Ambulance Dispatch Center Pilots Proactive Relocation Policies to Enhance Effectiveness

Martin van Buuren
Centrum Wiskunde & Informatica, Department of Stochastics, Amsterdam, the Netherlands
Vrije Universiteit Amsterdam, Faculty of Sciences, Amsterdam, the Netherlands, m.van.buuren@cwi.nl

Caroline Jagtenberg
Centrum Wiskunde & Informatica, Department of Stochastics, Amsterdam, the Netherlands, c.jagtenberg@cwi.nl

Thije van Barneveld, Rob van der Mei, Sandjai Bhulai
Centrum Wiskunde & Informatica, Department of Stochastics, Amsterdam, the Netherlands
Vrije Universiteit Amsterdam, Faculty of Sciences, Amsterdam, the Netherlands {t.van.barneveld@cwi.nl, r.d.van.der.mei@cwi.nl, s.bhulai@vu.nl}

In life-threatening emergency situations in which every second counts, the timely arrival of an ambulance can make the difference between survival and death. In practice, the response-time targets, defined as the maximum time between the moment an incoming emergency call is received the moment when onsite medical aid is provided, are often not met. A promising means to reduce late arrivals by ambulances is to *proactively relocate ambulances* to ensure good coverage by the available ambulances in real time. This paper evaluates two dynamic relocation policies that an ambulance service provider in the Netherlands modified for operational use and implemented in a software tool for real-time decision support. The policies were used in a pilot program within a dispatch center for 12 weeks. Based on the success of this pilot, our policies were adopted for ongoing use and permanent implementation. This paper describes the relocation methods, evaluates the pilot, provides statistics for efficiency improvements, and discusses the experiences of ambulance dispatchers and management.

Key words: ambulance; EMS; relocation; DAM; dynamic; management; operations.

Ambulance service providers (ASPs) worldwide must implement policies to improve efficiency, such as budget cuts or performance improvement programs. They can obtain efficiencies via changes to medical equipment, staff training, and the logistic domain. In this paper, we focus on the latter. The goal is to allocate the "right resources at the right time at the right place," such that the probability of meeting their response-time targets, within given budget constraints, is optimized.

The ambulance service provisioning process has several stages. When an emergency occurs, a dispatch center receives an emergency call, typically a 911 or 112 call (Stage 1). During this stage, an agent at the dispatch center performs triage (i.e., asks the caller a set of questions to assess the severity of the emergency). If the incident requires ambulance service, the agent immediately dispatches an ambulance—usually the closest available ambulance—to the scene of the emergency (Stage 2). The target response time, defined as the elapsed time between the moment that a call comes in and the moment that the ambulance arrives at the emergency scene, is country specific; in the Netherlands, the response-time target for high-emergency calls is 15 minutes. After performing onsite medical treatment (Stage 3), the emergency personnel on the ambulance may transfer the patient to a hospital (Stage 4). Upon completion of the patient transfer, the ambulance is available for handling the next emergency.

The traditional ambulance service provisioning paradigm is static and reactive. That is, each ambulance has a fixed base location (also referred to as a waiting site) from which it is dispatched in response to an incoming emergency call. When the ambulance becomes available again, it is sent back to either its base location or to service another emergency. This classic static and reactive approach to ambulance service provisioning is simple, but often highly inefficient, particularly in situations where multiple emergencies occur

simultaneously, and potentially leads to coverage problems, the late arrival of ambulances, and ultimately to loss of lives.

A promising and powerful means to boost efficiency is enforcing *proactive relocations* (i.e., proactively relocating ambulances to locations at which they can provide the 'best' coverage to the geographical ambulance region that each ambulance serves). In practice, one has to carefully balance the trade-off between coverage improvement and additional costs: over time, a relocation leads to additional fuel costs and wear and tear on the ambulances. Moreover, ambulance personnel are often reluctant to making relocations, unless they believe the relocations are absolutely necessary. That is, practitioners accept the enforcement of proactive relocations only if they improve efficiency and limit the number of relocations. Motivated by these factors, we developed several algorithms to optimize proactive relocations; that is, we developed methods that generate suggestions to the agents (dispatchers) in the ambulance dispatch centers about *when* to relocate, *which ambulance* to relocate, and *where* to relocate that ambulance. In practice, relocation suggestions are made by simply displaying an arrow on the dispatcher's monitor; the arrow indicates which ambulance to relocate to which base location. We emphasize that these arrows only give *suggestions* for relocations: the dispatching agent makes the final decision on whether to enforce a relocation.

To assess the practical usefulness and performance of our algorithms, we ran a pilot for 12 weeks. In doing so, we adapted these dynamic relocation policies so that they comply with local regulations and ease integration to the dispatchers' daily practice. In this study, we partnered with GGD Flevoland, an ambulance service provider (ASP), and CityGIS Homeland Security, a company that provided us with real-time data streams and navigation software. Based on the success of this pilot, GGD Flevoland and other dispatch centers adopted our policies for ongoing use and permanent implementation.

We organized the remainder of this paper as follows. In the *Literature Review* section, we review relevant literature. In the *Relocation Policies* section, we outline two relocation policies (and provide more details in Appendix A). Subsequently, in the *Adjustments for Use* section, we show how we adapt these two dynamic relocation models for use in practice. In the *Evaluation* section, we discuss the results of the pilot, which we ran in a real EMS dispatch center to evaluate performance statistics and practitioner experiences. In the *Conclusion* section, we provide concluding remarks and give recommendations.

