Ann Oper Res (2016) 236:131–148
DOI 10.1007/s10479-013-1487-0

Supporting decision making to improve the performance of an Italian Emergency Medical Service

Roberto Aringhieri · Giuliana Carello · Daniela Morale

Published online: 5 November 2013

Abstract An Emergency Medical Service (EMS) plays a fundamental role in providing good quality health care services to citizens, as it provides the first answer in distressing situations. Early response, one of the key factors in a successful treatment of an injury, is strongly influenced by the performance of ambulances, which are sent to rescue the patient. Here we report the research carried on by the authors on the ambulance location and management in the Milano area (Italy), as a part of a wider research project in collaboration with the EMS of Milano and funded by Regione Lombardia. The question posed by the EMS managers was clear and, at the same time, tricky: could decision making tools be applied, based on the currently available data, to provide suggestions for decision makers? To answer such a question, three different studies have been carried on: first the evaluation of the current EMS system performance through statistical analysis; then the study of operational policies which can improve the system performance through a simulation model; and finally the definition of an alternative set of posts through an optimization model. This paper describes the methodologies underlying such studies and reports on how their main findings were crucial to help the EMS in changing its organization model.

Keywords Emergency Medical Services · Ambulance · Simulation · Optimization · Decision making

R. Aringhieri (✉)
Dipartimento di Informatica, Università degli Studi di Torino, Turin, Italy
e-mail: roberto.aringhieri@unito.it

G. Carello
Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Milan, Italy
e-mail: giuliana.carello@polimi.it

D. Morale
Dipartimento di Matematica, Università degli Studi di Milano, Milan, Italy
e-mail: daniela.morale@unimi.it

1 Introduction

An Emergency Medical Service (EMS) is in charge of providing pre-hospital (or out-of-hospital) acute care to patients with illnesses and injuries. EMSs play a fundamental role in providing good quality health care services to citizens, as they provide the first answer in distressing situations. Besides, their importance is increasing due to the ageing of population. The key factors in a successful treatment of an injury are: early detection, early reporting, early response, good on scene care, care in transit, transfer to definitive care. Each factor has to be carefully managed in order to guarantee a suitable and quick response to citizens needs. In particular, the early response is strongly influenced by the performance of ambulances, which are sent to rescue the patients. The resources of an EMS, such as ambulances, are usually limited and therefore their management has a considerable impact on the overall system performance. Ambulances are usually deployed in order to provide a suitable coverage of the considered area, namely in order to reach each demand point within a limited time. Although different policies may be applied, according to a quite common policy the ambulances wait in a set of locations called *ambulance posts* or *posts*, a post being essentially a reserved parking. Such policy is applied by the EMS operating in the urban area of Milano, Italy, which is the subject of our study. Posts have been identified over the years without a clear coverage plan and without any decision support, except that of personnel experience.

In the past years the EMS of Milano has collected a huge amount of data about its every day activity, which however were never used to evaluate the possibility of improving the system performance and the management of the limited resources. Therefore the question arose if such huge amount of data could be exploited and decision making tools could be applied so as to provide suggestions for decision makers. This topic has been the subject of a research project in collaboration with the Emergency Medical Service of Milano and funded by Regione Lombardia.[1] Within the project, the authors carried on a research on the ambulance location and management in the Milano area, which is described in this paper. Three steps were developed: first the current EMS system performance was evaluated through statistical analysis based on the collected data; then a simulation model was developed in order to study operational policies which can improve the system performance; and finally an optimization model was studied with the purpose of defining alternative sets of posts. This paper describes the methodologies underlying such studies and reports on how their main findings were crucial to help the EMS in changing its organization model.

The paper is organized as follows. Sections 2, 3 and 4 illustrate the three studies, respectively. The findings which support the EMS management in reorganizing its process, leading to a new organizational model, are reported in Sect. 5. Section 6 closes the paper discussing some general insights regarding the EMS management in Italy and some new methodological aspects inspired by our collaboration with the EMS of Milano.

2 The performance of the actual EMS

As mentioned, the EMS of Milano collects, via the Operations Centre (OC), a huge amount of data describing the ambulance services or *missions*, from the instant in which a call is received by the operator to the instant in which the ambulance leaves the hospital and comes

[1] Regione Lombardia is the regional administrative district to which Milano belongs, and it is in charge of organizing emergency services.

Table 1 Frequencies of the ambulance requests. The first column lists the total requests of ambulances, which may or may not be served by prepaid ambulances; the second column shows the services covered by prepaid ambulances during the whole day; the last column the number of service requests covered by prepaid ambulances in the time period 7 a.m.–11 p.m

	Ambulance requests	Prepaid ambulances	7 a.m.–11 p.m.
Urgent calls	51413	41647	34663
Nonurgent calls	44681	36368	29808
	96094	78015	64471

back to an ambulance post. The operators at the OC are in charge of answering the calls and assigning a color code to each patient, based on the severity of injury, through a phase called *triage*. Here we refer to an urgent call as a patient with very severe injury to whom a red or yellow code is assigned: the Italian law states that the response to urgent calls has to be performed within a mandatory time of 8 minutes in the urban areas. From now on, we refer to this mandatory time as *LAW time*. After the triage phase the operator usually dispatches the nearest ambulance. Ambulance crew rescues the patient and, if necessary, transports him/her to a hospital. Ambulance crew is in charge of the patient until he/she is handed to the hospital staff. The crew looks after the patient until a bed is available; in the meanwhile the ambulance is not available.

The Milano EMS uses two types of ambulances, which differ in the kind of applied contract. The first set, composed of 29 ambulances, is always available and represents a fixed cost which does not depend on the number of performed missions; the second set can be summoned if needed and it is paid for each performed mission. Ambulances of the first set are located in the ambulance posts, while ambulances of the second one wait in the headquarters of the volunteering organizations which own them. We denote the ambulances in the first set as *prepaid ambulances*.

The EMS needed to evaluate, from a quantitative point of view, its capability of satisfying the emergency demand arising in different points of the urban area. This means that a statistical analysis of the available data with the goal of gathering together system performance and spatial information was needed. Therefore, we first analyzed the EMS activities in the time interval 7 a.m. to 11 p.m.; as a matter of fact, during this time period, the prepaid ambulances are placed in the corresponding ambulance posts, while during the night they wait in their headquarters, where the crew can have a rest. The results about the activities of year 2005 are reported in Table 1.

