BETA WORKING PAPER

Urban emergency medical service: dynamic model for dynamic cities

Marco Amorim[1*], Sara Ferreira[1], António Couto[1]

[1]*Research Centre for Territory, Transports and Environment, University of Porto, Faculty of Engineering, Porto, Portugal*

[*]Rua Dr. Roberto Frias, s/n, 4200-465 Porto – Portugal. E-mail address: mraul@fe.up.pt

Urban emergency medical service: dynamic model for dynamic cities

This paper presents a methodology to locate ambulance base stations and allocate ambulances using optimization and simulation models. The models allow us to better understand how city dynamics affect an urban emergency medical service response system (uEMS).

The methodology incorporates two steps. The first step uses scenario-based optimization and survival function theory to locate ambulance base stations, whereas the second step uses agent-based simulation to allocate ambulances to stations. The proposed models are tested for different situations and periods in the city of Porto.

The results of the sensitivity analysis of the models show the relevance of understanding the dynamics of cities and how they impact uEMS response systems. Useful insights regarding the number of stations and the average response time are addressed together with the minimum number of stations and ambulances required for different maximum response time limits and different survival coefficients.

Keywords: emergency medical service; scenario optimization; simulation; dynamic EMS; dynamic ambulance location

Introduction

Motivation and contribution

Post-crash response is pillar 5 of the WHO (2011) global plan for the decade of action for road safety 2011-2020. The post-crash response is divided into several activities; the last one, Activity 7, explicitly encourages research and development into improving post-crash response, pointing to the improvement of the response of emergency medical services.

Some researchers have worked to create models for planning emergency medical services, EMS, solely to assist road crashes in a city (Kepaptsoglou, Karlaftis, and Mintsis 2012) or in specific road networks (Zhu, Kim, and Chang 2012). However, emergency medical services usually respond to all types of medical emergencies, and no separate service may exist to assist just one type of medical emergence. Clearly, one can argue that there are moral issues regarding having emergency resources that cannot assist in a certain emergency because these resources are strictly allocated to other types of emergencies.

In recent works, the focus has been on dynamic EMS where ambulances are dynamically allocated, dispatched or routed to better prepare for the upcoming hours (Vasić et al. 2014; Zhang 2012; Panahi and Delavar 2009), accounting for the fact that traffic varies.

The aim of this work is to propose a methodology for an adequate EMS response plan in urban areas with the premise that land use dictates population dynamics, and population dynamics justify a dynamic service. More specifically, the present work argues that the locations of people and traffic are not static in an urban environment (Lam et al. 2015; Vasić et al. 2014), and these two variables are the most relevant ones when designing an urban EMS response plan. Whereas people in constant movement represent a possible dynamic demand (Krishnan, Marla, and Yue 2016; Wang et al. 2015), traffic represents the network load, which on one hand constrains how quickly an emergency vehicle can reach a medical emergency (Erkut et al. 2009; Kim 2016; Ingolfsson, Budge, and Erkut 2008; Budge, Ingolfsson, and Zerom 2010; Westgate et al. 2013) and on other hand is correlated with traffic accidents and injuries (Ferreira and Couto 2013).

To assess how city dynamics interfere with EMS, we propose an EMS response plan that consists of a long-term ambulance base station location scheme and a mid-term ambulance dynamic allocation plan. The response plan is contained in a time cycle. Within a cycle, the system must response to a dynamic city where people are in movement and traffic conditions vary. In sum, this framework provides an EMS response that is more flexible to fitting to the different changes occurring in urban areas enforced by the city's specific land use.

This work contributes to the literature in the following ways:

- Formalizing a methodology to plan a long-term EMS response solution prepared for a dynamic mid-term ambulance allocation planning;
- Objectively maximizing survival by accounting for urban dynamics, thereby implementing a scenario-based optimization model;

- Proposing a mid-term ambulance allocation dynamic planning assessment using simulation and dispatching heuristics;
- Presenting a case study where the proposed methodology and models are implemented;
- Assessing the impact of different EMS response and coverage parameters and the overall system performance.

EMS response models

The foundational stream of research for emergency response traces back to the year 1955 with the fire station location planning studies of Valinsky (1955). Additionally, Hogg (1968) together with Savas (1969) filled the base archetypes for this theme. However, the two most relevant works, which actually fermented the OR community interest in EMR, were those of Toregas et al. (1971) and Church and Velle (1974). Toregas et al. (1971) present a solution to solve the location set covering problem (LSCP), making sure all demand is covered within a maximum time or distance radius. Church and Velle (1974) note a solution for a maximal coverage location problem (MCLP) that attempts to overcome the resource limitations of the problem of Toregas et al. (1971).

The classical interpretation of the facility location problem was soon surpassed by uncertainty approaches, leading to double coverage, scenario approaches, stochastic and robust optimization problems and dynamic locations.

Focusing on the fact that once a facility is called for service demand, points under its coverage are no longer covered, Daskin and Stern (1981); (1983) and Hogan and ReVelle (1986); (1989) account for the business probability and reliability of facilities. The former solves the maximum expected covering location problem (MEXCLP), and the latter moves forward to a maximum availability location problem (MALP).

Maxwell, Henderson, and Topaloglu (2009) classified research on dynamic allocation problems into three categories depending on the following: the model is solved in real time each time a redeployment decision is to be made (Brotcorne, Laporte, and Semet 2003; Kolesar and Walker 1974; Gendreau, Laporte, and Semet 2001; Nair and Miller-Hooks 2006), solving the model involves computing optimal ambulance positions for every number of available ambulances via an integer programming formulation in an offline preparatory phase (Ingolfsson 2006; Gendreau, Laporte, and Semet 2005); or if one intends to incorporate system randomness into the model by modelling the problem as a Markov decision process (Berman 1981a; Berman 1981c, 1981b; Zhang, Mason, and Philpott 2008; Alanis, Ingolfsson, and Kolfal 2013; Berman and Odoni 1982; Jarvis 1981) or make decisions under particular system configurations (Andersson and Varbrand 2006; Andersson 2005).

In fact, when addressing dynamic location models, the bibliography tends to show its relation with multi-period location models, where time is discrete, which are much more useful than single period models, where time is continuous. This is proven by Miller et al. (2007) and supported by Boloori Arabani and Farahani (2012).

The concept of scenario-based approaches is also used when uncertainty is present. Serra and Marianov (1998) solved the p-median problem (PMP) under scenario-based demand uncertainty. When the number of facilities, or ambulances, is uncertain, Current, Ratick, and ReVelle (1998) propose a scenario-based approach and solve the problem with a general-purpose mixed integer programming (MIP) solver.

Moreover, with the advance of computer power and the availability of powerful personal computers, simulations have become a useful tool for researchers wanting to formulate more realistic and complex models, be it to assess solutions or to support optimization models (Restrepo, Henderson, and Topaloglu 2008; Maxwell et al. 2010; Yue, Marla, and Krishnan 2012; McCormack and Coates 2015; Iannoni, Morabito, and Saydam 2009; Su and Shih 2003). McCormack and Coates (2015) show that the simulation enhances the level of realism of EMS models, making it applicable to complex, real-life systems when proper data exist.

Nevertheless, in urban Emergency Medical Services (uEMS), contrary to non-emergency facility location problems, underestimated or overestimated solutions not only have a monetary impact but also carry a social impact, and a bad decision can lead to, e.g., higher response times, which may seriously reduce the survival probability of the victims to be rescued. For instance, Sánchez-Mangas et al. (2010) indicate that a reduction of 10 minutes in the emergency response time could result in a 30% reduction of traffic accident fatalities. Although this number can vary depending on many factors, it is obvious that a quicker medical response will result in improved medical assistance

(Blackwell and Kaufman 2002; Pons et al. 2005). In conformity, Erkut, Ingolfsson, and Erdogan (2008) note that the EMS response research direction is to substitute the covering concept with concepts that account for survival probabilities and for the heterogeneity of the victims. This type of concept has already been used in recent works (Knight, Harper, and Smith 2012; McCormack and Coates 2015).

McCormack and Coates (2015) prove the possibility of increasing cardiac arrest victims' survival without the need of additional resources; however, the proposed model accounts for only two types of medical emergencies: cardiac arrests and other types. In contrast, Kepaptsoglou, Karlaftis, and Mintsis (2012) focused their work on a uEMS model for the special case of road crashes disregarding other types medical emergencies. Knight, Harper, and Smith (2012) address the heterogeneity of medical emergencies in a more direct way. They propose a Maximal Expected Survival Location Model for Heterogeneous Patients, where a decaying survival function is used for cardiac arrests and step functions for other types of medical emergencies. Further, a weight parcel is added to capture emergency type priority

As a final remark, we highlight the three key techniques that were identified as the proper ones to be used in the pursuit of our claim. These are, first, scenario-based optimization to better achieve a station location plan for a dynamic environment; second, the use of an EMS simulation model to assess ambulance needs in a dynamic fashion; and third, victims' heterogeneity and the corresponding survival functions to provide a more realistic and better suited response to heterogeneous medical emergencies.

Content of the paper

In this work, we address the Emergency Medical Service, EMS, with a focus on the Urban Service, uEMS, and investigate city dynamics and how they might influence uEMS planning with the objective to cover and maximize the survival probability.