Literature Review

In general, ambulance allocation and relocation models are classified into two main categories: static location models and dynamic relocation models; see Brotcorne et al. (2003), Li et al. (2011), and Bélanger et al. (2015) for overviews of both types of models. Early proposed ambulance location models were integer linear programming (ILP) formulations, such as the set-covering location problem (SCLP), presented in Toregas et al. (1971), and the maximum-covering location problem (MCLP), which Church and ReVelle (1974) proposed. The goal of the MCLP is to find an allocation of ambulances to potential base locations to maximize the demand coverage. However, this model ignores the probabilistic aspects present in EMS systems, most notably that some demand points may no longer be covered once an ambulance is dispatched. This shortcoming was addressed by incorporating a *busy fraction* (i.e., the fraction of time an ambulance is unavailable) into the MCLP model. The resulting model, called the maximum expected covering-location problem (MEXCLP) and proposed in Daskin (1983), was one of the first probabilistic models for ambulance location. Although the MEXCLP model has some limitations, most notably the assumption that the ambulances are independent, it is still widely used as starting model, which can be extended. For example, in Batta et al.

(1989), the hypercube correction factors proposed in Larson (1974) were incorporated in the MEXCLP model to relax this independence assumption. In Erkut et al. (2008), the MEXCLP model is extended to a model that incorporates survival probabilities and probabilistic response times.

Static location models do not explicitly consider the state of the system following events, such as a change in the availability of an ambulance (when an ambulance has been dispatched or completed servicing a patient), the arrival of an ambulance at the scene of an emergency, or the departure of an ambulance for a hospital. In contrast, relocation executed in real time is the topic of many papers in the literature on dynamic models. In this literature, one can distinguish *offline* and *online* models. Offline models, which can be solved a priori, generate a look-up-table-like solution. Such a table provides a redeployment strategy for each possible system state. If the system state is described by the number of available ambulances (i.e., ambulances not busy with patient-related matters), such a table is called a *compliance table*. This table indicates the ideal locations for each possible number of available ambulances. Examples can be found in Gendreau et al. (2005), van Barneveld (2016), and Sudtachat et al. (2016).

Other offline dynamic models, not related to compliance tables, include the approximate dynamic programming (ADP) approaches proposed in Maxwell et al. (2009) and in Maxwell et al. (2010). In these papers, an approximate policy iteration is run offline to search for a good value function approximation. Once such a value function is obtained, the computation of a redeployment decision is fast and can be executed in real time. The computation of relocation and dispatching decisions by ADP is also the subject in Schmid (2012). Examples of other methods include stochastic programming (Naoum-Sawaya and Elhedhli 2013) and simulation-based optimization (Bjarnason et al. 2009).

In online models, no precomputation is executed. Based on the system state, a relocation decision is computed in real time, without using the results of an a priori computation. The first online relocation model, based on the double-standard model of Gendreau et al. (1997), is proposed in Gendreau et al. (2001). A tabu search heuristic and parallel computing are used to solve the relocation problem. Andersson and Värbrand (2006) use the notion of preparedness in the computation of redeployment decisions. Finally, our work is based on two online relocation models in Jagtenberg et al. (2015) and van Barneveld et al. (2016). These two policies have in common that they consider all redeployment options, compute a heuristic value for the benefit of each movement, and eventually propose the move with the best value. The policies differ, however, in how they define the value for a specific movement and handle the statuses of ambulances.

Few of the dynamic relocation policies described above have been implemented in practice. To our knowledge, only one company, Optima Corporation, had implemented repositioning models prior to the study we describe in this paper. It developed a commercial software package, Optima Live; however, because this package is commercial, all its details are not available to us. One feature that has been published is that Optima Live uses the real-time multiple-view generalized-cover repositioning model from Mason (2013). Other work that the Optima Corporation supports is presented in Richards (2007) and Zhang (2012); the latter includes more details on the software.

Relocation Policies

In our pilot, we tested two relocation policies in practice: (1) the *Dynamic Maximum Expected Coverage Location Problem* (DMEXCLP), and (2) the *Penalty Heuristic* (PH). The DMEXCLP policy, which is proposed in Jagtenberg et al. (2015), mandates that when a vehicle becomes idle after completing service for a patient, that vehicle goes to

the base location of choice within the region. This choice is made such that it maximizes the number of emergencies that are addressed within the response-time threshold, with an objective function and constraints similar to MEXCLP. We provide a full description of the DMEXCLP relocation policy in Appendix A.1.

The PH policy, which is proposed in van Barneveld et al. (2016), consists of two steps performed in sequential order. In Step 1, it computes the desired ambulance configuration (i.e., the distribution of the number of idle ambulances over the base locations). In the computation of this configuration, it uses the unpreparedness of each demand point (i.e., a reachability measure based on the response time of the closest available ambulance)—that the ambulance that can arrive at the emergency the quickest has the lowest unpreparedness score. The objective function of the policy minimizes the region-wide weighted unpreparedness. In Step 2, it calculates the actual movements of ambulances needed to reach the 'desired configuration' (determined in Step 1) in a minimal time, starting at the current configuration. The set of movements may include the use of *chain relocations*; that is, a limited number of simultaneous relocations are executed in order to achieve the desired configuration in minimal time. Appendix A.2 shows a summary of the DMEXCLP relocation policy. In Appendix A.3, we compare the two relocation policies.

