Ambulance location through optimization and simulation: the case of Milano urban area

Roberto Aringhieri* Giuliana Carello†
Daniela Morale‡
E–mail: roberto.aringhieri@unimi.it, carello@elet.polimi.it, daniela.morale@mat.unimi.it

15th October 2007

Abstract

In this paper, the problem of locating ambulance posts over an urban area is considered. A three steps approach is presented to deal with this problem, that combines different skills. First, the real life data on the considered system behavior are analyzed. Then, integer linear programming models are considered with the aim of finding new post locations. As such models represent a simplification and an abstraction with respect to the real life situation, the behavior of the proposed solutions is tested with a simulation framework, tailored on the considered problem features. The whole approach is tested over the Milano city area case, with the aim of pointing out the criticality of the system and providing suggestions for the emergency service management. Computational results are presented and discussed.

Keywords: Optimization, simulation, ambulance location, statistical analysis.

1 Introduction

An Emergency Medical Service system is a service providing pre-hospital (or out-of-hospital) acute care to patients with illnesses and injuries. It exists to fulfill the basic principles of First Aid, which are to Preserve Life, Prevent Further Injury and Promote Recovery. The key factors in a successful treatment of an injury are: early detection (a member of the public finds the incident), early reporting (the emergency services are summoned), early response (the emergency services get to scene quickly), good on scene care (appropriate treatment is given), care in transit

*Dipartimento di Tecnologie dell'Informazione, Università di Milano – Corresponding author.
†Dipartimento di Elettronica ed Informazione, Politecnico di Milano.
‡Dipartimento di Matematica, Università di Milano.

(the patient is looked after on the way to hospital), transfer to definitive care (the patient is handed to the care of a physician).

One essential step is the early response, i.e. the emergency service, via an ambulance, should get to the event scene very quickly. This is the reason why organizing and planning emergency medical service in an urban area is an important problem arising everywhere in the world. Different features have to be considered while managing similar services, such as ambulance location and relocation, ambulance routing, assignment of ambulances to each call. In many real life systems different problems are solved via the experience of the individual operator.

In this paper we deal with the problem of locating the ambulances in agreement with the official Emergency Medical Service, called also 118[1], over the urban area of Milano, Italy. Here the *operation center* collects a wide amount of data describing the services[2], from the instant in which a call is received by the *operator* to the time an ambulance leaves the hospital after the service. The categorization of patients, based on their severity of injury, is made by assigning to each call a *color code*. Here we refer to an *emergency call* as the one to which a red or yellow code is assigned (very severe injury). The Italian law states that the response to emergency calls has to be performed within a mandatory time of 8 minutes in the urban areas (LA time). At the present, as we show in Section 2, in Milano urban area only around sixty percent of the emergency calls are served within the LA time. To provide a better performance of the system, the limited resources must be carefully managed and mathematical techniques are worthy to be studied to provide decision aided tools to the emergency service management. The aim of this work is first to analyze the data on the system behavior to point out possible criticality and then to provide, through optimization and simulation, suggestions for the emergency service management. With the present work we investigate on the possibility that a different distribution of the ambulance over the territory and/or a different number of ambulances might improve the performance of the system. Combining different mathematical tools, the aim of the present work is to find an optimal location of the ambulances in agreement with the operation center. Starting from the statistical analysis of these data aimed to estimate some basic variables needed for possible models describing the activity, we solve optimization problems and performed simulations.

Although many optimization models on the ambulance location and relocation problem have been developed [1, 7], to the best of our knowledge none is related to the Italian legislation about emergency services. From the latter we get the constraint of covering the one hundred percent of the high priority calls. It is clear that this request seems to be too much. In literature, usually a weaker constraint is considered; however here we try to conform to Italian law.

[1]In Italy the emergency service phone number is 118.

[2]A sample of the collected data is available on the web site www.118milano.it.

The first part of the work is the analysis of the present situation, i.e. the capability of the system to serve all the emergency calls within the LA time. As already mentioned, the 118 emergency service is not capable to satisfied this requirement, as it happens in the most of the Italian towns. In Section 2, we see how the first problem comes from the stage of early reporting, since in the average the call lasts two minutes before the operator is able to summon an ambulance. Here we cannot deal with this problem, since it could often be solved thanks to a better instruction of the population. The second problem arises from the evidence that currently the ambulance are not able to cover all the urban area within the LA time. This could depend either from the bad distribution of ambulances over the urban area or from the low number of ambulances in agreement[3] with the 118 service.

We perform the analysis of the activities within the time interval from 7 a.m. to 11.00 p.m., since in Milano during this time, the ambulances in agreement with the 118 emergency service are placed in a set of locations (*ambulance post* or *post* in the following), where they wait until they are summoned up by the operation center for beginning of a new service. After the transfer to a definitive care, ambulances are relocated at one of the free posts. Notice that all these paths are unknown a priori and, in the common practice, the ambulance drivers do not follow the shortest distance path. This is due to the fact that each driver is in charge of the path choice and he usually based its choice on the experience. During the night all ambulances are located in specific headquarters, that cannot be changed; as a consequence the problem of optimal distribution of posts does not occur. We have information about the topology of the town, the current locations of the ambulance posts, and their number. Furthermore, from the collected data set many other information can be extrapolated, such as the spatial frequency of calls in different subareas, the averaged speed of the ambulances, and as a consequence the areas that can be reached within the given time limit, and so on.

The main part of the work consists of the resolution of optimization problems, by looking for possible different ambulance post distributions via Integer Linear Programming models, and of the performance of simulations. In Section 3 we consider and tested some standard static optimization models. Furthermore, we propose a new model which considers new constraints which fit for the reality under study. Both standard and the original models are tested using CPLEX [10].

All the previous models are static one. As a consequence they cannot take into account some of the peculiarities of the phenomenon under study, e.g. the different latent time that an ambulance operator spends in the hospital. Indeed, the transfer to a definitive care does not happen instantaneously as soon as a hospital is reached. As a consequence, they do not take into account the availability of an ambulance. This is the main reason why it seems that the static models in Section 3 are able to

[3]In Milano the 118 service has agreements with voluntary organizations according to which some of the organizations ambulances are available for the 118 service. We refer to such ambulances, together with those owned by 118 service management, as *in agreement ambulances*.

give a solution to the full coverage within the LA time. Furthermore they do not consider the fact that when an ambulance is available, it may start a new service, on the way back to a post. In the paper we refer to such ambulances as *smart ambulances*. In order to include in a model all these features, in Section 4 we present a simulation framework which takes into account the dynamics, and tries to reproduce the real situation. All the solutions provided by the models in the previous section are tested through the simulation framework. Results obtained on the case study are reported in Section 5. Conclusion and suggestions for further developments are given in Section 6.

2 The actual emergency service in Milano

In this section we analyze the situation of the emergency service in Milano; in particular we consider a set of data describing the activities of the 118 operation center in the urban area of Milano during the year 2005.

	Ambulance request	118 agreement	Assigned to posts
emergency calls	51413	41647	34663
non emergency calls	44681	36368	29808
Total	96094	78015	64471

Table 1: Frequencies of the ambulance requests in Milano.

The amount of calls received and of required ambulance were 145844 and 96094, respectively. As described in Table 1, out of the number of requests, 78015 services were carried out by ambulances in agreement with the 118 service; among them 64471 services are covered by the ambulances located at the posts, during the day (7a.m.-11p.m.). The remainder 18079 calls are covered by other ambulances, whose expenses are covered by the 118 management for each service.

This is the reason why the aim of the management is to cover the whole city with ambulances in agreement, since their services over the whole year are already paid by contract. As a consequence we consider the performance of the services of the ambulances in agreement located at the 29 ambulance posts. In Figure 1 the demand and the ambulance post distributions are shown.

