Real-Time Ambulance Relocation

Assessing real-time redeployment strategies for ambulance relocation

T.C. van Barneveld[*1,2], C.J. Jagtenberg[†1], S. Bhulai[‡2,1], and R.D. van der Mei[§1,2]

[1]Centrum Wiskunde & Informatica, Amsterdam, The Netherlands

[2]Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

February 11, 2016

Abstract

Providers of Emergency Medical Services (EMS) are typically concerned with keeping response times short. A powerful means to ensure this, is to dynamically redistribute the ambulances over the region, depending on the current state of the system. In this paper, we provide new insight in how to optimally (re)distribute ambulances. We study the impact of (1) the frequency of redeployment decision moments, (2) the inclusion of busy ambulances in the state description of the system, and (3) the performance criterion on the quality of the distribution strategy. In addition, we consider the influence of the EMS crew workload, such as (4) chain relocations and (5) time bounds, on the execution of an ambulance relocation. To this end, we use trace-driven simulations based on a real-life dataset of ambulance providers in the Netherlands. In doing so, we differentiate between rural and urban regions, which typically face different challenges when it comes to EMS. Our results show that: (1) taking the classical 0-1 performance criterion for assessing the fraction late arrivals only differs slightly from taking expert-opinion based S-curve for evaluating

[*]t.c.van.barneveld@cwi.nl

[†]c.j.jagtenberg@cwi.nl

[‡]s.bhulai@vu.nl

[§]r.d.van.der.mei@cwi.nl

the performance as a function of the response time, (2) adding more relocation decision moments is highly beneficial, particularly for rural areas, (3) considering ambulances involved in dropping off patients available for newly coming incidents only slightly reduces relocation times, and (4) simulation experiments for assessing move-up policies are highly favorable to simple mathematical models because of the inherent complexity and stochasticity.

Keywords— Ambulance redeployment; Response times; Workload; Simulation

1 Introduction

In emergency situations, ambulance providers need to respond to requests for ambulances to provide medical aid and transportation to a hospital quickly. It is of utmost importance that ambulances are on-site at emergency locations within a short period of time. Therefore, it is crucial to position ambulances throughout the region such that they occupy good locations with respect to expected demand. Moreover, it is important that a good distribution of vehicles is maintained when ambulances become busy. Hence, modern ambulance providers tend to *relocate* idle ambulances in order to achieve short *response times*: the time between the emergency call and the arrival of the ambulance at the emergency scene.

A commonly used quality measure for the performance of the ambulance service provider is the fraction of highest-urgency requests responded to within a certain time standard, usually between 8 and 15 minutes. Related to this time threshold is the concept of *coverage*. An area is said to be covered if it is reachable by an ambulance within the time threshold. One may interpret this coverage as the 'preparedness' of the system to respond to future calls, and therefore one may solve the ambulance relocation problem by relocating ambulances in such a way that an acceptable coverage level of the region is ensured.

1.1 Related Work

The literature on the ambulance relocation problem can roughly be divided in two categories: periodic redeployment and real-time ambulance relocation. The authors of [4] provide a comprehensive survey on both types of repositioning. In the first category, redeployment of ambulances is considered *preplanned* to anticipate time-dependent fluctuations in demand, travel times, and number of ambulances on duty. These models, extensively surveyed in [5]

and more recently in [16], effectively divide the planning horizon into discrete time periods, and then solve the static ambulance location problem[1] multiple times. An early model in the literature on preplanned redeployment is proposed in [21]. In that paper, the authors extend the maximum expected coverage location problem (MEXCLP), proposed by [8], to a location model with time-dependent variation in travel times and fleet size, hence its name TIMEXCLP. This model was applied to the EMS system of Louisville, Kentucky, and a decrease of 36% in response time was achieved. In [23], the focus is on preplanned repositioning as well, taking into account time-dependent travel times by extending the single-period double standard model proposed in [11] into a multi-period version. Minimization of the number of ambulance relocations over the planning horizon while maintaining a satisfactory coverage level, is the topic of [19], and a two-stage optimization model is proposed. Other papers in which periodic redeployment is considered include [9], [20], and [28].

In this paper, we focus on real-time ambulance relocation. In contrast to preplanned repositioning, real-time ambulance relocation bases its decisions on the actual state of the system as it is observed throughout the day. The real-time situation changes often, e.g., due to the arrival of a request whereupon an ambulance is dispatched, or a service completion of a patient. These events can trigger one or multiple ambulance relocations. Methods solving the real-time ambulance relocation problem, also known as 'move-up', can be divided in two categories: offline and online methods. A comprehensive study on both types of methods is conducted in [29].

In the offline approach, solutions to the ambulance relocation problem are precomputed for a variety of scenarios that may arise. Whenever such a scenario occurs in real time, the corresponding relocation is looked up and applied. The level of detail of these scenarios may differ. For instance, so-called *compliance table policies* base their decisions on the *number of idle ambulances* solely, and are therefore a category of policies with low detail about the state of the system. Compliance tables are simple to understand and to use by dispatchers, making this kind of policy a commonly used one. In [13], the maximum expected covering relocation problem (MECRP) for the computation of compliance tables is proposed. In [17], it is stated that computing compliance tables is just the first part of computing relocation decisions. The second part involves the actual assignment of ambulances to waiting sites, and two offline methods minimizing the total relocation time are proposed, based on compliance tables computed by MECRP. Such a decoupling is also present in [25], in which the MECRP model is extended

[1] in which each vehicle always returns to its own home base.

by addressing ambulance unavailability and general performance measures are considered. However, in contrast to the work done in [17], an online model for the actual assignment of ambulances to waiting sites is considered. In [1] a two-dimensional Markov chain is proposed to analyze the system performance of compliance table policies. This Markov chain is used in [24] as well. In this work, the steady-state probabilities serve as input to an integer program for the computation of nested compliance tables.

More sophisticated offline policies include additional information about ambulances and requests in the scenarios, (e.g., [18] and [22]). However, scalability issues arise when the number of scenarios is too large, yielding an intractable solution space. To overcome this problem, both papers present an approximate dynamic programming approach for the computation of ambulance relocation strategies.

In offline methods, the computation time is not an issue as the solution is computed beforehand. In contrast, in the online approach being able to calculate a relocation decision in real time is of utmost importance. Since obtaining a relocation suggestion quickly is desirable, the main focus in literature on the online approach is on fast heuristics. One of the first ambulance relocation methods is proposed in [12]. This model is based on the double standard model of [11] and it is solved via tabu-search. A dynamic relocation model called DYNAROC is presented in [2]. This article proposes a policy that includes both the ambulance dispatch as well as relocating idle ambulances, and uses a fast tree-search heuristic to solve DYNAROC. A one-step look-ahead heuristic is the considered in [26]. Several scenarios are constructed that may occur one time-step later and these scenarios are combined with each possible relocation decision to obtain a classification of these possible decisions. Finally, the online relocation models proposed in [14] and [27] are of the most importance to this work. These two methods are summarized in Section 3.

1.2 Contribution

This paper aims to thoroughly analyze the dynamic ambulance relocation process, also known as 'move-up', from a practical point of view. In some sense, it could be considered as a search for the 'best of both worlds' combination of [14] and [27]. The two methods proposed in these papers are easy to understand and to implement, and are therefore very suitable candidates to conduct further research on. Furthermore, unlike many other

move-up policies, these two methods have recently been tested[2] in practice. This combination of properties makes these paper a natural choice for our investigation.

Both methods have their strengths and shortcomings. A strength of the approach described in [14] is the ability to anticipate multiple emergency requests rather than just the first one, as done in [27]. However, in [27] a general performance measure, modelled by an expert-opinion based function of the response time, is considered, while the authors of [14] only use coverage as their performance measure. In Section 3.2 we discuss the differences between both approaches.

In this paper, we combine the methods developed in [14] and [27] to obtain practical insights on how an ambulance provider should implement a move-up strategy. We explore features of both algorithms, and their effects on various measures of the response time distribution. While our primary focus is on minimizing the fraction of late arrivals, other values – such as the average response time – are also reported.

Note that decision makers in practice may come to different conclusions based on the characteristics of their EMS region. For example, the size of the demand – as well as how it is spatially distributed, distances and overall workload have a great effect on the dynamics in the EMS system. These characteristics may affect the performance of a move-up policy, and a policy that performs well in one region, does not necessarily give the same result elsewhere. Since we aim to construct a robust algorithm with respect to region characteristics, we include case studies for two different types of regions: the rural region of Flevoland, and the urban region of Amsterdam, both in the Netherlands.

Although ambulance move-up methods can offer great performance improvements, the well-known downside is that the workload for the crew increases, combined with additional costs for the travelled distances. Thereto, we analyze the trade off between the number of move-ups, the total travel time needed for relocations, and the reduction in response times. Furthermore, we investigate whether move-up methods can benefit from taking into account vehicles that are currently dropping of a patient at a hospital. It is clear that these vehicles will become idle in the near future, but it is not trivial how one should model this, nor is it evident that this will have a positive effect on the performance. We show that taking ambulances at hospitals into account has hardly any effect on the response times, but it does slightly diminish relocation times – and thereby workload – for the crew.

[2]This resulted in very good performance on patient-related performance indicators such as the fraction of late arrivals and mean response times.

We also investigate the effect of long-distance relocations. The further we send an ambulance to, the longer it takes for the system to reach the desired configuration. Thereto, we analyze two options: 1) we *bound* the relocation time to a certain maximum, i.e., ignoring options that would take too long, and 2) introduce a 'chain movement' of multiple vehicles, thereby breaking up the long drive into several smaller ones, that may be executed simultaneously.

All our results are obtained from trace-driven simulations, that we consider to be an accurate representation of reality.