The paper is divided into an introductory chapter with the motivation, EMS models and this sub-chapter. Then, a methodological approach chapter follows, where the optimization and simulation models are presented. A real base case, the city of Porto, is presented afterwards in the application of the model chapter. The analysis of the results is offered afterwards in the discussion of the results chapter. The conclusion chapter finalizes the paper with the relevant conclusions and future developments.

Methodology

The present methodology intends to provide insight for a long-term plan and a mid-term plan of a uEMS response network composed of ambulance base stations and ambulances in a dynamic environment. The methodology is structured with the intention of offering an assessment of the impact of city dynamics - people movements and traffic changes - by developing a dynamic uEMS response system.

We view the uEMS response system as twofold: one, there is a long-term plan that defines a fixed number of ambulance base stations and their fixed location, and two, a mid-term plan defines the maximum number of ambulances required at each station during each period. To capture the cyclic dynamics of a city, we define a cycle as a time pattern that repeats over time and for which different time periods exist. The cycle captures the city routine in terms of population movements and traffic, whereas a period captures the static moment of a cycle.

We developed a two-step methodology to achieve the previously mentioned objectives. The first step of the proposed methodology entails a robust ambulance base stations location optimization model, whereas the second step entails a simulation model that computes the number of required ambulances to be allocated at each ambulance base station. The methodology uses a scenario-based approach where in each period, the traffic load and people location are distinct.

First step - Optimization Model

A system optimization requires a performance measure. In a uEMS response network, the literature shows that some of the most common measures of performance are coverage and system reliability. However, different types of emergencies have different requirements and priorities. Thus, the concept of maximum survival, first presented by Erkut, Ingolfsson, and Erdogan (2008), can measure the system in a more accurately fashion.

The performance P_i of a uEMS response to an event i of type k can be defined by a survival function that depends on the time between the event start and the arrival of the assistance team, r_i, as per equation (1). Therefore, a simple performance metric for an emergency system is the sum of all response performances. The average, minimum and mode could also be defined as an overall performance metric. Nevertheless, for simplicity and for this work, the sum of all uEMS response performances is the chosen metric to assess the system performance.

$$P_i = f^k\left(r_i\right) \tag{1}$$

We discussed that city behaves as a dynamic entity where traffic load and people location vary with time but repeat themselves in cycles. To meet our beliefs, a dynamic optimization model is essential to produce a solution that performs and adapts as best as possible through the system life time. As mentioned in the literature review, a possible way to implement such dynamic behaviour is to use a scenario-based approach. Scenario-based optimization is typically used to produce a robust solution that is prepared for different possible situations. This is usually the case where part of the model's inputs is unknown; thus the system designer predicts possible scenarios where a positive performance of the system is mandatory.

For us, the goal is slightly different. We know the system input, and each of our scenarios is a representation of a period from a defined cycle; thus our model aims to provide a solution that will perform as well as possible throughout the defined cycle.

This method allow us to create a static solution for an instant t_i, input $f(t_i)$, that varies with a cycle, C, of length T:

$f(t_0) = f(t_T) \neq f(t_1) \neq f(t_2) \neq \ldots \neq f(t_{T-1})$.

However, for short periods we assume a static behaviour. Thus, a cycle with length $p - 0 = T$ has a finite number of periods ($\#S_i = S$) so that $C = [S_0, S_1, \ldots, S_s]$ is a cycle C with periods S_0, S_1, …, and S_s, where S_0 is the period between 0 and n, S_1 is the period between n and m, S_s the period between l and p with $l > m$, thus:

$f(t_0) = f(t_1) = f(t_2) = \ldots = f(t_n) \neq f(t_{n+1}) = f(t_{n+2}) = \ldots = f(t_m) \neq \ldots \neq f(t_{l+1}) = f(t_{l+2}) = \ldots = f(t_p)$,

and $f(t_{p+a}) = f(t_0)$ with a as an infinitesimal.

The proposed model minimizes, in a cycle, the inverse of the performance metric by deciding where to locate ambulance base stations, *e*, and allocating the demand clusters nodes, *p*, in a nodal network:

$$\text{minimize} \sum_{s}\sum_{l}\sum_{p} y_{l,p} \times e_{s,p} \times \left[f^{k}\left(r_{s,l,p}\right)\right]^{-1}, \tag{2}$$

subjected to

$$\sum_{l} y_{l,p} \times r_{s,l,p} \le M_r \quad \forall p \text{ in } P, \forall s \text{ in } S \tag{3}$$

$$\sum_{l} y_{l,p} = 1 \;\; \forall p \text{ in } P \tag{4}$$

$$y_{l,p} \le x_l \quad \forall l \text{ in } L, \forall p \text{ in } P \tag{5}$$

$$\sum_{l} x_l \le M_l\,, \tag{6}$$

where

S is the set of periods s\{period 1, period 2, … , period s\} and $S = C$,

L is the set of possible ambulance base stations l\{station 1, station 2, … , station l\},

P is the set of demand cluster nodes p \{node 1, node 2, … , node p\},

$e_{s,p}$ is the number of events in demand cluster node p for period s,

$x_l = 1$ if an ambulance base station is located at l and 0 otherwise,

$y_{lp} = 1$ if ambulance base station l serves events at location p and 0 otherwise,

r_{slp} is the response time required for an ambulance located at l to arrive at p during s,

M_r is the maximum allowed response time, and

M_l is the maximum number of stations.

Equation (2) minimizes the sum of the inverse survival functions of each event occurring at each period of the defined cycle. If the survival function declines linearly with time, then the objective function is simply a minimization of the sum of the response times adorned by a constant specific for each type of medical emergency.

Equations (3), (4), (5) and (6) control the model properties. Equation (3) defines an upper bound for the response time. Decision variable y_{lp} and equation (4) are add to ensure that for every node, only one station is allocated. Finally, equation (5) ensures that if node p is served by a station at l, then a station must be located at l.

Furthermore, there is the problem in deciding how many stations should be deployed. This can be addressed if one can assess how much a new station is worth in terms of the gain in performance. Because this relation is not yet defined, equation (6) limits the number of stations, allowing the model to be run for different upper bounds.

To reduce the model size, there is a preparation step that merges sets L and P into an availability tuple a = {pair l^s p^s | *if* l can assist $p \forall s$ in S} from set $A = [a_1, a_2, \ldots, a_w]$. This transforms equations (3) and (4), respectively, into equations (7) and (8):

$$\sum_s \sum_l y_{l,p} \times r_{s,l,p} \le M_r \quad \forall a \text{ in } A \tag{7}$$

$$\sum_l y_{l,p} = 1 \ \ \forall p \text{ in } A \tag{8}$$

The preview modification implies a reduction in the number of decision variables as well as the number of constraints and sum parts of the objective function.

A limitation of this model can be identified. For different periods, each station is forced to serve the same nodes. To make the model more flexible and allow different stations to serve different nodes at different periods, decision variable $y_{l,p}$ should be transformed into $y_{s,l,p}$, which takes a value of 1 if during period s node p is served by a station located at l. Thus, the final model is represented by equation (9):

$$\text{minimize} \sum_s \sum_l \sum_p y_{s,l,p} \times e_{s,p} \times f^t\left(r_{s,l,p}\right)^{-1} \tag{9}$$

subjected to

$$\sum_l y_{s,l,p} \times r_{s,l,p} \le M_r \quad \forall\, p \text{ in } A, \forall\, s \text{ in } A \tag{10}$$

$$\sum_l y_{s,l,p} = 1 \ \ \forall\, p \text{ in } A, \forall\, s \text{ in } A \tag{11}$$

$$y_{s,l,p} \le x_l \quad \forall\, p \text{ in } A, \forall\, s \text{ in } A, \forall\, l \text{ in } L \tag{12}$$

$$\sum_{l} x_l \leq M_l \quad (13)$$

Although this increases the number of variables, it makes the model more flexible, leading to an increase in the performance of the final solution.

Second step – Simulation Model

Dynamic environments are in constant change; in cities, this translates into land demographic occupation changes, such as daily periods when people cluster in business and industrial areas, and night periods when people cluster in residential and nightlife areas. Moreover, when residential areas exist and cluster far from the business and industrial areas, traffic flow differs during commuting times; in the morning towards the city business, industrial and commercial areas and in the evening towards the residential areas.

With this principle, we propose a dynamic allocation of emergency medical vehicles, ambulances, in the network. The literature reviews several presented optimization models to allocate ambulances to base stations. Another option and straight forward solution is the use of simulation.

When studying ambulance allocation and dispatching, the assisting time is usually unknown. To cope with variables where their distribution is unknown or that vary in a random way, simulation allows us to introduce randomness in our model.

The main idea behind the proposed simulation model is to feed a simulated environment where a uEMS system exists with an infinite number of possible vehicles to be allocated to each station. Then, a record will keep track of how many vehicles are being used at each instant so that in the end, a statistic analysis can determine several ambulance indices detailed by station and period.

To simulate the system, an agent-based model is used, where an authority agent, the *city agent*, controls lower level agents: the *event agent*, road *network agent*, *ambulance agent*, and *node agent*. These *agents* coexist in an environment that simulates a spatial area defined by nodes, key locations, and a set of arcs connecting those nodes (Algorithm 1).

