European Journal of Operational Research 252 (2016) 257–269

Contents lists available at ScienceDirect

European Journal of Operational Research

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

Innovative Applications of O.R.

The effect of ambulance relocations on the performance of ambulance service providers

T. C. van Barneveld[a,b,*], S. Bhulai[b,a], R. D. van der Mei[a,b]

[a] Centrum Wiskunde en Informatica, Science Park 123, 1098XG Amsterdam, The Netherlands
[b] Vrije Universiteit Amsterdam, De Boelelaan 1081, 1081HV Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Received 20 February 2015
Accepted 14 December 2015
Available online 22 December 2015

Keywords:
OR in health services
Multiple criteria analysis
Dynamic smbulance management
Dynamic relocation
Response times

ABSTRACT

Dynamic Ambulance Management (DAM) is generally believed to provide means to enhance the response-time performance of emergency medical service providers. The implementation of DAM algorithms leads to additional movements of ambulance vehicles compared to the reactive paradigm, where ambulances depart from the base station when an incident is reported. In practice, proactive relocations are only acceptable when the number of additional movements is limited. Motivated by this trade-off, we study the effect of the number of relocations on the response-time performance. We formulate the relocations from one configuration to a target configuration by the Linear Bottleneck Assignment Problem, so as to provide the quickest way to transition to the target configuration. Moreover, the performance is measured by a general penalty function, assigning to each possible response time a certain penalty. We extensively validate the effectiveness of relocations for a wide variety of realistic scenarios, including a day and night scenario in a critically and realistically loaded system. The results consistently show that already a small number of relocations lead to near-optimal performance, which is important for the implementation of DAM algorithms in practice.

1. Introduction

In emergency situations, the location of ambulances has a huge impact on the response time to an incident, i.e., the total time between an incoming emergency call and the moment that an ambulance arrives at the emergency scene. The evaluation of ambulance services providers, judged by the authorities, heavily relies on their performance regarding these response times. For instance, in The Netherlands, the response time of an ambulance may not exceed 15 minutes in 95 percent of the highest priority emergency cases. To realize short response times, it is crucial to plan ambulance services well. This encompasses a variety of planning problems at the strategic, tactical, and operational level. At the strategic level, the locations of the ambulance base stations are determined. Then, at the tactical level, the number of ambulances and thus crews per base station is specified. At the operational level, real-time dispatching of ambulances to incidents and real-time relocation of ambulances is considered.

In this paper, we focus on the last part of the operational level: the relocation of ambulances. Ambulance vehicles are relocated in real-time, using dynamic and proactive relocation strategies, in order to achieve shorter response times to incidents. These relocation decisions are typically made when an event happens, e.g., when an ambulance is dispatched or when an ambulance is newly free after the service of a patient. However, whether relocations are allowed, and if so, to which locations, depend on regulatory rules. For instance, in Vienna, Austria, moving around ambulances unoccupied by a patient is not allowed, cf. Schmid (2012) as opposed to Edmonton, Canada, cf. Alanis, Ingolfsson, and Kolfal (2013). Moreover, the number of locations at which an ambulance is allowed to park up differs per country. This number can exceed the number of ambulances, as in Montreal, Canada, cf. Gendreau, Laporte, and Semet (2006). Many of these waiting sites are just street corners or different hot spots. In contrast, in The Netherlands, ambulances always must return to a base station, cf. Jagtenberg, Bhulai, and van der Mei (2015). This is a building with several facilities where the ambulance crew can spend its shift when idle. Another difference between countries is the average hospital transfer time. In North America this time can be very large, cf. Carter et al. (2015), as opposed to The Netherlands where the transfer time is usually short.

We consider the Dutch setting in this paper: short transfer times and the dispatcher is allowed to relocate ambulances unoccupied from base station to another one, but the number of

* Corresponding author at: Centrum Wiskunde en Informatica, Science Park 123, 1098XG Amsterdam, The Netherlands. Tel.: +31 205934010.
E-mail address: t.c.van.barneveld@cwi.nl (T.C. van Barneveld).

http://dx.doi.org/10.1016/j.ejor.2015.12.022
0377-2217/

locations on which an ambulance can idle, i.e., base stations, is rather small.

Ambulance relocations are not popular among ambulance crews, especially when the crew is idle at a base station and it is relocated to a different one. Instead, they prefer to spend their shift at a base station and not on the road. To keep the personnel motivated, the number of relocations they have to perform is not allowed to increase excessively. If the ambulance crews spend too much time on the road, the ambulance service provider probably will be condemned by an Occupational Safety and Health organization. Furthermore, costs for the ambulance service provider are associated with each relocation. Some ambulance service providers namely have the policy, especially at night, that the salary of the ambulance crew partly depends on their busy time in which their relocation time is included. Therefore, decision makers must make a consideration between the number of ambulance relocations and the effect of these relocations on the performance of the ambulance service provider.

As an alternative, one could also consider the relocation times. Especially if the above-mentioned payment structure is used, it can be cheaper to minimize these relocation times. However, in this paper, we treat the crew's perspective as our major critical success factor, instead of the financial aspect related to relocations. The number of relocations is a good measure for the crew's perspective, since crews in general prefer to perform one long relocation rather than several short ones. Of course, there is also a trade-off between number of relocations and relocation times. We will pay attention to this trade-off as well.

The relationship between performance and the number of ambulance relocations is complex. The consequences of moving an ambulance to a different base station are not known a priori, due to uncertainty that plays an important role in the process. It is usually not the case that 'more' is 'better', i.e., the more relocations are made, the better the performance of the ambulance service provider. But even if this was the case, there is still a trade-off: would one carry out extra ambulance relocations for only a small gain in performance? Opinions of different ambulance providers differ on this question and it is hard to set a standard concerning the execution of relocations. Therefore, useful insights about the relationship between performance and the number of ambulance relocations are desirable.

1.1. Related work

As stated before, the planning of ambulance services falls apart in three levels. Comprehensive studies of ambulance location and relocation models are done by Brotcorne, Laporte, and Semet (2003) and Li, Zhao, Zhu, and Wyatt (2011). In these papers several deterministic, probabilistic, and dynamic models and their solution procedures are reviewed. Another study on ambulance facility location problems is performed by Owen and Daskin (1998). The operational level falls apart in dispatching and relocation of ambulances. A dispatching algorithm based on the preparedness concept explained by Andersson and Värbrand (2007), is proposed by Lee (2011). Another dispatch method, based on the maximal covering location problem developed by Church and ReVelle (1974), is presented by Lim, Mamat, and Bräunl (2011) and it is shown by simulation that response times to urgent calls can be reduced.

A common way to solve the dynamic ambulance relocation problem is the offline approach: redeployment decisions are precomputed for different states of the system. For instance, compliance tables are computed, which prescribe desired locations for idle ambulances by Gendreau et al. (2006). With this purpose, the Maximal Expected Coverage Relocation Problem (MECRP) is proposed and solved, by formulating this problem as an integer linear program. It is stated by Maleki, Majlesinasab, and Sepehri (2014), that computing compliance tables is just the first part of the computation of relocation decisions. The second part involves the actual assignment of ambulances to base stations. Therefore, the Generalized Ambulance Assignment Problem (GAAP) and Generalized Ambulance Bottleneck Assignment Problem (GABAP) are proposed. Compliance tables are the subject of Alanis et al. (2013) as well: a two-dimensional Markov chain is proposed and analyzed to obtain optimal compliance tables. A two-stage stochastic optimization model for the ambulance redeployment problem that minimizes the number of relocations while maintaining an acceptable service level is presented by Naoum-Sawaya and Elhedhli (2013).

In addition to the offline approach, a large part of the ambulance literature focuses on the online computation of relocation decisions. Whenever an event occurs, e.g., an ambulance becomes available again, the dispatcher has the opportunity to control the system. Based on the information of the state of the system, one computes a relocation decision. Such a relocation decision needs to be obtained in a very short time, and thus is the main focus of this literature on heuristics. For instance, a heuristic called the Dynamic Maximum Expected Coverage Location Problem (DMEXCLP) is proposed by Jagtenberg et al. (2015). This problem, based on the MEXCLP presented by Daskin (1983), computes a new location for an ambulance that just finished service of a patient. Moreover, a parallel tabu search heuristic is used for the real-time redeployment of ambulances by Gendreau, Laporte, and Semet (2001). Andersson and Värbrand (2007) use the notion of preparedness. This preparedness is a measure for the ability to serve potential patients now and in the future. Moreover, a dynamic relocation model named DYNAROC and a heuristic to solve this model is presented. In addition, some papers use approximate dynamic programming for determining relocation strategies, for instance, Maxwell, Restrepo, Henderson, and Topaloglu (2010); Maxwell, Henderson, and Topaloglu (2013) and Schmid (2012). Relocation decisions are made at the time of call arrivals and when an ambulance becomes available again by Maxwell (2011). In this work, it is shown that making relocation decisions at such times is equivalent to the usage of a nested compliance table policy. At last, a comprehensive study on both online and offline redeployment is executed by Zhang (2012).

1.2. Our contribution

In this paper, we study the relationship between number of ambulance relocations and the performance of the ambulance service provider. Theretofore, we present an ambulance redeployment model, in which we are able to incorporate different performance criteria. We use a heuristic method that computes an action concerning the relocation of ambulances in such a way that the expected performance is maximized. This computation is done at decision moments: the time of occurrence of a new incident or the time of the idle report of an ambulance. We use a heuristic policy instead of the optimal one because computation of the optimal policy is very complex, if not impossible. Besides, even if it was possible to compute, the optimal policy is probably a complex one: it is not easy to understand and to execute by the dispatcher. Instead, we use a heuristic method that is not too far-fetched, while it is highly likely that this heuristic policy contains the same characteristics as the optimal one.

