Computers & Industrial Engineering 64 (2013) 917–928

Contents lists available at SciVerse ScienceDirect

Computers & Industrial Engineering

journal homepage: www.elsevier.com/locate/caie

Joint location and dispatching decisions for Emergency Medical Services

Hector Toro-Díaz [a], Maria E. Mayorga [a,*], Sunarin Chanta [b], Laura A. McLay [c]

[a] Department of Industrial Engineering, Clemson University, Clemson, SC 29634, USA
[b] Department of Industrial Management, King Mongkut's University of Technology North Bangkok (Prachinburi Campus), Prachinburi, 25230, Thailand
[c] Department of Statistical Sciences & Operations Research, Virginia Commonwealth University, Richmond, VA 23284, USA

ARTICLE INFO

Article history:
Received 27 July 2012
Received in revised form 17 December 2012
Accepted 14 January 2013
Available online 4 February 2013

Keywords:
Location/allocation in healthcare
Hypercube model
Genetic algorithm

ABSTRACT

The main purpose of Emergency Medical Service systems is to save lives by providing quick response to emergencies. The performance of these systems is affected by the location of the ambulances and their allocation to the customers. Previous literature has suggested that simultaneously making location and dispatching decisions could potentially improve some performance measures, such as response times. We developed a mathematical formulation that combines an integer programming model representing location and dispatching decisions, with a hypercube model representing the queuing elements and congestion phenomena. Dispatching decisions are modeled as a fixed priority list for each customer. Due to the model's complexity, we developed an optimization framework based on Genetic Algorithms. Our results show that minimization of response time and maximization of coverage can be achieved by the commonly used closest dispatching rule. In addition, solutions with minimum response time also yield good values of expected coverage. The optimization framework was able to consistently obtain the best solutions (compared to enumeration procedures), making it suitable to attempt the optimization of alternative optimization criteria. We illustrate the potential benefit of the joint approach by using a fairness performance indicator. We conclude that the joint approach can give insights of the implicit trade-offs between several conflicting optimization criteria.

1. Introduction

Emergency Medical Service (EMS) systems are a public service that provides out-of-hospital acute care and transport to a place of definitive care, to patients with illnesses and injuries that constitute a medical emergency. The ultimate goal of EMS systems is to save lives. The ability of these systems to do this effectively is impacted by several resource allocation decisions including location of servers, districting of demand zones and dispatching rules for the servers. Common objectives are minimizing the mean response time and/or maximizing coverage. The relationship between minimizing response time and improving survivability has been reported by several works such as Sanchez-Mangas, García-Ferrer, de Juan, and Arroyo (2010) and McLay and Mayorga (2010, 2011). A demand zone is said to be covered if there is at least one facility within a predefined distance/time threshold from the demand zone. The concept of coverage is related to the availability of a satisfactory facility rather than the best possible one (Farahani, Asgari, Heidari, Hosseininia, & Goh, 2012). Li, Zhao, Zhu, and Wyatt (2011) pointed out that the coverage maximization approach is the most widely used by practitioners, researchers and regulators.

Traditionally, location and dispatching decisions have been approached separately, even though various studies have shown that the servers' busy probabilities (and therefore the response time and coverage, among other performance indicators) are sensitive to the server locations and the choice of server dispatching strategies (Batta, Dolan, & Krishnamurthy, 1989; Larson & Odoni, 1981). Ambulance dispatch is the process of assigning a particular ambulance to answer an emergency call. An ambulance dispatch policy can be formed using various dispatch methods and there is no single policy that fits all systems (Li et al., 2011). The same authors emphasized that a dispatch policy has to be designed to fulfill the particular objectives and performance indicators defined by EMS providers and regulators. In our work we consider dispatch policies in which there is a single list associated with each demand zone that ranks the available servers (ambulances), or a subset of them, in order of dispatch preference. This type of list is commonly referred to as a contingency table.

The most common dispatching policy for EMS calls is rather simple in that the closest idle vehicle is usually dispatched to attend the call (Andersson & Varbrand, 2006; Goldberg, 2004). The rationale behind that policy is related to the idea of having the

* Corresponding author. Address: Department of Industrial Engineering, Freeman Hall, Box 340920, Clemson, SC 29634, USA. Tel.: +1 864 656 6919; fax: +1 864 656 0795.
E-mail addresses: htoro@clemson.edu (H. Toro-Díaz), mayorga@clemson.edu (M.E. Mayorga), snct@kmutnb.ac.th (S. Chanta), lamclay@vcu.edu (L.A. McLay).

0360-8352/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.cie.2013.01.002

objective of minimizing the mean system response time. The works on allocation of distinguishable servers by Jarvis (1981) and Katehakis and Levine (1986) pointed out that under light traffic conditions using a myopic allocation policy (i.e. assigning always the closest available sever) will lead to an optimal solution, when the objective is to minimize the long run average cost (response time). For heavy traffic the same works mentioned that the optimal policy can deviate from the myopic policy. However, even in the latter case using the myopic policy still might lead to solutions that are close the optimum (Katehakis & Levine, 1986). Related literature applied to EMS systems planning included arguments against the closest dispatching rule as a way to minimize the response time. Arguments were made originally by Carter, Chaiken, and Ignall (1972) and thereafter supported by Cuninghame-Green and Harries (1988) and Repede and Bernardo (1994). In the referred works the locations of the servers are assumed to be known. We have not found references addressing the relationship between a myopic dispatching policy and expected coverage. There is usually a trade-off between response time and coverage, so that improving one of them implies a sacrifice in the other.

In this work, first we present a mathematical model that integrates the location and dispatching decisions for an EMS system. It is a non-linear mixed integer optimization model in which even generating some of the equations is computationally intensive, therefore making it hard to solve. The Hypercube model is used providing an exact model of the stochastic queuing dynamics. The mathematical model is accompanied by the analysis of randomly generated small instances whose purpose is twofold: (i) given the small size it is possible to fully enumerate all feasible solutions hence also identifying the optimal, that can be used later for comparison purposes against faster/smarter solution strategies than enumeration; and (ii) after solving a variety of random instances it is also possible to point out some general trends observed in the optimal solutions (with respect to response time and coverage). Second, we present an optimization framework to solve the joint location and dispatching problem based on Genetic Algorithms (GAs). We present a heuristic solution procedure to solve the exact model of the system. Our work is different from previous approaches to the problem, for although we assume the general form of the dispatching policy, as a fixed preference list, we do not assume a priori any particular dispatching order (based on distance, for example). Instead, we model the location and dispatching decisions in a single mathematical model, and develop an optimization framework for its solution. In fact, since a district is the union of the demand zones assigned to a particular server, it can be said that an indirect result of our model is also a districting strategy: for each available server, all the zones having it as its first preferred server would form the server's district.

Our findings are that in fact the common dispatching rule based on the closest available server leads to the best solutions when the objective is minimizing the mean response time and locations are optimized simultaneously. Conversely, if the objective is maximizing expected coverage, then the optimal solution could deviate from the use of the closest dispatching rule. However, the best solutions based on coverage offer an increase of that indicator (with respect to the coverage attained by minimizing the mean response time) that is rather small (3.15% average increase – 95% CI: 2.75–3.55%) compared to the sacrifice in response time (65.2% average increase – 95% CI: 56.33–74.24%). Although these numbers correspond to the average results for the small instances, bigger instances showed similar behavior. The optimization procedure proposed has consistently obtained good solutions, i.e. within 1% gap compared to the best solutions obtained by full or partial enumeration procedures, which are computationally more intensive.

While our main goal was the development of the optimization framework for the solution of the joint location/dispatching problem, we discovered that little benefit can be gained from the integrated approach when using the two most commonly used criteria, namely response time and expected coverage. Thus we considered two additional criteria related to fairness, and we used one of them to illustrate the potential benefits of the joint approach. In particular we tested the variance of the individual response times as a measure of fairness from the point of view of the users of the system (demand zones). We found that in this case using a myopic policy would result in a potential deviation from the optimal policy aimed at reducing disparities, as measure by the variance of the response times. We also illustrate the trade-offs among the presented optimization criteria.

The rest of the paper is organized as follows. In Section 2 we provide the presentation of the problem as well as a review of related literature. Next, in Section 3 we introduce the mathematical model. Section 4 presents a small case study, as well as a summary of its results and implications. Section 5 provides a detailed description of the optimization framework based on GAs and Section 6 introduces the case studies to which the optimization procedure is applied, as well as the results obtained. The last two sections, 7 and 8 are the discussion of the results and the conclusions, respectively. As part of the conclusions possible extensions of the present work are mentioned.

