Dynamic Coordination of Ambulances for Emergency Medical Assistance Services

Holger Billhardt[a], Marin Lujak[a], Vicente Sánchez-Brunete[b], Alberto Fernández[a], Sascha Ossowski[a]

[a] *CETINIA, Universidad Rey Juan Carlos, Calle Tulipán s/n, 28933 Móstoles, Madrid, Spain*

[b] *Medical Emergency Service SUMMA 112, Comunidad de Madrid, C./ Antracita 2 Bis. 2, 28045 Madrid, Spain*

Abstract

The main objective of emergency medical assistance (EMA) services is to attend patients with sudden diseases at any possible location within an area of influence. This usually consists in providing "in situ" assistance and, if necessary, the transport of the patient to a medical centre. The potential of such systems to reduce mortality is directly related to the travel times of ambulances to emergency patients. An efficient coordination of the ambulance fleet of an EMA service is crucial for reducing the average travel times. In this paper we propose mechanisms that dynamically improve the allocation of ambulances to patients as well as the redeployment of available ambulances in the region under consideration. We test these mechanisms in different experiments using historical data from the EMA service of the Autonomous Region of Madrid in Spain: *SUMMA112*. The results empirically confirm that our proposal reduces the average response times of EMA services significantly.

Keywords: coordination of multiagent systems; emergency medical assistance systems; efficient fleet management.

*Corresponding author: Holger Billhardt (tel. +34 916647458

Email addresses: holger.billhardt@urjc.es (Holger Billhardt), marin.lujak@urjc.es (Marin Lujak), vicente.sanchezbrunete@salud.madrid.org (Vicente Sánchez-Brunete), alberto.fernandez@urjc.es (Alberto Fernández), sascha.ossowski@urjc.es (Sascha Ossowski)

1. Introduction

The domain of medical assistance in general, and in emergency situations in particular, includes many tasks that require flexible on-demand negotiation, initiation, coordination, information exchange and supervision among different involved entities (e.g., ambulances, emergency centres, hospitals, patients, physicians, etc.). Among such tasks the coordination of the available resources to provide assistance to emergency patients as fast as possible is of crucial importance for obtaining an efficient service. The main goal here is to improve one of the key performance indicators: the response time (time between a patient call and the moment an ambulance arrives and the patient can receive medical assistance).

There is a general understanding that shorter response times are an essential starting point to improve care and reduce mortality [1, 2]. This holds especially for severe injuries. In order to assure the quality of emergency services many countries and regions specify response time limits for EMA service provider organisations either by law or through contractual norms. In Europe and in the Unites States such limits usually lie between 8 and 15 minutes. In the UK, for instance, a national standard defines that at least 75 per cent of Category A (immediately life-threatening) calls should be responded within 8 minutes.

Even though response time standards are often fixed by norms, they can be rather considered as targets that EMA service providers continuously try to reach.

One way to reduce response time consists in reducing the part that depends on the logistic aspects of an EMA service: the travel or arrival times of ambulances to emergency patients. There are two main problems EMA managers are faced with in the logistic part of a service: allocation and redeployment of ambulances. The allocation problem consists in determining an ambulance that should be sent to assist a given patient. Redeployment consists in relocating available ambulances in the region of influence in a way that new patients can be assisted in the shortest time possible.

In this article we present a novel coordination model for ambulance fleets that combines a mechanism for dynamic redeployment of available ambulances with a dynamic allocation of ambulances to patients. Regarding ambulance redeployment, we propose a method, based on centroidal Voronoi tessellations, that tends to optimize the allocation of idle ambulances in each moment with respect to the probability distribution of possible emergency

cases. Regarding ambulance allocation, we propose to use a dynamic auction-based assignment of patients to ambulances that tends to optimize the sum of the expected arrival times in each particular moment. EMA services are highly dynamic; e.g., new emergency patients will have to be attended and previous missions will finish. We present an event-driven system that dynamically executes ambulance assignment and redeployment and, thus, continuously tends to optimize the situation of the ambulance fleet with regard to the changes that occur in an EMA service.

The outline of the rest of the paper is as follows. Section 2 presents related work and relates our approach to others. In Section 3 we provide a brief description of the operation of EMA services. Then we present our ambulance allocation and redeployment mechanisms, and we expose an event-driven architecture for employing both mechanisms dynamically in real time. In Section 4 we present an experimental evaluation of our proposal and compare it with the operation strategies currently used by SUMMA112, the EMA service provider organisation in the Autonomous Region of Madrid in Spain[1]. The experiments have been carried out in a simulated environment and using real patient data from Madrid. Finally, Section 5 gives some conclusions and points out some aspects of our current and future research.

2. Related work

There have been many proposals for the problem of coordinating ambulance fleets for EMA services. Brotcorne et al. provides a good review of ambulance allocation and redeployment strategies from the early 1970s through 2003 [3]. More recent reviews have been done by Li at all. [4] and Aboueljinane et al. [5]. Whereas the former concentrates on covering models and optimization techniques for facility location, the latter analyses the use of simulation models in emergency medical service operations.

To the best of our knowledge, most of the work has been dedicated to the redeployment or coverage problem, e.g., the optimal location of ambulances in an region such that all points can be reached within a predefined time standard. Early approaches concentrate on a static distribution of ambulances. The Location Set Covering Problem, proposed by Toregas et al. [6] tries to find the minimum number of emergency facilities and their locations to cover

[1] www.madrid.org/cs/Satellite?pagename=SUMMA112/Page/S112_home

all demand assuming that the demand occurs at a finite set of points. In the Maximal Covering Location Problem proposed by Church and Re Velle [7], the aim is to locate a fixed number of facilities in order to maximize the population covered within some service distance. Such static methods do not take into account the relation between mobility and coverage of ambulances. In particular, demand points will be uncovered if one or more ambulances are called for service. To overcome this problem, researchers have proposed to maximize the coverage of demand points by more than one ambulance or by using double standards for coverage (e.g., [8, 9]). Another trend has been to establish probabilistic models, that explicitly model the availability or the travel and assistance times of ambulances (e.g., [10, 11]).

More recent research on the covering problem has concentrated on the dynamic location of ambulances, where methods are proposed to redeploy ambulances during the operation of a service in order to take into account the intrinsic dynamism of EMA services. In [12], Gendreau et al. extend their Double Standard Model to reflect the dynamic nature of the problem. They propose to use tabu search heuristics and solve the model through a (non-exhaustive) pre-computation of redeployment scenarios. In [13], the same authors propose the Maximal Expected Coverage Relocation Problem and present a strategy for dynamically relocating idle ambulances that are located in low demand areas. Rajagopalan et al. [14] developed another dynamic model for redeploying ambulances to predictable demand fluctuations in time and space. The objective of the model is to determine the minimum number of ambulances and their locations for each time cluster in which significant changes in demand patterns might occur while meeting coverage requirement with a predetermined reliability. Whereas the previous methods require solving integer programs, in [15], Maxwell et al. propose to use an approximate dynamic programming approach for ambulance redeployment. To deal with the high-dimensional state space in the dynamic program, they construct approximations to the value function that are formulated in terms of the percentage of calls that are reached within a time standard. Naoum-Sawaya and Elhedhli [16] present a two-stage stochastic optimization model that minimizes ambulance relocations while maintaining acceptable service level. While many approaches are based on centralized optimization, the solution approach of Ibri et al. in [17] is decentralized. The authors propose a multi-agent system that integrates a dynamic ambulance dispatching and redeployment method. However, there are a couple of drawbacks to this method. To limit deviations of vehicles, they allow assigning a vehicle to

another call only if this latter has higher priority than the first one, thus, not leaving space for a real dynamic optimization of (current) travel times. Secondly, the vehicles are represented (and grouped) by the station agents and the redeployment is performed among a fixed number of stations, which in the case of an insufficient number and/or position of stations, may result in insufficient coverage.

Most proposals on dynamic redeployment of ambulances (like the ones mentioned before) only consider the possibility to relocate ambulances among different, predefined sites (stations). This requirement is relaxed in the work proposed by Andersson and Varbrand [18]. These authors propose a decision support tool that recommends the redeployment of a fixed number of ambulances to areas with less preparedness (a criteria for coverage) and ambulances can be relocated to any place in the region.

Regarding dispatching strategies for ambulances (the patient allocation problem), there has been less research that treats this problem explicitly. Most works use the "nearest available ambulance" rule for assigning ambulances to patients in a first-came first-served manner. Some works analyse priority dispatching strategies. For instance in [19], Baranda et al. analyse dispatching strategies that take into account the severity level of patients and evaluate the survival probability of patients for different strategies. Also in [18], for calls with the highest priority the vehicle with the shortest travel time is assigned, whereas for less severe patients a vehicle is dispatched that reaches the patient in a given time limit but harms less some coverage criterion. López et al. [20] propose a multiagent system where ambulances are also assigned based on the severity of the patients. Besides the distance, the system also takes into account a trust value, that reflects the belief that an ambulance can fulfil its obligations in time. In this sense, expert drivers will have higher trust values than novice drivers. A more complex approach is presented by Haghani et al. [21], where the system dynamically optimizes the total travel time (ambulances to patients, ambulances to base stations and ambulances to hospitals).