This paper differs from the mainstream literature in two respects

1. Most of the papers in the literature, e.g., Jagtenberg et al. (2015), assume that the computed action is always carried out. However, it may be the case that the expected gain in performance by taking this action is very small and that this benefit does not outweigh the disadvantages regarding the

number of additional ambulance relocations to achieve this gain. Therefore, we use the heuristic method to determine whether the redeployment action is really necessary, and we show results on several quantifications of 'really necessary'.

2. Another important difference between the mainstream literature and this paper is the way in which a redeployment action is carried out. We compute, using the heuristic method, a location that serves as origin, from which an ambulance needs to depart, and a base station serving as destination. However, it is not necessarily one particular ambulance that has to move from origin to destination, as assumed in most papers. Instead, we can use other idle ambulances, either driving or at a base station, in this relocation process in order to decrease the time required to attain the new ambulance configuration. However, this comes at the expense of extra relocations. We put restrictions on the number of other ambulances that may be relocated, and we show consequences on the resulting performance to obtain useful insights in the relationship between number of relocations and performance.

2. Model description

To investigate the above-mentioned relationship, we introduce the ambulance redeployment model described in this section. We consider the Dutch setting as explained in Section 1 and we model the region of interest as a directed graph. Geographical regions, e.g., neighborhoods, postal codes or streets, are represented by nodes. Moreover, arcs in this graph are weighted and the length of an arc represents the driving time (in seconds) between the nodes. These driving times are derived from a driving time table R, estimated beforehand and thus assumed to be given. Hospitals with an emergency department, to which patients can be transported are present in some nodes, i.e., geographical regions. A certain *demand probability* is associated with each node as well. This demand probability is defined as the probability that an incoming incident will occur in that specific node. We denote these *demand probabilities* by $p = \big(p(1), p(2), \ldots, p(N)\big)$, where N is the number of nodes.

2.1. System dynamics

The dynamics of our system closely mimic realistic situations. Incidents occur according to a Poisson point process, as by Matteson, McLean, Woodard, and Henderson (2011) and Maxwell et al. (2010). We assume that all incoming incidents are of the highest urgency. As a consequence, there is one universal *maximum allowed response time*. This assumption is justified by the fact that ambulance service providers are mostly judged on their performance regarding the highest priority incidents. This maximum allowed response time is set by the government or by the ambulance provider itself. Many ambulance service providers use the percentage of highest urgency incidents reached by an ambulance within this maximum allowed response time as their performance criterion. However, by doing this, there is no difference between a response time that is slightly below the maximum allowed one, and one that is really short. Something similar holds for the opposite case: a response time that is slightly above the maximum allowed response time and one that is really long are equally judged. However, it does matter for the patient. It is stated by Erkut, Ingolfsson, and Erdogan (2008), that the black-and-white nature of the coverage concept is an important limitation, and standard coverage models should not be used. Therefore, other performance measures, based on survival probabilities, are considered by Erkut et al. (2008). In this paper, we use *penalty functions* to model general performance measures regarding response times. It is possible to differentiate between different response times by using these penalty functions.

A penalty function f is a function of the response time solely, with domain $\mathbb{R}_{\geq 0}$. It can be used to incorporate different performance measures, such as minimizing the number of incidents for which the maximum allowed response time is exceeded, minimizing the average response time or measures related to survival probabilities, as studied by Erkut et al. (2008). Note that some performance measures, especially those related to average response times, do not use the maximum allowed response time. The amount of penalty generated by an incident solely depends on the response time to this incident. Hence, penalty functions are non-decreasing functions. An example of a penalty function is given in Section 4.1.

Note that the response time actually consists of the time between the emergency request comes in and the arrival of the ambulance at the emergency scene. We refer to Schmid (2012) for a graphical representation of this process. However, only the travel time of the ambulance to the incident is of interest to this paper. Hence, we assume both that the dispatch and turnout time, which is the time between the mission is received by the crew and the crew leaves the base station, are deterministic. We subtract these two times from the maximum allowed response time to obtain a *maximum allowed travel time*. However, turnout times are typically smaller when ambulances are already on the road. Therefore, the use of a maximum allowed travel time is a simplification of reality.

Since all the incidents are of the highest urgency, the closest[1] one is dispatched. This could be an idle ambulance at a base station or a driving idle ambulance. If none of the idle ambulances can reach the incident within the maximum allowed travel time, we have the possibility to interrupt an ambulance transferring a patient at a hospital. This may benefit the performance, as we can send an idle ambulance in the neighborhood away. This move-away-from-hospital avoids overlapping coverage around the hospital, cf. Zhang (2012).

However, we only preempt if this ambulance is already more than a target time T busy with the transfer of a patient. That is, T can be interpreted as the minimum time that an ambulance can be busy at the hospital without the possibility that it is preempted. The reason why this preemption is allowed is twofold. First, it often occurs that the ambulance crew already finished transferring the patient, but has not informed the dispatcher yet. Second, even if it may take longer than the target time for transferring the patient for whatever reason, there is enough personnel at the hospital that can take care of the patient, e.g., for the transport of the patient to the right room within the hospital. This kind of tasks does not necessarily have to be done by the ambulance crew. Hence, we consider an ambulance employable for a new incident, if it is already more than this target time busy with the transfer of a patient. Whether this interruption is allowed may differ per ambulance service provider, but this is the case for the considered service provider in the numerical study in Section 4.

Once an ambulance is dispatched to an incident, we assume that it immediately starts driving, since the turn-out time is part of the pre-travel time we subtracted. We distinguish the following five ambulance phases:

0. Phase 0-ambulances are the ambulances not currently involved in the service of a patient. They are either at a base station or executing a relocation.
1. An ambulance traveling to the emergency scene is in phase 1. Its travel time to the incident is given by the driving time table R.

[1] By 'closest', we mean closest in time, here and in the remainder. Note that this ambulance is not necessarily the closest one in space as well.

2. Ambulances busy at the emergency scene. The ambulance crew executes the treatment of the patient, which may consist of different kinds of medical assistance. During this treatment, the ambulance crew decides whether the patient needs to be transported to a hospital. If a patient does not need transportation, the ambulance becomes idle. Then, the dispatcher has to make a decision to which base location that ambulance should be send.
3. If transportation is needed, the ambulance enters phase 3 and the transportation of the patient is started. We assume that a patient is always transported to the closest hospital. This transportation time is given by the driving time table.
4. At the hospital, the ambulance is busy for a while transferring the patient. After that, the ambulance becomes idle at the hospital and a decision on where to send it to needs to be made.

2.2. Control of the system

The above-mentioned decisions on dispatching are assumed to be fixed, i.e., they serve as a rule to the dispatcher. However, the dispatcher has some freedom in the way he/she can control the system by making relocation decisions. We allow the dispatcher to make these decisions at the following moments:

1. when an ambulance is dispatched to an incoming incident, and
2. when an ambulance enters phase 0 again, either from phase 2 or phase 4.

We refer to these moments as decision moments of the *first* and *second* type, respectively. At both types of decision moments, the dispatcher is allowed to perform a so-called *ambulance motion*: a change in ambulance configuration in which *at most* one pair of base stations is affected. An ambulance motion has an *origin* and a *destination*. In the ambulance configuration the number of ambulances at the origin and destination is decreased and increased by 1, respectively. At a decision moment of the second type, the origin is given: this is the location of the ambulance that just finished service. In contrast, each base station with at least one ambulance in the ambulance configuration can serve as origin at decision moments of the first type.

The obvious way to execute an ambulance motion is to select an ambulance from the origin and to relocate it to the destination of the ambulance motion. However, the origin and destination are not necessarily close to each other and thus the travel time between them may be long. Such long trips are not desirable, since the new ambulance configuration must be attained as soon as possible.

A possibility to avoid long trips is the usage of multiple phase 0-ambulances, either driving or at a base location, in a motion. Instead of moving just one ambulance, it could be beneficial to break up the ambulance motion in two or more separate *ambulance relocations* to ensure that the new ambulance configuration is attained earlier. We refer to Fig. 1 for an example.

Example 1. In this small illustration, the ambulance motion is (1, 5) and there are ambulances in 1 and 2. In addition, one ambulance is traveling from 4 to 3, and it is currently in node 6. The obvious way would be to relocate the ambulance from 1 to 5. However, it takes 1,548 seconds before the motion is completely performed (Fig. 1a). If one uses the ambulance at 2, this time can be reduced to 1,402 seconds, at the expense of one extra relocation (Fig. 1b). In addition, if *redirection* is allowed, which we assume, one can use the driving ambulance to decrease the time in which the new ambulance configuration is attained to 975 seconds (Fig. 1c).