Algorithm 1 General simulation algorithm

Definitions:

T = simulation period

t = timestamp

j = step

j = 60 s

t = 0

While $t < T$

1. Update city
   - Sets the environment conditions, s, from possible status $S = \{s_1, s_2, ..., s_n\}$, where $s = f(time)$
   - Move events from events waiting list $E^w = \{e_1, e_2, ..., e_m\}$ to events active list E^a if the timestamp of event $e_m(t) < time$, and generate assisting time required, e_n^{atime}
2. For all ambulances in the network:
   - Ambulance time to destination, a_d^t, is updated → $a_d^t = a_d^t - j$
   - If $a_d^t = 0$ → transfer ambulance to destination
3. For all active events $e_n^a \in E^a$:
   - if no ambulance is allocated → run Ambulance dispatching algorithm, **Algorithm 2**
   - If ambulance is at the occurrence location → Update assisting timer, $e_n^{atime} = e_n^{atime} - j$
   - If $e_n^{atime} \leq 0$, assisting time ended → run Ambulance to hospital routing algorithm, **Algorithm 3**
4. Update nodes of type Hospital
   - If ambulance arrived → Transfer event to hospital
   - Ask network to return ambulance to its station → set new a_d^t
5. Update results dictionary, $R_{\{i\}\{j\}}$, with i = t and j = a^g
   - For all ambulances in the network → if not in original station, a^g, $R_{\{i\}\{j\}} = R_{\{i\}\{j\}} + 1$, with $i = t$ and $j = a^g$
6. If $t < T$ go back to 1.

Algorithm 2 Ambulance dispatching algorithm

Definitions:

Station $s_p \in S = \{s_1, s_2, ..., s_p\}$

$S_p^a = \{ s_p^1, s_p^2, ..., s_p^a\}$ is a list of ambulances parked at s_p

$C = \{t_v, t_{v+j}, ..., t_b\}$ is a set of timestamps t

e_m^{max} is the maximum allowed response time for e_m

$Time(s_p, e_m)_c$ is the minimum time travel between station s_p and e_m at scenario $c \in C = \{c_1, c_2, ..., c_m\}$

1. For all c *in* C: if t *in* $c \rightarrow q = c$
2. For all s_p *in* S: order S by $Time(s_p, e_m)_q$ in ascending order
3. For all s_p *in* S: if $S_p^a \neq \{\emptyset\}$ and $Time(s_p, e_m)_q \leq e_m^{max} \rightarrow a = s_p^1$, proceed to 5
4. Select $s_1 \rightarrow$ create s_1^1, $a = s_1^1$
5. Allocate a to e_m and return to **Algorithm 1**

Algorithm 3 Ambulance to hospital routing algorithm

Definitions:

Node of type Hospital $h_r \in H = \{h_1, h_2, \ldots, s_r\}$

$C = \{t_v, t_{v+j}, \ldots, t_b\}$ is a set of timestamps t

$Time(e_m , h_r)_c$ is the minimum time travel between e_m and h_r at scenario c from list $C = \{c_1, c_2, \ldots, c_m\}$

a is the ambulance allocated to e_m, and a^d is the destination of ambulance a

1. For all c *in* C: if t *in* $c \rightarrow q = c$
2. For all h_r *in* H: order H by $Time(e_m , h_r)_q$ in ascending order
3. Select $h_1 \rightarrow a^d = h$
4. Return to **Algorithm 1**

The *city agent* is responsible for generating and dispatching ambulances when required and activating the events at the right time. The city is also accountable for storing all other *agents* and gives update orders to them.

The *event agent* is responsible for feeding the *city agent* with events and informing the *city agent* of its current state, asking for an *ambulance agent* to be allocated when it is activated (Algorithm 2). When being assisted, the *event agent* is responsible for generating a random assisting time and when this time terminates it will request the *network* agent to be transported to the closest *node agent* of type hospital (Algorithm 3).

Algorithm 2 step 3 goes through a list of ordered stations and chooses the one with an inactive ambulance if the time between this station and the event is lower than the maximum time allowed to assist the event. When there is no available ambulance, step 4 creates a new ambulance at the station that requires the least amount of time to arrive at the event.

Algorithm 3 simply chooses the closest hospital (in terms of trip time) by ordering a vector of available hospitals, step 2, and then selecting the first member of the ordered vector, step 3.

The *network agent* is responsible for routing all *ambulance agents* and choosing the closest hospital when an *ambulance agent* is transporting an *event agent*. It is also responsible for computing the fastest real time OD route.

The *ambulance agent* keeps track of its position in the *network agent* and informs the network when it arrives at any destination. It travels to the node where the event occurs, assists the event, brings the event to the closest hospital and returns to its base. It is completely dependent on orders given by other agents.

The node agent has three types: node, hospital and station. This agent assists the network and city agents by storing ambulances and events.

Model application

The proposed methodology was tested with real data from Porto during the period between May 2012 and May 2013. The daily uEMS response network operation was divided into three periods of equal length: The morning period (6:00 am to 2:00 pm), the afternoon period (2:00 pm to 10:00 pm), and the night period (10:00 pm to 6:00 am). These periods are eight hours long, which is the usual daily working time across many countries. The network operation also differentiates weekdays (Monday through Friday) from weekend days (Saturday and Sunday). Accordingly, a total of 5 periods are formed: Period 1 Weekday 6 am to 2 pm, Period 2 Weekday 2 pm to 10 pm, Period 3 Weekday 10 pm to 6 am, Period 4 Weekend 6 am to 10 pm and Period 5 Weekend 10 pm to 6 am. The weekend morning and afternoon periods were joined together due to their similarities in terms of traffic conditions.

For the maximum response time, it is known, within reasonable simplifications, that without any sort of intervention, the survival rate of a cardiac arrest victim drops, linearly, to zero after 10 minutes (Eisenberg et al. 1990). Moreover, Valenzuela et al. (1997) indicate that the time interval needed for EMTs or paramedics to attach a defibrillator and clear the patient for defibrillation once CPR is in progress is estimated to be 2 minutes past EMT arrival or 1 minute past the time of initiation of CPR by EMTs. This leads to a threshold of 8 minutes for the medical team to arrive at the event scene. Cardiac arrests are assumed to be the most demanding type of medical emergency; thus, the maximum allowed response time for each node is assumed to be 8 minutes.

Whereas the influence of the response time to cardiac arrests is very well defined in the literature, other types of medical emergencies do not have such survival functions. To simplify, we assumed that every type of emergency survival function follows a linear law represented by a survival coefficient $(\alpha^t)^{-1}$ with inverse α^t. This transforms equation (9) into equation (14):

$$\text{minimize} \sum_s \sum_l \sum_p y_{l,p} \times e_{s,p} \times \alpha^k \times r_{s,l,p} \tag{14}$$

where K is the set of type of events k {cardiac arrest, car crash, others}

The types of events chosen are justified by two assumptions. The first is that cardiac arrests have the quickest response time requirement. The second is that car crashes have a direct impact in the network traffic. This has straightforward implications in the time travel to other events occurring in the meantime. Thus, there is an indirect effect of the rescue time of car crash victims and the survival rate of other types of emergencies. All the other types of emergencies are considered to cluster in groups of similar behaviour.

The events occurred in Porto were collected from the INEM (National Institute of Medical Emergency of Portugal) database containing information on the type of emergency, timestamp and address of the crash spot. There are a total of 33 736 events in a one-year period. The addresses were converted into coordinates using a python script that connects with the Google Maps API for geocoding. The care-assisting time on the crash scene of each event is unknown, so a uniform distribution between 1 and 30 minutes was assumed and picked at random for each event.

For the optimization model, the city network was converted into a nodal network where each node is the centroid of the city census sub zones, for a total of 87 nodes. Using a radial-distanced based cluster algorithm, each event was allocated to the corresponding node.

The ambulance base stations were assumed to be possibly located in any of the 87 nodes. Afterward, a python script was created to use the Google Directions API and calculate the OD matrix of time travels for the different periods. This script asks Google Directions API for the fastest travel time, by car, between two coordinates for the morning peak hour (8 am), the afternoon peak hour (6 pm), the weekend peak hour (3 pm), and the free flow speed travel time. The free flow speed times were allocated to the night periods.

All the data were processed and stored in an SQL Database using Python, SQLite3 and DB Browser for SQL. Later, the data were prepared to be used by the optimization and the simulation models. To reduce the number of calls to the SQL database, the time travel matrix and the availability set were compiled into python raw files, which reduce the data processing time when running the models. The two models were also programmed in Python. For the optimization model, the Gurobi Optimizer python library, a state-of-the-art math programming solver, was used.

The optimization model was run for this study case, followed by the simulation model. Sensitivity analysis was processed by changing the relevant parameters to understand their implications on the methodology formulation.

Results and discussion

Computing resources

To support our claims, we propose a through sensitive analysis regarding the spatial and temporal dynamics that may influence how the EMS system is planned.

With the optimization model, we tested the impacts of the maximum number of stations, M_l, the maximum response time, M_r, and the inverse survival coefficients per type of emergency, α^k. With the simulation model, we tested the impacts of different uEMS network configurations from the optimization model.

Each models' runs were computed on a machine with an intel quad core processor at 1.73GHz and 8GB of memory ram in a WIN10 64bits operative system. Python v.2.7.8 and Gurobi v.6.5.2 were used, both in 64 bits. The computing time for each model run was under 1 minute.