Adjustments for Use

The two relocation models described above cannot be used in practice without modifications. During the implementation phase, we encountered a number of practical constraints that required us to modify the models to ensure they are applicable to an operational dispatch center. This section describes the adaptations we made.

Appendix A.5 includes an in-depth description of these adjustments; however, prior to

the implementation, we had to determine which input parameters to select for the pilot study.

Input Parameters in the Original Models

We discussed the policy input parameters with ASP management, including discretization of demand, and which base locations and chain relocation settings to select. These input parameters are the same for both policies.

Discretezation of the area in demand nodes: Both the DMEXCLP and the PH models assume that the service area is discretized and partitioned into, for example, N subareas (i.e., *demand nodes).* Thus, the next incident will occur at exactly one of these demand nodes; the probability that the next incident will occur at node i is denoted by a vector of probabilities p_i $(i = 1, \ldots, N)$ that sum up to 1. Thus, the demand location is modeled by a vector $(p_1, \ldots, p_N)$, which can be estimated or forecast based on historical data. For both policies, we aggregated to four-digit postal code numbers, each with an average of 4,000 inhabitants, and used the normalized number of inhabitants as the demand.

Base locations and travel times: Both policies require information about the locations (i.e., the nodes) of the existing base locations, and the expected driving time between each base location and each demand node. Based on discussions with ASP management, we decided to only use the base locations, and excluded relocation destinations that dispatchers used only occasionally (e.g., parking lots). Driving times between base locations and nodes were all precalculated and available in a database. We used navigation software to calculate the driving times from ambulances that were available, but not located at a base location.

Chain relocations: In discussions with dispatchers and management, we determined that a relocation chain may contain at most two simultaneous ambulance movements, and that we would use such a chain only when we could ensure a relocation duration gain of at least 10 minutes. When multiple relocation chains are possible, we use the one with the minimal relocation duration; see Appendix A.4.

Adjustment Series 1

In the implementation and system integration stage of the pilot, we had to make four adjustments to the original models to simplify the dispatchers' daily work.

First, we extended the DMEXCLP policy with chain relocations similar to the PH policy, which is a fairly straightforward process.

Second, we realized that the theoretical definition of relocation and the definition used in practice were different. In the theoretical policies, every ambulance that finishes a call receives a relocation instruction back to a base location. In practice, every ambulance has a default location to which it returns when it becomes available. Because entering a relocation requires dispatcher time and effort, and because the dispatcher will more willingly accept the relocation module's recommendation when it is in accord with the ASP's historical procedure (i.e., the procedure the ASP has used for many years), we changed the original policies to respect that all ambulances have historically defined default behaviors. Each shift has a default base location assigned. We assume that all available ambulances that are not actively performing a relocation must move to their default base location. For example, the D1 shift has as default base location, Dronten, and its hours are defined as 07:30 am to 16:30 pm. When an ambulance becomes available or when its shift starts, we assume that the vehicle moves to this base location. After 16:30 pm, we label this ambulance "in overwork." Only when a dispatcher enters a

relocation for an ambulance with the D1 shift into the system, the destination of that ambulance changes based on the information the dispatcher has entered.

Third, if the ambulance region is in a rural area, some shifts include a sleeping stage. During a night shift, the emergency medical technicians (EMTs) are allowed to sleep and may only be contacted when they are assigned to an incident. Hence, EMTs who are sleeping cannot be relocated, and assigning another ambulance to a base location at which they are sleeping is also prohibited. Base locations that are located in cities include some night shifts during which these emergency personnel must stay awake.

Fourth, we added decision moments at which a new relocation could be proposed. The DMEXCLP policy had only one decision moment: when an ambulance became available again, it should be optimally included in the fleet (i.e., optimal with respect to the objective function of the chosen policy). Discussions with the ASP management motivated us to include five additional decision points, which we list below, to these policies. These decision points are used in both policies, but only at the moments that relocation decisions are made.

1. **Start of shift:** EMTs contact the dispatch center when they start their shifts, and a dispatcher assigns the employees to ambulances and also specifies shift codes. In practice, ambulances, EMTs, and shifts are coupled as units, and can therefore be considered as interchangeable for the purposes of our work. When the dispatcher has completed entering a new shift into the system, the relocation module is activated. In the DMEXCLP model, the ambulance to which a shift code has just been assigned will be the origin of the relocation proposal.

2. **End of shift:** A shift can end in one of two ways. The most common way is when the scheduled end time of the shift is reached. In the case of overtime (i.e., an EMT has

exceeded his (her) scheduled number of working hours), the relocation module excludes all vehicles on which the EMTs are working overtime, because they cannot be dispatched to a new emergency. Alternatively, a dispatcher can manually end a shift.

3. **Ambulance dispatch:** When an ambulance is coupled to an incident, the relocation module proposes a new relocation.

4. **Ambulance availability:** When an ambulance is coupled to an incident, the relocation module updates its relocation proposal.

