Relocation Algorithms
for Emergency Medical Services

Thije Christiaan van Barneveld
Relocation Algorithms for Emergency Medical Services
ISBN: 978-94-028-0457-7

The research described in this dissertation is supported by the Dutch Technology Foundation STW, which is part of the Netherlands Organisation for Scientific Research (NWO), and partly funded by the Ministry of Economic Affairs (project number 11986).

CWI

VU VRIJE UNIVERSITEIT AMSTERDAM

STW
connecting innovators

Printed by: Ipskamp Printing, Enschede, The Netherlands

Cover design: Ubel Smid | ubelsmid.nl (based on artwork by Jazzberry Blue)

VRIJE UNIVERSITEIT

Relocation Algorithms for Emergency Medical Services

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor aan
de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. V. Subramaniam,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de Faculteit der Exacte Wetenschappen
op woensdag 25 januari 2017 om 11.45 uur
in de aula van de universiteit,
De Boelelaan 1105

door

Thije Christiaan van Barneveld

geboren te Leiderdorp

promotoren:	prof.dr. R.D. van der Mei
	prof.dr. S. Bhulai

Voorwoord

Het leven is een reis, vaak weet je niet waarheen. Soms reis je met elkaar, en soms ook weer alleen. Het heeft waarschijnlijk een reden dat de eerste paar regels van dit gedichtje mij te binnen schieten bij het schrijven van dit voorwoord. Want wát een reis heb ik gemaakt. Een reis van vier jaar lang! Vier jaren van promotieonderzoek aan het Centrum Wiskunde en Informatica. Vier jaren van onzekerheid over waar je nou eigenlijk heen gaat. Vier jaren van dagelijks vallen en weer opstaan, werkend aan dit belangrijke bijproduct. Ik zeg bewust bijproduct, want meer dan een bewijs van het ondernemen van deze reis, is dit proefschrift niet. Veel meer is het leven van een promovendus een verhaal van persoonlijke groei, van doorzettingsvermogen, van continu werken aan jezelf. Een reis die je vaak alleen aflegt, maar gelukkig ook vaak met anderen: velen hebben mijn pad gekruist de afgelopen jaren. Ik dank eenieder van hen voor zijn of haar bijdrage aan deze reis. Maar dit proefschrift zou niet compleet zijn zonder een aantal van hen met naam en toenaam te noemen.

Sandjai, als ik de metafoor van promotietraject als reis voortzet, dan was jij mijn mentor, mijn acharya, mijn gids. Jij sleurde me door de diepste rivieren en over de hoogste bergen wanneer ik die steun nodig had. Jouw wijze lessen, of het nou gaat over het doen van onderzoek, over levensovertuigingen, of over Hindoestaanse mythologie, hebben mij telkens verder gebracht met het vinden van mijn eigen weg. En ondanks dat je het vaak vraagt, weet je denk ik best wel waarom ik 'nou weer lach', namelijk omdat ik blij ben je te zien. Rob, in de eerste plaats dank ik jou voor het feit dat jij het überhaupt voor mij mogelijk maakte om deze reis te beginnen, maar ook tijdens dit traject ben je voor mij van onschatbare waarde geweest. Van het veranderen van de scope van een artikel tot het ad hoc in elkaar draaien van een presentatie op een Oostenrijkse skipiste, jij stond altijd klaar om me te helpen. Je zult het vaker horen, maar jouw relativeringsvermogen, positieve instelling en enthousiasme zijn eigenschappen om te koesteren.

Goede herinneringen bewaar ik aan mijn medepromovendi van de REPRO-groep. Pieter, jouw vermogen om precies de juiste vragen te stellen, is voor een onderzoeker een waardevolle kwaliteit. Ook heb ik samen met jou het grappigste moment in mijn promotietraject beleefd, op die bewuste zomerdag in de Eeuwige Stad. Martin, jouw kennis over de wereld van de ambulancezorg is groot, en ik ben je er dankbaar voor dat ik vaak van die kennis gebruik mocht maken. Ook geldt dat voor je uitgebreide vaardigheden op het gebied van de IT. Caroline, jij was misschien wel mijn meest naaste collega, zowel in letterlijke als figuurlijke zin. Letterlijk, omdat we wekelijks een flink aantal uren tegenover elkaar doorbrachten als kamergenoten, en figuurlijk vanwege het feit dat het onderwerp van ons promotieonderzoek eigenlijk nauwelijks verschilde. Je gezelschap was voor mij best wel een gemis tijdens jouw tijd in Nieuw-Zeeland.

Binnen REPRO dank ik ook het RIVM en de verschillende ambulancediensten die bij het project betrokken waren, met name voor het delen van hun data. In het bijzonder wil ik André van Breukelen en alle andere medewerkers van de RAV Flevoland bedanken voor het bieden van de kans om op locatie onderzoek te doen. Dit heeft mij erg geholpen de praktische relevantie niet uit het oog te verliezen.

Daarnaast ben ik dank verschuldigd aan mijn promotiecommissie, bestaande uit Karen Aardal, Floske Spieksma, Geert-Jan van Houtum, Johann Hurink en Ger Koole. Ik dank hen niet alleen voor het lezen, begrijpen en becommentariëren van mijn proefschrift, maar ook voor hun bijdrages van andere aard, zowel voor als tijdens mijn promotietraject. Laatstgenoemde wil ik expliciet danken voor de genoten gastvrijheid binnen de OBP groep op de VU.

Maar misschien ben ik nog wel het meest dankbaar voor al die fantastische collega's van zowel de Stochastics groep op het CWI als de OBP groep op de VU. Het is met name jullie toe te schrijven dat ik het altijd erg naar mijn zin heb gehad tijdens de werkuren. Of het nou gaat om tafeltennissen, theepauzes of samen dineren tijdens een conferentie, ik mag mij van geluk prijzen jullie op mijn pad te hebben getroffen. In het bijzonder wil ik Sihan, Dirk en Bart bedanken voor de geweldige trip door de VS die we na afloop van een conferentie hebben gemaakt. Laatstgenoemde wil ik ook danken voor het prettige gezelschap tijdens de bijna dagelijkse tocht naar en van Amsterdam. Met jou als mijn carpoolmaatje had ik altijd een vrolijk begin en einde van de werkdag.

Nu begint mijn levensreis aan een nieuw hoofdstuk. Het is tijd om nieuwe wegen in te slaan. Waar deze mij zullen brengen? Dat is iets wat de toekomst uit zal wijzen. Wat ik wél weet, is dat mijn ouders, familie en vrienden mij onvoorwaardelijk zullen steunen, welk pad ik ook kies. Hiervoor wil ik jullie dan ook uit de grond van mijn hart bedanken. Met de wetenschap van jullie begeleiding, gezelschap en ondersteuning maak ik mij dan ook geen zorgen over de bestemming van mijn reis. Dit wetende moet het tweede zinnetje van bovenstaand gedichtje misschien wel aangepast worden: *Soms reis je alleen, maar meestal met elkaar...*

Thije van Barneveld
Amsterdam, augustus 2016

Contents

Voorwoord **v**

1 Introduction **1**

1.1 EMS Process 2

1.2 Ambulance Care in the Netherlands 3

1.3 Literature Review 7

1.4 Outline 17

I Online Approaches **21**

2 A Dynamic Ambulance Management Model for Rural Regions **23**

2.1 Introduction 23

2.2 Model 25

2.3 Heuristic Solution 37

2.4 Numerical Results 47

2.5 Concluding Remarks 56

3 The Penalty Heuristic and the Impact of Ambulance Relocations **59**

3.1 Introduction 59

3.2 Model 61

3.3 Penalty Heuristic 65

3.4 Experimental Setup 70

3.5 Numerical Results 73

3.6 Concluding Remarks 82

4 Insights into the Implementation of a Relocation Policy **85**

4.1 Introduction 85

4.2 Model 87

4.3 Algorithms and Features 87

4.4 Numerical Results 93

4.5 Concluding Remarks 106

II Offline Approaches 107

5 MEXPREP for Ambulance Compliance Tables 109
5.1 Introduction . . . 110
5.2 Model . . . 111
5.3 Assignment Problem . . . 117
5.4 Computational Study . . . 118
5.5 Concluding Remarks . . . 129

6 Compliance Tables for Two Vehicle Types 131
6.1 Introduction . . . 132
6.2 System Dynamics . . . 133
6.3 Model . . . 135
6.4 Computational Study . . . 144
6.5 Concluding Remarks . . . 157

7 A Unified View on the Online and Offline Approach 159
7.1 Introduction . . . 159
7.2 Methods . . . 160
7.3 Numerical Results . . . 161
7.4 Concluding Remarks . . . 169

List of Acronyms 171

Bibliography 173

Summary 183

Samenvatting 187

1

Introduction

It is generally believed that Dominique Jean Larrey (1766–1842) was the first to use the word 'ambulance' (Skandalakis et al., 2006). As a surgeon of Napoleon Bonaparte's Imperial Guard, he developed a plan for rapid evacuation of wounded soldiers from the battle field during combat using flexible medical units. The term ambulance was born. The first types of these units were pulled by horses and were used for the transportation of injured people from the battle field and for the provision of first aid. Nowadays, approximately 200 years later, ambulances have become common in our streets. Everybody knows what an ambulance is. However, few people are aware of the underlying processes that play a role in the planning of emergency medical services (EMS). Due to limited budgets and resources, efficient planning of ambulance services is crucial, in the medical as well as in the logistic domain.

This dissertation is concerned with the latter one. To be more specific, we regard ambulance repositioning as a tool to achieve cost-effective quality of emergency care without increasing the number of ambulances on duty. To that end, we consider the ambulance relocation problem in which units may be relocated to ensure that the ability to respond to emergencies quickly is maintained in periods of decreased resource availability, i.e., when ambulances become busy. In this context, short *response times*, i.e., the time between the moment the emergency request is reported and the arrival of the ambulance at the emergency scene, are of utmost importance. After all, providing medical aid quickly can make the difference between survival or death.

In many countries, governments use strict response-time targets. The fraction of highest emergency calls responded to within some *time threshold* is widely used as perhaps the most important quantitative performance indicator for the evaluation of ambulance service providers. Strongly related to this performance measure is the *coverage* concept. Coverage utilizes a time standard (also called coverage radius) for service delivery. All demand areas that can be reached by an ambulance within this threshold are considered to be covered. One may interpret this coverage as the 'preparedness' of the EMS system to respond to future calls, and therefore one may solve the ambulance relocation problem by relocating ambulances in such

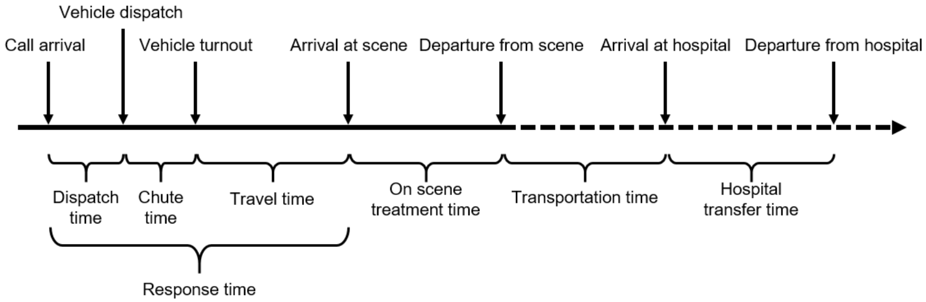


FIGURE 1.1: EMS process.

a way that an acceptable coverage level of the region is ensured.