Our redeployment approach differs from others in the sense that we do not try to maximize the zones in a region that are covered with respect to some time limits. Instead, we use an approach based on geometric optimization [22] that tends to optimize in each moment the positions of all ambulances that are still available such that the expected arrival time to potential new emergency patients is minimized. Using centroidal Voronoi tessellations, that are scalable with the number of agents in the network [23], we compute

optimal ambulance positions dynamically in real time. The latter takes into account the probability distribution of emergency cases in the region (at different times of the day), based on historical data. Furthermore, in our approach, ambulances can be redeployed to any point in the region and all idle ambulances (and not only a limited number) are dynamically redeployed whenever the changes in the system indicate that a better allocation may exist.

With regard to the allocation of patients to ambulances, our approach is similar to the one proposed by Haghani et al. [21] We also propose a dynamic approach. However, instead of optimizing the global travel times of all ambulance movements (including transfers to hospitals or base stations) as they do, we concentrate only on the sum of the arrival times of ambulances to the pending emergency patients. We use assignment based on computational optimization auctions to minimize this sum in each particular moment. In practice, in the case of severe patients (immediately life-threatening) it is this arrival time that is often crucial for saving lives. In our work we concentrate only on severe patients and, thus, we explicitly do not treat the problem of priority dispatching.

3. Coordination Model for Ambulances

EMA services are based on flexible and complex interactions between people playing different roles in diverse contexts of high responsibility. Even though EMA services might have different ways of operation, there are some main lines of emergency management common to all of them. The assistance procedure typically starts when a patient calls an Emergency Coordination Centre asking for assistance. The call is received by an operator who gathers initial data from the patient. The operator, possibly with the help of a physician, assigns one of several levels of severity to incoming calls. These levels are directly related to the priority that should be given to each emergency patient. There exist several different triage systems, for both, pre-hospital emergency medical services and emergency departments at hospitals [24]. In the case of Madrid, SUMMA112 employs its own system containing four levels of severity: level zero, urgent life-threatening calls; level one, urgent but not life-threatening calls; level two, less urgent calls, and level three representing non-urgent calls. According to the evaluation of the severity of a call, a specific type of ambulances is assigned, taking into account their availability, distance, and the estimated time to reach the patient. EMA ser-

vices typically work with at least two types of ambulances: basic life support (BLS) and advanced life support (ALS) units; where the latter are normally assigned to the most severe patients. When the ambulance arrives at the patient's location, the crew provides first aid and in some cases "in situ" assistance. According to the conditions of the patient, he/she is transported or not to hospital.

In the following two subsections we present our coordination model for EMA services based on dynamic allocation and redeployment. We concentrate only on the assistance of the most severe patients, with advanced life support units, that is, we do not consider the problem of priority dispatching or dispatching of different types of ambulance.

We use the following notation to describe the problem and to present our solution. The set of ambulances of an EMA service is denoted by $A = \{a_1, \ldots, a_n\}$ where n is the cardinality of A. Each ambulance has a position and an operational state which vary during time. $pos(a_i)$ and $state(a_i)$ denote the current position and the current state of ambulance a_i, respectively. The position refers to a geographical location and the state can be one of the following:

- *assigned*: An ambulance that has been assigned to a new patient and is moving to the patients location.
- *occupied*: An ambulance that is occupied either attending a patient "in situ" or transferring him/her to a hospital.
- *idle*: An ambulance that has no mission in this moment. The ambulance is either waiting at a base station for a new mission, or returning to its base from a previous mission.

Regarding the patients, $P = \{p_1, \ldots, p_m\}$ denotes the current set of patients that have to be attended and are waiting for an ambulance, where m is the cardinality of P. Each patient $p_j \in P$ has a position (denoted by $pos(p_j)$). We assume that patients do not move while they are waiting for an ambulance. That is, $pos(p_j)$ is constant meanwhile the patient p_j belongs to the set of unattended patients. Once the attendance of a patient starts, after the arrival of an ambulance, the patient is removed from the set of unattended patients P.

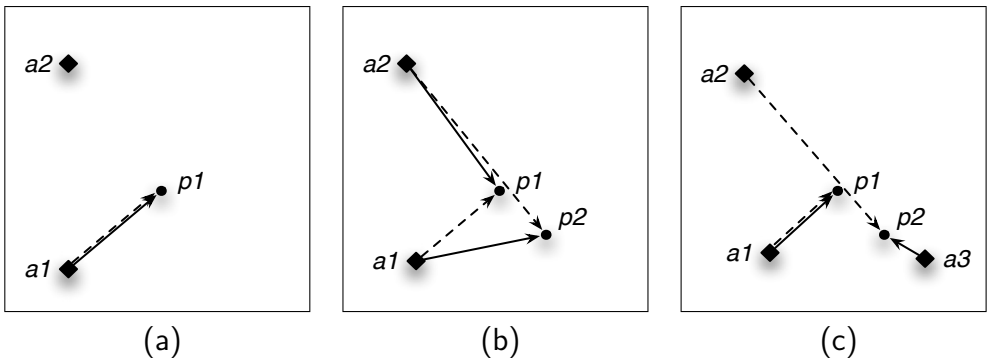

Figure 1: Example of ambulance assignment in dynamic environments. Solid lines represent the optimal solution and dotted lines assignment obtained with the fixed FCFS assignment strategy.

3.1. Dynamic Ambulance Allocation Mechanism

The ambulance allocation problem consists in finding an assignment of (available) ambulances to the emergency patients that have to be attended in each moment such that the expected arrival time of the ambulances to the patients' locations is as short as possible. In practice, many real-world EMA services, like in the case of SUMMA112 in Madrid, use a priority dispatching strategy. Patients with the highest severity level (in our case level-0 patients) are assigned to ambulances before other patients with a lower severity level. Furthermore, for level-0 patients, a first-call first-served (FCFS) rule applies, e.g., the patients that called first are also assigned first to an ambulance. Normally the closest available ALS unit is chosen, where an ambulance is considered available if it is either waiting at a station or it is on the way to a station after finishing an assistance mission. Once a level-0 patient has been assigned to an ambulance, this assignment is usually fixed, e.g., no reassignments take place. An exception is the case where a patient was assigned to a BLS unit (because no ALS vehicle was available) and is reassigned to a newly available ALS unit. We do not consider this latter case in this paper.

Considering the fact that several level-0 patients may have to be attended at the same time, this fixed FCFS approach is not always optimal from a global perspective. In particular, it does not always assure an assignment of ambulances to patients that minimizes the average arrival time. Fig. 1 presents an example to clarify this fact. In Fig.1a a patient p_1 has to be

attended and two ambulances a_1 and a_2 are available. a_1 is slightly closer to p_1 than a_2, and, thus, assigning a_1 to p_1 would be the (locally and globally) optimal choice. Now let's suppose that a new patient p_2 appears some moments later (Fig.1b). Also for this patient, a_1 is the closest ambulance, whereas a_2 is much further away. In this case, the globally optimal choice would be reassigning a_1 to p_2 and assigning a_2 to p_1. In this solution, the expected travel time to patient p_1 would be slightly worse, but at the benefit of a much shorter time required to reach patient p_2. That is, the average expected travel time would be reduced. Using the fixed FCFS assignment approach, p_2 would be assigned to a_2 (the dotted lines in the figure) and the overall solution would be worse. If we carry on with the example (Fig.1c), let's now suppose that a few moments later another ambulance, a_3, has finished a previous mission and is becoming available again. a_3 is very close to the location of patient p_2 (even closer than a_1). Thus, in this moment the optimal choice would be assigning a_3 to p_2 and a_1 again to p_1. a_2 would not need to be assigned to any patient. The solution obtained with the fixed FCFS method (dotted lines) is clearly worse when considering the average required arrival times.

The example indicates that in order to reduce the average arrival time in the dynamic environment of an EMA service, the assignment of ambulances to patients has to be recalculated whenever relevant events take place. Based on this idea, we propose a dynamic (re)assignment mechanism of ambulances to patients. In particular, whenever a new patient appears or an ambulance becomes available again after finishing a mission, we start the (re)assignment of all unattended patients to ambulances (including patients that have been already assigned, but where the ambulance did not yet reach the patient). This set of unattended patients is given by P. The set of available ambulances for assignment is the set $A_{av} = \{a_i \in A | state(a_i) \in \{assigned, idle\}\}$. That is, we consider all ambulances that are either idle or already assigned (but not attending a patient yet).

At a given moment in time an optimal assignment of ambulances to patients is a one-to-one relation between the sets A_{av} and P, that is, a set of pairs $AS = \{< a_k, p_l >, < a_s, p_q >, ...\}$ such that the ambulances and the patients are all distinct, and that fulfils the following conditions:

- The maximum possible number of patients is assigned to ambulances, that is:

$$\forall p_j \in P : \exists < a_i, p_j > \in AS \quad if \quad n \geq m$$
$$\forall a_i \in A_{av} : \exists < a_i, p_j > \in AS \quad if \quad n < m$$

- The total expected travel time of the ambulances to their assigned patients is minimized, that is:

$$\sum_{<a_i, p_j> \in AS} ETT(pos(a_i), pos(p_j)) \text{ is minimal}$$

  where $ETT(x, y)$ denotes the expected travel time for the fastest route from one geographical location x to another location y.