During the year 2005, the OC received 145844 calls. Among them, 96094 calls required an ambulance service. As described in Table 1, out of the number of requests, 78015 services were carried out by prepaid ambulances, 64471 of which were covered by the ambulances located at the posts, during the 7 a.m.–11 p.m. interval. In Table 2 the number of served requests is detailed for each post. Here we focus on the performance of the 29 prepaid ambulances, located in the 29 ambulance posts. The rationale is to guarantee that the EMS serves the largest number of requests with prepaid ambulances, with the aim of both minimizing the service cost and guaranteeing the same quality of service to every citizen.

The first aim of the study was to evaluate whether or not the actual post location covers all the emergency requests within LAW time. In order to estimate the area covered by each ambulance post within the mandatory time, we first performed a statistical analysis upon the random variable "*OC performing time*" describing the time needed by the OC to assign a call to a specific ambulance. We estimated an average OC performing time of 2.328 minutes with

Table 2 Frequencies per ambulance post of the 34663 urgent services covered by prepaid ambulances in the time period 7 a.m–11 p.m

Post	# Call	Post	# Call	Post	# Call	Post	# Call
1	894	8	1493	15	2129	22	260
2	884	9	1055	16	903	23	1128
3	1546	10	1847	17	1807	24	815
4	1666	11	492	18	1241	25	616
5	2019	12	2096	19	1909	26	190
6	1284	13	681	20	1659	27	912
7	912	14	1389	21	1444	28	534
						29	858

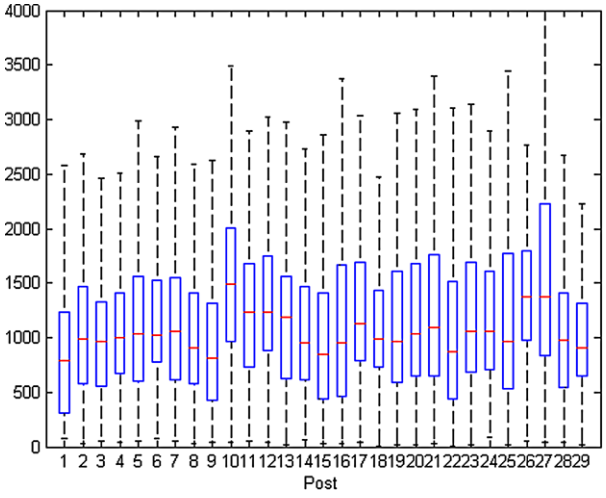


Fig. 1 Box-Whisker plot for the Euclidean distance (in meter) covered by the ambulances within the LAW time, starting from each of the 29 posts

a 95 % confidence interval, given by [2.32, 2.34], in the case of urgent calls. For nonurgent calls, the average OC performing time is higher, i.e., 4.58 minutes, with a 95 % confidence interval [4.53, 4.63]. We observed that both intervals are rather tight. Hereafter, we consider the allowed travel time as the difference between LAW time and the average time needed by the OC to alert an ambulance for a high priority call.

In order to estimate the area covered by the ambulances, we consider a random variable which describes the Euclidean distance between the post and the scenes traveled within the allowed travel time, according to the collected data detailed in Table 2.

The use of Euclidean distance is due to the fact that there is a lack of information on the trajectory of each ambulances: actually, we do not have any information about the routes followed by the ambulance drivers and, by consequence, GIS distances can not be measured. To verify if it is possible to consider the Euclidean distance instead of the real one, we have performed a regression among the two distances. It comes out that it is statistically significant to consider the linear relation $d_{GIS} = 1.4 * d_E$ (p-value < 0.05), where d_{GIS} is the GIS distance while d_E is the Euclidean distance.

Figure 1 depicts the Box-Wisker plot for the distance in meters covered by ambulances starting from each of the 29 assigned posts: the y-axis denotes the distance in meters covered

Table 3 Percentage of demands served within the LAW time for each post

Post	%	Post	%	Post	%	Post	%
1	62.1 %	8	65.2 %	15	57.5 %	22	51.5 %
2	59.1 %	9	61.1 %	16	69.3 %	23	71.3 %
3	61.1 %	10	54.0 %	17	58.4 %	24	48.6 %
4	63.0 %	11	64.9 %	18	66.2 %	25	66.8 %
5	60.1 %	12	63.4 %	19	53.5 %	26	57.4 %
6	61.4 %	13	66.3 %	20	61.4 %	27	46.4 %
7	65.2 %	14	59.0 %	21	61.7 %	28	57.6 %
						29	49.4 %

by an ambulance starting from the post reported in the x-axis, considering all the 34663 urgent missions served by prepaid ambulances from 7 a.m. to 11 p.m. Note that the average covered distance may be different from post to post. The sample size per each post is shown in Table 2.

From the above analysis, we deduced that the urban area of Milano is not completely covered by the posts within the LAW time. The estimated percentage of the demand served within the LAW time per ambulance post is shown in Table 3, while the average over all posts is 60.1 % with a 95 % confidence interval given by [56.13 %, 64.06 %]. It would be certain important to include in the reported analysis a distinction among days in a week or hours in a day but here we are interested in focusing on the covering capability of the target area. A wider analysis for estimating and forecasting the demand of ambulance service in the area of Milano is reported in Micheletti et al. (2010).

3 Actions for improving the EMS performance

The analysis reported have shown that there is room for improving the performance of the EMS system. To achieve this goal, two different actions are often taken into account, that is to increase the average ambulance speed and to add a new ambulance. A further action, suggested by the preliminary analysis, is meant to increase the time availability of ambulances. These actions, especially the first one, require a huge investment without any guarantee of return in terms of improving the performance.

To overcome this limitation, a simulation model has been developed in order to evaluate the behavior of the EMS system when a critical parameter, such as speed or number of ambulances, changes. In this section, we first describe a new simulation model adopted in the present analysis, and then we report about its use to evaluate the above actions.

3.1 The ABS-EMS simulation model

One of the most critical issues to be addressed in developing a simulation model for an EMS is how to model the movement of an ambulance in the system. The simulation models already proposed in literature (see, e.g., Goldberg et al. 1990a, 1990b; Henderson and Mason 2004; Ingolfsson et al. 2003; Wu and Hwang 2009; Van Buuren et al. 2012; Zaki et al. 1997) are usually based on a discrete event simulation (DES) approach. In a DES framework, the movement of an ambulance from a place to another one is usually represented by a new event, whose occurrence is set after a given time interval from the occurrence of the

event modelling the beginning of the movement. The interval represents the time needed by the ambulance to reach the destination. This time can be computed by using a travel time model (Henderson and Mason 2004) or exploiting a third part route planning software (Van Buuren et al. 2012) based on an accurate speed estimation. Basically, the actual movement of the ambulance is not an active part of the simulation model.