5. **Relocation entry into the system:** If an ambulance receives relocation instructions, the relocation module is updated with the new instructions. If the dispatcher follows the relocation proposal, as he (she) usually does, the system assumes that the optimal configuration has been achieved and does not provide any additional recommendations. Therefore, the relocation arrows on the dispatcher's monitor disappear. If a dispatcher makes a different relocation decision, the relocation module considers the relocation entered and generates a counter proposal (i.e., another relocation).

6. **Sleep interval beginning:** When a sleep interval starts, all EMTs who are allowed to sleep go to their assigned night bases, and other EMTs who do not have permission to sleep are requested to leave all base locations where colleagues will go to bed. We model this by sending all these ambulances back to their default bases. Consequently, a new relocation recommendation is calculated to optimally redistribute the ambulances.

Adjustment Series 2

After the first six weeks of policy evaluation, we discussed updates that would improve the performance and could be implemented within a week. One update addressed ambulances that are outside the ambulance region.

In some cases, EMTs on an ambulance must drive a patient outside their ambulance region and for multiple hours; for example, a patient that needs basic life support (BLS)

might require transportation to a hospital that has that medical specialization. When the ambulance becomes available again, the previous adjustment includes the current destination of the ambulance. Using coordinates to determine when the ambulance reenters its own region is difficult. Therefore, we use the navigation software to determine if a base location is within a 20-minute drive from any base location in its own region. If we do not find such a base location, we label the ambulance as "outside the region." When the ambulance is within a 20-minute drive from any base location within its region, the ambulance is marked as "inside the region" and the relocation module is updated to show that this ambulance is available.

Technical Details

We wrote the relocation module using the C++11-framework, TIFAR, which is an interface designed for ambulance research. CityGIS Homeland Security provided the navigation software and the communications interface for the system state we designed for this paper, which includes the location and status of each ambulance. This navigation software is the standard for emergency services in the Netherlands and includes all roads and travel speeds that EMS personnel use. The National Institute for Public Health and the Environment (RIVM), which also uses this navigational software, provided a look-up table for travel times between each pair of postal codes. Statistics Netherlands (CBS) provided demographic data for each postal code during the year 2013. The initial start-up of the program takes between 10 and 20 seconds because of cache creation; the program usually computes run-time relocation recommendations almost instantaneously—but sometimes may take up to a few seconds if other systems also require processing time from shared resources.

Table 1 The Stages of the Pilot in 2015 Are Listed with Their Corresponding Week Numbers

Stage	Week no.	Policy
	1, . . . , 35	Implementation and system integration at dispatch center.
1	36, 37, 38	Evaluating **Adjusted DMEXCLP 1**
	39	Changing the policy.
2	40, 41, 42	Evaluating **Adjusted PH 1**
	43	Fixed out of region ambulances.
3	44, 45, 46	Evaluating **Adjusted DMEXCLP 2**
	47	Changing the policy.
4	48, 49, 50	Evaluating **Adjusted PH 2**

Evaluation

We evaluated both policies for six weeks in the dispatch center of Flevoland, a Netherlands' ambulance region. Table 1 lists the various stages. In the first stage, we tested the adjusted DMEXCLP 1 policy for three weeks, spent a week switching policies, and then evaluated the Adjusted Penalty Heuristic 1 (PH 1) for three weeks. During the next week, we implemented and tested the Adjustment Series 2 for both policies. During the second half of the pilot, we evaluated three weeks of Adjusted DMEXCLP 2, one week of switching policies, and one week of Adjusted Penalty Heuristic 2 (PH 2). In this paper, we omitted data for the three weeks during which we switched policies (i.e., weeks 39, 43, and 47).

During the pilot, dispatchers were required to follow our relocation proposals, unless they had information not available to the system; examples include when an ambulance will be required for a BLS call in the near future or when a shift will end. In 2015, no policy or operational changes were made, other than the use of the relocation decision support software.

Table 2 The Table Shows the Performance and Volume that ASP GGD Flevoland Achieved from 2010 through 2015

Year	2015	2014	2013	2012	2011	2010
Call volume	24.136	23.337	22.459	22.427	21.521	20.884
Response time $\leq$ 15 minutes	95%	94%	94%	93%	93%	92%
High-urgency volume	10.288	11.006	9.863	10.707	10.245	9.534

The pilot evaluation of the dispatching policies included both quantitative and qualitative aspects, as we discuss below.

Quantitative Analysis

We start the quantitative analysis by analyzing long-term patterns. Table 2 shows the results over 2015, which GGD Flevoland provided, and the preceding five years (see Boers 2015). As the table shows, the total call volume increased by 3–4 percent per year, which is approximately the national average; 2013 was the only exception. The primary performance indicator for Dutch ambulance care is the percentage of late arrivals for high-urgency calls. Measured over the calendar year, Dutch ambulance law requires that for high-urgency calls each ASP must meet a response-time requirement of 95 percent of the calls within 15 minutes; the timer starts when a representative at the regional EMS dispatch center answers the telephone and stops when an ambulance arrives at the incident. Various improvements by GGD Flevoland have provided a steady increase of performance over the past years.

In 2014, only 7 of the 24 ambulance regions met this requirement; thus, a national average of 93 percent of high-urgencies calls was met on time (Boers 2015). In 2015, the year of this pilot, GGD Flevoland met the 95 percent on-time criterion for the first time in its history.