In our work we propose to calculate the optimal assignment of a set of ambulances to a set of patients with Bertsekas' auction algorithm [25, 26]. We use Bertsekas' optimization algorithm, instead of other methods (e.g. the Hungarian method [27]), because it has a naturally decentralized character. In the emergency medical assistance domain, this characteristic may be of interest since it may allow to accomplish the assignment of patients locally among different ambulances. However, in this paper, we do not analyse such a decentralized assignment approach.

Algorithm 1 (*getOptimalAssignment*) summarizes the adaptation of Bertsekas' algorithm to our problem. The general idea is that patients bid for the ambulances in an auction process. The input to the algorithm consists of the current set of available ambulances A_{av} and the current set of unattended patients P. First, the prices of all ambulances are initialized to 0 (lines 1 to 3) and the global assignment (AS) is initialized to the empty set. Then, the auction process starts (steps 5 to 18). In each iteration, a bidding and an assignment phase take place. During the bidding phase (lines 9 to 12), each patient p_j that is not currently assigned to any ambulance (not included in the global assignment AS) determines the ambulance a_i and a_k with the least cost ($c1$) and second least cost ($c2$), respectively. The cost of an ambulance a_s for patient p_j is computed as the expected travel time for a_s to reach patient p_j plus the current price of a_s. Then, patient p_j issues a bid for its best ambulance (a_i), where the bid value is the difference between the cost of the second best and the best ambulance for p_j plus a constant ϵ. The rationale behind this bid value is that, at the current prices and up to a price increment of $c2 - c1$ for ambulance a_i, patient p_j would prefer this ambulance with respect to its second choice (a_k). ϵ is a (positive) constant (the

Algorithm 1 *getOptimalAssignment* for ambulance assignment

Require: A_{av} - the set of available ambulances
Require: P - the set of unattended patients
1: **for all** $a_i \in A_{av}$ **do**
2: $price_{a_i} \leftarrow 0$
3: **end for**
4: $AS \leftarrow \emptyset$
5: **repeat** {Auction process}
6: **for all** $a_i \in A_{av}$ **do**
7: $Bids_{a_i} \leftarrow \emptyset$
8: **end for**
9: **for all** $p_j \in P$ with $< \_, p_j > \notin AS$ **do** {Bidding phase}
10: Determine the ambulances a_i and a_k with the least cost $c1$ and second least cost $c2$ for p_j, where:
$c1 = min_{a_s \in A_{av}}\{ETT(pos(a_s), pos(p_j)) + price_{a_s}\}$ and
$c2 = min_{a_s \in A_{av} \wedge a_s \neq a_i}\{ETT(pos(a_s), pos(p_j)) + price_{a_s}\}$.
11: Patient p_j issues a bid for ambulance a_i:
$Bids_{a_i} \leftarrow Bids_{a_i} \cup \{bid_{j,i}\}$, where $bid_{j,i} = c2 - c1 + \epsilon$.
12: **end for**
13: **for all** $a_i \in A_{av}$ with $Bids_{a_i} \neq \emptyset$ **do** {Assignment phase}
14: Determine the highest bid $bid_{j,i}$ for a_i:
$bid_{j,i} = max\{bid \in Bids_{a_i}\}$.
15: Assign ambulance a_i to the highest bidder p_j:
$AS \leftarrow AS - \{< a_i, \_ > \in AS\}$
$AS \leftarrow AS \cup \{< a_i, p_j >\}$
16: Increment the price for ambulance a_i:
$price_{a_i} \leftarrow price_{a_i} + bid_{j,i}$
17: **end for**
18: **until** $\forall p_j \in P : \exists < \_, p_j > \in AS$
19: **return** AS

minimum price increment), necessary to assure termination of the auction process. After all unassigned patients have issued their bids, the assignment phase takes place (lines 13 to 17). Each ambulance a_i that received a bid, is assigned to the patient p_j that issued the highest bid for that ambulance (line 15). If a_i was already assigned to another patient, it is deassigned previously. Finally, the price of a_i is incremented by the highest bid value. The

bidding and assignment phases are repeated until all patients are assigned to an ambulance.

The expected travel time between two locations (ETT) can be calculated using a normal route service. Furthermore, assuming ETT's to be integers (they can be scaled up to integers if necessary) and selecting $\epsilon < 1/m$, it is assured that the algorithm finds an optimal global assignment. In the worst case, where $|P| = |A_{av}|$ (e.g., the number of unattended patients is equal to the number of available ambulances), the algorithm has a pseudo-polynomial time complexity of $O(m^3C/\epsilon)$ if ϵ is kept constant and where $m = |P| = |A_{av}|$ and C denotes the maximum expected travel time between any ambulance/patient pair. This complexity can be reduced to $O(m^3 log(mC))$ for a particular implementation that uses ϵ-scaling [25].

Algorithm 1 is defined for the case where the number of available ambulances is greater or equal to the number of patients. If this is not the case, e.g. there are more patients than available ambulances, the roles of patients and ambulances in the auction process have to be changed. That is, in such a case ambulances bid for patients.

Let AS be the optimal global assignment of unattended patients to available ambulances calculated at a given moment t by algorithm 1. The dynamic nature of an EMA service implies that such an assignment may not be optimal at a later point in time t' ($t' > t$). The following situations have to be considered regarding AS:

1. One or more new patients require assistance: In this case, the set of patients that have to be attended changes and the current assignment AS may not be optimal any more.
2. Some ambulances that have been occupied at time t have finished their mission and are idle at time t': This implies that the set of available ambulances changes and a better assignment than the current one may exist.
3. None of the two situations above happens: It can be observed that in this case the assignment AS is still optimal at time t'. For simplicity, here we assume that the service for calculating the shortest expected travel times is consistent and ambulances always move with the velocity corresponding to the expected travel times. Furthermore, we exclude certain unforeseen events like ambulance break down, etc.
4. An ambulance a_i has reached the location of the assigned patient p_j: the pair $< a_i, p_j >$ can be eliminated from the assignment AS and if

Algorithm 2 *ambulanceReallocation* for ambulance (re)allocation

Require: AS - the current global assignment
Require: A - the set of ambulances
Require: P - the current set of patients to be attended
Require: E - set of received events
1: **for all** $ambOccupiedEvent(a_i) \in E$ **do**
2: $AS \leftarrow AS - \{< a_r, p_q > \in AS | a_r = a_i\}$
3: **end for**
4: $AS_{old} \leftarrow AS$
5: **if** $\exists\, newPatientEvent(p_j) \in E$ or
$\exists\, ambFinishedEvent(a_i) \in E$ **then**
6: $A_{av} \leftarrow \{a_i \in A | state(a_i) \in \{assigned, idle\}\}$
7: $AS \leftarrow getOptimalAssignment(A_{av}, P)$ \\execution of Algorithm 1
8: **else**
9: $AS \leftarrow AS_{old}$
10: **end if**
11: **return** AS

none of the conditions in 1 and 2 take place, the resulting assignment is still optimal at time t'.

Based on this analysis, we define algorithm 2 (*ambulanceReallocation*) to dynamically re-calculate the global assignment AS whenever a better solution may exist. In particular, the algorithm is executed whenever any of the following events are received:

- $newPatientEvent(p_j)$: a new patient (p_j) has to be attended.
- $ambFinishedEvent(a_i)$: an ambulance (a_i) has finished a patient assistance mission and has changed its state from *occupied* to *idle*.
- $ambOccupiedEvent(a_i)$: an ambulance (a_i) that was assigned to a patient has reached the patient's location. Thus it has changed its state from *assigned* to *occupied*.

The algorithm *ambulanceReallocation* assures that the assignment AS is the optimal assignment at each moment in time. It first eliminates pairs $< a_i, p_j >$ for all ambulances a_i that have reached the assigned patients (line 1 to 3). Then, if a new patient has appeared or if an ambulance has become

idle after finishing a mission (line 5), it recalculates the global assignment AS by calling algorithm 1 with the current set of unattended patients P and the current set of available ambulances A_{av} (line 7). If there are no new patients nor new available ambulances, the global assignment does not need to be recalculated since it is still optimal (line 9).

3.2. Dynamic Ambulance Redeployment Mechanism

Besides an optimal allocation of ambulances to patients, the average arrival time of an EMA service can be reduced by efficiently deploying idle ambulances in the region of interest. In particular, the deployment objective is to place ambulances such that the expected travel time to appearing future emergency patients is minimized. We tackle this problem by using Voronoi tessellations [28, 29].

A Voronoi tessellation (or Voronoi diagram) is a partition of a space into a number of regions based on a set of sites such that for each site there will be a corresponding region. Each region consists of all points in the space that are closer to the site of the region than to any other site. Formally, in a two dimensional space, let $\Omega \subseteq \mathbb{R}^2$ denote a bounded space and let $S = \{s_1, ..., s_k\}$ denote a set of sites in Ω. The Voronoi region V_i corresponding to the site s_i is defined by

$$V_i = \{y \in \Omega : |y - s_i| < |y - s_j| \text{ for } \ j = 1, ...k, j \neq i\}$$

where $|\cdot|$ denotes the Euclidean norm. The set $V(S) = \{V_1, ..., V_k\}$ with $\bigcup_{i=1}^{k} V_i = \Omega$ is a Voronoi tessellation of S in Ω.