Stations location

In the first analysis, we test different values of M_l and M_r and from the produced results, we assess the impact of the number of stations on the average response time and the station network requirements for different thresholds of the maximum response time. Furthermore, we propose a base case that will serve as an overall solution for the presented optimization problem. We also assess the singular solutions for each proposed period and compare them with the overall solution, indicating the impact of dynamic cities on a uEMS response system. Finally, we assess the spatial dynamics of the emergency type and test the sensibility of the inverse survival coefficients per type of emergency, α^k, justifying the importance of considering the heterogeneity of medical emergencies.

Average response time

The optimization model was run for different threshold of ambulance base stations, M_l, with equals α^t.

The minimum number of stations for a feasible solution is 8. Figure 1 show these results, where the objective function result was converted into the average travel time.

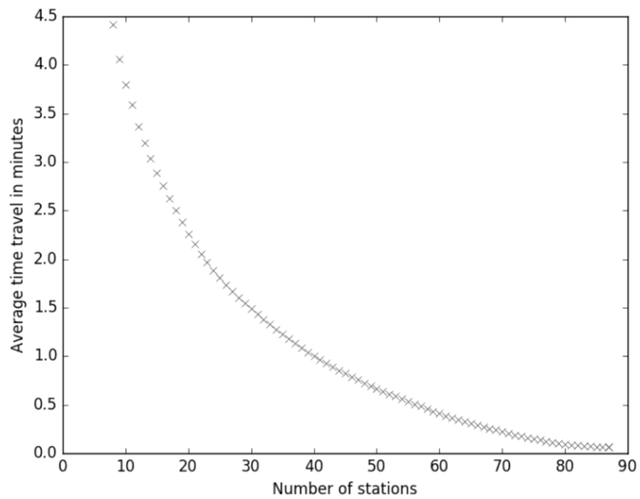

Figure 1. Average time travel for different number of implemented stations

As the number of stations increases, the average response time quickly drops in the first few additional stations and then slows down as the number of stations approach the number of nodes. It is important to remember that events were clustered into nodes; thus a station implemented in a certain node will respond to the events of that node instantly. It is also important to understand that the response time is only the driving time; it does not account for the time the emergency call is being processed and the time for the paramedic team to prepare the victim for any necessary intervention.

Moreover, Figure 1 shows an apparent correlation between the average response time and the number of stations implanted. To the naked eye, there seems to be a hyperbolic or exponential relationship between both variables. Nevertheless, it is important to note that as the number of station increases, theoretically, the average response time will never reach zero. In addition, when the average time grows, towards infinite, the number of stations required will be lower and lower but will never be null.

In a further analysis, we tested several types of fitting curves and several variable transformations to assess a possible law between the number of stations and the average travel time. Figure 2 groups the best-found relations.

The analysis assessment leads to the identification of two different correlations. One occurs in the first 7 observations, sample 1, and the other occurs in the remaining observations, sample 2. Undoubtedly, a power law explains the sample 1 correlation, whereas the sample 2 correlation is better described by an exponential law, or, if we transform $x \rightarrow 1/(x + 10)$ by a linear law.

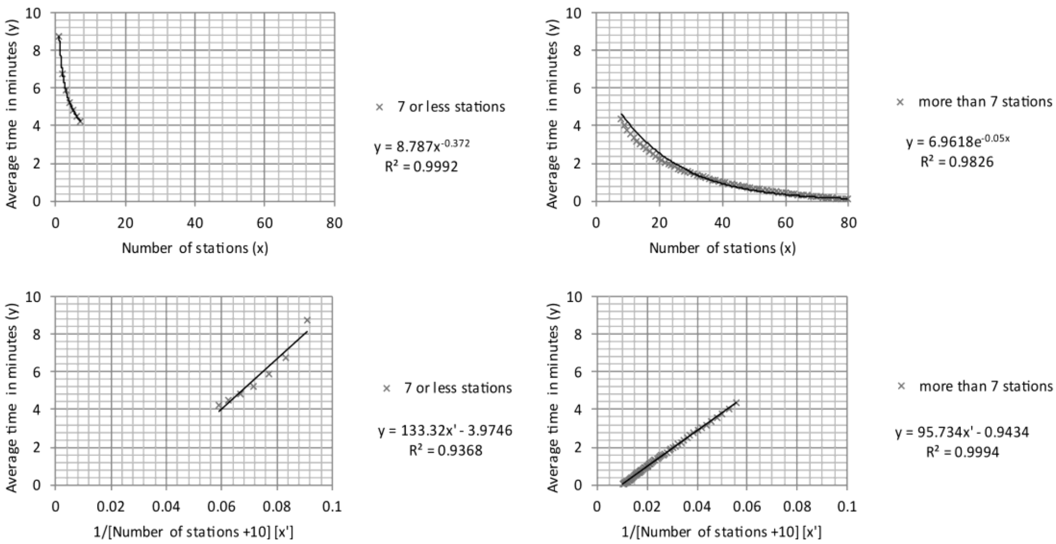

Figure 2. Correlation analysis between average time and number of stations

Clearly, there is a disruption at the 7th observation, corresponding to 7 stations implanted in the network. When adhering to the x transformation, the samples behave differently. The sample 1 average time drops more than 30% faster than sample 2 when $1/(n+10)$ decreases (number of stations increase), pointing to differences in the network behaviour at the macro scale (few stations try to support the whole network) and micro scale (many stations in the network, allowing each of them to focus in specific city areas) due to possible dynamic effects.

Maximum response time

The maximum response time is one of the key parameters in an EMS optimization system. The response time defines the quality of an EMS system; nevertheless, a shorter response time require more stations.

Figure 3 shows the decrease in required stations when the maximum response time is increased. For 5 minutes of the maximum response time, 24 stations are required, but as soon as this limit is extended by a half minute, the requirements drop to 18 stations. When increasing the time by one-third (from 5 minutes to approximately 8 minutes), the required stations drop to one-third (from 24 to approximately 8). After the 8^{th} minute, the number of required stations drops in a less significant way. With an increase of 5 minutes (total of 13 minutes), the number of required stations drops from 8 to 3 stations. The critical maximum times are 6.5 minutes and 9 minutes. These seem to be the boundaries of a quick but costly response system (<6.5 minutes response time and >13 stations required), a standard response system (between 6.5 minutes and 9 minutes, and between 13 stations and 6 stations), and a slow but cheap response system (>9 minutes response time and <6 stations required).

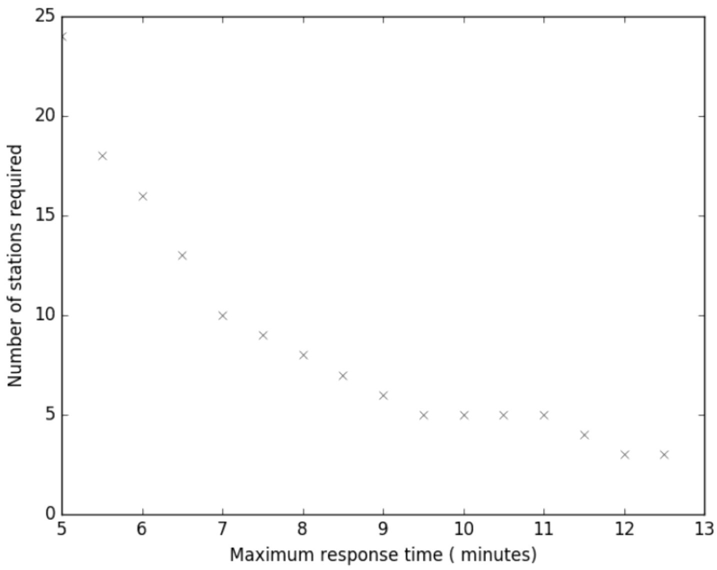

Figure 3. Number of minimum stations required for different maximum response times

These results show that a maximum response time of approximately 7 to 8 minutes can better equilibrate both the number of stations (10 to 8 stations) and the quality of the uEMS service. In fact, from 10 to 11 minutes, the number of required stations is the same as when the limit is set to 9.5 minutes

Nevertheless, this value is tightly connected with the road network configuration and land use.

Base case

We define the base case as the solution where the maximum number of stations is 10, the alphas are 1 for cardiac arrests and road accidents and null for the other cases, and the maximum response is 8 minutes for cardiac arrests and 12 minutes for any other medical emergency. The base case solutions, plus the solution for each isolated period, are represented in Figure 4.

