INFORMS Journal on Computing

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

The Minimum Expected Penalty Relocation Problem for the Computation of Compliance Tables for Ambulance Vehicles

Thije van Barneveld

To cite this article:
Thije van Barneveld (2016) The Minimum Expected Penalty Relocation Problem for the Computation of Compliance Tables for Ambulance Vehicles. INFORMS Journal on Computing 28(2):370-384. http://dx.doi.org/10.1287/ijoc.2015.0687

INFORMS Journal on Computing
Vol. 28, No. 2, Spring 2016, pp. 370–384
ISSN 1091-9856 (print) | ISSN 1526-5528 (online)

http://dx.doi.org/10.1287/ijoc.2015.0687

The Minimum Expected Penalty Relocation Problem for the Computation of Compliance Tables for Ambulance Vehicles

Thije van Barneveld
Centrum Wiskunde and Informatica, 1098 XG Amsterdam, Netherlands; and VU University Amsterdam, 1081 HV Amsterdam, Netherlands, t.c.van.barneveld@cwi.nl

We study the ambulance relocation problem in which one tries to respond to possible future incidents quickly. For this purpose, we consider compliance table policies: a relocation strategy commonly used in practice. Each compliance table level indicates the desired waiting site locations for the available ambulances. To compute efficient compliance tables, we introduce the minimum expected penalty relocation problem (MEXPREP), which we formulate as an integer linear program. In this problem, one has the ability to control the number of waiting site relocations. Moreover, different performance measures related to response times, such as survival probabilities, can be incorporated. We show by simulation that the MEXPREP compliance tables outperform both the static policy and compliance tables obtained by the maximal expected coverage relocation problem (MECRP), which both serve as benchmarks. Besides, we perform a study on different relocation thresholds and on two different methods to assign available ambulances to desired waiting sites.

Keywords: ambulance relocation; compliance tables; general performance measures; integer linear program
History: Accepted by Allen Holder, Area Editor for Applications in Biology; received March 2015; revised August 2015, October 2015; accepted October 2015. Published online April 21, 2016.

1. Introduction

In life-threatening situations where every second counts, the ability of ambulance service providers to arrive at the emergency scene within a few minutes to provide medical aid can make the difference between survival or death. The location of ambulances has a huge impact on the *response time* to an incident, i.e., the total time between an incoming emergency call and the moment that an ambulance arrives at the emergency scene. To realize short response times, it is crucial to plan ambulance services efficiently. This encompasses a variety of planning problems at the strategic, tactical, and operational level. At the strategic level, the locations of the ambulance *base stations* are determined. A base station is a structure or other area set aside for idle ambulances. In addition, stations often also have a crew room and other facilities for the ambulance personnel. Ambulance staff may be summoned for emergencies by siren, radio, or pagers, depending on the station. When not busy serving patients, the crew usually spends its shift at a base station. At the tactical level, the number of ambulances per base station is specified and, as a direct consequence, the number of ambulance crews per base station. In this paper, we focus on the operational level: the real-time dispatching of ambulances to incidents and the real-time relocation of ambulances.

At certain moments in time, crews may be required to park up at a waiting site away from the base station, to increase coverage of the region. Such a relocation decision is usually made when an event happens, i.e., when a change of the system occurs. Examples of event types are, for instance, a change in availability of ambulances (when an ambulance is dispatched to an incident or when an ambulance finishes service), the arrival of an ambulance at the emergency scene, or the departure to a hospital. However, whether relocations are allowed and if so, to which potential other waiting sites, depends on regulatory rules. For instance, in Vienna, repositioning idle ambulances (apart from sending them back to a waiting site) is not allowed (Schmid 2012), as opposed to Edmonton, Alberta (Alanis et al. 2013). Moreover, the number of allowed potential waiting sites may differ. The number of allowed potential waiting sites may exceed the number of ambulances, as in Gendreau et al. (2006), but this is certainly not the case in general. The setting we consider is the Dutch setting: the dispatcher is allowed to relocate an ambulance from one waiting site to another, but the number of waiting sites is relatively small.

The most common measure on which ambulance service providers base their performance is the fraction of calls reached within some response time or

coverage radius: a demand point is covered if an idle ambulance is present within this coverage radius. Note that this coverage radius is expressed in time, e.g., 12 minutes. At the strategic level, at the tactical level, and at the operational level, the focus is on the search for the best possible coverage, based on the coverage radius.

At the operational level, a commonly used policy structure for the ambulance relocation problem is the use of *compliance tables* (Alanis et al. 2013, Sudtachat et al. 2016, Gendreau et al. 2006). Each row in a compliance table indicates, for a given number of available ambulances, the desired waiting sites for these ambulances. If these ambulances are at their desired waiting sites, the system is *in compliance*. The number of available ambulances changes when a request arrives or when an ambulance becomes available again. Then, each idle ambulance may be assigned to a different waiting site.

A strength of the compliance table policy is that it is simple to explain to and to use by dispatchers: the state of the Emergency Medical Services (EMS) system is only described by the number of available ambulances. Another strength is the ability to calculate compliance tables offline, as opposed to real-time methods which need to compute new relocation decisions whenever an event occurs. After all, there may not be enough time to compute such decisions between two events, although efficient methods exists (Jagtenberg et al. 2015).

1.1. Related Work

Surveys on ambulance location and relocation models are provided by Brotcorne et al. (2003) and Li et al. (2011). In these papers, several deterministic, probabilistic, and dynamic models are reviewed, of which the maximum coverage location problem (MCLP) (Church and ReVelle 1974), and the maximum expected coverage location problem (MEXCLP) (Daskin 1983) are of most interest to this paper. The MEXCLP is an extension of the MCLP in the sense that the MEXCLP takes ambulance unavailability into account. Multiple extensions to the MEXCLP exist; for instance, the ones considered in Batta et al. (1989) and Repede and Bernardo (1994). As an alternative to the MCLP, the p-median problem, proposed in ReVelle and Swain (1970), can be used to determine ambulance locations if the objective is to minimize weighted average response times to incidents.

Ambulance location models for performance measures other than coverage and average response times are considered in Erkut et al. (2008). In this paper, performance is based on the survival probability of a patient suffering from a cardiac arrest. Survival functions proposed in Larsen et al. (1993), Maio et al. (2003), Valenzuela et al. (1997), and Waaelwijn et al. (2001) are incorporated in the MCLP and the MEXCLP. It is shown empirically that survival-maximizing location models are better suited for ambulance location than models based on coverage. In some of these models, probabilistic response times are incorporated based on the work by Ingolfsson et al. (2008). Moreover, McLay and Mayorga (2010) propose a methodology for evaluating the performance of response time thresholds in terms of resulting patient survival rates. In this paper, the model proposed in Larsen et al. (1993) is used, which results in a patient survival probability that is a function of the response time.

A common way to solve the ambulance relocation problem is the online approach: whenever an event occurs, most often when an ambulance becomes available again, the dispatcher has the opportunity to control the system. That is, the current state of the system is observed, and based on that information, a relocation decision is computed. Since these problems typically need to be solved in real-time, the main focus is on heuristics. For instance, in Jagtenberg et al. (2015), a dynamic version of the MEXCLP is proposed to compute a new waiting location for an ambulance that just finished service of a patient. Moreover, in Gendreau et al. (2001), a parallel tabu search heuristic is used for the real-time redeployment of ambulances. In Andersson and Värbrand (2007), the notion of preparedness is used. Preparedness is a measure for the ability to serve potential patients now and in future. Moreover, a dynamic relocation model named DYNAROC and a heuristic to solve this model is presented. In addition, Maxwell et al. (2010, 2013), and Schmid (2012) use approximate dynamic programming for determining relocation strategies.

In contrast to the real-time computation of relocation decisions, many ambulance service providers use prescribed rules or compliance tables as their policy. Real-time computation of such policies is not necessary. In some papers, repositioning is considered preplanned and provides ambulance locations for every time interval on the planning horizon (Rajagopalan et al. 2008, Schmid and Doerner 2010, van den Berg and Aardal 2015). Some literature focuses on optimization of compliance tables, e.g., in Alanis et al. (2013) a two-dimensional Markov chain is proposed to analyze the system performance of compliance table policies. This Markov chain is also used in Sudtachat et al. (2016), in which an integer programming model for the computation of nested compliance tables is proposed, using steady-state probabilities of this Markov chain model as input parameters. The objective is to maximize expected coverage in a system with a single type of ambulance and a single type of call priority.

The model that is of most importance to this paper is the maximal expected coverage relocation problem (MECRP), proposed in Gendreau et al. (2006),

which can be used to compute compliance tables. In Maleki et al. (2014), it is stated that computing compliance tables is just the first part of the computation of relocation decisions. The second part involves the actual assignment of ambulances to waiting sites, and two models minimizing relocation travel times are proposed, based on compliance tables computed by MECRP.