GGD Flevoland provided us with a database that includes details of call records; this enabled us to calculate the performance indicators that we list in Table 3. Analyzing the four stages yielded the following insights.

Table 3 The Table Provides an Overview of the Three Key Performance Indicators for Each of the Four Pilot Stages, and the Three-Week Average for 2015 before the Pilot Began: the Fraction of Late Arrivals at High-Urgency Calls, the number of High-Urgency Calls in Three Weeks, and the Number of Relocations

Stage	before	1	2	3	4
Response time $\leq$ 15 minutes	94.4%	97.2%	97.3%	96.0%	96.8%
High-urgency volume	579.2	596	619	668	682
Number of relocations	422.3	480	353	360	328

First, in all stages, we met 96.0 percent to 97.3 percent of the high-urgency calls on time, significantly exceeding the 95 percent requirement. Thus, the results we obtained during the pilot period compensated for the lower score of 94.4 percent achieved during the first months of the year; as we state above, this was the first year that GGD Flevoland met the legal response-time requirement.

Second, in the first stage, the dispatch center followed all of our relocation proposals, which resulted in 480 relocations in a three-week period; historically, approximately 420 relocations are normal for this ambulance region. Based on feedback we received and follow-up discussions with the EMS management, we determined that we would have to assign a lower bound on the contributed value of a relocation (i.e., the improvement on the objective function of the chosen policy). We omitted relocations that provide a lower contributed value. Because the EMTs went on strike during the final three stages, thus causing a data transfer delay of several months, we could only directly adjust the bound after the first stage. The results show that we significantly reduced the number of relocations.

Third, we see that performance in the third and fourth stages was worse than in the first two. We can explain this by the increase in high-urgency call volumes and the decrease in the number of relocations.

Finally, the number of ambulances remained almost constant over the last six years. An additional ambulance was provided in 2013 only, resulting in a total of 14 ambulances during the day. With an increase in call volume, a decrease in the number of relocations, an equal number of ambulances, and an increased fraction of on-time arrivals, we can conclude that our work resulted in more efficient relocations.

Qualitative Analysis

We observed that the dispatchers adopted our relocation proposals as much as they could. To accommodate legislation limitations or an approaching end of shift, or for another valid reason, they were allowed to ignore our relocation proposals. If a dispatcher decided to enforce a relocation, it always matched our relocation proposal. In the feedback that dispatchers gave us, they mentioned that the relocation proposals often coincided with their own insights. In some instances, our policies were counter to the dispatchers' intuition; however, when they applied our policies, they agreed these policies were better than their former ways of working (e.g., by intuition). Based on the new insights from our work, they have changed their daily routines. Other than the situations we discuss below, we are not aware of instances in which dispatchers strongly disagreed with our proposal. At the start of the pilot, some dispatchers did not like the concept of a relocation tool, which they believed would tell them how to perform their work. Their opinions changed during their weeks of use, and the dispatchers realized that the tool was supporting them—not controlling their work. Dispatchers always had the final say in each relocation decision. At the completion of the pilot, they rated their overall user experiences as very good.

In their previous method of working, the dispatchers used a small offline relocation look-up table, which told them where the first 5 ambulances, of the 13 available during the day, had to be positioned. This left many degrees of freedom for the dispatcher. Our relocation tool put in place a uniform policy that depends less on human decisions, ensures that the relocation decision is not dispatcher dependent, provides good coverage, and improves communications among the ambulance teams. Although during the first stage of the pilot, the EMTs told us that there were too many relocations, in the later

stages, because of the low number of relocations, they sometimes asked if the pilot had been terminated early and the system implemented in production mode.

In situations with many concurrent incidents, dispatchers know that the priority is communicating medical information to the healthcare professionals; because of this necessity to communicate, relocations have a high added value in ensuring optimal coverage. Another advantage of a relocation tool is that it reduces the time required to provide a relocation when the dispatcher is time constrained and under stress.

Using our software, the dispatchers could see the location of all available ambulances projected on one map. They were not given unnecessary information, which could cloud their ability to make decisions. It gave them a good overview of ongoing incidents and the status of the ambulances; in some situations, dispatchers were not able to determine the location and status of an ambulance prior to looking at our screen.

Ambulance service is a field where the logistic requirements change constantly. The causes include demands from local governments, agreements between neighboring ambulance regions, and new ambulance-management insights. Using a relocation tool provides an opportunity for management to modify the way that dispatchers work.

The dynamic ambulance management (DAM) policies mentioned in this paper leave room for improvement. Working-hour legislation dictates that employees on 24-hour and 16-hour shifts may be working 13 hours and 9 hours, respectively. During the remainder of the time on these shifts, they must relax at their home base locations. Our implementation does not address these issues; hence, the dispatcher must ignore some relocation proposals. Only a few locations have shifts of this length.

The implementation does not ensure that an ambulance is back at its home location when its shift ends. Dispatchers using our software must always keep this in mind. A

fairly straightforward solution that the dispatchers use is to instead send another ambulance that has sufficient remaining shift time and is also available at the same base location as the ambulance that the software suggests. We noticed that overtime by the ambulance teams did not increase during the pilot.