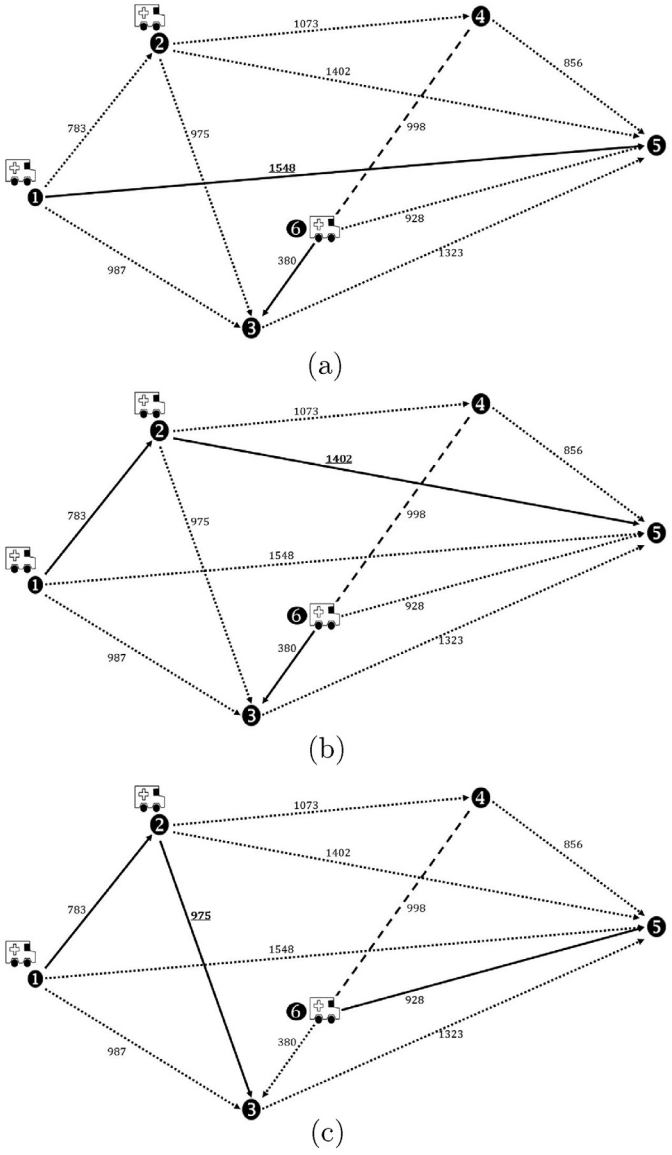

Fig. 1. Illustration of the usage of multiple ambulances per motion. The motion is (1, 5), and full arcs denote the way in which ambulances are relocated. The numbers next to the arcs are the driving times in seconds.

We assume that the turnout time of a relocated ambulance is zero, and the decision is made instantaneously after the decision moment. At a decision moment of the second type, the ambulance that just finished service needs to be relocated to a base station. If it is relocated to the closest one, this does not count as a relocation. After all, this does not inconvenience the ambulance personnel as they can idle as quickly as possible to recover from the patient-related work they just carried out. Moreover, an ambulance redirection, as in Fig. 1c, does not count as a relocation, as the crew is already en route. As Fig. 1 shows, there is a trade-off between the number of relocated ambulances and the time it takes to attain the new ambulance configuration.

2.3. Problem description

At decision moments the dispatcher usually faces three problems:

1. Is an ambulance motion necessary? At decision moments of the first type, it may be the case that the resulting

configuration after the dispatch is still satisfactory, in terms of expected response times to future incidents. That is, it may not be beneficial to execute a motion by reasons mentioned in Section 1. This question does not arise at decision moments of the second type, since the dispatcher is always required to perform an ambulance motion for the ambulance that just became idle.
2. Which ambulance motion should be executed? The dispatcher must select two base locations: one serving as origin, one as destination. A heuristic method for calculating the best ambulance motion is described in Section 3.
3. How to execute this ambulance motion? As stated in the previous section, the dispatcher has multiple options to ensure that the new configuration is attained by performing a sequence of ambulance relocations.

In the next section, we will present a heuristic method concerning these three problems.

3. Heuristic method

For the evaluation of the usefulness of ambulance motions and relocations, we present a heuristic that can easily handle several types of restrictions on the decisions of the dispatcher. First, we describe the heuristic method. Then, we will provide a more detailed explanation regarding the incorporation of these constraints.

The key idea of this method is as follows: at a decision moment, the dispatcher observes the current state of the system. Given this information, the dispatcher executes the motion that minimizes the *unpreparedness*. This is a measure regarding the configuration of ambulances. We explain this concept extensively in the next section.

3.1. Unpreparedness

The concept of unpreparedness plays an important role in the heuristic method. This term can have several interpretations, depending on the use of penalty function. For instance, if a linear penalty function is used, one focuses on minimization of the average response time. Penalty and response time are equivalent then and the unpreparedness has the interpretation of being an approximation of the expected time required to respond to the next emergency request, for a given ambulance configuration. That is, the heuristic method tries to minimize the expected response time to the next call. However, if one uses a general penalty function, this interpretation generalizes to being an approximation of the expected penalty the next emergency request generates, for a given configuration. We proceed with a formal definition of unpreparedness of an ambulance configuration.

Let s be the current state of the system: the current location or destination of ambulances and the phases they are in. We define the set of ambulances by $\mathcal{A}$, where $A := |\mathcal{A}|$. Moreover, we define $\mathcal{A}^k(s)$ as the set of ambulances in phase k if the state of the system is s.

To define unpreparedness formally, we need some additional definitions. Consider node i, $1 \le i \le N$. Let $des(j, s)$ denote the *destination* of ambulance j if the state of the system is s, and R is the driving time matrix. We define

$$r_i^0(s) = \min_{j \in \mathcal{A}^0(s)} R\big(des(j,s), i\big),$$

as the driving time between the destination of the closest phase 0-ambulance and node i. The destination equals the current location of the ambulance if the ambulance is not on the road. The reason that we use the destination instead of the current location, is twofold

1. If we had used the actual location of the driving phase 0-ambulances, one might think that one can quickly respond to an incident in the area in which the ambulance is currently driving. However, we are uncertain about the time of the next incident. If the next incident happens in that particular area after some time, it may take long to respond to this incident, since the ambulance has left that area.
2. In addition, a relocated ambulance may still be far away from its destination. Hence, the area around this destination will be classified as vulnerable if one uses the current location of the ambulance. As a consequence, the method may decide to send another ambulance to that area. This is probably useless, since an ambulance is moving towards that area already.

As explained in Section 2.1, phase 4-ambulances can respond to incoming incidents if their service has already lasted for at least T seconds. Similar to $r_i^0(s)$, we define $r_i^4(s)$ to be the expected time until the closest phase 4-ambulance is able to be present at i:

$$r_i^4(s) = \min_{j \in \mathcal{A}^4(s)} \left\{ \left[T - \tau(j,s) \right]^+ + R\big(loc(j,s), i\big) \right\},$$

where $\tau(j, s)$ is the service time that has past already, $[\cdot]^+$ denotes the positive part and $loc(j, s)$ denotes the location of ambulance j if the state of the system is s.

Let $p(i)$ denote the demand probability: the probability that an incoming incident will occur in node i. Now we have all the ingredients to define the unpreparedness of the configuration of ambulances, denoted by $U(s)$ if the current state of the system is s:

$$U(s) := \sum_{i=1}^{N} f\Big(\min\left\{ r_i^0(s), r_i^4(s) \right\} \Big) p(i),$$

where f is the penalty function.

Example 2. Consider the system in Fig. 1a. Assume each node has the same demand probability: $p(i) = \frac{1}{5}, i = 1, \ldots, 5$. Moreover, suppose we use the penalty function corresponding to the minimization of the average response time: $f(t) = t, t \ge 0$. That is, the heuristic method tries to minimize the expected response time to the next call. Note that there are no phase 4-ambulances, so $r_i^4(s) = 0, i = 1, \ldots, 5$. We compute $r_1^0(s) = r_2^0(s) = 0$, since ambulances are present at nodes 1 and 2. Moreover, $r_3^0(s) = 0$ as well, because node 3 is the destination of a driving ambulance. The closest ambulance to node 4 is in node 2, since the ambulance traveling from 4 to 3 is assumed to be at its destination. Therefore, $r_4^0 = 1073$, and $r_5^0(s) = 1323$. At last, the computed unpreparedness is $\frac{3}{5} \times 0 + \frac{1}{5} \times 1073 + \frac{1}{5} \times 1323 = 479.2$. This is the expected time required to respond to the next incident for the configuration 1,2,3.

We did not consider the ambulances in phase 1, 2 or 3, for specific reasons. The expected remaining busy time of phase 1-ambulances and phase 3-ambulances is probably too large, and thus they are not considered. Although phase 3-ambulances are dispatchable to an incident after their remaining transportation time plus T seconds, we assume that T is set in such a way that it is never beneficial to wait for an ambulance that is still in phase 3 for the response to an incident. Expected remaining busy times for phase 2-ambulance are shorter, but highly uncertain since it is not known whether a patient needs transportation in advance.

Note that there are several differences between the unpreparedness defined here and the preparedness introduced by Andersson and Värbrand (2007). First, ambulances that are busy at a hospital are not included in the definition of preparedness. Moreover, unpreparedness has the nice physical interpretation of the expected penalty to the next incident. After all, no artificial contribution factor is incorporated in the computation. Besides, the definition of

preparedness is based on travel times solely, while in the unpreparedness definition a general penalty function is incorporated.

3.2. Evaluation of the ambulance motions

At a decision moment of the first type, determining the unpreparedness of the state of the system is the first step in the heuristic. That is, the motion in which none of the ambulances move except for the ones on the road. We refer to this motion as the *static motion*, denoted by m_0. For the remainder, we denote the unpreparedness if m_0 is carried out by $U(s_0)$. Subsequently, we evaluate ambulance motions. Denote the remaining possible ambulances motions by $m_1, m_2, \ldots, m_K$, enumerated by $1, \ldots, K$. Moreover, let s_k denote the state of the system as if m_k was carried out instantaneously and all driving phase 0-ambulances would be at their destinations. Then, we compute $U(s_k)$ for $1 \leq k \leq K$ to obtain a classification of the ambulance motions. The best motion is the ambulance motion that minimizes the unpreparedness. That is, we select the motion m_l for which

$$s_l = \arg\min_{k=0,\ldots,K} U(s_k).$$

For decision moments of the second type, we do something similar. However, the ambulance that just finished service of a patient, either at scene or at a hospital, has to be relocated anyway. This is a consequence of the restriction that each ambulance has to return to a base location. Therefore, we cannot define the static motion as before, in which this ambulance would keep its position. Alternatively, we define our static motion to be equal to the motion in which the just finished ambulance is relocated to the nearest base station. We denote this static motion by m_0.