2. Problem presentation and related literature

In Goldberg's review of models for deployment of EMS vehicles (Goldberg, 2004), it is mentioned that little work had been done on dispatching of ambulances. Similar opinion is shared by Lee (2011), mentioning that the contributions in ambulance dispatching are sparse. In turn, Galvao and Morabito (2008) and Iannoni, Morabito, and Saydam (2011) mention as an interesting extension of their work the use of different dispatch preference lists, instead of assuming that for a given set of locations the dispatching order is based on the closest dispatching rule.

The most widely used dispatching rule under a fixed preference scheme is to send the closest unit, looking to minimize the response times (Andersson & Varbrand, 2006). The first argument against the use of such a policy was made by Carter et al. (1972). They present a case where two units, A and B, have equally large areas of responsibility, but A's area has a significantly higher call frequency. In those conditions, the mean response time will decrease if B is allowed to respond to some of the calls for which A is the closest unit. The result was generalized for cases involving more than two units by Cuninghame-Green and Harries (1988). Repede and Bernardo (1994) also supported the argument. The works by Jarvis (1981) and Katehakis and Levine (1986) studied the optimal allocation of distinguishable servers on Markovian queuing systems, reaching a different conclusion. For a given location of the servers, so that the cost of assigning a server to a particular customer is known (the cost in EMS planning is usually related to the amount of time that it takes for the EMS system to effectively respond to a call), these two works showed that under light traffic conditions (traffic is measured by the ratio between the mean arrival rate and the mean total service rate) the use of a myopic policy always would lead to an optimal solution, i.e. minimizing the long run average cost (response time). For heavy traffic the use of a myopic policy will deviate from the optimal, however the deviation is rather small (2–3%). Katehakis and Levine (1986) used 0.38 as an indicator of light traffic and 1.94 for the case of heavy traffic.

We propose a mathematical model that combines location and dispatching decisions for EMS vehicles, initially looking for optimal solutions according to maximum coverage or minimum response time. The dispatching decisions are modeled as a fixed preference

scheme, meaning that there is a single list associated with each costumer that ranks the available servers (ambulances) ir order of dispatch preference. That list does not change as a result of changes in the state of the system. However, the particular unit that will be dispatched to attend each call from a demand zone is not known in advance, since the assignment depends on the availability of the servers (system's state) when the call is received. Katehakis and Levine (1986) pointed out some results from Markov Decision Theory indicating that, when the number of states of the system as well as the number of actions available to perform in every state (allocation of the servers) are finite, it suffices to consider only deterministic policies; a deterministic policy is one which, whenever the system is in particular state, the set of available actions to perform is deterministic and depends only of the actual state (in our case, which servers are busy, and which are idle).

The servers in a typical EMS system are: (i) spatially distributed in the region; (ii) share the system workload due to cooperation among them and (iii) have different operational characteristics, such as different preferential regions (Galvao & Morabito, 2008). Those characteristics have been progressively included in different approaches used for planning EMS systems. Congestion is also a typical phenomena related to EMS systems. According to Galvao, Chiyoshi, and Morabito (2005) the volume of calls for service may keep ambulances busy from 20% to 30% of the time.

Brotcorne, Laporte, and Semet (2003) provided a review focused on location models and their particular application to EMS. They classified the location models that evolved over the past 30 years into two main categories, deterministic and probabilistic, recognizing that the most recent models were more concerned with the representation of the stochastic nature of the systems. Location models were also distinguished in *coverage* and *median* type problems. The first class attempts to locate the servers so as to maximize the fraction of the demand that has at least one server unit within a predefined maximum distance or time. The latter type minimizes the average or total travel time/cost between servers and demand zones.

The two seminal attempts to develop basic coverage models were the set covering location problem (SCLP) by Toregas, Swain, ReVelle, and Bergman (1971) and the maximal coverage location problem (MCLP) by Church and ReVelle (1974). Extensions to those basic models were developed later. TEAM and FLEET models, by Schilling, Elzinga, Cohon, Church, and ReVelle (1979), considered several types of servers; Marianov and ReVelle (1992) improved the MCLP model. Multiple coverage of demands were considered in BACOP1 and BACOP2 by Hogan and Revelle (1986) and other extensions, DSM and DDSM were added by Gendreau, Laporte, and Semet (1997, 2001). The *p-median* problem was introduced by Hakimi (1964). The use of the *p-median* model in the planning and location of facilities for EMS can be found in Carbone (1974) and Carson and Batta (1990).

Basic location models are deterministic in nature and therefore do not represent the system accurately (Brotcorne et al., 2003; Jia, Ordóñez, & Dessouky, 2007a). Basic coverage models might make sense when the location of facilities are fixed, but in the case of an EMS system, as soon as a unit leaves its home base to attend a request for service, other demand points that are supposed to be covered by that unit may no longer be covered. The work by Snyder (2004) reviewed several models that address variations in the inputs, such as demands and travel times, as a way to take uncertainty into account. The same work pointed out the importance of addressing congestion. Daskin (1983) developed the maximum expected coverage location model (MEXCLP) including the modeling of congestion elements. Hogan and Revelle (1986) developed the maximal availability location problem (MALP I and II) and later Marianov and ReVelle (1996) improved it. Farahani et al. (2012) present an extensive up to date review on covering problems in facility location. Arabani and Farahani (2012) developed a survey on facility locations dynamics.

It was the work by Larson (1974) that first used queueing theory elements in facility location models by introducing the hypercube model. Larson (1975) later developed an approximation for the hypercube model due to the fact that exact calculations were prohibitive. Chiyoshi, Galvao, and Morabito (2001) pointed out, after comparing several models, that the hypercube was the only one with the capabilities for an accurate representation of the system. There is a variety of applications and extension of the hypercube model to EMS system such as the works by Brandeau and Chiu (1989), Mendona and Morabito (2001), Atkinson, Kovalenko, Kuznetsov, and Mykhalevych (2008), Iannoni and Morabito (2007), Iannoni, Morabito, and Saydam (2008), Galvao and Morabito (2008) and Geroliminis, Karlaftis, and Skabardonis (2009), among others. It is well documented that the hypercube model is a descriptive tool allowing scenario analysis, not designed as an optimization model. However, it is possible to embed it into an optimization framework. Batta et al. (1989) combined MEXCLP with the hypercube into an iterative, local search algorithm. Aytug and Saydam (2002) replaced the local search by a genetic algorithm. Iannoni and Morabito (2007) as well as Iannoni et al. (2008) and Geroliminis, Kepaptsoglou, and Karlaftis (2011) have embedded the hypercube model into genetic algorithms to solve the location problem. In this paper we also use the hypercube model as it exactly models the system. While hypercube approximations (Jarvis, 1985; Larson, 1975) may lead to faster solution procedures, they do not provide an exact solution. Thus, as mentioned earlier, our approach is to find a heuristic solution to an exact problem as opposed to an exact solution to an approximate problem. Future work, related to scalability issues of the proposed method is mentioned in Section 8.

3. Mathematical model

Our model is different from existing literature in that we integrate location and dispatching decisions into a single framework, whereas the mentioned references assumed the use of a priori dispatching policy, particularly based on the closest relationship.

3.1. Assumptions

It is assumed that the system provides service to a certain geographical region **J** that is partitioned into service regions -demand zones, cells or atoms are other terms that have been use for these partitions. A given number of servers are located at points $i \in \mathbf{I} \subset \mathbf{J}$. Demands occur solely at the center of each service region by time homogeneous Poisson requests for service and are attended at exponential service rates. Larson and Odoni (1981) have shown that reasonable deviation from this last assumption do not significantly alter the accuracy of the model.

Each service region j generates a fraction f_j of the total demand ($\sum_j f_j = 1$). The total demand is then λ and the demand of each zone is $\lambda_j \equiv \lambda f_j$. A server's primary response area (*district*) consists of those service regions to which the server would be dispatched if available. When a request for service arrives, if the primary responsible server is available, it is dispatched immediately. The server travels to the place of the incident, spends some time at scene and then returns to its base location before being assigned to the next request. If the responsible server is busy when a request for it arrives, another server will be assigned, following a fixed priority list with respect to the servers for each demand zone. The priority list can include all the servers available in the system (total backup) or only a subset of them (partial backup). If all the servers are busy, the request is considered to be lost (this typically means that

it will be served by an external system). The basic model also assumes that the servers are identical and that the service time of any response unit for any call for service has an exponential distribution with mean $1/\mu$ (This assumption is reasonable if the travel times are short compared to the total service time, which is usually the case in urban areas). The service time for a call includes the set up time, the travel time from the base to the incident location, the on-scene time, a possible related follow up-time and the travel time back to the base. The response time interval is the time from when an ambulance is dispatched until it arrives at the scene.