Conclusion

In this study, we put two dynamic ambulance-management policies into practice at an EMS dispatch center in Flevoland, an ambulance region in the Netherlands. We observed that the effectiveness of relocations improved when using a dynamic relocation policy, compared to previous years in which relocation algorithms were not used. One advantage we perceived is less latency, that is, the number of service calls for which the response time exceeds the threshold set, for a similar demand volume and number of relocations. The EMS region met the response-time requirement of 95 percent within 15 minutes for the first time in its history in 2015. The results indicate that both DAM policies perform comparably.

Other advantages are (1) all dispatchers work in a consistent way, (2) relocation decisions can be made faster during busy times at the dispatch center, (3) new policies can be introduced more rapidly, (4) dispatchers have a better overview of the available ambulances, and (5) the use of scientifically proven policies instead of dispatcher intuition improves efficiency and enables management to provide better oversight. Overall, the advantages of our policies strongly outweigh the disadvantages. Our implementation does not address the start and ends of shifts, or working-hour legislation; each would be an appropriate topic for further research. As a result of this study, however, multiple ASPs have adopted our policies for ongoing use and permanent implementation.

Acknowledgments

We thank the management and all dispatchers of GGD Flevoland (`www.ggdflevoland.nl`) for their enthusiastic participation in this research. We are also indebted to CityGIS (`www.citygis.nl`) for its live data streams, which include vehicle locations and status, and for allowing us to use of its navigation services. We thank the Police Department of Flevoland for its support when implementing our tool and for providing access to its systems, and RIVM (`www.rivm.nl`) for providing us with the look-up table with driving times between each two postal codes in the Netherlands. This research is supported by the Dutch Technology Foundation STW, which is part of the Netherlands Organization for Scientific Research (NWO), and which is partly funded by the Ministry of Economic Affairs.

Appendix A: Outline of the Original Relocation Policies

This appendix describes the two relocation policies that we tested in practice. In Appendix A.1, we discuss the DMEXCLP policy proposed in Jagtenberg et al. (2015), and in Appendix A.2, we discuss the PH policy proposed in van Barneveld et al. (2016). We compare the two relocation methods in Appendix A.3. Appendix A.5 provides details on how we modified the two methods for use in a real environment.

Appendix A.1: DMEXCLP

The DMEXCLP policy moves an ambulance when it becomes idle after having completed service for a patient, and directs it to a base location of choice within the region. Its sole objective is to maximize the number of incidents that are addressed within the time threshold. We first describe which aspects of the current state of the ambulance system should be used as input for the policy, and then explain how to make the relocation decision based on this input.

At a decision moment, the current state of the ambulance system may be observed. The DMEXCLP policy disregards all information about ambulances that are busy, and focuses purely on the set of idle ambulances. As we mention above, it uses the *destination* of the ambulances, rather than the actual location. For ambulances that are idle at a

Table 4 This Notation Listing Provides an Overview of the Variables We Use Most Frequently in This Paper and the Meaning of Each

V	The set of demand points.
W	The set of base locations, $W \subseteq V$.
T	The time threshold.
d_i	The fraction of demand in i, $i \in V$.
τ_{ij}	The expected driving time between i and j with siren turned on, $i, j \in V$.
n_i	The number of idle ambulances that have destination i, $i \in W$.

base, the destination equals the current location. This information is captured by the variables n_i: the number of idle ambulances that have destination i ($i \in W$). Table 4 provides an overview of the notation.

We next describe how the DMEXCLP algorithm computes the recommended relocation based on the previously described information. In some sense, we can regard this policy as a dynamic version of the maximum expected covering location problem—hence, its name. MEXCLP was designed to calculate an optimal static distribution of ambulances over base locations, by calculating the *coverage* of the region using an integer liner programming (ILP) formulation. The DMEXCLP policy reuses this definition of coverage, but it computes it for relocation purposes (without resorting to ILP solvers). The MEXCLP, as published in Daskin (1983), defines the coverage of a region in terms of a busy fraction, which it denotes as q. The busy fraction can be estimated by dividing the expected load of the system by the total number of available ambulances. This busy fraction is predetermined, and is assumed to be the same for all vehicles. Furthermore, vehicles are assumed to operate independently. Consider a node $i \in V$, which is within the range of k ambulances. Using the expected travel times τ_{ij}, $i, j \in V$, we can directly determine this number k. The travel times should be taken as estimates for movements, which are faster because ambulance sirens are on. The probability that at least one of these k ambulances is available at any point in time is then given by $1 - q^k$. If we let d_i be the demand at node i, the expected covered demand of this vertex is $E_k = d_i(1 - q^k)$.

The MEXCLP policy positions the ambulances in such a way that the total maximal expected covered demand, summed over all demand vertices, is reached.

The DMEXCLP policy proposes to send the ambulance that just became idle to the base, such that this allocation results in the greatest coverage according to the MEXCLP model. This is equivalent to choosing the base that gives the largest *marginal* coverage over all demand. This marginal coverage can be interpreted as the added value of having a kth ambulance nearby, and is given by $E_k - E_{k-1} = d_i(1-q)q^{k-1}$. The base that gives the largest marginal coverage over the entire region, and hence the destination that DMEXCLP proposes, can be expressed as follows:

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

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

Here, $\mathbb{1}$ denotes the indicator function. The expression for k in Equation (2) simply counts the number of idle ambulances that have a destination within the range of demand point i, assuming that the ambulance that is up for relocation will be sent to w. That is, it counts the number of ambulances that, in the near future, may respond in a timely manner to an incident in i. Because the number of base locations is typically small, we compute the maximization in Equations (1) and (2) by brute force (i.e., we iterate over all possible base locations and select the best location).