1.1 EMS Process

The core of EMS operations is the EMS process. This process consists of several subsequent steps (see Figure 1.1).

When idle, ambulances have to wait for future requests at designated *waiting sites*. These are usually *base stations*: structures set aside for idle ambulances, although different types of waiting sites exist as well, e.g., parking lots where crews may be required to park up temporarily to increase the coverage of the region. Base stations often have a crew room and other facilities for the ambulance personnel. Ambulance staff may be summoned for emergencies by siren, radio, or pagers, depending on the station.

When the emergency services number (112 in Europe) is called after an incident has occurred, the call is answered by an emergency control center agent who assists the caller in first aid, inquires the condition of the patient (also called triage) and determines the level of urgency. Meanwhile, the dispatcher consults the dispatching system about which ambulance is most suitable to respond to the patient. To this end, most emergency control centers have access to modern technologies like a global positioning system (GPS) and computer-aided dispatch (CAD), which provide the agent a detailed overview of the current location and status of the ambulances and suggestions for dispatching, respectively.

After selecting an appropriate ambulance, the dispatcher informs the ambulance crew about the location, urgency and condition of the patient. The crew is usually present at a base station and departs for the emergency scene as soon as possible. It might also be that an idle ambulance is on the road, heading towards a base after the transportation of a patient, for instance. If this is the case, the crew is expected to reroute to the emergency scene immediately, without visiting a base station first. During the travel time to the patient, the ambulance has certain privileges: the crew can use emergency lanes, can turn on optical and sound signals to make other traffic aware, and it is allowed to exceed the maximum speed limit to achieve a faster response.

When the ambulance arrives at the emergency scene, the professional medical treatment can start. For this reason, most ambulances are equipped with technologies such as an automated external defibrillator (AED), an electro-cardiograph and respiration equipment, but also with a broad range of medicine to treat malfunctions of heart, lungs and blood vessels in an early stage. The crew, or at least one crew member, is fully qualified to work with this equipment. During the provision of first aid, the crew decides whether transportation of the patient to a hospital to receive specialized care not able to be carried out at the emergency scene, is necessary. The choice of the hospital usually depends on several factors, like the location of the emergency scene, preferences of the patient or hospital specializations. When the on-scene treatment has finished, the patient is placed on a stretcher and loaded into the ambulance.

During the transit, one crew member usually continues to provide appropriate medical care, if necessary. Meanwhile, the driver travels to the selected hospital as fast as possible. At the hospital, the ambulance crew unloads the patient and takes her/him to a suitable department, usually the emergency department or intensive care. After this drop-off, the crew informs the emergency control center that it has become idle. At that moment, the dispatcher assigns the ambulance to another task, or it tells the crew that it can travel to the base station the agent has selected. During the course of this procedure, the ambulance crew informs the dispatcher each time it changes status by pushing a designated button in the ambulance.

1.2 Ambulance Care in the Netherlands

In this dissertation, which is based on the research pursued as part of the Dutch REPRO project (From Reactive to Proactive Planning of Ambulance Services), we consider the Netherlands as our test bed. In this section, we describe how ambulance care in the Netherlands is organized.

The first law concerning ambulance care in the Netherlands was adopted in 1971. Up to then, EMS care was poorly organized in the Netherlands, which was painfully demonstrated in 1962 at the Harmelen train disaster: each town had its own emergency services number, ambulances were accommodated at local garages and the medical knowledge of the personnel was very limited. This resulted in extremely long response times, and hence, 93 deceased and 52 wounded people. From 1971 on, the “Wet Ambulancevervoer”[1] regulated the organization of EMS and its funding. This law was replaced by the “Tijdelijke Wet Ambulancezorg”[2] (temporary law on ambulance care), which is in effect between 2013 and 2018. In this law it is stated that in each of the 24 EMS regions in the Netherlands (see Figure 1.2) only one ambulance service provider is allowed to organize the EMS care, including the emergency control center. In addition, ambulance care may only be conducted on behalf of the emergency control center. Furthermore, this law includes a standard on accessibility: an ambulance must arrive at the

[1] Published online, `http://wetten.overheid.nl/BWBR0002757/2010-10-01` (in Dutch).

[2] Published online, `http://wetten.overheid.nl/BWBR0031557/2013-01-01` (in Dutch).

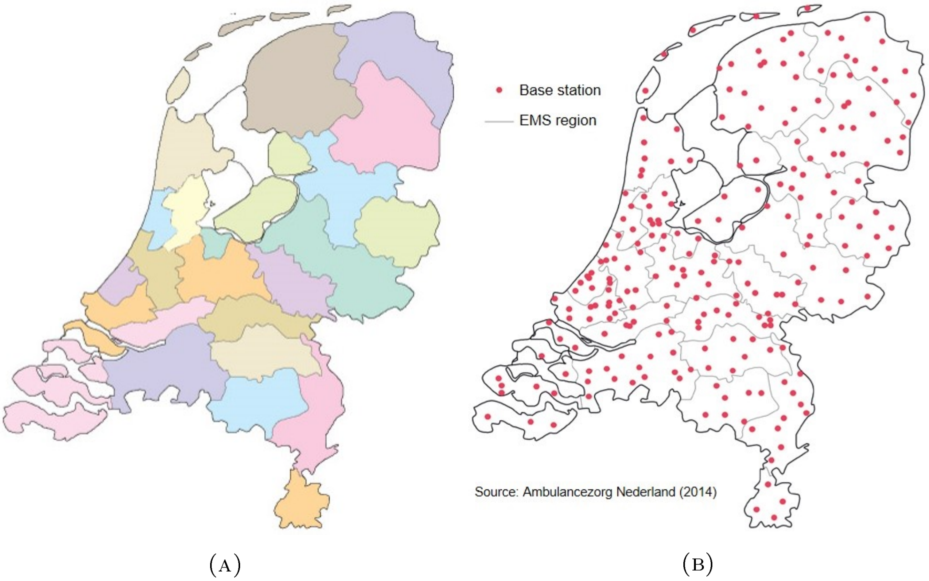


FIGURE 1.2: EMS regions and base stations in the Netherlands in 2016.

emergency scene within 15 minutes, starting at the moment the call is answered, in case of a life-threatening situation. This type of emergency is classified as an *A1-call*.

In the Netherlands, calls are classified according to three different call priorities. This categorization is assigned by the emergency call center agent. We already mentioned the life-threatening A1-calls, including calls for which a serious health risk for the patient exists as well. For A2-calls a fast response is desirable, but these are generally not life-threatening or serious health risk inducing. The optical and sound signals are usually turned off for this call priority. Dispatchers strive for a response time within 30 minutes to A2-calls, but this is not strictly enforced by law. In addition to the urgent A1- and A2-calls, ordered transport in the Netherlands has its own classification: B-calls. These are taxi-type calls for interfacility transport or transport from a patient's house to a hospital, or vice versa. A part of the calls of this type can be scheduled in advance, as the time between the call is made and the desired pick-up moment is long.

The fleet mix in the Netherlands is quite diverse. Each ambulance service provider chooses the type of response units it prefers to work with, in addition to the regular ambulance vehicles. For instance, some Dutch EMS regions use special ambulances for the B-calls. These vehicles contain less equipment than normal ambulances and are therefore not suitable for urgent response; they are only able to provide *Basic Life Support* (BLS). In contrast, regular ambulances can also provide *Advanced Life Support* (ALS). *Rapid Responder Ambulances* (RRAs) are used for fast first response to an emergency request. In the Netherlands, RRAs are usually cars or motor cycles, although bikes are used in the larger cities as

	2014	2013	2012	2011
Number of ambulances	755	744	725	711
Number of base stations	231	215	207	206
Total budget ambulance care (€)	500M	486M	439M	438M
Number of A1-calls	579,784	541,164	500,835	478,331
Mean response time A1-calls (m:s)	9:29	9:39	9:23	9:32
A1 responded to within 15 min. (%)	93.4	92.6	92.9	93.3
Number of A2-calls	288,924	274,907	273,692	263,257
Mean response time A2-calls (m:s)	14:56	15:26	15:15	15:25
A2 responded to within 30 min. (%)	96.7	96.1	96.3	96.0
Number of B-calls	321,612	328,709	325,892	342,838

TABLE 1.1: EMS statistics in the Netherlands.

well. These units are staffed by a highly educated person equipped with the same gear the regular ambulance personnel takes inside a patient's house, and they can provide ALS care. Basically, there are two differences between RRAs and regular ambulances: RRAs are faster, but they lack the ability to transport a patient to a hospital if necessary. Other unit types are the Mobile Intensive Care Unit (MICU), which is a truck used for the transport of intensive care patients, and the trauma helicopters, of which there are four in the Netherlands.

Table 1.1 shows some statistics about EMS operations in the Netherlands over the last years. These numbers are retrieved from the report Ambulancezorg Nederland (2014). Such a report is composed anually by the organization "Ambulancezorg Nederland", based on data provided by the RIVM (Rijksinstituut Volksgezondheid en Milieu; National Institute for Public Health and the Environment). The reference date is December 31 in each corresponding year. In approximately 75% of all calls the patient is transported to a hospital. These rides can be invoiced by the ambulance service provider at the health insurance company of the patient. In addition, 20% of the emergency patients do not need transportation. In these cases, the ambulance crew provides first aid but decides that the patient does not need to visit a hospital, possibly in consultation with the patient. For the remaining 5% of the calls the turnout of a medical response unit is unnecessary, i.e., upon arrival at the (supposed) emergency scene, there is no need for medical aid or transport.

Note that the demand for ambulance care has increased over the considered years. This is mainly due to the increase in A1-calls. Possible explanations for this growth are both the aging population and an increase in population in general. Moreover, the number of A1-calls as a fraction of the total demand has gradually increased from 44% in 2011 to 49% in 2014, but this is probably due to the decrease in the number of B-calls. If one considers the number of life-threatening A1-calls as a fraction of all urgent calls (A1 and A2), this percentage is around 65% for all years.