The first question we want to answer to is whether or not the actual post distribution on the urban area is suitable to cover all the emergency demands within the mandatory time. In order to estimate the area covered by each ambulance location, we first perform a statistical analysis of the random variable describing the time needed by a 118 operator to assign the call to a specific ambulance, from the time when he answered to the call. We call this variable ”118 performing time”. We have estimated an averaged 118 performing time of 2.328 minutes with a 95%

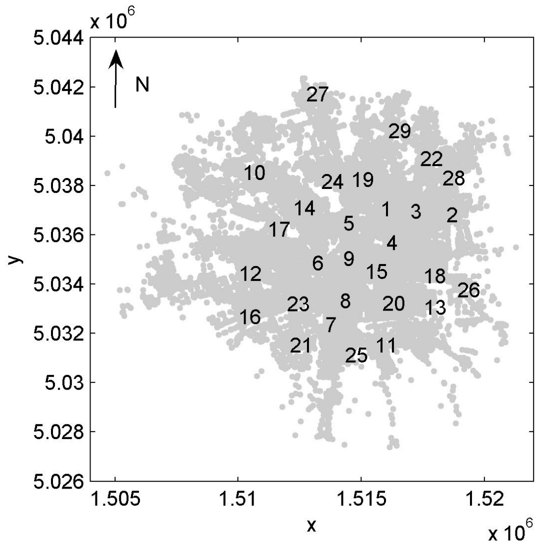

Figure 1: Spatial distribution of the request of ambulances in Milano (grey spots) and the ambulance posts (numbers). The positions are described by the Gauss-Boaga coordinate system.

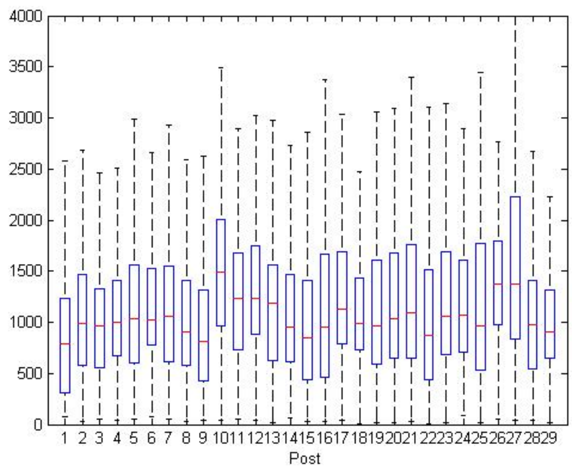

Figure 2: Box-Whisker plot for the distance covered by the ambulances within the LA time (emergency calls).

confidence interval, given by [2.32, 2.34]. We observe that the interval is rather tight. We recall that now we consider only the emergency calls. Indeed for the low priority calls (green code) the 118 performing time is much higher, i.e. 4.58 minutes, with a 95% confidence interval [4.53, 4.63]. From now on we consider as LA time the mandatory time minus the averaged time needed by the operation center to alert an ambulance for a high priority calls.

Post	mean radius	st. dev.	Post	mean radius	st. dev.
1	912.81	737.21	16	1191.06	852.95
2	1086.03	699.75	17	1330.48	942.88
3	1078.12	759.13	18	1195.63	761.97
4	1123.90	720.76	19	1189.59	841.25
5	1172.10	811.49	20	1220.30	812.21
6	1197.85	727.19	21	1229.68	776.43
7	1201.34	756.30	22	1159.82	992.76
8	1051.60	682.56	23	1236.45	798.10
9	1013.60	815.68	24	1245.59	925.65
10	1517.76	795.56	25	1310.55	1099.48
11	1288.56	799.61	26	1540.21	1053.64
12	1434.16	917.69	27	1648.05	1108.41
13	1242.67	986.16	28	1068.27	686.50
14	1190.31	853.27	29	1058.16	663.94
15	1022.78	765.21			

Table 2: Mean and standard deviation for the distance covered by a post within the LA time (meters).

In order to estimate the area covered by the ambulances within the LA time for emergency calls, we consider a random variable which describes the Euclidean distance between the post and the scenes reached within the LA time, for each post. In Figure 2 and in Table 2 a Box-Whisker plot and basic statistics for the distance covered by the ambulances within the LA time are shown, respectively. From the data one can deduce that the urban area of Milano is not covered by the emergency service within the LA time. The estimated percentage of the demands served within the LA time per ambulance post is shown in Table 3, while the average over all posts is 60,1%.

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

Table 3: Percentage of demands served within the LA time for each post.

Hence, it seems to be necessary to optimize either the spatial distribution of the ambulance posts or the number of ambulances. Currently, at each post only one ambulance is located. By means of an analysis of variance, by considering as dependent variable the mean radius per post, and as independent variables the hours of a day, we try to estimate homogeneous daily time intervals. We have divided the daily time (from 7 to 23) in seven time intervals: 7-9; 9-11; 11-14; 14-16; 16-19; 19-21; 21-23.

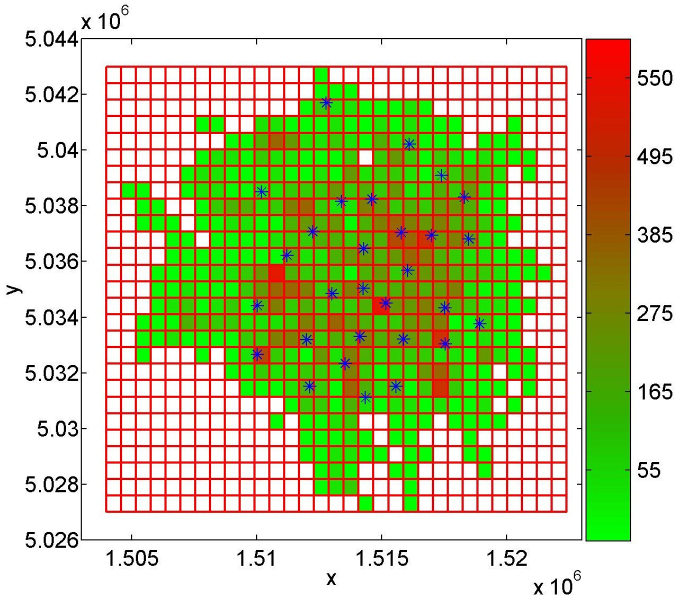

Figure 3: Spatial distribution estimates of the demand in the urban area of Milano: demands with hospitalization. The space is described by the Gauss-Boaga coordinate system.

The third step of the analysis is meant to built up some interesting fields, useful for the application of the optimization models. In particular, we estimate the spatial distribution of the demand using a square grid with a side per element of about 593 meters. From Figures 3 and 4, it is clear how a higher number of demands is located in the center of the town with low tails in the suburb. The same trend is highlighted both for the calls which needed a hospitalization and for the other ones. We have also carried out the estimate of this field for the seven relevant time intervals.