1.2. Contribution

In this paper we present the minimum expected penalty relocation problem (MEXPREP), which is an extension of the MECRP proposed in Gendreau et al. (2006), to compute compliance tables. Although the MECRP (summarized in Section 2.1) is a good and applicable model to compute compliance tables, it has some major limitations:

1. An area is covered if an idle ambulance is present within a certain coverage radius: multiple idle ambulances within the coverage radius do not contribute to the coverage of the area. Especially in an EMS system with a high call arrival rate, it may happen that another incident occurs before the idle ambulances reach the locations to which they are assigned, according to the compliance table. The MECRP does not take this into account—it only focuses on the next future emergency request.

2. There are at least as many waiting site locations as ambulances. This is a rather strong assumption and not generally true in practice. After all, it may be dictated by law that ambulances are allowed to idle at designated ambulance base stations only.

3. The capacity of each waiting site location equals one. This may be true for designated ambulance parking spaces, but in general not for base stations.

4. Only a performance measure related to coverage can be incorporated.

As a consequence of limitations 1 and 3, each waiting site location occurs at most once in each compliance table level. However, it could be beneficial to locate multiple ambulances at a waiting site—e.g., at a waiting site in the middle of a densely populated area with a high call arrival rate—to anticipate a possible rapid succession of incidents occurring in that area. In addition, we are forced to do this in a system in which limitation 2 does not hold. We extend the MECRP in such a way that within a compliance table level, a waiting site can occur multiple times. We do this by incorporating the objective function of the maximum expected coverage location problem (MEXCLP), presented in Daskin (1983), into the objective function of the MECRP.

The last limitation is related to coverage. As pointed out in Maio et al. (2003), the most common EMS standard is to respond to 90% of all urgent calls within eight minutes. Many EMS systems use the percentage of calls covered as a performance measure. However, as stated in Erkut et al. (2008), the black-and-white nature of the coverage concept is an important limitation, and standard coverage models should not be used for ambulance location. First, coverage can result in large measurement errors because of their limited ability to discriminate between different response times. Second, these measurement errors are likely to result in large optimality errors when one uses covering models to locate ambulances instead of a model that takes survival probabilities into account. The difference between "coverage" and "survival" is demonstrated by an artificial example in Erkut et al. (2008), and it is shown that covering models can result in arbitrarily poor location decisions for ambulances.

In the MECRP, only the performance measure of coverage can be incorporated. The MEXPREP we propose in this paper is an extension of the MECRP in which a general performance measure can be incorporated, including the concept of survival previously mentioned. We do this by introducing a penalty function, which is a nondecreasing function that solely depends on the response time (hence the name minimal expected penalty relocation problem).

The remainder of this paper is organized as follows: in Section 2.1 we explain the MECRP of Gendreau et al. (2006). In Sections 2.2 and 2.3, we treat the limitations mentioned above, resulting in the formulation of the MEXPREP in Section 2.4. In Section 3, we consider two models for the assignment problem, which needs to be solved to obtain an assignment of available ambulances to the waiting sites corresponding to the compliance table level. We conclude the paper by a numerical study in Section 4.

2. Mathematical Model

One method used to compute compliance tables is solving MECRP, presented in Gendreau et al. (2006). In this section, we will extend MECRP. Next, we proceed with a summary of this problem.

2.1. Maximal Expected Coverage Relocation Problem

The MECRP is defined on a directed graph $G = (V \cup W, A)$ representing the region of interest. The region is discretized into demand zones, e.g., postal codes, in which V is the vertex set of these demand points. Moreover, W is the vertex set of potential waiting sites for n emergency vehicles and A is a set of arcs defined on $(V \cup W)^2$. A travel time is associated to each arc $(i, j) \in A$ and d_i denotes the demand at vertex $i \in V$. This d_i may, for instance, correspond to the population of demand zone i, or to the probability that an incoming emergency call occurs in demand zone i, which can be estimated by analyzing historical data. A vertex i is said to be covered by a vertex $j \in W$ if the

expected travel time from j to i, denoted by τ_{ji}, is less than a given coverage radius r, expressed in time. We denote by W_i the subset of vertices of W covering i.

In MECRP, the *busy fraction* p plays an important role. This is the probability that an ambulance is busy, i.e., responding to an emergency call, or serving or transporting a patient. This busy fraction could be computed by $p = \lambda/(n\mu)$, where λ is the call arrival rate, μ is the average service rate and the number of ambulances is n. This busy fraction may also be estimated by analysis of historical data. The probability of being in a situation with k available ambulances, denoted by q_k, is easily computed by means of a binomial distribution:

$$q_k = \binom{n}{k}(1-p)^k p^{n-k}, \quad k = 0, \ldots, n. \tag{1}$$

As was pointed out in Gendreau et al. (2006), a simple relaxation procedure for the MECRP consists of solving MCLP (presented in Church and ReVelle 1974) for each compliance table level $k = 1, \ldots, n$. This procedure produces a compliance table, but it ignores constraints on waiting site changes at each event. To incorporate such constraints, it is useful to view the system as being in a succession of states k over time, where k is the number of available ambulances. In the remainder of the paper, we will call the row of the compliance table level with k waiting sites the kth level of the compliance table, which indicates the desired waiting sites for k available ambulances. This compliance table level k is described by binary variables x_{jk} equal to 1 if and only if an ambulance is located at $j \in W$, and by binary variables y_{ik} equal to 1 if demand point i is covered by at least one ambulance in compliance table level k. Moreover, a bound α_k is imposed on the number of waiting site changes between compliance table levels k and $k+1$, where $1 \le k \le n-1$. As a consequence, binary variables u_{jk} are defined, which equal 1 if and only if $j \in W$ ceases to be a waiting site in compliance table level $k+1$, starting from level k. The MECRP is formulated as follows:

MECRP:

$$\text{Maximize} \quad \sum_{k=1}^{n}\sum_{i\in V} d_i q_k y_{ik} \tag{2}$$

$$\text{Subject to:} \quad \sum_{j\in W_i} x_{jk} \ge y_{ik} \quad i \in V, k = 0, 1, \ldots, n \tag{3}$$

$$\sum_{j\in W} x_{jk} = k \quad k = 0, 1, \ldots, n \tag{4}$$

$$x_{jk} - x_{j,k+1} \le u_{jk} \quad j \in W, k = 1, \ldots, n-1 \tag{5}$$

$$\sum_{j\in W} u_{jk} \le \alpha_k \quad k = 1, \ldots, n-1 \tag{6}$$

$$x_{jk} \in \{0,1\} \quad j \in W, k = 0, 1, \ldots, n \tag{7}$$

$$y_{ik} \in \{0,1\} \quad i \in V, k = 0, 1, \ldots, n \tag{8}$$

$$u_{jk} \in \{0,1\} \quad j \in W, k = 1, \ldots, n-1. \tag{9}$$

In this model, the objective function (2) maximizes the expected coverage. Constraints (3) induce that vertex $i \in V$ is covered only if at least one ambulance is located in at least one of the waiting sites in W_i, in compliance table level k. Constraints (4) ensure that exactly k waiting sites are occupied in compliance table level k. Constraints (5) and (6) control the number of waiting site changes between compliance table levels k and $k+1$. The designated waiting sites at compliance table level k are given by decision variables x_{jk}. Although $k = 0$ is included in the original MECRP by Gendreau et al. (2006), it is not necessary to include this case.

2.2. Expected Covered Demand

In the MECRP, the objective function for a given compliance table level k is to maximize the demand covered within the response time threshold. Then, each level is weighted according to q_k, the probability of being in a situation with k available ambulances, which can be computed using Equation (1). As stated in Gendreau et al. (2006), the MECRP reduces to the MCLP with k ambulances if $q_k = 1$. After all, k ambulances are always available, because $q_i = 0$ for $i \ne k$.

Although the MCLP is a useful method for determining ambulance base locations, it has a major shortcoming: it assumes there is always an ambulance available at a base location. In practice, this is not true, since ambulances may be busy serving a patient. The fraction of duty time an ambulance is busy serving a patient is the definition of the earlier mentioned busy fraction p. As a consequence of this limitation, it makes no sense in the MCLP to locate multiple ambulances at one location. This shortcoming was addressed in Daskin (1983) by proposing the maximum expected coverage location problem (MEXCLP), which was one of the first probabilistic models for ambulance location.

In the MEXCLP, the busy fraction is incorporated as follows: if vertex $i \in V$ is covered by k ambulances, the expected covered demand is $d_i(1-p^k)$. Moreover, the marginal contribution of the kth ambulance equals $d_i(1-p)p^{k-1}$. This expression is incorporated in the objective value of MEXCLP:

MEXCLP:

$$\text{Maximize} \quad \sum_{i\in V}\sum_{k=1}^{n} d_i(1-p)p^{k-1} z_{ik} \tag{10}$$

$$\text{Subject to:} \quad \sum_{j\in W_i} x_j \ge \sum_{k=1}^{n} z_{ik} \quad i \in V \tag{11}$$

$$\sum_{j\in W} x_j \le n \tag{12}$$

$$x_j \in \{0, 1, \ldots, n\} \quad j \in W \tag{13}$$

$$z_{ik} \in \{0, 1\} \quad i \in V,\ k = 1, \ldots, n. \tag{14}$$

Here, $z_{ik} = 1$ if and only if vertex i is covered by at least k ambulances. Note that constraint (12) is an inequality, while its MCLP counterpart is an equality. This is due to the concavity of the objective function in k for each i, which implies that if $z_{ik} = 1$, then $z_{i1} = z_{i2} = \cdots = z_{ik} = 1$ and if $z_{il} = 0$, then $z_{i,l+1} = z_{i,l+2} = \cdots = z_{in} = 0$. Moreover, the objective is to be maximized. Hence, constraint (12) will be satisfied at equality.