2 Problem Description

In this section we describe the general EMS process. When idle, ambulance crews spend their shift at designated waiting sites. These could be *base stations*: structures set aside for parking idle ambulances with a crew room and other facilities for the ambulance personnel. However, if the situation requires, the ambulance crew may also be asked to park up at other waiting sites away from the base station, e.g., parking lots, fuel stations or other hot spots. This practice tends to become more and more common in North America, and although our models allow for this situation, we focus our evaluation on the emergency system in The Netherlands where the number of ambulances on duty usually exceeds the number of waiting sites. Hence, multiple ambulances are typically present at a waiting site.

At a certain moment in time, a request for an ambulance arrives at the emergency control center. This call is answered by a dispatcher who assists the caller in first aid, inquires the condition of the patient and determines the urgency based on the answers. Meanwhile, the dispatcher consults the dispatching system which ambulance is most suitable to respond to the patient, taking into account the current location and status of the ambulances. For calls of the highest urgency, usually the closest idle ambulance is assigned to perform this task.

After selecting an appropriate ambulance, the dispatcher informs the ambulance crew about the location, urgency and condition of the patient. Note that the ambulance is usually present at a *base station*. However, it could also be the case that an idle ambulance is on the road, headed towards a base after the transportation of a patient for instance. The ambulance crew is expected to leave for the emergency scene immediately, and does not need to return to base first.

After driving some time, with or without optical and sound signals depending on the urgency, the ambulance arrives at the scene and starts the

medical treatment of the patient. During this treatment, it is decided whether the patient needs transportation to a hospital. If so, the patient is loaded into the ambulance and brought to a hospital. The dispatcher does not have influence on the selection of an appropriate hospital since it depends on the wishes of the patient, the type of the incident and the emergency location.

At the hospital, the ambulance crew unloads the patient and takes her/him to a suitable department, in consultation with the hospital personnel. When the ambulance crew finished the transfer of the patient, it informs the emergency control center that it is free for service again. If there is no other request to be responded to, a new destination for the ambulance needs to be selected.

2.1 Model

In this section, we describe the mathematical model and we introduce the notation used throughout this paper. We model the region of interest as a weighted complete directed graph $G = (V, A, (\tau^{(1)}, \tau^{(2)}))$. The region is discretized into geographical demand zones, e.g., municipalities, neighborhoods, postal codes or streets. We define V as the vertex set of these demand points. The fraction of demand occuring in node $i \in V$ is denoted by d_i, and we assume that incidents take place in a Poisson manner with rate λ. Hence, the arrival rate of incidents for node i equals λd_i. Let W be the set of potential waiting sites, $W \subseteq V$, and the number of ambulances is denoted by n. The road-network of the region is modelled by arcs $(i, j) \in A$, where $i, j \in V$. Two different travel times are associated to each arc: $\tau_{ij}^{(1)}$ denotes the expected travel time between nodes i and j when driving with optical and sound signals turned on, typically used while responding to an emergency or the transportation of a patient to a hospital. If the ambulance is not performing patient-related duties, such as the return to a waiting site, the optical and sound signals are not turned on. This yields a longer travel time, denoted by $\tau_{ij}^{(2)}$. As in many papers, we consider a single type of ambulance and a single type of demand priority, inducing a single threshold or target, denoted by T, for the response time.

3 Algorithms and Features

In this section, we first explain the DMEXCLP method as published in [14] and the penalty heuristic of [27]. Both methods have in common that it is only allowed to relocate vehicles to existing waiting sites. Such a relocation decision may only be taken at discrete *decision moments* in time, which we

will define later. The decision is then computed by brute force in real time. Moreover, both methods incorporate the location of idle ambulances in the same way: for a travelling idle ambulance they pretend that it is already at its destination instead of at its current location. This choice has two advantages: first of all, for a real-life system it is typically easier to keep track of destinations since they change less often than current locations. Second, there is a methodological advantage: for a moving ambulance, its current location is only relevant for a very short time, while our relocation decision should be beneficial to the system for a longer time. In Section 3.3 we will describe the incorporation of several aspects considered in [27] into the DMEXCLP method and into the simulation used for obtaining results.

3.1 Summary of DMEXCLP

In its original form, the DMEXCLP method moves a vehicle when it becomes idle after finishing service of a patient. At such so-called *decision moments* it relocates this ambulance to an appropriate waiting site within the region. The sole objective of DMEXCLP is to maximize the number of incidents that can be reached within the time threshold T. In that sense, DMEXCLP is closely related to the Maximum Expected Covering Location Problem (MEXCLP), formulated as an ILP in [8]. This problem was designed to compute an optimal static distribution of vehicles over waiting sites, by calculating the *coverage* of the region. It is often used as the basis for an extension to more complicated models, like the Adjusted MEXCLP presented in [3].

MEXCLP defines the coverage of a region in terms of a 'busy fraction' q. This busy fraction is predetermined, and assumed to be the same for all vehicles. It can be estimated by dividing the expected load of the system by the total number of available ambulances. Furthermore, ambulances are assumed to operate independently. Consider a demand point $i \in V$ that is within the time threshold T of k ambulances. We can straightforwardly determine this number k using the expected travel times $\tau_{ij}^{(1)}$, $i, j \in V$. The probability that at least one of these k ambulances is available at any point in time, is then given by $1 - q^k$. If we let d_i be the demand at node i, the expected covered demand of this vertex is $E_k = d_i(1 - q^k)$. The MEXCLP positions the ambulances in such a way that the total maximal expected covered demand, summed over all demand points, is reached.

DMEXCLP, or Dynamic MEXCLP, reuses this definition of coverage, but computes it for relocation purposes each time when an ambulance becomes available. At such a decision moment, the current state of the system is observed. DMEXCLP disregards all information about ambulances that are busy, and focuses purely on the set of idle vehicles. As mentioned, we only

consider the destination of idle ambulances. (If an ambulance is standing at a waiting site, we define its destination to be its current location.) Information regarding the destination of each ambulance is captured by variables n_j: the number of idle ambulances that have waiting site j as destination, $j \in W$. In addition, DMEXCLP requires information on $(d_i)_{i \in V}$ and $(\tau_{ji}^{(1)})_{j \in W, i \in V}$.

At a decision moment, the DMEXCLP method proposes to send the ambulance, that just became idle, to the waiting site that results in the largest coverage according to the MEXCLP model. This is equivalent to choosing the waiting site that maximizes the *marginal* coverage over all demand. This marginal coverage can be interpreted as the added value of having a k^{th} ambulance nearby, and is given by $E_k - E_{k-1} = d_i(1-q)q^{k-1}$. The waiting site that results in the largest marginal coverage over the entire region can be computed by

$$\arg\max_{w \in W} \sum_{i \in V} d_i (1-q) q^{k(i,w,n_1,\ldots,n_{|W|})-1} \cdot \mathbb{1}_{\{\tau_{wi} \leq T\}}, \tag{1}$$

where

$$k(i, w, n_1, \ldots, n_{|W|}) = \sum_{j=1}^{|W|} n_j \mathbb{1}_{\{\tau_{ji} \leq T\}} + \mathbb{1}_{\{\tau_{wi} \leq T\}} \tag{2}$$

expresses the number of idle ambulances that have a destination within range of demand point i, assuming that the ambulance of consideration will be relocated to waiting site w. That is, it counts the number of ambulances that in the near future may respond timely to an incident in i.

3.2 Comparison to Penalty Heuristic

In this section, we highlight differences between the penalty heuristic, presented in [27], and the DMEXCLP method as published in [14]. As mentioned above, similarities exist between both methods. Both papers differ on the following five major aspects:

1. **Coverage:** The penalty heuristic as presented in [27] uses a different notion of coverage: an area is either covered or not covered. It therefore ignores multiple vehicle coverage and ambulance unavailability. In the penalty heuristic, the closest ambulance defines the coverage of a demand point solely. This so-called *single coverage* comes down to a MEXCLP model with $q = 0$. That is, MEXCLP may be interpreted as a generalization of single coverage.

2. **Number of decision moments:** As we have seen, [14] proposes a relocation only when an ambulance becomes available. This choice

has to do with the fact that DMEXCLP was originally designed for busy regions, in which vehicles often become idle[3]. In [27], however, a relocation may also be executed immediately after the dispatch of an ambulance to an incident.

3. **Busy ambulances:** As mentioned in Section 3.1, busy ambulances do not contribute to the coverage in [14]. In contrast, in [27] ambulances at hospital also may provide coverage: they consider an ambulance as dispatchable if its transfer time at a hospital exceeds a predefined standard $\bar{\tau}$. That is, after some time, the transfer may be interrupted if necessary. This influences the coverage of the region, as now a busy ambulance covers the direct neighborhood of the hospital.

4. **Chain relocations:** Whereas in [14] a new waiting site is suggested for an ambulance that just finished service, it is not necessarily this particular ambulance that is redeployed there in [27]. Instead, a *chain relocation* is set up in order to attain the desired ambulance configuration in less time. The, otherwise possibly long, trip may be split into two or more trips, in which multiple ambulances are involved. We refer to [27] for a graphical illustration. Note that this extension does not influence the calculation of which waiting site should receive one additional vehicle: it can be regarded as a second step, executed after the computation of the new ambulance configuration.

5. **Objective:** The focus is on minimization of late arrivals solely in [14]: one incurs a penalty of 1 each time the response time to an incident exceeds T. In contrast, this objective is generalized in [26] by the definition of a *penalty function*, hence the name penalty heuristic. This is a non-negative non-decreasing function on $\mathbb{R}_{\geq 0}$ relating a certain penalty to each possible response time. (Note that the objective of DMEXCLP can be easily modelled by the penalty function $\Phi(t) = \mathbb{1}_{\{t>T\}}$.) However, the authors of [27] question the dichotomous nature of this objective, as medical outcomes are completely ignored, (cf. [10]). Instead, they use a different penalty function, in which the primary goal is to maximize coverage as before, but there is more distinction between different response times. This function is given by

$$\Phi(t) = \begin{cases} \frac{1}{\beta(1+e^{-\alpha(t-T)})} & 0 \leq t \leq T, \\ \frac{\beta-1}{\beta} + \frac{1}{\beta(1+e^{-\alpha(t-T)})} & t > T, \end{cases} \tag{3}$$

and displayed in Figure 1 for $\alpha = 0.008$, $\beta = 5$, and $T = 720$.

