4

Practical Insights into the Implementation of a Relocation Policy

At the end of the previous chapter we mentioned the implementation of adjusted versions of the DMEXCLP method (Jagtenberg et al., 2015) and the penalty heuristic (Van Barneveld et al., 2016a) in a real-time decision support software tool in the Flevoland emergency control center (Van Buuren et al., 2016). This illustration of a successful application of academic research to practice motivated the developers of the DMEXCLP method and us, the designers of the penalty heuristic, to further enhance both algorithms. This chapter reports about the findings in our cooperation, in which we thoroughly analyze the dynamic ambulance relocation problem from a practical point of view. In some sense, it could be considered as a search for the 'best of both worlds' combination of the work done by Jagtenberg et al. (2015) and that by Van Barneveld et al. (2016a), the latter serving as the basis for Chapter 3. 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, recall that unlike many other relocation policies, these two methods have recently been tested in practice. This combination of properties makes these algorithms a natural choice for our investigation.

This chapter is based on the work by Van Barneveld et al. (2016b).

4.1 Introduction

Both the DMEXCLP method and the penalty heuristic have their strengths and shortcomings. This chapter is concerned with some interesting issues on the implementation of both relocation methods in practice, in order to improve the efficiency from a patient, but also, from a crew perspective. After all, although ambulance relocation 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. Therefore, we analyze the trade-off between the number of relocations, the total travel time needed for relocations, and the reduction in response times. We study the following topics:

The frequency of redeployment decision moments. Many papers consider a regime in which it is only allowed to relocate a vehicle at the end of a mission. However, if such a regime is adapted there is limited ability to control the system, which may cause only marginal performance improvements with respect to the so-called *static* policy in which an ambulance is always relocated to its home station. This especially holds for rural regions with a small incident arrival rate, and hence, a lower frequency at which ambulances become idle. On the other side, allowing dispatchers to relocate ambulances too often may lead to crew annoyance.

The inclusion of busy ambulances in the state description of the system. We investigate whether ambulance repositioning 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.

The performance criterion on the quality of the relocation strategy. It is commonly accepted to judge the ambulance service provider on the fraction of calls responded to within the time threshold. However, the limited discrimination in different response times is an important limitation (Erkut et al., 2008). Possibly, deviation from the generally adopted coverage criterion, considering a performance criterion that can be arbitrarily chosen through the selection of an appropriate penalty function, may result in better performance.

The use of chain relocations. The further we send an ambulance to, the longer it takes for the system to reach the desired configuration. To that end, we consider chain relocations in which multiple vehicles take part, thereby breaking up the long drive into several smaller ones, that may be executed simultaneously (see Figure 3.2).

Time bounds on the relocation time. As an alternative to chain relocations, which may inconvenience ambulance crews as their workload increases due to the number of extra relocations, we consider time bounds on the relocation time. That is, we ignore options that would take excessively long.

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 relocation 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. All our results are obtained from trace-driven simulations. While our primary focus is on minimizing the fraction of late arrivals, other metrics, such as crew related performance indicators, are also reported.

The remainder of this chapter is organized as follows. Section 4.2 introduces the model and the used notation, which largely coincides with that of the previous chapter. In Section 4.3 we summarize the DMEXCLP method and we modify this algorithm by the incorporation of features considered by Van Barneveld et al. (2016a) in the algorithm. We conclude this chapter with an extensive numerical study regarding the aspects treated in Section 4.3, for both the EMS region of Flevoland and Amsterdam.

4.2 Model

The ambulance redeployment model used in this chapter largely coincides with the one described in Section 3.2.1, except for two deviations: (1) no service preemption of the hospital transfer time is allowed, and (2) we distinguish travel times for both emergency and relocation purposes. The reasons behind these adjustments are of a practical nature: it is not generally adopted that ambulance crews can be forced to interrupt the hospital drop-off in practice, and units are typically not allowed to exceed the maximum speed limit for relocation purposes.

As in the previous chapters, 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. We model the region as a doubly weighted complete directed graph $G = (V \cup W, A, (\tau^{(1)}, \tau^{(2)}))$, in which V, W and A are as defined in Section 3.2.4. 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, then the optical and sound signals are not turned on. This yields a longer travel time, denoted by $\tau_{ij}^{(2)}$. Obviously, it holds that $\tau_{ij}^{(2)} \geq \tau_{ij}^{(1)}$. For an overview of notation we refer to Tables 3.1 and 4.1, the latter summarizing the newly introduced notation in this chapter.