Analogous to the extension of the MCLP to the MEXCLP, we will extend the MECRP to address the first three shortcomings of the MECRP mentioned in Section 1.2. This is done by replacing the objective function of the MECRP, expression (2), by the following objective function:

$$\text{Maximize} \sum_{i \in V} \sum_{k=1}^{n} \sum_{l=1}^{k} d_i q_k (1-p) p^{l-1} z_{ikl}, \tag{15}$$

where $z_{ikl} = 1$ if and only if, in compliance table level k, vertex i is covered by at least l ambulances. Otherwise, $z_{ikl} = 0$. Moreover, constraint (3) is replaced by

$$\sum_{j \in W_i} x_{jk} \geq \sum_{l=1}^{k} z_{ikl}, \quad i \in V,\ k = 1, \ldots, n. \tag{16}$$

This constraint is satisfied at equality by the same reasons as before. None of the other constraints of the MECRP change, except for constraints (7) and (8), which become $x_{jk} \in \{0, 1, \ldots, n\}$ and $z_{ikl} \in \{0, 1\}$, where $j \in W$, $i \in V$, $k = 1, \ldots, n$ and $l = 1, \ldots, k$. Moreover, constraint (9) is changed into $u_{jk} \in \{0, 1, \ldots, n\}$, where $j \in W$ and $k = 1, \ldots, n-1$.

2.3. General Performance Measures

As stated in Section 1.2, another limitation of the MECRP is the inability to incorporate EMS performance measures other than coverage, such as patient survival. This is a limitation of the MCLP and the MEXCLP as well. In this section, we demonstrate how to incorporate different objectives in the MECRP. Similar to van Barneveld et al. (2015), we do this by introducing a non-negative nondecreasing penalty or cost function Φ, which is a function of the response time solely, with domain $\mathbb{R}_{\geq 0}$. A penalty function assigns a penalty to each different response time, and thus several performance measures related to response times can be incorporated. The commonly used EMS performance measure of coverage can be translated into the penalty function $\Phi(t) = \mathbb{1}_{\{t > r\}}$, where t denotes the response time and r the coverage radius. Other examples of objectives could be minimizing the average response time or minimizing the average lateness, modeled by penalty functions $\Phi(t) = t$ and $\Phi(t) = \max\{0, t - r\}$, respectively. In Erkut et al. (2008), survival functions are considered, which we can use as penalty function as well (see Section 4).

To incorporate penalty functions, and thus general performance objectives in the MECRP framework, we must be aware of the fact that coverage does not play a role here: we cannot use the set W_i defined in our model formulation. After all, even an ambulance positioned at a location for which the travel time between this location and vertex i exceeds the coverage radius, has an effect. This effect gets larger and larger if fewer ambulances are available. Hence, all available ambulances are of influence on the ability to respond to a request for each vertex. In contrast, ambulances outside the coverage radius of a certain vertex i are treated as nonexistent ones for this vertex, if one uses the 0-1 nature of coverage.

As a consequence, constraint (3) of the MECRP needs to be replaced by a different constraint, which is able to take all available ambulances for each vertex into account. That is, for each vertex i, we need an ordering of ambulances according to their expected travel time to i, because we incorporated ambulance unavailability in our model: with probability $1-p$ the closest ambulance will respond to the request, generating a certain penalty $\Phi(t_1)$, and with probability $(1-p)p$ the second closest ambulance will respond, generating penalty $\Phi(t_2) \geq \Phi(t_1)$, and so on up to the kth ambulance for compliance table level k. Moreover, to specify $\Phi(t_1), \Phi(t_2), \ldots, \Phi(t_k)$ for compliance table level k, we need to incorporate the expected travel times $t_1, t_2, \ldots, t_k$ in our model, because the penalty function relies on these.

As previously stated, the expected travel time from waiting site $j \in W$ to demand point $i \in V$ is denoted by τ_{ji}. If $\tau_{ji} \leq \tau_{j'i}$, then it holds that $\Phi(\tau_{ji}) \leq \Phi(\tau_{j'i})$, from the definition of the penalty function. Moreover, for the ordering of ambulances, we define $z_{ijkl} = 1$ if and only if for compliance table level k, the lth closest ambulance to vertex i is at waiting site j. We need to introduce the constraint $\sum_{j \in W} z_{ijkl} = 1$ to ensure that at compliance table level k, there is exactly one ambulance that is the lth closest to i. Now we have all the ingredients to formulate the minimal expected penalty relocation problem (MEXPREP).

2.4. Minimal Expected Penalty Relocation Problem

The MEXPREP is formulated as follows:

$$\text{Minimize} \sum_{k=1}^{n} \sum_{l=1}^{k} \sum_{i \in V} \sum_{j \in W} q_k d_i (1-p) p^{l-1} \Phi(\tau_{ji}) z_{ijkl} \tag{17}$$

$$\text{Subject to:} \sum_{l=1}^{k} z_{ijkl} = x_{jk} \quad i \in V,\ j \in W,\ k = 1, \ldots, n \tag{18}$$

$$\sum_{j\in W} z_{ijkl} = 1 \qquad i\in V, k=1,\dots,n, l=1,\dots,k \quad (19)$$

$$\sum_{j\in W} x_{jk} = k \quad k=1,\dots,n \quad (20)$$

$$x_{jk} - x_{j,k+1} \le u_{jk} \qquad j\in W, k=1,\dots,n-1 \quad (21)$$

$$\sum_{j\in W} u_{jk} \le \alpha_k \quad k=1,\dots,n-1 \quad (22)$$

$$x_{jk} \in \{0,1,\dots,n\} \quad j\in W, k=1,\dots,n \quad (23)$$

$$z_{ijkl} \in \{0,1\} \quad i\in V, j\in W, k=1,\dots,n, \; l=1,\dots,k \quad (24)$$

$$u_{jk} \in \{0,1,\dots,n\} \qquad j\in W, k=1,\dots,n-1. \quad (25)$$

Note that there is only a contribution to the objective value if $z_{ijkl} = 1$, i.e., if for compliance table level k, the lth closest ambulance to vertex i is at waiting site j. The marginal contribution of this lth closest ambulance to vertex i is $d_i(1-p)p^{l-1}\Phi(\tau_{ji})$ for given vertex i, waiting site j, and compliance table level k. That is, with probability $(1-p)p^{l-1}$, the lth closest ambulance to vertex i is the closest available one, inducing a penalty of $d_i\Phi(\tau_{ji})$. Like in the MECRP, each compliance table level k is weighted according to the probability that the system is in a situation with k available ambulances, as computed in (1).

Constraints (18) and (19) take over the role of constraint (3) in the MECRP formulation. In constraint (18), both the left- and the right-hand side represent the number of ambulances at waiting site j for compliance table level k. Note that no i-index is present in the right-hand side. Since constraint (18) holds for each $i \in V$, it is immediately forced that

$$\sum_{l=1}^{k} z_{i_1jkl} = \sum_{l=1}^{k} z_{i_2jkl}, \quad i_1, i_2 \in V, j\in W, k=1,\dots,n. \quad (26)$$

This should hold in a feasible solution to the problem, since for level k all the ambulances at waiting site j contribute to the penalty induced by each demand point in the objective function. As stated before, constraint (19) ensures that at compliance table level k, there is exactly one ambulance that is the lth closest to i. All the other constraints are the same as the constraints in the MECRP formulation, except for the integer and binary constraints. Note that since the objective is to be minimized and the penalty function Φ is non-decreasing, we do not require constraints related to the ordering of ambulances.

2.5. Adjusted MEXPREP

In the MEXCLP-formulation of Daskin (1983), some simplifying assumptions are made: ambulances operate independently; each ambulance has the same busy fraction; and ambulance busy fractions are invariant with respect to the ambulance locations. Moreover, the MEXPREP formulation, like the formulations of MEXCLP and MECRP, all assume that the busy fraction is an input. However, in reality, the busy fraction p is an output as the service rate that is needed to calculate the busy fraction depends on the allocation of ambulances to waiting sites. The use of a universal busy fraction is a rough approximation of reality, since the actual busy fractions depend on the compliance table itself and on the dispatch policy.

Batta et al. (1989) consider an adjustment of the objective function in MEXCLP, relaxing the assumptions on busy fractions. In this problem, called AMEXCLP, correction factors $Q(n,p,k)$, $k = 0,\dots,n-1$, derived in Larson (1975), are incorporated in the objective function of MEXCLP. We extend MEXPREP to AMEXPREP by incorporating the correction factors $Q(n,p,k-1)$ in Equation (17), where

$$Q(n,p,k) = \frac{\sum_{j=k}^{n-1} \frac{(n-k-1)!(n-j)}{(j-k)!} \frac{n^j}{n!} p^{j-k}}{(1-p)\sum_{i=0}^{n-1} \frac{n^i}{i!} p^i + \frac{n^n p^n}{n!}}, \qquad k=0,\dots,n-1, \quad (27)$$

analogous to the work done by Batta et al. (1989). In Section 4.6, we will explore the differences between MEXPREP and AMEXPREP.