[3]Although the authors state that the method can be easily adjusted for usage at other

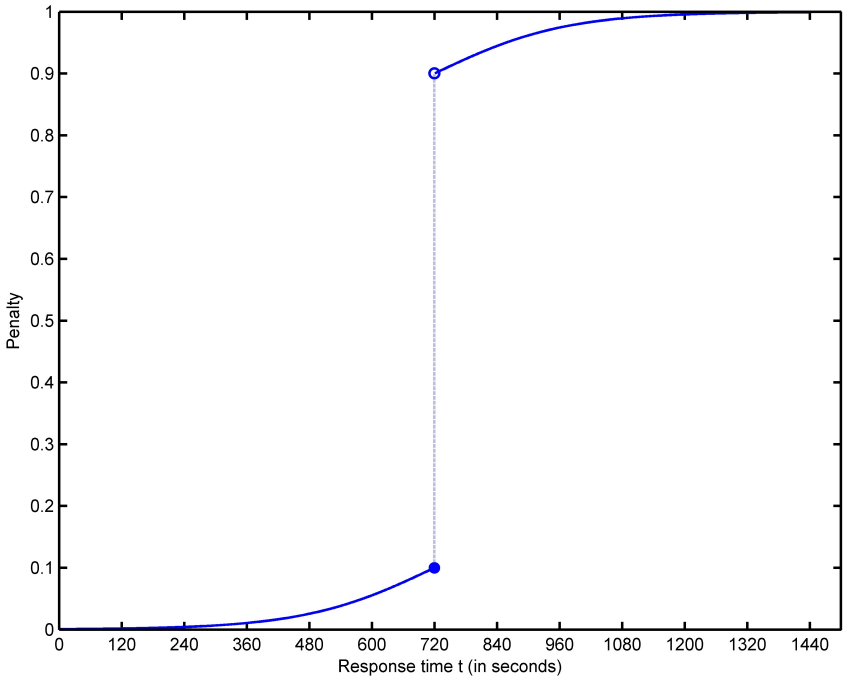


Figure 1: Penalty function used in [27].

We conclude that in one way DMEXCLP is richer than the penalty heuristic, as the multiple and non-integer MEXCLP coverage is a generalization of the penalty heuristic's single coverage. On the other points, the assumptions made in [14] are generalized in [27]. In the next section, we will explain how we modify the original DMEXCLP method by incorporating a number of features related to the five aspects described above.

3.3 Modification of DMEXCLP

In this section we address some features considered in [27]. We explain the incorporation of these into the DMEXCLP method in this section. Moreover, we introduce a new feature, neither considered in [14] nor in [27]: a bound on the relocation time. One by one, we discuss the incorporation of these features.

Decision Moments. At the added decision moment – when a vehicle is dispatched – it is not clear from which waiting site an ambulance should be

types of decision moments, it is not clear which ambulance should be relocated.

relocated to. This is easily computed, however, by the following modification of Equations (1) and (2):

$$\begin{aligned}\underset{(w_1,w_2)\in W^2:n_{w_1}>0}{\arg\max} &\sum_{i\in V} d_i(1-q)q^{k(i,w_2,n_1,\ldots,n_{|W|})-1}\cdot \mathbb{1}_{\{\tau_{w_2 i}\leq T\}} \\ &- \sum_{i\in V} d_i(1-q)q^{k(i,w_1,n_1,\ldots,n_{|W|})-1}\cdot \mathbb{1}_{\{\tau_{w_1 i}\leq T\}},\end{aligned} \tag{4}$$

in which w_1 and w_2 denote the old origin and new destination of the vehicle to relocate, and $k(i, w, n_1, \ldots, n_{|W|})$ as defined in Equation (2). In Equation (4) each possible waiting site pair with at least one ambulance at the origin, is evaluated. Since the number of waiting sites is typically small, the maximization in Equation (4) can be computed by brute force.

Busy Ambulances. Although the authors of [27] allow transfer time interruptions if the transfer at a hospital has lasted for at least $\bar{\tau}$ seconds already, we do not in this paper. After all, the allowance of these preemptions is a specific rule for their region of interest, but not universally adopted. We take into account these busy ambulances in a different way. We assume that the hospital transfer time follows a probability distribution. Let

$$R(a, \tau(a)) := \mathbb{E}\{B(a) \mid B(a) > \tau(a)\} - \tau(a) \tag{5}$$

denote the expected remaining transfer time of ambulance a if its transfer already lasted for $\tau(a)$ time. Moreover, let $h(a) \in V$ denote the demand zone in which the hospital where ambulance a is busy is located. Let $\mathcal{A}$ be the set of ambulances currently dropping off a patient at a hospital. We adjust Equation (2) as follows:

$$k(i, w, n_1, \ldots, n_{|W|}) = \sum_{j=1}^{|W|} n_j \mathbb{1}_{\{\tau_{ji}\leq T\}} + \sum_{a\in\mathcal{A}} \mathbb{1}_{\{R(a,\tau(a))+\tau_{h(a),i}\leq T\}} + \mathbb{1}_{\{\tau_{wi}\leq T\}}. \tag{6}$$

That is, ambulance a contributes to the coverage of demand point i if the sum of its expected remaining transfer time and the travel time of the current location to i does not exceed T.

Chain relocations. As stated before, the use of chain relocations is not a modification of the DMEXCLP method, but the calculation of this chain is a subsequent step: the expression of Equation (1) is not modified. In [27], the *Linear Bottleneck Assignment Problem* is considered for this computation. We refer to [6] for an extensive discussion on this problem. This approach assumes all ambulances as eligble for participation in a chain relocation. The

authors of [27] conclude that the benefit to the patient-based performance of a chain relocation consisting of more than two links is very small. They observe a large performance gain, however, if chains consisting of two links are used. The crew-based performance decreases if chains consist of more than two links, as a consequence of an inflation in number of relocations. As the regions considered in the numerical study of this paper are the same as in [27], we follow their conclusion and restrict that at most two ambulances may take part in a chain relocation. The computation of these chains can be done by brute-force.

Relocation time bounds. At a decision moment, the DMEXCLP method searches for the waiting site for which the expected coverage is maximized, without taking into account the current location of the ambulance. However, from both patient and crew perspective, it might be beneficial to steer the system towards a good, but not the best, configuration that can be attained quickly. After all, driving to a waiting site, although best classified by DMEXCLP, may take long. In order to study the behaviour of the performance if the focus is on good local configurations, we impose an upper bound B on the relocation time of an ambulance. That is, we do not allow the relocation of an ambulance to a waiting site for which the driving time between its current location and destination exceeds B time-units. Let c be the current location of the ambulance under consideration. Then we modify Equation (1) as follows:

$$\arg\max_{w \in W : \tau_{cw} \leq B} \sum_{i \in V} d_i (1-q) q^{k(i,w,n_1,\ldots,n_{|W|})-1} \cdot \mathbb{1}_{\{\tau_{wi} \leq T\}}. \tag{7}$$

That is, we evaluate only the waiting sites that can be reached within B time-units from the current location of the ambulance in the maximization. In Section 4.6 we analyze the behaviour of the system on both patient and crew-based performance for different values of B.

Performance criteria. The incorporation of a different performance criterion, such as the one considered in Equation (3) and Figure 1, requires more effort than the previous features: one can no longer simply count the number of ambulances within range of demand node i. After all, each idle ambulance contributes to the coverage of i, no matter how far away. Due to the notion of MEXCLP coverage, this contribution levels off the farther away an ambulance: with probability $1-q$ the closest one to i is available and responds to an incident occuring there, inducing a penalty of $\Phi(\tau_{ji})$ if the closest ambulance to i is located at waiting site j. With probability $(1-q)q$

the second closest responds, generating $\Phi(\tau_{j'i})$ penalty if this ambulance is at j', and so on.

Let $c(w, n_1, \ldots, n_{|W|})$ denote the configuration in which each idle ambulance is at its destination, assuming that w is selected as destination for the ambulance that just became free. We define $z_{(c(w,n_1,\ldots,n_{|W|}),i,j,l)} := 1$ if and only if the l^{th} closest available ambulance to demand node i is at waiting site j according to configuration $c(w, n_1, \ldots, n_{|W|})$, and 0 otherwise. Let A be the number of available ambulances. Then, we compute w by

$$\arg\min_{w \in W} \sum_{i \in V} \sum_{j \in W} \sum_{l=1}^{A} d_i (1-q) q^{l-1} \Phi(\tau_{ji}) z_{(c(w,n_1,\ldots,n_{|W|}),i,j,l)}. \tag{8}$$

Note that Equation (8) is a minimization problem, as penalty functions are non-decreasing in the response time.