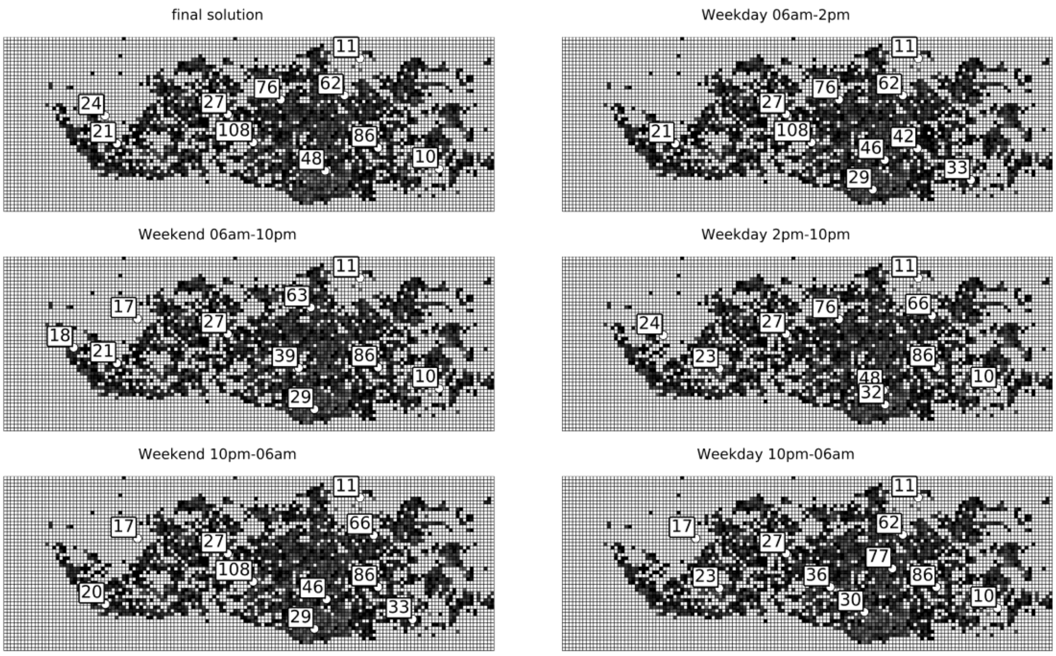

Figure 4. Stations locations for the different periods and final solution.

During the afternoon of the weekdays, the middle stations concentrate in the city centre, whereas during the night time the two west stations shift towards the map centre. During the weekend, the west area is occupied by 1 additional station; this might be due to the recreational areas, such as shopping malls, the seaside and the city park, present there. During the night period, there is an apparent placement of the stations closer to the nightlife spots during the weekend and to the residential areas during the weekday. Overall, the final solution (all periods are accounted) disperses the stations in a more even fashion, supporting the fact that although all emergencies calls are clustered together in a unique solution, having no special weights for any specific period might lead into a good final solution.

There are clear differences of needs during the system cycle (in this case an average week), which indicate that dynamic solutions have a role to play in emergency systems. Further support and reasons are stressed in the next analyses.

Alpha sensitivity test for cardiac and road crash events

To analyse the influence of the parameter alpha, α^k, for cardiac arrests and road crashes, a batch of test cases were computed varying each alpha by 2^n with $n = \{0, 1, 2, 3, 4, 5, 6\}$. The idea underneath is to understand how the solutions behaves in terms of spatial occupation. For each test case, a centroid is calculated by averaging the position of the stations' optimal location. In Figure 5, the centroids for each tested case are presented.

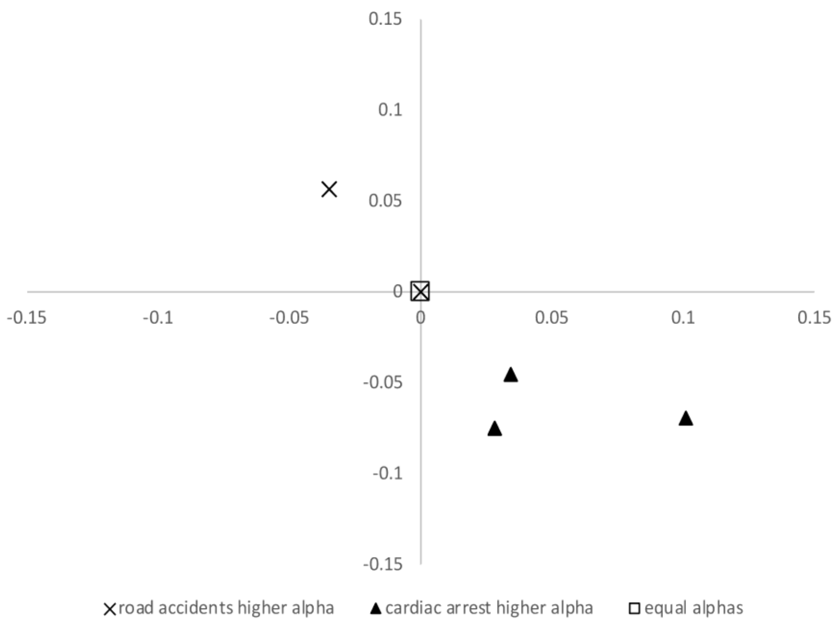

Figure 5. Percentage towards the city limit of displacement of each solution centroid, for different combinations of alpha, with respect to the centroid where the solution represents the optimal location when road accidents and cardiac arrest alphas are equal.

The coordinate (0, 0) shows the centroid when alpha is equal for both types. When the parameter alpha for cardiac arrests grows, compared to the road crash alpha, the solution centroid moves towards the city centre, from northwest to southeast. In the opposite situation, the road crash alpha grows relatively to the cardiac arrest alpha, and the solution centroid moves in the opposite direction, from southeast to northwest, and towards the outerbound of the city. This indicates that both road crashes and cardiac arrests are correlated with space and depend on the city land use. This supports our assumption that cities are dynamic and different types of events occur in different places of the city (Figure 5) and at different times (Figure 4).

Number of active ambulances

The simulation model was run for one year using the base case optimization solution (Figure 4). Using the results agent, we kept track of the minimum, maximum and average number of ambulances in use at each instant (hour steps). Within the simulation, we allow a maximum of 12 minutes for an ambulance to respond to any event but a cardiac arrest. For the latter, the maximum response time was 8 minutes.

Several station configurations were tested to assess the uEMS sensitivity in terms of the number of ambulances and their location. These configurations were computed using the optimization model and by varying the number of stations, the periods and the alpha parameters.

To compare the simulation results of the final solution with the different period solutions, we propose a comparative graph. This graph shows the additional number of ambulances required by a certain individual solution in comparison with the final solution for different percentiles of served hours.

It is important to highlight that individual solutions only respect the maximum response times during the period they were computed for. This means that when an individual solution is implemented to be used during the overall cycle not all emergencies will be answered within the proposed limits.

Base case

One of the most important questions in an EMS response system is the number and location of ambulance stations. Nevertheless, defining the station capacity is also a crucial design task.

The number of ambulances required in an EMS system is an important planning decision; thus understanding the system needs is fundamental.

We run the simulation model for the base case and assess the ambulance needs for one year (in 1 hour steps on a total of 8760 steps). We use different station configurations (Figure 4) corresponding to the different analysed periods in the station location analysis.

The results from the simulation model, Figure 6, show that in a period of one year, there is a maximum of 19 ambulances simultaneously being used. However, when we consider the percentile 99.98% (exclude the most loaded period), this number drops to 17 ambulances. At 95%, the number of ambulances reduces drastically to 7. The next reduction, 6 ambulances, occurs for 90%. This would mean that with 6 ambulances, 90% of the time, every request would be fulfilled within the maximum response times defined.

It is evident that there are few hours in the year where the system has outlier behaviour. Further investigations show that the most loaded hours correspond to the period between 1 am and 4 am on day of São João. On this city day, a festival is thrown at the city centre and most of the EMS calls in this period are because of intoxication and trauma. A clear overflow of calls is identified originating from a specific event. In these cases, a specific day plan is advised rather than overestimating the EMS system with ambulances that would be used only once per year.