3. Assignment Problem

Determining the compliance table is just the first part of the ambulance relocation problem. The second part is related to the actual assignment of the k available ambulances to the k waiting sites occurring in compliance table level k. This problem is studied extensively in Maleki et al. (2014), and two models for determining the assignment of ambulances to the waiting sites in compliance table level k (as computed via solving the MECRP) are proposed. In each of these two models, called the generalized ambulance assignment problem (GAAP) and the generalized ambulance bottleneck assignment problem (GABAP), a different, yet related, objective is incorporated: GAAP minimizes the total travel time travelled by all ambulances to attain the configuration of the compliance table level, while GABAP minimizes the maximum travel time. Both, like the MECRP, are offline methods, computing assignments beforehand. However, scalability issues are present, since the number of combinations between hospitals/waiting sites and waiting sites grows very rapidly.

As opposed to the offline approach of Maleki et al. (2014), we use an online approach in our computations, by modeling the assignment problem as either a minimum weighted bipartite matching problem (MWBMP) or a linear bottleneck assignment problem (LBAP), similar to van Barneveld et al. (2015). By modeling the problem as a MWBMP, we aim to find an assignment of available ambulances to the designated waiting sites in the compliance table that minimizes the total travel time. However, in the assignment, it may happen that one ambulance needs to make a very long trip. Hence, the area around the waiting site to which this ambulance is assigned is vulnerable for a long time. It may be advantageous to minimize the maximum travel time, and thus the time until the system is in compliance. This can be done by modeling the assignment problem as an LBAP.

In contrast to the computation of compliance tables, fast methods exist for solving MWBMP and the LBAP, e.g., the Hungarian method of complexity $\mathcal{O}(n^3)$ for MWBMP and the threshold algorithm of complexity $\mathcal{O}(n^{2.5}/\sqrt{\log n})$ for LBAP (Burkhard et al. 2009). Hence, this can be done in real time and an offline solution is not necessary. After all, this would require a complex state dependent policy, which shows relocation moves for every realized state of the system. Moreover, an online implementation of the assignment problem takes into account the actual locations of driving ambulances and hence a redirection of ambulances to different waiting sites. Therefore, we recommend computing compliance tables offline, and the assignment problem online. In Section 4.4, we will explore the differences in MWBMP and LBAP.

4. Computational Study

MEXPREP computes compliance tables taking into account ambulance unavailability, general performance measures, and a restriction on the number of waiting site changes. We apply MEXPREP to an EMS region in The Netherlands, particularly the capital of Amsterdam and its surrounding region. Results are generated by simulation using historical data.

4.1. Experimental Setup

The EMS region of Amsterdam and the surrounding areas is an amalgamation of two former EMS regions: the semirural Zaanstreek-Waterland (North) and the urban Amsterdam-Amstelland (South). The region is displayed in Figure 1(a). This region covers approximately 630 km^2 and is home to 1.2 million inhabitants, of which 68% live in Amsterdam itself. Ambulance waiting sites are at the 17 nodes in Figure 1, of which locations 2–4, 6–8, 11, 13, and 16 are hospitals. Hence, $|W| = 17$. The numbers in brackets denote the actual waiting site capacities. These restrictions could be incorporated into all methods in Section 2, but in the computational study we do not consider these, apart from Section 4.7.

We aggregate the region into 162 demand points based on four-digit postal codes, hence, $|V| = 162$. In our computations, we use two different travel times. For the average emergency travel times between the vertices, we use travel times estimated by the RIVM,[1] which provided us a 162×162 table of travel times between the four-digit postal codes in this region. We refer to Kommer and Zwakhals (2008) for a more detailed description of this travel time model, which we summarize in the online supplement (available as supplemental material at http://dx.doi.org/10.1287/ijoc.2015.0687). The relocation travel times were computed by multiplying the emergency travel times by a factor 10/9.

Moreover, historical data on emergency requests in the year 2011 was provided by Ambulance Amsterdam, which runs the emergency medical services in this region. We only consider the time-period between 7 A.M. and 6 P.M., because during the evening and night a different number of ambulances is on duty. During the considered time period, 33 ambulances are present in the system. However, of these 33 ambulances, many are busy with ordered transport: taxi-type transport of patients not able to travel to the hospital themselves, usually scheduled in advance. Therefore, we assume a fleet size of 21 in our computations.

In 2011 between 7 A.M. and 6 P.M., the total number of emergency requests was 44,966, yielding an hourly arrival rate of 11.2 requests. Only 44,520 of these requests are useful, because historical data of the remainder was not complete. We use this historical data to compute the busy fraction by dividing the total patient-related work during these 4,015 hours by the total duty time of 21 ambulances to obtain a busy fraction of $p = 0.43047$. The average busy time (excluding relocation time after transferring the patient at the hospital) of an ambulance is 0.82 hours. The annual number of emergency requests ranges between two (in a postal code somewhere between waiting sites 9 and 13) and 1,545 (in the city center of Amsterdam, near waiting site 1), with an average demand of 275 per node. We define d_i as the probability that an incoming request occurs in vertex i, computed by normalization of the number of emergency requests.

We assume a deterministic dispatch time of 120 seconds and a deterministic pre-trip delay of 60 seconds for ambulances at a waiting site. There is no pre-trip delay if the dispatched ambulance is already on the road. Moreover, the pre-trip delay for moving

[1] Rijksinstituut Volksgezondheid en Milieu (National Institute for Public Health and the Environment).

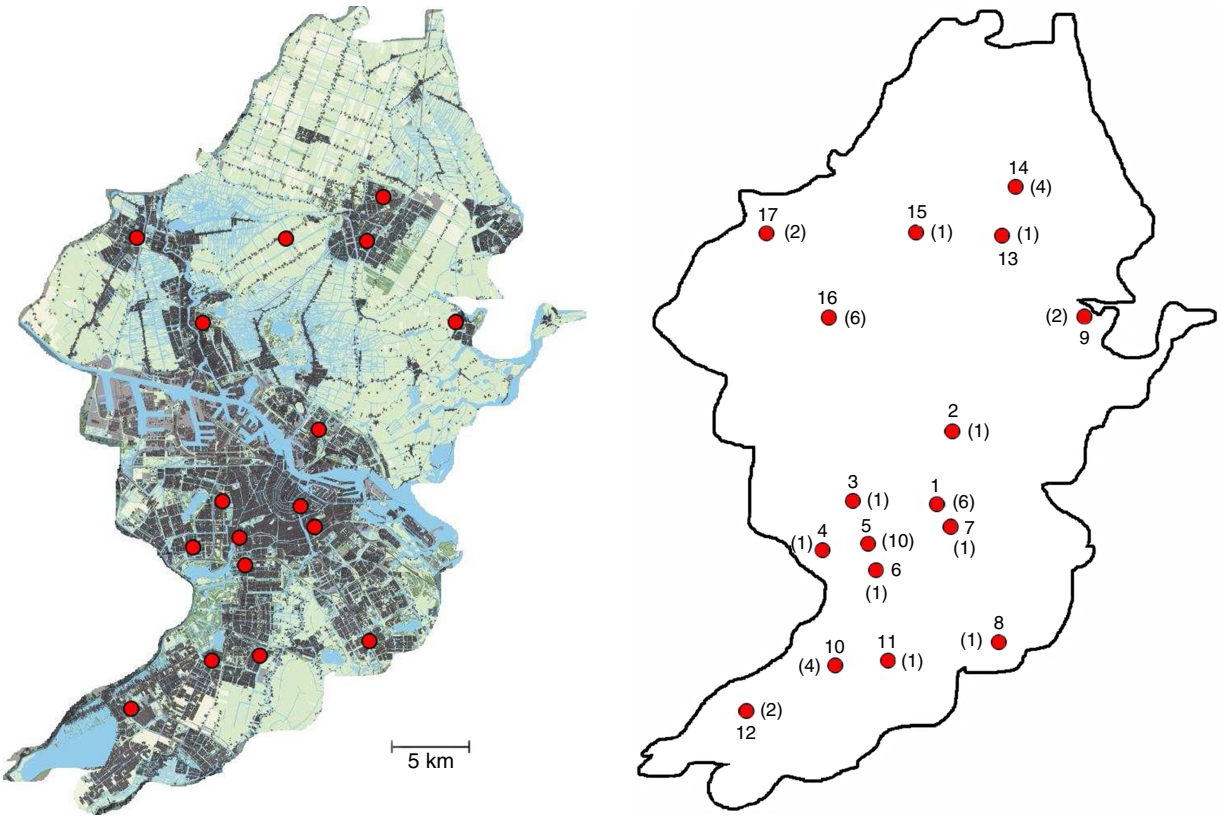

Figure 1 **(Color online) EMS Region of Amsterdam**

an ambulance from a waiting site to another one is assumed to be 180 seconds. The ambulance that can be present fastest at the emergency scene is always dispatched to the request.