As already mentioned, the Euclidean metric (d_E) is used as spatial distance. The last significant issue is the relation between this metric and the distance d_{GIS} calculated using a geographical information system, GIS. In order to understand if it possible to consider the Euclidean distance instead of the real one we have performed a regression among the two distance. It comes out the it is statistically significant to consider the linear relation $d_{GIS} = 1.4 * d_E$. As discussed later on,

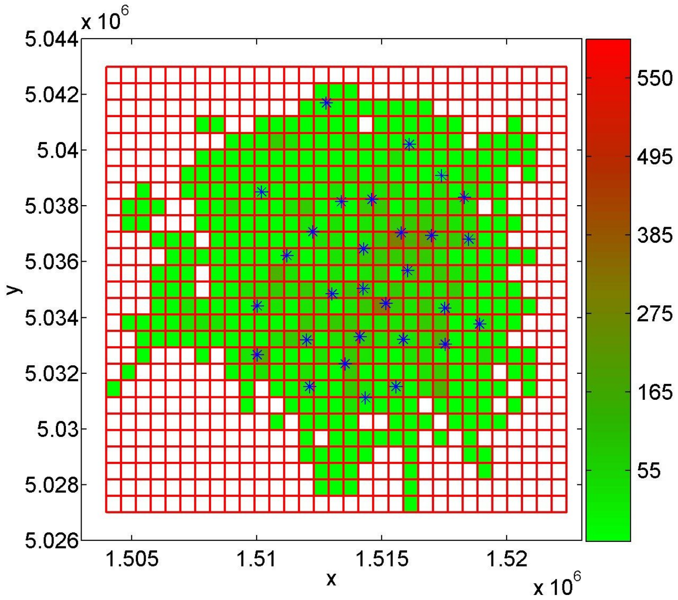

Figure 4: Spatial distribution estimates of the demand in the urban area of Milano: demands without hospitalization. The space is described by the Gauss-Boaga coordinate system.

we may consider the Euclidean distance qualitatively correct since we do not have any information about the true route followed by the ambulance.

3 Optimization Models

There is a widespread literature on the vehicles post location problem and in particular on the problem of optimal locating ambulance posts. A recent survey is given in [1], where static models (such as in [16, 2, 3]) and dynamical ([4]) models are discussed. One aspect we would like to highlight here is that the static ones do not describe the relocation of ambulances after the end of a service. Furthermore both deterministic (such as in [16, 2, 3]) and probabilistic (such as in [6, 11, 12, 13]) descriptions of the phenomenon have been studied. Indeed, there are many random parameters involved, as the instant of a new call occurring at the emergence service, the time of response, the waiting time in a hospital. etc. This randomness influences the availability of a free ambulance, that becomes a random variable itself. In deterministic models, in which this randomness is neglected, the optimization of post distribution imposes constraints on the coverage.

In this paper we focus on static deterministic models. In this section we first consider some well known basic models, able to capture some features of our case study. Then we propose a new model, including some peculiar aspects.

3.1 Standard optimization models

The first model (LSCM) we discuss has been proposed in [16] and it is a location set covering model. The only constraint taken into account is the coverage of the considered area. Each area of the city must be covered by at least one ambulance post. Let us recall that we say that a post *covers* an area if the distance between the post and the area is less than a given threshold, or equivalently, the mean time needed to reach the area starting from that post is less than a threshold time. The objective is to minimize the total number of ambulance posts. If it is assumed that there is one ambulance in each post, the total number of ambulances is minimized, as well.

Let $\mathcal{W}$ be the set of candidate locations for the posts. Let $\mathcal{V}$ be the set of points to be covered. In order to represent the city area with a set of points, the city is divided into grid squares, as mentioned in Section 2. Each grid square represents a subarea of the city such that every part of the subarea is covered by the same subset of candidate post locations. Thus, by guaranteeing that each $i \in \mathcal{V}$ is covered, we guarantee that any possible origin of an emergency call is covered by at least one chosen post, and therefore served within LA time. For each grid square $i \in \mathcal{V}$, let $\mathcal{W}_i$ be the set of candidate posts covering i. Let $x_j, \forall j \in \mathcal{W}$, be binary variables such that

$$x_j = \begin{cases} 1 & \text{if an ambulance post is set in site } j, \\ 0 & \text{otherwise.} \end{cases}$$

The model can be formulated as follows:

(LSCM)

$$\min \quad \sum_{j \in \mathcal{W}} x_j, \tag{3.1a}$$

$$\text{s.t.} \quad \sum_{j \in \mathcal{W}_i} x_j \geq 1, \qquad \forall i \in \mathcal{V}; \tag{3.1b}$$

$$x_j \in \{0, 1\}, \qquad \forall j \in \mathcal{W}. \tag{3.1c}$$

The LSCM may be consider as a first simple description of the problem under study. Since the traffic conditions, and therefore the average speed of the ambulances, and the coverage radius of each candidate post change during the day, then different instances of the problem can be derived by considering time partition in such a way that the mean distance covered by each post in each time interval may be considered the same by hypothesis. These data have been estimated, as described in Section 2. In this first model ambulance availability and amount of calls are not considered. Therefore, although the ambulances provided by the optimal solution can cover the whole city area, they might not be able to serve all the emergency calls. Thus the optimal number of ambulances given by the model represents a lower bound of the needed ambulances.

The next model (BACOP1), first proposed in [9], considers the importance of covering twice at least a fraction of the demand, in order to guarantee at least one backup ambulance for a fraction of the demand. An additional binary variable is introduced

$$u_i = \begin{cases} 1 & \text{if } i \text{ is covered twice,} \\ 0 & \text{otherwise.} \end{cases}$$

The importance of double coverage is represented by the objective function, aiming at maximizing the amount of demand covered by at least two posts, while the covering by one post is guaranteed. The model is the following:

(BACOP1)

$$\begin{aligned}
\max \quad & \sum_{i \in \mathcal{V}} d_i u_i, & & \text{(3.2a)} \\
\text{s.t.} \quad & \sum_{j \in \mathcal{W}_i} x_j - u_i \geq 1, & \forall i \in \mathcal{V}; \quad & \text{(3.2b)} \\
& \sum_{j \in \mathcal{W}_i} x_j = p, & & \text{(3.2c)} \\
& x_j \in \{0, 1\}, & \forall j \in \mathcal{W}; \quad & \text{(3.2d)} \\
& u_i \in \{0, 1\}, & \forall i \in \mathcal{V}. \quad & \text{(3.2e)}
\end{aligned}$$

Parameter d_i, defined as the number of emergency calls in grid square i, can be easily estimated. Parameter p represents the given number of ambulance posts to be located.

As LSCM, also BACOP1 does not take into account the fact that a post might not be able to serve a call, if the ambulance is already busy because of a previous call. Therefore this model provides a lower bound of the needed ambulances. Furthermore, post dimensioning, i.e. considering a number of ambulances located in the same post larger than one, is not considered.

A slightly different model (BACOP2) is also presented in [9]. Its objective function aims at maximizing a combination, using parameter θ, of the demands covered once and twice. The coverage of all urban area is not required in this case. Actually, in many real life systems the one hundred percent coverage is not guaranteed. A new binary variable y_i is introduced, for each subarea $i \in \mathcal{V}$, such that

$$y_i = \begin{cases} 1, & \text{if } i \text{ is covered,} \\ 0, & \text{otherwise.} \end{cases}$$

The model reads as the following

(BACOP2)

$$
\begin{aligned}
\max \quad & \theta \sum_{i \in \mathcal{V}} d_i y_i + (1-\theta) \sum_{i \in \mathcal{V}} d_i u_i, && && (3.3a)\\
\text{s.t.} \quad & \sum_{j \in \mathcal{W}_i} x_j - y_i - u_i \geq 0, && \forall i \in \mathcal{V}; && (3.3b)\\
& u_i - y_i \leq 0, && \forall i \in \mathcal{V}; && (3.3c)\\
& \sum_{j \in \mathcal{W}_i} x_j = p, && && (3.3d)\\
& x_j \in \{0,1\}, && \forall j \in \mathcal{W}; && (3.3e)\\
& u_i \in \{0,1\}, && \forall i \in \mathcal{V}; && (3.3f)\\
& y_i \in \{0,1\}, && \forall i \in \mathcal{V}. && (3.3g)
\end{aligned}
$$

3.2 A new optimization model

As some of the features of the considered real life case are not present in the above described models, we developed a new model tailored on the Milano case. However, the developed model is a static one and does not consider the arriving calls from a dynamic point of view. The model considers the dimensioning of the ambulance posts, i.e. the number of ambulances to be located in each post, to satisfy the total demand. We consider both the emergency (red and yellow code) and the low priority calls (green calls). As regards the green code calls, the law only mentions that they have to be served in a *reasonable* time; so we consider a second time limit, less tight than the LA time. Furthermore, since they have a non trivial impact on the ambulance availability, their assignment to an emergency vehicle is to be optimized as well.