Each server can be busy or free (idle), generating 2^N possible states for the system (where N = number of servers); the states can be mapped to the vertices of a hypercube (strictly a cube for the case of exactly three servers) named $B_j (j = 1,2,\ldots,2^N)$ of dimension N. Each vertex, or state, is denoted by an ordered set of N one digit binary numbers taking the value of 1 if the server is busy and 0 if not ($B_j \equiv \{b_1, b_2, \ldots, b_N\}$). It is assumed that only one step transitions occur, i.e. two servers cannot be assigned simultaneously. Using the convention proposed by Larson (1974), transitions are only allowed between states with Hamming distance equal to 1, where the Hamming distance d_{ij} between two vertices B_i and B_j is the number of digits by which the two vertices differ (or the 'right angle' distance between two vertices of the hypercube). The terms "upward" and "downward" Hamming distance, d_{ij}^+ and d_{ij}^-, refer to the number of binary digits switching from 0 to 1 and 1 to 0. The model of the system corresponds to a finite-state continuous time Markov process. Steady-state probabilities are determined from equations of detailed balance that express a conservation of flow between consequent states. This set of balance equations depends on both, the location of the servers and the dispatching policy.

3.2. Formulation

In the following formulation $\mathbf{J}$ represents the service regions; $\mathbf{I}$ are the potential location sites, $|\mathbf{I}| \leqslant |\mathbf{J}|$; N is the total number of response units (servers); t_{nj} is the mean response time for server n to reach region j, when available; λ is the total network-wide demand (requests/unit time); f_j is the fraction of network-wide workload generated from region $j \in \mathbf{J}$; E_{nj} is the set of states where server n is the preferred server for region j; C_N are the vertices of a N-dimensional hypercube; d_{ij}^-, d_{ij}^+ are the "downward" and "upward" Hamming distances between vertices B_i and B_j, $\left(d_{ij}^- + d_{ij}^+ = d_{ij}\right)$ and λ_{ij}, μ_{ij} are the upward and downward mean rates at which transitions are made from state i to state j, corresponding to vertices B_i and B_j, given that the system is in state i. Finally, we have the decision variables:

$$x_i = \begin{cases} 1 & \text{if we locate a vehicle at potential site } i \\ 0 & \text{otherwise} \end{cases}$$

$$y_{ij}^l = \begin{cases} 1 & \text{if vehicle located at } i \text{ has priority } l \text{ to zone } j \\ 0 & \text{otherwise} \end{cases}$$

The following are auxiliary decision variables: ρ_{nj} is the fraction of dispatches sending unit n to region j, $n = 1,2,\ldots,N$; $P(B_k)$ is the steady-state probability of state corresponding to vertex B_k, $k = 1,2,\ldots,2^N$.

Eq. (1) is the objective function, Mean Response Time (MRT); constraint (2) determines the number of servers to be located and constraint (3) is the integrality constraint for the decision variable x_i. Constraint (4) calculates the fraction of all dispatches that send server n to region j using standard queueing theory arguments and assuming a zero-line capacity system (calls that arrive when all the servers are busy are lost); Eq. (5) represents detailed balance equations determining steady-state probabilities of the finite-state continuous time Markov process model with N servers. Note that even though it was assumed that the service rate is equal for all servers since they were identical, the general expression given by this equation allows for different service rates for different servers. For details on calculating λ_{ij} and μ_{ij}, see Geroliminis et al. (2009).

Constraint (6) ensures that the sum of probabilities is equal to one. Eq. (7) is the integrality constraint for the decision variable y_{ij}^l; constraint (8) states the logical relationship between the location decision and the assignment of a location within the priority list of a demand zone and finally, constraints (9) and (10) assure that there is a complete priority list for each demand zone, and that within the priority list of each demand zone each server appears only once. The full model given by (1)–(10) represents the basic optimization problem in which the location of the servers and the dispatching rule for each demand zone are the decisions to be made. Also note that the steady-state probabilities are auxiliary variables that change for every full combination of location and dispatching decision.

The optimization problem is formulated as:

$$\text{Minimize } MRT = \sum_{n=1}^{N}\sum_{j=1}^{J} \rho_{nj} t_{nj} \tag{1}$$

s.t:

$$\sum_{i=1}^{I} x_i = N \tag{2}$$

$$x_i \in \{0,1\}, \quad i \in \mathbf{I} \tag{3}$$

$$\rho_{nj} = f_j \frac{\sum_{B_i \in E_{nj}} P(B_i)}{1 - P(B_{2^N})} \quad n = 1,\ldots,N; \ j \in \mathbf{J} \tag{4}$$

$$P(B_j)\left[\sum_{\substack{i \\ B_i \in C_N : d_{ij}^+ = 1}} \lambda_{ij} + \sum_{\substack{i \\ B_i \in C_N : d_{ij}^- = 1}} \mu_{ij}\right] = \sum_{\substack{i \\ B_i \in C_N : d_{ij}^- = 1}} \mu_{ij} P(B_i) + \sum_{\substack{i \\ B_i \in C_N : d_{ij}^+ = 1}} \lambda_{ij} P(B_i) \quad j = 1,\ldots,2^N \tag{5}$$

$$\sum_{i=1}^{2^N} P(B_i) = 1 \tag{6}$$

$$y_{ij}^l \in \{0,1\} \quad i \in \mathbf{I}; j \in \mathbf{J}; l = 1,\ldots,N \tag{7}$$

$$x_i \geqslant y_{ij}^l \quad i \in \mathbf{I}; j \in \mathbf{J}; l = 1,\ldots,N \tag{8}$$

$$\sum_{l=1}^{N} y_{ij}^l = 1 \quad i \in \mathbf{I}; j \in \mathbf{J} \tag{9}$$

$$\sum_{i=1}^{I} y_{ij}^l = 1 \quad j \in \mathbf{J}; l = 1,\ldots,N \tag{10}$$

Formally, the presented model corresponds to an NP-Hard problem (Geroliminis et al., 2009). It is a non-linear mixed integer programming model that has embedded a queuing sub-model corresponding to the finite-state continuous time Markov process. Given a particular set of locations for the servers available and a preference list for each demand zone with respect to the same servers, it is necessary to first solve the flow balance equations given by (5), (6), before being able to calculate the value of the objective function. Although compactly written, those equations are neither easily determined nor easily solved. The number of flow balance equations equals 2^N, therefore the number of equations to solve the subproblem grows exponentially with respect to the number of servers. It has been mentioned by Galvao and Morabito (2008) that in fact the computer time required to generate the coefficients of the linear

system may be even higher than the time required to solve it. That is because of the complex relationships imposed by the combined location and dispatching decisions. The flow balance equations lead to a linear system of equations, whose exact solution requires the calculation of the inverse for the matrix of coefficients. The size of this matrix grows exponentially, therefore the time that it takes to perform a single iteration to evaluate a candidate solution can be prohibitive.

It was mentioned that maximizing coverage is the most commonly used approach to planning EMS systems. Instead of the standard coverage, we use the concept of expected coverage as presented by Ingolfsson, Budge, and Erkut (2008), which takes into account the congestion of the system and potentially the variability in responses times. The following equations details how to calculate the expected coverage:

$$\text{Ex. Cov} = \sum_{j=1}^{J} f_j \sum_{n=1}^{N} P_{j,j(n)} (1 - P_{j(n)}) \prod_{u=1}^{n-1} P_{j(u)} \tag{11}$$

where $P_{j,i}$ is the probability that station i covers node j, P_i corresponds to the busy probability of the ambulance in station i and $j(n)$ refers to the nth preferred station for demand node j. Note that $P_{j,i}$ can be used as a binary variable, indicating whether or not the coverage threshold is satisfied by the available servers, but it can also be used as the probability of that coverage being possible within the given threshold, accounting for variability in travel times. In this particular case it has been used as a binary variable. Eq. (11) replaces Eq. (1) in the optimization model for the cases in which we are looking at maximum expected coverage.

4. Toy case study

In this section we introduce a case study that is small enough that we can enumerate all the possible solutions, also identifying the optimal. Because of the small size we also use the exact solution for the embedded hypercube model. For this example we use a square region on the cartesian plane and assume that we have five demand zones, that are also candidate locations for three available servers. The locations of the demand zones can be in any integer ordered pair within a grid, starting at (0,0) and extending up to (10,10) on the plane. The demand for each zone ranges between 1 and 20 calls/period-time.