Of special interest are centroidal Voronoi tessellations. A centroidal Voronoi tessellation is a Voronoi tessellation that has the property that each site s_i is itself the mass centroid of its Voronoi region w.r.t. some positive density function ρ. That is, for each s_i it holds

$$s_i = \frac{\int_{y \in V_i} y\rho(y)dy}{\int_{y \in V_i} \rho(y)dy}$$

A centroidal Voronoi tessellation $V(S)$ is a necessary condition for minimizing the cost function

$$\mathcal{F}(S) = \sum_{V_i \in V(S)} \int_{y \in V_i} \rho(y)|y - s_i|^2 dy \tag{1}$$

For a detailed description of centroidal Voronoi tessellations, the interested reader is referred to [28, 29].

Applied to our application scenario, S represents the positions of idle ambulances and ρ is a function that should reflect the density of predicted future emergency cases, then minimizing (1) is a reasonable approximation for minimizing the expected distance and, thus, the arrival time, to future emergency patients. As mentioned in section 2, most other works on ambulance redeployment try to optimize the allocation of ambulances to fixed bases or to predefined zones. In contrast, our approach based on centroidal Voronoi tessellations is a geometric optimization technique. That means, ambulances can be positioned at any point in the region and not only to a set of fixed places. This allows more freedom for better ambulance allocations. Furthermore, reasonable approximations of centroidal Voronoi tessellations can be calculated very fast, what allows for a dynamic recalculation of ambulance positions whenever the current situation has changed, as we propose in this paper.

A common approach to calculate centroidal Voronoi tessellations and, thus, to minimize $\mathcal{F}$ is the algorithm proposed by Lloyd [30]. The algorithm is an iterative gradient descent method that finds a new set of points S that minimizes $\mathcal{F}$ in each iteration and converges to a local optimum. Lloyd's algorithm performs the following steps:

1. Select an initial set S of k sites in Ω
2. Generate the Voronoi tessellation $V(S)$
3. Compute the mass centroids of all Voronoi regions in $V(S)$ w.r.t. the density function ρ. These centroids compose the new set of points S.
4. If some termination criteria is fulfilled, finish; otherwise return to step 2.

With regard to the density function ρ, in our work this function represents the forecasting of the positions of future emergency cases. In particular, we divide the region of interest in a set of equally sized cells $C = \{c_1, ..., c_u\}$, where u is the cardinality of C. Then, we estimate for each cell c_i the conditional probability that an emergency patient will appear in that cell, given that a new emergency case happens. We denote the probabilities by p_{c_i} and it holds that $\sum_{c_i \in C} p_{c_i} = 1$. p_{c_i} can be obtained by tracking historical data on emergency cases. The historical emergency occurrence dynamics can have many attributes like the season of the year (summer, fall, winter, spring),

the day of the weak and the hour of the day, etc. Based on these types of attributes and on the geographical coordinates, different emergency estimation models can be obtained for different situations. In our experiments we defined a different estimation model based on historical data and for each particular day and hour of the day.

Algorithm 3 presents the adaptation of Lloyd's method to our scenario. The objective of the algorithm is to find positions of all currently idle ambulances such that the response time for future emergency patients is minimized. Ω denotes the two-dimensional geographical region in which the EMA service operates and C is a partition of Ω in equally sized cells as described before.

The algorithm returns the recommended positions of idle ambulances (set AP) that have been obtained after a fixed number of iterations ($maxIterations$). It starts with the current real positions of all ambulances ($pos(a_i)$) and iteratively encounters new positions (pos_i).

We use the Euclidean norm as a distance measure to generate the Voronoi regions for the ambulances. While in a real traffic scenario, as it is our case, the Euclidean distance is a rather imprecise approximation of real distances on the road network, from a global perspective, and assuming a rather homogeneous connection between different points of the region of interest (as it is usually the case in many urban areas), the Euclidean norm seems to work reasonably well for our purposes. Furthermore, using "road-network distances" when calculating the Voronoi regions is a rather complicated task that would increase the computation complexity considerably.

We use the probability distribution $\{p_{c_1}, ..., p_{c_u}\}$ to compute the mass centroid of each Voronoi region V_i. We estimate the centroid of V_i as the weighted average of the centre coordinates of the cells (denoted by y_{c_j}) belonging to V_i, weighted by the proportion of the probability density of each cell that corresponds to the region V_i (denoted by $prop(p_{c_j}, V_i)$). In the calculation we consider all those cells that at least partially belong to V_i, e.g., cells c_j for which $c_j \bigcap V_i \neq \emptyset$ holds. If a cell belongs to more than one Voronoi region, its probability density is distributed proportionally.

Obviously, the size of the cells influences the precision of the calculation of the Voronoi tessellation and the solution to the minimization problem. Smaller cell sizes will lead to more precise results, whereas bigger cell sizes result in less computational costs.

In general, Lloyd's algorithm is not assured to find the global minimum for the cost function. Moreover, if the density function is discrete, as it is

Algorithm 3 *ambulanceRedeployment* for calculating positions of idle ambulances

Require: A_{idle} - the current set of idle ambulances
Require: $\{p_1, ..., p_u\}$ - a probability distribution over the cells in C

for all $a_i \in A_{idle}$ **do**
  $pos_i \leftarrow pos(a_i)$
end for
for $i = 1$ to $maxIteration$ **do**
  for all $a_i \in A_{idle}$ **do**
    Generate the Voronoi region V_i in Ω corresponding to the ambulance's position pos_i
  end for
  for all $a_i \in A_{idle}$ **do**
    Compute the mass centroid s_i of V_i:

$$s_i \leftarrow \frac{\sum_{c_j \in V_i} y_{c_j} prop(p_{c_j}, V_i)}{\sum_{c_j \in V_i} prop(p_{c_j}, V_i)}$$

    $pos_i \leftarrow s_i$
  end for
end for
$AP \leftarrow \emptyset$
for all $a_i \in A_{idle}$ **do**
  $AP \leftarrow AP \bigcup \{< a_i, pos_i >\}$
end for
return AP

in our case, after a certain point, the iterations may oscillate between different local minima. However, the algorithm finds good solutions very fast – after a few iterations. To illustrate this fact, we analysed the convergence of algorithm 3 (*ambulanceRedeployment*) in our particular setting of allocating ambulances in a geographical area – a rectangle of 125 133 kilometres that covers the region of Madrid. The area is split into cells of about 1300 1300 meters (the set $C = \{c_1, ..., c_u\}$) and the probabilities p_{c_i} correspond to the probabilities of emergency patients appearing Mondays between 9:00 and 10:00 (obtained from statistical data for the year 2009). We executed algorithm 3 100 times with different randomly chosen initial positions of 29

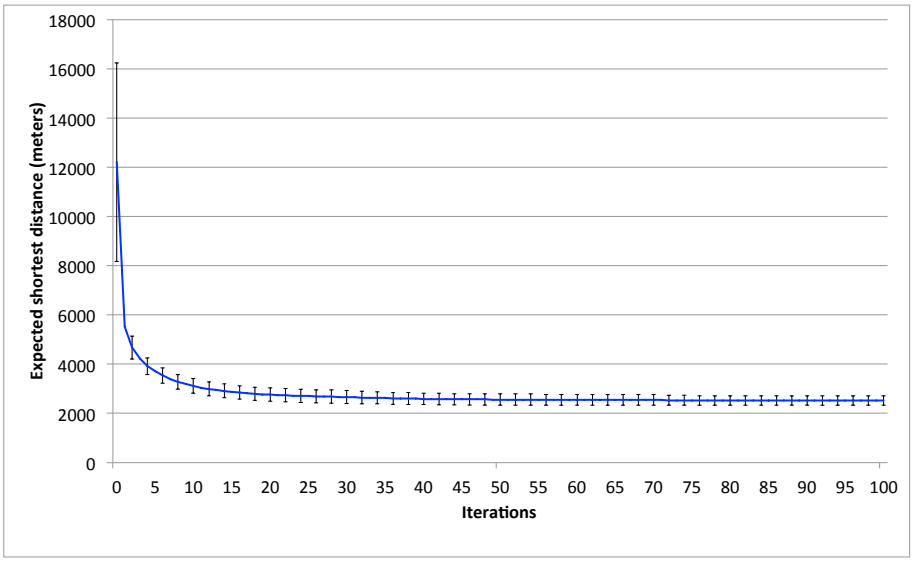

Figure 2: Improvement of expected shortest distance over iterations of the *ambulanceRedeployment* algorithm.

ambulances (corresponding to the ALS units used by SUMMA112). In each execution, we measure the evolution of the expected (Euclidean) distance of the closest ambulance to possible upcoming emergency patients (following the given probability distribution) corresponding to the positions calculated in each iteration of the algorithm.