The problem is to assign the red and yellow demand of each grid square to the posts, conditioned to the constraints on the time limit, while the green demand can be assigned to any post. However, to guarantee a good service, at least a given fraction of the low priority calls must be served within a reasonable, although less tight, time limit. For each post the number of needed ambulances is to be set, with respect to the demand assigned to the post.

Although the model is a static one and does not represent the dynamic behavior of the system, we want to take into account somehow the availability of the ambulances. An ambulance can afford a limited number of missions in a given time interval. This average value depends on both the time needed for the rescue and the time spent at the hospital and can be computed on the base of the data provided by the 118 service. Then it is possible use such number as an ambulance *capacity*. By guaranteeing that the ambulance capacity is not exceeded, i.e. by guaranteeing that the number of missions assigned to each ambulance is smaller than the computed limit, the availability of ambulances is taken into account, although in a static approximation. Even though the assignment provided by the model solution cannot be applied in real life, the model has to assign calls to posts to take into account the ambulance capacity.

As the standard models, the model aims at providing a lower bound on the number of the needed ambulances, as the model gives a simplified representation of the problem, e.g. the calls are considered as evenly distributed along the considered time horizon.

For each subarea $i \in \mathcal{V}$ of the city the amount of red and yellow demand are denoted with d_i^r and d_i^y respectively, while the green demand are denoted with d_i^g. All the demands are computed on the base of the available data over the whole year 2005. For each subarea i the set of candidate posts covering i within the given time limit is denoted by W_i^1. A second, less tight, time limit is given, according to which for each i a subset $W_i^2 \subseteq \mathcal{W}$ is defined. Finally, let k_j denote the number of missions that an ambulance located in post j can afford in the time horizon considered by the problem.

Two continuous variables w_{ij} and z_{ij} are defined, z_{ij} representing the fraction of red/yellow demand of grid square i assigned to post j and w_{ij} representing the fraction of green demand of subarea i assigned to post j. An integer variable x_j is defined for each post j representing the number of ambulances to be assigned to j.

The model (Lower-Priority Calls Coverage) can be formulated as follows.

(LPCC)

$$\min \quad \sum_{j \in \mathcal{W}} x_j, \tag{3.4a}$$

$$\text{s.t.} \quad \sum_{j \in \mathcal{W}_i^1} x_j \geq 1, \qquad \forall i \in \mathcal{V}; \tag{3.4b}$$

$$\sum_{j \in \mathcal{W}_i^1} z_{ij} = 1, \qquad \forall i \in \mathcal{V}; \tag{3.4c}$$

$$\sum_{j \in \mathcal{W}} w_{ij} = 1, \qquad \forall i \in \mathcal{V}; \tag{3.4d}$$

$$\sum_{i \in \mathcal{V}} \sum_{j \in \mathcal{W}_i^2} d_i^g w_{ij} \geq q \sum_{i \in \mathcal{V}} d_i^g, \tag{3.4e}$$

$$\sum_{i \in \mathcal{V}} (d_i^r + d_i^y) z_{ij} + d_i^g w_{ij} \leq k_j x_j, \qquad \forall j \in \mathcal{W}; \tag{3.4f}$$

$$x_j \in \mathbb{Z}_+, \qquad \forall j \in \mathcal{W}; \tag{3.4g}$$

$$w_{ij} \in [0, 1], \qquad \forall i \in \mathcal{V}, j \in \mathcal{W}; \tag{3.4h}$$

$$z_{ij} \in [0, 1], \qquad \forall i \in \mathcal{V}, j \in \mathcal{W}; \tag{3.4i}$$

The objective function aims at minimizing the number of needed ambulances. The first constraint (3.4b) states that there must be at least one ambulance close enough to each subarea, guaranteeing the coverage of the whole city. Constraint (3.4c) guarantees that all the emergency calls are served within the given time limit, while constraint (3.4d) guarantees that all the low priority calls are served from any post and (3.4e) forces at least a given percentage q of the low priority demand to be served within a second, less tight, time limit. Constraint (3.4f) states that the number of missions assigned to a post must not exceed the number of missions that the post can afford. This value is given by the number of ambulances assigned to the post (y_j) multiplied by the capacity of each ambulance. Finally, (3.4g), (3.4h) and (3.4i) set the variables domain.

In Section 5 we provide a comparison of the models by the estimation of the parameters and subsets in the case study of Milano area.

4 Simulation Framework

The models presented in Section 3 provide solutions, given estimates of the involved parameters, i.e. averaged behaviors are considered. Furthermore, as discussed in the previous Sections, only a static description has been considered, and, as a consequence, some features have been neglected. In order to test the solutions given by the static optimization models in a dynamics environment, a simulation framework has been developed to model the behavior of the emergency service with respect to the dynamic call arriving. We consider an agent based simulation model for the case under study, developed using the multi-paradigm simulation tool AnyLogic [14].

4.1 The Agent based Simulation Model

An agent based model allows to track the behavior of each individuals acting in the simulated environment. A set of rules describes the agent behavior and its interactions with the environment; as a consequence, the state of the each agent is determined. Different agents communicate via messages or signals [15]. Note that in AnyLogic, the rules are described by using a statechart, as those reported in Figures 5 and 6, in which a box represents an agent state whilst an arc represents a transition between two states.

Whenever a call is generated, a color code is assigned and the operator at the call center selects a free ambulance. Then the latter starts the service (*mission*) and get to the scene. After the appropriate care is given, if required, the patient is taken to the nearest hospital. So after the patient is handed to the care of a physician, the ambulance moves to a post, waiting for the next service assignment.

Therefore, a crucial point to be taken into account by the model is the movement of the ambulance. In literature a discrete event approach is often considered: in this framework, the movement of an ambulance from a place to another place is represented by a new event, whose occurrence is set in the time horizon after a given interval from the occurrence of the current event. The time interval represents the time needed by an ambulance to reach a given place and is computed using a travel time model such as the one proposed in [8]. In the agent based model we present here, the ambulance movement is continuous and is represented by a continuously varying state of the agent related to the ambulance itself. The state is given by a set of features of the agent, such as position, speed, direction, etc. This makes the model more flexible in testing different ambulance management policies. For instance, it allows to change the destination of an ambulance while it is moving, if a more serious disease occurs.

The main agents considered in our models are the call, the operator at the operation center and the ambulance.

The Call. Although our model is able to generate randomly the calls following a suitable probabilistic distribution, we have implemented a trace-driven simulation in which each call is generated by using the real data of a given day. In this way, the performance – number of not served calls – of the post location computed by the models can be tested on the set of data representing a given day. In a trace-driven simulation, each call agent state is given by spatial coordinates, the color code, and the hospital where the patient is moved.

The Operation Center. The operator at the call center is the decision core of the whole system. He takes two important decisions: which ambulance has to serve a given call and the time when it happens. The target is to serve the emergency calls as fast as possible, trying to keep the whole urban area covered. Clearly, these two targets contrast when the number of calls increases for an extended period, e.g. epidemic flu, summer hot days, etc. Currently, the operation center of Milano adopts the following simple strategy: it serves all the emergency calls quickly assigning the service to the nearest available ambulance, whereas the green calls queue in the case the number of available ambulances is below a threshold.

The agent modelling the operation center follows the set of rules shown in Figure 5: the operation center assigns an ambulance to the calls, ranked by their color code, if at least one ambulance is available; when two or more ambulances are available, the nearest one is selected.