To generate different instances we use random numbers as follows. The coordinates (x,y) for each of the five demand zones are obtained by generating uniform integers between 0 and 10. For each one of the demand zones the demand is obtained by generating uniform integers between 1 and 20. The distances between demand zones correspond to right angle distances. Optimal locations are nevertheless insensitive to the choice of a distance metric (Benveniste, 1985). The service rate for the servers is assigned based on assuming a particular value for the overall utilization of the system, namely $\rho = \lambda/3\mu$. As in the works by Budge, Ingolfsson, and Erkut (2009) and Chiyoshi et al. (2001), where ρ is varied between 0.1 and 0.9, three different scenarios of utilization are evaluated for each combination of location of demand zones and demands, by using $\rho = 0.1, 0.5, 0.9$. The server's speed is assumed to be 1.0 distance-units/time. The maximum threshold used for coverage was 7.0 distance units. 100 different set of locations are generated, and since for each one of them three scenarios are considered for the service rates, we generate 300 different instances.

One of such randomly generated problem (Table 1) and its respective optimal solution for MRT (Table 2) after enumeration is detailed next. It is worth noticing the number of possible solutions: 77,760. There are $\left(\binom{5}{3} = 10\right)$ possible locations. Each demand zone has an ordered list of the servers, and since there are three servers, each customer can have 3! unique lists. The total number of solutions is then $10 \times 3!^5 = 77{,}760$. Note that the inclusion of another demand zone would increase the number of solutions to 466,560. In other words, the number of possible solutions increase by a factor of 3!. Hence, the search space for a real size problem is huge, and enumeration is no longer an alternative.

In Table 2, St. Cov refers to the basic calculation of coverage, hence each demand zone is considered covered if there is at least one ambulance located at a distance of 7.0 or less distance units; St. Cov = 1.0 means full coverage. However, this definition of coverage does not take into account the congestion of the system, hence we used Ex. Cov (Eq. (11)). MRT (Eq. (1)) is the mean system response time. The 5th column corresponds to the probability of the system being busy (all the servers are attending calls), and therefore new calls would be rejected. The last column indicates the optimal locations of the servers.

At first sight results in Table 2 correspond to what was expected. On one hand, the increase of the overall utilization, which basically means reducing the service rates while keeping the call rates constant, causes an increase in the expected response time, as well as an increase in the busy probability. On the other hand, we can see that the standard coverage is not able to take into account the congestion phenomena. The expected coverage given by (11) is clearly affected by the increase of the overall utilization. The more congested the system, the lower the expected coverage.

We have solved by enumeration a total of 300 small size problems, for both minimum response time and maximum expected coverage. As expected according to the arguments expressed in Section 2, for each one of the 300 problems, the optimal solution that minimizes MRT was the same as a solution where the locations were optimized and a dispatching list based on the closest vehicle was used. We have also observed that, when there are ties (several servers are at the exactly same minimum distance from a given demand zone), only one of the combinations leads to the optimal solution, although the use of other dispatching ranking, which would still be based on the closest rule (because of the ties), causes an increase on the objective function value that in the worse case is below 2%. Note that the change of the preference list of a single demand zone, even if for that demand zone several servers' locations are tied, changes the overall performance indicators of the system.

Next we looked for maximum expected coverage (Ex. Cov), as given by Eq. (11). Once again, we enumerated all the possible solutions for all 300 instances to be able to identify the ones that gen-

Table 1
Spatial locations and demand - one random instance.

	Locations		
Index	*x-Coord*	*y-Coord*	Demand
1	10	5	20
2	1	1	18
3	7	9	12
4	2	7	8
5	6	1	6

Table 2
Optimal (MRT) solution information.

	Performance indicator				
ρ	St. Cov.	Ex. Cov.	MRT	$P[111]$	Optimal locations
0.1	1.0	0.954	2.123	0.003	1–2–3
0.5	1.0	0.721	4.340	0.134	1–2–3
0.9	1.0	0.517	5.355	0.309	1–2–4

erate the maximum coverage. In this case we have noticed that the optimal solutions do not follow the closest dispatching rule. We have also observed that there are several solutions that exhibit the same maximum coverage for a particular instance, and that the associated response time of those solutions present great variation. Since minimizing the response time is also important, in cases where there were several optimal solutions with respect to coverage we have selected the one with the minimum associated response time. Although the optimal solutions with respect to expected coverage do not follow a myopic allocation policy we have also noticed that the use of such a policy would cause a decrease in expected coverage that in all cases is below 3.8%. In fact for half of the cases it is below 1.0%.

Since both objectives are important, we compare the optimal solutions obtained with each optimization criteria. We noticed that for every instance of the problem, the maximum coverage identified is in fact greater than the coverage associated with the minimum response time solution. However, the increase in coverage is small on average, ranging from 0.18% to 17.4%, with a mean increase of 3.15% (95% CI: 2.75–3.55%). On the other hand, the increase in coverage (obtained by changing the objective function) comes as a result of worsening the response time. The increase in MRT ranges from 1.5% to 117%, averaging 65.2 (95% CI: 56.33–74.24%)%. As expected, there is a trade-off between response time and expected coverage. These results seem to indicate that focusing on minimizing the response time yields solutions that are robust with respect to the expected coverage. There were only 4 cases (out of 300) in which the proportional improvement in coverage (when maximizing coverage) was in fact higher than the corresponding increase in response time. In all the other cases the trade-off between improved coverage but increased response time is not appealing. These results are aligned with those reported by Geroliminis et al. (2009), who mentioned that the optimal locations obtained by using MCLP (a coverage maximization approach) performed up to 40% worse when the response time was evaluated by using the hypercube model. In the next section we introduce an optimization framework that allows us to solve bigger size problems, hence allowing us to check if the observed behavior of the small instances holds for more real-world sized problems.

5. Genetic Algorithm based optimization framework

Next we develop an optimization framework to solve the combined location and dispatching decision problem for EMS systems. The optimization is based on GAs. In his review, Goldberg (2004) suggest that the use of spatial queuing (hypercube model) or simulation procedures embedded within a heuristic search offers the greatest utility for real world EMS planning applications. Aytug and Saydam (2002) also comment on the success of GAs in solving combinatorial problems, which make them strong candidates to solve the ambulance location/allocation problem. Iannoni and Morabito (2007) as well as Iannoni et al. (2008) and Geroliminis et al. (2011) have also embedded the hypercube model into genetic algorithms to solve the location problem. As mentioned by Geroliminis et al. (2009) the objective function MRT, as a function of the location space, has many local minima, making it suitable for a global search procedure such as GAs. Jia, Ordóñez, and Dessouky (2007b) proposed a GA to solve the problem of locating facilities to attend large scale emergencies. Shariff, Moin, and Omar (2012) used a GA for solving the MCLP problem applied to healthcare facility location in Malaysia.

Genetic Algorithms were first introduced by Holland (1975) and popularized later by (Goldberg, 1989). GAs are general-purpose, population based search algorithms that resemble the natural selections survival of the fittest. Particular coded schemes (solutions representations) corresponding to chromosomes represent population members. At each iteration individual solutions are evaluated and assigned a fitness value (related to the objective function being optimized). According to their fitness values, solutions are selected to construct the next generation by applying genetic operators: selection, crossover and mutation. Current members of the population are probabilistically selected based on their fitness values, where a high fitness value yields a higher chance of being selected for the next generation. After selection, current solutions may be carried to the next generation without altering (selection), or they may be crossed-over to generate the next set of solutions. Crossover is an operator by which two solutions mutually interchange their current genes. Mutation is an operator that randomly alters the value of a gene of a selected solution. Following the work by Aytug and Saydam (2002), there are five key issues in designing a GA algorithm: (1) Selecting an appropriate solution representation, (2) an effective mutation operator, (3) an effective crossover operator, (4) a feasible initialization and (5) appropriate crossover and mutation rates as well as population size.

5.1. Solution representation

The present work uses the idea of a composite chromosome. That is, a chromosome that is in fact composed of several chromosomes. This representation makes sense given the nature of the problem in which there are two decisions to be made, a location decision and a dispatching decision. Therefore, those two decisions are coded in separate sub-chromosomes. Furthermore, since the dispatching decision is in fact one decision per each demand zone, that gives rise to the idea of having separate chromosomes to represent each demand zone. Fig. 2 shows the composite chromosome for a case in which there are three servers to be located among five candidate locations to attend five demand zones (every demand node is a candidate to locate a server). Note that the chromosome has been divided into sub-chromosomes. The first one deals with the location decision, and therefore has size 5, with the three first components storing the location of a server. The location sub-chromosome stores more information than required, since only a subset of locations will have a server. However, it is kept that way to facilitate feasibility checking as well as the mutation operation, described later. The remaining sub-chromosomes have size 3, representing the order in which every server is ranked to attend a particular demand zone. Note that the sub-chromosome for the location decision corresponds to a permutation of the candidate locations and any sub-chromosome for the dispatching decision corresponds to a permutation of rankings.