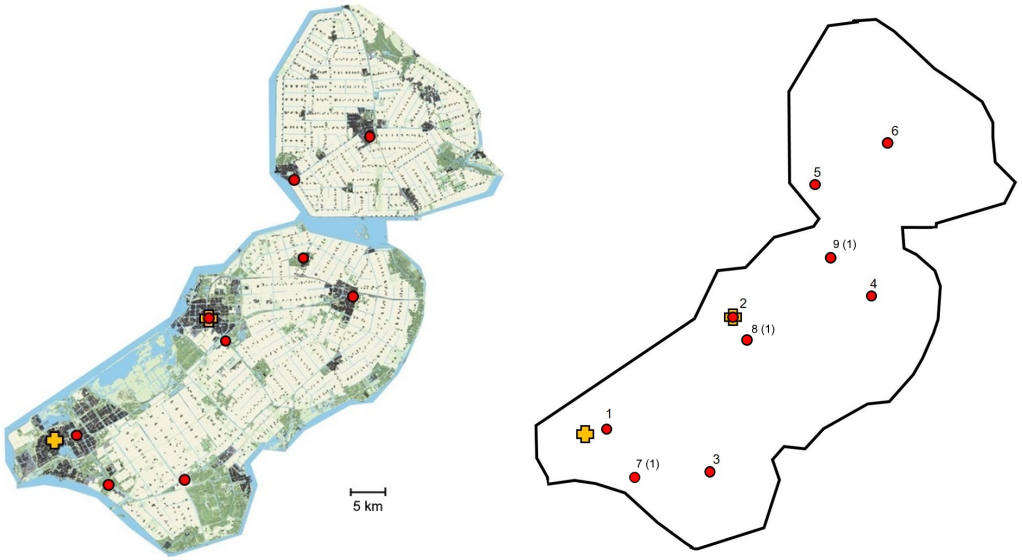
4 Numerical Study

The purpose of this section is to show computational results on the performance regarding the in- and exclusion of the described features in the algorithms explained in Section 3. Results are obtained by trace-driven simulations using historical data for two EMS regions in The Netherlands.

4.1 Experimental Setup

We base our computations on two different EMS regions in The Netherlands: the EMS regions of Flevoland and Amsterdam. These regions are opposites of each other in terms of size and population. Flevoland is a large yet sparsely populated region, according to Dutch standards. On the other hand, Amsterdam is small but urban. Next, we will describe the regions in more detail. We refer to Figures 2 and 3 for a geographical representation of Flevoland and Amsterdam, respectively.

Flevoland. Flevoland covers approximately 1,400 km^2 and is home to nearly 400,000 people. Almost half of the total population of Flevoland lives in the city indicated with a '1' in Figure 2b. The remaining population is mainly concentrated in one of the five other towns, although a couple of small villages exist as well, especially in the north-east. An ambulance waiting site, indicated by a dot in Figure 2, is located in or near each of the six major towns. There are three additional waiting sites, with a capacity of one ambulance, located at strategic places in the region. The crosses in this



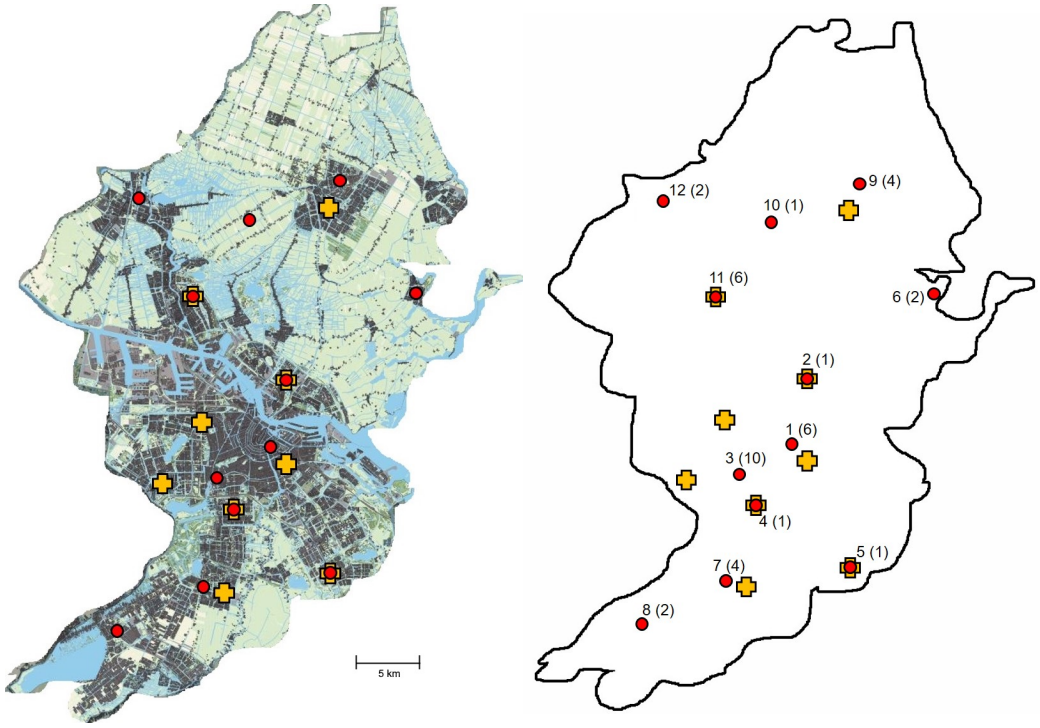
(a) (b)

Figure 2: EMS region of Flevoland.

figure mark the two hospitals in Flevoland. We aggregate the region into 93 demand nodes, based on 4-digit postal codes. Note that the postal code corresponding to the dot indicated by a '2' contains both a waiting site and a hospital.

Amsterdam. The EMS region containing the city of Amsterdam and its surroundings is approximately 630 km^2. However, the population of Amsterdam is three times larger than that of Flevoland: 1.2 million inhabitants. Approximately 68% lives in Amsterdam itself, while the northern part of the region is less densely populated. Ambulance waiting sites and hospital are present at the dots and crosses in Figure 3, respectively. The numbers in brackets denote the actual waiting site capacities. The region is aggregated into 162 postal codes, which serve as demand points. Moreover, both a waiting site and a hospital are present in the postal codes corresponding to dots 2, 4, 5, and 11. Approximately 73% of the patients needs transportation to a hospital.

Historical data on emergency requests in the year 2011 was provided by *GGD Flevoland* and *Ambulance Amsterdam*, the ambulance service providers of Flevoland and Amsterdam, respectively. We built two traces based on this data and simulate them in a discrete-event simulation. The trace is constructed as follows. We consider all emergency requests occuring between 7 AM and 6 PM, generally the busiest time of the day. In the trace, we



(a) (b)

Figure 3: EMS region of Amsterdam.

include the following incident related information:

- Time of occurence, i.e., the time of the emergency call;
- Location of occurence (postal code);
- Time spent on-scene by the ambulance;
- Hospital transfer time.

Emergency requests of which above data is not complete or infeasible are ignored. We are interested in an algorithm that performes well for *most* days. Therefore, we classify the days for which the number of incidents falls outside the interval $[\mu-2\sigma, \mu+2\sigma]$ as outliers, where μ and σ denote the mean number of requests per day and the standard deviation, respectively. This results in an exclusion of two days for both regions. Moreover, we remove the last 12 days of the year because the fleet capacity was inadequate. We connect the remaining 352 days such that 6 PM is followed directly by 7 AM the next day to ensure that the ambulance system is in continuous operation. This avoids

that the system becomes empty over night, and thereby our aproach allows us to obtain measurements that are close to 'steady state', which is what we are interested in. In the resulting trace 7,632 resp. 41,996 incidents occur in Flevoland and Amsterdam, respectively. This yields an hourly arrival rate of 1.97 resp. 10.84 emergency requests. Moreover, around 87% resp. 73% of the patients needs transportation to a hospital. The average busy time of an ambulance is 0.74 resp. 0.73 hours, excluding relocation time after the transfer. In order to ensure an out-of-sample validation, we estimate the demand probabilities per postal code based on the year 2010, and not 2011.

In our simulation, the closest idle ambulance always responds to the incident. If no ambulance is available, the call enters a queue. Once an ambulance becomes available from service again, it is immediately dispatched to the longest waiting request. Moreover, if a patient needs transportation to a hospital, the closest hospital is selected. In the simulation model, we use travel times estimated by the RIVM[4], which provided us tables containing travel times between each pair of postal codes in the regions of consideration. We refer to [15] for a more detailed description on the travel time model used for the estimation of these travel times. We interpret the travel times in these tables as the arc lengths $\tau^{(1)}$. The travel times $\tau^{(2)}$ are obtained by multiplying $\tau^{(1)}$ with a multiplication factor of $\frac{10}{9}$. We do not simulate a dispatch time or pre-trip delay.

We test the performance of the methods considered on the following seven statistics:

1. Percentage on time: the fraction of requests responded to within the response time threshold of 12 minutes. Actually, the statutory threshold in The Netherlands is 15 minutes, but typically 3 minutes are reserved for handling the phone call and the pre-trip delay. We also provide confidence intervals.

2. Mean response time.

3. Number of relocations. This number includes the relocation of an ambulance that just finished service as well.

4. Average relocation time. Note that this number is solely based on the travel times $\tau^{(2)}$ since it is not allowed to perform a relocation with optical signals and sirens turned on.

5. Total relocation time.

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

6. Mean single coverage. Each time a relocation decision is made in the simulation, the distribution of ambulance vehicles over waiting sites changes. At that moment, we compute the coverage of the region as if each idle ambulance was already at its destination, based on the assumption that a demand point is covered if it is covered by at least one ambulance (single coverage). This coverage value lasts until the time of the next event: the arrival or completion of a call. The reported percentage is a time-average over the complete simulation horizon.

7. Mean MEXCLP coverage. The computation of this value is similar to the computation of the mean single coverage, but we use the MEXCLP coverage instead.

The number of ambulances we assume to be on duty is smaller than the number in reality. This is because we focus on the urgent transports, while the ambulance providers in practice sometimes also respond to non-urgent requests using the same vehicles. These non-urgent requests are a taxi-like transports of patients that are not able to travel to the hospital themselves. These requests are of a different nature, since they can usually be scheduled in advance, and therefore we do not wish to mix the two cases in our analysis. In our implementation, we choose a fleet size such that a 'good' policy gives a performance of a magnitude that is realistic for practical purposes: 10 resp. 18 ambulances for Flevoland resp. Amsterdam. Busy fractions $q = 0.1716$ resp. $q = 0.4991$ are computed by dividing the total patient-related work by the total duty time of all ambulances.