We perform simulations using the computed compliance tables and the actual emergency requests in the region during the daytime of the year 2011. We use the following historical data of these requests: time and place (based on four-digit postal codes) of occurrence, the on-scene time of the ambulance, whether the patient needed transportation to a hospital, and the hospital time of the ambulance. No randomness is involved, as we use the actual historical data (trace-driven). The simulation model is coded in MATLAB. Computation of the assignment of ambulances to waiting sites is done online by solving either the MWBMP or LBAP during the simulation. We test performance according to six statistics:

1. Percentage requests responded to within the response time threshold (720 seconds)
2. Average penalty per request
3. Average response time
4. Average number of relocations per ambulance per day. A move of an ambulance only counts as relocation if this move is induced by carrying out the compliance table policy.
5. Average relocation time
6. Computation time to solve the model, run with CPLEX 12.6 on a 2.2 GHz Intel(R) Core(TM) i7-3632QM laptop with 8 GB RAM

In our computations, we consider five different penalty functions. Three of them are based on survival functions, considered in Maio et al. (2003), Valenzuela et al. (1997), and Waaelwijn et al. (2001). These three functions all relate a survival probability to a response time, in the case of a cardiac arrest. However, these survival probabilities depend on additional factors rather than just the response time, e.g., whether the collapse of a patient was witnessed by the ambulance crew, the duration from collapse to defibrillation, and the duration from collapse to cardiopulmonary resuscitation (CPR). These three survival functions are considered in Erkut et al. (2008), and assumptions on these factors are made. We follow these assumptions to obtain a survival function solely depending on the response time (in seconds). The considered penalty functions are as follows:

$$\Phi_1(t) = \mathbb{I}_{\{t>720\}} \tag{28}$$

$$\Phi_2(t) = t \tag{29}$$

$$\Phi_3(t) = 1 - (1 + e^{0.679+0.0044t})^{-1} \tag{30}$$

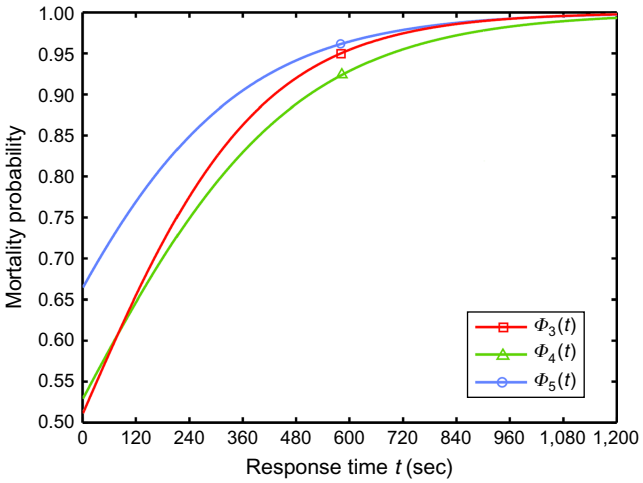

Figure 2 **(Color online) Mortality Probabilities as a Function of the Response Time**

$$\Phi_4(t) = 1 - (1 + e^{0.113+0.0041t})^{-1} \tag{31}$$

$$\Phi_5(t) = 1 - (1 + e^{0.04+0.005t})^{-1}. \tag{32}$$

Function Φ_1 is based on coverage, in which we consider a response time threshold of 720 seconds (12 minutes). Function Φ_2 represents the penalty function focusing on the objective of minimizing the average response time. Functions Φ_3, Φ_4, and Φ_5 represent the survival functions of Maio et al. (2003), Valenzuela et al. (1997), and Waaelwijn et al. (2001), respectively, in a penalty function (mortality) setting. A graphical representation of Φ_3, Φ_4, and Φ_5 is given in Figure 2.

4.2. Comparison of MEXPREP with MECRP

First, we compare the compliance tables obtained by MEXPREP with the ones obtained by MECRP, following the formulation proposed in Gendreau et al. (2006). We do this for the coverage-based penalty function $\Phi(t) = \mathbb{1}_{\{t>r\}}$, since MECRP cannot take other penalty functions into account. We use a coverage radius of $r = 720$ seconds (12 minutes), and compute compliance tables for different values of α_k. Due to the inability of the MECRP to consider systems with more ambulances than waiting sites, which is the case here, we compare MEXPREP with MECRP on two different settings: a setting with 17 ambulances instead of 21; and a setting in which we have 21 ambulances, but the compliance table will be carried out only if 17 or fewer ambulances are available. If more than 17 ambulances are available, ambulances that finish service of a patient return to their home waiting site. In the first setting, the busy fraction is 0.53175, whereas in the second setting the busy fraction equals 0.43047 as mentioned before.

We only display the compliance tables for the $\alpha_k = 0$ case, since these compliance tables are nested and thus can be represented efficiently. We represent such a nested compliance table by a one-dimensional vector, where compliance table level k is given by entries 1 up to k. The computed MECRP and MEXPREP compliance tables for $\alpha_k = 0$, for the two different settings are displayed in (33) and (34), respectively. Note that none of these four compliance tables equals another, although the two MECRP-tables are very similar. Simulation results, using MWBMP as assignment policy, for these compliances tables are listed in Tables 1 and 2, respectively. These tables include 95% confidence intervals around the percentage of requests responded to within 720 seconds.

$$\begin{aligned} \text{MECRP:}\ & (1, 16, 12, 14, 5, 9, 17, 8, 11, 15, \\ & 3, 10, 4, 13, 2, 6, 7) \\ \text{MEXPREP:}\ & (1, 1, 6, 16, 6, 15, 2, 10, 16, 14, 1, \\ & 10, 15, 9, 6, 17, 12) \end{aligned} \tag{33}$$

$$\begin{aligned} \text{MECRP:}\ & (1, 16, 12, 14, 5, 9, 17, 8, 10, 15, \\ & 11, 3, 4, 13, 2, 6, 7) \\ \text{MEXPREP:}\ & (1, 6, 16, 1, 15, 10, 2, 14, 6, 16, \\ & 10, 9, 17, 2, 12, 14, 5) \end{aligned} \tag{34}$$

Table 1 **Simulation Results for 17 Ambulances and Penalty Function $\Phi(t) = \mathbb{1}_{\{t>720\}}$, Based on 44,520 Requests in 2011**

Method	Performance indicators	$\alpha_k = 0$	$\alpha_k = 1$	$\alpha_k = \lceil k/2 \rceil$	$\alpha_k = k$
MECRP	Percentage on time (%)	86.55	86.29	86.62	86.60
	Lower bound 95%-CI (%)	86.24	85.97	86.31	86.28
	Upper bound 95%-CI (%)	86.87	86.60	86.94	86.92
	Average response time (s)	473	476	474	474
	Mean no. of relocations	1.62	2.14	3.86	3.72
	Average relocation time (s)	646	576	451	457
	Computation time (s)	<1	<1	<1	<1
MEXPREP	Percentage on time (%)	88.23	88.18	88.18	88.34
	Lower bound 95%-CI (%)	87.93	87.88	87.88	88.04
	Upper bound 95%-CI (%)	88.53	88.48	88.48	88.64
	Average response time (s)	461	461	461	460
	Mean no. of relocations	1.30	1.31	1.31	1.54
	Average relocation time (s)	625	616	616	571
	Computation time (s)	76	85	85	77

Note that in Table 1 as well as in Table 2, the MEXPREP significantly outperforms the MECRP on the most important performance indicator: the percentage of requests responded to within the response time threshold of 720 seconds. We observe improvements on this criterion between 0.7% (second setting, $\alpha_k = 0$) and 1.89% (first setting, $\alpha_k = 1$). Moreover, this performance gain is achieved with fewer relocations, although the average relocation time is longer for MEXPREP. A small disadvantage of the MEXPREP compared to the MECRP is the computation time. However, as stated in Section 1, the computation time of the MEXPREP compliance tables is of less importance, since the problem can be solved offline.

Table 2 **Simulation Results for 21 Ambulances, Compliance Tables Up to Level 17 and Penalty Function $\Phi(t) = \mathbb{I}_{\{t>720\}}$, Based on 44,520 Requests in 2011**

Method	Performance indicators	$\alpha_k = 0$	$\alpha_k = 1$	$\alpha_k = \lceil k/2 \rceil$	$\alpha_k = k$
MECRP	Percentage on time (%)	94.39	94.27	94.11	94.09
	Lower bound 95%-CI (%)	94.17	94.06	93.89	93.87
	Upper bound 95%-CI (%)	94.60	94.49	94.33	94.31
	Average response time (s)	415	417	418	418
	Mean no. of relocations	2.64	3.79	4.16	4.17
	Average relocation time (s)	509	444	420	420
	Computation time (s)	<1	<1	<1	<1
MEXPREP	Percentage on time (%)	95.09	95.09	95.09	95.17
	Lower bound 95%-CI (%)	94.89	94.89	94.89	94.97
	Upper bound 95%-CI (%)	95.30	95.30	95.30	95.37
	Average response time (s)	416	416	416	412
	Mean no. of relocations	1.53	1.53	1.53	2.88
	Average relocation time (s)	675	675	675	515
	Computation time (s)	67	67	67	72

Observing Tables 1 and 2, we note that the benefit of allowing non-nested compliance tables is very marginal with respect to the percentage of requests, for which the response time threshold is achieved, and to the average response time. In some cases it is even disadvantageous to allow more than zero waiting site changes. Besides that, in the second setting the MEXPREP computes the same compliance tables for $\alpha_k = 0$, $\alpha_k = 1$, and $\alpha_k = \lceil k/2 \rceil$. However, the effect on the number of relocations is large if one uses the compliance tables with no restrictions on waiting site changes rather than compliance tables with restrictions. The question arises whether this marginal performance improvement outweighs this increase in number of relocations. In line with Gendreau et al. (2006), the average relocation time decreases if more waiting site changes are allowed, as expected.