Note that the number of possible motions is $\mathcal{O}(AB)$, where A and B are the number of ambulances and base locations, respectively. For decision moments of the second type, the number of ambulance motions is $\mathcal{O}(B)$, since the dispatcher has to decide on a new location only for the ambulance that just finished service. Note that the computation of the unpreparedness can be done in $\mathcal{O}(NA)$ time, since for N demand points we have to determine which of the A ambulances is the closest phase 0- and phase 4-ambulance. Therefore, the total complexity of the algorithm is $\mathcal{O}(NA^2B)$, which is polynomial in the number of demand points, fleet size and number of base locations.

3.3. From motions to relocations

Let m_l be the best ambulance motion, and assume $m_l = (b_1^l, b_2^l)$ is the pair of base stations, where b_1^l is the origin and b_2^l the destination. Once the ambulance motion is determined, the dispatcher needs to make a decision concerning the exact execution of this motion. To be more specific, the number of additional ambulances and which ones involved in carrying out this motion need to be determined. We do this by solving a *Linear Bottleneck Assignment Problem* (LBAP). The formal definition of the LBAP is: given two sets V and W, together with a weight function $c: V \times W \to \mathbb{R}$. Find a bijection g: $V \to W$ such that the cost function $\max_{v \in V} c\big(v, g(v)\big)$ is minimized. The LBAP can be solved to optimality in polynomial time, for instance by methods presented by Burkhard, Dell'Amico, and Martello (2009).

In our setting, this is equivalent to the computation of an assignment of phase 0-ambulances to the base locations that have to be occupied by an ambulance in the new configuration, in such a way that the maximum driving time of an ambulance is minimized. To be more specific, if we denote the set of destinations for phase 0-ambulances by D_0, we define the set $W = \big\{D_0 \cup \{b_2^l\}\big\} \backslash \{b_1^l\}$. The set V consists of the current locations of the phase 0-ambulances. When there are multiple ambulances per location, we specify the elements corresponding to this location with subindices in either V or W. Therefore, $|V| = |W|$. Let c be the function describing the driving time between elements of V and elements of W, obtained from the driving time matrix R.

We can interpret the solution to the LBAP in our setting as follows: it is the minimal time required to perform the ambulance motion. Since we base the ambulance motion on the state of the system as it is at the decision moment, apart from the fact that we assume driving phase 0-ambulances to be at their destination, it is desirable that the new ambulance configuration is attained quickly. There is an obvious relationship between the number of additional ambulances participating in an ambulance motion, and the completion time of the ambulance motion: the more ambulances are allowed to be relocated, the faster the new ambulance configuration may be attained. However, it may occur that the number of extra ambulance relocations only has a small impact on the performance, since the gain of participation of additional ambulances in a motion may be limited. Therefore, in the next section we will restrict the dispatcher to relocate a limited number of additional ambulances. Moreover, we compare the performance and the number of ambulance relocations to the case in which all ambulances are allowed to take part in the motion.

3.4. Constraints on decisions

We restrict the dispatcher in two ways

1. The dispatcher is only allowed to perform the best motion if the gain in unpreparedness with respect to the static motion is substantial.
2. The dispatcher is not allowed to relocate more than M phase 0-ambulances in a motion.

In order to get a feeling about the necessity of the best motion, m_l, we compare it to the static motion m_0, defined as above. To be more specific, we compute

$$q := \frac{U(s_0) - U(s_l)}{U(s_0)},$$

where $U(s_0)$ and $U(s_l)$ denote the unpreparedness of the state of the system when, respectively, the static and best motion are performed. Note that $U(s_l) \leq U(s_0)$, since the best motion may equal the static motion. We define Q to be the *motion threshold*: the dispatcher may carry out the best motion only if $q > Q$. Note that $0 \leq q \leq 1$. If we set $Q = 1$, the dispatcher is restricted to the execution of the static motion solely. In contrast, if $Q = 0$, he/she is always allowed to perform the best motion, even if it results in just a small gain in unpreparedness. Note that we prefer to assess the performance using a relative metric as opposed to an absolute metric. The latter makes sense when a strict 0–1 penalty function is used, however, since we allow for general penalty functions the former is preferable.

The second type of restriction is closely connected to the third question at the end of Section 2.3: the way in which an ambulance motion is carried out, i.e., the number of ambulances used to perform an ambulance motion. The above-mentioned M is a hard constraint that holds for both types of decision moments and $1 \leq M \leq A$. Remember that a dispatcher may at any time redirect an ambulance if it is already on the road, since this does not count as an extra relocation. Thus, the number of redirected ambulances is not restricted by M.

In short, the restrictions are given by (Q, M). A summarizing diagram with the different steps of the method is displayed in Fig. 2. In the next section, we show some results regarding the performance of the system and the number of relocations as function of Q and M.

Remember that we only consider the closest ambulance. If each base location is the destination of at least one phase 0-ambulance

1. Consider the system as if each ambulance is at its destination.
2. For each combination of origin and destination:
   (a) Remove one ambulance from the origin.
   (b) Add one ambulance to the destination.
   (c) Compute the unpreparedness of the resulting configuration.
3. Select the best motion and compare it to the static motion.
4. If $q > Q$: Solve LBAP with at most M ambulances.

Fig. 2. Summary of the approach.

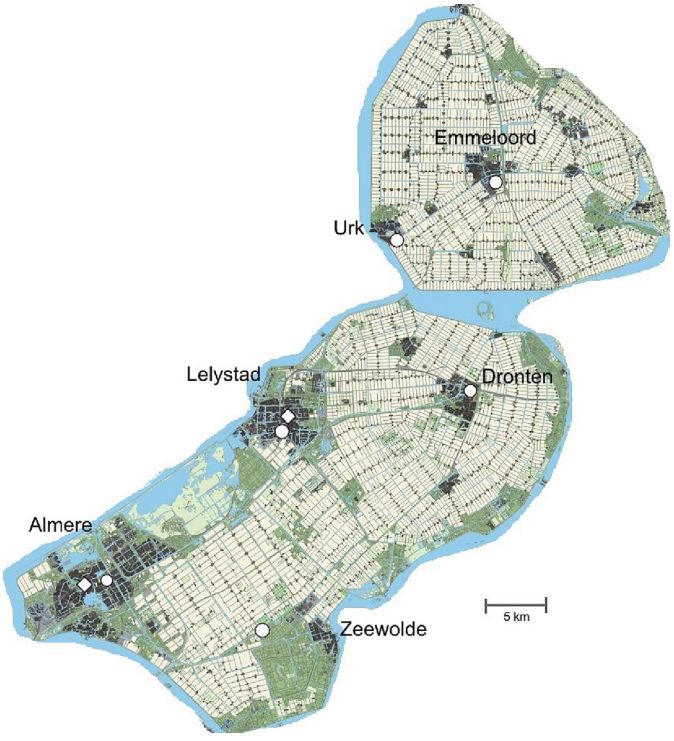

Fig. 3. Flevoland.

at a decision moment of the second type, all motions are evaluated as equally good. Similarly, for decision moments of the first type, it could occur that the best motion is not unique as well in such a situation. If this is the case, we create scarceness in the number of phase 0-ambulances by ignoring exactly one ambulance of each base station, and we compute the best motion based on this configuration. If each base location is occupied twice, that is, each base location is the destination of at least two ambulances, then we always carry out the static motion. However, for the regions and situations we studied, this was hardly the case.

4. Numerical case study

4.1. Experimental setup

In this section, we show results for Flevoland, displayed in Fig. 3. Flevoland is a region in the Netherlands and covers approximately 2,500 km^2. The number of inhabitants is nearly 400,000 of which 49 percent lives in the largest city: Almere. The remaining percentage of the population is mainly concentrated in one of the other five major towns, while only approximately 15,000 people live outside these six towns. All of these cities have exactly one base location at the dots in Fig. 3. Moreover, hospitals are present in Almere and Lelystad, displayed by diamonds.

Flevoland is divided in 93 different postal codes, for which between any pair of postal codes the driving time is given in a driving time table R. These driving times were estimated by the RIVM.[2] This driving time table was constructed in two steps. First, ambulance emergency speeds were estimated from a large amount of data, for 22 different road types. These average speeds were entered in a routeplanner, computing an estimate of the driving time for each pair of postal codes. We refer to Kommer and Zwakhals (2008) for a more detailed description of the computation of the driving time table.

We model Flevoland as a directed complete graph with 93 nodes, in which each arc is weighted according to R. To keep track of the actual locations for the driving ambulances, we need the *route* between each pair of postal codes. Theretofore, we define a postal code-incidence graph in which nodes are only connected by an arc if the corresponding postal codes are adjacent. In this incidence graph, the present arcs are weighted according to the driving time table. We use a shortest-path algorithm to compute all shortest paths, and we obtain both a sequence of lengths and the actual paths. Note that the computed shortest path length between the start- and endpoint is not necessarily equal to the driving time obtained from the driving time table R, as a consequence of the triangle-inequality. Therefore, we scale the whole sequence of times according to the driving time in the driving time table. Now, we obtain an estimate on the arrival time of the ambulance at a certain intermediate node on the route between start- and endpoint. In the determination of the actual location of an ambulance, we consider the driving time already past, and round it to the nearest number in the scaled sequence of times to estimate the actual location.