On the contrary, in our preliminary work (Aringhieri et al. 2008), and in its extension (Aringhieri 2010), we proposed an agent based model (ABS-EMS) in which the ambulance movement is a crucial part of the model: as a matter of fact, the agent modelling the ambulance replicates its movement on the Euclidean space or on the GIS map. This characteristic makes the model more flexible when testing different ambulance management policies: for instance, it naturally allows to reroute an ambulance while it is moving if a more serious emergency request occurs nearby. An agent based model allows to track the behavior of each individual acting in the simulated environment (Gilbert 2008). A set of rules describes the agent behavior and its interaction with the environment; as a consequence, the state of each agent is determined (Gilbert and Terna 2000). Here we report a simplified description of the whole model. Further details are reported in Aringhieri (2010). Even if it has been developed for the case of Milano EMS, ABS-EMS can be simply generalized by adjusting, if needed, the statecharts reported in the following.

ABS-EMS is intended to evaluate the EMS performance starting from a set of posts. Due to difficulty of having a reliable emergency demand generator in terms of both spatial and temporal distribution, each emergency request is generated by using the real data of a given day. This choice is also motivated by the need of the EMS managers to evaluate the system performance during some selected critical days, which are representative of typical emergency scenario such as, for instance, a day with a large number of missions and, among them, a large number of urgent requests. ABS-EMS is composed of two types of agents, that is "Operation Centre" and "Ambulance". Regarding the ambulance agent description, two different models will be reported, that is the "standard ambulance" and the "smart ambulance".

The Operation Centre The Operation Centre is in charge of two important decisions: which ambulance has to serve a given call and the time within which it happens. The target is to serve the emergency requests as fast as possible, trying to keep the whole urban area covered. Clearly, these two targets, namely serving each request as soon as possible and keeping a good area coverage with the available ambulances, contrast when the number of calls increases for an extended time period.

Currently, OC adopts the following simple strategy: it serves all the urgent calls quickly by assigning the service to the nearest available ambulance, whereas the nonurgent calls are queued if the number of available ambulances is below a given threshold. Practically, EMS adopts a nearest neighbor policy (Cuninghame-Greene and Harries 1988) which has been proven to perform, on the average, uniformly better than the other dispatching rules studied in Larsen et al. (2002). The agent modelling the OC follows the set of rules shown in Fig. 2.

Standard ambulance If not busy, an ambulance waits in a post until it is activated by the OC for a new service. Then it starts its task as depicted in Fig. 3. We refer to such type of ambulance as *standard ambulance*. The speed assigned to each ambulance is a function of the time of the day and of the area in which the ambulance is currently located. As already discussed in Sect. 2, Euclidean distances can be considered in place of GIS distances. Furthermore, this assumption reduces the running time of the simulation: actually, a GIS based simulator should compute a large number of shortest paths between two points in

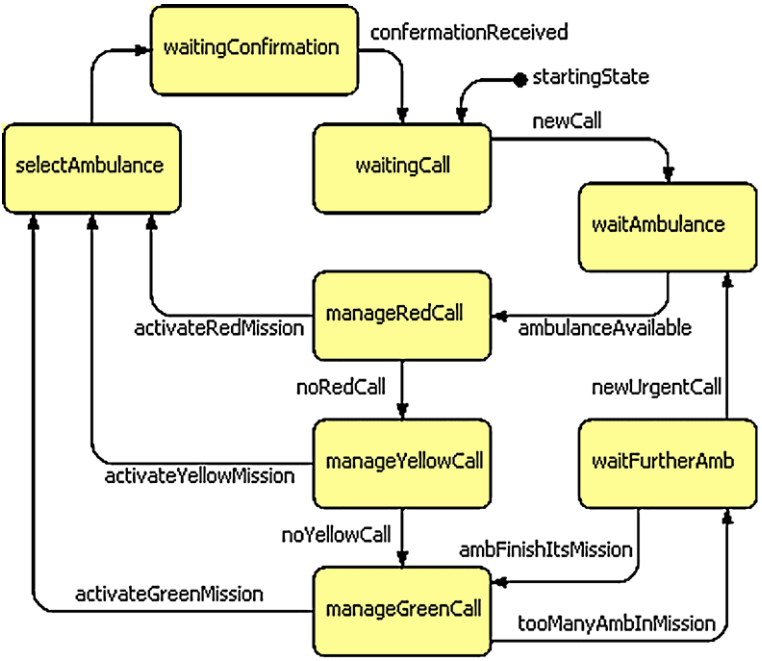


Fig. 2 Behavior of the agent "Operation Centre"

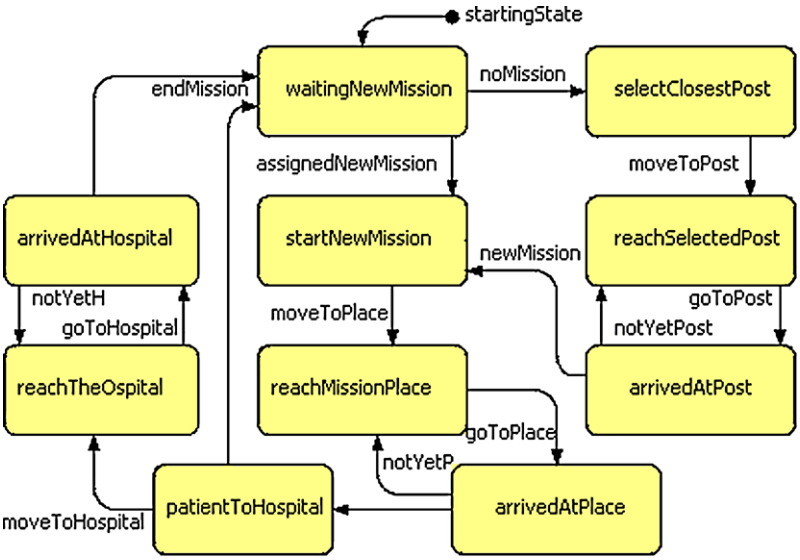


Fig. 3 Behavior of the agent "standard ambulance"

the graph representation of the GIS map, as discussed in Aringhieri (2010), which is more computational expensive than computing the Euclidean distance.