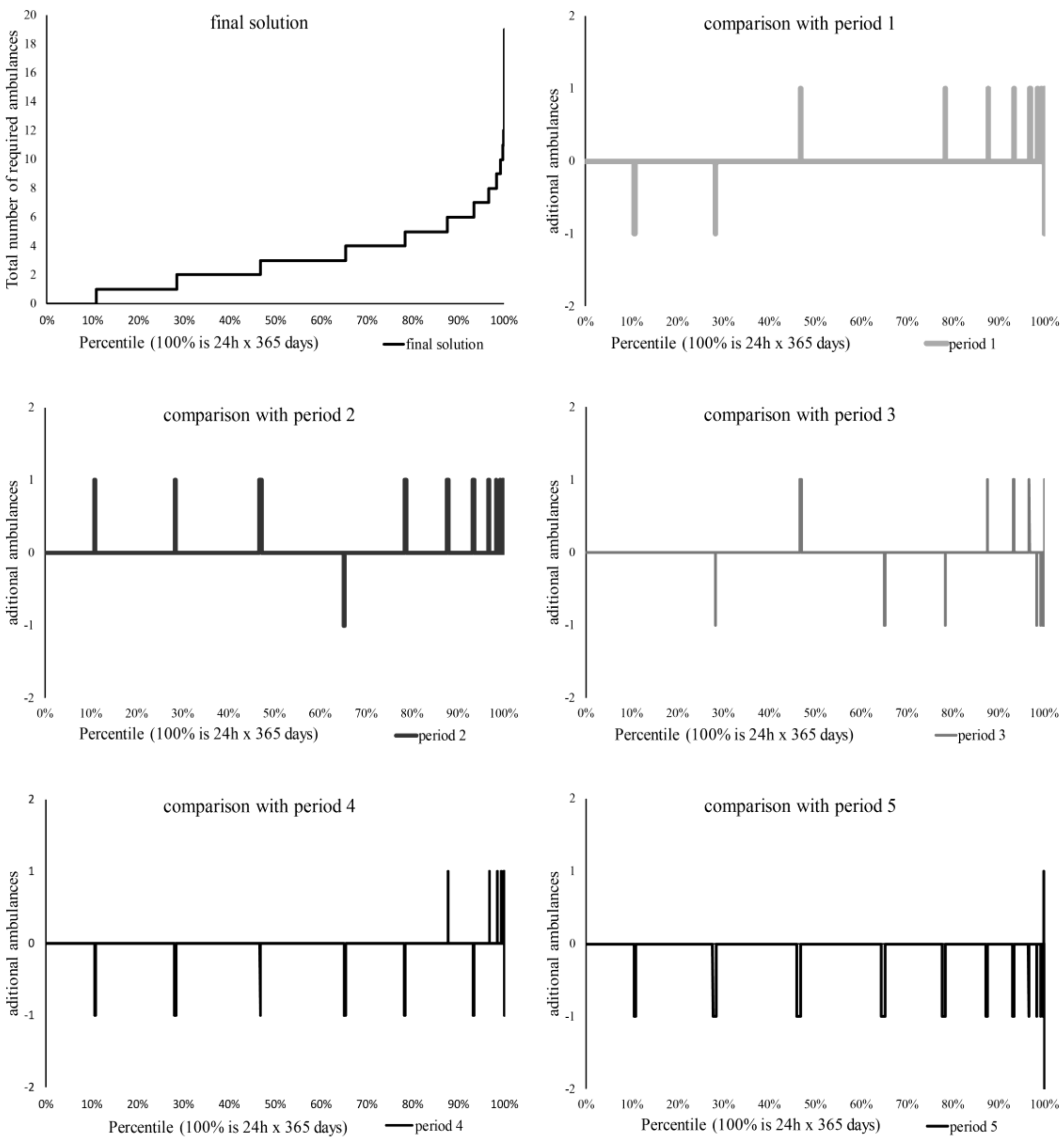

Figure 6. Number of active ambulances by percentile for one year in the all scenario facility solution, and comparison with the individual scenario solution.

Moreover, we compared the final solution requirements with the 5-single period solution. This will help us to determinate if our station location model is a good fit for a dynamic ambulance allocation.

Overall, the final solution has a clear fit to the dynamic allocation of ambulances. However, in a few occasions for the higher percentiles, this solution underperforms by requiring 1 more ambulance than the periods that account for the peak hours. Moreover, when compared with period 2 (afternoon peak hour), the solution is confirmed to be less robust, requiring one more ambulance in some occasions.

Nevertheless the final solution is the only that assures all emergencies are answered within the maximum response times proposed.

Although there is no clear evidence of a better fit than the static station location model, the scenario-based optimization model seems to have a good response towards dynamic systems. It is important to notice that this is a macro analysis that accounts for all emergency calls, without stressing the advantages of properly responding to cardiac arrests and road accidents.

Sensitivity test of the number of stations on the number of active ambulances

In the optimization model, we defined the number of stations as the minimum required to fulfil the coverage requirements. Nevertheless, this does not assure us the optimal solution in terms of ambulances required. To assess this, we computed the optimization model for different limits of stations, ran the simulation model and compared it with the base case (Figure 7).

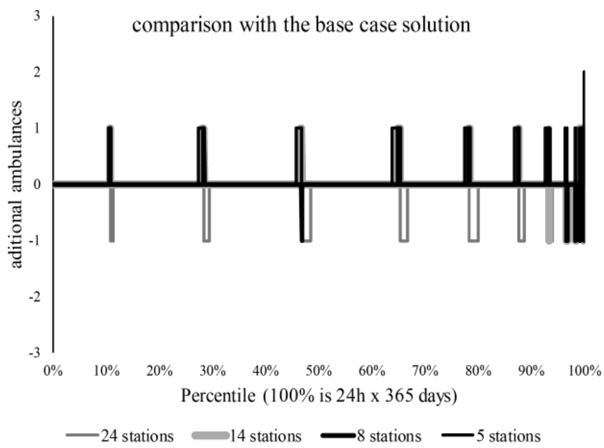

Figure 7. Number of additional active ambulances for different number of stations compared to the base case solution.

When using a solution with only 5 stations the system performs poorly, requiring additional ambulances in certain periods, even for the lowest percentiles. When using 8 stations, for the minimum required to fulfil the entire network within the maximum response time, the system behaves in a similar fashion but in fewer periods and never requiring more than 1 ambulance. The situation clearly inverts only when the number of stations jumps to as high as 24. 14 stations present a clear advantage for high percentiles (>90%). The reason the solution with 10 stations performs better than the solution with 14 stations for percentiles < 90% might be that additional stations tend to improve out of ordinary days, concentrating in areas where peaks of uEMS calls occasionally occur.

Alpha sensitivity test on the number of active ambulances

As per the optimization model, we also test how the alpha parameter influences the number of ambulances by influencing the stations location and compare this with the base case (equal alpha for cardiac and road accident calls and zero for other types) (Figure 8).

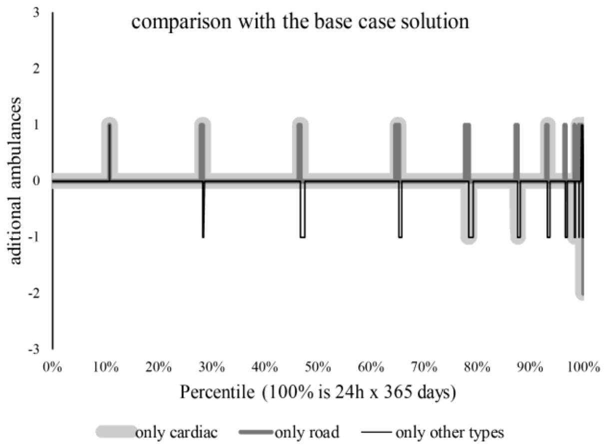

Figure 8. Number of additional active ambulances for different number of stations compared to the base case solution.

When the focus of the stations location is road accidents, the system responds faster to road accidents but becomes slower to other type of requests. This leads to a requirement of one additional ambulance in several periods. A similar situation occurs when the focus turns to the cardiac arrest. However, there are some periods where the system requires one fewer ambulance, which coincides with the percentile where the system would need one more ambulance if the focus were on road accidents. This is evidence of the fact that road accidents and cardiac arrests are distinct in time and space, at least during some periods at certain levels, providing clear proof that dynamic systems can make a difference.

Nevertheless, we also compared the base case with the solution where no focus is given to cardiac or road accidents. This incorporates most of the uEMS calls; thus the stations network are better positioned to respond to most of the cases leading to a requirement of fewer ambulances in several cases when the percentile is > 25%. For the future, effort should be made to score the survival gain rather than simply comparing the number of required ambulances.

Ambulances per station requirement for the base scenario and worst case for each hour

Finally, to obtain a clear understanding of the city dynamics and how a dynamic EMS system is justified over a static one, we assess the ambulance requirements per station for 24 h with the most loaded occurrence for each of the day hour (Figure 9).

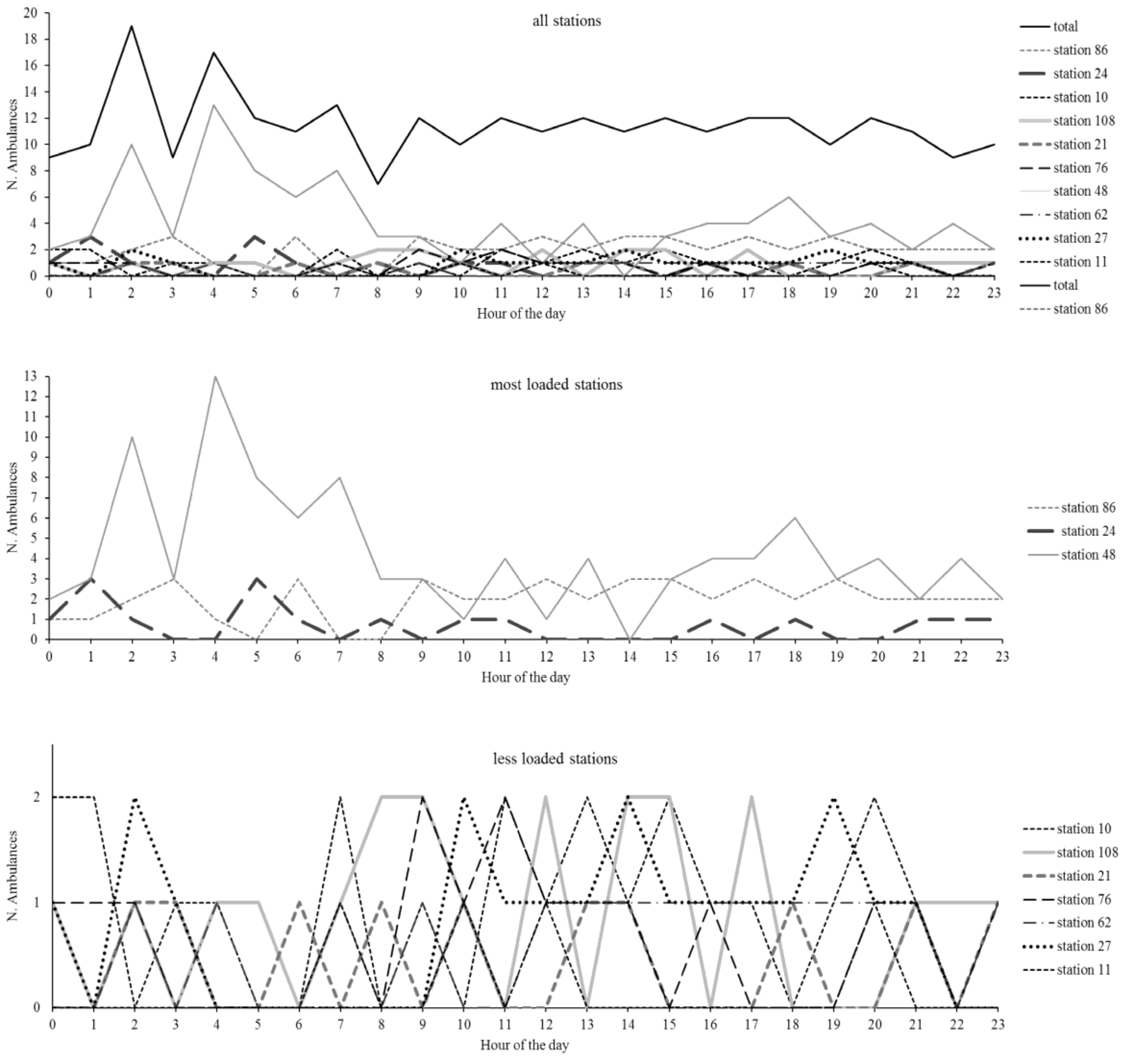

Figure 9. Maximum number of required ambulances per hour at each station for the base case.