Recall that one of the goals of the Ministry of Public Health regarding ambulance care is to respond within 15 minutes to life-threatening calls in 95% of the cases. However, in none of the displayed years this percentage is achieved, although some ambulance service providers do. The response time to A1-calls consists of approximately 19% dispatch time (1:48 minutes), 10% chute time (0:56 minutes) and 71% driving time (6:41 minutes). For A2-calls this distribution is similar.

Ambulance care in the Netherlands is for a large part funded by the health care insurance companies. Table 1.1 shows an annual increase in the amount of money spent on ambulance care. Apparently, this increase in budget, and consequently, in the number of ambulances, is necessary to maintain the ability to offer top quality ambulance care. With the expected increase in demand for ambulance care in mind, and hence, the required resources (e.g., vehicles, base stations, personnel), it is important to use the current resources efficiently to ensure that the costs of ambulance care do not grow out of proportion in the future.

A highly promising development that is gaining momentum in the ambulance sector is the emergence of *Dynamic Ambulance Management* (DAM). The basic idea of DAM is that ambulance vehicles are proactively relocated to achieve a good coverage of the EMS region in real time. Throughout this dissertation, we consider models and methods for DAM, based on the Dutch EMS setting. Next, we describe three key characteristics of EMS operations in the Netherlands.

Number of Waiting Sites

It tends to be more and more common in the US and Canada to park up (temporarily) at a street corner or other strategic hotspot. However, this is not the case in the Netherlands yet. The number of potential waiting sites typically exceeds the number of ambulances on duty. As a consequence, multiple ambulances, and hence, crews, are usually present at each base station, especially during peak hours. Note that this does not contribute to the coverage level of the region; after all, each of the ambulances present at the same location has the same coverage radius. This concept of coverage is referred to as *single coverage*: an area is said to be covered if and only if at least one ambulance can reach that area within the time threshold. A more elaborate notion of coverage is *probabilistic coverage*: this notion of coverage takes into account the fact that ambulances might be busy, and hence, not available. Therefore, instead of an area being covered (1) or not (0) the coverage level of a certain area takes fractional values depending on the number of ambulances present within the coverage radius. Hence, positioning multiple ambulances at the same location may be beneficial; not for the single, but for the probabilistic coverage level. We will discuss some single and probabilistic coverage models at a later stage.

Repositioning Idle Vehicles

In some countries it is prohibited by law to reposition idle ambulances, apart from sending them back to a waiting site, e.g., in Austria (Schmid, 2012). In the Dutch EMS system this is *not* the case: dispatchers are allowed to relocate idle ambulances, even between base stations. However, there are some restrictions concerning repositioning. The “Arbeidstijdenwet” [3] (Working Hours Act) does not allow that ambulance crews are too long or too frequently away from their home base station, either for service of a patient or due to repositioning. This holds especially for long shifts with more than nine working hours. Furthermore, if ambulance crews spend too much time on the road due to frequent relocations, the ambulance service provider will probably be condemned by an Occupational Safety and Health organization, which regulates the enforcement of the “Arbeidsomstandighedenwet” [4] (law on working conditions). Therefore, to keep the personnel motivated, the number of relocations and relocation time must be kept at a minimum.

Hospital Transfer Times

In the Netherlands, the hospital transfer times are relatively short compared to other countries, especially to North America (Carter et al., 2015). Usually, no crowding takes place at the emergency department in the hospital. In practice, an ambulance that is busy with the drop-off of a patient for already ten minutes is considered as being idle in the emergency control center. Hence, it can be assigned to a new task. This avoids that the ambulance personnel spends too much time in the hospital.

1.3 Literature Review

Nowadays, the literature on EMS planning in the field of Operations Research (OR) and Management Science (MS) is quite rich, although this subfield is relatively young: to the best of our knowledge, the first paper on ambulance planning was published in 1969 by Savas (1969). This work describes a computer simulation used to analyze the possible improvements in ambulance services that would result from proposed changes in the number and location of ambulances for New York City. The author highlighted that this was the first time that computer simulation was utilized to aid decision-making in the city of New York. After this pioneering publication on EMS planning, many would follow. Not surprisingly, this is due to the wide variety of problems that occur in the planning of ambulance services, most of them devoted to optimally locating medical units. In this literature review, we will focus mostly on these types of problems, models and methods as these are the most relevant for this dissertation. However, we will address other EMS problems not related to ambulance positioning as well.

The literature on ambulance location can roughly be divided into two categories: problems, models and methods devoted to (1) *static location*, and to (2) *dynamic relocation* of ambulances. The key difference between papers from both categories is the way decisions are made, either in *non-real-time* or in *real-time*. Therefore, static location models are of a strategic and tactical nature, while dy-

[3] Published online, `http://wetten.overheid.nl/BWBR0007671/2016-01-01` (in Dutch).

[4] Published online, `http://wetten.overheid.nl/BWBR0010346/2016-01-01` (in Dutch).

namic relocation is done in an operational fashion in general. Despite the fact that this dissertation is primarily devoted to relocation, we review the most relevant literature on static location planning of ambulance services as well. After all, static location models form the basis of many relocation models. For comprehensive surveys on static location models, we refer to ReVelle (1989), Owen and Daskin (1998), Brotcorne et al. (2003), Green and Kolesar (2004), Goldberg (2004), Li et al. (2011), and Bélanger et al. (2015).

1.3.1 Static Location

In the literature on static location models, one can distinguish two main subcategories: (1) deterministic location models, and (2) probabilistic location models. Başar et al. (2012) present a more comprehensive taxonomy for ambulance location models. Deterministic coverage models assume that a medical unit is always available if an emergency request arrives. However, ambulance availability is not always ensured, since ambulances get busy due to the response to patients in reality. Probabilistic models take this unavailability into account: to ensure a high probability of having at least one unit available nearby, the number of ambulances that can respond quickly a certain area is of importance.

Deterministic Location Models

In the earliest deterministic location models, the concept of single coverage plays an important role as this notion of coverage is perhaps the most intuitive one due to its 0-1 nature: an area is either covered or not, depending on whether an ambulance is positioned nearby. The first deterministic location model was the location set covering problem (LSCP) proposed by Toregas et al. (1971). This model aims to find the minimum number of ambulances needed to cover all demand areas. The LSCP is formulated as binary integer program and it decides on both the number of ambulances needed and their location. However, in the LSCP no distinction in importance of demand areas is present. An LSCP solution ensures total coverage of the region, although it might be the case that there is no need, and no budget, to cover each demand area, as some of them may be sparsely populated. For this reason, Church and ReVelle (1974) proposed the famous maximum coverage location problem (MCLP), formulated as binary linear program. This model aims to maximize the fraction of the population covered given a certain fleet size, and it optimizes the location of the ambulances. In addition to the problem formulation, the authors of this work provide a heuristic approach to solve the MCLP, which was quite a challenge in those days.

Despite its simplicity, the MCLP has deserved a lot of attention both in practice and in theory. Much has been published about solution techniques for MCLP, including a Lagrangean heuristic (Galvão and ReVelle, 1996), a decomposition heuristic (Pereira et al., 2010) and, more recently, a swap local search algorithm (Kerkkamp and Aardal, 2016). Moreover, the MCLP is frequently used as a basis for more sophisticated and realistic facility location models. For instance, the tandem equipment allocation model (TEAM) and facility-location equipmentemplacement technique (FLEET), proposed by Schilling et al. (1979), are both extensions of the MCLP. Although the authors focus on the location of two types of fire fighter equipment, the model is also appropriate for ambulance location in an EMS system with multiple types of medical response units. After all, differentiation in ambulance vehicle types exists as well, for instance, in the level of care they can provide: either Advanced (ALS) or Basic Life Support (BLS). Charnes and Storbeck (1980) use this classification of medical response units, and they develop a goal programming model, incorporating two types of demand as well.

A location model related to the MCLP, which deserves attention here, is the p-median problem, formulated by ReVelle and Swain (1970). This model selects locations, for instance, for ambulances, according to a different criterion than coverage. Instead, the focus is on minimization of the weighted average response time. Although the p-median problem is somewhat older, one could regard this as a generalization of the MCLP. Distances in an instance of the p-median problem can be modified to binary values depending on whether a facility is within the coverage radius for a certain demand area or not. Solving this p-median problem is equivalent to solving the MCLP. However, the MCLP is a faster model to solve due to the more complex nature of the p-median problem. After all, in the p-median problem one considers the distance of each of the possible facilities to a certain demand area, while it suffices in the MCLP to consider the set of possible locations which are within range of this area, reducing the number of variables.

A similarity between MCLP and the p-median problem is that for each demand area only the closest ambulance is of influence on the objective of the model. The other ambulances are treated as nonexistent ones for a particular demand area. In other words, both the MCLP and the p-median problem assume that always the closest ambulance responds to a call, although it might be unavailable. After all, an ambulance may not be able to respond to an emergency request if the time between two successive calls occurring in the same area is short.

For the abovementioned reason, Daskin and Stern (1981) considered *multiple coverage*: a certain area is covered if a predefined number of ambulances is present within the coverage radius. The authors incorporated a hierarchical objective to maximize the number of demand points covered more than once. Other well-known multiple coverage models are the backup coverage models (BACOP1 and BACOP2), formulated by Hogan and ReVelle (1986). Both models are extensions of the MCLP and maximize the demand covered twice. BACOP2 is a generalization of BACOP1 in the sense that one can balance single and double coverage in BACOP2. The last multiple coverage model we will discuss is the double standard model (DSM) by Gendreau et al. (1997). A novel ingredient in this model is the introduction of two different time thresholds. The DSM requires all demand to be covered within the least strict threshold, while a certain fraction of demand must be covered within the most tight threshold. Then, the DSM maximizes the demand covered twice within the most tight time threshold. The model is solved by Tabu Search. The ambulance location plan of the DSM is applied in several countries, including Austria (Doerner et al., 2005), Belgium and Canada, as reported by Laporte et al. (2009). Moreover, this model forms the basis for one of the first relocation models, proposed by the same authors (Gendreau et al., 2001).

Probabilistic Location Models

Although multiple coverage models address a crucial shortcoming of single coverage models, namely, they extend the 0-1 coverage to 0-1-2-... coverage, ambulance unavailability is not modelled explicitly. This drawback of multiple coverage was addressed in the early 80s by the introduction of the so-called *busy probability* or *busy fraction*: the fraction of time a single ambulance is busy and hence not dispatchable to an incoming emergency request.