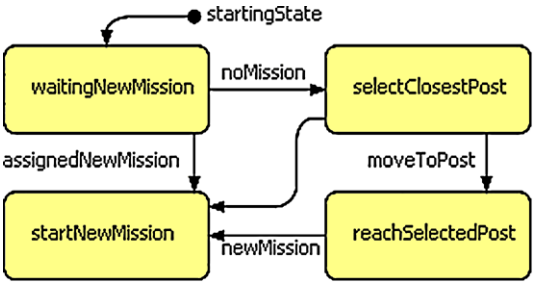


Fig. 4 Behavior of the agent "smart ambulance" (detail)

Table 4 Scenarios

Day	Total	Number of missions			
		Hospitalization	Not hospitalization	Urgent	Nonurgent
Jan 25	256	221	35	153	103
Feb 02	298	251	47	172	126
Mar 08	270	224	46	170	100
Apr 20	268	236	32	146	122
May 20	303	251	52	204	99
Jun 07	284	235	49	156	128
Sep 05	250	219	31	142	108

Smart ambulance Preliminary analysis showed that the time lost by an ambulance in the transfer to a post at the end of a service might be not negligible. This was also confirmed by the EMS management and by the senior OC operators. Therefore we evaluate the possibility of summoning an ambulance before it reaches the post after finishing a mission.

With respect to the statechart in Fig. 3, we note that a transition to "*startNewMission*" state is also possible from "*selectClosestPost*" and "*reachSelectedPost*" states. This capability, depicted in Fig. 4, might depend on both technological and human factors. We refer to this case as *smart ambulance*. Notice that this feature is clearly an advantage of ABS-EMS due to the flexibility of agent based methodology. We also observe that the implementation of this new feature strongly depends on the fact that the model simulates an effective ambulance movement.

3.2 Model validation and actions assessment

In order to carry out our analysis, we selected 7 different scenarios. Each scenario is selected to be representative of different and critical levels of emergency load and different compositions of the emergency demand. Each scenario represents a day from 7 a.m. to 11 p.m. reporting all the required information to perform the simulation, such as time instants, call coordinates and triage codes. Table 4 describes the scenarios in terms of calls. Hereafter, we consider an *experiment* as the execution of the simulator on all of the 7 scenarios. For each experiment, we report the percentage of urgent (**U**) and nonurgent (**nU**) calls **not served** within LAW time. Although a nonurgent call is not subject to LAW time constraint, it is important to evaluate them in terms of quality of service. Note that each experiment requires about 1 minute of running time, on average.

Concerning the ambulance speed, our study showed that it is usually very close to its average value of 25.8 km/h in the time period 7 a.m.–11 p.m. Note that this behavior is

Table 5 ABS-EMS validation: percentage of calls not served within LAW time for each scenario (last column reports the average percentage)

	Jan 25	Feb 02	Mar 08	Apr 20	May 20	Jun 07	Sep 05	Avg.
U	29.41 %	48.26 %	32.35 %	29.45 %	45.59 %	43.59 %	23.94 %	36.08 %
nU	38.83 %	64.29 %	41.00 %	41.80 %	39.39 %	50.00 %	28.70 %	43.43 %

Table 6 Action 1: percentage of calls not served within LAW time for each scenario and for each average ambulance speed tested (last column reports the average percentage)

Speed		Jan 25	Feb 02	Mar 08	Apr 20	May 20	Jun 07	Sep 05	Avg.
20.8	U	54.25 %	64.53 %	54.71 %	48.63 %	75.98 %	66.03 %	45.77 %	58.56 %
	nU	61.17 %	88.10 %	66.00 %	56.56 %	72.73 %	72.66 %	50.00 %	66.74 %
30.8	U	16.99 %	28.49 %	21.18 %	17.81 %	38.24 %	26.28 %	14.79 %	23.40 %
	nU	28.16 %	47.62 %	26.00 %	31.15 %	28.28 %	29.69 %	17.59 %	29.78 %

the same empirically observed by ambulance drivers in their experience. As a consequence, in agreement with EMS managers, we set the speed of the ambulance equal to its average value. Ambulances are modeled as standard ambulances.

Model validation The validation of a simulation model requires a quite complex analysis. This is particularly true in the case of ambulance simulation (Goldberg et al. 1990a, 1990b; Henderson and Mason 2004). Since we are interested in the evaluation of calls not served within the LAW time, we focus our validation process on this value. Table 5 reports the percentage of calls not served within the LAW time obtained running the simulation model over the 7 test scenarios starting from the actual post location using 29 ambulances. The last column reports the average among those values.

Table 5 reports an average number of calls not served within LAW time equal to 36.08 % corresponding, conversely, to the 63.92 % of urgent calls served within the LAW time. As reported in Sect. 2, the average value obtained over all the posts is 60.1 % with a 95 % confidence interval [56.13 %, 64.06 %]. Since the average value 63.92 % belongs to the estimated confidence interval, we can consider the simulation outcomes enough representative of the EMS behavior.

Action 1: average speed analysis We consider the variation of the ambulance average speed. By decreasing the speed to 20.8 km/h, we represent the case in which traffic jam increases in the urban area. On the contrary, by increasing the speed to 30.8 km/h, we represent the case in which the municipality operates against traffic jam, that is, for instance, by arranging reserved lanes and green wave for ambulances on the main streets.

By comparing the results in Table 6 with those reported in Table 5, we observe that the worsening and the improvement of the EMS performance are not proportional to the decrement and to the increment of the average speed. Note that the average speed in Milano is slightly decreasing along the years, because of traffic congestion.

Action 2: adding a new ambulance The simulation experiment consists in evaluating the impact of adding one ambulance but keeping the same number of posts (29) and their corre-

Table 7 Action 2: percentage of calls not served within LAW time for each scenario (last column reports the average percentage)

		Jan 25	Feb 02	Mar 08	Apr 20	May 20	Jun 07	Sep 05	Avg.
+1	U	32.03 %	44.77 %	38.24 %	31.51 %	48.53 %	48.72 %	28.87 %	38.95 %
	nU	33.01 %	57.94 %	38.00 %	35.25 %	37.37 %	47.66 %	28.70 %	39.70 %

Table 8 Action 3: percentage of calls not served within LAW time for each scenario (last column reports the mean percentage)

	Jan 25	Feb 02	Mar 08	Apr 20	May 20	Jun 07	Sep 05	Avg.
U	17.65 %	28.49 %	21.18 %	26.71 %	25.49 %	26.28 %	15.49 %	23.04 %
nU	30.10 %	50.00 %	30.00 %	23.77 %	19.19 %	35.94 %	22.22 %	30.17 %

sponding location. This scenario models the solution that would be implemented temporarily by EMS management until a new post is activated.

From the results reported in Table 7, we observe that the average performance of the system slightly decreases if compared with that in Table 5. Adding a new ambulance maintaining the same number of posts implies that each post should host, if needed, two ambulances at the same time, instead of one. We also remark that the system currently implements a nearest neighbor policy. It is well known in literature (Channouf et al. 2007; Setzler et al. 2009) that 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. Analysis confirms the demand swings hour by hour (Micheletti et al. 2010), while no meaningful difference has been showed during the day of the year except for special days such as, for instance, August holidays or snowing days (Righini et al. 2011). This implies that, during the simulation, it might happen that some posts are uncovered whilst some other posts are covered by two ambulances as ambulances tend to follow the emergency demand. Basically, ambulances tend to be gathered in few posts determining an unbalanced global coverage.