As mentioned before, the most loaded hours corresponded to a specific event and occurred between 1 am and 4 am. Clearly, these occurrences occur in a specific area of the city, as only one station seems to be overloaded. Nevertheless, when looking for the total number of stations, the morning and afternoon periods seem to stabilize around 12 and 10 active ambulances. In a first view, this would indicate that a dynamic management of the fleet is not required. In a closer view, when focusing on the three most loaded stations, we recognize load changes between stations 86, 48 and 24, from the outskirts of the city, where business and industry focus, to the city centre, where commerce and old residential areas coexist. It is important to add that the area that station 86 covers is characterized by heavy traffic periods and fast roads.

In the less loaded stations (between 0 and 2 active ambulances), there is an evident interchange in the stations with active stations. For instance, there is a clear equilibrium between station 108 and stations 10 and 27, i.e., when the former is more loaded the latter are free and vice versa.

This analysis gives clear evidence of the advantages of the dynamic allocation of ambulances and how it can reduce the total number of required ambulances by dynamically reallocating them to a proper station, reducing the service time and consequently reducing the busy time of each ambulance.

Conclusions

This work opens doors to the study of city dynamics and its influences in the management of a uEMS response system

We defined a performance metric for the EMS response by summing the survival score of each rescued victim. Afterwards, we proposed a scenario-based optimization model where the scenarios are exchanged by static day periods to capture city dynamics.

An agent-based model simulation is offered to assess uEMS needs in terms of ambulances and stress the importance of a dynamic system.

The models were validated and, after minor simplifications, performed quickly, allowing for several cases to be tested within a reasonable time. The validation and sensitive tests were performed in a real case city, Porto, with data from one year with a total of 33 736 events from 10 May 2012 to 9 May 2013.

The presented models allow decision maker to better rationalize the number of stations and the average response time of the system. Nevertheless, the number of stations shows low variance in relation to the total number of active ambulances as shown by the simulation model.

The cycle division in periods is a simple and efficient way to deal with city dynamics and is proven to be relevant in the positioning of the stations. The city has its own dynamics concentrating people and traffic in different parts of its network throughout the day as proven by the individual period optimization analysis. Moreover, road accidents and cardiac arrests were proven to have different time and space behaviours, once more supporting our assumptions.

In terms of ambulances, it was shown that there are a few occasions when the system requires almost double the amount of ambulances compared to what would be required 95% of the time. Evidently, a supporting plan should be designed for these specific periods, such as large city events, thereby releasing the main uEMS system from this burden.

Although the number of active ambulances during the day is reasonably even, as per the analysed situation, there is evidence of the existence of a main station supported by a couple of other stations from where most of the calls are responded to. The remaining ones serve as area-specific stations that are requested at certain periods of time.

Overall the scenario-based optimization model was proved to be fit for a following dynamic allocation of ambulances.

It was revealed that the location of stations is impacted by the city dynamics and the survival functions, stressing further developments in the study of these functions for road accidents and other types of meaningful (survival or system related) emergency events. In addition, the use of realistic survival functions would allow a better assessment of the sensitive analyses provided here and possibly achieve clearer and more eloquent proofs.

Different period sizes should be tested, and the simulation model should be relaxed to allow ambulances to be reallocated to different stations. The simulation model should also allow ambulances returning from a hospital to be allocated to an active event without the need to return to their base.

Additionally, we propose a study of the impact of road crashes in the uEMS response time for other occurrences so that, first, a better performance function can be added to the optimization model, and, second, the simulation model is improved accordingly. Adding double coverage to the optimization is also worth investigating.

We acknowledge the support of FCT (Portuguese national funding agency for science, research and technology) under the grant PD/BD/52355/2013 during the development of this work. We also express gratitude to INEM (Portuguese Institute of Medical Emergency) for the information and data that they provided us.

References

Alanis, Ramon, Armann Ingolfsson, and Bora Kolfal. 2013. "A Markov Chain Model for an EMS System with Repositioning." *Production and Operations Management* 22 (1):216-31. doi: 10.1111/j.1937-5956.2012.01362.x.

Andersson, T., and P. Varbrand. 2006. "Decision support tools for ambulance dispatch and relocation." *J Oper Res Soc* 58 (2):195-201.

Andersson, Tobias. 2005. "Decision support tools for dynamic fleet management." *Doktorsavhandling, Linköpings Universitet, Sverige*.

Berman, O. 1981a. "Repositioning of Two Distinguishable Service Vehicles on Networks." *Systems, Man and Cybernetics, IEEE Transactions on* 11 (3):187-93. doi: 10.1109/TSMC.1981.4308651.

Berman, Oded. 1981b. "Dynamic Repositioning of Indistinguishable Service Units on Transportation Networks." *Transportation Science* 15 (2):115-36. doi: 10.1287/trsc.15.2.115.

Berman, Oded. 1981c. "Repositioning of distinguishable urban service units on networks." *Computers & Operations Research* 8 (2):105-18. doi: http://dx.doi.org/10.1016/0305-0548(81)90038-1.

Berman, Oded, and Amedeo R. Odoni. 1982. "Locating mobile servers on a network with markovian properties." *Networks* 12 (1):73-86. doi: 10.1002/net.3230120106.

Blackwell, Thomas H., and Jay S. Kaufman. 2002. "Response Time Effectiveness:Comparison of Response Time and Survival in an Urban Emergency Medical Services System." *Academic Emergency Medicine* 9 (4):288-95. doi: 10.1197/aemj.9.4.288.

Boloori Arabani, Alireza, and Reza Zanjirani Farahani. 2012. "Facility location dynamics: An overview of classifications and applications." *Computers & Industrial Engineering* 62 (1):408-20. doi: http://dx.doi.org/10.1016/j.cie.2011.09.018.

Brotcorne, Luce, Gilbert Laporte, and Frédéric Semet. 2003. "Ambulance location and relocation models." *European Journal of Operational Research* 147 (3):451-63. doi: http://dx.doi.org/10.1016/S0377-2217(02)00364-8.

Budge, Susan, Armann Ingolfsson, and Dawit Zerom. 2010. "Empirical analysis of ambulance travel times: the case of Calgary emergency medical services." *Management Science* 56 (4):716-23.

Church, Richard, and Charles R. Velle. 1974. "THE MAXIMAL COVERING LOCATION PROBLEM." *Papers in Regional Science* 32 (1):101-18. doi: 10.1111/j.1435-5597.1974.tb00902.x.

Current, John, Samuel Ratick, and Charles ReVelle. 1998. "Dynamic facility location when the total number of facilities is uncertain: A decision analysis approach." *European Journal of Operational Research* 110 (3):597-609. doi: http://dx.doi.org/10.1016/S0377-2217(97)00303-2.

Daskin, Mark S. 1983. "A Maximum Expected Covering Location Model: Formulation, Properties and Heuristic Solution." *Transportation Science* 17 (1):48-70. doi: doi:10.1287/trsc.17.1.48.

Daskin, Mark S., and Edmund H. Stern. 1981. "A Hierarchical Objective Set Covering Model for Emergency Medical Service Vehicle Deployment." *Transportation Science* 15 (2):137-52. doi: doi:10.1287/trsc.15.2.137.

Eisenberg, Mickey S., Bruce T. Horwood, Richard O. Cummins, Robin Reynolds-Haertle, and Thomas R. Hearne. 1990. "Cardiac arrest and resuscitation: A tale of 29 cities." *Annals of Emergency Medicine* 19 (2):179-86. doi: http://dx.doi.org/10.1016/S0196-0644(05)81805-0.

Erkut, E., A. Ingolfsson, and G. Erdogan. 2008. "Ambulance location for maximum survival." *Naval Research Logistics* 55 (1):42-58. doi: 10.1002/nav.20267.

Erkut, Erhan, Armann Ingolfsson, Thaddeus Sim, and Güneş Erdoğan. 2009. "Computational Comparison of Five Maximal Covering Models for Locating Ambulances." *Geographical Analysis* 41 (1):43-65. doi: 10.1111/j.1538-4632.2009.00747.x.