4.3 Algorithms and Features

In this section, we explain the DMEXCLP method as published by Jagtenberg et al. (2015) and highlight the differences with the penalty heuristic presented

T	Response time threshold.
$\tau_{ij}^{(1)}$	Expected emergency travel time between nodes i and j.
$\tau_{ij}^{(2)}$	Expected relocation travel time between nodes i and j.
p	Busy fraction.
n_j	Number of ambulances having waiting site j as destination.

TABLE 4.1: Notation

in Chapter 3. 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 traveling 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 4.3.3 we describe the incorporation of several aspects considered by Van Barneveld et al. (2016a) into the DMEXCLP method and into the simulation used for obtaining results.

4.3.1 Outline of DMEXCLP

In its original form, the DMEXCLP method moves a vehicle when it becomes idle after finishing service of a patient: the algorithm relocates this ambulance to an appropriate waiting site within the region. The sole objective of the DMEXCLP method 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 MEXCLP, formulated as an ILP by Daskin (1983). This problem was designed to compute an optimal static distribution of vehicles over waiting sites, by calculating the (probabilistic) coverage of the region.

MEXCLP defines the coverage of a region in terms of a so-called *busy fraction* p. 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 $1 - p^k$. If we let d_i be the demand at node i, the expected covered demand of this vertex is $E_k = d_i(1 - p^k)$. The MEXCLP positions the ambulances in such a way that the total maximum expected covered demand, summed over all demand points, is obtained.

DMEXCLP, or dynamic MEXCLP, reuses this definition of probabilistic 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 an 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 $E_k - E_{k-1} = d_i(1-p)p^{k-1}$. The waiting site that results in the largest marginal coverage over the entire region can be computed by

$$\arg\max_{w \in W} \sum_{i \in V} d_i (1-p) p^{k(i,w,n_1,\ldots,n_{|W|})-1} \cdot \mathbb{1}_{\{\tau_{wi}^{(1)} \leq T\}}, \tag{4.1}$$

where

$$k(i, w, n_1, \ldots, n_{|W|}) = \sum_{j=1}^{|W|} n_j \mathbb{1}_{\{\tau_{ji}^{(1)} \leq T\}} + \mathbb{1}_{\{\tau_{wi}^{(1)} \leq T\}} \tag{4.2}$$

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 at node i.

4.3.2 Comparison to Penalty Heuristic

In this section, we highlight differences between the penalty heuristic, presented by Van Barneveld et al. (2016a), and the DMEXCLP method as published by Jagtenberg et al. (2015). As mentioned above, similarities exist between both methods. Both papers differ on the following five major aspects:

Coverage: The penalty heuristic 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 $p = 0$. That is, MEXCLP may be interpreted as a generalization of single coverage.

Number of decision moments: As we have seen, Jagtenberg et al. (2015) propose 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. Although the authors state that the method can be easily adjusted for usage at other types of decision moments, it is not clear which ambulance should be relocated. In addition, Van Barneveld et al. (2016a) allow a relocation to be executed immediately after the dispatch of an ambulance to an incident.

Busy ambulances: As mentioned in Section 4.3.1, busy ambulances do not contribute to the coverage in the work by Jagtenberg et al. (2015). In contrast, Van Barneveld et al. (2016a) consider an ambulance as dispatchable if its transfer time at a hospital exceeds a predefined standard Δ. 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.

Chain relocations: Jagtenberg et al. (2015) do not consider chain relocations, in contrast to Van Barneveld et al. (2016a). 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. 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.

Objective: The focus is on minimization of late arrivals solely in Jagtenberg et al. (2015): one incurs a penalty of 1 each time the response time to an incident exceeds T. In contrast, this objective can be generalized by the definition of a *penalty function*. 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 modeled by the penalty function $\Phi(t) = \mathbb{1}_{\{t>T\}}$. However, Van Barneveld et al. (2016a) question the dichotomous nature of this objective, as medical outcomes are completely ignored. 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 Equation (3.2) and displayed in Figure 3.5 for $\alpha = 0.008$, $\beta = 5$, and $T = 720$, in the previous chapter.

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 by Jagtenberg et al. (2015) are generalized by Van Barneveld et al. (2016a). In the next section, we explain how we modify the original DMEXCLP method by incorporating a number of features related to the five aspects described above.

4.3.3 Modification of DMEXCLP

In this section we incorporate the abovementioned features into the DMEXCLP method. Moreover, we introduce a new feature, neither considered by Jagtenberg et al. (2015) nor by Van Barneveld et al. (2016a): a bound on the relocation time. One by one, we discuss the fusion of the DMEXCLP framework with the features of the penalty heuristic.