The Ambulance. If not busy, an ambulance waits in a post until the operation center activates it for a new service. Then it starts its task: first it reaches the scene; then, if necessary, it brings the patient to a hospital. In our simulation framework the type of diseases is not considered: as a consequence a patient is carried to the closest hospital. At the end of the service, the ambulance moves to the closest empty post. The speed assigned to each ambulance is an estimated mean speed.

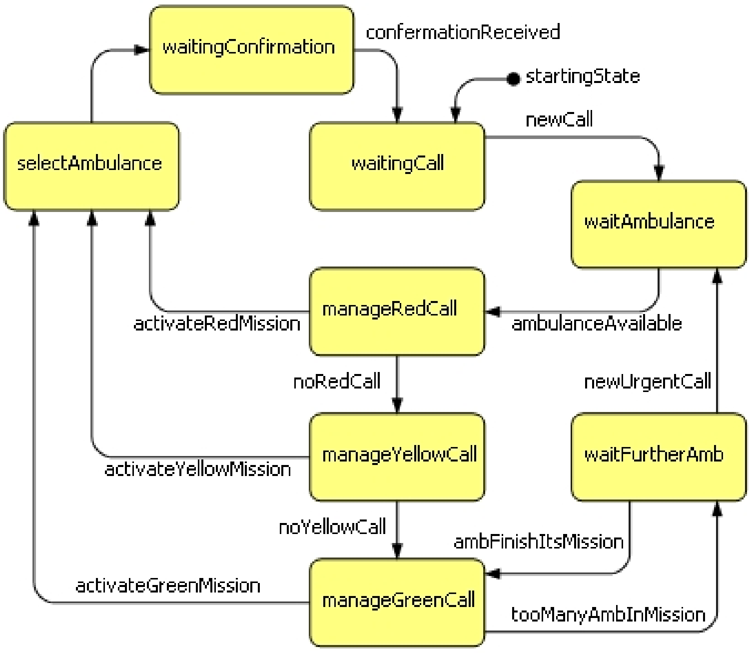

Figure 5: The statechart describing the behavior of the agent modeling the operation center.

Distances are considered as Euclidean: in fact such hypothesis reduces the running time of the simulation, with respect to a GIS based one. In Section 2 we have estimated a linear relation between the GIS and Euclidean distances. Moreover, we cannot calculate the distances a priori, since a priori we might not know the starting point of the ambulance (e.g. when a new service starts before it has reached a post). Finally, it allows a fair comparison, since coverage radius of the models are computed according to the Euclidean distance. Figure 6 shows the rules followed by the ambulance agent.

Notice that a transition to "startNewMission" is possible from several different states. This fact models the capability of the operation center to use an ambulance when it is moving back to a post after the end of a service. Such a capability might depend on both technological and human factors. We refer to this case as *smart ambulance*. Notice that this feature is clearly and advantage of the simulation model improving its flexibility. We also observe that the implementation of this feature depends on the fact that the model simulates an effective movement of the ambulance. Hereafter, we refer to the other case as *standard ambulance*.

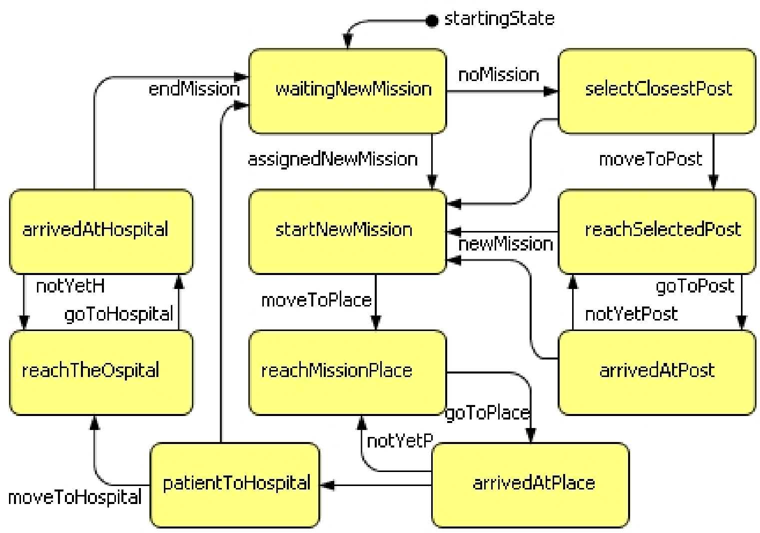

Figure 6: A simplified version of the statechart describing the behavior of the agent modeling the ambulance.

5 Case study

The models described in Section 3 have been tested on the data available discussed in Section 2. The urban area has been partitioned into subareas via a squared grid: the grid is built in such a way that each point of each element of the grid is covered within the LA time by the same subset of ambulance posts. Each subarea represents a demand point.

Following the analysis in Section 2, we have considered six homogeneous daily time intervals (intervals 9a.m -11a.m. and 11a.m-2p.m. have been collapsed into one for computational reasons), such that in each interval the mean radius covered by a post, within the LA time, may be considered equal, for each post. As a consequence, seven different instances can be derived for each model, one for each interval and one related to the whole day. Data referred to each instance represent the values of the considered data over one year. Each of these instance has been solved with CPLEX 8.1 on an Intel Xeon, at 2.80 GHz, with 2Gb of RAM memory. The models are solved on the seven instances. The tightest solution found, i.e. the one requiring the maximum number of ambulances, is validated with the simulation framework.

5.1 Results for the standard optimization models

The model LSCM provides the minimum number of posts to guarantee a coverage of the whole urban area within the given time limit. The constraint of the model guarantees that the whole city area is covered by any feasible solution of LSCM. However, the availability of ambulances is not taken into account nor is the number of calls: a solution provides a coverage of the city area but does not guarantee that each call is served. Therefore, a feasible solution represents a lower bound of the number of needed ambulances.

In Table 4 the minimum number of needed ambulance posts, together with the computational time needed to solve the model, is given for each time interval and for the instance over the whole day.

Time interval	Minimum number of posts	CPU time [sec]
7a.m.-9a.m.	20	2.43
9a.m.-2p.m.	17	1.24
2p.m.-4p.m.	18	2.60
4p.m.-7p.m.	18	1.27
7p.m.-9p.m.	19	0.57
9p.m.-11p.m.	20	2.45
whole day	17	0.49

Table 4: Results for LSCM model.

The ambulance speed, as well as the coverage radius of each candidate post, varies during the day, due to the different traffic conditions. Therefore the minimum number of ambulance posts changes as well. For instance, the minimum number of ambulances is twenty during the time interval 7a.m.-9a.m., while it is seventeen for the whole day instance. This is due to the fact that the average speed is low in the early morning, since the traffic level is higher due to the moving mass of workers, and therefore the coverage radius of the posts is small. On the other hand, in the whole day instance the speed is averaged out over the whole day and also the higher speed values of other time intervals contribute to its value. Then the average speed computed over the whole day is higher than the one computed over the interval 7a.m.-9a.m.: the coverage radii are larger. Beside, the optimal location of ambulance posts changes for different time interval; this may suggest the opportunity of relocating ambulances during the daytime. This may be seen as a first step toward a dynamical approach.