4.3. Relocation Thresholds

The number of relocations in Table 2 is quite large. For instance, for the MEXPREP with $\alpha_k = 0$, the average number of relocations per day is 32. This is because of the large number of changes in availability of ambulances. After all, each time an ambulance is dispatched or finishes service, relocations may be performed. However, one could argue the effect of ambulance relocations if enough ambulances are still available. As an example, it probably makes no sense to relocate ambulances if $n-1$ instead of n ambulances are available, since frequent movements may inconvenience ambulance crews. A way to address this is the introduction of a *relocation threshold*, denoted by K. If the number of available ambulances is below this threshold, we use the compliance table policy. However, if this is not the case, we carry out the *static policy*: we perform no relocations if an ambulance is dispatched, and we send a newly finished ambulance back to its home waiting site. If a transition from level K to $K+1$ occurs, each ambulance is sent back to its home waiting site. Note that these ambulance movements do *not* contribute to the number of relocations, as it is beneficial from the crew's perspective to be present at the home waiting site.

The determination of the ideal level of this relocation threshold is an interesting topic. If it is too high, it is possible that too many relocations are performed. On the other hand, a low threshold may result in a worse performance of an ambulance service provider. To investigate the behavior of different relocation thresholds K, we compute compliance tables by MEXPREP for $K = 7$, $K = 14$, and $K = 21$, for the five different penalty functions of (28)–(32), where $\alpha_k = 0$. That is, we change n in the MEXPREP formulation to K and compute K compliance table levels. Except for the fact we do not change the q_k values in the objective function, we compute MEXPREP as if there were K ambulances instead of n.

In addition, we compute an initial configuration of the $n = 21$ ambulances by an ordinary location problem, which is a modification of MEXPREP. In MEXPREP, we set $k = 21$ in all constraints and in the objective function. Moreover, we discard constraints (21), (22), and (25), as well as q_k in the objective function. Note that for penalty function Φ_1, this modification of MEXPREP is equivalent to MEXCLP.

Then, we simulate our system for $K = 0$ (the static policy), $K = 7$, $K = 14$, and $K = 21$, starting in the initial configuration. This initial configuration also determines the home waiting site of each ambulance. In the simulation, we solve the MWBMP to obtain a solution to the assignment problem. Results are listed in Table 3.

As expected, the patient-based performance indicators (which are fractions on time, average penalty, and average response time) increase as K increases. Specifically, the compliance tables obtained by MEXPREP outperform the static policies, which in addition to the MECRP compliance tables could also serve as a benchmark policy on all penalty functions. However, this comes at the expense of additional ambulance relocations.

Interestingly, fewer ambulance relocations are performed when a relocation threshold $K = 21$ is used instead of $K = 14$. This behavior is easily explained by the following observation: the majority of the ambulance relocations are done when a transition from level $K+1$ to K occurs. If $K = n = 21$, there are no transitions from level $K+1$ to level K. Due to the nesting of the compliance table, relatively few ambulance relocations are performed. However, for $K = 14$, there are many transitions from level 15 to 14. Together with the fact that level 14 is not in general

Table 3 **Simulation Results for Several Levels of K, $n = 21$ Ambulances, and $\alpha_k = 0$, Based on 44,520 Requests in 2011**

Function	Performance indicators	$K = 0$	$K = 7$	$K = 14$	$K = 21$
Φ_1	Percentage on time (%)	90.41	91.36	95.02	95.19
	Average penalty	0.10	0.09	0.05	0.04
	Average response time (s)	462	456	422	415
	Mean no. of relocations	0	1.09	2.07	1.40
	Average relocation time (s)	—	707	676	678
	Computation time (s)	—	5	38	470
Φ_2	Percentage on time (%)	93.54	94.09	95.47	95.66
	Average penalty	433	429	405	403
	Average response time (s)	433	429	405	403
	Mean no. of relocations	0	1.13	2.30	1.60
	Average relocation time (s)	—	604	595	647
	Computation time (s)	—	6	35	172
Φ_3	Percentage on time (%)	93.26	93.92	95.08	95.13
	Average penalty	0.9124	0.9114	0.9052	0.9043
	Average response time (s)	431	426	405	402
	Mean no. of relocations	0	1.02	2.41	1.75
	Average relocation time (s)	—	632	574	603
	Computation time (s)	—	7	68	455
Φ_4	Percentage on time (%)	93.26	93.89	95.08	95.11
	Average penalty	0.8464	0.8447	0.8351	0.8341
	Average response time (s)	431	426	405	403
	Mean no. of relocations	0	1.02	2.41	1.76
	Average relocation time (s)	—	633	574	608
	Computation time (s)	—	5	50	548
Φ_5	Percentage on time (%)	93.26	93.89	95.05	95.09
	Average penalty	0.8741	0.8726	0.8632	0.8614
	Average response time (s)	431	426	405	402
	Mean no. of relocations	0	1.02	2.39	1.78
	Average relocation time (s)	—	633	575	600
	Computation time (s)	—	4	107	713

nested in the ambulance configuration with 15 ambulances, many ambulance relocations are carried out. This behavior is also reflected in Figure 3, where the total number of relocations and mean penalty as function of K is displayed. It is not a surprise that the peak of the number of relocations is at $K = 12$. After all, the mean number of available ambulances is between 12 and 13, so many transitions from a situation with 13 to a situation with 12 available ambulances take place.

Note that for the static policy $K = 0$, the performance indicators differ for the considered penalty functions in general, although no compliance table policy is carried out. This is a direct consequence of the differences in the initial configurations. Moreover, it is worth noting that the coverage penalty function Φ_1 is outperformed by the average response time penalty function Φ_2 on the percentage on time criterion, despite the fact that Φ_1 focuses on maximizing this percentage. This underlines the conclusion of Erkut et al. (2008) about the weakness of models based on coverage.

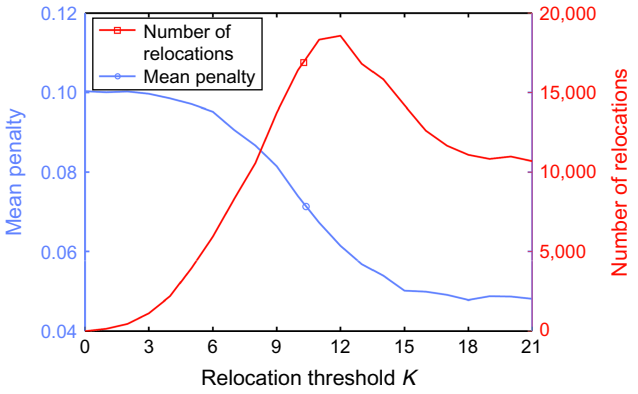

Figure 3 **(Color online) Total Number of Relocations and Mean Penalty as a Function of the Relocation Threshold K, for 21 Ambulances, $\alpha_k = 0$, and Penalty Function $\Phi(t) = \mathbb{1}_{\{t>720\}}$**

4.4. Assignments

We proceed this numerical study with a comparison of the two models for solving the assignment problem, mentioned in Section 3, namely the MWBMP and the LBAP. Results are displayed in Table 4.

The results in Table 4 show that using the LBAP for the assignment problem results in a slightly better performance regarding the patient-based performance indicators. This small increase is explained by the observation that LBAP minimizes the maximum travel time of a relocated ambulance. As a consequence, the ambulance configuration corresponding to the new compliance table level is attained faster. Hence, as expected, the average relocation time per ambulance decreases drastically. After all, using the LBAP, a long trip of one ambulance is split into multiple shorter trips, thus reducing the average relocation time per ambulance. However, the total number of relocations is approximately quadrupled with respect to the usage of the MWBMP as assignment problem. This is probably not acceptable from the crew perspective. It is up to the ambulance service provider to decide whether this tremendous increase of number of relocations outweighs the benefits of the increase in patient-based performance.

4.5. Expected Number of Survivors

Another interesting indicator that provides insight into the performance of the compliance tables is the expected number of survivors. This expected number is easily computed by the summation of the 44,520 penalties for the survival functions Φ_3, Φ_4, and Φ_5. Moreover, we perform cross-comparisons of these functions: we evaluate the compliance table corresponding to the solution of MEXPREP for one specific penalty function (rows) using the other ones (columns), for both MWBMP and LBAP. Results are listed in Table 5.