This innovation induced a shift from deterministic to probabilistic, or expected, coverage. One of the first probabilistic coverage models, the maximum expected location problem (MEXCLP), was proposed by Daskin (1982, 1983). This model, formulated as an integer linear program, is an extension of the MCLP. The objective of MEXCLP is akin to that of MCLP: maximization of the (expected) coverage. However, due to the rational values the busy fraction may take, the coverage of a certain area takes fractional values in contrast to single and multiple coverage. In the MEXCLP formulation, the busy fraction is assumed to be known. Moreover, the same busy fraction is used for each ambulance, regardless of its location. A heuristic solution was presented by Daskin (1983) to solve the MEXCLP. An alternative non-linear formulation of the MEXCLP was presented by Saydam and McKnew (1985). Other early well-known probabilistic coverage location models worth mentioning are the maximum availability location problems (MALP I and MALP II) by ReVelle and Hogan (1989). These models, relaxing the assumption that the busy fraction is the same for each base station, maximize the demand covered with a given probability α. Galvão et al. (2005) present a unified view of the MEXCLP and the MALP.

The simple yet powerful concept of busy fraction unleashed a breakthrough in ambulance location models, and the two mentioned papers by Daskin are among the most cited ones in the literature on ambulance location and relocation models. Moreover, the MEXCLP model serves, both directly and indirectly, as the basis for many extensions and modifications, both in the literature on static location and on dynamic relocation. However, Batta et al. (1989) state some simplifying assumptions concerning busy fractions of the MEXCLP: ambulances operate independently, ambulances have the same busy fraction and busy fractions are invariant with respect to the ambulance locations. To address these issues, Batta et al. (1989) used the celebrated Hypercube model developed by Larson (1974) to compute performance measures regarding a given ambulance location plan, e.g., busy fractions. This model was used to compute the expected coverage in a single node substitution heuristic. Moreover, "correction factors" for computing the probability that the j^{th} selected ambulance is the first available one, computed by Larson (1975), are embedded in the MEXCLP formulation to obtain an adjusted version: AMEXCLP. From then on, many probabilistic static ambulance location models used Hypercube models for estimating EMS system performance characteristics. As a consequence, the Hypercube model has been extended multiple times to take more realistic features into account (Jarvis, 1985; Budge et al., 2009).

Over the years, several interesting features in ambulance location models have emerged. We list some of these in the remainder of this subsection. In addition to the uncertainty related to ambulance availability, some papers on the static location problem consider EMS vehicle travel times to be stochastic. To this extent, a coverage probability is used in existing models, e.g., the MCLP (Karasakal and Karasakal, 2004), the MALP (Marianov and ReVelle, 1996) or the MEXCLP (Goldberg et al., 1990; Ingolfsson et al., 2008; van den Berg et al., 2014). Some papers also focus on the estimation of ambulance travel times and, hence, coverage probabilities, e.g., Budge et al. (2010) and Westgate et al. (2013, 2016). Erkut et al. (2009) perform a computational comparison between five versions of the MCLP and the MEXCLP in which in some probabilistic response times and station-specific busy fractions are incorporated. They conclude that models that incorporate this type of uncertainty yield coverage estimates.

Vehicle Types

Differentiation in vehicle type is another interesting aspect in the literature on static probabilistic ambulance location, although this was first done in a fire fighter setting (Marianov and ReVelle, 1992) before EMS systems became of interest (Jayaraman and Srivastava, 1995). Concerning this stream of literature, almost all models with multiple unit types make a distinction in the level of care an ambulance can provide: either Advanced (ALS) or Basic Life Support (BLS), and ambulances are classified as such. For instance, Mandell (1998) considers a two-tiered model (TTM) with two types of vehicles, ALS and BLS, and two response time standards. The objective is to maximize expected coverage, based on the number of ALS vehicles that cover a certain demand area within the tightest and least strict response time threshold and the number of BLS vehicles within the least strict threshold.

Marianov and Serra (2001) also consider two vehicle types. They present two models, extensions of the LSCP and the MCLP, and require that a demand point is covered if both types of ambulances are within prescribed response time thresholds and the patient does not queue with more than a prespecified number of other patients due to congestion. In addition to two vehicle types, ALS units for first response and BLS units for transportation, call urgencies are considered by McLay (2009). She proposes an extension of MEXCLP for two types of ambulances (called MEXCLP2) that locates both types of units maximizing the total number of expected highest priority calls covered within the coverage radius, bearing in mind that units may become busy due to patients of less urgency type.

Performance Measures

A part of the literature on static location models also focuses on different response-time related performance measures than the commonly used concept of coverage. We already mentioned the p-median problem that minimizes the weighted average response time, but more sophisticated models exist as well. For instance, Rajagopalan and Saydam (2009) present two variants of a model named the minimum expected response location problem (MERLP), which are both extensions of the classic p-median problem.

Erkut et al. (2008) openly question the use of coverage models in ambulance location due to their limited ability to discriminate between different response times. Instead, they advocate to relate the response time of an EMS vehicle to a patient to the survival probability of the patient. To this end, Erkut et al. (2008) studied published research in the medical domain related to survival rates and found that almost all this literature focuses on survival after a cardiac arrest. They also formulated the maximum survival location problem (MSLP) and maximum expected survival location problem (MEXSLP). These are extensions to the MCLP and the MEXCLP, respectively, in the sense that survival can be incorporated, and the authors considered several of such survival functions. One of these, the one by Larsen et al. (1993), was used by McLay and Mayorga (2010) in a model to evaluate different response time thresholds in terms of their resulting patient survival rates.

In addition, Knight et al. (2012) present an important extension of the work by Erkut et al. (2008) by permitting multiple survival functions in order to accommodate heterogeneous patient classes and reflect different outcome measures within the population served by the EMS. The Maximal Expected Survival Location Model for Heterogeneous Patients (MESLMHP) they propose aims to maximize the overall expected survival probability of multiple-classes of patients.

Preplanned Redeployment

None of the abovementioned models take variations over time in input parameters into account, e.g., time variations in demand, travel times, busy fractions or fleet size. To address these issues, part of the literature on ambulance location incorporates time variation and computes location plans for multiple time-periods. At prespecified moments in time, vehicles are redeployed. Although this class of models could also be classified as relocation models, we do not, since this type of redeployment is preplanned and happens at times known a priori, in contrast to dynamic relocation.

One of the first to incorporate variations in demand patterns and fleet size over time were Repede and Bernardo (1994) by extending the MEXCLP to a time-dependent variant: TIMEXCLP. Van den Berg and Aardal (2015) added an extra dimension to this model in the sense that costs are induced by the redeployment of ambulances and by opening base stations between two different time periods. They intend to balance coverage and costs, taking variations in travel times throughout the day into consideration as well. The latter was also done by Schmid and Doerner (2010), who proposed a multi-period version of the DSM. The DSM is also an important ingredient in the work done by Başar et al. (2011), who combined the DSM with BACOP to take time dependency of input parameters into account. Other models on preplanned ambulance redeployment include the dynamic available coverage location model by Rajagopalan et al. (2008), its extension by Saydam et al. (2013), and the model by Degel et al. (2015), which bases the preplanned location plan on a empirically determined required coverage-level.

1.3.2 Dynamic Relocation

The literature on dynamic relocation of ambulances is less comprehensive than the literature on the static location problem. To our knowledge, the first known dynamic relocation model in the area of emergency logistics was proposed by Kolesar and Walker (1974). Their work describes a computer-based method for the relocation of outside fire companies when all of the urban ones are engaged in fighting fires in New York City. The authors provide a mathematical programming formulation of their problem and solve it via a heuristic algorithm. Some years later, Berman (1981a,b,c) was the first to consider the dynamic ambulance relocation problem. The author provided an exact dynamic programming approach to the ambulance relocation problem, although his formulation was tractable only in an oversimplified version of the problem.

Two decades after Berman (1981a,b,c) published his work, dynamic relocation of ambulances became of interest to the EMS planning community. The reason that this took so long is probably explained by the complexity of the problem. As stated by Brotcorne et al. (2003), the ambulance relocation problem is difficult to solve since solutions have to be generated at very short notice. With the development of advanced computer technologies, tackling the dynamic ambulance relocation problem in a realistic setting became possible. Gendreau et al. (2001) were the first to propose a model with this purpose: they extended their DSM formulation to a dynamic version: redeployment problem at time t (RP^t). In this model, practical considerations regarding the frequency and length of relocations are taken into account: excessively long or repeated round trips between the same two stations are penalized in the objective function, in addition to maximizing the double coverage. Each time an emergency request is reported, a solution to the RP^t is computed using a tabu search heuristic, taking into account specific information about the state of the EMS system. To be more specific, the history regarding relocations per ambulance is captured by the RP^t. The island of Montreal (Canada) was used as test bed for the proposed model.

Ambulance relocation models and methods can be classified according to the amount of computational work carried out in real-time and a priori. If most of the computations are done beforehand, we speak of an *offline* approach. However, the RP^t mentioned above is an example of the *online* approach: most work is done in a real-time fashion, i.e., when a decision moment occurs. Therefore, online methods can handle very detailed information about the current state of the EMS system. In contrast, offline methods store computed relocation decisions for each possible state a priori. If the system is in a certain state, the corresponding relocation decision is retrieved or computed very fast and applied immediately. To keep the number of states manageable, typically a low-level state-space description is used in the offline approach. In the remainder of this subsection, we provide an overview of both online and offline relocation models and methods.

Online Approaches

In the early years of this millennium, solving the RP^t exactly within a short period of time was not possible due to the lack of computational power. That is why Gendreau et al. (2001) resorted to a tabu search heuristic and parallel computing. One decade later, ILP-solvers could easily handle the RP^t and this problem regained interest from Moeini et al. (2014). They formulate the dynamic relocation problem (DRP^t) by slightly changing the objective function of the RP^t into one in which the double coverage of some demand nodes is given more importance than that of others. The authors have tested and verified the model on data sets belonging to the county of Val-de-Marne, France. Moreover, they performed numerical simulations which show improvement in coverage levels if their model is used instead of the original RP^t.

Another online relocation model that is similar to the RP^t is presented by Mason (2013). This real-time multi-view generalized-cover repositioning model (Rt-MvGcRM) is solved every time a relocation decision is desired. Like in the online relocation models earlier mentioned, ambulance crew unfriendly actions, e.g., moving idle ambulances, redirecting en-route vehicles, are penalized. Furthermore, all input parameters are assumed to depend on the vehicle positions, call arrival rates and road speeds at the time the model is solved. Unfortunately, Mason (2013) does not provide the solution technique used for solving this model. This is probably due to the fact that this model is implemented in the commercial EMS Management software Optima Live, used to aid ambulance dispatchers in real-time relocation decisions and developed by the Optima Corporation. Other work supported by this corporation is presented by Richards (2007) and Zhang (2012).

Andersson and Värbrand (2007) use a performance measure that differs from the previously mentioned models. Instead of focusing on coverage, they define a quantifiable measure for preparedness, which evaluates the ability to serve potential patients with ambulances now and in the future. The preparedness of a certain demand area increases if an ambulance is moved towards that area. If the preparedness for one or more demand points drops below some level, a decision moment occurs. That is, a model, called DYNAROC, is solved using a tree-search heuristic. This model aims to minimize the maximum travel time for any of the relocated ambulances in order to ensure a certain preparedness level for each demand node. To this extent, ambulances can park up in each demand area. The authors simulate the EMS system of Stockholm (Sweden) using the DYNAROC algorithm and conclude that a high level of preparedness is helpful in reaching the response time target set by the authorities.