Decision moments

At the added decision moment – when a vehicle is dispatched – it is not clear from which waiting site an ambulance should be relocated. This is easily computed, however, by the following modification of Equation (4.1):

$$\begin{aligned}\underset{(w_1,w_2)\in W^2:n_{w_1}>0}{\arg\max} &\sum_{i\in V} d_i(1-p)p^{k(i,w_2,n_1,\ldots,n_{|W|})-1}\cdot \mathbb{1}_{\{\tau^{(1)}_{w_2 i}\leq T\}} \\ &- \sum_{i\in V} d_i(1-p)p^{k(i,w_1,n_1,\ldots,n_{|W|})-1}\cdot \mathbb{1}_{\{\tau^{(1)}_{w_1 i}\leq T\}},\end{aligned} \tag{4.3}$$

in which w_1 and w_2 denote the old origin and new destination of the vehicle to relocate, and $k(i, w, n_1, \ldots, n_{|W|})$ is as defined in Equation (4.2). In Equation (4.3) 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 (4.3) can be computed by brute force.

Busy ambulances

Although Van Barneveld et al. (2016a) allow transfer time interruptions if the transfer at a hospital has lasted for at least Δ seconds, we do not in this chapter for reasons stated above. However, we do take into account these busy ambulances in a different way. To this end, we assume that the hospital transfer time follows a probability distribution. Let

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

denote the expected remaining transfer time of ambulance a if its transfer already lasted for $\tau(a)$ units of 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 (4.2) as follows:

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

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 DMEX-CLP method, but the calculation of this chain is a subsequent step: the expression of Equation (4.1) is not modified. As mentioned in the previous chapter, the linear bottleneck assignment problem is considered for this computation. In Chapter 3, we concluded that the benefit to the patient-based performance of a chain relocation consisting of more than two links is very small. We observed a large performance gain, however, if chains consisting of exactly two links are used, instead of

a regime in which no chain relocations are allowed. 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 Chapter 3, we follow this conclusion and restrict that at most two ambulances may take part in a chain relocation.

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 necessarily the best, configuration that can be attained quickly. After all, driving to a waiting site, although best classified by DMEXCLP, may take long. 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 (4.1) as follows:

$$\underset{w \in W : \tau^{(2)}_{cw} \leq B}{\arg\max} \sum_{i \in V} d_i (1-p) p^{k(i,w,n_1,\ldots,n_{|W|})-1} \cdot \mathbb{1}_{\{\tau^{(1)}_{wi} \leq T\}}. \tag{4.4}$$

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 4.4.6 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 (3.2) and Figure 3.5, 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 probabilistic coverage, this contribution levels off the farther away an ambulance: with probability $1-p$ the closest one to i is available and responds to an incident occuring there, inducing a penalty of $\Phi(\tau^{(1)}_{ji})$ if the closest ambulance to i is located at waiting site j. With probability $(1-p)p$ the second closest responds, generating $\Phi(\tau^{(1)}_{j'i})$ 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

$$\arg\min_{w \in W} \sum_{i \in V} \sum_{j \in W} \sum_{l=1}^{n} d_i (1-p) p^{l-1} \Phi(\tau_{ji}^{(1)}) z_{(c(w,n_1,\ldots,n_{|W|}),i,j,l)}. \tag{4.5}$$

Note that Equation (4.5) is a minimization problem, as penalty functions are non-decreasing in the response time.

4.4 Numerical Results

In this section we show computational results on the performance regarding the in- and exclusion of the described features in the algorithms explained in Section 4.3. Results are obtained by trace-driven simulations using historical data for two EMS regions in the Netherlands.

4.4.1 Experimental Setup

We base our computations on two different on the EMS regions of Flevoland and Amsterdam. We refer to Section 3.4.1 and Section 3.4.2 for an extensive description of these regions. Unlike in the previous chapter, we assume that ambulances may idle at any of the red or green nodes in Figure 3.3 (9 waiting sites) and Figure 3.4 (12 waiting sites), respectively.

Historical data on emergency requests in the year 2011 was used for our analyses. 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 include the following incident related information:

- Time of occurence, i.e., the time of the emergency call;
- Location of occurence (4-digit 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. 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 simulations, 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, which provided us tables containing travel times between each pair of postal codes in the regions of consideration. We refer to Section 3.4.3 and Kommer and Zwakhals (2008) 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 factor of $\frac{10}{9}$. Moreover, we use the framework described in Section 3.4.3 for the computation of the travel routes, in order to keep track of the actual location of a moving vehicle. 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.
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 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 and Amsterdam, respectively. 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.