Figure 2 presents the results, averaged over the 100 executions. The error bars reflect the standard deviation over the 100 executions. As this figure shows, the positions of ambulances improve very fast during the first 20 iterations and there is almost no improvement after 50 iterations. Therefore, we set $maxIterations$ to 50 in the experiments presented in section 4. This fast improvement suits well with our particular application, where we need to calculations ambulance positions in almost real time. Furthermore, the fact that the algorithm obtains only suboptimal solutions is not really critical. In fact, geographical and road network restrictions imply that it does not make sense to calculate the very optimal positions of ambulances. The latter is due to the fact that ambulances waiting for new missions can not be placed at every possible geographical location; they need an appropriate parking space, preferably at a location that provides good road connections to the surrounding area. In this sense, we are rather interested in finding the

approximate positions where idle ambulances should wait for new missions. The idea is that, once given such an area, the ambulance driver will decide which is the most appropriate waiting location in that area.

Similar to the ambulance allocation problem, the dynamic nature of an EMA service implies that the optimal positions of the idle ambulances may change when changes in the environment occur. In particular, the optimal positions may change if the set of idle ambulances changes (e.g., a previously occupied ambulance is becoming idle or an idle ambulance is assigned to assist a patient), or if the environmental situation implies a change in the probability distribution $\{p_1, ..., p_u\}$.

In order to cope with such changes, algorithm 3 has to re-calculate ambulance positions in a dynamic manner. In order to do that, it is executed whenever any of the following events occur:

- $ambIdleEvent(a_i)$: An ambulance (a_i) that was assigned to a patient has been deassigned. It has changed its state from *assigned* to *idle*
- $ambFinishedEvent(a_i)$: an ambulance (a_i) has finished a patient assistance mission and has changed its state from *occupied* to *idle*.
- $ambAssignedEvent(a_i)$: An ambulance (a_i) that was *idle* has been requested to assist a patient. It changes its state from *idle* to *assigned*.
- $probDistChangeEvent()$: A different probability distribution has to be used to compute ambulance positions.

3.3. Event-Driven Architecture

Our coordination model is dynamic in the sense that ambulance allocation and redeployment are revised in real time whenever changes in the system may imply the existence of a better assignment or redeployment strategy. As explained before, such changes are captured by events, which trigger the execution of the allocation and redeployment mechanisms (algorithms 2 and 3). Fig. 3 presents the global architecture and summarizes the event-driven dynamic nature of our approach.

The architecture contains two basic layers. The top layer contains the ambulances, modelled as agents. The bottom layer represents the EMA coordination centre. It includes a fleet coordination module and possibly other components that are necessary for the normal operation of EMA services (e.g., components for monitoring, call management, call operators, etc.). In

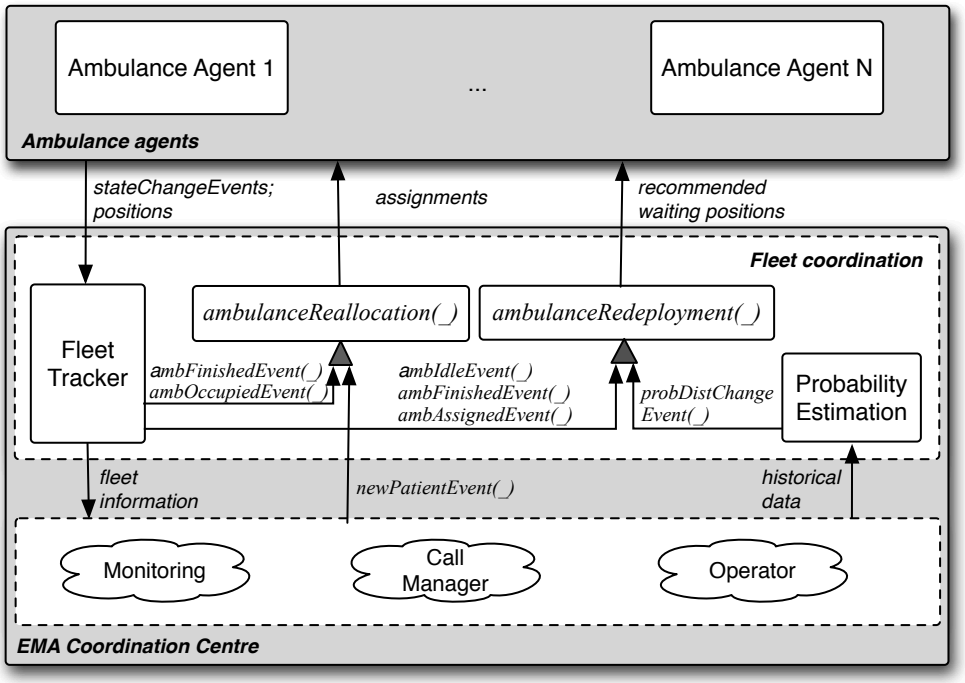

Figure 3: Event-driven architecture for EMA ambulance coordination.

the fleet coordination module contains a *Fleet Tracker* keeps track of the current operational state and the positions of ambulances. We assume that ambulances send their current positions on a regular basis and inform about any changes in their operational states ($stateChanegEvents$).

The ambulance allocation mechanism (algorithm 2) is executed either through the *Fleet Tracker*, if an $ambFinishedEvent$ or an $ambOccupiedEvent$ has been received, or if a call operator has generated a $newPatientEvent$. The algorithm recalculates the optimal assignment for all pending patients and taking into account all idle and already assigned ambulances. If the optimal assignments change, the affected ambulances are informed about such changes. In particular, an ambulance a_i has to be informed in the following situations:

- a_i was assigned to patient p_j and it is now assigned to a different patient p_k
- a_i was *idle* and is now assigned to patient p_j; The state of a_i changes

from *idle* to *assigned*.

- a_i was assigned to patient p_j and it is deassigned from p_j. The state of a_i changes from *assigned* to *idle*.

In the latter two cases, the ambulances would generate new events – *ambAssignedEvent* and *ambIdleEvent*, respectively –, that would trigger a new execution of the *ambulanceRedeployment* algorithm.

The ambulance redeployment mechanism (algorithm 3) is executed through the *Fleet Tracker* if an *ambIdleEvent*, *ambFinishedEvent* or an *ambAssignedEvent* has been received, or if the *Probability Estimation Module* issues an *proDistChangeEvent*. The latter module, uses past patient data to maintain estimations of the probability distribution of emergency patients for different situations (e.g., different days or hours). It generates an *proDistChangeEvent* if a different probability distribution of emergency patients should be used for ambulance redeployment. When executed, algorithm 3 recalculates recommended waiting positions for all idle ambulances and, if those positions have changed, the affected ambulances are informed.

In practice, the redeployment mechanism may be executed quite frequently what leads to rather continuous changes in the recommended waiting positions of ambulances. In order to avoid very small and almost continuous movements, we establish that ambulances should not move, if the new recommended position is within a certain threshold U. In our experiments, we set $U = 300$ meters.

4. Empirical Evaluation

In order to evaluate the effectiveness of our ambulance fleet coordination model we tested it in different experiments simulating the operation of SUMMA112, the EMA service provider organisation in the Autonomous Region of Madrid in Spain. For the experiments we used a simulation tool that allows for a semi-realistic simulation of intervals of times of normal operation of an EMA service. The tool reproduces the whole process of attending emergency patients, from their appearance and communication with the emergency centre, the schedule of an ambulance, the "in situ" attendance and, finally, the transfer of the patients to hospitals. The appearance of new patients can either be generated randomly, or by using a file with historical patient data. The simulator operates in a synchronized manner based on a step-wise execution, with a step frequency of 5 seconds. That is, every 5

seconds, the activities of all agents are reproduced leading to a new global state of the system.

In the experiments we are mainly interested in analysing the movements of ambulances and the subsequent arrival times to the patients. The movements are simulated using direction services. In particular, we use Mapquest's Open Directions API Web Service to reproduce semi-realistic movements on the actual road network with a velocity adapted to the type of road. External factors, like traffic conditions or others, are ignored. In fact, such factors are not so relevant for ambulances, as the fact shows that our simulated travel times are similar to the real travel times observed by SUMMA112. In the simulations, based on information provided by EMA experts, other time intervals that play a role in the attendance process are simplified as follows:

- The duration of the phone call between a patient and the emergency centre is set to 2 minutes.
- The time for attending a patient"in situ", after an ambulance has arrived at his/her location, is set to 15 minutes.

4.1. Experimental Setup

As the area of consideration we used a rectangle of 125 × 133 kilometres that covers the whole region of Madrid. We used 29 hospitals (all located at their real positions) and 29 ambulances with advanced life support (as currently used by SUMMA112).

We simulate the operation of the service for different days (24 hour periods) with real patient data from 2009. In particular, data of the most severe patients from 10 different days have been selected. Included is the day with the highest number of severe patients in 2009 – the 21 of January (221 patients) – and the day with the lowest number of such patients – the 17 of October (96 patients). The rest of the days were chosen to have a representation of high, medium and low work loads. The total number of analysed patients on the 10 days is 1609.

We compare the performance of the following coordination models:

- **C–SUMMA112**: this is the current coordination model used by SUMMA112. Level-0 patients are assigned to the closest AVL ambulances using a fixed FCFS strategy. No re-assignment of AVL ambulances takes place. Furthermore, idle ambulances are positioned on fixed stations (at the hospitals) and return to their home station after

the completion of a mission. The distribution of the stations has been optimized based on population densities and on geographic parameters.