Appendix A.2: Penalty Heuristic

In this section, we summarize the penalty heuristic proposed in van Barneveld et al. (2016). The policy consists of two sequential steps. First, we compute the desired ambulance configuration (i.e., the number of idle ambulances per waiting site). Next, we

assign ambulances to the desired waiting sites according to the computed ambulance configuration.

Based on the observed information, the destinations of idle ambulances, and the location and elapsed service time of ambulances at hospitals, an ambulance configuration minimizing the *unpreparedness* is suggested. Unpreparedness is a measure of the (in)ability to quickly respond to incoming emergency calls, based on the configuration of ambulances. We refer to van Barneveld et al. (2016) for a formal definition of this concept. We prefer to talk about unpreparedness instead of coverage due to the objective criterion of interest: we define a *penalty function* by assigning a specific penalty to each realized response time. Note that this induces a generalization of the coverage concept: one can incorporate the commonly used performance criterion of coverage by defining a 0-1 function. Other performance criteria, such as response time, lateness minimization, or maximization of survival probabilities, can also be incorporated.

To compute the unpreparedness level of the region, we consider for each demand point the ambulance $\ell \in \mathcal{A}^{hosp}(s)$ that can be onsite as quickly as possible. This ambulance could be idle, but it is also possible that none of the ambulances can respond to such an incident in a timely manner. In that case, an ambulance currently busy with the transfer of a patient at the hospital may be asked to wrap up its task and depart for the emergency scene as quickly as possible. We are only allowed to preempt if the hospital transfer time has already lasted for a substantial amount of time (e.g., 10 minutes).

Let t_i denote the expected travel time to demand point i of the ambulance that can arrive the fastest; this can be an idle ambulance or an ambulance that is at a hospital. Note that in the latter case, we must add the remainder of a 10-minute allowed transfer time to the ambulance's driving time: $t_i = \tau_{\ell i} + \max\{0,\ 10\ minutes\ -\ passed\ transfer\ time\}$.

The unpreparedness is defined as the weighted sum of these t_i, that is, $\sum_{i \in V} d_i f(t_i)$, where d_i denotes the demand probability of point i, and $f(t_i)$ denotes the penalty value that corresponds to a minimal response time t_i for demand point i.

There are two decision moments: (1) when an ambulance has just been dispatched, and (2) when an ambulance becomes available after servicing a patient. At decision moments of the first type, the ambulance configuration, which is the resulting configuration if each idle ambulance is at its destination, may be changed at *at most* one pair of waiting sites. That is, one waiting site is selected as *origin* and one as *destination*. An ambulance leaves the origin and one arrives at the destination. Using brute force, we compute the unpreparedness among all allowed configurations. For a decision moment of the second type, the origin is given. This concludes the first part of the policy.

In the second step, we compute the optimal move to obtain the desired ambulance configurations, which is based on the current location of the ambulances, not the destinations. Quickly attaining this configuration is important. Therefore, we solve a *Linear Bottleneck Assignment Problem*. In this problem, one aims to find an assignment of ambulances to waiting sites that minimizes the maximum travel time to attain the desired configuration. Note that relocating multiple vehicles is allowed if this reduces the time until compliance. We refer to van Barneveld et al. (2016) for an illustration.

Table 5 The Table Contains a Summary of Properties for Each Relocation Method, and Indicates Whether the Property is Included (X) or Optional

	DMEXCLP	PH
Uses destinations of idle vehicles	X	X
Uses time until busy vehicle becomes idle	-	X
Focuses solely on one response-time target	X	optional
Uses multiple coverage	X	-
Allows relocation when vehicle becomes idle	X	X
Allows relocation when vehicle becomes busy	optional addition	X
Relocates multiple vehicles per decision moment	-	X
Computes solution in real time using brute force	X	X

Appendix A.3: Comparison of the Relocation Methods

In this section, we compare the two relocation policies that we discussed in the two preceding sections. From Table 3, we see that these DAM policies perform comparably for the pilot region. In Table 5, we compare properties of the relocation methods. The main difference is that the PH provides an optimal spread of ambulances over the safety region, while the DMEXCLP focuses on multiple coverage. As a result, rural areas tend to get more ambulances with the PH, and large cities get fewer ambulances. In contrast, DMEXCLP keeps ambulances near the cities, and only provides coverage to rural areas when a sufficient number of idle ambulances is available.

An additional difference is that the PH considers ambulances that are in a hospital. This slightly favors rural areas, because hospitals are often located in cities. An ambulance can be sent out of a city when another becomes idle at a hospital on short notice.

The teleportation assumption

At a decision moment, both policies use the locations of idle ambulances. Some of these ambulances are typically waiting at a base location, while others are driving toward a base location. Instead of keeping track of their true locations, we only store their destinations (also referred to as their teleportations). This choice has two important advantages. First, in a real-time system, keeping track of destinations is typically easier because they change than current locations less frequently. Second, there is a strategic 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. Hence, using their destinations can in some sense be regarded as taking a snapshot of the future.