In Figures 7 and 8 the position of selected posts is shown together with the current ones for the interval 7a.m.-9a.m. and for the whole day. The demand distribution is shown as well. Both the solution move the post locations towards the suburbs of the city, with respect to the current posts. The solution for the interval 7 a.m.-9 a.m. locates more posts in the city center, due to the smaller coverage radius. Besides, both solution locate post on the city suburb even if the amount of demand is low on the suburb. This is due to the fact that the whole coverage of the area is required, without taking into account the amount of demand. Results show that the needed

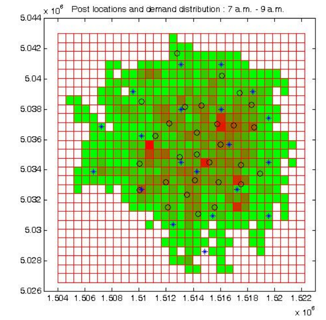

Figure 7: Current and computed post locations for the interval 7 a.m.-9 a.m. Current post locations are represented by circle, computed ones by star. The space is described by the Gauss-Boaga coordinate system.

computational time is always negligible, and this allows, if necessary, to test the model under different assumptions on the data.

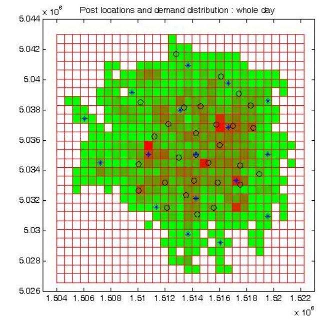

Figure 8: Current and computed post locations for the whole day. Current post locations are represented by circle, computed ones by star. The space is described by the Gauss-Boaga coordinate system.

The model BACOP1 allocates posts and guarantees the coverage of the whole city, while maximize the amount of demand covered by at least two posts. As in the first described model, the availability of the ambulances is not taken into account, and the number of ambulances per post is not dimensioned. First the number of posts is set to twenty-nine, as the actual one. In Table 5 BACOP1 model results are shown. It comes out that with the actual number of posts, almost the whole demand is covered twice: in fact even in the worst case about 95% of the demand is covered twice. However, as the ambulance availability is not considered, the solution does not guarantee that when an emergency call occurs there is always one ambulance available to serve it. Results show that computational time is reasonable for all the considered time intervals, rising up to about 17 minutes in the worst case, while being about 500 sec. in the average.

Time interval	Demands covered twice	Total demands	Percentage	CPU time [sec]
7a.m.-9a.m.	3939	4129	95.40	1044.08
9a.m.-2p.m.	14830	14838	99.95	11.33
2p.m.-4p.m.	4869	4887	99.63	493.33
4p.m.-7p.m.	7885	7904	99.76	701.58
7p.m.-9p.m.	4985	5013	99.44	212.05
9p.m.-11p.m.	5399	5526	97.70	835.05
whole day	49955	50071	99.77	156.55

Table 5: Results for BACOP1 model. Number of posts: 29.

Afterwards, we tested model BACOP1 both by setting the number of posts equal to the minimum obtained by LSCM, for each time interval, (i.e. twenty for the interval 7a.m.-9a.m.), and the minimum value increased by one (i.e. twenty-one for the interval 7a.m.-9a.m.). Results are shown in Table 6 and 7, respectively: in the first case the percentage of demands covered twice is always less than or equal to 50%; in the second case, by adding one more post, the percentage of double covered demand increases between 53% and 65%. These results show that if only one ambulance is placed in each post, the minimum number of post does not provide a robust coverage. It is necessary to take into account the availability of ambulances. Computational times are always limited, rising up to about 30 sec. only in one case.

Time interval	Demand covered twice	Total demand	Percentage	CPU time [sec]
7a.m.-9a.m.	1596	4129	38.65%	21.59
9a.m.-2p.m.	7362	14838	49.62%	23.55
2p.m.-4p.m.	2475	4887	50.64%	1.56
4p.m.-7p.m.	3757	7904	47.53%	13.18
7p.m.-9p.m.	2380	5013	47.48%	1.38
9p.m.-11p.m.	2371	5526	42.91%	33.93
whole day	16832	50071	33.62%	14.93

Table 6: Results for BACOP1 model - minimum number of posts estimated by LSCM.

The model BACOP2 does not guarantee the coverage of the whole urban area, but it maximizes a combination of the amount of demand covered once and twice. In Table 8 the amount of demand covered once and twice, according to different values of the parameter θ in equation (3.3a), is given for the whole day instance, together with the needed computational time. The number of posts is set equal to

Time interval	Demand covered twice	Total demand	Percentage	CPU time [sec]
7a.m.-9a.m.	2188	4129	52.99%	18.92
9a.m.-2p.m.	9643	14838	64.99%	14.39
2p.m.-4p.m.	2987	4887	61.12%	16.9
4p.m.-7p.m.	4788	7904	60.58%	15.54
7p.m.-9p.m.	3054	5013	60.92%	1.84
9p.m.-11p.m.	3161	5526	57.20%	20.32
whole day	25306	50071	50.54%	10.58

Table 7: Results for BACOP1 model - minimum number of posts estimated by LSCM plus one.

the current number, twenty-nine.

θ value	Demand covered once	Demand covered twice	CPU time [sec]
0.1	50046	50004	5470.52
0.3	50046	50003	797.29
0.5	50043	50003	144.215
0.7	50046	50004	105.92

Table 8: Results for BACOP2 model - whole day, total demand 50071.

From these results it seems that, even for different values of parameter θ, it is possible to reach a good single and double coverage. Computational time are reasonable for some values of θ, such as 0.5 or 0.7, but they may be quite high for others. As $\theta = 0.7$ seems to provide a good trade off between quality of the results and needed computational effort, in Table 9 results are given for all the time intervals for $\theta = 0.7$. For each interval the percentage of demand covered once and twice is given, together with the needed computational time. Almost complete coverage is obtained for all the time intervals, except the one from 7a.m. to 9a.m. However, as for the previously discussed models, also this model does not guarantee that every emergency call is served. The CPU time is limited, as it rises up to at most about 6 minutes even in the worst case.

5.2 Results for the new optimization model

The model LPCC guarantees that every demand point is assigned to one or more posts so that all its calls can be served. It takes into account the availability of the ambulances through the definition of the ambulance capacity, although assuming

Time interval	Demand covered once [%]	Demand covered twice [%]	CPU time [sec]
7a.m.-9a.m.	94.62	97.65	38.66
9a.m.-2p.m.	99.93	99.93	4.89
2p.m.-4p.m.	99.75	99.75	0.43
4p.m.-7p.m.	99.94	99.68	98.43
7p.m.-9p.m.	99.92	99.66	11.57
9p.m.-11p.m.	99.76	98.71	349.08
whole day	99.95	99.87	105.92

Table 9: Results for BACOP2 model - $\theta = 0.7$.

that calls are evenly distributed during the considered time interval. Further, it guarantees that at least a percentage of the low priority calls is served within a reasonable time. The number of needed ambulances is minimized. The percentage of the low priority calls to be served within small time is set to 50% and the less tight time limit is set to 30 minutes.

In Table 10 the objective function and the computational time for each time interval are given. Results shown in Table 10 are obtained assuming that the calls are evenly distributed over the considered time interval. The capacity is computed as the sum of the calls served in the considered interval along one year. The assumption on the distribution of the calls is a strong one. The time arrival of calls is not considered and, as a consequence, when a call occurs it is assumed that there is at least one available ambulance. Results show that, with these assumptions, the number of ambulances is the same given by the solution of LSCM. In real life, calls are not evenly distributed and it happens quite often that a call occurs when all the ambulances are already busy. In fact, in real life, a low priority call may wait for a considerable time if ambulances are all busy. Thus, to provide a robust solution, that can deal even with a not even call distribution and with critical situation, we test the model also on instances in which the computed ambulance capacity is reduced by a factor 0.5. Results are shown in Table 11.