4.4.2 Original DMEXCLP method

In this section, we report results for both regions of interest, Flevoland and Amsterdam, of the original DMEXCLP method, as explained in Section 4.3.1. 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 4.2.

A large performance improvement in terms of late arrivals can be observed in Table 4.2 for the Amsterdam region. This quantity decreases 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 (Jagtenberg et al. (2015), 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 policy are much smaller for Flevoland. This was already foreseen by Jagtenberg et al. (2015), 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 4.2 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 4.2, the mean single coverage is an optimistic ap-

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

proximation 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 coverage of multiple ambulances is very small: the distances between the 6 large towns generally exceed the time threshold. 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.

4.4.3 Decision Moments

As explained in Section 4.3.3, we allow the dispatcher to make an ambulance relocation decision if the number of available ambulances decreases, just after the dispatch. As a consequence, the number of opportunities to steer the EMS system is multiplied by two. Results are displayed in Table 4.3. In this table and the forthcoming ones, the default policy is the DMEXCLP policy explained in Section 4.3.1, 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 4.2 underlines.

Performance Indicators	Flevoland		Amsterdam	
	Default	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 4.3: Simulation results for Flevoland and Amsterdam, based on 7,632 and 41,966 incidents in 2011, with 10 and 18 ambulances, respectively.

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 decreases 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 relocation method (which was 2.1%). For Amsterdam, the default method 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 the following: since Flevoland is a rural region, an ambulance traveling between two waiting sites provides no or very little coverage. After all, few people live in the areas between the cities, c.f., Figure 3.3. In contrast, a large part of the Amsterdam region is urban, c.f., Figure 3.4. 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 Chapter 3, 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.

4.4.4 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 4.3.3, 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, as we have learned from discussions with dispatchers and management. 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, c.f., Carter et al. (2015).

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 Maxwell et al. (2010). 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 4.4, we list 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 observed, 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 4.3.3

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 4.4: Simulation results for Flevoland and Amsterdam for different hospital regimes, based on 7,632 and 41,966 incidents in 2011, with 10 and 18 ambulances, respectively.

does not significantly influence the performance.

Alternatively, the Weibull distribution used for the estimation of the transfer time may perhaps be a poor approximation. To test whether this indeed is 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 4.3.3), 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 due to 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, or at least, not as much as the default algorithm would have. 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.

4.4.5 Chain Relocations

In Chaper 3, 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 two links instead of one, but using 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 4.5.

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 4.5: Simulation results regarding chain relocations, for Flevoland and Amsterdam, based on 7,632 and 41,966 incidents in 2011, with 10 and 18 ambulances, respectively.

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 4.5. This contradicts the findings of Van Barneveld et al. (2016a). This is probably due to the fact that they allow extra decision moments, as considered in Sections 4.3.3 and 4.4.3. In Section 4.4.7, we 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.

4.4.6 Relocation Time Bounds

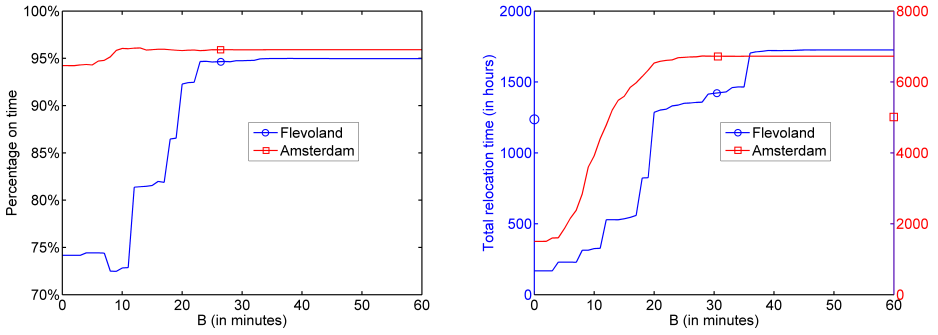
As explained in Section 4.3.3, we impose different bounds on the relocation time of an ambulance. This bound is given by the variable B. If there is no waiting site that can be reached within B minutes, the ambulance travels to the nearest waiting site. For $B = 0$, the obtained policy is equivalent to this 'nearest base'-policy. In Figures 4.1a and 4.1b we show results on the most important patient- and crew-related performance indicators: percentage on time and total relocation time, as a function of B. In Tables 4.6 and 4.7, results on all performance indicators are displayed for $B = 0, 10, 20, 30$ minutes.