- **DAAM**: The dynamic ambulance allocation mechanism is used to dynamically assign and eventually re-assign patients to ambulances. Each ambulance maintains its (fixed) base station and returns to that station after it has completed a mission.

- **DARM**: Ambulances do not have a fixed base station. Instead, the dynamic ambulance redeployment mechanism is used to dynamically re-calculate adequate positions of available ambulances such that the expected travel time to new patients is minimized. The assignment of ambulances is done using the fixed FCFS strategy.

- **DAAM+DARM**: Both mechanisms, dynamic ambulance allocation and dynamic redeployment, are combined.

In the case of C–SUMMA112 and DAAM, the (fixed) base stations of the 29 ambulances are located at the 29 hospitals. In the case of DARM and DAAM+DARM, only the initial positions of ambulances are at the 29 hospitals. Afterwards the ambulances move to the positions recommended by algorithm 3. Furthermore, in these two models the centroidal Voronoi tessellation is calculated each time for 50 iterations.

In order to estimate the probability of appearance of patients in the DARM approach (as described in section 3.2) we split the region into a grid of cells of about 1300 $\times$ 1300 meters. The patient appearance probabilities are obtained from statistical data (patient data from the whole year 2009). A different probability distribution is calculated for each day of the week and each hour.

4.2. Comparing DAAM with C–SUMMA112

In the first set of experiments we analyse the effectiveness if the dynamic reallocation of ambulances (DAAM) in comparison to the C–SUMMA112 approach.

It should be noted that, in most cases, the DAAM approach provides exactly the same solution as a fixed FCFS assignment strategy (the closest available ambulance is assigned). As described in section 3.1, only in some occasions a different and necessarily better solution can be found, basically if more than one patient has to be attended at the same time or a newly

Day	21/01/09	28/05/09	2/07/09	30/09/09	5/10/09	
#patients	221	152	199	124	137	
#affected pat.	40	28	25	7	4	
C–SUMMA112	16:57	16:44	13:07	16:14	15:42	
DAAM	15:30	14:30	11:44	17:36	14:45	
Improvement	1:27 (8.5%)	2:14 (13.3%)	1:23 (10.6%)	-1:22 (-8.4%)	0:58 (6.1%)	
Day	6/10/09	17/10/09	25/10/09	16/11/09	30/11/09	10 days
#patients	175	96	160	144	201	1609
#affected pat.	21	0	16	11	16	168
C–SUMMA112	13:38	na	13:34	13:45	11:47	14:51
DAAM	13:10		12:57	11:37	10:27	13:34
Improvement	0:27 (3.4%)	na	0:37 (4.5%)	2:05 (15.5%)	1:20 (11.3%)	1:16 (8.6%)

Table 1: Comparison of average arrival times for patients affected by the DAAM mechanism.

available ambulance is closer to an already assigned patient than the assigned ambulance. In the case of the 10 analysed days, 168 out of the 1609 patients are affected by DAAM assignment. The rest of the patients have exactly the same arrival times in both approaches. Table 4.2 presents the average arrival times (in minutes) of the DAAM approach in comparison to C–SUMMA112 for the affected patients. In general, it can be observed that as the workload of the service (number of patients to be attended) increases, more patients are affected by the DAAM model. For instance, the highest number of affected patients is on the 12 of January (40 out of 221) and there are no patients affected on the 17 of October (with a total number of only 96 patients).

The average improvement of the arrival time for the patients affected by DAAM in comparison to C–SUMMA112 is 1 minute and 16 seconds (about 8.6 %). This shows that the DAAM approach does have a positive effect on the arrival times. It should be noted that this improvement has no extra cost and is obtained just through a better assignment of ambulances to patients. In fact, comparing the average distances ambulances have to cover during a 24 hour period, they are slightly lower with DAAM than with C–SUMMA112 (94 km vs. 95.8 km on the 10 analysed days).

The table also shows that there is one day for which DAAM provides

worse results than C–SUMMA112 (the 30 of September). Looking in detail at the 7 affected patients for this day, there is one patient that obtains a significantly worse arrival time in the DAAM approach (33:00 minutes versus 14:10 minutes). This particular patient is actually not directly affected by a reassignment in the DAAM approach. What happens is that prior to the appearance of this particular patient, DAAM assigns an ambulance that is fairly close to the location of the patient to another case (in order to optimize the global assignment at that particular moment). Thus the ambulance moves away and no other ambulance is sufficiently close when the patient appears. This can happen because, even though the DAAM approach produces an optimal assignment on a static snapshot of the problem, in the more dynamic scenario it just constitutes a (sophisticated) heuristic. However, as the results show, a significant improvement can be obtained on the average case of the affected patients.

Fig. 4 compares the distribution of arrival times (in minutes) of C–SUMMA112 and DAAM on the affected 168 patients. For each curve, the patients are ordered by increasing arrival time. The curves show that both distributions are similar but, in general, DAAM obtains a better behaviour. The exception is in the seven highest values, which are slightly worse with DAAM than with C–SUMMA112. We believe that this is just due to the stochastic nature of the appearance of patients.

As we mentioned before, the DAAM approach is in general more effective in higher workload situations, e.g., when the number of patients to be attended at the same time is high. Based on this observation, we where interested in analysing how the approach would perform on a rather extreme work load. To do that, we tested the approach on patient data for an artificially generated day with a much higher number of patients. In particular, we merged the patient data of four days (5/10/2009, 12/01/2009,16/11/2009 and 30/11/2009) into a data set for a single day. At the whole, the number of patients on this"extreme day" was 703. In the simulation with this day, 562 patients where affected by the DAAM approach as compared to C–SUMMA112. Only 141 patients had exactly the same arrival time. For the 562 patients, the average arrival time with C–SUMMA112 was 20:47 min and with DAAM 13:51 min. This is an improvement of the average arrival time for the affected patients of 6:56 min (about 33.4%). On the whole data set, considering all 703 patients, the improvement is still very high: 5:33 min (29.1%). Fig. 5 shows the distribution of the arrival times (in minutes) of C–SUMMA112 and DAAM on the affected 562 patients of this experiment. The

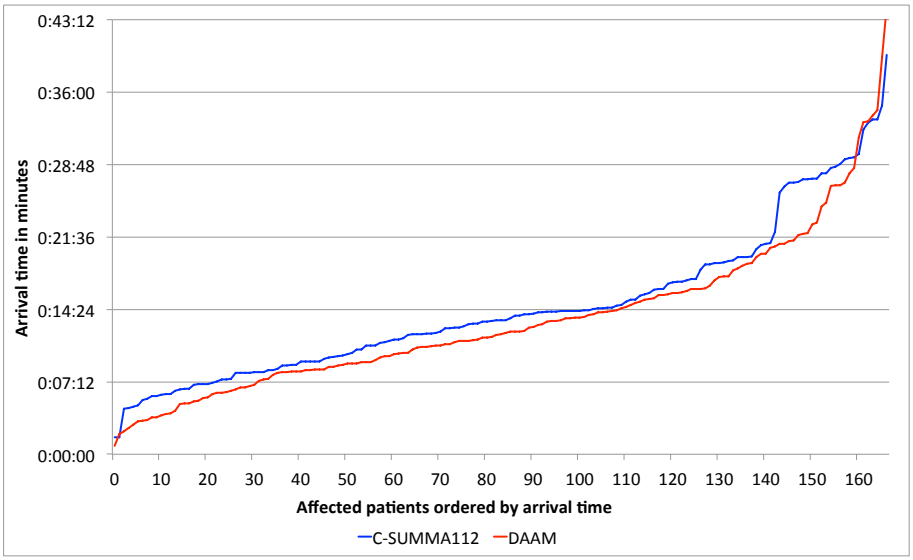

Figure 4: Comparison of the distribution of arrival times among the 168 patients affected by DAAM versus C–SUMMA112.

figure shows a clear improvement on all arrival time ranges and especially on the patients with higher arrival times.

4.3. Comparing DARM and DAAM+DARM with C–SUMMA112

In the next set of experiments we analysed the effect of the dynamic ambulance redeployment mechanism (DARM) and its combination with DAAM (DAAM+DARM) in comparison to C–SUMMA112. Table 4.3 presents the average arrival times (in minutes) obtained with these three models in simulations for the 10 selected days and over all 10 days (last column).

As the results presented in table 4.3 show, the use of the DARM approach provides a considerable improvement on the average arrival times for all 10 days with an average reduction of 1 minute and 41 seconds (about 14.4%). The best performance is obtained when both mechanisms are combined (DAAM+DARM), where the average reduction of the arrival time is on the 1609 patients is 1 minute and 52 seconds (about 15.8%). A reduction of the arrival times of this magnitude is considered by the managers from SUMMA112 as very significant.