Action 3: ambulance time availability The current system does not allow the assignment of ambulances when they are traveling to an ambulance post, as their position is unknown since the prepaid ambulances are not equipped with a Global Positioning System (GPS). Despite the clear advantages determined by such a system, the past years were characterized by a lack of political will to support the new organization regarding especially the need of a secure communication link between ambulances and OC.

In the current experiment, we consider smart ambulances instead of the standard one: as a matter of fact, as depicted in Fig. 4, a new mission can be assigned during the path toward an ambulance post.

Table 8 reports the results of the experiment. We observe a performance improvement with respect to the same case without smart ambulances (see Table 5). This fact indicates that increasing the time availability of an ambulance, catching up the time spent when they are en route to a location while they are not serving a call, can significantly increase the EMS performance. Furthermore, if we consider the scenario "May 20", we can observe a large percentage reduction (about the 20 %) of urgent calls not served within LAW time. This scenario represents the EMS *heavy day*, i.e., day with a large number of total missions and, among them, a large number of urgent calls. The same trend is confirmed by the results of scenarios "Feb 02" and "Jun 07".

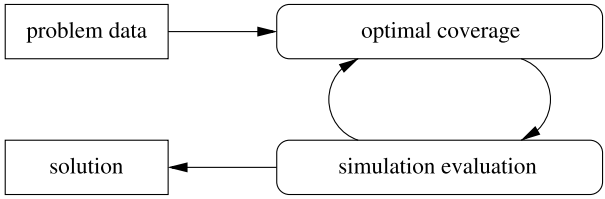


Fig. 5 The optimization simulation process

We also observe that the improvement obtained adopting smart ambulances is equivalent, on average, to that obtained by increasing the average speed. Notice that the technologies required by smart ambulances are cheaper than the cost of road upkeeping to increase average speed. Moreover, the average speed depends not only on the road status but also, for instance, on the weather. Therefore, the investment for smart ambulances seems to be more trustworthy than the one needed to increase the average speed.

4 An alternative set of ambulance posts

The simulation model discussed in Sect. 3 allows to evaluate different actions whose target is to improve the operational efficiency of the current EMS system. Clearly, the improvement of such actions strongly depends on the current set of ambulance posts: the strategic decision on the ambulance posts location has an impact on the operational efficiency. But what happens if a different set of ambulance posts is used? Is the current set of posts really efficient? This section is devoted to the third step of the study, namely the problem of finding an efficient set of ambulance posts in the urban area of Milano.

Uncertainty arises inherently in many parameters describing an EMS system, such as the instant of a new call occurring, the time of response, the traveling time, the waiting time in a hospital, and therefore the ambulance availability is itself a random variable. However, the randomness is more relevant at the operational level than at the strategic one. The methodology proposed here is a combined approach of optimization and simulation models: the optimization model takes into account the emergency demand and the required coverage level to determine an optimal set of posts while the simulation model evaluates the actual coverage dealing with the randomness of parameters. The proposed approach follows the line of the main finding reported in Aringhieri et al. (2013) for which health care optimization problems often require to adopt unconventional solution methodologies: the optimization simulation process tries to determine a final solution through the iterated solution of the above models and the integration of their main results, as depicted in Fig. 5.

The remaining of the section is organized as follow. First we describe the optimization model, then we describe how to combine it with the simulation model discussed in Sect. 3.

4.1 The Low-Priority Calls Coverage optimization model

There is a widespread literature on the ambulance location models as evidenced by the number of surveys available in literature (Brotcorne et al. 2003; Goldberg 2004; Li et al. 2011; Marianov and ReVelle 1995; ReVelle and Hogan 1989). Typically, optimization models are classified in static models (such as in Church and ReVelle 1974; Gendreau et al. 1997) and dynamical models (see, e.g., Bélanger et al. 2012; Gendreau et al. 2001; Laporte and Louveaux 2010): dynamical models deal also with the relocation of ambulances after the end of a service instead of only dealing with location, as in static models. Furthermore, both deterministic (such as in Church and ReVelle 1974; Gendreau et al. 1997; Toregas et al. 1971)

and probabilistic (such as in Goldberg et al. 1990b; Iannoni et al. 2008; Larson 1974, 1975; Mandell 1998) descriptions of the phenomenon have been studied. An interesting taxonomy is proposed in Başar et al. (2012) while alternative approaches are recently discussed in Chanta et al. (2011), Noyan (2010).

Many strategic optimization models have been proposed in the literature (see, e.g., Toregas et al. 1971; Hogan and ReVelle 1986). However, some features of the Milano EMS, such as the different kinds of demands which can be served in different response time, are not usually considered. We developed a model which captures the management of nonurgent calls requiring the use of an ambulance: although it has been developed for the Milano case it can be applied to the more general problem of locating ambulance posts in a given area.

An ambulance can perform a limited number of missions during a given time interval. Such number may be also represented as an *ambulance capacity* and it depends on the ambulance position. The main advantage of the introduction of a capacity parameter is to take into account the ambulance availability in a static deterministic optimization model.

Coverage models usually take into account only the urgent calls because they require a response within a mandatory time. In those models, nonurgent calls are discarded even if they may require an ambulance mission. Besides, patients with nonserious disease may wait for considerable time in the hospital before they are assigned to a physician, thus reducing the ambulance availability. In the proposed model the impact of nonurgent calls on the ambulance capacity is taken into account together with the need of providing a good quality of service also to nonurgent patients.

Let $\mathcal{V}$ and $\mathcal{W}$ be the set of points to be covered and the set of candidate post locations, respectively. For each point $i \in \mathcal{V}$, d_i^h denotes the amount of urgent or *high priority* (h) demands arising in demand point i, while d_i^ℓ denotes the amount of nonurgent or *low priority* (ℓ) demands.

The capacity associated to each post $j \in W$ is denoted by k_j. Let $\mathcal{W}_i^h$ be the set of candidate posts from which demand point $i \in \mathcal{V}$ can be reached within the LAW time. To model the coverage of the nonurgent demand arising in $i \in \mathcal{V}$, we introduce a second time limit, which is less tight and models the quality of service requirement for nonurgent calls. According to such time limit, let $\mathcal{W}_i^\ell \subseteq \mathcal{W}$ be the set of posts covering $i \in \mathcal{V}$.