Appendix A.4: Relocation Chains

We implemented a postprocessor for both models, such that a long relocation distance is 'cut' into multiple simultaneous ambulance movements, forming a *relocation chain*. Relocation chains provide a means to quickly reach a desired configuration of an ambulance, given the current ambulance configuration. For example, consider the situation where a relocation policy determines that an ambulance must be relocated from A to C, which takes 30 minutes. If base location B, located half way through this route, also contains an ambulance, simultaneously relocating one ambulance from A to B and another ambulance from B to C is a better option. Using this chain, the relocation duration is decreased from 30 to 15 minutes.

Appendix A.5: Adjustments to the DMEXCLP and the Penalty Heuristic

In this section, we provide an in-depth description on the adjustments made to the algorithms we used in the pilot to ensure that they are applicable in practice. Some of the changes are similar in both methods.

In the case in which an ambulance is redirected to its own base location using a relocation algorithm, we say that no relocation is necessary, because view this as default behavior. If no relocation is necessary, dispatchers are not required to contact the EMTs on ambulance team to tell its them that they can move to their own base location. Ambulance teams that may sleep during a night shift are always relocated to their home base; that is, when such an ambulance becomes available, the program shows that no relocation is required. We ensure that no other ambulance can go to a base at which an ambulance team is asleep; we provide details below.

When an ambulance that is out of the region becomes available (i.e., it is more than 20 minutes of driving time away from any base location in the pilot region), no relocation

recommendation is calculated. Instead, the program monitors the ambulance and calculates a relocation proposal when it enters the region in the same way it does when the ambulance becomes available.

DMEXCLP-specific adjustments: We used the parameter value $q = 0.3$ for the busy fraction, which was a realistic value for the pilot region. Base locations that host sleeping shifts are excluded from $w \in W$ in the argmax-argument of Equation (1), such that no relocation is recommended to a base at which people are sleeping. Ambulances that are out of region are not considered in calculating variable n_i ($i \in W$). At other relocation moments, we remove an ambulance a from the system state, and calculate its relocation recommendation to be *when it would become available* at its current location. We repeat this procedure for each available ambulance, and suggest to the dispatcher the one that provides the highest contribution to the coverage. Relocation chains are formed similar to the Penalty Heuristic.

PH-specific adjustments: An out-of-region ambulance is never considered to be the closest ambulance to respond to an incident. Furthermore, such an ambulance is not counted as driving to a base location. Base locations with sleeping ambulance teams are not included in the set of destinations. In PH, bases that have at least one ambulance are not considered as destinations. We allowed a base to have a second ambulance when all bases are filled. This is computed by first teleporting ambulances to their destination, removing one ambulance from each base, and computing the PH on the resulting state space.

References

Andersson T, Värbrand P (2006) Decision support tools for ambulance dispatch and relocation. *J. Oper. Res. Soc.* 58(2):195–201.

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

Bélanger V, Ruiz A, Soriano P (2015) Recent advances in emergency medical services management. Accessed September 25, 2017, https://www.cirrelt.ca/DocumentsTravail/CIRRELT-2015-28.pdf.

Bjarnason R, Tadepalli P, Fern A, Niedner C (2009). Simulation-based optimization of resource placement and emergency response. Accessed September 25, 2017, https://www.aaai.org/ocs/index.php/IAAI/IAAI09/paper/viewFile/255/1018.

Boers I (2015) Ambulances in-zicht 2014. Accessed September 25, 2017, https://www.ambulancezorg.nl/static/upload/raw/95d94429-86cf-4ab2-8763-8a0e663b5ba4/ambulances-in-zicht-2014.pdf.

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

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

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

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

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

Gendreau M, Laporte G, Semet F (2005) The maximal expected coverage relocation problem for emergency vehicles. *J. Oper. Res. Soc.* 57(1):22–28.

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

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

Larson RC (1974) A hypercube queuing model for facility location and redistricting in urban emergency services. *Comput. Oper. Res.* 1(1):67–95.

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

Mason AJ (2013) Simulation and real-time optimised relocation for improving ambulance operations. Denton B, ed. *Handbook of Healthcare Operations: Methods and Applications* (Springer Nature, New York), 289–317.

Maxwell MS, Henderson SG, Topaloglu H (2009) Ambulance redeployment: An approximate dynamic programming approach. *Proc. Winter Simulation Conf.* (IEEE, NY), 1850–1860.

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

Naoum-Sawaya J, Elhedhli S (2013) A stochastic optimization model for real-time ambulance redeployment. *Comput. Oper. Res* 40(8):1972–1978.

Richards DP (2007) Optimised ambulance redeployment strategies. Master's thesis, University of Auckland, New Zealand.

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

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

Toregas C, Swain R, ReVelle C, Bergman L (1971) The location of emergency service facilities. *Oper. Res.* 19(6):1363–1373.

van Barneveld TC (2016) The minimum expected penalty relocation problem for the computation of compliance tables for ambulance vehicles. *INFORMS J. Comput.* 28(2):370–384.

van Barneveld TC, Bhulai S, van der Mei RD (2016) The effect of ambulance relocations on the performance of ambulance service providers. *Eur. J. Oper. Res.* 252(1): 257–269.

Zhang L (2012) Simulation optimisation and Markov models for dynamic ambulance redeployment. Doctoral dissertation. The University of Auckland, NZ.