5.2. Mutation operator

The standard mutation operator randomly selects a chromosome from the pool, and then goes through every one of its genes changing them randomly with a given probability. Since we are using a composite chromosome, once a chromosome has been selected for mutation the operation should analyze every one of its sub-chromosomes. For we are working with sub-chromosomes that are a permutation, the standard mutation operator is replaced by a swapping operator. It randomly interchanges the positions of two genes within the chromosome, as depicted in Fig. 1. Note that for the location sub-chromosome the swap is done such that the interchange occurs between an assigned location in the current

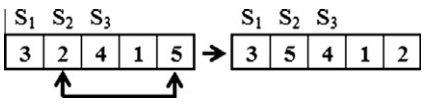

Fig. 1. Swapping mutation operator.

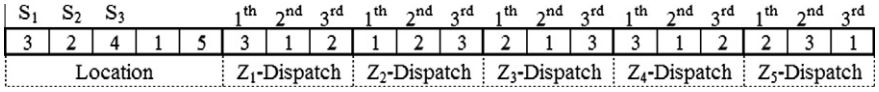

Fig. 2. Composite chromosome.

solution, and a candidate location not yet selected. That is in order to avoid swaps that do not affect the solution.

5.3. Cross-over operation

A single point cross-over operation is used on the implementation of the GA. The recombination of genes is done at sub-chromosomes level, which means that the candidate cross-over points correspond to sub-chromosomes as well. To better understand the way it operates an illustrative example is given in Fig. 3. **A** and **B** are the two parents. **O1** and **O2** represent the two offsprings that it is possible to generate. Parent **A** has been shadowed so that it is possible to trace where the genes of it are going to be after the cross-over operation. The possible crossover points are represented by vertical dashed lines.

5.4. Population initialization

It is usually the case that the initialization is done randomly. The existence of constraints might require us to develop initialization routines that produce feasible solutions. In this case the population of the GA can be randomly initialized, since any permutation for any sub-chromosome will generate a feasible solution. However, it is also possible to use an initialization procedure to create 'good' initial solutions, using the available knowledge about the problem. Initial tests of the GA implementation were done with a random generated population. Based on the result from the enumeration procedure for the small case study presented in Section 4, a better initialization procedure is devised. Since the use of the closest dispatching rule seems to effectively helps in minimizing the response time also providing good coverage, it makes sense to use that information as part of the initialization process. In fact, when solving mid-size problems, the locations are generated randomly during the initialization, but the dispatching is based on the use of the closest servers first.

5.5. Cross-over and mutation rates – population size

In order to test the performance of the GA values of mutation (P_m) and cross-over (P_c) rates are required. Iannoni et al. (2008) used $P_c = 0.5$, and $P_m = 0.06$, while the population size was set to $S = 100$ individuals. In turn, Aytug and Saydam (2002) suggested $P_c = 0.6$ and $P_m = 0.03$, while the population size was set according to $S = \max(100; 0.75n)$, where n is the number of nodes in the problem being solved. The authors argued that for objective functions with potential multiple local optima, there is a trade-off between mutation and cross-over, and that large population sizes are generally favorable, at the cost of computation time. It is also mentioned that the rule of thumb $P_m = 1/L$, where L refers to the length of the chromosome could yield good results. Instead of selecting arbitrarily values for these GA parameters, in the next section we introduce an experimental design to tune-in the parameters of the GA before using it. The implementation of the GA has been done using the Java GA framework developed by Meffert, Meseguer, DMartf, Jerry, and Rotstan (2012).

6. Computational results

6.1. Tuning the GA

A tuning procedure was carried out to find adequate values for several parameters of the GA. The purpose of any experiment is to get the maximum amount of information with the minimum expenditure of resources. A Central Composite Experimental Design (CCD) was used, which according to Montgomery (2008) is widely used because it is highly efficient and flexible. A CCD is normally used to fit a second order polynomial model of a variable of interest. In our case we are not trying to fit a polynomial model. However the combination of factors' values suggested by the CCD provide a good exploration of trade-offs between the different parameters of the GA and its general performance.

There are three parameters that need to be set up: mutation rate, cross-over rate and population size. In experimental design the parameters are called factors, and for each one of these three factors it is necessary to specify a minimum and a maximum value. The minimum and maximum values to be tested for each parameter have been selected according to general recommendations of designing GAs from previous works (Aytug & Saydam, 2002; Iannoni et al., 2008). The CCD also uses the midpoint of the factor (given the minimum and maximum values), as well as the so-called axial points. Axial points correspond to values of the factors that assure that the predicted values of the fitted response surface have the same variance, if the predicted points are at the same distance from the center of the design region (Montgomery, 2008). For the case of three factors a standard CCD requires 20 runs. The first 14 runs correspond to different combinations of the factor's levels, while the last 6 runs correspond to experiments in which each factor is set to its midpoint. A standard CCD does not uses replication. We do use it (30 runs for each combination of factors), as a way to improve the statistical significance of the tests. Instead of the 6 last runs each with one replication, we have a single run setting the factors to their midpoints and we replicate it 30 times. We explore 15 combinations of factors, detailed in Table 3.

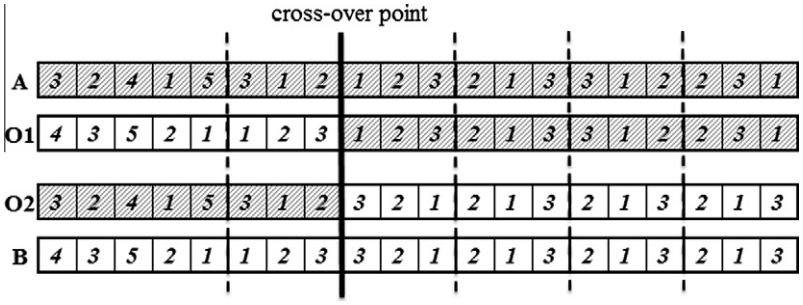

Fig. 3. Single point cross-over for the composite chromosomes.

Table 3
Combination of factors for experimental design.

Combination	P_m	P_c	Pop. size
1	0.02	0.4	30
2	0.05	0.4	30
3	0.02	0.6	30
4	0.05	0.6	30
5	0.02	0.4	100
6	0.05	0.4	100
7	0.02	0.6	100
8	0.05	0.6	100
9	0.01	0.5	65
10	0.059	0.5	65
11	0.036	0.332	65
12	0.036	0.668	65
13	0.036	0.5	6
14	0.036	0.5	123
15	0.036	0.5	65

The minimum and maximum values for the mutation rate are 2% and 5% (the values used in the experiments are then those two, plus the center point, 3.5% and the axial points 1% and 6%). For the Cross-over the minimum and maximum values are 40% and 60% and the population size varies between 30 and 100 individuals. In all the runs of the GA while tuning the parameters, the number of evolutions is set up so that the total number of individuals being evaluated remains constant (approximately equal to 10.000). For example, if the population size is set to 30 then 334 evolutions are performed.

The results from the tuning procedure are given by the box plot graph shown in Fig. 4. It corresponds to the tuning for MRT optimization. As it was mentioned before, for each combination of factors given in Table 3, the GA was run 30 times, applied to different instances and each time using a different random seed. In each case 100 evolutions of the GA were allowed. We have noticed that allowing more evolutions did not further improve the objective value. In order to have a comparison point to tune the GA we enumerate only the location solutions for the case study $\left(\binom{30}{3} = 4060\right)$ possible location decisions. It is not possible to also enumerate the dispatching decisions, which would be computationally prohibitive. For each possible location solution we use the closest rule to set the priority dispatching list of each demand zone. We then compare the performance of the GA (GASol) against the best solution found (BestSol) following the enumeration procedure just described. The *Gap* is calculated as ($Gap = (BestSol - GASol)/BestSol$).

Negative values of the Gap indicate that the GA obtained a solution with a worse objective function value than the best solution coming from the enumeration procedure. If Gap = 0 it basically means that the GA was able to find a solution with the same objective function value. Positives values of the Gap would indicate that the combination of dispatching and location decisions was useful in getting a better value for the objective function. Recall that the gap reported is the average over 30 runs. Out of the 15 combinations of factor's levels under consideration, combinations 7, 8 and 14 showed the best overall performance. All have a Gap close to zero, and exhibited low variability. We ran normality tests on the selected combinations and could not verify the normality of the data. Therefore, we performed a non-parametric test, the Wilcoxon Signed Rank Test, to obtain the Confidence Intervals (CIs) for the three candidate combinations. Table 4 shows the results from the non-parametric test. As it can be seen, the overlapping CIs indicate that statistically there is no difference between the parameters combinations. We decided to use combination 8 which has the smaller CI.