Two continuous variables are defined: y_{ij}, representing the fraction of emergency demand of point i served by an ambulance located in post j, and w_{ij}, representing the fraction of nonurgent demand of point i served by an ambulance located in post j. An integer variable x_j is defined for each post $j \in \mathcal{W}$, representing the number of ambulances assigned to the post.

The Low-Priority Calls Coverage (LPCC) model aims at providing a lower bound on the number of the ambulances needed to serve the demand, and it can be formulated as follows.

$$\min \quad z = \sum_{j \in \mathcal{W}} x_j, \tag{1a}$$

$$\text{s.t.} \quad \sum_{j \in \mathcal{W}_i^h} x_j \geq 1, \qquad \forall i \in \mathcal{V} \tag{1b}$$

$$\sum_{j \in \mathcal{W}_i^h} y_{ij} = 1, \qquad \forall i \in \mathcal{V} \tag{1c}$$

$$\sum_{j \in \mathcal{W}} w_{ij} = 1, \qquad \forall i \in \mathcal{V} \tag{1d}$$

$$\sum_{i\in\mathcal{V}}\sum_{j\in\mathcal{W}_i^\ell} d_i^\ell w_{ij} \geq q \sum_{i\in\mathcal{V}} d_i^\ell, \tag{1e}$$

$$\sum_{i\in\mathcal{V}} \left(d_i^h y_{ij} + d_i^\ell w_{ij}\right) \leq k_j x_j, \qquad \forall j \in \mathcal{W} \tag{1f}$$

$$x_j \in \mathbb{Z}_+,\ w_{ij} \in [0,1],\ y_{ij} \in [0,1], \quad \forall i \in \mathcal{V},\ \forall j \in \mathcal{W} \tag{1g}$$

The objective function (1a) aims to minimize the number of required ambulances. The first constraints (1b) state that there must be at least one ambulance close enough to each demand point, guaranteeing the coverage of the whole city. Constraints (1c) guarantee that all the urgent calls are served within the given time limit (LAW time). Constraints (1d) guarantee that all the low priority calls are served from any post and (1e) force at least a given percentage q of such demand to be served within the second time limit. Constraints (1f) state that the number of missions assigned to a post must not exceed the number of missions $k_j x_j$ that the post can afford in the considered time horizon. Finally, (1g) define the variables domain.

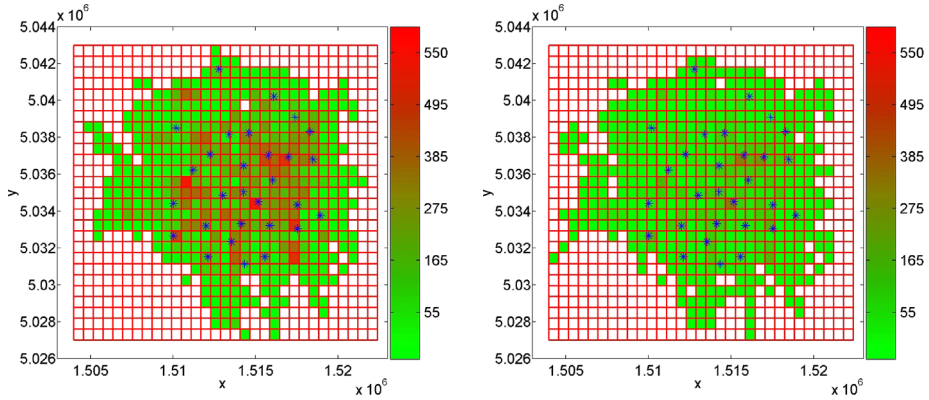
4.2 Combining LPCC and ABS-EMS

Combining optimization and simulation is a promising methodology as discussed in Fu (2002), Fu et al. (2005). Here we propose an iterative greedy procedure to compute an alternative set of posts for the urban area of Milano.

The procedure starts from the optimal solution computed by LPCC. This solution provides a lower bound on the number of the ambulances needed to cover the area under deterministic assumption. Such a solution is then evaluated via ABS-EMS determining a ranking of the ambulance posts with respect to their utilization. If the number of posts is less than the number of available ambulances, the procedure adds a new post in such a way to decrease the highest utilization value: let $p_{\max}$ be the post with the highest utilization value and let p' and p'' be the posts having larger utilization value among those near to $p_{\max}$; the new post is the point equidistant from $p_{\max}$, p' and p''. The procedure iterates from the ABS-EMS evaluation and the post is located for each available ambulance.

Finally, we observe that the iterative procedure can be used to design a set of posts in order to guarantee a given overall system performance: instead of stopping when available ambulances are finished, the procedure continues by adding ambulances until the performance of the system reaches a given threshold.

The solution for the case of Milano In order to determine the set of parameters for LPCC, the city is divided into 493 grid squares representing the demand points in the LPCC model. Each grid square represents a subarea such that every part of the subarea is covered by the same subset of candidate post locations. Thus, by guaranteeing that each subarea is covered, we guarantee that any possible origin of an urgent call is covered by at least one chosen post, and therefore served within LAW time. We estimate the distribution of the emergency demand via a statistical spatial distribution analysis. Figure 6 shows urgent (a) and nonurgent (b) estimated emergency requests for each subarea. We observe that the profiles are roughly the same. Furthermore, one can see that the demand is higher in the city center than in the suburb both for urgent and nonurgent demands. Finally, the urgent requests seems to be concentrated in the city center more than nonurgent requests. In the following, we refer to a square grid or subarea as a single point.



(a) urgent demand (b) non urgent demand

Fig. 6 Estimated spatial distribution profiles of the demand divided in urgent (**a**) and nonurgent (**b**) requests. Points represent the current ambulance posts (Gauss-Boaga coordinate system)

Table 9 Alternative ambulance posts evaluation: percentage of calls not served within LAW time for each scenario (last column reports the average percentage)

		Jan 25	Feb 02	Mar 08	Apr 20	May 20	Jun 07	Sep 05	Avg.
LPCC	U	73.86 %	76.16 %	74.71 %	67.81 %	82.35 %	77.56 %	68.31 %	74.39 %
	nU	96.12 %	97.62 %	97.00 %	90.98 %	80.81 %	92.19 %	73.15 %	89.69 %
LPCC	U	26.80 %	46.51 %	37.06 %	31.51 %	37.25 %	34.62 %	35.21 %	35.57 %
ABS-EMS	nU	42.72 %	60.32 %	45.00 %	34.43 %	49.49 %	47.66 %	35.19 %	44.97 %

The initial solution computed by LPCC is obtained by setting $\mathcal{W} = \mathcal{V}$, the amount of demand and the capacity are computed with respect to the time interval 7 a.m.–11 p.m., $q = 0.5$ and the time limit for nonurgent demand is set to 30 minutes. The optimal solution value is equal to 25. This means that the iterative procedure should allocate the remaining 4 ambulances.