Fig. 6 compares the distribution of arrival times for the different approaches for all 1609 patients of the 10 selected days. Again, the patients in each curve are ordered by increasing arrival time. Whereas the curves for

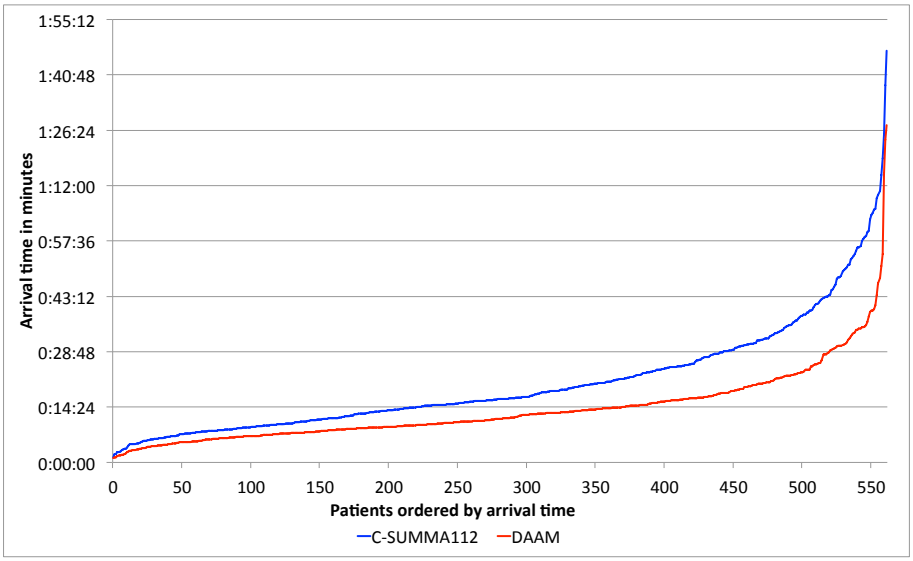

Figure 5: Comparison of the distribution of arrival times for the 562 affected patients on the "extreme work load day".

DARM and DAAM+DARM are very similar, a clear difference can be observed between both methods with respect to the current operation model of SUMMA112. The results are clearly better for almost all arrival time ranges. Furthermore, the most important improvements can be observed in the range of higher arrival times. This is a very positive effect for EMA services because it assures that more patients can be attended within given response time objectives. For example, out of the 1609 patients, 1163 (72.3%) are reached within 14 minutes with C–SUMMA112, whereas this number increases to 1356 patients (84.3 %) with DAAM+DARM.

One negative side effect of the dynamic ambulance redeployment approach is that available ambulances have to change their positions frequently in order to improve their locations regarding new potential emergency patients. Table 4.3 compares the average distances ambulances have to cover on each of the analysed days and on average in the C–SUMMA112 and the DAAM+DARM approaches. As it can be observed, the average distance increases approximately by a factor of 3. It will depend on each particular application whether the significant improvement regarding the arrival times justifies the associated higher operation costs of the DAAM+DARM approach.

Day	21/01/09	28/05/09	2/07/09	30/09/09	5/10/09	
# patients	221	152	199	124	137	
C–SUMMA112	11:43	11:52	11:03	11:23	11:50	
DARM	10:03	10:00	09:38	10:20	10:11	
Improvement	1:40 (14.2%)	1:52 (15.7%)	1:25 (12.8%)	1:03 (9.3%)	1:39 (14.0%)	
DAAM+DARM	09:46	09:50	09:29	09:39	10:09	
Improvement	1:57 (16.6%)	2:02 (17.1%)	1:33 (14.1%)	1:44 (15.3%)	1:41 (14.2%)	
Day	6/10/09	17/10/09	25/10/09	16/11/09	30/11/09	10 days
# patients	175	96	160	144	201	1609
C–SUMMA112	12:30	12:51	12:42	10:11	11:49	11:45
DARM	10:53	9:50	10:13	08:59	10:23	10:04
Improvement	1:37 (12.9%)	3:00 (23.4%)	2:29 (19.6%)	1:12 (11.8%)	1:26 (12.1%)	1:41 (14.4%)
DAAM+DARM	10:51	09:48	10:05	09:05	10:08	09:54
Improvement	1:38 (13.1%)	3:02 (23.7%)	2:38 (20.7%)	1:06 (10.8%)	1:41 (14.3%)	1:52 (15.8%)

Table 2: Comparison of average arrival times for 10 different days.

Day	21/01/09	28/05/09	2/07/09	30/09/09	5/10/09	
# patients	221	152	199	124	137	
C–SUMMA112	129.87	88.54	111.15	69.53	82.83	
DAAM+DARM	384.63	271.86	329.81	259.17	280.78	
Day	6/10/09	17/10/09	25/10/09	16/11/09	30/11/09	10 days
# patients	175	96	160	144	201	1609
C–SUMMA112	117.49	62.89	106.92	68.54	120.5	95.84
DAAM+DARM	312.2	201.09	317.24	265.98	376.92	299.97

Table 3: Comparison of average distances (in kilometres) ambulances have to cover on a 24 hour period.

5. Conclusions

This paper has reported on a piece of *applied research*, looking into the use of ICT for emergency medical assistance (EMA) services. In particular, we

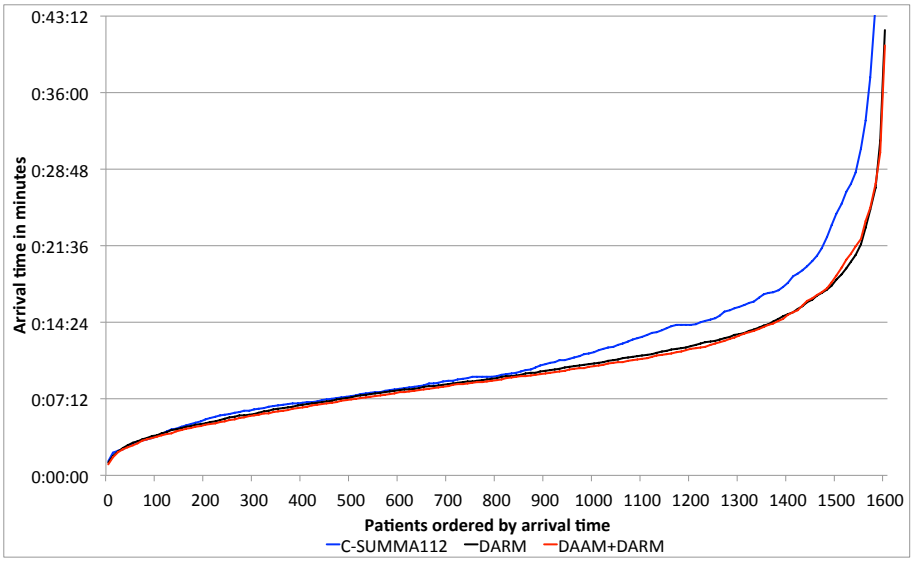

Figure 6: Comparison of the distribution of arrival times among the 1609 patients of the 10 analysed days.

have put forward a dynamic coordination model for ambulance fleets of EMA services, which combines an ambulance allocation mechanism (DAAM) with an ambulance redeployment mechanism (DARM). The DAAM is in charge of assigning available ambulances to patients that have to be attended, dynamically calculating the optimal assignment from a global perspective, minimising the sum of the expected travel times to all pending patients. Setting out from historical data, the DARM determines appropriate positions for available ambulances such that upcoming patients can be attended faster. We evaluated the coordination model in a simulated environment and compared it against an actually deployed strategy for a real-world scenario (patients, ambulances, ambulance and hospital positions) provided by SUMMA112, the EMA service provider in Madrid, Spain. The results empirically confirm significant improvements of arrival times over SUMMA112's current mode of operation, a fact that, especially for severe patients, can potentially be life-saving.

The main contributions of our work can be summarised as follows:

1. We have chosen two existing algorithms (Bertsekas auction and Lloyds algorithm for centroidal Voronoi tessellation) and instantiated them to the ambulance allocation and ambulance redeployment problems.

2. The resulting algorithms have been adapted so as to allow for their smooth integration, while taking into account the *dynamicity* of the ambulance fleet coordination problem. In particular, DAAM continuously optimises the assignment of ambulances to patients, and also supports the re-assignment of ambulances, an option that SUMMA112 has not considered so far. Furthermore, DARM uses a geometric method for *dynamic* ambulance redeployment, i.e. we calculate new (nearly) optimal ambulance positions for all currently available ambulances, whenever the fleet situation changes. To the best of our knowledge, this is a novel approach to the redeployment problem, insofar as ambulances can be positioned at any point in the region.
3. We have defined an event-based architecture and characterised a set of events that provide the "runtime glue" among ambulance allocation (DAAM) and redeployment algorithms (DARM). The software implementation of this architecture is the backbone of a knowledge-based system prototype that, combined with the simulation tool that we developed, allows us to perform realistic quantitative evaluations of the real-world scenario provided by SUMMA112.
4. The analysis of our approach is carried out with real data for different days and load situations provided by SUMMA112. Its performance is compared to the coordination model that is currently in use in Madrid. Especially in "high load" situations, when multiple patients have to be attended at the same time, an important improvement can be obtained by the use of DAAM – in the experiments on average about 8.6% on the affected patients. Furthermore, the over 14% improvement in the arrival times obtained with the DARM mechanism is certainly significant. On average, both approaches together manage to reduce arrival times by almost 16%.

A key assumption underlying our approach is that it is suitable and feasible to repeatedly determine optimal assignments of ambulances to patients in a dynamic scenario. The experimental results obtained confirm that, at least for the case of ambulance coordination in Madrid, this heuristic obtains good results. Furthermore, scalability is not a predominant issue for typical EMA scenarios, as the number of ambulances and hospitals, as well as patient arrival rate, are usually rather low. For the Madrid case (29 ambulances and less than 250 severe patients per day), for each trigger (event) our system showed runtimes of a few seconds.