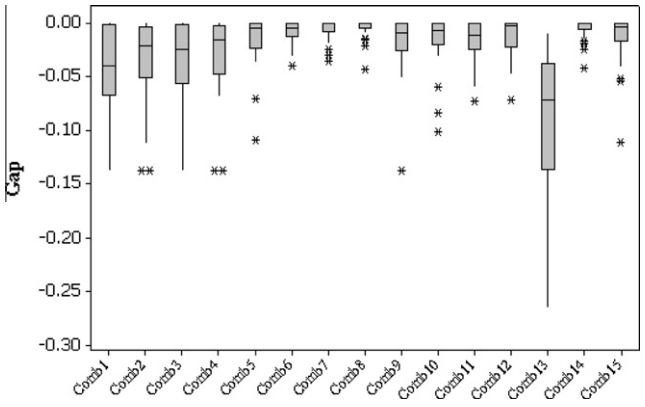

Fig. 4. Results of GA experimental design tuning for MRT minimization.

6.2. Mid-size case study

We solved a bigger, mid-size problem, proposed as an instance of MCLP (http://www.lac.inpe.br/lorena/correa/Q_MCLP_30.txt) (Correa, Chaves, & Lorena, 2007). We analyze several scenarios, varying the number of servers to be located, considering 3 and 4 ambulances. Because of the small number of ambulances we use again an exact solution for the hypercube model. The server rates are obtained by selecting particular values for the overall utilization factor, $\rho = \lambda/(N \times \mu)$. In fact, ρ is varied between 0.1 and 0.9, with increases of 0.1. For the scenarios having three servers we use full backup, which means that any zone can be attended by any of the available servers. In the case of four servers we use partial backup, therefore each demand zone is only allowed to be served by 3 of the available servers. There are two reasons to proceed this way, that have also been suggested by Geroliminis et al. (2009): (i) from a practical perspective, allowing servers that are ranked as 4th and up for a particular demand zone is not desirable, because the overall efficiency of the system would likely decrease; (ii) the calculation of transition rates for the embedded hypercube model becomes very tedious.

For each instance we ran the GA using the tuned parameters and initially minimizing the MRT. In each case we also enumerate the location solutions, and combine them with the use of the closest dispatching policy to have a full solution. That gives us a comparison point. The GA was allowed to run for 100 evolutions. We have noticed that allowing more evolutions does not improve the results. The performance of the GA is compared to the best solution coming from the enumeration procedure. Table 5 shows the results of applying the GA, in each case running it 30 times starting with different initial solutions. The experiments have been run on a PC executing Windows 7-64 Bit, with an Intel® Core 2 Duo processor running at 2.13 GHz and 2 GB of RAM. All the programming was done in Java. The average running time of the GA for the three servers scenarios was 20 s, while the average for the case of four servers was 55 s.

As expected given the low-medium traffic (Jarvis, 1981; Katehakis & Levine, 1986), for the mid-size problem the results suggest

Table 4
Wilcoxon test for obtaining CIs.

		Conf. interval	
Combination	Median gap (%)	Lower	Upper
Comb7	−0.191	−0.823	−0.058
Comb8	−0.208	−0.476	−0.049
Comb14	−0.271	−0.712	−0.109

Table 5
Mid-size case study – MRT results.

	3 Servers scenarios			4 Servers scenarios		
ρ	MRT	Gap (%)	CV	MRT	Gap (%)	CV
0.1	0.587	−0.26	0.0039	0.485	−0.51	0.0054
0.2	0.649	−0.07	0.0015	0.518	−0.27	0.0043
0.3	0.679	−0.08	0.0022	0.536	−0.37	0.0050
0.4	0.703	−0.02	0.0009	0.542	−0.53	0.0041
0.5	0.722	−0.02	0.0001	0.544	−0.78	0.0068
0.6	0.737	−0.03	0.0012	0.551	−0.46	0.0071
0.7	0.750	−0.14	0.0054	0.591	−1.46	0.0160
0.8	0.759	−0.04	0.0014	0.602	−0.98	0.0117
0.9	0.767	0.00	0.0000	0.621	−1.00	0.0113

that a policy that focuses on appropriately selecting locations in combination with dispatching the closest server minimizes mean system response time. These results serve as a validation of the general structure of the mathematical model as well as for the correctness of the optimization procedure. Note that the MRT is smaller when we have an additional server available. Also, for the same number of servers, increasing the utilization of the system also increases the MRT. For the mid-size case studies we have also observed that the expected coverage associated with the solution that minimizes the MRT is smaller (7.2% smaller, on average. 95% CI: 6.14–8.31%) than the maximum observed after the enumeration procedure.

Next we approached the optimization of the system maximizing the expected coverage. A procedure similar of that described in Section 6.1 was followed to tune the GA to be used with the new objective function, Expected Coverage. In this case the combination 7 (from Table 3) showed the best results and therefore was selected as the values for the GA parameters. The enumeration procedure of the location decisions together with a myopic dispatching policy was used again to identify the solution with the highest expected coverage. The performance of the GA was compared against the solution from enumeration. The overall average Gap of the GA compared to the enumeration procedure was −0.87%. The overall mean coefficient of variation of the maximum coverage was 0.0136. These performance measures of the GA show that the algorithm was consistently able to get to the same or to a very close solution from the best found by the enumeration procedure. Compared to the solution that minimizes the response time, the average improvement in coverage is 7.9% (95% CI: 6.64–9.09%). However, this increase comes at a price, a sacrifice of MRT that on average increased by 19.2% (95% CI: 16.35–21.97%). Recall that the expected coverage of the solution minimizing MRT was on average 7.2% smaller than the maximum obtained with enumeration, while the increase on MRT would be on average 19.2% as a result of maximizing coverage. The joint location/allocation approach was not able to improve the solution found by combining the enumeration of locations and the closest dispatching policy.

6.3. Hanover County case study

Here we introduce a case study using real data from the Hanover County Fire/EMS department (Hanover, VA). The county has been divided for planning purposes in 122 demand zones. There are 16 candidate locations for five ambulances. The total demand rate has been estimated in 1.2 calls/h. The average service time per call has been estimated to be 74 min and it is assumed to be independent of the demand zone being served. The system covers 474 square miles and a population nearing 100,000 individuals. For this case study we use partial backup, allowing every demand zone to be served only by three out of the five available servers. The utilization factor for this real system is $\rho = 0.2$. However we consider two variations, increasing the demand by a factor of 1.5 and 2 respectively (which increases the overall utilization, ρ). For this case study we are also using the exact procedure to solve the hypercube model (with five servers the number of states is 32). The average running time of the GA for each scenario of this case study was 280 s. Table 6 shows the results for the Hanover scenarios. The gap reported in Table 6 (third column) is a comparison of the MRT obtained by using the joint location/allocation approach versus the best solution found by solving a location only problem for which the dispatching rule is always sending the closest available vehicle available. As the gap values indicate, the joint approach is able to produce results for the MRT that are around 1% of the same criterion obtained by the location only approach. The fourth column shows the expected coverage associated with the solution that minimizes the response time, while the last column shows the maximum possible coverage (obtained by solving the model with the objective of maximizing expected coverage).

Table 6
Hanover case study – MRT and exp. coverage.

ρ	Min. MRT	Gap MRT (%)	Exp. Cov.	Max. Ex. Cov.
0.2	4.53	−1.48	0.88	0.89
0.3	4.82	−1.08	0.83	0.85
0.4	4.97	−1.03	0.78	0.81

Table 7
Mid-size case study – workloads and individual MRT.