Results of Table 11 show that, taking into account availability of ambulances, the number of posts must be increased or more than one ambulance must be assigned to some of the posts. In the former models basic coverage constraints are considered, according to which a post is sufficient to cover a demand point, no matter how many calls occur in the considered area. In this model, on the other hand, a post may not be enough to serve an area, if the ambulance capacity is low. Thus, more posts are to be located or more than one ambulance is to be placed at some posts. As for the other models, the number of posts is influenced by the coverage radius, that may change during the day, according to traffic conditions. That, together with the varying amount and distribution of demands, causes the changing of the objec-

Time interval	# ambulances	CPU time [sec]
7a.m.-9a.m.	20	243.06
9a.m.-2p.m.	17	909.68
2p.m.-4p.m.	18	889.14
4p.m.-7p.m.	18	979.24
7p.m.-9p.m.	19	344.62
9p.m.-11p.m.	20	235.26
whole day	17	442.66

Table 10: Results for LPCC model.

tive function value. However, the increasing in the number of ambulances is not the same for all the time interval. This may be caused by the different distribution of the calls. If in a time interval the calls are spread over the city area the number of ambulances is forced by the coverage constraint. For such time intervals the number of ambulances is great even with the non reduced capacity, due to the coverage constraint, and therefore reducing the ambulance capacity does not cause a significant increase in the number of ambulances. On the other hand, if in a time interval the capacity constraint is tighter than the coverage one, reducing the ambulance capacity causes a significant increase. The number of ambulances varies from 20, in the time interval 9 p.m. to 12p.m., to 25, for the whole day instance. The change in the solutions in different time intervals suggests that a better representation of the systems may be provided by dynamic models. In Figure 9 posts located by the solution are shown (star) together with current posts (circle). The solution locates more posts on the city suburbs than those currently located to guarantee full city area coverage. Besides, the solution locates more posts in the city center than those located by LSCM (Figure 8), due to the ambulance capacity and to the distribution of demand. As a consequence, the number of needed ambulance for the whole day is slightly higher than the number required for the interval 9 a.m. - 2 p.m..

Time interval	#ambulances used	CPU time [sec]
7a.m.-9a.m.	21	7828.87
9a.m.-2p.m.	24	13490.1
2p.m.-4p.m.	21	1842.96
4p.m.-7p.m.	21	20911.6
7p.m.-9p.m.	21	33072.4
9p.m.-11p.m.	20	2247.5
whole day	25	2545.16

Table 11: Results for LPCC model – half capacity.

All the considered instances can be solved to optimality, but the computational

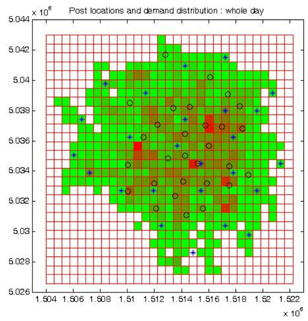

Figure 9: Current and computed post locations for the whole day. Current post locations are represented by circle, computed ones by star.

time rises up to about 9 hours in the worst case. However, since such model is to be solved out of line, the required effort is reasonable.

5.3 Simulation analysis

The results for LPCC provides a more robust solution with respect to those guaranteeing only coverage. Nevertheless, it is a simplification of the real life problem, as it does not represent the dynamic behavior of the system and does not take into account the simultaneity of the calls. Due to their static nature and to the assumption on which they are based, the optimization models discussed in Section 3 provide a set of lower bounds of the number of ambulances needed to guarantee a good service. To understand how tight such lower bounds are, the behavior of the tightest solution provided is tested with the simulation framework.

In order to carry out the simulation test, we selected 7 different instances. Each instance represents a day from the available data and reports all the required information to perform the trace-driven simulation such as time instants, call coordinates and triage codes. Tables 12 reports about the instances composition in terms of calls.

Hereafter, for each simulation experiment, we report the percentage of emergency (**E**) and non emergency (**nE**) calls not served within LA time. Note that each experiment requires 1 minute of running time on average. The average speed of an ambulance is set to 25.8 km/h.

Model validation

Generally, the validation of a simulation model requires a quite complex analysis. This is particularly true in the case of ambulance simulation [8, 6, 5]. Since we are

day	total # of calls	# of calls			
		hospitalization	not hospitalization	emergency	non emergency
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

Table 12: Description of instances used to perform simulation analysis.

interested in the evaluation of not served calls within the LA time, we focus our validation process on this value.

Table 13 reports the percentage of calls not served within the LA time running the simulation model over the 7 test instances starting from the actual post location using 29 ambulances and posts. From these results, we derive that the 63.92% of the emergency calls are served within the LA time.

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

Table 13: Simulation results of the current post location.

Table 3 reports the results from the analysis of the calls served within the LA time for each current post. As reported in Section 2, the average value obtained over all the posts is 60.1% with a 99.95% 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 real emergency system behaviour.

Optimization models comparison

We now compare the results obtained by the simulation of the solutions provided by the optimization models discussed in Section 3.

Table 14 reports the results obtained applying the post locations provided by the optimization models except those for LSCM which are always above the 98% of emergency calls not served within the LA time since it uses just 17 ambulances confirming the remarks derived from the results in Table 6 and 7. The solution of both BACOP1 and BACOP2 needs 29 ambulances and posts: their average results

		Jan 25	Feb 02	Mar 08	Apr 20	May 20	Jun 07	Sep 05	avg.
BACOP1	E	50.98%	60.47%	51.76%	45.21%	73.04%	53.85%	43.66%	54.14%
	nE	60.19%	81.75%	61.00%	48.36%	65.66%	63.28%	44.44%	60.67%
BACOP2	E	42.48%	58.14%	47.65%	50.68%	68.14%	55.13%	45.77%	52.57%
	nE	52.43%	76.19%	66.00%	56.56%	64.65%	61.72%	44.44%	60.28%
LPCC	E	73.86%	76.16%	74.71%	67.81%	82.35%	77.56%	68.31%	74.39%
	nE	96.12%	97.62%	97.00%	90.98%	80.81%	92.19%	73.15%	89.69%

Table 14: Simulation results of the solution provided by the optimization models.

are similar but worse than those obtained by the current post location (Table 13). Finally, the LPCC solution seems the worst one with respect to BACOP models and to the actual post location. These results are essentially due to the fact that LPCC uses 4 ambulances less than the other post locations.

In order to have a fair comparison, we tested three extended LPCC solutions. Two solutions are obtained by fixing in the model the number of posts to 29: the first solution (**F**) is just a feasible solution whilst the second (**M**) is computed by forcing the model to maximize the number of low priority calls served within a given time limit, i.e. 30 minutes. Finally, the third solution (**R**) is obtained by simply adding four current posts randomly chosen among those whose location does not overlap the posts computed by LPCC. The three solutions uses 29 ambulances and posts except for **F** which uses 28 posts.

		Jan 25	Feb 02	Mar 08	Apr 20	May 20	Jun 07	Sep 05	avg.
F	**E**	58.82%	79.07%	50.59%	54.79%	84.31%	71.15%	56.34%	65.01%
	nE	57.28%	94.44%	65.00%	63.93%	75.76%	78.13%	61.11%	70.81%
M	**E**	47.06%	68.60%	51.18%	51.37%	78.92%	58.97%	47.18%	57.61%
	nE	55.34%	88.89%	57.00%	52.46%	70.71%	71.88%	44.44%	62.96%
R	**E**	26.80%	46.51%	37.06%	31.51%	37.25%	34.62%	35.21%	35.57%
	nE	42.72%	60.32%	45.00%	34.43%	49.49%	47.66%	35.19%	44.97%

Table 15: Simulation results of the extended LPCC solutions.

Solutions provided by **F** and **M** improve the initial LPCC solution: the percentage of the emergency calls not served within the LA time decreases down to 16.78% whilst the low priority decreases down to 26.73%. A little bit surprisingly is the great improvement obtained by the **R** solution: emergency calls decreases down to 38.83% whilst non emergency down to 44.72%. Moreover, it is the only one solution comparable with the current location posts.