The last online ambulance relocation model we want to discuss here is the dynamic MEXCLP (DMEXCLP) proposed by Jagtenberg et al. (2015). When an ambulance becomes available after serving a patient, a new destination for this ambulance is decided by determining the relocation that maximizes the coverage of the region. Since it shares the same coverage concept with the MEXCLP, one can regard this method as its dynamic counterpart version. The DMEXCLP computes relocation decisions very fast. After all, the number of possible moves is bounded by the number of waiting sites, and hence, the computation can be done by brute-force. The authors compare their method to the static policy in which each ambulance always returns to its home station, for the EMS region of Utrecht in the Netherlands. They show that the DMEXCLP easily outperforms the static policy on the fraction of late arrivals.

Offline Approaches

As stated before, offline methods generally use little information on the state of the EMS system. A state description popular in both research and practice is by the number of available units: every time this number changes, due to the assignment of an ambulance to a request or when a vehicle becomes idle again, the corresponding location plan is applied. These location plans are usually summarized in a table, the so-called *compliance table*. Gendreau et al. (2006) were the first to our knowledge to conduct research on this type of policy, although their study was motivated by the problem of relocating physician cars, instead of ambulances, in the EMS region of Montreal (Canada). They formulate the maximum expected coverage relocation problem (MECRP) as an integer linear program, which is an extension of the MCLP. This model computes the desired distribution of ambulances throughout the region (called *ambulance configuration* in the remainder) for each state of the system. The states are weighted according to the expected steady-state probabilities. Moreover, the number of vehicles that is required to change location is restricted in the MECRP.

The MECRP does not specify the movement of ambulances among stations and from hospitals to stations, only the ambulance location plans. Gendreau et al. (2006) suggest that a transportation model can be applied to determine this. This observation inspired Maleki et al. (2014) to propose two assignment problems for the actual assignment of ambulances to waiting sites, when the desired configuration is known. These models, the generalized ambulance assignment problem (GAAP) and generalized ambulance bottleneck assignment problem (GABAP), are offline approaches in which these assignments can be computed in advance for every possible state transition and ambulance configuration. The models differ in the sense that the GAAP minimizes the total travel time of the ambulances that move between two configurations, while the GABAP focuses on minimization of the longest travel time of an ambulance, and hence, the time until the system is in compliance. The authors tested the MECRP, GAAP, and GABAP on data obtained from the EMS region of Isfahan (Iran).

Sudtachat et al. (2016) propose another compliance table model. To be more specific. They consider a special class: *nested compliance tables*, which restrict the number of relocations that can occur simultaneously. The foundation to this work is the paper by Alanis et al. (2013), who propose and analyze a tractable two-dimensional Markov model of an EMS system that repositions ambulances using a compliance table policy. This model has the same data requirements and can produce the same output as the Hypercube model, but it also takes relocations into account. Furthermore, the authors develop procedures to estimate the parameters needed in the model and they show that outcomes of the Markov model serve as a good approximation to several performance measures obtained by simulation. The computed steady-state probabilities serve as input for the integer linear program of Sudtachat et al. (2016). The authors demonstrate the efficiency of their nested-compliance table policy compared to the static policy induced by the AMEXCLP based on data collected from an EMS department in Hanover County, Virginia, on several performance indicators.

Offline approaches that require solving an integer linear program in advance, but not related to compliance tables, are the topic of both Nair and Miller-Hooks (2009) and Naoum-Sawaya and Elhedhli (2013), although the first did not present their solution technique. The proposed models are multi-objective in the sense that they aim to balance both patient and cost-related criteria. The resulting location plans are applied to the Canadian EMS regions of Montreal and Waterloo, respectively.

The last class of offline models differs from the integer linear programming models treated above. Maxwell et al. (2010) efficiently apply approximate dynamic programming (ADP) for redeployment of ambulances that finish service of a patient. The authors use an elaborate state space description, especially compared to policy structures with low detail about the state of the system, like compliance tables. The problem is formulated as a dynamic program. Using basis functions that keep essential information about the state of the EMS system, e.g., the uncovered and missed call rate now and in the future, they parameterize the value function to obtain an approximation. The authors use least squares regression within an approximate policy iteration procedure to tune these parameters. The policy evaluation within this procedure is done through simulation, which is computationally heavy. However, if a good parameterization of the value function is obtained, it takes very short time (less than one second in their case study) to compute the relocation decision. In another paper, the authors show how to use direct search methods to tune the parameters in a value function approximation (Maxwell et al., 2013). Moreover, they construct a lower bound on the long-run fraction of late arrivals that holds for nearly any ambulance redeployment policy, involving the solution of integer linear programs and simulation of multiserver queues (Maxwell et al., 2014).

Schmid (2012) also uses ADP to solve the ambulance relocation problem. In her model, relocation decisions can be made when a busy ambulance becomes available again, similar to the model of Maxwell et al. (2010). This is a direct consequence of the fact that in the region of interest in her case study (Vienna, Austria) repositioning of idle ambulances is not allowed, apart from sending them back to a base station after a service completion. However, the same model is used for the dispatching decision, so two different events trigger a decision. The objective is to minimize the average response time, in contrast to all previously mentioned offline approaches in which coverage is the patient-related performance criterion of interest. Schmid (2012) also incorporates time-dependent parameters, e.g., travel times and call arrival rates, in her model.

1.3.3 Other Topics

In addition to the literature on positioning ambulances, either in real-time or in non-real-time, many papers focus on other issues present in planning EMS services. For instance, prediction of ambulance call volumes for different urgencies has received considerable attention, e.g., by Channouf et al. (2007); Setzler et al. (2009); Matteson et al. (2011)). Other topics in the EMS literature include scheduling ambulance crews (Erdoğan et al., 2009); the vehicle mix decision (Chong et al., 2015); scheduling ordered patient transportation (Van den Berg, 2016); and EMS district design (Mayorga et al., 2013; Ansari et al., 2015).

Dispatching

Another interesting topic in the EMS literature, closely related to the relocation problem in the sense that both are at the operational level, is the dispatching problem. The most common rule is to send the closest vehicle to an incident, but several papers question whether this is optimal. For instance, Lee (2011) investigates the ambulance dispatching algorithm proposed by Andersson and Värbrand (2007), and finds that dispatching the closest vehicle yields the lowest average response time. However, Jagtenberg et al. (2016) present a Markov Decision Problem (MDP) and a heuristic to solve the dispatching problem, and show that the number of calls not responded to within the response time threshold can be greatly reduced. Bandara et al. (2012) show something similar: they compute dispatch policies for different urgencies that maximize patient survival probabilities by using an MDP model (Bandara et al., 2012) and simulation (Bandara et al., 2014). An MDP is also used by McLay and Mayorga (2013b), who compare optimal dispatching policies under different strategies regarding the classification of patient priorities. Various publications on this subject are (co-)authored by McLay and Mayorga, e.g., they present a dispatching model that balances efficiency and equity, the latter both from a patient as well as a crew perspective (McLay and Mayorga, 2013a), they propose a model that integrates the location and dispatching decisions (Toro-Díaz et al., 2013, 2014) and they consider dispatching vehicles under multitiered response (Sudtachat et al., 2014). A dispatch model based on the MCLP is presented by Lim et al. (2011).

Simulation

At last, much has been published about simulation of EMS systems. After all, simulation is a powerful tool to support decision making as changes in policy can be evaluated without influencing practice (what-if scenarios). It is possible that policies yielding good theoretical results perform worse in practice compared to ones with inferior theoretical results, and vice versa. Therefore, simulation is a necessary tool in the design and evaluation of policies. Aboueljinane et al. (2013) provide an extensive review on nineteen EMS simulation models. The authors classify these models according to the types of decisions they are used for (e.g., relocation, shift scheduling), the performance measures of consideration, demand related data and dispatching rules. We refer to this work for a comprehensive overview.

1.4 Outline

In the following chapters we present several methods and models for solving the ambulance relocation problem. In all chapters the Dutch EMS setting as explained in Section 1.2 is considered, and results are based on a case study of EMS regions in the Netherlands. Chapters 2, 3, and 4 are concerned with the online approach, and Chapters 5 and 6 describe two offline models.

In Chapter 2, we develop an MDP formulation for the ambulance relocation (and dispatching) problem so as to maximize a measure of system-wide response-time performance. The formulation discretizes time and discretizes the transportation network into arcs with travel times of one time unit. We solve the formulation heuristically, using a one-step look-ahead method. This heuristic is based on the enumeration of possible actions and on selecting the one providing the best metric value over a set of scenarios. We focus on rural EMS regions, which are generally different from the urban EMS regions due to the smaller number of events, smaller number of ambulances, higher fluctuation of demands and smaller coverage provided by ambulances when traveling between two high-demand areas. We test the formulation and heuristic using data from Flevoland, a rural EMS region in the Netherlands. The performance of the heuristic solution is compared to compliance table policies. Chapter 2 is based on Van Barneveld et al. (2015).

In Chapter 3, we focus on the trade-off between two conflicting criteria in the ambulance relocation problem: timely response to emergency requests and workload of the crew. Proactive ambulance relocations are an effective tool in reducing response times, but are tiresome for the crews as they have to deal with increasing workloads. Therefore, it is of great interest to determine the marginal benefits of additional moves. For this purpose, we develop a penalty heuristic for solving the ambulance relocation problem. A penalty function, which is a function of the response time, is used to compute the expected impact of an ambulance relocation on a system-wide performance measure. A change in the ambulance location plan may only take place if it induces a substantial gain in the ability to respond to emergency requests timely. We test different thresholds and study how these impact the system-wide performance measure, which can be arbitrarily chosen through the choice of penalty functions. Moreover, we study the effect of changing the number of ambulances that may be relocated simultaneously. We simulate a real-life data set of two Dutch EMS regions, the rural region of Flevoland and the urban Amsterdam region, divided in day and night scenarios and we consider fleet sizes. Chapter 3 is based on Van Barneveld et al. (2016a).

In Chapter 4, we combine the penalty heuristic explained in Chapter 3 and the DMEXCLP method developed by Jagtenberg et al. (2015). The two methods are similar, but differ in some interesting aspects: the notion of coverage, the performance criterion and the inclusion of busy ambulances in the state description of the EMS system. We study the impact of these features on several EMS performance indicators. In that sense, the work presented in this chapter could be regarded as a search for the 'best of both worlds' combination of DMEXCLP and the penalty heuristic from a practical point of view. In addition, we consider the influence of the frequency of redeployment decision moments, chain relocations, and relocation time bounds on the EMS crew workload. As we aim to obtain insights which are robust with respect to the characteristics of the EMS region, we include case studies for the two different types of regions mentioned above. We carry out simulations of the developed class of relocation strategies to test the effect of the mentioned aspects and features. Chapter 4 is based on Van Barneveld et al. (2016b).