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 4.6: Simulation results for Flevoland based on 7,632 incidents in 2011, with 10 ambulances. Results on relocation bounds 0, 10, 20, and 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 4.7: Simulation results for Amsterdam based on 41,966 incidents in 2011, with 18 ambulances. Results on relocation bounds 0, 10, 20, and 30 minutes are displayed.



(A) (B)

FIGURE 4.1: Percentage on time and total relocation time as a function of B.

In Figure 4.1a 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, 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 the vast majority of the ambulances becomes available there, relocations to waiting sites 3, 4, 5 and 6 (in the enumeration of Figure 3.3) 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 4.6 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 4.1b 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

Performance Indicators	Flevoland			
Combination:	1	2	3	4
Percentage on time	96.24%	96.24%	96.27%	94.22%
Lower Bound 95%-CI	95.79%	95.77%	95.82%	93.64%
Upper Bound 95%-CI	96.69%	96.71%	96.71%	94.80%
Mean response time	292 s	292 s	292 s	288 s
Number of relocations	24,747	24,408	23,481	22,047
Average relocation time	774 s	766 s	766 s	599 s
Total relocation time	5,318 h	5,196 h	4,997 h	3,671 h
Mean single coverage	97.34%	97.34%	97.34%	97.43%
Mean MEXCLP coverage	94.61%	94.60%	94.58%	93.24%

Table 4.8: Simulation results for different combinations for Flevoland, based on 7,632 incidents in 2011, with 10 ambulances.

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 4.7: the coverage levels corresponding 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 $B = 30$.

4.4.7 Combinations

In this section, we combine different promising features and test the resulting methods for both regions. Moreover, we compare the performance to the penalty heuristic as presented in Chapter 3. 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 4.4.4.
3. Similar to 2, but now we have perfect information about the transfer times.
4. Penalty heuristic (see Chapter 3).

The results are displayed in Tables 4.8 and 4.9. Although allowing chain relocations initially did not result in better performance regarding the percentage on time criterion, as observed in Table 4.5, it is a valuable addition if it is combined with the allowance of extra decision moments, for both regions. If we compare Table 4.3, which shows the best performance concerning this criterion up to now,

Performance Indicators	Amsterdam			
Combination:	1	2	3	4
Percentage on time	97.23%	97.21%	97.26%	97.10%
Lower Bound 95%-CI	96.82%	96.77%	96.84%	96.68%
Upper Bound 95%-CI	97.64%	97.66%	97.67%	97.51%
Mean response time	303 s	302 s	302 s	283 s
Number of relocations	132,918	132,530	127,467	129,988
Average relocation time	440 s	439 s	424 s	457 s
Total relocation time	16,258 h	16,172 h	15,026 h	16,486 h
Mean single coverage	99.12%	99.11%	99.13%	99.34%
Mean MEXCLP coverage	96.79%	96.78%	96.75%	95.62%

TABLE 4.9: Simulation results for different combinations for Amsterdam, based on 41,966 incidents in 2011, with 18 ambulances.

with the first columns in Tables 4.8 and 4.9, we see that performance improvements of 0.64 and 0.88 percentage points are achieved for Flevoland and Amsterdam, respectively. That is, the number of late arrivals decreases with 14.55% and 24.11%. This behaviour is probably explained by the following observation: 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 4.8 and 4.9, we barely see any differences in patient-based performance. This underlines the observations in Section 4.4.4. Results on crew-based performance are similar to those obtained in Section 4.4.4 as well.

The DMEXCLP method with its features is quite consistent in its 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. 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 (3.2). One can observe in Tables 4.8 and 4.9 that minimizing the average response time is included in this penalty function

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

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 Jagtenberg et al. (2015) 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.

4.4.8 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 (3.2), which is a compromise between minimizing late arrivals and minimizing average response times.

Results are displayed in Table 4.10. One might 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 Chapter 2.

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.

4.5 Concluding Remarks

In this chapter, we studied the implementation of several aspects and features presented by Van Barneveld et al. (2016a) in the dynamic relocation method proposed by Jagtenberg et al. (2015). Next, we draw conclusions and we make recommendations.

Based on the results in Table 4.10, 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 4.3. For urban regions, we consider this an optional addition, 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 4.5 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, c.f., Tables 4.8 and 4.9.

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 relocation strategy somewhat harder to implement (and explain), while it does not benefit the patients. An exception to this rule could be, when an ambulance service providers struggles with the workload of EMS crews: in that case, including the ambulances at hospital could be worthwhile, because it slightly reduces the relocation times (as seen in Table 4.4).