Table 4 **Simulation Results for $n = K = 21$ and $\alpha_k = 0$, Based on 44,520 Requests in 2011**

Method	Performance indicators	Φ_1	Φ_2	Φ_3	Φ_4	Φ_5
MWBMP	Percentage on time (%)	95.19	95.66	95.13	95.11	95.09
	Average penalty	0.04	403	0.9043	0.8341	0.8614
	Average response time (s)	415	403	402	403	402
	Average no. of relocations	1.40	1.60	1.75	1.76	1.78
	Average relocation time (s)	678	647	603	608	600
	Computation time (s)	470	172	455	548	713
LBAP	Percentage on time (%)	95.62	95.71	95.23	95.26	95.27
	Average penalty	0.04	394	0.9017	0.8300	0.8579
	Average response time (s)	408	394	395	395	396
	Average no. of relocations	6.11	6.33	6.47	6.52	6.51
	Average relocation time (s)	394	387	365	363	361
	Computation time (s)	467	172	440	552	710

Table 5 **Expected Number of Survivors for $n = 21$ and $\alpha_k = 0$, Based on 44,520 Requests in 2011**

	Φ_3		Φ_4		Φ_5	
Evaluation:	MWBMP	LBAP	MWBMP	LBAP	MWBMP	LBAP
Φ_1	4,033	4,163	7,056	7,248	5,803	6,003
Φ_2	4,228	4,350	7,355	7,537	6,106	6,294
Φ_3	4,261	4,378	7,404	7,577	6,159	6,339
Φ_4	4,250	4,372	7,387	7,567	6,142	6,329
Φ_5	4,268	4,371	7,413	7,565	6,170	6,328

Table 6 **MEXPREP Objective Values and Simulated Penalties for $n = 21$, Based on 44,520 Requests in 2011**

	Performance indicators	Φ_1	Φ_2	Φ_3	Φ_4	Φ_5
$\alpha_k = 0$	MEXPREP objective value	0.0572	443	0.9172	0.8533	0.8817
	MWBMP simulated penalty	0.0438	403	0.9043	0.8341	0.8614
	LBAP simulated penalty	0.0438	395	0.9012	0.8293	0.8573
$\alpha_k = k$	MEXPREP objective value	0.0571	443	0.9172	0.8533	0.8817
	MWBMP simulated penalty	0.0439	403	0.9038	0.8328	0.8608
	LBAP simulated penalty	0.0426	396	0.9013	0.8292	0.8566

If one considers the rows corresponding to Φ_3, Φ_4, and Φ_5 in Table 5, one may observe that the differences within these columns are very small: the numbers differ at most by 0.5%. We conclude that the chosen survival function is not of influence on the maximization of survivors. In contrast, the number of survivors differs for the compliance tables induced by the penalty functions based on coverage and average response times; Φ_1 and Φ_2, respectively. Especially for Φ_1, this difference is around 5% compared to the survival functions. However, the difference between the survival functions and Φ_2 is relatively minor. As a consequence, it seems that the average response time is a better approximation for survival than coverage.

As can be observed in Table 5, there are differences between MWBMP and LBAP. For instance, the expected number of survivors using LBAP increases with approximately 2.6% with respect to the case in which the MWBMP is used as assignment problem for Φ_3. This was largely as expected due to the increase in performance of LBAP with respect to MWBMP, as can be observed in Table 4. The expected number of survivors is smallest when the compliance tables are evaluated using penalty function Φ_3. This is explained by the fact that Φ_3 is the most pessimistic survival function (see Figure 2).

4.6. AMEXPREP

In Section 2.5, we discussed some limitations and assumptions on busy fractions. These assumptions may result in an objective value of MEXPREP that differs from the values computed through simulation. In Table 6, objective and simulated values are listed for the two extremes $\alpha_k = 0$ and $\alpha_k = k$ for both MWBMP and LBAP.

From Table 6, we conclude that MEXPREP's estimation of the system performance is somewhat too pessimistic. This is most evident in Φ_1, in which the relative gap between objective value and simulated values is largest. Moreover, we observe a difference only in the fourth digit in the objective values for $\alpha_k = 0$ and $\alpha_k = k$ for Φ_1. From this observation, one could draw the conclusion that nested compliance tables are already very close to optimal. This is also underlined by the simulated values. In all cases, the simulated values using MWBMP are closer to the objective values than in the simulation that uses LBAP as the assignment problem. This is as expected, since the use of LBAP results in better patient-based performance (see Table 4).

As opposed to the objective values of MEXPREP, the AMEXPREP presented in Section 2.5 provides an optimistic estimation of the system performance, as can be observed in Table 7. For the penalty functions based on survival, Φ_3, Φ_4, and Φ_5, the objective value of AMEXPREP differs more from the simulated values than is the case for MEXPREP. Surprisingly, for Φ_1 and Φ_2, it is the opposite. At last, it is worth noting that AMEXPREP performs slightly better than MEXPREP on the penalty criterion in general.

Table 7 **AMEXPREP Objective Values and Simulated Penalties for $n = 21$ and $\alpha_k = 0$, Based on 44,520 Requests in 2011**

	Performance indicators	Φ_1	Φ_2	Φ_3	Φ_4	Φ_5
MEXPREP	Objective value	0.0572	443	0.9172	0.8533	0.8817
	MWBMP simulated penalty	0.0438	403	0.9043	0.8341	0.8614
	LBAP simulated penalty	0.0438	394	0.9017	0.8300	0.8579
AMEXPREP	Objective value	0.0371	380	0.8127	0.7539	0.7794
	MWBMP simulated penalty	0.0435	400	0.9032	0.8323	0.8600
	LBAP simulated penalty	0.0423	395	0.9014	0.8293	0.8574

4.7. Base Station Capacities

In this section, we solve MEXPREP taking into account the actual waiting site capacities depicted in Figure 1. These restrictions can easily be incorporated in MEXPREP by introducing constraints of the type

$$x_{jk} \le c_j \quad j \in W,\ k = 1, \ldots, n-1, \tag{35}$$

where c_j denotes the capacity of waiting site $j \in W$. We compute the restricted version of MEXPREP for $\alpha_k = 0$. We compare the obtained compliance table to the actual capacities. The number of deviations is reported in the columns $c1$ in Table 8. For instance, for Φ_1, the number of capacity violations is 13 for the whole compliance table, and these violations occur in levels nine up to 21. In addition, columns $c2$ report the numbers for the restricted compliance table compared to the unrestricted one. Note that the compliance tables consist of 231 numbers in total.

Only for Φ_1 the computation of restricted MEXPREP results in a different objective value compared to the unrestricted MEXPREP: 0.0576. For the other penalty functions, the objective values do not differ in the first four digits, although different compliance tables were generated, as can be observed in Table 8. From this observation, one could draw the conclusion that minor differences in compliance tables are hardly noticed in the objective value: there are many compliance tables that are near-optimal. It is also interesting to see that there are deviations in lower levels for penalty functions Φ_3, Φ_4, and Φ_5 with respect to the restricted compliance table, while these are not present in the middle levels.

In addition, we simulate the restricted compliance tables. The differences in average penalties between restricted and unrestricted compliance tables are very small for all penalty functions and not worth reporting. According to this analysis, one might conclude that the current capacity is not a limiting factor.

4.8. Computation Times

We conclude this section with an investigation on computation times of MEXPREP. Unfortunately, we are not able to investigate the increase in computation time by choosing a different demand aggregation for the considered case, since we only have access to travel times between four-digit postal codes. As an alternative, we create an artificial problem instance: we pick $|V|$ demand nodes out of a grid of size 100×100, for different values of $|V|$, and assign demand probabilities to them. Travel times between nodes are calculated by the Manhattan metric. For the base locations, we select $|W| = 15$ points, and we consider $n = 20$ ambulances. Then, we solve MEXPREP for the extremes $\alpha_k = 0$ and $\alpha_k = k$, and for Φ_2 and Φ_5, since in Table 4 the computation time of these penalty functions is shortest and longest, respectively. Results on computation times, as well as number of variables and constraints (Equations (18)–(22)) are listed in Table 9.

For large values of $|V|$, it takes more time to obtain a solution for $\alpha_k = 0$ compared to $\alpha_k = k$, as can be observed in Table 9. The explanation of this phenomenon is probably in the method CPLEX uses to compute a solution. From Tables 3 and 4 one may conclude that the use of Φ_2 and Φ_5 induce the shortest and longest computation times, respectively. However, Table 9 shows that Φ_2 did not consistently result in shorter computation times than Φ_5.

5. Concluding Remarks

In this paper, we presented the minimum expected penalty relocation problem (MEXPREP) to compute compliance tables. The MEXPREP is an extension of the maximal expected coverage relocation problem (MECRP) formulated in Gendreau et al. (2006) in two directions. First, we incorporated the objective function of the maximum expected covering location problem (MEXCLP) into the objective function of the MECRP to anticipate multiple future emergency requests beyond a first request. Then, we introduced penalty functions in order to focus on performance measures other than coverage, including survival probabilities. Moreover, based on the assumptions and limitations of busy fractions, we introduced an adjusted version of MEXPREP. In this adjusted version, called AMEXPREP, correction factors proposed in Batta et al. (1989) were incorporated. Additionally, we considered both the minimum weighted bipartite matching problem (MWBMP) and the linear bottleneck assignment problem (LBAP) as assignment problems for the assignment of available ambulances to the waiting sites indicated by the compliance table level.