As objective, we use a compromise between minimizing the average response time and the number of incidents for which the response time exceeds the maximum allowed one. In The Netherlands, this maximum allowed response time is 15 minutes, but as mentioned before, this time includes dispatch and turn-out time. We assume that this dispatch and turn-out time is 3 minutes, which induces 12 minutes (720 seconds) as maximum allowed travel time. The penalty function we use is

$$f(t) = \begin{cases} \dfrac{1}{5e^{-0.008(t-720)}+5} & 0 \le t \le 720, \\ \dfrac{4}{5} + \dfrac{1}{5e^{-0.008(t-720)}+5} & t > 720. \end{cases} \quad (1)$$

This function is displayed in Fig. 4, and was composed in consultation with a policy officer of the ambulance service provider of Flevoland. Note that the focus in this penalty function is on minimizing the number of late arrivals rather than on minimizing the average response time. After all, an incident reached within the maximum allowed response time induces a penalty between 0 and 0.1, while an incident for which the maximum response time is exceeded, induces a penalty between 0.9 and 1.

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

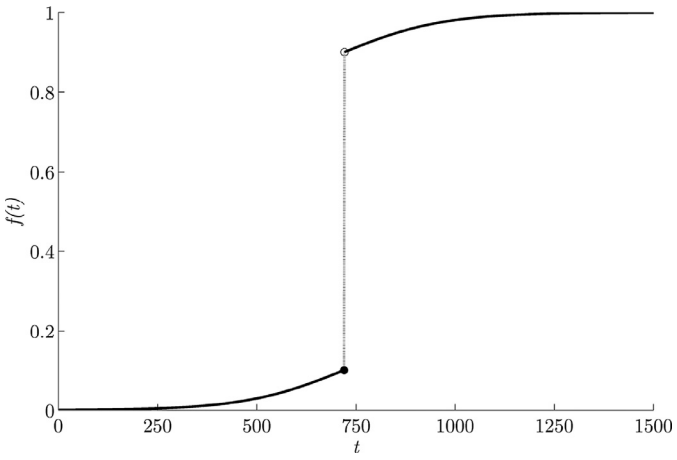

Fig. 4. Graphical representation of the penalty function of (1).

The ambulance service provider of Flevoland uses a target of 10 minutes for the hospital transfer time. After these 10 minutes, the ambulance is considered as idle by the dispatchers and it may be dispatched to another call. Therefore, we set $T = 600$.

We generate results by a discrete-event simulation using historical data. We have access to the following information of incidents: time and place (based on postal-code level) of occurrence, the on-scene time of the ambulance, whether the patient needed transportation to a hospital and the hospital time of the ambulance. At night, the mean on-scene time and mean hospital time are 1,170 seconds and 938 seconds, with standard deviations of 756 seconds and 661 seconds, respectively. Moreover, 71 percent of the patients needs to be transported to a hospital. During day time, the means are 1,090 seconds and 1,536 seconds, with standard deviations of 680 seconds and 631 seconds. In addition, 75 percent needs transportation to a hospital. We also use the historical data for the computation of the demand probabilities $p(i), i = 1, \ldots, N$, by dividing the number of requests at i by the total number of requests, for day and night separately.

No randomness is involved, since we use the actual historical data (trace-driven). The simulation evolves according to the system dynamics described in Section 2.1. When an ambulance just got freed from service and there are still requests waiting because no ambulances were available, the ambulance will immediately respond to the one that is longest in the system.

We consider two different situations

1. a critical situation, in which available ambulances are scarce, and
2. a realistic situation.

As mentioned before, the redeployment of ambulances may be beneficial if there is scarceness in the number of available ambulances. If we apply the heuristic method described in Section 3, we implicitly assume available ambulances are scarce. After all, the contribution of each node to the unpreparedness depends on one ambulance solely, namely the closest one. Therefore, in one of the situations that we consider, we assume that there is scarceness, i.e., the probability that there are no available ambulances for an incoming incident, is around 1 percent. To achieve this, we decrease the number of ambulances. We do this in such a way that the blocking probability (using the Erlang blocking formula) is around 1 percent. We call the outcome the critical situation.

In addition to the critical situation, we consider a realistic situation in which we use a more realistic number of ambulances. We adjust the actual number of ambulances on duty, since many of them are busy with ordered transport as well.

Table 1
Columns I and II represent the gain in performance and the increase in number of relocations for Q_{min} compared to $M = 1$, respectively, where Q_{min} is the value at which the minimum of the graphs in Figs. 5a and 6a is attained. Column III represents the gain in performance for Q_{min} with respect to $Q = 1$.

	Critical night situation			Realistic night situation		
	I	II	III	I	II	III
$M = 1$	–	–	27.1%	–	–	33.6%
$M = 2$	11.3%	32.5%	35.6%	6.2%	32.9%	37.7%
$M = A$	11.3%	37.5%	35.6%	8.0%	41.2%	38.9%

We simulate our system according to the historical data, which runs between January 2008 and September 2012. We make a distinction between day (07:30–17:00) and night (00:00–07:30). We do not consider the evening (17:00–00:00), since the extremes (day and night) are more interesting to serve as illustration. The total number of incidents during day and night in the data is 37,844 and 11,579, respectively. There are 1,704 natural days in our dataset, so on average there are approximately 22 and 6 incidents per day and night, respectively. When a day (night) is over, we reset our system to the initial state and proceed with the next day (night).

4.2. Critical night situation

In the critical night situation, we assume there are $A = 4$ ambulances. Moreover, during night time, 71 percent of the patients needs to be transported to a hospital. In Fig. 5a, we display the penalty per night as function of the motion-threshold Q, for $M = 1, 2, A$. However, since our system only contains four ambulances, the graphs for $M = 1$ and $M = A$ hardly differ. Note that the largest gap between $M = 1$ and $M = A$ is at $Q = 0$, i.e., the dispatcher is always allowed to perform the motion. This gap is approximately 11.3 percent, as observed in Table 1. Thus, there is a significant gain in performance if more than one ambulance is used in performing a motion. However, this performance gain comes at the price of extra ambulance relocations. This number, as function of Q for $M = 1, 2, A$ is displayed in Fig. 5b. We observe approximately six additional ambulance relocations per night.

If we compare the cases $Q = 1$ and $Q = 0$, that is, always the static motion and always the best motion is performed, respectively, we see a gain in performance as well, as observed in column III of Table 1. However, additional ambulance relocations were needed to achieve this gain, as observed in Fig. 5b. Furthermore, it is worth noting that although the graphs for $M = 2$ and $M = A$ coincide in Fig. 5a, this is not the case for the number of relocations. The participation of more than two ambulances in a motion has no effect on the performance here.

The increase just before $Q = 0.5$ in Fig. 5a and the corresponding downfall in Fig. 5b is explained by a geographical reasoning. Remember that there are two hospitals in the two largest cities: Almere and Lelystad. These two cities together are inhabited by 68 percent of the total population of Flevoland. From the base locations in these cities, none of the other four major towns can be reached within 720 seconds. From the base location in Emmeloord, 16 percent of the demand can be reached within 720 seconds though. However, for $Q \geq 0.5$, the gain related to performing the best motion, which sends an ambulance to Emmeloord, is too small and thus the static motion is always performed. Since the majority (71 percent) of the ambulances finishes the treatment of a patient at a hospital, it hardly occurs that an ambulance becomes deployable again at one of the four other towns. Therefore, it often occurs that an ambulance entering phase 0