In Chapter 5, we shift focus to the offline approach of solving the ambulance relocation problem. We present an integer linear programming model, the minimum expected penalty relocation problem (MEXPREP), that extends the MECRP of Gendreau et al. (2006), to obtain compliance tables for ambulance relocation. The new model removes capacity limitations for base locations and incorporates the possibility that an ambulance that should be available according to the compliance table is not available, using an approach that is borrowed from the MEXCLP of Daskin (1983). A computational study compares the MEXPREP to the MECRP and to a static solution in which each ambulance returns to its home station after a task has been performed. Moreover, based on the EMS region of Amsterdam, we investigate the impact of *relocation thresholds*. If the number of available ambulances is below this threshold, no relocation takes place. In addition, we compare two methods for assigning ambulances to bases in order to reach compliance by simulation. This chapter is based on Van Barneveld (2016).

The computation of ambulance compliance tables is the topic of Chapter 6 as well. A crucial difference with the previous chapter is the inhomogeneity of the fleet: we consider an EMS system with both rapid responder ambulances (RRAs) and regular transport ambulances (RTAs). The key difference between both types of units is that RRAs are faster, but they lack the ability to transport a patient. Therefore, if transportation is required, a subsequent dispatch of an RTA has to be carried out. An EMS system with two types of ambulances brings forth additional complexity to the compliance table problem, as now a two-dimensional state description is needed, and hence, a two-dimensional compliance table. In this chapter, we present an integer linear program to compute such two-dimensional compliance tables, based on the MEXCLP2 of McLay (2009). In this program, we incorporate two interesting constraints: we force some degree of nestedness in the compliance table, and we restrict the maximum trip length a unit may have to carry out. We apply the two-dimensional compliance tables to the EMS region of Flevoland in a discrete-event simulation to obtain practically relevant results and insights. Chapter 6 is based on Van Barneveld et al. (2017).

This thesis is concluded by Chapter 7 in which we present a unified view on the online and offline approach of the ambulance relocation problem. To that end, we select two relocation methods considered in this thesis: one representant of the online, and one of the offline approach. We simulate these representants in a discrete-event simulation based on historical data, for both the EMS region of Flevoland and Amsterdam. Comparing the results of this simulation study yields some interesting insights.

Part I

Online Approaches

2

A Dynamic Ambulance Management Model for Rural Regions

In this chapter, we consider the dynamic relocation problem, in which ambulances can be redeployed *proactively* throughout the region, for rural regions specifically. This type of region differs in a few aspects with respect to the urban case, e.g., the fleet size, the number of events and the spatial demand distribution. We construct a discrete-time Markov decision process (MDP) to model the Dynamic Ambulance Management (DAM) problem in an EMS system. Therefore, we model the road network of the region of interest as an equidistant graph and we take into account the current status of both the system and the ambulances in a state. We do not require ambulances to return to a base station: they are allowed to idle at any node in the graph. Instead, a policy is sought that specifies for each ambulance that is not busy with or en-route to a call whether to move to an adjacent node.

Since the MDP model is not tractable for large model instances, we present a heuristic approach to compute such redeployment actions. We construct several scenarios that may occur one time-step later and combine these scenarios with each feasible action to obtain a classification of actions. We show that on both patient- and crew-related performance indicators, the heuristic policy significantly outperforms a commonly used relocation policy structure in practice: the compliance table policy. Moreover, we compare the heuristic to the optimal policy for small-scaled instances of the problem, i.e., for an EMS system with few nodes and few ambulances.

The work in this chapter is based on Van Barneveld et al. (2015).

2.1 Introduction

The focus on rural regions has several important implications. In most papers on DAM, the numerical results section, in which the performance of the proposed methods is validated, is based on ambulance service providers operating in urban EMS regions, i.e., in large cities. However, there are substantial differences between urban and rural regions. We list three of them.

1. In rural regions, the number of ambulances is small compared to urban regions. Therefore, the effect of one ambulance fewer available, for instance, if this ambulance is busy, is more noticeable in rural regions with a limited number of ambulances. In contrast, in urban regions one ambulance fewer available probably only has a small impact on the ability to respond to emergency requests quickly. Thus, in rural regions, one has to be more careful about how to (re)deploy ambulances.

2. Besides, in rural EMS regions, the fluctuation in demand per area is much higher. There are areas with practically no demand, while in other areas, especially in cities or towns, the demand is high. As a consequence, an ambulance driving from a high-demand area to another high-demand area usually traverses an area of low demand, providing only very marginal coverage when it is en route. This is typically not the case in urban areas, in which an ambulance is always supplying coverage to a large amount of population, wherever it is. In this sense, relocating ambulances between areas of high demand involves more risks regarding the timely response to a patient.

3. A last difference between rural and urban regions is the number of events. Most papers, e.g., the work done by Gendreau et al. (2006), by Maxwell et al. (2010) and by Schmid (2012), assume that relocation decisions are taken only at the time of events, e.g., when the number of available ambulances changes due to a vehicle dispatch or service completion. This may work well for urban areas: after all, there are a lot of events due to the large number of demand requests. Hence, the moments at which dispatchers have the possiblity to adjust the ambulance location plan, are numerous as well. In contrast, the number of events in rural regions is low, resulting in fewer opportunities to do this. Summarizing, urban and rural regions differ much from each other, and thus, they should be approached differently.

The structure of this chapter is as follows. In Section 2.2, we provide an MDP formulation for the ambulance relocation problem. In this problem, ambulances can be present at designated locations in the region: nodes in the graph representing the region of interest. The objective is to find a good ambulance configuration: a distribution of ambulances throughout the region in such a way that one is able to respond to an incoming request quickly. This ambulance configuration can be achieved by moving ambulances over the graph. We decide on how we should move these ambulances, given the state of the system. Moreover, a certain penalty is associated with each possible response time. This penalty is defined using penalty functions. The proposed MDP-formulation is not tractable for large problem instances, so we resort to a heuristic, which is the topic of Section 2.3. The general idea of the heuristic is to consider scenarios that may occur one time step later. We combine these scenarios with each possible change in ambulance configuration to obtain a potentially new state. In this state, we consider the minimal expected

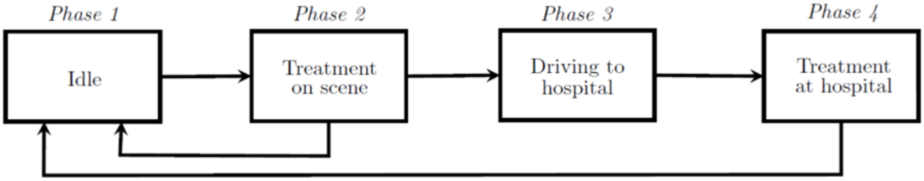


FIGURE 2.1: Simplified EMS process.

penalty related to the response to additional requests and classify the movement, based on these expectations and the probability that this particular scenario occurs. We conclude this chapter by the numerical study in Section 2.4. This study is based on simulation results for an ambulance service provider in a rural EMS region in the Netherlands: Flevoland.

2.2 Model

In this chapter, we make some assumptions on the general EMS process as described in Section 1.1 in order to fit into our modeling framework. Each incoming request of a patient needs an ambulance to attend to. We assume that the level of priority of requests for an ambulance is equal for each request. That is, we only consider emergency requests of the highest urgency: the life-threatening A1-calls. This assumption is justified by the fact that ambulance service providers are mostly judged on their performance regarding the highest priority incidents. Upon arrival at the emergency scene, the ambulance crew decides whether the patient needs transportation to a hospital. We assume that this decision is made quickly after arrival at the emergency scene, since the crew is already informed of the severity of the request by the emergency control center agent. With probability r, $0 \leq r \leq 1$, a patient needs transportation. If so, the ambulance crew treats the patient for a random number of time units at the emergency scene: the *treatment time on scene*. Then, he/she is transported to the *nearest* hospital. There, the ambulance transfers the patient for some random time, which we will call *treatment at hospital*. We assume no queueing takes place at the hospital: emergency departments have infinite capacity. This assumption is justified by the fact that we focus on rural regions with a small number of incoming requests per hour. Summarizing, our simplified EMS process is as follows: when the ambulance arrives at the emergency scene, the remaining time the ambulance is busy consists of a stochastic treatment time on scene, a deterministic transportation time, and a stochastic treatment time at the hospital. We refer to these stages as phase 1 to phase 4, see Figure 2.1.

Note that we do not include a phase for ambulances that are on their way to respond to a patient. The reason for this is that such an ambulance, although initially assigned, may not be the one to provide service. This kind of behaviour

$\mathcal{N}$	Node set.
N	Number of nodes.
A	Number of ambulances.
L	Length of longest path that any ambulance might take.
$\mathcal{H}$	Subset of nodes with a hospital.
p_i	Parameter of Poisson distribution that models the arrival of requests at node i.
r	Probability that a patient needs transportation to a hospital.
B_k^j	Treatment time of ambulance j at node k, $k \in \mathcal{N}$.
ρ_k^j	Probability that unit j at node k finishes its treatment one time step later, $k \in \mathcal{N}$.
a_i^s	Change in number of ambulances at location i, induced by action a^s.
d_i	Number of ambulances that start treatment of new patient at location i.
$\mathcal{F}(s)$	Set of feasible actions in state s.
F	Number of idle ambulances.
X	Number of requests not served by an ambulance yet.

Table 2.1: Notation.

occurs if a second ambulance, located closer to the request, becomes idle when the first one is en route. Therefore, as long as an ambulance is on its way to an emergency request, it is regarded as if it is idle. If a patient does not need transportation to a hospital, the busy time of the ambulance only consists of the stochastic treatment time on scene.

We model the region of interest as a graph, with $\mathcal{N}$ as its node set. The nodes of the graph serve as *demand locations*: locations where an incident might occur. There are two types of nodes: nodes *with* and *without* a hospital. Let these disjoint sets be denoted by $\mathcal{H}$ and $\bar{\mathcal{H}}$, respectively, where $\mathcal{N} = \mathcal{H} \cup \bar{\mathcal{H}}$. For simplicity, we enumerate the nodes in such a way that there is a hospital at the first $|\mathcal{H}|$ nodes in the enumeration, so

$$\mathcal{N} = \{1, 2, \ldots, |\mathcal{H}|, |\mathcal{H}| + 1, \ldots, |\mathcal{H}| + |\bar{\mathcal{H}}|\}.$$

The road network is modelled by edges, that can be either one- or bidirectional, depending on whether a U-turn is allowed on the specific road. This is typically not the case on highways. We assume that the length of each edge equals 1, so it takes one time step to traverse an edge. Therefore, time is discretized in time steps of Δt. As a consequence, it takes an ambulance Δt time units (e.g., 5 minutes) to cross an edge. In realistic situations, the graph is constructed in a way that Δt is fine enough to model ambulance movements. To model more realistic situations, one could decrease Δt, but then the graph should contain more nodes and edges. Therefore, for $\Delta t \to 0$, this model becomes continuous in both time and space.