4.2 Original DMEXCLP method

In this section, we report results for both regions of interest, Flevoland and Amsterdam, of the original DMEXCLP method, as proposed in [14]. Moreover, we compare these results to the static policy according to the MEXCLP solution: each ambulance returns to its home base station when newly idle. Results are listed in Table 1.

A large performance improvement in terms of late arrivals can be observed in Table 1 for the Amsterdam region. This quantity decreased from on average 6.19% to 4.10%, a difference of 2.09 percentage point and a decrease of 33.76%, even outperforming the performance gain reported in the original article ([14], for the region of Utrecht). However, the performance gain regarding this criterion is small for Flevoland: a difference of 0.11 percentage point, which is a decrease of only 2.1%. Moreover, the confidence bounds for this region overlap almost entirely. In addition, the gaps in mean single coverage and mean MEXCLP coverage between the static and DMEXCLP

Performance Indicators	Flevoland		Amsterdam	
	Static	DMEXCLP	Static	DMEXCLP
Percentage on time	94.86%	94.97%	93.81%	95.90%
Lower Bound 95%-CI	94.28%	94.45%	93.21%	95.40%
Upper Bound 95%-CI	95.45%	95.49%	94.43%	96.41%
Mean response time	304 s	303 s	371 s	329 s
Number of relocations	7,632	7,632	41,311	41,391
Average relocation time	437 s	814 s	384 s	585 s
Total relocation time	927 h	1,726 h	4,410 h	6,725 h
Mean single coverage	96.26%	96.63%	97.64%	98.81%
Mean MEXCLP coverage	93.24%	93.57%	93.43%	95.78%

Table 1: Simulation results for the static and DMEXCLP policy, based on 7,632 and 41,966 incidents in 2011, with 10 and 18 ambulances, respectively.

policy are much smaller for Flevoland. This was already foreseen in [14], and a possible explanation for this phenomenon is given: the DMEXCLP method is designed for busy areas in particular. The hourly arrival rate of incidents in Flevoland is much smaller compared to the urban Amsterdam region. As a consequence, there are fewer relocation moments, inducing a smaller performance improvement. (In the next subsection, we allow additional decision moments.)

In contrast to Flevoland, the number of ambulance relocations in Amsterdam does not equal the number of incidents. This is explained by the fact that in Amsterdam sometimes the situation occurs that none of the ambulances is available for a reported incident. As soon as an ambulance finishes service of a patient, it is immediately dispatched to a waiting call. This is not recorded as a relocation and hence, the number of relocations does not necessarily equal the number of incidents. Based on Table 1 one can compute that the total number of incidents for which no ambulance was immediately available, equals 655 and 575 for the static and DMEXCLP policy, respectively.

Note that both the mean single and MEXCLP coverage performance indicators serve as an estimate of the number of calls for which the response time treshold is achieved. As observed in Table 1, the mean single coverage is an optimistic approximation of this quantity for both policies, as expected. After all, ambulance unavailability is not taken into account in the concept of single coverage. The relative gap between mean single coverage and percentage on time is smaller for Flevoland, compared to Amsterdam, for both policies. This is not very surprising, since in Flevoland the overlap in cover-

age of multiple ambulances is very small: the distances between the 6 large towns generally exceed the time threshold. Only multiple ambulances parked at one and the same waiting site do provide overlapping coverage. Furthermore, the busy fraction in Flevoland is relatively low. Therefore, the error made when ignoring ambulance unavailability will also be small.

Even for Flevoland, the mean MEXCLP coverage over time turns out to be a more accurate approximation for the on time arrivals, although there is still a small gap. Note that for Amsterdam the mean MEXCLP coverage is closer to the observed percentage on time. We conjecture that this is probably due to the way in which the coverage is computed. As explained earlier, we compute this based on the configuration in which each ambulance is at its destination. For Amsterdam, the time until the desired ambulance configuration is attained is much shorter as a consequence of both a smaller area and a larger number of waiting sites, compared to Flevoland. Therefore, the mean MEXCLP coverage is a more accurate estimate on the percentage on time for Amsterdam than for Flevoland.

4.3 Decision Moments

As explained in Section 3.3, we allow the dispatcher to perform an ambulance relocation if the number of available ambulances decreases, just after the dispatch. As a consequence the number of opportunities to steer the system is multiplied by 2. Results are displayed in Table 2. In this table and the forthcoming ones, the default policy is the DMEXCLP policy explained in Section 3.1, without any additional features. This policy outperforms the static policy, commonly used as benchmark policy in ambulance literature, on the most important performance indicators, as Table 1 underlines.

For the percentage on time criterion, we observe an increase of 0.63 and 0.45 percentage point for Flevoland and Amsterdam, respectively. That is, the number of late arrivals decreased with 12.53% and 10.98%. We conclude that for Flevoland, the effect of adding additional relocation moments is much larger than the original effect of changing from static ambulance planning to the default move-up method (which was 2.1%). For Amsterdam, the default move-up already had a large effect, hence the added benefit of additional relocation moments seems smaller in comparison.

Surprisingly, the results on mean response times do not concur with those on the late arrivals criterion: in Flevoland, a performance gain of only 1.64% is achieved. In contrast, the mean response time in Amsterdam decreases with 7.44%. A possible explanation for this behaviour is as follows: since Flevoland is a rural region, an ambulance travelling between two waiting sites provides no or very little coverage. After all, few people live in the areas

Performance Indicators	Flevoland		Amsterdam	
	Default[5]	Moments	Default	Moments
Percentage on time	94.97%	95.60%	95.90%	96.35%
Lower Bound 95%-CI	94.45%	95.06%	95.40%	95.87%
Upper Bound 95%-CI	95.49%	96.14%	96.41%	96.83%
Mean response time	303 s	299 s	329 s	306 s
Number of relocations	7,632	13,308	41,391	76,161
Average relocation time	814 s	1,367 s	585 s	730 s
Total relocation time	1,726 h	5,054 h	6,725 h	15,453 h
Mean single coverage	96.63%	97.34%	98.81%	99.10%
Mean MEXCLP coverage	93.57%	94.61%	95.78%	96.76%

Table 2: Simulation results for Flevoland and Amsterdam, based on 7,632 and 41,966 incidents in 2011, with 10 and 18 ambulances, respectively.

between the cities, cf. Figure 2. In contrast, a large part of the Amsterdam region is urban, cf. Figure 3. In an urban area, an ambulance performing a relocation drives through a densely populated area, being able to respond to an incoming call in that area quickly. As the number of ambulance relocations almost doubles for both regions, this effect will be largest in Amsterdam, resulting in a relative large decrease in mean response time.

In the crew-related performance indicators, we observe both an increase in number of relocations and average relocation time. As a consequence, the total relocation time is more than doubled. A trade-off between patient- and crew-based performance, which is the subject of [27], is clearly visible here as well. The question arises whether this large increase outweighs the gain in patient-based performance. It is up to the ambulance service provider to decide on this, but we suspect that the answer depends on the daily workload of the crew. As this is typically lower in rural regions, we expect those EMS providers to be more open to additional relocation moments.

Note that for Amsterdam the mean MEXCLP coverage is now an optimistic estimate for the number of calls responded to within the time threshold, if more decision moments are allowed. We conjecture that this is due to the 'intended configuration', on which the computation of the mean MEXCLP coverage is based, changes so often that only a small fraction of these configurations is actually attained. That is, the steering towards the intended ambulance configuration is often interrupted by a new decision moment, which results in a different desired configuration.

[5]In this table and the forthcoming ones, the default policy is the DMEXCLP policy explained in Section 3.1, without any additional features.

4.4 Hospitals

In this section, we explore the differences in performance if ambulances transferring patients at hospitals are taken into account. We do this in two ways. First, we consider the data obtained via the ambulance service providers and fit a distribution on the busy times of an ambulance at a hospital. As mentioned in Section 3.3, we plug in the expected remaining service time in the formula given the hospital time already elapsed. As an alternative approach, we simulate the system in which we have 'perfect information' regarding the hospital transfer time. We assume that we know this time when an ambulance arrives at the hospital, which results in a deterministic remaining service time. This approach clearly is a rather optimistic approach, and it can be interpreted as a bound on the knowledge that one can have on the remaining service time. However, this approach is more realistic than one might expect at first glance, as ambulance crews and dispatchers in The Netherlands are able to estimate the hospital transfer time rather accurately[6]. In particular, hospitals in The Netherlands do not suffer from queues building up at an emergency department, in contrast to North America where the average transfer time can be very large and highly variable, cf. [7].

We estimate the service time at a hospital by a Weibull distribution, for both regions. In our experience, this distribution provides a rather accurate approximation. Moreover, a Weibull distribution for this quantity was also used in both [18] and [26]. The means of the fitted distributions are 966 seconds and 1,160 seconds for Flevoland and Amsterdam, respectively. The differences in mean are probably explained by the fact that the hospitals in Amsterdam are typically larger, and thus the ambulance personnel spends more time on the transport of the patient to the appropriate department within the hospital. Based on the Weibull distributions, we calculate the expected remaining transfer time for each possible value of service time already elapsed.

In Table 3, we listed simulated results on the assumption of Weibull distributed transfer times and perfect information, and we compare those to the default policy explained above. We observe neither an increase nor a decrease in the patient-related performance indicators in the Weibull case. A small decrease in average relocation time can be noted, which has a small effect on the total relocation time as well. Based on these observations, one might conclude that the inclusion of ambulances busy at a hospital in the algorithm in the way described in Section 3.3 does not influence the performance.

Alternatively, the Weibull distribution used for the estimation of the

[6] as we have learned from discussions with dispatchers and management.