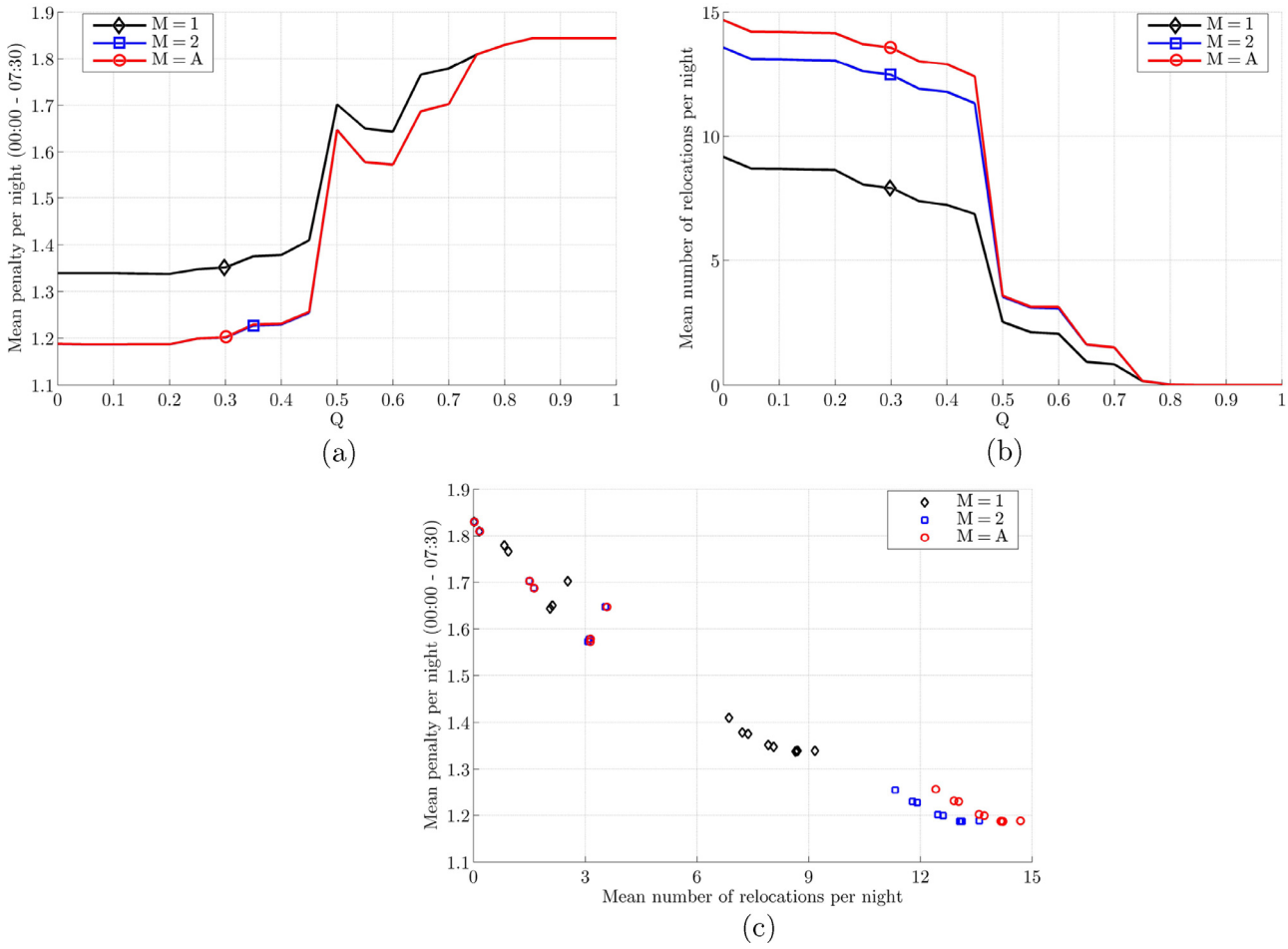

Fig. 5. The mean penalty (Fig. 5a) and number of relocations per night (Fig. 5b) as function of the motion-threshold Q for the critical night situation with $A = 4$. The accuracy, based on 95 percent-confidence intervals, is approximately 0.08 for Fig. 5a and 0.2 for Fig. 5b. Fig. 5c displays the relation between penalty and number of relocations per night.

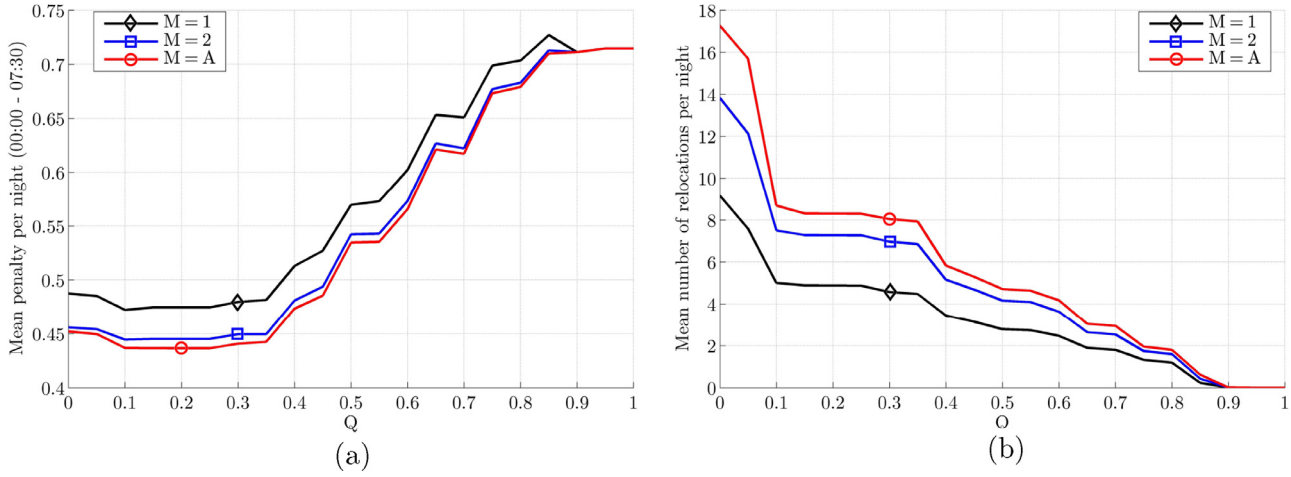

Fig. 6. The mean penalty (Fig. 6a) and number of relocations per night (Fig. 6b) as function of the motion-threshold Q for the realistic night situation with $A = 7$. The accuracy, based on 95 percent-confidence intervals, is approximately 0.04 for Fig. 6a and 0.2 for Fig. 6b.

again at one of the largest two cities, is not relocated to Emmeloord, because the static motion is performed. This results in a decrease in both performance and number of ambulance relocations.

The peak at $Q = 0.5$ in Fig. 5a can be explained by a similar reasoning. For $Q > 0.5$, the static motion is performed if an incident occurs in Almere and an ambulance is present in Emmeloord. That is, no ambulance is redeployed from Emmeloord to Almere. However, at $Q = 0.5$, the best motion is performed in this situation, in which Emmeloord is the origin and Almere the destination. The time to perform this motion for a single ambulance is 32 minutes, while it takes at least 20 minutes when multiple ambulances participate in the motion. Since many ambulances finish their service in Almere, it often occurs that an ambulance finishes

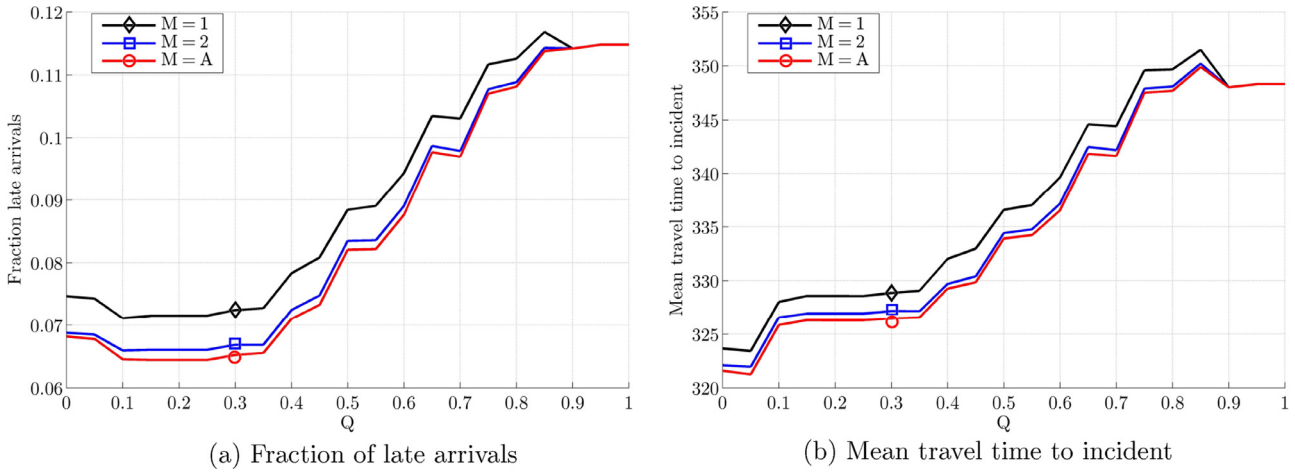

Fig. 7. The fraction of incidents for which the maximum allowed travel time of 720 seconds is exceeded, and the mean response time as function of the motion-threshold Q for the realistic night situation with $A = 7$. The accuracy, based on 95 percent-confidence intervals, is approximately 0.04 for Fig. 7a and 2 for Fig. 7b.

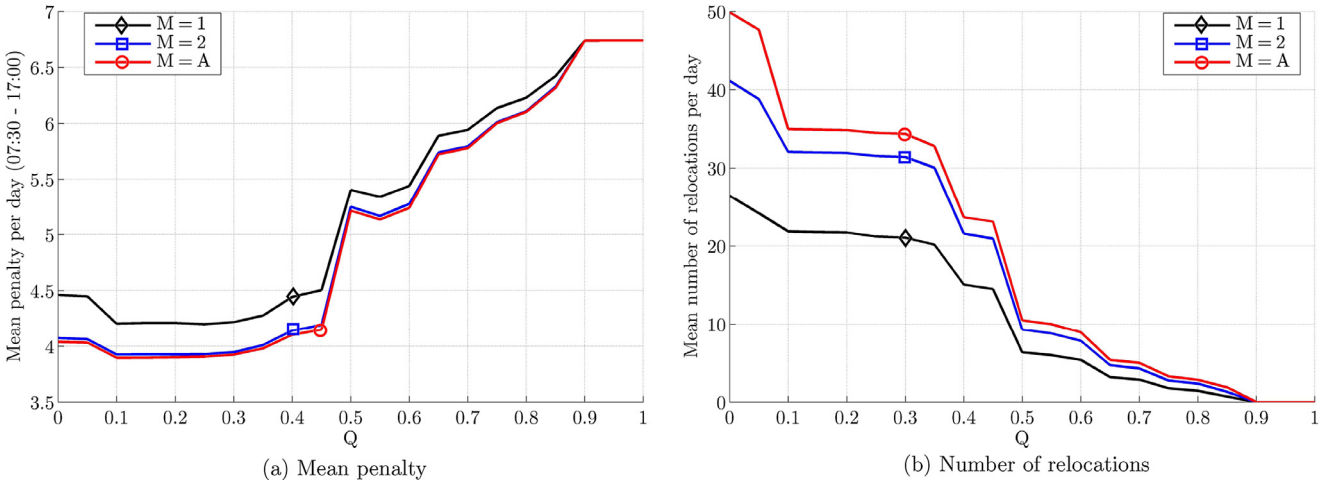

Fig. 8. The mean penalty (Fig. 8a) and number of relocations per day (Fig. 8b) as function of the motion-threshold Q for the critical day situation with $A = 6$. The accuracy, based on 95 percent-confidence intervals, is approximately 0.2 for Fig. 8a and 0.3 for Fig. 8b.

service before a relocated ambulance arrives there. For Almere, this is beneficial and a better performance can be observed there. However, this gain in performance does not outweigh the loss in Emmeloord. After all, there is no ambulance in the neighborhood for a possible large amount of time in a relatively large part of the region. This effect vanishes for $Q < 0.5$ by the reasoning described above. Concluding, in general, performing the best motion does not always result in a better performance compared to the static motion.