Still, it must be acknowledged that, as it is based on Bertsekas' auction algorithm, our assignment method shows an asymptotic runtime of $O(m^3 log(mC))$ for a particular implementation that uses ϵ-scaling and where m available ambulances have to be assigned to m unattended patients. C is the maximum expected travel time between any ambulance and any patient [25]. To this respect, it should be noted that the Hungarian method, the fastest known method for solving the general assignment problem, shows a slightly better asymptotic runtime of $O(m^3)$ [27]. Nevertheless, we preferred using Bertsekas' method, firstly because of its distributed character, giving it the potential to support coordination for EMA services that are organised in a more decentralised manner; and secondly because it allows us to conceive the coordination process as a sequence of interactions ("auctions") among "agents" (ambulances and patients), making it easier to convey its basic functioning to stakeholders in the domain.

Furthermore, in order to effectively solve the ambulance assignment problem for the case of real-world EMA services based on the aforementioned heuristic, the ambulance travel times within a town's road network at a given moment must be known for all ambulance-patient pairs. This can be computed in $O(mE + mVloglogV)$ for directed graphs where m is again the number of unattended patients/available ambulances, V is the number of vertices in the graph (intersections in the road network), and E is the number of edges (roads segments) [31]. In our settings, E and V are by far greater than m. This means that the complexity of determining the routes is higher than the one of the proper assignment algorithm, and therefore the overall asymptotic time complexity would usually not be improved if we had chosen the Hungarian method.

Regarding ambulance redeployment, alternative approaches in the literature direct ambulances to a set of fixed base stations or to predefined zones, and tend to optimize the coverage of the region of interest (e.g., try to assure that every point can be reached within given time limits). By contrast, for DARM we selected Lloyds algorithm, firstly because it is a geometric optimization technique that allows locating ambulances at any point in the region; and secondly because it gives us a natural way to minimize the expected travel time to future emergency patients based on historical distributions of emergency incidents. Again, our empirical results have shown that redeployment based on DARM leads to significant performance improvements as mentioned above.

Whereas the DAAM mechanism has no direct additional cost, the DARM

approach, however, relies on rather frequent movements of available ambulances in order to adapt the coverage to the particular situation in each moment. This increases the distances an ambulance has to cover each day. In our experiments, for the case of Madrid, the distance grew from about 100 kilometres a day to about 300 kilometres. This increase implies higher operation and maintenance costs. It belongs to the realm of politics whether the obtained improvements compensate the higher costs. Nevertheless, it should be noted that a common way of reducing arrival times for EMA services is to increment the number of ambulances. In this sense, the coordination approach proposed in this paper may constitute a less expensive alternative.

Regarding our future work, we plan to improve the estimation of expected travel times by including information about other influencing factors, e.g., more fine-grained information on traffic and weather conditions. This can be done either using real-time information provided by web services, or by a statistical analysis of past ambulance missions. We also plan to look into the problem of event generation and low-level event processing. Currently, it is the ambulance crew who informs the coordination centre about changes in the operational state of their ambulance. Elsewhere [32] we reported on a preliminary approach for automatically detecting such changes based on different types of sensors combined with complex-event processing (CEP) software [33]. In the future, we plan to integrate the aforementioned approach to event-generation and low-level processing based on CEP with the higher-level, knowledge-based coordination approach reported in this article.

Acknowledgement

This work has been partially supported by the Spanish Ministry of Science and Innovation through the projects OVAMAH (grant TIN2009-13839-C03-02; co-funded by Plan E) and "AT" (grant CSD2007-0022; CONSOLIDER-INGENIO 2010), and by the Spanish Ministry of Economy and Competitiveness through the project iHAS (grant TIN2012-36586-C03-02). We also would like to thank the people from SUMMA112 for their help and for providing us historical data of emergency patients.

References

[1] R. Elvik, A. Hoye, T. Vaa, M. Sorensen, The Handbook of Road Safety Measures, second edition Edition, Emerald Group Publishing Limited, 2009.

[2] E. T. Wilde, Do emergency medical system response times matter for health outcomes?, Health Economics 22 (7) (2013) 790–806.

[3] L. Brotcorne, G. Laporte, F. Semet, Ambulance location and relocation models, European Journal of Operational Research 147 (3) (2003) 451–463.

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

[5] L. Aboueljinane, E. Sahin, Z. Jemai, A review on simulation models applied to emergency medical service operations, Computers & Industrial Engineering 66 (4) (2013) 734 – 750.

[6] C. Toregas, R. Swain, C. ReVelle, L. Bergman, The location of emergency service facilities, Operations Research 19 (6) (1971) 1363–1373.

[7] R. Church, C. ReVelle, The maximal covering location problem, Papers in Regional Science 32 (1) (1974) 101–118.

[8] K. Hogan, C. ReVelle, Concepts and applications of backup coverage, Management Science 32 (11) (1986) 1434–1444.

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

[10] M. S. Daskin, A maximum expected covering location model: Formulation, properties and heuristic solution, Transportation Science 17 (1) (1983) 48–70.

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

[12] M. Gendreau, G. Laporte, F. Semet, A dynamic model and parallel tabu search heuristic for real-time ambulance relocation, Parallel Computing 27 (12) (2001) 1641–1653.

[13] M. Gendreau, G. Laporte, F. Semet, The maximal expected coverage relocation problem for emergency vehicles, Journal of the Operational Research Society 57 (1) (2006) 22–28.

[14] H. K. Rajagopalan, C. Saydam, J. Xiao, A multiperiod set covering location model for dynamic redeployment of ambulances, Computers & Operations Research 35 (3) (2008) 814 – 826.

[15] M. S. Maxwell, M. Restrepo, S. G. Henderson, H. Topaloglu, Approximate dynamic programming for ambulance redeployment, INFORMS Journal on Computing 22 (2) (2010) 266–281.

[16] J. Naoum-Sawaya, S. Elhedhli, A stochastic optimization model for real-time ambulance redeployment, Computers & Operations Research 40 (8) (2013) 1972–1978.

[17] S. Ibri, M. Nourelfath, H. Drias, A multi-agent approach for integrated emergency vehicle dispatching and covering problem, Engineering Applications of Artificial Intelligence 25 (3) (2012) 554 – 565.

[18] T. Andersson, P. Varbrand, Decision support tools for ambulance dispatch and relocation, Journal of the Operational Research Society 58 (2) (2007) 195 – 201.

[19] D. Bandara, M. E. Mayorga, L. A. McLay, Priority dispatching strategies for EMS systems, Journal of the Operational Research Society, published on-line September 2013. doi:http://dx.doi.org/10.1057/jors.2013.95.

[20] B. López, B. Innocenti, D. Busquets, A multiagent system for coordinating ambulances for emergency medical services, Intelligent Systems, IEEE 23 (5) (2008) 50–57.

[21] A. Haghani, H. Hu, Q. Tian, An optimization model for real-time emergency vehicle dispatching and routing, in: 82nd annual meeting of the Transportation Research Board, Washington, DC., 2003.

[22] V. Boltyanski, H. Martini, V. Soltan, Geometric Methods and Optimization Problems, Vol. 4 of Combinatorial Optimization, Springer Verlag, 1999.

[23] F. Bullo, J. Cortés, S. Martínez, Distributed Control of Robotic Networks: A Mathematical Approach to Motion Coordination Algorithms, Princeton University Press, 2009.

[24] I. Lidal, H. Holte, G. Vist, Triage systems for pre-hospital emergency medical services - a systematic review, Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine 21 (1) (2013) 28.

[25] D. Bertsekas, The auction algorithm: a distributed relaxation method for the assignment problem, Annals of Operations Research 14 (1) (1988) 105–123.

[26] D. Bertsekas, Auction algorithms for network flow problems: A tutorial introduction, Computational Optimization and Applications 1 (1) (1992) 7–66.

[27] J. Munkres, Algorithms for the assignment and transportation problems, Journal of the Society for Industrial & Applied Mathematics 5 (1) (1957) 32–38.

[28] Q. Du, V. Faber, M. Gunzburger, Centroidal voronoi tessellations: applications and algorithms, SIAM Review 41 (4) (1999) 637–676.

[29] A. Okabe, B. Boots, K. Sugihara, S. N. Chiu, Spatial Tessellations: Concepts and Applications of Voronoi Diagrams, 2nd Edition, Probability and Statistics, Wiley, NYC, 2000.

[30] S. Lloyd, Least squares quantization in PCM, IEEE Transactions on Information Theory 28 (2) (1982) 129–137.

[31] S. Pettie, A new approach to all-pairs shortest paths on real-weighted graphs, Theoretical Computer Science 312 (1) (2004) 47–74.

[32] H. Billhardt, R. Bruns, J. Dunkel, M. Lujak, S. Ossowski, Intelligent event processing for emergency medical assistance, in: Proceedings of the 29th Annual ACM Symposium on Applied Computing ACM-SAC2014, Gyeongju, Korea, 2014, pp. 200–206.

[33] L. D., The Power of Events: An Introduction to Complex Event Processing in Distributed Enterprise Systems, Addison-Wesley, 2002.