Ferreira, Sara, and António Couto. 2013. "Hot-Spot Identification: A Categorical Binary Model Approach." *Transportation Research Record: Journal of the Transportation Research Board* 2386 (-1):1-6. doi: 10.3141/2386-01.

Gendreau, M., G. Laporte, and F. Semet. 2005. "The maximal expected coverage relocation problem for emergency vehicles." *J Oper Res Soc* 57 (1):22-8.

Gendreau, Michel, Gilbert Laporte, and Frédéric Semet. 2001. "A dynamic model and parallel tabu search heuristic for real-time ambulance relocation." *Parallel Computing* 27 (12):1641-53. doi: http://dx.doi.org/10.1016/S0167-8191(01)00103-X.

Hogan, Kathleen, and Charles ReVelle. 1986. "Concepts and Applications of Backup Coverage." *Management Science* 32 (11):1434-44. doi: doi:10.1287/mnsc.32.11.1434.

Hogg, Jane M. 1968. "The Siting of Fire Stations." *J Oper Res Soc* 19 (3):275-87.

Iannoni, Ana Paula, Reinaldo Morabito, and Cem Saydam. 2009. "An optimization approach for ambulance location and the districting of the response segments on highways." *European Journal of Operational Research* 195 (2):528-42. doi: http://dx.doi.org/10.1016/j.ejor.2008.02.003.

Ingolfsson, A. 2006. The impact of ambulance system status management. Paper presented at the Presentation at 2006 INFORMS Conference.

Ingolfsson, Armann, Susan Budge, and Erhan Erkut. 2008. "Optimal ambulance location with random delays and travel times." *Health Care Management Science* 11 (3):262-74. doi: 10.1007/s10729-007-9048-1.

Jarvis, James P. 1981. "Optimal assignments in a markovian queueing system." *Computers & Operations Research* 8 (1):17-23. doi: http://dx.doi.org/10.1016/0305-0548(81)90028-9.

Kepaptsoglou, Konstantinos, Matthew Karlaftis, and George Mintsis. 2012. "Model for Planning Emergency Response Services in Road Safety." *Journal of Urban Planning and Development ASCE* 138 (1):18-25.

Kim, Byeongsun. 2016. "Exploring emergency areas for medical service using microscopic traffic simulation model." *Spatial Information Research* 24 (2):75-84. doi: 10.1007/s41324-016-0008-z.

Knight, V. A., P. R. Harper, and L. Smith. 2012. "Ambulance allocation for maximal survival with heterogeneous outcome measures." *Omega* 40 (6):918-26. doi: http://dx.doi.org/10.1016/j.omega.2012.02.003.

Kolesar, Peter, and Warren E. Walker. 1974. "An Algorithm for the Dynamic Relocation of Fire Companies." *Operations Research* 22 (2):249-74. doi: 10.1287/opre.22.2.249.

Krishnan, K., L. Marla, and Y. Yue. 2016. Robust ambulance allocation using risk-based metrics. Paper presented at the 2016 8th International Conference on Communication Systems and Networks (COMSNETS), 5-10 Jan. 2016.

Lam, Sean Shao Wei, Ji Zhang, Zhong Cheng Zhang, Hong Choon Oh, Jerry Overton, Yih Yng Ng, and Marcus Eng Hock Ong. 2015. "Dynamic ambulance reallocation for the reduction of ambulance response times using system status management." *The American Journal of Emergency Medicine* 33 (2):159-66. doi: http://dx.doi.org/10.1016/j.ajem.2014.10.044.

Maxwell, Matthew S, Mateo Restrepo, Shane G Henderson, and Huseyin Topaloglu. 2010. "Approximate dynamic programming for ambulance redeployment." *INFORMS Journal on Computing* 22 (2):266-81.

Maxwell, Matthew S., Shane G. Henderson, and Huseyin Topaloglu. 2009. "Ambulance redeployment: an approximate dynamic programming approach." In *Winter Simulation Conference*, 1850-60. Austin, Texas: Winter Simulation Conference.

McCormack, Richard, and Graham Coates. 2015. "A simulation model to enable the optimization of ambulance fleet allocation and base station location for increased patient survival." *European Journal of Operational Research* 247 (1):294-309. doi: http://dx.doi.org/10.1016/j.ejor.2015.05.040.

Miller, TanC, TerryL Friesz, RogerL Tobin, and Changhyun Kwon. 2007. "Reaction Function Based Dynamic Location Modeling in Stackelberg–Nash–Cournot Competition." *Networks and Spatial Economics* 7 (1):77-97. doi: 10.1007/s11067-006-9013-4.

Nair, R, and E Miller-Hooks. 2006. A case study of ambulance location and relocation. Paper presented at the Presentation at 2006 INFORMS Conference.

Panahi, S, and MR Delavar. 2009. "Dynamic Shortest Path in Ambulance Routing Based on GIS." *International journal of Geoinformatics* 5 (1).

Pons, Peter T., Jason S. Haukoos, Whitney Bludworth, Thomas Cribley, Kathryn A. Pons, and Vincent J. Markovchick. 2005. "Paramedic Response Time: Does It Affect Patient Survival?" *Academic Emergency Medicine* 12 (7):594-600. doi: 10.1197/j.aem.2005.02.013.

Restrepo, Mateo, Shane G. Henderson, and Huseyin Topaloglu. 2008. "Erlang loss models for the static deployment of ambulances." *Health Care Management Science* 12 (1):67-79. doi: 10.1007/s10729-008-9077-4.

ReVelle, Charles, and Kathleen Hogan. 1989. "The Maximum Availability Location Problem." *Transportation Science* 23 (3):192-200. doi: doi:10.1287/trsc.23.3.192.

Sánchez-Mangas, Rocío, Antonio García-Ferrrer, Aranzazu de Juan, and Antonio Martín Arroyo. 2010. "The probability of death in road traffic accidents. How important is a quick medical response?" *Accident Analysis & Prevention* 42 (4):1048-56. doi: http://dx.doi.org/10.1016/j.aap.2009.12.012.

Savas, E. S. 1969. "Simulation and Cost-Effectiveness Analysis of New York's Emergency Ambulance Service." *Management Science* 15 (12):B-608-B-27. doi: doi:10.1287/mnsc.15.12.B608.

Serra, Daniel, and Vladimir Marianov. 1998. "The p-median problem in a changing network: the case of Barcelona." *Location Science* 6 (1–4):383-94. doi: http://dx.doi.org/10.1016/S0966-8349(98)00049-7.

Su, Syi, and Chung-Liang Shih. 2003. "Modeling an emergency medical services system using computer simulation." *International Journal of Medical Informatics* 72 (1–3):57-72. doi: http://dx.doi.org/10.1016/j.ijmedinf.2003.08.003.

Toregas, Constantine, Ralph Swain, Charles ReVelle, and Lawrence Bergman. 1971. "THE LOCATION OF EMERGENCY SERVICE FACILITIES." *Operations Research* 19 (6):1363-73.

Valenzuela, Terence D, Denise J Roe, Shan Cretin, Daniel W Spaite, and Mary P Larsen. 1997. "Estimating effectiveness of cardiac arrest interventions a logistic regression survival model." *Circulation* 96 (10):3308-13.

Valinsky, David. 1955. "Symposium on Applications of Operations Research to Urban Services—A Determination of the Optimum Location of Fire-Fighting Units in New York City." *Journal of the Operations Research Society of America* 3 (4):494-512. doi: doi:10.1287/opre.3.4.494.

Vasić, Čedomir, Bratislav Predić, Dejan Rančić, Petar Spalević, and Dženan Avdić. 2014. "Dynamic Relocation of Emergency Ambulance Vehicles Using the AVL Component of the GPS/GPRS Tracking System." *Acta Polytechnica Hungarica* 11 (9).

Wang, Tianbing, Fei Wang, Xiaofeng Yin, Na Han, Peixun Zhang, Yuhui Kou, and Baoguo Jiang. 2015. "Changes and trends of pre-hospital emergency disease spectrum in Beijing in the past decade (from 2003 to 2012)." *Journal of Local and Global Health Science* 2015 (2):34. doi: 10.5339/jlghs.2015.itma.34.

Westgate, Bradford S., Dawn B. Woodard, David S. Matteson, and Shane G. Henderson. 2013. "Travel time estimation for ambulances using Bayesian data augmentation."1139-61. doi: 10.1214/13-AOAS626.

WHO. 2011. "Global Plan for the Decade of Action for Road Safety 2011-2020." *World Health Organization*.

Yue, Yisong, Lavanya Marla, and Ramayya Krishnan. 2012. An Efficient Simulation-Based Approach to Ambulance Fleet Allocation and Dynamic Redeployment. Paper presented at the AAAI.

Zhang, Lei. 2012. "Simulation optimisation and Markov models for dynamic ambulance redeployment." ResearchSpace@ Auckland.

Zhang, O, AJ Mason, and AB Philpott. 2008. Simulation and optimisation for ambulance logistics and relocation. Paper presented at the Presentation at the INFORMS 2008 Conference.

Zhu, Shanjiang, Woon Kim, and Gang-Len Chang. 2012. "Design and benefit-cost analysis of deploying freeway incident response units: case study for Capital Beltway in Maryland." *Transportation Research Record: Journal of the Transportation Research Board* (2278):104-14.