Performance Indicators	Flevoland			Amsterdam		
	Default	Weibull	Perfect	Default	Weibull	Perfect
Percentage on time	94.97%	94.97%	95.00%	95.90%	95.85%	95.91%
Lower Bound 95%-CI	94.45%	94.46%	94.47%	95.40%	95.35%	95.40%
Upper Bound 95%-CI	95.49%	95.48%	95.52%	96.41%	96.34%	96.42%
Mean response time	303 s	304 s	304 s	329 s	329 s	330 s
Number of relocations	7,632	7,632	7,632	41,391	41,383	41,394
Average relocation time	814 s	806 s	777 s	585 s	583 s	551 s
Total relocation time	1,726 h	1,709 h	1,647 h	6,726 h	6,702 h	6,341 h
Mean single coverage	96.63%	96.62%	96.62%	98.81%	98.81%	98.82%
Mean MEXCLP coverage	93.57%	93.56%	95.55%	95.78%	95.77%	95.75%

Table 3: Simulation results for Flevoland and Amsterdam, based on 7,632 and 41,966 incidents in 2011, with 10 and 18 ambulances, respectively.

transfer time may perhaps be a poor approximation. To test whether this indeed may be the case, we simulate the system in which we have perfect information about the transfer time to exclude this source of randomness. However, we do not observe an improvement in the patient-related performance indicators. Based on these results, we claim that taking into account ambulances busy at a hospital in the way we did (as explained in Section 3.3), has no effect on the patient-related performance, regardless the distribution used.

In contrast, the assumption of perfect information leads to a shorter average relocation time of 4.5% and 5.8% for Flevoland and Amsterdam, respectively, while the number of relocations stays equal. As a consequence, the relocations are shorter. This is probably explained by the fact that ambulances at hospitals contribute to the coverage in the near surroundings of that hospital. Therefore, decisions made while the ambulance was in the hospital, would typically *not* have sent idle vehicles towards this hospital area[7]. When the ambulance eventually becomes available, it is therefore more likely that it is needed to provide coverage in the area close to the hospital.

4.5 Chain relocations

In [27], it is stated that it is beneficial to use chain relocations: the break-up of a certain long lasting relocation into multiple short relocations by different ambulances. Moreover, their computational results - based on the same regions considered in this paper - show substantial benefit when using

[7]or at least, not as much as the default algorithm would have

Performance Indicators	Flevoland		Amsterdam	
	Default	Chains	Default	Chains
Percentage on time	94.97%	94.89%	95.90%	95.89%
Lower Bound 95%-CI	94.45%	94.39%	95.40%	95.35%
Upper Bound 95%-CI	95.49%	95.39%	96.41%	96.43%
Mean response time	303 s	306 s	329 s	331 s
Number of relocations	7,632	11,619	41,391	64,998
Average relocation time	814 s	563 s	585 s	415 s
Total relocation time	1,726 h	1,816 h	6,726 h	7,490 h
Mean single coverage	96.63%	96.57%	98.81%	98.78%
Mean MEXCLP coverage	93.57%	93.51%	95.78%	95.72%

Table 4: Simulation results for Flevoland and Amsterdam, based on 7,632 and 41,966 incidents in 2011, with 10 and 18 ambulances, respectively.

two links instead of one, but more than two links appears to be redundant. We simulate the system according to this regime: a relocation is decomposed into a chain relocation of length two if this reduces the time until the new configuration is attained. Results are displayed in Table 4.

Although the time until the desired configuration is attained is decreased, we do not observe a gain on the patient-related performance criteria. Instead, even a slight deterioration can be seen in Table 4. This contradicts the findings of [27]. This is probably due to the fact that in [27] extra decision moments are allowed, as considered in Sections 3.3 and 4.3. In Section 4.7, we will study the effect of the combination of extra decision moments and chain relocations.

As expected, the number of relocations increases a lot in a regime in which chain relocations are allowed. In approximately 52% of the times an ambulance becomes available, an additional ambulance is relocated in Flevoland. This percentage for Amsterdam is approximately 56%. One would expect this percentage for Amsterdam to be much higher, as more waiting sites and ambulances are present in Amsterdam. Hence, there are more possibilities to set up a chain relocation. However, the distances between waiting sites in this region are shorter, whereby the gain of chain relocations is probably smaller. This is also reflected in the average relocation time. Of course, this quantity decreases tremendeously for both regions, but the relative decrease for Flevoland is much larger, as a consequence of the longer distances between waiting sites.

4.6 Relocation time bounds

As explained in Section 3.3, we impose different bounds on the relocation time of an ambulance. This bound is given by the variable B. If there is no waiting site that can be reached within B minutes exists, the ambulance travels to the nearest waiting site. For $B = 0$, the obtained policy is equivalent to this 'nearest base'-policy. In Figures 4 and 5 we show results on the most important patient- and crew-related performance indicators: percentage on time and total relocation time, as function of B. In Tables 5 and 6 results on all performance indicators are displayed for $B = 0, 10, 20, 30$ minutes.

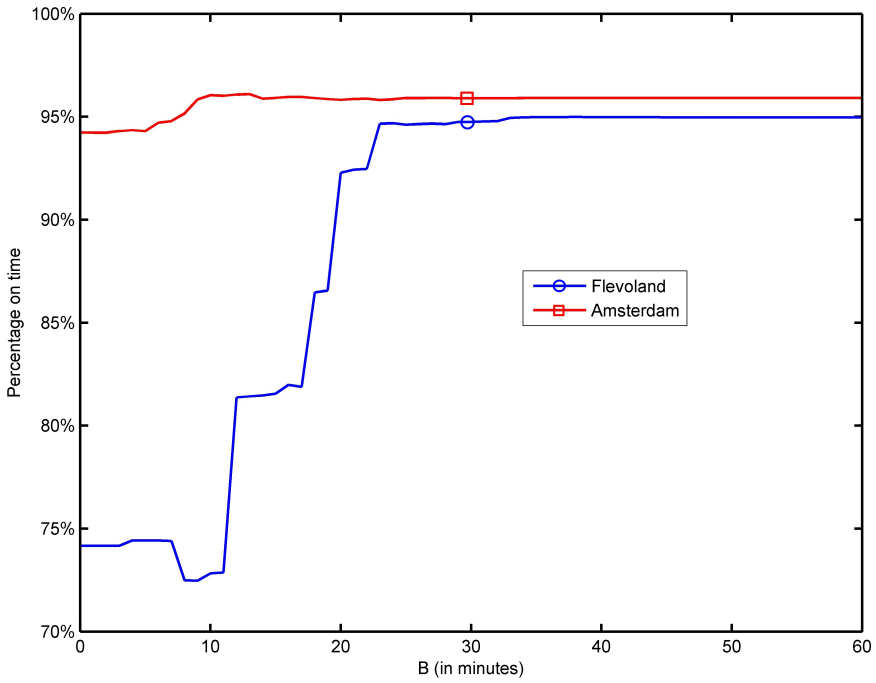


Figure 4: Percentage on time as function of B.

In Figure 4 we observe a large difference in the system's behaviour. For Amsterdam, the bound B is of little influence only: the percentage of calls reached within the time threshold is close to 95% for all levels of B. In contrast, we see a huge improvement in performance for larger values of B in Flevoland: for $B < 12$ the percentage on time is below 75% and this increases up to approximately 95%. This phenomenon has a simple explanation: it is a consequence of both the size and the number of waiting sites and hospitals in Flevoland. The mean distances between two waiting sites are much larger,

Performance Indicators	$B = 0$ min	10 min	20 min	30 min
Percentage on time	74.17%	72.83%	92.28%	94.75%
Lower Bound 95%-CI	73.00%	71.46%	91.49%	94.16%
Upper Bound 95%-CI	75.32%	74.19%	93.08%	95.33%
Mean response time	495 s	496 s	335 s	308 s
Number of relocations	7,632	7,632	7,632	7,632
Average relocation time	79 s	153 s	607 s	670 s
Total relocation time	168 h	325 h	1,286 h	1,420 h
Mean single coverage	75.59%	74.87%	94.19%	96.42%
Mean MEXCLP coverage	74.61%	73.16%	91.17%	93.33%

Table 5: Simulation results for Flevoland based on 7,632 incidents in 2011, with 10 ambulances. Results on relocation bounds 0, 10, 20, 30 minutes are displayed.

Performance Indicators	$B = 0$ min	10 min	20 min	30 min
Percentage on time	94.23%	96.05%	95.82%	95.90%
Lower Bound 95%-CI	93.72%	95.55%	95.29%	95.40%
Upper Bound 95%-CI	94.74%	96.54%	96.35%	96.40%
Mean response time	323 s	322 s	330 s	329 s
Number of relocations	41,398	41,388	41,390	41,391
Average relocation time	131 s	341 s	568 s	585 s
Total relocation time	1,504 h	3,919 h	6,535 h	6,726 h
Mean single coverage	97.69%	98.63%	98.80%	98.81%
Mean MEXCLP coverage	93.60%	95.55%	95.75%	95.78%

Table 6: Simulation results for Amsterdam based on 41,966 incidents in 2011, with 18 ambulances. Results on relocation bounds 0, 10, 20, 30 minutes are displayed.