	3 Servers – CV		4 Servers – CV	
ρ	Workload	Ind. MRT	Workload	Ind. MRT
0.1	0.387	0.594	0.357	0.605
0.2	0.250	0.590	0.105	0.529
0.3	0.039	0.565	0.318	0.489
0.4	0.029	0.541	0.335	0.436
0.5	0.038	0.549	0.325	0.512
0.6	0.018	0.539	0.317	0.501
0.7	0.031	0.531	0.612	0.556
0.8	0.028	0.525	0.609	0.551
0.9	0.025	0.520	0.605	0.548

6.4. Non-efficiency criteria

Thus far we have introduced an optimization framework for the joint location/allocation problem, however we have noticed that for the two most common objectives the joint approach is not adding value, since the use of a myopic policy seems to suffice to get to the optimal or near optimal solution. Hence we have turned our attention to calculating other performance indicators for the system. In particular, other works have mentioned the importance of finding solutions in which the total workload is evenly distributed among the available servers, and some others have mentioned that it would be desirable to have individual response times (the mean response time for each demand zone) that do not vary too much among the demand zones. Both performance indicators are associated to the idea of fairness, either from an internal or external point of view. In Table 7 we present the coefficient of variation (CV) for both, mean individual workloads and mean individual response times, resulting from the solutions that optimize mean response time. In this table several instances of high CV values (for example $\geqslant 0.5$) are observed, which implies high variability among server's workloads or demand zones' response time.

Among the several instances of the case study it is possible to notice that variability on individual response times tends to be higher (see Table 7), hence we attempted to improve that perfor-

Table 8
Mid-size case study – optimizing CV resp. times.

	Min CV Ind. RT		Delta	Trade-offs (%)	
ρ	Enum.	Loc./Disp.	CV (%)	MRT	Ex. Cov.
0.1	0.489	0.364	−25.577	37.853	−11.980
0.2	0.504	0.355	−29.652	32.423	−10.461
0.3	0.496	0.367	−25.948	24.485	−4.603
0.4	0.478	0.375	−21.571	19.698	−3.981
0.5	0.465	0.380	−18.285	16.345	−3.911
0.6	0.458	0.384	−16.154	13.764	−3.497
0.7	0.454	0.389	−14.300	11.954	−3.472
0.8	0.481	0.392	−18.636	11.051	−3.098
0.9	0.476	0.397	−16.534	9.612	−3.022

mance indicator by using the optimization approach already developed. Eq. (1) gives the total average response time for the system and in doing so it includes the response time for each demand zone. We use the coefficient of variation (CV) of the individual response times as the new optimization criteria. Once again it was necessary to tune the GA with the new objective function. The GA parameters that perform the best were the same as for the case of MRT minimization. We used the enumeration procedure of the location decisions to get a reference point of the minimum CV for the response times and the compare those solutions with the ones obtained by the joint location/allocation approach. Table 8 shows the results for several instances of the problem, all of them using three servers.

The second column in Table 8 shows the minimum CV for individual response times that was attainted by using the enumeration procedure for the location decisions in combination with the closest dispatching rule. The third column shows the CV that was possible to achieve by using the proposed optimization approach while the fourth column compares the two previous, showing the relative improvement that was possible thanks to the joint approach. The last two columns show the sacrifices in MRT and Expected Coverage that come as a result of the reduction in response times variability across demand zones. We see that there are both an increase in response time and a reduction in coverage. The size of the trade-offs depends upon the utilization (ρ) of the system. The trade-offs were calculated using the solution that minimizes response time as a reference point. For instance, for $\rho = 0.4$ there is a reduction of 21.5% in response time variability, as measured by the coefficient of variation, as well as in increase of about 20% in response time and a reduction of 4% in coverage.

7. Results summary and discussion

We have done extensive computational experiments using 300 small case studies (enumerating more that 70,000 solutions for each instance). We were looking for a better understanding of the potential benefits when location and dispatching decisions are made together for an EMS system. The instances have been generated randomly therefore not favoring any particular result in terms of the decisions being made. Although previous literature had suggested that the existence of demand zones with very different demand rates could lead to situations in which the dispatching based on the closest rule was not optimal, our results were in agreement with some other references showing that using a myopic policy can lead to optimal solutions. We have allowed the demand rates to vary between 1 and 20, therefore introducing differences in the demand rates. What we have found is that if the dispatching policies are designed as a fixed priority list associated to each demand zone, then focusing on finding good locations, and combining them with the use of the closest dispatching rule, yields the desired result of minimizing the mean response time.

In terms of coverage, which is also a common objective to optimize in EMS system, we have used an expected version of coverage, since previous works have made it clear that the standard coverage, which does not take into account the congestion phenomena, overestimates the real coverage. For the small instances we have found that the solutions that maximize the coverage did not use the dispatching policy based on the closest rule. However, we have also noticed that the improved coverage that comes as a result of its maximization, causes a deterioration in the mean response time. As pointed out in Section 4, optimizing the coverage increases it less than 5% (compared to the coverage obtained by the best solution with respect to MRT), while the sacrifice in MRT would be greater than 60%. Those results basically suggest that optimizing the MRT is a better strategy, and that in fact the results in coverage when optimizing MRT are robust, in the sense that the coverage is only 1 or 2% below its optimal value. These results about coverage were also validated with a mid-size real case study found and adapted from previous literature. For the mid-size cases the best coverage was reached when using the closest policy. The average improvement in coverage was again smaller than the average increase in response time.

Results from alternative performance indicators such as those depicted in Table 7, suggest some other observations. Given the values for the CVs it is not surprising that in some cases there are some demand zones with a MRT that doubles that of other zones, or one ambulance having a much heavier workload than the others. Solutions that are good from the point of view of system wide mean response time, can have other performance indicators affected negatively. Since the optimization has been done with a single objective in mind, there is no guarantee of good performance with respect to other criteria. Our results have shown that optimizing the MRT also yields good values for expected coverage. That is convenient since those two are the most common performance indicators used for planning purposes of EMS systems. We illustrated the potential benefits of the joint approach by considering a fairness performance indicator from the user point of view, namely coefficient of variation for individual response times. In this case, the joint approach was able to find better solutions than those that could be reached by using a myopic allocation policy. Of course, the improvement of a fairness objective like the one we have used has consequences, altering other performance indicators such as MRT and coverage. It would be up to the decision maker to balance those trade-offs.

8. Conclusions

Our main goal was to develop an optimization framework for the joint location/allocation problem for EMS systems. We combined the mathematical model and a heuristic solution procedure based on Genetic Algorithms, to be able to solve bigger instances in which enumeration is no longer an option. We were able to validate our approach. The GA has been consistently able to find the same or a pretty close solution to that obtained by full or partial enumeration procedures. In terms of MRT minimization or Expected Coverage maximization we have noticed that the integrated approach do not offer tangible benefits. A more simpler approach considering only the location decision combined with a myopic allocation of the servers based on closest distance would be enough.

One general explanation of the observed behavior is that MRT and Expected Coverage are in fact a function of the distance (time) between servers and demand zones. Hence, locations that reduce the overall distance between servers and costumers tend to dominate the optimization procedure. Although in this case we could just have proposed an optimization procedure in which the deci-

sions are the optimal locations, combined with the use of the closest dispatching rule, we have kept both sets of decisions as part of the optimization framework. We believe that it is important because it gives us the opportunity to attempt the optimization of other performance indicators, so that we can see the trade-offs that are being made as a result focusing on minimizing the response time or maximizing coverage. The fact that solutions that minimize response time offer at the same time a good expected coverage is convenient, since those two criteria are the most commonly used. There is another important consideration: regulations are usually imposed as coverage thresholds, which leads to coverage maximization as preferred optimization criteria. However our results show that the coverage maximization approach can lead to sacrifices in response time that do not compensate for the gains achieved in coverage. On the contrary, minimizing response time offers a good trade-off with respect to the maximum coverage.

We have illustrated two alternative criteria, in particular variability on individual response time, as well as variability on ambulances workloads. Those criteria can be seen as fairness performance indicators from the perspective of internal and external customers. We used individual response time variability as a optimization criteria. As in the case of maximizing the expected coverage, when focused on reducing response time variability among demand zones it is the case that the best solutions do not follow the use of the closest dispatching rule. Furthermore, the improvements that can be made on variability are important, and not only marginal as in the case of maximizing coverage. The proposed optimization framework, already proven to work correctly, can be used to analyze the EMS system from other perspectives, gaining insight into the design of better operation strategies.

As future research directions we will attempt to identify other performance indicators of EMS systems for which the joint location and dispatching problem can yield substantial gains. Another potential area for future research deals with the issue of scalability. We are aware of the limitations of our approach in terms of applying the joint model and its solution procedure to real-sized case studies, basically because the exact solution of the hypercube model will likely require extensive computation time (recall that the exact solution to the hypercube model requires solving a linear system of equations that grows exponentially in size with respect to the number of servers available in the system). However, available approximation procedures that have been suggested in the literature could be embedded in the meta-heuristic optimization framework proposed, hence reducing the computational burden and allowing the solution of bigger instances.

Acknowledgements

The last author was supported by the National Science Foundation (CMMI-1054148). The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the National Science Foundation.