The above results suggest a way to improve the solution quality of LPCC adopting a two step optimization approach. The first step consists in the computation of the minimum number of posts. Then, the second step consists in adding the remaining available ambulances to new posts in such a way to maximize the number of non emergency demands assigned to them. The idea is simply to free the posts, computed at the first step, from serving low priority demands in order to improve their emergency demand coverage.

Performance analysis

Considering both the current location post (**A**) and the one provided by the **R** version of LPCC, we evaluate them varying the average speed, the number of ambulances and the use of smart ambulances.

The first experiment consists in the evaluation of the two solutions when the average speed changes. First we tested a decrement of the average speed to 20.8 km/h which means to evaluate what happens when traffic jam increases. Then we tested an increment to 30.8 km/h which represents the case in which the municipality arranges reserved lanes for ambulances on the main streets. Table 16 reports the simulation experiment results: on average, the extended LPCC solution seems to be a little more efficient than the current post location with respect to the speed variation.

		Jan 25	Feb 02	Mar 08	Apr 20	May 20	Jun 07	Sep 05	avg.
A 20.8	E	54.25%	64.53%	54.71%	48.63%	75.98%	66.03%	45.77%	58.56%
	nE	61.17%	88.10%	66.00%	56.56%	72.73%	72.66%	50.00%	66.74%
R 20.8	E	16.99%	28.49%	21.18%	17.81%	38.24%	26.28%	14.79%	23.40%
	nE	28.16%	47.62%	26.00%	31.15%	28.28%	29.69%	17.59%	29.78%
A 30.8	E	54.25%	61.05%	48.24%	50.68%	63.73%	60.90%	49.30%	55.45%
	nE	54.37%	89.68%	56.00%	59.02%	59.60%	69.53%	52.78%	63.00%
R 30.8	E	11.11%	28.49%	17.65%	14.38%	33.33%	21.79%	15.49%	20.32%
	nE	28.16%	46.03%	30.00%	27.87%	26.26%	28.91%	19.44%	29.52%

Table 16: Performances changing the average speed.

The second experiment is carried out by adding one ambulances but keeping the same number of posts and their location. The idea is to evaluate the impact of a new single ambulances.

		Jan 25	Feb 02	Mar 08	Apr 20	May 20	Jun 07	Sep 05	avg.
A+1	E	32.03%	44.77%	38.24%	31.51%	48.53%	48.72%	28.87%	38.95%
	nE	33.01%	57.94%	38.00%	35.25%	37.37%	47.66%	28.70%	39.70%
R+1	E	39.87%	54.07%	46.47%	43.15%	55.88%	46.79%	48.59%	47.83%
	nE	54.37%	60.32%	54.00%	45.08%	49.49%	53.13%	38.89%	50.75%

Table 17: Performances adding one new ambulance.

Table 17 reports the results of the experiment. We observe that in both cases the average performance of the system slightly decreases. Notice that to add a new ambulance, we need to impose that all posts can host two ambulances at the same time instead of just one. Since an ambulance is addressed to the closest free post location, during the simulation it could happens that some posts are uncovered whilst some other posts are covered by two ambulances. By consequence, it could happen that the overall coverage is unbalanced, i.e. ambulances are gathered in few posts. This experiment suggests the need to individuate a proper location for each new ambulance. Ideally, every time the number of ambulances increases, all the ambulance posts should be changed.

The third experiment consists in the evaluation of the smart ambulances. In the previous experiments, the operation center does not have the capability to assign a new mission to an ambulance which is coming back to a post. From a technological point of view, smart ambulances means the use of a Global Positioning System (GPS) receiver for each ambulance and a secure communication link between ambulances and the operation center.

		Jan 25	Feb 02	Mar 08	Apr 20	May 20	Jun 07	Sep 05	avg.
A	E	17.65%	28.49%	21.18%	26.71%	25.49%	26.28%	15.49%	23.04%
	nE	30.10%	50.00%	30.00%	23.77%	19.19%	35.94%	22.22%	30.17%
R	E	18.95%	33.72%	25.88%	23.97%	26.96%	23.72%	21.83%	25.01%
	nE	22.33%	48.41%	30.00%	29.51%	25.25%	35.94%	25.00%	30.92%

Table 18: Performances using smart ambulances.

Table 18 reports the results of the experiment. We observe a performance improvement with respect to the same solution without smart ambulances. This indicates the need of a more accurate use of the available ambulances when they are enroute to a location while they are not serving a call. For instance, the case in which an ambulance serving a low priority call can be reassigned to a new emergency call merits further investigation.

6 Conclusions

The emergency service is an important aspect in the life of every city and due to limited resources requires a careful management. Mathematical tools may be useful in dealing with such a problem. In this paper, the authors combine statistical analysis, linear programming model and simulation to solve the ambulance location problem, with the aim of providing a decision aided tool. The models have been tested on the case of Milano urban area. Computational results show that the approach might provide suggestions and guidelines, while keeping reasonable the needed computational time. The analysis of the available data point out the main problems of the system. Then integer linear programming models, both standard and tailored on the considered case, are used to derive new possible solutions. Such models cannot capture all the aspects of the problem, especially those related to the dynamic behavior of the system. Thus the obtained solutions behavior is tested, through simulation, over a real life situation. Moreover, the simulation model allows to evaluate how the performances of the emergency system can be improved adopting, for instance, smart ambulances. Currently, the results are subject of discussion with the emergency service management of Milano.

Acknowledgments

The authors wish to thank the Milano 118 Emergency Service Management for fruitful discussion and for providing us the data set and allowing their use in testing the developed approach. Besides, the author wish to thank the students Perego and Romantini for their help for the numerical results.

References

[1] L. Brotcorne, G. Laporte, and F. Semet. Ambulance location and relocation models. *European Journal of Operations Research*, 147:451–463, 2003.

[2] R. Church and C. ReVelle. The maximal covering locational problem. *Papers of the Regional Science Association*, 32:101–108, 1974.

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

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

[5] Goldberg, Dietrich, Chen, Mitwasi, Valenzuela, and Criss. A simulation model for evaluating a set of emergency vehicle base locations. *Socio-economic Planning Sciences*, 24:125–141, 1990.

[6] Goldberg, Dietrich, Chen, Mitwasi, Valenzuela, and Criss. Validating and applying a model for locating emergency medical vehicles in tucson, AZ. *European Journal of Operational Research*, 49:308–324, 1990.

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

[8] S.G. Henderson and A.J. Mason. Ambulance service planning: Simulation and data visualisation. In *Operations Research and Health Care: A Handbook of Methods and Applications*, volume 70 of *International Series in Operations Research & Management Science*, chapter 4, pages 77–102. Springer New York, 2004.

[9] K. Hogan and C.S. ReVelle. Concepts and applications of backup coverage. *Managment Science*, 34:1434–1444, 1986.

[10] Ilog. *Ilog CPLEX 8.0. User's Manual*, July 2002.

[11] R.C. Larson. A hypercube queueing model for facility location and redistricting in urban emergency service. *Computers and Operations REsearch*, 1:67–75, 1974.

[12] R.C. Larson. Approximating the performance of urban emergency service systems. *Operation Research*, 23:845–868, 1975.

[13] M. Mandell. Covering models for two-tiered emergency medical services systems. *Location Science*, 6(1):355–368, 1998.

[14] XJ technologies. *AnyLogic 5.5 user's manual*, 2005.

[15] P. Terna, R. Boero, M. Morini, and M. Sonnessa. *Simulazione, modelli ad agenti e scienze sociali*. il Mulino, 2006.

[16] C.R. Toregas, R. Swain, C.S. ReVelle, and L.Bergman. The location of emergency service facilities. *Operations Research*, 19:1363–1373, 1971.