the possible high negative impact on green code patients to obtain a slight time saving for the other patients. This confirms the fact that priority dispatching policies can improve the performance for urgents at the price of a worsening those of non-urgents, as reported in Bandara et al. (2014).

Table 7: Results: impact of the D-CPQ policy varying the threshold (Scenario 2).

id	threshold	r	f_g	f_{ry}	u	u^+	w_g	w_y	id	threshold	r	f_g	f_{ry}	u	u^+	w_g	w_y
6t	5%	11.0	86.3	47.1	40.2	44.4	7.6	2.4	0st	5%	8.1	93.3	64.4	37.7	43.7	10.6	2.3
6t	15%	11.2	85.3	46.6	40.4	44.5	6.6	2.0	0st	15%	8.1	93.8	63.7	37.8	43.8	10.7	2.4
6t	25%	11.3	84.7	48.2	40.2	44.3	6.8	2.3	0st	25%	8.1	92.0	64.7	37.4	43.3	10.5	2.3
6t	30%	11.7	81.8	47.5	40.4	44.5	6.7	2.1	0st	30%	8.4	89.9	64.8	37.6	43.6	9.5	2.2
6t	35%	13.6	75.2	49.3	40.1	44.2	7.1	2.3	0st	35%	9.5	82.2	65.4	37.9	43.8	9.1	2.3
6t	40%	14.3	70.6	50.3	39.9	44.1	6.9	2.2	0st	40%	11.4	73.7	64.9	38.3	44.2	9.4	2.2
6t	45%	23.4	54.0	50.2	40.8	44.9	6.0	2.1	0st	45%	16.3	57.5	65.9	38.8	44.5	8.3	2.1
6t	50%	40.2	40.5	52.0	40.6	44.7	6.3	2.0	0st	50%	30.7	34.3	67.0	39.8	45.5	8.8	2.1

6 CONCLUSIONS

In this paper, several DRRP have been presented and analyzed for the ambulance real-time management. We provided a comprehensive analysis of the EMS system that allows us to make an extensive comparison among different policies. In particular, we provided a general analysis of the smart assignment policy, which confirms the significant results reported by Aringhieri (2010) and Aringhieri et al. (2016) for the EMS of Milano, Italy. The impact of such a policy on sufficiently crowded scenario is huge and allows us to have performance better than using any other combination of policies. Regarding the other policies, results shown a trade-off among their impact on a fast arrival of the ambulance and the waiting times for the admission in the ED confirming the insight reported by McLay and Mayorga (2013) for which incorporating equity might lead to a lower service or negative outcomes. More generally, the trade-off among the outcomes of the different policy combinations justify the need of a modeling approach to support decision making in the EMS management.

REFERENCES

Aboueljinane, L., E. Sahin, and Z. Jemai. 2013. "A Review on Simulation Models Applied to Emergency Medical Service Operations". *Computers & Industrial Engineering* 66(4):734–750.

Aringhieri, R. 2010. "An Integrated DE and AB Simulation Model for EMS Management". In *2010 IEEE Workshop on Health Care Management*, 1–6. Venice, Italy: IEEE.

Aringhieri, R., M. Bruni, S. Khodaparasti, and J. van Essen. 2017. "Emergency Medical Services and Beyond: Addressing New Challenges through a Wide Literature Review". *Computers and Operations Research* 78:349–368.

Aringhieri, R., G. Carello, and D. Morale. 2016. "Supporting Decision Making to Improve The Performance of an Italian Emergency Medical Service". *Annals of Operations Research* 236(1):131–148.

Aringhieri, R., D. Dell'Anna, D. Duma, and M. Sonnessa. 2018. "Evaluating The Dispatching Policies for a Regional Network of Emergency Departments Exploiting Health Care Big Data". In *International Conference on Machine Learning, Optimization, and Big Data*, edited by G. Nicosia, P. Pardalos, G. Giuffrida, and R. Umeton, Volume 10710 of *Lecture Notes in Computer Science*, 549–561. Cham, Switzerland: Springer International Publishing.

Bandara, D., M. Mayorga, and L. McLay. 2014. "Priority Dispatching Strategies for EMS Systems". *Journal of the Operational Research Society* 65(4):572–587.

Bélanger, V., A. Ruiz, and P. Soriano. 2018. "Recent Optimization Models and Trends in Location, Relocation, and Dispatching of Emergency Medical Vehicles". *European Journal of Operational Research*. Published Online: March 7, 2018. DOI: 10.1016/j.ejor.2018.02.055.

Channouf, N., P. L'Ecuyer, A. Ingolfsson, and A. Avramidis. 2007. "The Application of Forecasting Techniques to Modeling Emergency Medical System Calls in Calgary, Alberta". *Health Care Management Science* 10(1):25–45.

Cuninghame-Greene, R., and G. Harries. 1988. "Nearest-neighbour Rules for Emergency Services". *Zeitschrift fur Operations Research* 32(5):299–306.

Duma, D., and R. Aringhieri. 2017. "Mining The Patient Flow Through an Emergency Department to Deal with Overcrowding". In *3rd International Conference on Health Care Systems Engineering*, Volume 210 of *Springer Proceedings in Mathematics and Statistics*, 49–59. Cham, Switzerland: Springer International Publishing.

George, F., and K. Evridiki. 2015. "The Effect of Emergency Department Crowding on Patient Outcomes". *Health Science Journal* 9(1):1–6.