The performance of the final solution obtained by the iterative greedy procedure is then evaluated using ABS-EMS in order to compare them with the current set of posts. Table 9 reports the performance evaluation of the set of posts computed by the iterative procedure (third and fourth rows). First and second rows report the performance of the initial solution computed by solving LPCC. First we remark the evident improvement gained by the iterative procedure with respect to the initial solution. Then, comparing these results with those in Table 5, we observe that such a solution is comparable, in terms of system performance, with the current location posts. The same remark holds when evaluating actions 1–3 starting from the new set of posts.

5 The answers and the EMS reorganization process

Although they are able to provide a lot of information about the system (see, e.g., their website http://www.118milano.it for real time information), the Milano EMS management

was not able to link together system performance and spatial information. In other words, they are not able to evaluate, from a quantitative point of view, their capability to satisfy the emergency demand coming from the urban area.

The studies reported in Sects. 2–4 have shown that there is room for improving the performance of the system following one of the three actions proposed and evaluated. Furthermore, they support the idea that the inefficiencies were due to the demand peaks during the day while less impact has the current post location. This claim is supported by the following remarks. The first one is the difficulty to find an alternative set of posts: the solution reported in this paper is the best one among 7 different solutions tested in Aringhieri et al. (2008) and obtained applying different strategic optimization models. Then, we observe that the LPCC solution determines a lower bound of the number of ambulances which is 25 against the 29 currently in service. Note that such a lower bound is the solution value of a deterministic optimization model in which the capacity parameter models the ambulance capacity assuming that emergency demands occurring in the same post are not simultaneous, which is not true especially when an emergency demand peak arises.

To improve the EMS performance, the more promising action to be taken seemed that of increasing the ambulance time availability. Furthermore, this action seemed to be more reliable in terms of final improvement. Therefore, the EMS management decided to introduce the concept of smart ambulance within the EMS organizational model. As already mentioned, the introduction of smart ambulances requires an innovation in terms of both technological and human factors determining a change in the EMS organizational model.

The main change concerned the introduction of the so called "logistic operator". The introduction of logistic operators allows to assign ambulances to missions when traveling to a post. It determines a change in the emergency request management. The classical OC operators are still in charge of answering the calls and of the triage procedure, but are no longer responsible for dispatching ambulances. Logistic operators are in charge of assigning ambulances to emergency requests. They can select the nearest ambulance among all the ambulances, that is to say, not only those waiting in a post, but also those traveling to a post. Furthermore, if there are many ambulances able to serve the request within LAW time, the logistic operator can select, based on his/her experience, the ambulance minimizing the loss of coverage. To allow the use of traveling ambulances, the system must be equipped with a complete GPS-based tracking system based on secure link connections. If such tracking system is not available, the logistic operator is also in charge of manually tracking the position of the ambulances. The transition to the new organizational model has been made easier allowing special training session for logistic operators employing the interactive simulator discussed in Pinciroli et al. (2010).

During the collaboration, a web site (http://118.dti.unimi.it) was maintained in order to spread the main project findings and to allow participation. Although it was open to everyone for consultation, the target audience of the OC operators and ambulance crews, i.e., people interested in the possible organizational changes. The success of the collaboration expanded the audience also to people working on other EMS of Regione Lombardia.

6 Conclusions

The emergency service is an important aspect in the life of every city, and, due to limited resources, requires a careful management. In Italy, the organizational model of emergency medical services is not uniquely defined and we can observe several and different models implemented by Italian EMSs. For instance, the Milano EMS deploys the ambulances in a set of posts so as to provide a coverage of the emergency demand.

Despite the availability of many Information Technology and mathematical decision support tools, in Italy EMSs usually locate and manage their ambulances based on operators' experience rather than on quantitative tools. To the best of our knowledge, the Milano EMS is the only one who systematically collects data about its every day activity. Nevertheless, such huge amount of data was not exploited to evaluate the system performance or to suggest new management strategies. In this paper we report a study which proves that statistical modelling, simulation and mathematical programming can be successfully applied to an EMS, in order to evaluate its current performance and to provide suggestions to improve it. The study shows that the Milano EMS provides a high level of performance, and yet it can benefit from the policy analytics techniques to improve the quality of the delivered service and to better exploit limited and expensive resources.

Regarding the organization of an EMS, we can gather some general insights. The first insight is that the ambulance management model in use at EMS of Milano could be exported in the other Italian large urban area adopting the iterative procedure depicted in Sect. 4.2. The second one concerns the impact of practical parameters, such as the average ambulance speed and the number of ambulances, when they may vary: the action assessment study shows, for instance, how to manage the introduction of a new ambulance in the system. The last one concerns the method for improving the performance of the whole system, i.e., the introduction of the required technologies and communication systems for smart ambulances. Furthermore, the smart ambulance can be used to implement more sophisticated dispatching policies: for instance, it allows to re-route an ambulance, while serving a low priority call, to serve a medical emergency having higher priority located nearby.

The close cooperation with EMS management has also determined some novelty in terms of methodological contributions. The need of dealing with nonurgent demands encouraged the development of the LPCC optimization model while the idea of increasing the ambulance time availability inspired the ABS-EMS model. Finally, difficulties in finding an alternative set of posts determined the development of the iterative procedure combining optimization and simulation. We remark that the developed procedure can be used to design from scratch a set of posts guaranteeing a given overall system performance.

Acknowledgements The authors wish to thank the Milano 118 Emergency Service Management for the fruitful collaboration and for providing us the data set and allowing their use in this paper. Besides, the author wish to thank the students S. Delfa, S. Perego and C. Romantini for their help for the numerical results. Finally, the authors wish to thank the anonymous referees for their comments which helped in improving the paper.

References

Aringhieri, R. (2010). An integrated DE and AB simulation model for EMS management. In *2010 IEEE workshop on health care management, WHCM 2010*. ISBN 978-1-4244-4998-9.

Aringhieri, R., Carello, G., & Morale, D. (2008). A simulation based tool for ambulances management evaluation. In *Operations research for health care delivery engineering, proceeding of the 33rd international conference on operational research applied to health service (ORAHS 2007)* (pp. 379–392).

Aringhieri, R., Tànfani, E., & Testi, A. (2013). Operations research for health care delivery. *Computers & Operations Research*, *40*(9), 2165–2166.

Başar, A., Çatay, B., & Ünlüyurt, T. (2012). A taxonomy for emergency service station location problem. *Optimization Letters*, *6*(6), 1147–1160.