We also observe a small decrease in the number of relocations between $Q = 0$ and $Q = 0.05$ in Fig. 5b, although this is not reflected in the performance in Fig. 5a. If we consider the values of the two objectives intertwined in the penalty function, separately, we observe that at $Q = 0$ the fraction of late arrivals is 0.225 for $M = 1$ and 0.197 for $M = A$. The mean travel times to an incident are 450 seconds and 433 seconds, respectively.

4.3. Realistic night situation

In the realistic night situation, seven ambulances are on duty. In the initial state, each base location is occupied by one ambulance. The remaining ambulance is located at the base location in Almere. The graphs for the mean penalty and the number of relocations as function of Q are displayed in Fig. 6. In Fig. 6a, the confidence intervals overlap, but we are more interested in the patterns and the relation between the different lines. Note that a gap exists between the graphs for $M = 2$ and $M = A$. At $Q = 0.2$, this gap is approximately 6.2 percent and 8.0 percent for $M = 2$ and $M = A$ with respect to $M = 1$, as in column I of Table 1. Thus, by allowing the dispatcher to use more than two ambulances in performing a motion, the performance improves. However, this improvement is small compared to the performance gain if one allows two ambulances to participate in the motion instead of one. In the last column of Table 1, results on the comparison between $Q = 0.2$ and $Q = 1$ are displayed.

If we compare $Q = 0$ and $Q = 0.1$, we observe a tremendous decrease in number of relocations in Fig. 6, and the penalty decreases as well, albeit to a lesser extent. This behavior is explained by the choice of the penalty function. Results for the two objective functions compromised in the penalty function separately are displayed in Fig. 7. Usually, these two objectives are not really conflicting. However, a decrease in the fraction of late arrivals can be observed between $Q = 0$ and $Q = 0.1$, while the mean response time increases. This is due to one particular motion. Urk can be reached within 720 seconds from Emmeloord only. For $Q = 0.1$, the gain in unpreparedness is too small if the best motion is performed, so we do not send an ambulance to Urk. For $Q = 0$, always the best motion is performed, which sends an ambulance to

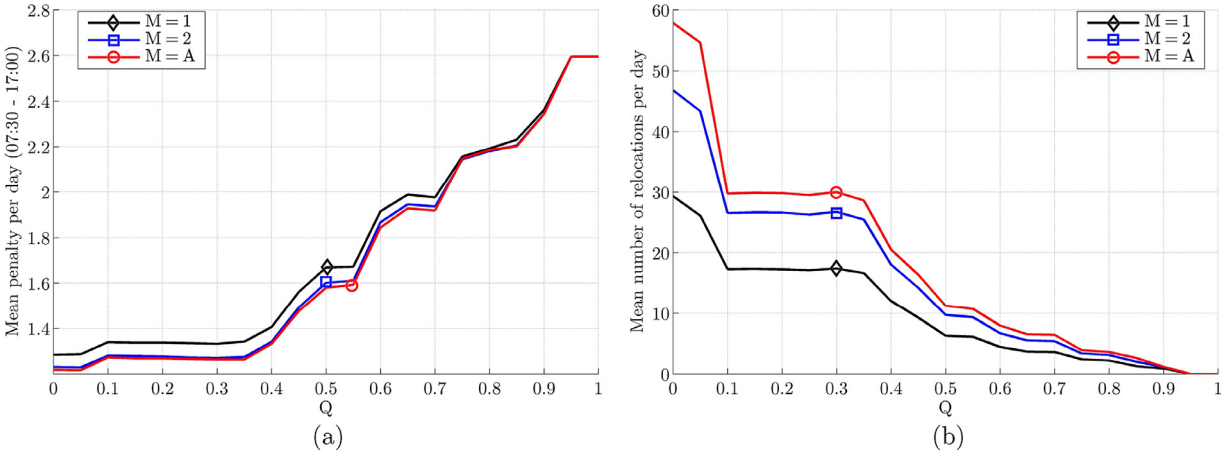

Fig. 9. The mean penalty (Fig. 9a) and number of relocations per day (Fig. 9b) as function of the motion-threshold Q for the realistic day situation with $A = 12$. The accuracy, based on 95 percent-confidence intervals, is approximately 0.07 for Fig. 9a and 0.4 for Fig. 9b.

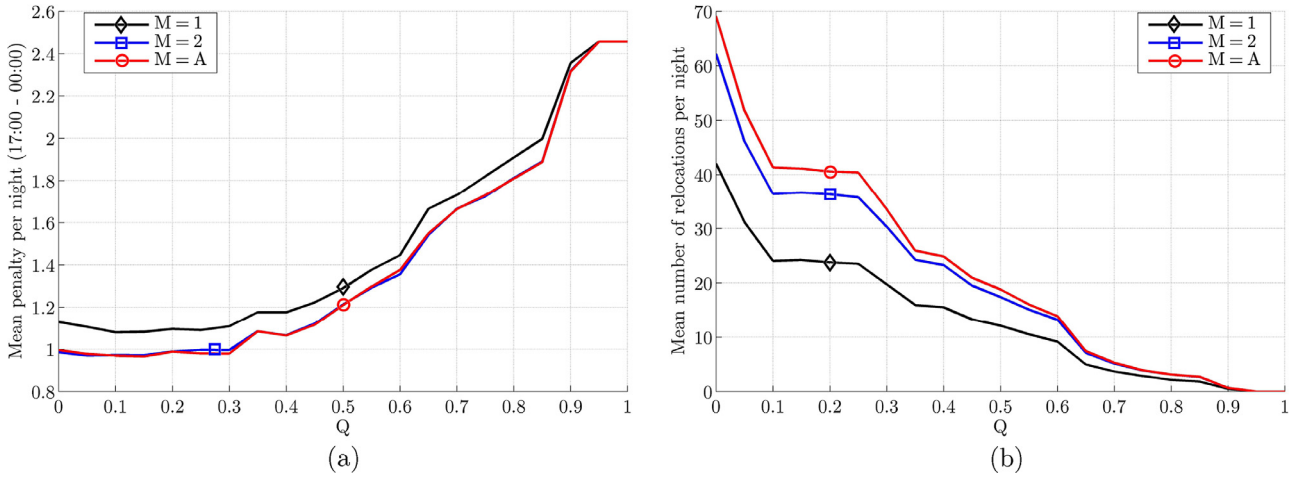

Fig. 10. The mean penalty (Fig. 10a) and number of relocations per night (Fig. 10b) as function of the motion-threshold Q for the realistic night situation with $A = 15$. The accuracy, based on 95 percent-confidence intervals, is approximately 0.12 for Fig. 10a and 1.5 for Fig. 10b.

Urk. However, performing this motion is of influence on the mean response time only and not on the late arrivals. The performance loss can be explained by the fact that one ambulance is send to Urk, where it is actually not really needed. This underlines the statement that always performing the best motion does not necessarily result in a better performance.

A similar reasoning holds for the peak around $Q = 0.85$ in Fig. 7, especially for $M = 1$. It takes much time to perform the best motion, as an ambulance has to move from Urk to Almere.

If we compare the critical and realistic situation in Table 1, we observe that the benefit of using more than one ambulance in a motion is larger for the critical situation than for the realistic setting. However, the benefit of doing relocations at all is larger in the realistic situation, as column III indicates.

4.4. Critical day situation

During daytime, the maximum number of ambulances needed to ensure that we are always in the critical situation is six. We locate an ambulance at each base station in the initial state. Compared to the night situation, slightly more patients need transportation to a hospital: this percentage is 75 percent. Results for this situation are displayed in Fig. 8 and Table 2.

We observe clear similarities between the night and day situation. For instance, the peak at $Q = 0.5$ is still present, although we use different demand probabilities for the night and day situation. Moreover, we again observe the drop between $Q = 0$ and $Q = 0.1$, which is explained by the same reasoning as in the realistic night situation. We conclude from Table 2 that the benefit of using more ambulances in a motion has decreased, compared to the critical night situation. However, the gain in performance compared to the case in which no relocations are performed, is larger.

Table 2
Columns I and II represent the gain in performance and the increase in number of relocations for Q_{min} compared to $M = 1$, respectively, where Q_{min} is the value at which the minimum of the graphs in Figs. 8a and 9a is attained. Column III represents the gain in performance for Q_{min} with respect to $Q = 1$.