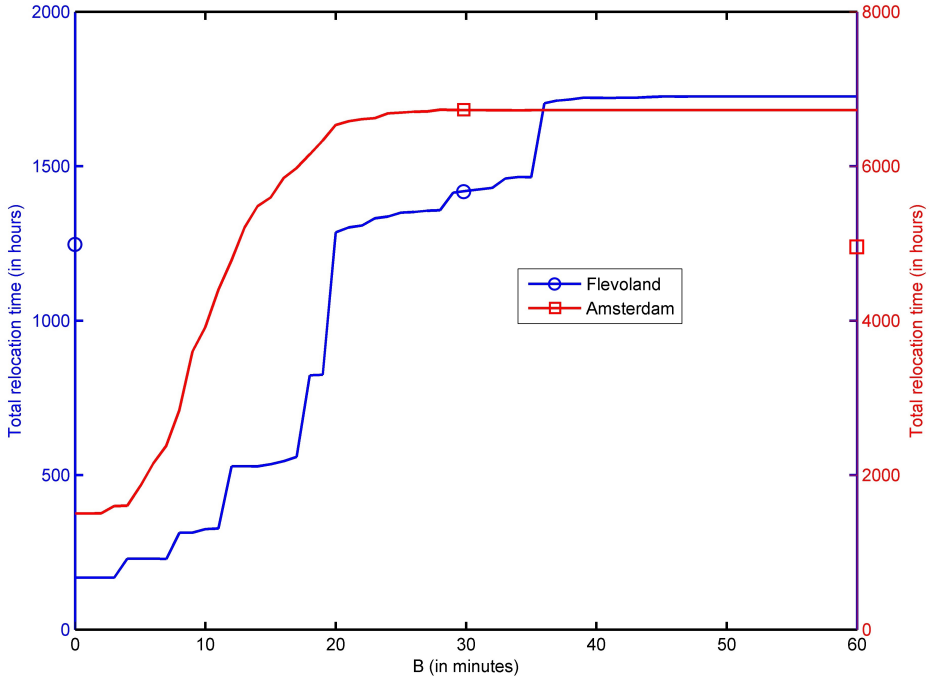


Figure 5: Total relocation time as function of B.

so for small values of B there are few possibilities for the destination of an ambulance after a service completion. Moreover, since there are only two hospitals in the region and approximately 75% of the ambulances become available there, relocations to waiting sites 3, 4, 5 and 6 do not take place.

Another interesting point is the drop between $B = 7$ and $B = 8$ for Flevoland. This behaviour is due to one relocation in particular: the relocation time for an ambulance between the hospital in city 1 and waiting site 7 is exactly 7.5 minutes. Thus, for $B = 7$, an ambulance becoming free at this hospital moves to waiting site 1, regardless of the number of ambulances already present there. In contrast, for $B = 8$, this ambulance travels to waiting site 7, if unoccupied. The benefit of covering the southeastern part is outweighed by the performance loss in city 1. This aspect can be observed in the coverages displayed in Table 5 as well.

All large jumps are easily explained as well: the jump at $B = 12$ is due to the allowance of a relocation from 2 to 9; the one at $B = 18$ is due to the relocation from 1 to 3. If $B = 20$, it is now allowed to relocate an ambulance from 2 to both 4 and 5 as well. Finally, waiting site 6 can be reached from 2 if B exceeds 23 minutes. These jumps are largely visible in Figure 5 as well.

Moreover, the large increase in total relocation time at $B = 36$ is due the fact that relocations from 1 to 4 and 6 both are acceptable now.

The pattern for Amsterdam is of different shape: the best performance is achieved for $10 \leq B \leq 13$, although the differences are minor. Apparently, it is beneficial to the performance if one chooses a relatively close waiting site if an ambulance is newly free. That is, a local optimum that can be reached quickly performs better than a global one for which it takes long until that configuration is attained. A possible explanation for this phenomenon is the large number of events and thus decision moments in Amsterdam. This behaviour is also reflected in Table 6: the coverage levels belonging to $B = 30$ are higher than for $B = 10$, although $B = 10$ yields a larger percentage on time. Note that there is also a reduction in mean response time of approximately 2.1% for $B = 10$ compared to $B = 30$.

4.7 Combinations

In this section, we will combine different highly promising features and test the method for both regions. Moreover, we compare the performance with two other policies: the penalty heuristic of [27] summarized in Section 3.2 and a *compliance table policy*. A compliance table indicates the desired configuration for each number of available ambulances. We test the following combinations and methods:

1. DMEXCLP with extra decision moments, with chain relocations, without taking into account ambulances busy at hospitals.

2. DMEXCLP with extra decision moments, with chain relocations; busy time at the hospital follows the Weibull distribution considered in Section 4.4.

3. Similar to 2, but now we have perfect information about the transfer times.

4. Compliance table: to obtain the desired configurations per number of available ambulances, we solve multiple MEXCLP problems. The computed compliance tables are displayed in Table A1. We do not allow chain relocations.

5. The same compliance table is used, but we allow chain relocations now.

6. Penalty heuristic, (see Section 3.2).

Performance Indicators	Flevoland					
Combination:	1	2	3	4	5	6
Percentage on time	96.24%	96.24%	96.27%	95.15%	95.41%	94.22%
Lower Bound 95%-CI	95.79%	95.77%	95.82%	94.55%	94.82%	93.64%
Upper Bound 95%-CI	96.69%	96.71%	96.71%	95.76%	96.01%	94.80%
Mean response time	292 s	292 s	292 s	305 s	307 s	288 s
Number of relocations	24,747	24,408	23,481	29,518	49,466	22,047
Average relocation time	774 s	766 s	766 s	991 s	688 s	599 s
Total relocation time	5,318 h	5,196 h	4,997 h	8,126 h	9,447 h	3,671 h
Mean single coverage	97.34%	97.34%	97.34%	97.24%	97.18%	97.43%
Mean MEXCLP coverage	94.61%	94.60%	94.58%	94.33%	94.27%	93.24%

Table 7: Simulation results for different combinations for Flevoland, based on 7,632 in 2011, with 10 ambulances.

Performance Indicators	Amsterdam					
Combination:	1	2	3	4	5	6
Percentage on time	97.23%	97.21%	97.26%	95.60%	95.36%	97.10%
Lower Bound 95%-CI	96.82%	96.77%	96.84%	95.11%	94.82%	96.68%
Upper Bound 95%-CI	97.64%	97.66%	97.67%	96.09%	95.90%	97.51%
Mean response time	303 s	302 s	302 s	322 s	325 s	283 s
Number of relocations	132,918	132,530	127,467	315,629	414,782	129,988
Average relocation time	440 s	439 s	424 s	456 s	372 s	457 s
Total relocation time	16,258 h	16,172 h	15,026 h	40,009 h	43,169 h	16,486 h
Mean single coverage	99.12%	99.11%	99.13%	99.10%	99.00%	99.34%
Mean MEXCLP coverage	96.79%	96.78%	96.75%	96.60%	96.60%	95.62%

Table 8: Simulation results for different combinations for Amsterdam, based on 41,966 in 2011, with 18 ambulances.

Results are displayed in Tables 7 and 8. Although allowing chain relocations initially did not result in better performance regarding the percentage on time criterion, as observed in Table 4, it is a valuable addition if it is combined with the allowance of extra decision moments, for both regions. If we compare Table 2, which shows the best performance concerning this criterion up to now, with the first columns in Tables 7 and 8, we see that performance improvements of 0.64 resp. 0.88 percentage points are achieved for Flevoland and Amsterdam, respectively. That is, the number of late arrivals decreased with 14.55% and 24.11%. This behaviour is probably explained by the following reason: it is more likely that a poor ambulance configuration arises just after the dispatch than when an ambulance becomes available. Therefore, at that decision moment, it is more important to attain the desired configuration quickly. This is achieved by using chain relocations, explaining the difference in performance.

If we compare columns 1, 2 and 3 in Tables 7 and 8, we barely see any differences in patient-based performance. This underlines the observations in Section 4.4. Results on crew-based performance are similar to those obtained in Section 4.4 as well.

Note that the DMEXCLP method in which extra decision moments and chain relocations are allowed (columns 1, 2 and 3 in Tables 7 and 8) performs significantly better than the MEXCLP compliance table policy on the percentage on time criterion. Moreover, it also outperforms the compliance table on the crew-related performance indicators. We conclude that although allowing for chain relocations in the compliance table policy (column 5) reduces the average relocation time, this effect is outweighed by the dramatic increase in number of relocations.

Both the DMEXCLP method with its features and the MEXCLP compliance table policies are quite consistent in their behaviour for both regions, although the regions of consideration differ heavily. The penalty heuristic, however, shows different performance: it performs comparably to the DMEXCLP method for Amsterdam, while for Flevoland it is outperformed even by the compliance table policy. A simple explanation for this phenomenon has its roots in the concept of single coverage: the method tries to maximize the demand covered at least once. This results in the relocation of ambulances to each outskirt of the region in Flevoland. As a consequence, it 'misses' a second call occuring shortly after a first one in one of the two large cities, in which approximately 75% of the incidents occur: ambulances located in the towns 3, 4, 5, and 6 are not able to arrive in cities 1 and 2 within the time threshold, resulting in a worse performance. In contrast, the distances from waiting sites to postal codes are much shorter in Amsterdam, and as a side effect, a postal code is typically automatically multiple covered, even

the algorithm focuses on maximizing single coverage.

Note that the penalty heuristic does not focus on coverage solely, but it uses the penalty function of Equation (3). One can observe in Tables 7 and 8 that minimizing the average response time is included in this penalty function as well, as this method yields the shortest mean response time for both regions. In addition, the single coverage concept is used in the penalty heuristic. As a consequence, the mean single coverage levels are highest for the penalty heuristic, at the expense of a lower mean MEXCLP coverage.

If we modify the DMEXCLP method of [14] in such a way that extra decision moments and chain relocations are allowed, we observe an improvement over other policies on most performance indicators if the coverage penalty function is used. In the next section, we consider different penalty functions and explore the performance of the DMEXCLP method with additional features.

4.8 Different performance criteria

For the study of different penalty functions we have chosen the DMEXCLP method in which we assume that the hospital transfer time follows a Weibull distribution (method 2 in the previous section). We consider the following penalty functions:

- $\Phi_1(t) = \mathbb{1}_{\{t>720\}}$: the coverage penalty function, with a time threshold of 720 seconds.
- $\Phi_2(t) = t$: this penalty function focuses on minimization of the average response time.
- $\Phi_3(t)$: the penalty function of Equation (3), which is a compromise between minimizing late arrivals and minimizing average response times.