Haghani, A., Q. Tian, and H. Hu. 2004. "Simulation Model for Real-Time Emergency Vehicle Dispatching and Routing". *Transportation Research Record: Journal of the Transportation Research Board* 1882(1):176–183.

Hwang, U., and J. Concato. 2004. "Care in the Emergency Department: How Crowded is Overcrowded?". *Academic Emergency Medicine* 11(10):1097–1101.

Jagtenberg, C., P. van den Berg, and R. van der Mei. 2017. "Benchmarking Online Dispatch Algorithms for Emergency Medical Services". *European Journal of Operational Research* 258(2):715–725.

Larsen, A., O. Madsen, and M. Solomon. 2002. "Partially Dynamic Vehicle Routing-Models and Algorithms". *The Journal of the Operational Research Society* 53(6):637–646.

Lee, S. 2014. "Role of Parallelism in Ambulance Dispatching". *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 44(8):1113–1122.

Maxwell, M. S., S. G. Henderson, and H. Topaloglu. 2009. "Ambulance Redeployment: An Approximate Dynamic Programming Approach". In *Proceedings of the 2009 Winter Simulation Conference*, edited by M. D. Rossetti, R. R. Hill, B. Johansson, A. Dunkin, and R. G. Ingalls, 1850–1860. Piscataway, New Jersey: IEEE.

McLay, L., and M. Mayorga. 2013. "A Dispatching Model for Server-to-customer Systems that Balances Efficiency and Equity". *Manufacturing & Service Operations Management* 15(2):205–220.

Nafarrate, A., J. Fowler, and T. Wu. 2010. "Bi-criteria Analysis of Ambulance Diversion Policies". In *Proceedings of the 2010 Winter Simulation Conference*, edited by B. Johansson, S. Jain, J. Montoya-Torres, J. Hugan, and E. Ycesan, 2315–2326. Piscataway, New Jersey: IEEE.

Nasrollahzadeh, A., A. Khademi, and M. Mayorga. 2018. "Real-Time Ambulance Dispatching and Relocation". *Manufacturing & Service Operations Management*. Published Online: April 11, 2018. DOI: 10.1287/msom.2017.0649.

Ni, E., S. Hunter, S. Henderson, and H. Topaloglu. 2012. "Exploring Bounds on Ambulance Deployment Policy Performance". In *Proceedings of the 2012 Winter Simulation Conference*, edited by C. Laroque, J. Himmelspach, R. Pasupathy, O. Rose, and A. M. Uhrmacher, 1–12. Piscataway, New Jersey: IEEE.

Paul, S., M. Reddy, and C. Deflitch. 2010. "A Systematic Review of Simulation Studies Investigating Emergency Department Overcrowding". *Simulation* 86(8-9):559–571.

Ramirez-Nafarrate, A., A. Hafizoglu, E. Gel, and J. Fowler. 2012. "Comparison of Ambulance Diversion Policies via Simulation". In *Proceedings of the 2012 Winter Simulation Conference*, edited by C. Laroque, J. Himmelspach, R. Pasupathy, O. Rose, and A. M. Uhrmacher, 967–978. Piscataway, New Jersey: IEEE.

Reuter-Oppermann, M., P. L. van den Berg, and J. L. Vile. 2017. "Logistics for Emergency Medical Service Systems". *Health Systems* 6(3):187–208.

Setzler, H., C. Saydam, and S. Park. 2009. "EMS Call Volume Predictions: A Comparative Study". *Computers & Operations Research* 36(6):1843–1851.

van Barneveld, T., C. Jagtenberg, S. Bhulai, and R. van der Mei. 2018. "Real-time Ambulance Relocation: Assessing Real-time Redeployment Strategies for Ambulance Relocation". *Socio-Economic Planning Sciences* 62:129–142.

van Buuren, M., R. van der Mei, K. Aardal, and H. Post. 2012. "Evaluating Dynamic Dispatch Strategies for Emergency Medical Services: TIFAR Simulation Tool". In *Proceedings of the 2012 Winter Simulation Conference*, edited by C. Laroque, J. Himmelspach, R. Pasupathy, O. Rose, and A. M. Uhrmacher, 509–519. Piscataway, New Jersey: IEEE.

Yoon, S., and L. A. Albert. 2017. "An Expected Coverage Model with a Cutoff Priority Queue". *Health Care Management Science*. Published Online: July 19, 2017. DOI: 10.1007/s10729-017-9409-3.

AUTHOR BIOGRAPHIES

ROBERTO ARINGHIERI is Assistant Professor at the University of Turin, Italy. He holds a M.S. in Computer Science and a Ph.D. in Mathematics for Economic Decisions and Operations Research from the University of Pisa. His main research interest focuses on simulation and optimization algorithms applied to health care management. He is officer and member of the board of the EURO Working Group on Operational Research Applied to Health Services. His email address is roberto.aringhieri@unito.it.

SIMONE BOCCA is a M.S. student of Computer Science at University of Turin, Italy. He holds a B.S. in Computer Science from the University of Turin. His email is simone.bocca@edu.unito.it.

LUIGI CASCIARO is a recent former student at the Department of Computer Science of the University of Turin, Italy. He holds a B.S. and a M.S. in Computer Science from the University of Turin. His email is luigi.casciaro@edu.unito.it.

DAVIDE DUMA is a Ph.D. student at Computer Science Department of University of Turin, Italy. He holds a B.S. in Mathematics and Computer Science from University of Salento and a M.S. in Mathematics from University of Turin. His main research interest focuses on quantitative analysis and decision support systems for health care management. His email address is davide.duma@unito.it.