References

Andersson, T., & Varbrand, P. (2006). Decision support tools for ambulance dispatch and relocation. *Journal of the Operational Research Society, 58*, 195–201.

Arabani, A. B., & Farahani, R. Z. (2012). Facility location dynamics: An overview of classifications and applications. *Computers and Industrial Engineering, 62*, 408–420.

Atkinson, J. B., Kovalenko, I. N., Kuznetsov, N., & Mykhalevych, K. V. (2008). A hypercube queueing loss model with customer-dependent service rates. *European Journal of Operational Research, 191*, 223–239.

Aytug, H., & Saydam, C. (2002). Solving large-scale maximum expected covering location problems by genetic algorithms: A comparative study. *European Journal of Operational Research, 141*, 480–494.

Batta, R., Dolan, J. M., & Krishnamurthy, N. N. (1989). The maximal expected covering location problem: Revisited. *Transportation Science, 23*, 277–287.

Benveniste, R. (1985). Solving the combined zoning and location problem for several emergency units. *The Journal of the Operational Research Society, 36*, 433–450.

Brandeau, M. L., & Chiu, S. S. (1989). An overview of representative problems in location research. *Management Science, 35*, 645–674.

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

Budge, S., Ingolfsson, A., & Erkut, E. (2009). Technical note approximating vehicle dispatch probabilities for emergency service systems with location-specific service times and multiple units per location. *Operations Research, 57*, 251–255.

Carbone, R. (1974). Public facility location under stochastic demand. *INFOR, 12*, 261–270.

Carson, Y. M., & Batta, R. (1990). Locating an ambulance on the amherst campus of the state university of New York at buffalo. *Interfaces, 20*, 43–49.

Carter, G. M., Chaiken, J. M., & Ignall, E. (1972). Response areas for two emergency units. *Operations Research, 20*, 571–594.

Chiyoshi, F., Galvao, R. D., & Morabito, R. (2001). Modelo hipercubo: Analise e resultados para o caso de servidores na-homogeneos. *Pesquisa Operacional, 21*, 199–218.

Church, R., & ReVelle, C. R. (1974). The maximal covering location problem. *Papers in Regional Science, 32*, 101–118.

Correa, F., Chaves, A.-A., & Lorena, L. A. N. (2007). Hybrid heuristics for the probabilistic maximal covering location-allocation problem. *Operational Research, 7*, 323–343.

Cuninghame-Green, R.A., & Harries, G., (1988). Nearest-neighbour rules for emergency services. Emmitsburg, MD: National Emergency Training Center.

Daskin, M. S. (1983). A maximal expected covering location model: Formulation, properties and heuristic solution. *Transportation Science, 17*, 48–70.

Farahani, R. Z., Asgari, N., Heidari, N., Hosseininia, M., & Goh, M. (2012). Covering problems in facility location: A review. *Computers and Industrial Engineering, 62*, 368–407.

Galvao, R. D., Chiyoshi, F. Y., & Morabito, R. (2005). Towards unified formulations and extensions of two classical probabilistic location models. *Computers and Operations Research, 32*, 15–33.

Galvao, R. D., & Morabito, R. (2008). Emergency service systems: The use of the hypercube queueing model in the solution of probabilistic location problems. *International Transactions in Operational Research, 15*, 525–549.

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

Gendreau, M., Laporte, G., Semet, F., MontrTal, U. D., Centre-ville, S., J, M. H., et al. (2001). A dynamic model and parallel tabu search heuristic for real-time ambulance relocation. *Parallel Computing, 27*, 1641–1653.

Geroliminis, N., Karlaftis, M. G., & Skabardonis, A. (2009). A spatial queuing model for the emergency vehicle districting and location problem. *Transportation Research Part B: Methodological, 43*, 798–811.

Geroliminis, N., Kepaptsoglou, K., & Karlaftis, M. G. (2011). A hybrid hypercube genetic algorithm approach for deploying many emergency response mobile units in an urban network. *European Journal of Operational Research, 210*, 287–300.

Goldberg, D. (1989). *Genetic Algorithms in search, optimization, and machine learning*. Addison-Wesley Professional.

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

Hakimi, S. L. (1964). Optimum locations of switching centers and the absolute centers and medians of a graph. *Operations Research, 12*, 450–459.

Hogan, K., & Revelle, C. (1986). Concepts and applications of backup coverage. *Management Science, 32*, 1434–1444.

Holland, J. H. (1975). *Adaptation in natural and artificial systems*. University of Michigan Press.

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

Iannoni, A. P., & Morabito, R. (2007). A multiple dispatch and partial backup hypercube queuing model to analyze emergency medical systems on highways. *Transportation Research Part E: Logistics and Transportation Review, 43*, 755–771.

Iannoni, A. P., Morabito, R., & Saydam, C. (2011). Optimizing large-scale emergency medical system operations on highways using the hypercube queuing model. *Socio-Economic Planning Sciences, 45*, 105–117.

Ingolfsson, A., Budge, S., & Erkut, E. (2008). Optimal ambulance location with random delays and travel times. *Health Care Management Science, 11*, 262–274.

Jarvis, J. P. (1981). Optimal assignments in a markovian queueing system. *Computers and Operations Research, 8*, 17–23.

Jarvis, J. P. (1985). Approximating the equilibrium behavior of multi-server loss systems. *Management Science, 31*, 235–239.

Jia, H., Ordóñez, F., & Dessouky, M. (2007a). A modeling framework for facility location of medical services for large-scale emergencies. *IIE Transactions, 39*, 41–55.

Jia, H., Ordóñez, F., & Dessouky, M. M. (2007b). Solution approaches for facility location of medical supplies for large-scale emergencies. *Computers and Industrial Engineering, 52*, 257–276.

Katehakis, M. N., & Levine, A. (1986). Allocation of distinguishable servers. *Computers and Operations Research, 13*, 85–93.

Larson, R. C. (1974). A hypercube queuing model for facility location and redistricting in urban emergency services. *Computers and Operations Research, 1*, 67–95.

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

Larson, R. C., & Odoni, A. R., (1981). *Urban operations research*. Prentice Hall. <http://web.mit.edu/urban_or_book/www/book/>.

Lee, S. (2011). The role of preparedness in ambulance dispatching. *Journal of the Operational Research Society, 62*, 1888–1897.

Li, X., Zhao, Z., Zhu, X., & Wyatt, T. (2011). Covering models and optimization techniques for emergency response facility location and planning: A review. *Mathematical Methods of Operations Research*, 1–30.

Marianov, V., & ReVelle, C. (1992). The capacitated standard response fire protection siting problem: Deterministic and probabilistic models. *Annals of Operations Research, 40*, 303–322.

Marianov, V., & ReVelle, C. (1996). The queueing maximal availability location problem: A model for the siting of emergency vehicles. *European Journal of Operational Research, 93*, 110–120.

McLay, L. A., & Mayorga, M. E. (2010). Evaluating emergency medical service performance measures. *Health Care Management Science, 13*, 124–136.

McLay, L. A., & Mayorga, M. E. (2011). Evaluating the impact of performance goals on dispatching decisions in emergency medical service. *IIE Transactions on Healthcare Systems Engineering, 1*, 185–196.

Meffert, K., Meseguer, J., DMartf, E., Jerry, V., & Rotstan, N., (2012). JGAP – java genetic algorithms and genetic programming package. <http://jgap.sf.net>.

Mendona, F. C., & Morabito, R. (2001). Analysing emergency medical service ambulance deployment on a brazilian highway using the hypercube model. *The Journal of the Operational Research Society, 52*, 261–270.

Montgomery, D. (2008). *Design and analysis of experiments*. Hoboken, NJ: Wiley.

Repede, J. F., & Bernardo, J. J. (1994). Developing and validating a decision support system for locating emergency medical vehicles in Louisville, Kentucky. *European Journal of Operational Research, 75*, 567–581.

Sanchez-Mangas, R., García-Ferrer, A., de Juan, A., & Arroyo, A. M. (2010). The probability of death in road traffic accidents. how important is a quick medical response? *Accident Analysis and Prevention, 42*, 1048–1056.

Schilling, D., Elzinga, D. J., Cohon, J., Church, R., & ReVelle, C. (1979). The team/fleet models for simultaneous facility and equipment siting. *Transportation Science, 13*, 163–175.

Shariff, S. R., Moin, N. H., & Omar, M. (2012). Location allocation modeling for healthcare facility planning in Malaysia. *Computers and Industrial Engineering, 62*, 1000–1010.

Snyder, L. V. (2004). Facility location under uncertainty: A review. *IIE Transactions, 38*, 547–564.

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