Moreover, another assumption made in this model is that the number of in-

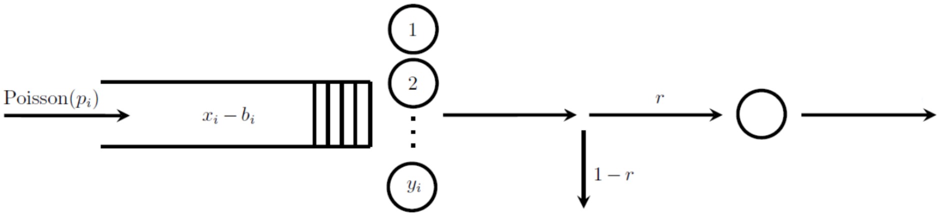


Figure 2.2: "Life cycle" of a request arriving at node i.

x_i	Number of patients at location i.
y_i	Number of ambulances at location i.
b_i	Number of busy ambulances at location i.
f_i	Number of idle ambulances at location i.
z_i	Number of ambulances at location i that need to transport a patient.
$Z(k, j)$	Elapsed treatment time of ambulance j at node k, $k \in \mathcal{N}$.
$D(h, t)$	Number of occupied ambulances that will arrive at hospital h in t time units.

Table 2.2: State space variables.

coming calls at each demand location per unit of time is Poisson distributed with parameter $p_i(\Delta t)$ for node i, $i \in \mathcal{N}$. These parameters can easily be estimated using historical data. In reality, these parameters vary over time, but here we assume that these are fixed for the sake of simplicity. Moreover, this is not really a limitation, since one can use different parameter values for different times of the day. For modeling issues, we assume that no external requests arrive at hospitals, so $p_h(\Delta t) = 0$ for $h \in \mathcal{H}$. The total number of ambulances in the system is A and all ambulances are of same type. The fleet size is homogeneous, constant and does not vary over time. Although $p(\Delta t) = (p_i(\Delta t))_{i \in \mathcal{N}}$ depends on the chosen time step size Δt, we will omit this dependence in the remainder for readibility issues. The notation is summarized in Table 2.1.

2.2.1 State Space

There are four major sources of randomness in the EMS process model considered in this chapter: the arrival of requests, the possible need for transportation to a hospital, the service time on scene, and the time an ambulance spends at the hospital, see Figure 2.2. The state of our system in our MDP formulation is given by five components, which we describe in detail below. For an overview of the notation of the state space variables, we refer to Table 2.2.

1. The number of patients per demand location. Due to the spatial aggregation, there can be multiple patients waiting for an ambulance at the same time in the same area. This number is denoted by a vector $x = (x_1, x_2, \ldots, x_N)$ of length $N = |\mathcal{N}|$, where $x_i \in \mathbb{N}_0$ for $1 \leq i \leq N$. We assume that each patient needs an ambulance and an ambulance cannot serve more than one patient at a time.

2. The number of ambulances either in phase 1, 2 or 4 per demand node. This is similarly denoted as the previous state component by $y = (y_1, y_2, \ldots, y_N)$ of N, where $y_i \in \mathbb{N}_0$, $1 \leq i \leq N$. Moreover, $\sum_{i=1}^{N} y_i \leq A$, since the fleet size cannot be exceeded. If there is both a patient and an ambulance at a certain node, we assume that this ambulance is treating this patient: the vector $b = (b_1, b_2, \ldots, b_N)$ of busy ambulances (ambulances either in phase 2 or phase 4) is given by $b = \min(x, y)$. In addition, $f = y - b$ denotes the vector of idle ambulances, i.e., the ambulances in phase 1.

3. The number of ambulances per demand location required to transport patients. That is, the number of phase 2 ambulances that once the treatment on scene has finished, will make a transition to phase 3. We denote this by a vector $z = (z_1, z_2, \ldots, z_N)$, where $0 \leq z_i \leq b_i$ for each node i. Moreover, an ambulance at a hospital does not have to transport a patient, so $z_h = 0$ for $h \in \mathcal{H}$.

4. The elapsed service time of ambulances in phases 2 and 4. We denote this by a matrix Z with $|\mathcal{N}|$ rows and A columns, where $Z(k, j_1)$ denotes the elapsed service time of ambulance j_1 at node k, where $j_1 \leq b_k$. Moreover, $Z(k, j_2) = -1$ for $b_k < j_2 \leq A$. Hence, $\sum_{j=1}^{A} \mathbb{1}_{\{Z(k,j) \geq 0\}} = b_k$. We assume that each row of Z is sorted in non-increasing order, in order to simplify the description of the computations in Sections 2.2.3 and 2.3.1. Rows $k \leq |\mathcal{H}|$ and the remaining rows correspond to ambulances treating at hospitals and ambulances treating on scene, respectively.

5. Destinations and remaining driving times of ambulances in phase 3. We denote these quantities by a matrix D. Let $D(h, t)$ describe the number of phase-3 ambulances that will arrive in $t \geq 1$ time units at hospital $h \in \mathcal{H}$. Note that $\sum_{h \in \mathcal{H}} \sum_{t=1}^{L} D(h, t) + \sum_{i=1}^{N} y_i = A$, where L denotes the length of the longest path that any ambulance might take.

A state s is now defined by the tuple $s = (x, (b, f), z, Z, D)$, or equivalently: $s = (x, y, z, Z, D)$, where $y = b + f$.

2.2.2 Actions

We will now describe the control process of the MDP. An action set belongs to each state s. Actions describe the change in *configuration* of *idle* ambulances: we can either dispatch an idle ambulance to one of its neighbouring nodes, or we can let it hold its current position. An action belonging to the action set of state s is denoted by

$$a^s = (a_1^s, a_2^s, \ldots, a_N^s),$$

where $a_i^s \in \mathbb{Z}$ denotes the change in y_i, i.e., the change in the number of ambulances present at node i. It is possible that $a_i^s = 0$, while ambulances are moving from/to node i. This occurs when the number of incoming ambulances equals the number of outgoing ambulances at location i. To keep track of the exact movement of ambulances, we can decompose a^s into an $(a^s)^-$- and an $(a^s)^+$-part, where $(a^s)^-$ and $(a^s)^+$ denote the number of outgoing and incoming ambulances per node, respectively. Naturally, $(a_i^s)^-, (a_i^s)^+ \in \mathbb{N}_0$ for $i \in \mathcal{N}$ and $a^s = (a^s)^+ - (a^s)^-$. Action a^s satisfies the condition $-a_i^s \leq f_i$, since no more than f_i ambulances can be removed from node i. Similarly, no more than the total number of idle ambulances can be sent to location i, so $(a_i^s)^+ \leq \sum_{j \neq i} f_j$. All edges have length 1, so it takes exactly one time step to carry out an action. Therefore, it holds that

$$\sum_{i=1}^{N} \Big((a_i^s)^+ - (a_i^s)^- \Big) = \sum_{i=1}^{N} a_i^s = 0,$$

since the number of departing idle ambulances equals the number of arriving idle ambulances. Furthermore, since the actions are configuration-based rather than based on each ambulance separately, idle ambulances are *indistinguishable*.

Note that actions are only defined for idle ambulances. Busy ambulances, which are ambulances either treating a patient at an emergency scene or at a hospital, continue their service. There are actions that are not reasonable to take, but still allowed: actions in which the response time to a request is unnecessarily delayed. We want to exclude these actions since these are suboptimal in the model and in reality they are not even considered. We call these actions *infeasible*. The question arises on how to define the set of *feasible* actions, which we denote by $\mathcal{F}(s)$ for state s. To compute $\mathcal{F}(s)$ in state s, we solve either a *minimum weighted bipartite matching* (MWBM) problem or a *linear bottleneck assignment problem* (LBAP). Both problems differ in objective, but they share the same modeling framework, which we describe next in the context of the ambulance relocation model considered in this chapter.

Assume $s = (x, (b, f), z, Z, D)$. Let $F = \sum_{i=1}^N f_i$ and $X = \sum_{i=1}^N (x_i - y_i)^+$ denote the number of idle ambulances and the number of requests that are not served by an ambulance yet, respectively. We introduce a weighted complete bipartite graph $K_{F,X} = (V_1 \cup V_2, E, l)$, where V_1, V_2 are the two node sets, E the edge set and l a function assigning weights to edges. The node set V_1 corresponds to the locations of the F idle ambulances: for each ambulance we introduce a node indexed by its location. In a similar way, we define the node set V_2, but these nodes correspond to the location of patients waiting. If there are more ambulances or patients waiting at a particular location, then we specify the nodes belonging to this location with a subindex. Let $v_1 \in V_1$, $v_2 \in V_2$. The weight $l((v_1, v_2))$ of edge (v_1, v_2) equals the length of the shortest path between v_1 and v_2. This corresponds to the required number of time units to travel from the corresponding locations of v_1 and v_2 in the original graph representing the region of interest. Therefore,

$l : E \to \mathbb{N}_0$. Moreover, we define a *matching* as a set of edges without common nodes. A node is matched if it is an endpoint of one of the edges in the matching. A matching is called *maximal* if all nodes in V_1 or all nodes in V_2 are matched. Both the MWBM and the LBAP aim to find an optimal matching. We explain both assignment problems next.

Minimum Weighted Bipartite Matching

It seems obvious to always dispatch the nearest ambulance to a request. However, this action can be suboptimal in our model, because it possibly delays the response time to a different request. By modeling the assignment problem by an MWBM, the total response time to all requests that are not served yet is minimized. If there is only one such request, this assumption is equivalent to the policy in which the nearest ambulance is assigned to a request. A *minimum weighted bipartite matching* is defined as a maximal matching M where the sum of the weights of the edges in M has a minimal value. That is, our objective criterion is

$$\min_{M \in \mathcal{M}} l(M) = \min_{M \in \mathcal{M}} \sum_{e \in M} l(e) \tag{2.1}$$

and $\mathcal{M}$ is the set of all maximal matchings. The assignment problem is solved by the *Hungarian Algorithm*, which runs in $\mathcal{O}((|V_1|+|V_2|)^2|E|)$ time (Schrijver, 2003). Note that finding a minimum weighted bipartite matching in $K_{F,X}$ is equivalent to finding an assignment of ambulances to requests with respect to Equation (2.1).