We concluded this paper with a numerical study, based on 44,520 emergency requests in 2011 in the

Table 8 Deviations of Unrestricted MEXPREP Compliance Tables with Respect to Actual Capacities and Restricted MEXPREP for $\alpha_k = 0$

	Φ_1		Φ_2		Φ_3		Φ_4		Φ_5	
	$c1$	$c2$	$c1$	$c2$	$c1$	$c2$	$c1$	$c2$	$c1$	$c2$
Deviations	13	16	4	4	5	9	5	10	11	14
Levels	9–21	9–21	18–21	18–21	17–21	2, 3, 18–21	17–21	2–4, 18–21	17–21	2–4, 18–21

Table 9 Computation Times for the Artificial Problem Instance

	$\lvert V\rvert = 100$	$\lvert V\rvert = 200$	$\lvert V\rvert = 300$	$\lvert V\rvert = 400$	$\lvert V\rvert = 500$	$\lvert V\rvert = 600$
No. of variables	3.2×10^5	6.3×10^5	9.5×10^5	1.3×10^6	1.6×10^6	1.9×10^6
No. of constraints	3.1×10^5	6.2×10^5	9.2×10^5	1.2×10^6	1.5×10^6	1.8×10^6
Computation time Φ_2, $\alpha_k = 0$ (s)	53	168	348	695	1,119	1,770
Computation time Φ_5, $\alpha_k = 0$ (s)	38	196	387	808	1,182	1,689
Computation time Φ_2, $\alpha_k = k$ (s)	63	197	459	595	1,049	1,594
Computation time Φ_5, $\alpha_k = k$ (s)	47	189	349	576	997	1,650

region of Amsterdam and its surroundings. In this study, we compared the MEXPREP compliance tables to both the MECRP compliance tables and the static policy, and we observed that the MEXPREP outperforms both of them on most performance indicators. We also carried out a comparison between several restrictions on waiting site changes. Moreover, we considered several relocation thresholds, and compared the resulting performance when using LBAP and MWBMP as assignment problems. In addition, we compared the objective values with the simulated values for both MEXPREP and AMEXPREP. Studies regarding computation times of MEXPREP and the effect of base station capacities were conducted as well.

There are several extensions that can be made to improve the realism of the MEXPREP model. For instance, we assumed travel times to be deterministic, while in reality these are stochastic. Moreover, we used one universal busy fraction p, which included some limitations. For instance, in reality, this busy fraction probably differs per base location. Another interesting research topic is a modification of MEXPREP in which only certain designated levels of the compliance table are computed, rather than the whole compliance table, and how this kind of policy affects the performance. With regard to survival probabilities, we only considered survival functions based on a cardiac arrest, while other types of emergency requests occur in practice as well. However, survival functions for several types of emergency requests could be combined in one survival function using weights corresponding to the frequency of different request types (if this could be quantified, as pointed out in Erkut et al. 2008). The MEXPREP model to compute compliance tables presented in this paper forms a good basis for these extensions and modifications.

Supplemental Material

Supplemental material to this paper is available at http://dx.doi.org/10.1287/ijoc.2015.0687.

Acknowledgments

The author thanks the ambulance service provider of the EMS region of Amsterdam, Ambulance Amsterdam, for providing data. In addition, the author thanks the RIVM for providing the travel times for ambulances in the EMS region considered in the case study. The author also thanks the review team for their useful comments. The author thanks Sandjai Bhulai and Rob van der Mei for their valuable support regarding this work. This research was financed in part by Technology Foundation STW [Contract 11986], which the author gratefully acknowledges.

References

Alanis R, Ingolfsson A, Kolfal B (2013) A Markov chain model for an EMS system with repositioning. *Production Oper. Management* 22(1):216–231.

Andersson T, Värbrand P (2007) Decision support tools for ambulance dispatch and relocation. *J. Oper. Res. Soc.* 58(2):195–201.

Batta R, Dolan JM, Krishnamurthy NN (1989) The maximal expected covering location problem: Revisited. *Transportation Sci.* 23(4):277–287.

Brotcorne L, Laporte G, Semet F (2003) Ambulance location and relocation models. *Eur. J. Oper. Res.* 147 (3):451–463.

Burkhard RE, Dell'Amico M, Martello S, eds. (2009) Other types of linear assignment problems. *Assignment Problems*, Chapter 6 (SIAM, Philadelphia), 171–202.

Church R, ReVelle C (1974) The maximal covering location problem. *Papers Regional Sci. Assoc.* 32(1):101–118.

Daskin MS (1983) A maximum expected covering location model: Formulation, properties and heuristic solution. *Transportation Sci.* 17(1):48–70.

Erkut E, Ingolfsson A, Erdoğan G (2008) Ambulance location for maximum survival. *Naval Res. Logist.* 55(1):42–58.

Gendreau M, Laporte G, Semet F (2001) A dynamic model and parallel tabu search heuristic for real-time ambulance relocation. *Parallel Comput.* 27(2):1641–1653.

Gendreau M, Laporte G, Semet F (2006) The maximal expected coverage relocation problem for emergency vehicles. *J. Oper. Res. Soc.* 57(1):22–28.

Ingolfsson A, Budge S, Erkut E (2008) Optimal ambulance location with random delays and travel times. *Health Care Management Sci.* 11(3):262–274.

Jagtenberg CJ, Bhulai S, van der Mei RD (2015) An efficient heuristic for real-time ambulance redeployment. *Oper. Res. Health Care* 4:27–35.

Kommer G, Zwakhals S (2008) Referentiekader spreiding en beschikbaarheid ambulancezorg. RIVM Briefrapport 270192001/2008. Accessed March 27, 2016, https://www.ambulancezorg.nl/download/downloads/1538/referentiekaderspreiding-en-beschikbaarheid-2008.pdf.

Larsen M, Eisenberg M, Cummins R, Hallstrom A (1993) Predicting survival from out-of-hospital cardiac arrest—A graphic model. *Ann. Emergency Medicine* 22(11):1652–1658.

Larson RC (1975) Approximating the performance of urban emergency service systems. *Oper. Res.* 23(5):845–868.

Li X, Zhao Z, Zhu X, Wyatt T (2011) Covering models and optimization techniques for emergency response facility location and planning: A review. *Math. Methods Oper. Res.* 74(3):281–310.

Maio VD, Stiell I, Wells G, Spaite D (2003) Optimal defibrillation for maximum out-of-hospital cardiac arrest survival rates. *Ann. Emergency Medicine* 42(2):242–250.

Maleki M, Majlesinasab N, Sepehri MM (2014) Two new models for redeployment of ambulances. *Comput. Indust. Engrg.* 78:271–284.

Maxwell M, Henderson S, Topaloglu H (2013) Tuning approximate dynamic programming policies for ambulance redeployment via direct search. *Stochastic Systems* 3(2):322–361.

Maxwell MS, Restrepo M, Henderson SG, Topaloglu H (2010) Approximate dynamic programming for ambulance redeployment. *INFORMS J. Comput.* 22(2):266–281.

McLay L, Mayorga M (2010) Evaluating emergency medical service performance measures. *Health Care Management Sci.* 13(2):124–136.

Rajagopalan H, Saydam C, Xiao J (2008) A multiperiod set covering location model for dynamic redeployment of ambulances. *Comput. Oper. Res.* 35(3):814–826.

Repede J, Bernardo J (1994) Developing and validating a decision support system for locating emergency medical vehicles in Louisville, Kentucky. *Eur. J. Oper. Res.* 75(3):567–581.

ReVelle C, Swain R (1970) Central facilities location. *Geographical Anal.* 2(1):30–42.

Schmid V (2012) Solving the dynamic ambulance relocation and dispatching problem using approximate dynamic programming. *Eur. J. Oper. Res.* 219(3):611–621.

Schmid V, Doerner K (2010) Ambulance location and relocation problems with time-dependent travel times. *Eur. J. Oper. Res.* 207(3):1293–1303.

Sudtachat K, Mayorga ME, McLay LA (2016) A nested-compliance table policy for emergency medical service systems under relocation. *OMEGA* 58:154–168.

Valenzuela T, Roe D, Cretin S, Spaite D, Larsen M (1997) Estimating effectiveness of cardiac arrest intervention—A logistic regression survival model. *Circulation* 96(10):3308–3313.

van Barneveld T, Bhulai S, van der Mei R (2015) A dynamic ambulance management model for rural areas. *Health Care Management Sci.*, ePub ahead of print October 3, http://dx.doi.org/10.1007/s10729-015-9341-3.

van den Berg P, Aardal K (2015) Time-dependent MEXCLP with start-up and relocation cost. *Eur. J. Oper. Res.* 242(2):383–389.

Waaelwijn R, de Vos R, Tijssen J, Koster R (2001) Survival models for out-of-hospital cardiopulmonary resuscitation from the perspectives of the bystander, the first responder, and the paramedic. *Resuscitation* 51(2):113–122.