Results are displayed in Table 9. One may expect that the number of late arrivals and average response time are positively correlated. However, the results contradict this hypothesis: an increase of 6.00% resp. 9.42% in late arrivals is observed if one uses Φ_2 instead of Φ_1, for Flevoland and Amsterdam, respectively. In contrast, the average response time is reduced with 5.82% and 11.59%, respectively. Similar behaviour was also observed in [26].

Concerning the mean response time, the results clearly indicate that Φ_3 is a compromise between Φ_1 and Φ_2. This is not reflected in the percentage on time, however: surprisingly, the incorporation of Φ_3 into the DMEXCLP method with additional features performs slightly better than Φ_1, which focuses on maximizing this quantity. (Although it should be noted that the confidence intervals largely overlap.)

Performance Indicators	Flevoland			Amsterdam		
	$\Phi_1(t)$	$\Phi_2(t)$	$\Phi_3(t)$	$\Phi_1(t)$	$\Phi_2(t)$	$\Phi_3(t)$
Percentage on time	96.24%	95.96%	96.31%	97.21%	96.92%	97.32%
Lower Bound 95%-CI	95.77%	95.48%	95.84%	96.77%	96.53%	96.95%
Upper Bound 95%-CI	96.71%	96.45%	96.77%	97.66%	97.32%	97.70%
Mean response time	292 s	275 s	285 s	302 s	267 s	282 s
Number of relocations	24,408	24,287	26,122	132,530	134,113	134,162
Average relocation time	766 s	727 s	744 s	439 s	418 s	424 s
Total relocation time	5,197 h	4,907 h	5,401 h	16,173 h	15,580 h	15,813 h
Mean single coverage	97.34%	97.31%	97.35%	99.11%	98.99%	99.15%
Mean MEXCLP coverage	94.60%	94.09%	94.59%	96.78%	96.24%	96.80%

Table 9: Simulation results for Flevoland and Amsterdam, based on 7,632 and 41,966 incidents in 2011, with 10 and 18 ambulances, respectively.

5 Conclusion

In this paper, we studied the implementation of several aspects and features present in [27] in the dynamic relocation method proposed in [14]. Next, We draw conclusions and make recommendations.

Based on the results in Table 9, we would suggest to use $\Phi_3(t)$ in a DMEXCLP environment. However, we want to note that $\Phi_1(t)$ makes for a fine alternative, as the results only differ slightly (7 to 20 seconds for the average response time). A reason to choose $\Phi_1(t)$ could be to make it easier to explain the behaviour of the system to EMS management and/or crew.

Adding extra decision moments (i.e., also relocating when a vehicle is dispatched to an incoming incident) is something we highly recommend in rural regions. We draw this conclusion based on the results in Table 2. For urban regions, we consider this an optional extra, that may be implemented if the region is willing to increase the crew's workload. Moreover, we recommend the use of chain relocations only if these extra decision moments are added. After all, Table 4 shows that no performance gain is achieved, while the workload on the crew is much higher. In contrast, if extra decision moments are added, the effect of chain relocations on the performance is much larger, cf., Tables 7 and 8.

When it comes to ambulances involved in a drop-off at a hospital, our initial recommendation is to ignore them (in terms of coverage provided). The reason for this, is that including them makes the move-up somewhat harder to implement (and explain), while it does not benefit the patients. An exception to this rule could be, when an EMS crew struggles with their workload: in that case, including the ambulances at hospital could be worth-

while, because it slightly reduces the relocation times (as seen in Table 3).

Before implementing any ambulance move-up policy, we have one final – and very important – recommendation. Perform simulation experiments in order to get a realistic idea of what effect the move-up policy has on response times. Keep in mind that every region is different, and that it is very hard to predict effects in a system as complex and stochastic as ambulance services. Mathematical models should be used with care in complex systems in practice: in our opinion simulation is an important tool that can truly capture the behaviour of the system.

Acknowledgements

We would like to thank the ambulance service providers of the EMS regions of Flevoland, GGD Flevoland, and Amsterdam, Ambulance Amsterdam, for providing data. In addition, we are grateful to the RIVM for providing the travel times for ambulances in the EMS regions considered in the numerical study. This research was financed in part by Technology Foundation STW under contract 11986, which we gratefully acknowledge.

References

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

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

[3] R. Batta, J. Dolan, and N. Krishnamurthy. The maximal expected covering location problem: Revisited. *Transportation Science*, 23(4):277–287, 1989.

[4] V. Bélanger, A. Ruiz, and P. Soriano. Recent advances in emergency medical services management. 2015.

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

[6] R.E. Burkhard, M. Dell'Amico, and S. Martello. *Assignment Problems*, chapter 6. SIAM, Philadelphia, 2009.

[7] A. Carter, J. Gould, P. Vanberkel, J. Jensen, J. Cook, S. Carrigan, M. Wheatley, and A. Travers. Offload zones to mitigate emergency medical services EMS offload delay in the emergency department: a process map and hazard analysis. *Canadian Journal of Emergency Medicine*, pages 1–9, 2015.

[8] M. Daskin. The maximal expected covering location model: Formulation, properties, and heuristic solution. *Transportation Science*, 17:48–70, 1983.

[9] D. Degel, L. Wiesche, S. Rachuba, and B. Werners. Time-dependent ambulance allocation considering data-driven empirically required coverage. *Health Care Management Science*, 18(4):444–458, 2015.

[10] E. Erkut, A. Ingolfsson, and G. Erdogan. Ambulance location for maximum survival. *Naval Research Logistics*, 55(1):42–58, 2008.

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

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

[13] M. Gendreau, G. Laporte, and F. Semet. The maximal expected coverage relocation problem for emergency vehicles. *Journal of the Operations Research Society*, 57:22–28, 2006.

[14] C.J. Jagtenberg, S. Bhulai, and R.D. van der Mei. An efficient heuristic for real-time ambulance redeployment. *Operations Research for Health Care*, 4:27 – 35, 2015.

[15] G. Kommer and S. Zwakhals. Referentiekader spreiding en beschikbaarheid ambulancezorg 2008, 2008.

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

[17] M. Maleki, N. Majlesinasab, and M. Mehdi Sepehri. Two new models for redeployment of ambulances. *Computers & Industrial Engineering*, 78:271–284, 2014.

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

[19] J. Naoum-Sawaya and S. Elhedhli. A stochastic optimization model for real-time ambulance redeployment. *Computers & Operations Research*, 40:1972–1978, 2013.

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

[21] J.F. Repede and J.J. Bernardo. Developing and validating a decision support system for location emergency medical vehicles in Louisville, Kentucky. *European Journal of Operational Research*, 75(3):567–581, 1994.

[22] V. Schmid. Solving the dynamic ambulance relocation and dispatching problem using approximate dynamic programming. *European Journal of Operational Research*, 219:611–621, 2012.

[23] V. Schmid and K. Doerner. Ambulance location and relocation problems with time-dependent travel times. *European Journal of Operational Research*, 207(3):1293–1303, 2010.

[24] K. Sudtachat, M.E. Mayorga, and L.A Mclay. A nested-compliance table policy for emergency medical service systems under relocation. *Omega*, 58:154–168, 2016.

[25] T.C. van Barneveld. The minimum expected penalty relocation problem for the computation of compliance tables for ambulance vehicles. *INFORMS Journal on Computing*. To appear.

[26] T.C. van Barneveld, S. Bhulai, and R.D. van der Mei. A dynamic ambulance management model for rural areas. *Health Care Management Science*, 2015. doi 10.1007/s10729-015-9341-3.

[27] T.C. van Barneveld, S. Bhulai, and R.D. van der Mei. The effect of ambulance relocations on the performance of ambulance service providers. *European Journal of Operational Research*, 2016. doi 10.1016/j.ejor.2015.12.022.

[28] P.L. van den Berg and K. Aardal. Time-dependent MEXCLP with start-up and relocation cost. *European Journal of Operational Research*, 242(2):383–389, 2015.

[29] L. Zhang. *Simulation Optimisation and Markov Models for Dynamic Ambulance Redeployment*. PhD thesis, The University of Auckland, 2012.

Appendix A MEXCLP Compliance Tables

Region	Level	Compliance Table
Flevoland	1	7
	2	7-8
	3	7-8-9
	4	6-7-8-9
	5	1-6-7-8-9
	6	1-3-6-7-8-9
	7	1-2-3-6-7-8-9
	8	1-2-3-4-6-7-8-9
	9	1-2-3-4-6-6-7-8-9
	10	1-1-2-3-4-6-6-7-8-9
Amsterdam	1	3
	2	2-3
	3	3-4-11
	4	2-3-4-11
	5	2-3-3-4-11
	6	2-3-4-7-10-11
	7	2-3-3-4-7-10-11
	8	2-3-4-7-7-10-11-11
	9	2-3-4-7-7-9-10-11-11
	10	2-3-3-4-7-7-9-10-11-11
	11	2-3-3-4-7-7-9-9-10-11-11
	12	1-2-3-4-7-7-7-9-9-10-11-11
	13	1-2-3-4-7-7-7-9-9-10-11-11-11
	14	2-3-3-4-6-7-7-7-9-9-10-11-11-11
	15	1-2-3-4-6-7-7-7-7-9-9-10-11-11-11
	16	1-2-3-4-6-7-7-7-7-9-9-10-11-11-11-11
	17	1-2-3-3-4-6-7-7-7-7-9-9-10-11-11-11-11
	18	1-2-3-3-4-6-7-7-7-7-9-9-9-10-11-11-11-11

Table A1: MEXCLP Compliance Tables