Brotcorne, L., Laporte, G., & Semet, F. (2003). Ambulance location and relocation models. *European Journal of Operations Research*, *147*, 451–463.

Bélanger, V., Ruiz, A., & Soriano, P. (2012). Deployment and redeployment of ambulance vehicles in the management of prehospital emergency services. *INFOR. Information Systems and Operational Research*, *50*(1), 1–30.

Channouf, N., L'Ecuyer, P., Ingolfsson, A., & Avramidis, A. (2007). The application of forecasting techniques to modeling emergency medical system calls in Calgary, Alberta. *Health Care Management Science*, *10*, 25–45.

Chanta, S., Mayorga, M. E., & McLay, L. A. (2011). Improving emergency service in rural areas: a bi-objective covering location model for ems systems. *Annals of Operations Research*. Article in press.

Church, R., & ReVelle, C. (1974). The maximal covering locational problem. *Papers of the Regional Science Association*, *32*, 101–108.

Cuninghame-Greene, R., & Harries, G. (1988). Nearest-neighbor rules for emergency services. *Zeitschrift fur Operations Research*, *32*(5), 299–306.

Fu, M. C. (2002). Optimization for simulation: theory vs. practice. *INFORMS Journal on Computing*, *14*(3), 192–215.

Fu, M., Glover, F., & April, J. (2005). Simulation optimization: a review, new developments, and applications. In M. Kuhl, N. Steiger, F. Armstrong, & J. Joines (Eds.), *Proceedings of the 2005 winter simulation conference* (pp. 83–95).

Gendreau, M., Laporte, G., & Semet, F. (1997). Solving an ambulance location model by tabu search. *Location Science*, *5*(2), 75–88.

Gendreau, M., Laporte, G., & Semet, F. (2001). A dynamic model and parallel tabu search heuristic for real-time ambulance relocation. *Parallel Computing*, *27*, 1641–1653.

Gilbert, N. (2008). *Agent-based models*, *Quantitative applications in the social sciences* (Vol. 153). Thousand Oaks: Sage.

Gilbert, N., & Terna, P. How to build and use agent-based models in social science. *Mind & Society*, pp. 57–72 (2000).

Goldberg, J. (2004). Operations research models for the deployment of emergency services vehicles. *EMS Management Journal*, *1*(1), 20–39.

Goldberg, J., Dietrich, R., Chen, J. M., Mitwasi, M., Valenzuela, T., & Criss, E. (1990a). A simulation model for evaluating a set of emergency vehicle base locations. *Socio-Economic Planning Sciences*, *24*, 125–141.

Goldberg, J., Dietrich, R., Chen, J. M., Mitwasi, M., Valenzuela, T., & Criss, E. (1990b). Validating and applying a model for locating emergency medical vehicles in Tucson, AZ. *European Journal of Operational Research*, *49*, 308–324.

Henderson, S., & Mason, A. (2004). Ambulance service planning: simulation and data visualisation. In *Operations research and health care: a handbook of methods and applications*, *International series in operations research & management science* (Vol. 70, pp. 77–102). New York: Springer.

Hogan, K., & ReVelle, C. (1986). Concepts and applications of backup coverage. *Management Science*, *34*, 1434–1444.

Iannoni, A., Morabito, R., & Saydam, C. (2008). A hypercube queueing model embedded into a genetic algorithm for ambulance deployment on highways. *Annals of Operations Research*, *157*(1), 207–224.

Ingolfsson, A., Erkut, E., & Budge, S. (2003). Simulation of single start station for Edmonton EMS. *Journal of the Operational Research Society*, *54*(7), 736–746.

Laporte, G., Louveaux, F., Semet, F. d., & Thirion, A. (2010). *Applications of the double standard model for ambulance location*. *Lecture notes in economics and mathematical systems* (Vol 619, pp 235–249).

Larsen, A., Madsen, O., & Solomon, M. (2002). Partially dynamic vehicle routing-models and algorithms. *Journal of the Operational Research Society*, *53*(6), 637–646.

Larson, R. (1974). A hypercube queueing model for facility location and redistricting in urban emergency service. *Computers & Operations Research*, *1*, 67–75.

Larson, R. (1975). Approximating the performance of urban emergency service systems. *Operational Research*, *23*, 845–868.

Li, X., Zhao, Z., Zhu, X., & Wyatt, T. (2011). Covering models and optimization techniques for emergency response facility location and planning: a review. *Mathematical Methods of Operations Research*, *74*(3), 281–310.

Mandell, M. (1998). Covering models for two-tiered mergency medical service systems. *Location Science*, *6*, 355–368.

Marianov, V., & ReVelle, C. (1995). Siting emergency services. In Z. Drezner (Ed.), *Facility location, springer series in operations research and financial engineering* (pp. 199–223). Berlin: Springer.

Micheletti, A., Morale, D., Rapati, D., & Nolli, P. (2010). A stochastic model for simulation and forecasting of emergencies in the area of Milano. In *2010 IEEE workshop on health care management, WHCM 2010*. ISBN 978-1-4244-4998-9.

Noyan, N. (2010). Alternate risk measures for emergency medical service system design. *Annals of Operations Research*, *181*(1), 559–589.

Pinciroli, A., Righini, G., & Trubian, M. (2010). An interactive simulator of emergency management systems. In *2010 IEEE workshop on health care management, WHCM 2010*. ISBN 978-1-4244-4998-9.

ReVelle, C., & Hogan, K. (1989). The maximum availability location problem. *Transportation Science*, *23*(3), 192–200.

Righini, G., Gozzini, D., & Zorzi, G. (2011). Forecasting ambulance missions in Milan. In *Proceedings of ORAHS 2011 "OR informing national health policy"*.

Setzler, H., Saydam, C., & Park, S. (2009). Ems call volume predictions: a comparative study. *Computers & Operations Research*, *36*(6), 1843–1851.

Toregas, C., Swain, R., ReVelle, C., & Bergman, L. (1971). The location of emergency service facilities. *Operations Research*, *19*, 1363–1373.

Van Buuren, M., Van Der Mei, R., Aardal, K., & Post, H. (2012). Evaluating dynamic dispatch strategies for emergency medical services: TIFAR simulation tool. In *Proceedings—winter simulation conference*.

Wu, C. H., & Hwang, K. (2009). Using a discrete-event simulation to balance ambulance availability and demand in static deployment systems. *Academic Emergency Medicine*, *16*(12), 1359–1366.

Zaki, A., Cheng, H., & Parker, B. (1997). A simulation model for the analysis and management of an emergency service system. *Socio-Economic Planning Sciences*, *31*(3), 173–189.