	Critical day situation			Realistic day situation		
	I	II	III	I	II	III
$M = 1$	–	–	37.5%	–	–	50.5%
$M = 2$	6.7%	32.0%	41.7%	4.2%	37.3%	52.6%
$M = A$	7.3%	37.7%	42.1%	5.2%	49.3%	53.1%

4.5. Realistic day situation

In the realistic day situation, 12 ambulances are present in the system. Results for this case are listed in Table 2 and Fig. 9. There are some differences compared to the situations before. For

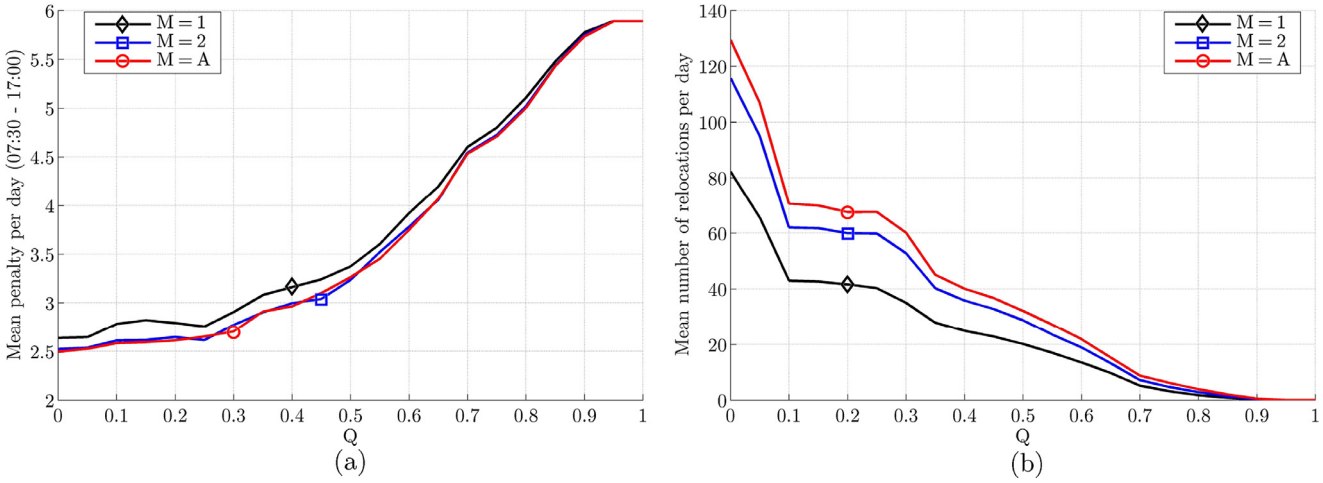

Fig. 11. The mean penalty (Fig. 11a) and number of relocations per night (Fig. 11b) as function of the motion-threshold Q for the realistic day situation with $A = 24$. The accuracy, based on 95 percent-confidence intervals, is approximately 0.15 for Fig. 11a and 2.3 for Fig. 11b.

Table 3
Columns I and II represent the gain in performance and the increase in number of relocations for Q_{min} compared to $M = 1$, respectively, where Q_{min} is the value at which the minimum of the $(M = 1)$-graphs in Figs. 10a and 11a is attained. Column III represents the gain in performance for Q_{min} with respect to $Q = 1$.

	Realistic night situation			Realistic day situation		
	I	II	III	I	II	III
$M = 1$	–	–	56.0%	–	–	55.2%
$M = 2$	11.1%	51.2%	60.5%	4.5%	40.8%	57.2%
$M = A$	11.5%	71.6%	60.7%	5.7%	57.4%	57.6%

instance, there is now an increase in penalty between $Q = 0$ and $Q = 0.1$, as observed in Fig. 9a. This is explained by the fact that the number of ambulance in the rest of the region is enough, and we can send an ambulance to Urk. This benefits the average response time, while the fraction of late arrivals is not influenced by this. Moreover, the gap between $M = 1$ and $M = 2$ has further narrowed.

4.6. Amsterdam

In addition to the results on the relatively rural region of Flevoland, we provide a short numerical study on one of the most crowded regions in the Netherlands: Amsterdam and its surroundings. This region covers approximately 630 km^2 and is home to 1.2 million inhabitants, of which 68 percent lives in Amsterdam itself. There are 162 postal codes, and the 162 × 162 table of driving times was provided by the RIVM, cf. Kommer and Zwakhals (2008). Moreover, there are eight base locations in this region, and the number of hospitals equals eight as well. We again use (1) as the penalty function and we retain the parameters corresponding to the maximum allowed response time and the dispatch and turnout time as in the Flevoland case. Moreover, historical data of the year 2011 serves as the basis for our computations, and we distinguish a day and a night situation. The total number of incidents in 2011 was 12,362 and 38,784 during night and day, respectively. This results in 34 and 106 incidents on average per night and day.

We consider both the realistic night and day situation with 15 and 24 ambulances, respectively. Results are displayed in Figs. 10 and 11, and Table 3. Since Amsterdam is a smaller region than Flevoland and there are more base locations in Amsterdam, the driving times between base locations are smaller. Moreover, a lot more incidents occur in Amsterdam. However, these differences are not reflected in the results: many of the results carry over to Amsterdam. We highlight one difference:

In the Flevoland cases, $M = 2$ results in a higher penalty than $M = A$ in general. However, for Amsterdam, these two graphs are intertwined, as can be observed in Figs. 10 and 11. For some $Q-$values, the usage of only two ambulances in a motion results in a better performance than the unlimited case. This can be explained by both the difference in area of the two regions, and the difference in number of base locations. As a consequence, the driving times between base locations in Amsterdam are shorter compared to Flevoland. Therefore, it makes less sense to break up an ambulance motion in multiple parts to reduce the time required to perform the motion.

5. Summary and conclusion

In this paper, we analyzed the effect of ambulance relocations on the performance of the ambulance service provider. Theretofore, we described an ambulance redeployment model, in which a performance measure related to the response time can be chosen by the ambulance service provider by defining a corresponding penalty function. Moreover, we presented a heuristic for computing ambulance motions and relocations at decision moments. In this heuristic, we restricted the number of ambulance relocations in two ways: the first one is related to the necessity of the ambulance motion, and for the second we set bounds on the number of ambulance relocations within a motion. We used historical data of two regions in the Netherlands to simulate the system, and we showed results for one particular penalty function suggested by an ambulance service provider for both regions. We distinguished a day and night scenario, and we made a distinction between a realistic situation and a critical situation, in which there is always undercapacity in the number of idle ambulances.

The presented results all imply that there is a significant improvement if ambulances are relocated, compared to the static policy in which always the static motion is performed ($Q = 1$). Moreover, this decrease in penalty is largest if only a few ambulance relocations are allowed instead of zero. However, this behavior levels off: it gets harder and harder to increase the performance by executing additional ambulance relocations. Even allowing too many relocations may result in a worse performance. We observed that this could also be a consequence of the chosen penalty function: performance measures that seem to be strongly related to each other, can be conflicting. The graphs presented in this paper can be very useful for ambulance service providers to gain

insights in the relationship between performance and number of relocations.

We end this paper with a short note on further research. In this paper, we restricted the dispatcher at a decision moment of the first type to change the ambulance configuration at at most two points: the origin and the destination. However, it could be beneficial for the performance if this restriction would be relaxed, but this probably comes at the expense of more relocations. Moreover, the relation between performance or number of relocations and number of decision moments is interesting as well: what would happen if one decreases (e.g., only when an ambulance is newly free) or increases (e.g., every minute) the number of decision moments? One could also impose a bound on the relocation time of an ambulance per relocation. This will influence the performance and number of relocations as well. In addition, the balance of number of relocations per ambulance vehicle is an interesting topic, especially from the crew's perspective. A policy in which only a few ambulances need to relocate several times while others do never, is probably not desirable. This paper can be used as a basis for these interesting research topics. At last, the method presented in this paper will be tested in a real-life pilot in the emergency control room of Flevoland to support the dispatchers in their decisions regarding ambulance relocations.

Acknowledgements

This research was financed in part by Technology Foundation STW under contract 11986, which we gratefully acknowledge. We also would like to thank the ambulance service providers of the regions of Flevoland and Amsterdam for providing data, and the RIVM for providing the driving time tables.

References

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

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

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

Burkhard, R., Dell'Amico, M., & Martello, S. (2009). *Assignment problems*. Philadelphia: SIAM.

Carter, A., Gould, J., Vanberkel, P., Jensen, J., Cook, J., Carrigan, S., Wheatley, M., & Travers, A. (2015). 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, 17*(6), 670–678.

Church, R., & ReVelle, C. (1974). The maximal covering location problem. *Papers Regional Science Association, 32*(1), 101–118.

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

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

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

Gendreau, M., Laporte, G., & Semet, F. (2006). The maximal expected coverage relocation problem for emergency vehicles. *Journal of the Operations Research Society, 57*, 22–28.

Jagtenberg, C., Bhulai, S., & van der Mei, R. (2015). An efficient heuristic for real-time ambulance redeployment. *Operations Research for Health Care, 4*, 27–35.

Kommer, G. Zwakhals, S. (2008). Referentiekaderspreiding en beschikbaarheid ambulancezorg. RIVM Briefrapport 270192001/2008, 2008. https://www.ambulancezorg.nl/download/downloads/1538/referentiekader-spreiding-en-beschikbaarheid-2008.pdf

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

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

Lim, C., Mamat, R., & Bräunl, T. (2011). Impact of ambulance dispatch policies on performance of emergency medical services. *IEEE Transactions on Intelligent Transportation Systems, 12*, 624–632.

Maleki, M., Majlesinasab, N., & Sepehri, M. M. (2014). Two new models for redeployment of ambulances. *Computers & Industrial Engineering, 78*, 271–284.

Matteson, D., McLean, M., Woodard, D., & Henderson, S. (2011). Forecasting emergency medical service call arrival rates. *The Annals of Applied Statistics, 5*(2B), 1379–1406.

Maxwell, M. (2011). *Approximate Dynamic Programming Policies and Performance Bounds for Ambulance Redeployment.* Ph.D. thesis. Graduate School of Cornell University. Chapter 4

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

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

Naoum-Sawaya, J., & Elhedhli, S. (2013). A stochastic optimization model for real-time ambulance redeployment. *Computers & Operations Research, 40*, 1972–1978.

Owen, S., & Daskin, M. (1998). Strategic facility location: a review. *European Journal of Operational Research, 111*, 423–447.

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

Zhang, L. (2012). *Simulation Optimisation and Markov Models for Dynamic Ambulance Redeployment.* Ph.D. thesis. The University of Auckland. Chapter 5