Linear Bottleneck Assignment

The objective of minimizing the total mean response time to all patients that are waiting seems reasonable. However, one can argue about it. Possibly, an ambulance responds to the majority of these patients within a short amount of time, while for one patient it takes a very long time before an ambulance arrives at the emergency scene. It seems fairer to divide the total response time equally over all requests waiting. A way to achieve this is to model the matching problem as a linear bottleneck assignment problem (LBAP). Instead of minimizing the total sum of the edges in the matching as before, the LBAP aims to find a maximal matching with the property that the maximum weight of the edges in the matching is minimized. That is, the objective criterion is

$$\min_{M \in \mathcal{M}} l(M) = \min_{M \in \mathcal{M}} \max_{e \in M} l(e),$$

and $\mathcal{M}$ is the set of all maximal matchings. This problem and several of its solution methods are treated in detail in Burkhard et al. (2009), in which a polynomial-time algorithm is proposed. Moreover, if the set of such matchings contains more than one such matching, this algorithm finds the matching with minimal total weight in this set. In the context of dynamic ambulance management, this translates to obtaining an assignment of idle ambulances to requests, such that the maximum response time is minimized and given this maximum response time, the total response time is minimized.

ω_i^1	Number of arriving requests at location i.
$\omega_i^2(s)$	Number of treatment completions on scene in state s at location i.
$\omega_h^3(s)$	Number of treatment completions at hospital h in state s.
$\omega_i^4(s, \omega_i^2(s))$	Number of ambulances departing for a hospital from node i in state s.
$\omega_i^5(s, \omega_i^1, \omega_i^4(s, \omega_i^2(s)))$	Number of patients at location i decided to be transported.

Table 2.3: Types of randomness in the evolution of the system.

Now, the construction of the set of feasible actions in state s, $\mathcal{F}(s)$, is as follows: we solve either an MWBM or an LBAP to obtain a matching in which ambulances are assigned to requests. If the number of patients waiting is smaller than the number of idle ambulances, there are ambulances that are not assigned to requests. For these ambulances, we have a choice where to send them to. For these states, the set of feasible actions contains *more* than one action: if $X < F$ in state s, we have to decide for $F - X$ ambulances how to relocate them, where ambulances that are not assigned to a request can either be relocated to one of the neighbouring nodes or they can keep their position. However, if the number of patients waiting exceeds the number of idle ambulances, we can not respond to all requests. Then, we have only one possible action: the action induced by the matching. This is also the case when we have an equal number of idle ambulances and patients waiting.

2.2.3 Evolution

In this section, we describe the underlying dynamics of the MDP model of the EMS system considered in this chapter. If our current state is $s = (x, y, z, Z, D)$ and action $a^s \in \mathcal{F}(s)$ is taken, the system evolves according to random variables related to number of arriving requests (denoted by ω_1), number of treatment completions (ω_2 and ω_3), number of ambulances departing to a hospital (ω_4) and the number of patients for which it is decided that they need transportation (ω_5). These random variables are summarized in Table 2.3. We describe the dynamics per state component. Let $s' = (x', y', z', Z', D')$ denote the next state.

1. The number of patients per demand location. We distinguish between nodes with and nodes without a hospital. Consider node $h \in \mathcal{H}$. The number of requests at hospital h in the next state, x'_h, depends on two processes: arrival of occupied ambulances at node h and the completion of treatments by an ambulance at hospital h. Recall that $p_h = 0$ for $h \in \mathcal{H}$, so there are no new arrivals. The number of arriving ambulances at location $h \in \mathcal{H}$ in the next time step equals $D(h, 1)$. For the number of completions, we use the Poisson binomial distribution, which is the discrete probability distribution of a sum of independent Bernoulli trials that are not identically distributed (Wang, 1993). The probability of having κ successful trials out of a total of n trials can be written as the sum

$$\mathbb{P}\{K=\kappa\} = \sum_{U \in \mathcal{U}_\kappa(n)} \prod_{i \in U} \rho^i \prod_{j \in \bar{U}} (1-\rho^j), \tag{2.2}$$

where $\rho^1, \rho^2, \ldots, \rho^n$ denote the success probabilities, $\mathcal{U}_\kappa(n)$ is the set of all subsets of κ integers selected from $\{1, 2, \ldots, n\}$. The ordinary binomial distribution is a special case where all the success probabilities are equal. The number of completions, denoted by $\omega_h^3(s)$, is a Poisson binomially distributed number on b_h trials. The success probability ρ_h^j, which is the probability that ambulance j at h will have finished its treatment at the next step, depends on the elapsed service time of ambulance j, which is $Z(h,j)$. That is,

$$\rho_h^j = \mathbb{P}\{B_h^j = Z(h,j)+1 | B_h^j > Z(h,j)\}, \tag{2.3}$$

where B_h^j is the treatment time of ambulance j at hospital h. The b_h probabilities needed for the Poisson binomial distribution are given by the vector $(\rho_h^1, \rho_h^2, \ldots, \rho_h^{b_h})$. Now,

$$x'_h = x_h + D(h,1) - \omega_h^3(s), \;\; h \in \mathcal{H}.$$

If $i \in \bar{\mathcal{H}}$, i.e., if there is no hospital at node i, then x'_i is defined differently, since it depends on two other processes: the arriving requests at location i and the number of treatment completions on scene at location i. These numbers are denoted by ω_i^1 and $\omega_i^2(s)$. Note that ω_i^1 does not depend on s and is Poisson distributed with parameter p_i. In contrast, $\omega_i^2(s)$ does depend on s: this number is Poisson binomially distributed with success probability $(\rho_i^1, \rho_i^2, \ldots, \rho_i^{b_i})$.

2. The number of ambulances either in phase 1, 2 or 4 per demand location. For the evolution of $y = f + b$, we also distinguish between hospital locations and other locations. We consider the case that $i \in \bar{\mathcal{H}}$ first. The number of ambulances y'_i in the next state at location i depends on the current number of ambulances y_i, the action a_i^s and the number of ambulances that completes service on scene and departs for a hospital. Let this random number, which depends on the number of completions on scene, be denoted by $\omega_i^4(s, \omega_i^2(s))$. Then, we find that

$$y'_i = y_i + a_i^s - \omega_i^4\left(s, \omega_i^2\left(s\right)\right), \;\; i \in \bar{\mathcal{H}}.$$

The quantity $\omega_i^4(s, \omega_i^2(s))$ is determined as follows. We have b_i ambulances that are serving a patient on scene. Of these b_i ambulances, z_i ambulances need to go to a hospital and $\omega_i^2(s)$ ambulances complete their service on scene now. Therefore, $\omega_i^4(s, \omega_i^2(s))$ is hypergeometrically distributed on a population size of b_i of which z_i are of one type, and $b_i - z_i$ of the other type. Moreover, the number of draws is $\omega_i^2(s)$.

If $h \in \mathcal{H}$, y'_h depends on the current number of ambulances at location h, the action a^s, and the number of occupied ambulances that arrive at hospital h. Hence,

$$y'_h = y_h + a_h^s + D(h,1), \;\; h \in \mathcal{H}.$$

3. The number of busy ambulances per demand location required to transport patients. When considering the number of busy ambulances at hospitals required to transport patients, it is clear that $z_h = 0$ for $h \in \mathcal{H}$. If i does not correspond to a hospital location, z_i' is obtained as follows. It depends on $\omega_i^4(s, \omega_i^2(s))$ defined before and the number of new patients for which it is decided that they need transportation. Let this last random quantity be denoted by $\omega_i^5(s, \omega_i^1, \omega_i^4(s, \omega_i^2(s)))$. Note that this number depends on the number of arriving and completed requests ω_i^1 and $\omega_i^2(s)$. Then,

$$z_i' = z_i - \omega_i^4\left(s, \omega_i^2(s)\right) + \omega_i^5\left(s, \omega_i^1, \omega_i^4\left(s, \omega_i^2(s)\right)\right).$$

Note that this number is bounded by the number of ambulances that start a treatment of a new patient at location i, denoted by d_i. Then,

$$d_i = \min\{\omega_i^2(s) - \omega_i^4\left(s, \omega_i^2(s)\right) + f_i + a_i^s,\ \omega_i^1 + x_i - b_i\}, \tag{2.4}$$

where $\omega_i^2(s) - \omega_i^4(s, \omega_i^2(s)) + f_i + a_i^s$ equals the number of ambulances that start a new treatment: there were already f_i idle ambulances and we add the $\omega_i^2(s)$ ambulances that complete service. However, $\omega_i^4(s, \omega_i^2(s))$ of these ambulances leave for a hospital and cannot start a new treatment. If $a_i^s > 0$, we have arrivals of ambulances, which can all start a new service, so we add that number as well. If $a_i^s < 0$, some of these idle ambulances leave for a different location and these ones cannot start a treatment at location i. Note that

$$\omega_i^2(s) - \omega_i^4(s, \omega_i^2(s)) + f_i + a_i^s \geq 0,$$

since $\omega_i^2(s) - \omega_i^4(s, \omega_i^2(s)) \geq 0$ and $f_i + a_i^s \geq 0$. However, not all of these $\omega_i^2(s) - \omega_i^4(s, \omega_i^2(s)) + f_i + a_i^s$ ambulances can start a new service if there are not that many requests waiting at i. This quantity is given by $\omega_i^1 + x_i - b_i$: there were $x_i - b_i \geq 0$ patients without an ambulance treating them, and ω_i^1 additional requests arrive. Then, $\omega_i^5(s, \omega_i^1, \omega_i^4(s, \omega_i^2(s)))$ is binomially distributed on d_i ambulances that start a new treatment, and with probability r each of these patients requires transport.

4. The elapsed service time of ambulances in phases 2 and 4. Let $h \in \mathcal{H}$ and let $Z(h)$ denote the h-th row of Z. The evolution of $Z(h)$ depends on two processes: the completion of the service time of patients and the arrival of occupied ambulances. The number of completions at hospital h is $\omega_h^3(s)$. Each busy ambulance j completes its service with probability ρ_h^j defined in (2.3), where $j \leq b_h$.

Thus, in total there are $I = \binom{b_h}{\omega_h^3(s)}$ options, which we enumerate by the variable i, for the new configuration of busy ambulances at h. Each of these options has positive probability. To calculate these probabilities, we need to enumerate all options. Define $\mathcal{U}_{b_h}(\omega_h^3(s))$ as the set of subsets of $\omega_h^3(s)$ integers that can be selected from $\{1, 2, \ldots, b_h\}$. Moreover, let $U^i \in \mathcal{U}_{b_h}(\omega_h^3(s))$ be the set of ambulances that remain busy in the i-th option, where $|U^i| = b_h - \omega_h^3(s)$ and $1 \leq i \leq I$. Then we define $\pi(U^i)$ as the probability that only the ambulances in U^i remain busy. These probabilities are calculated by

$$\pi(U^i) = \prod_{j_1 \in U^i} (1 - \rho_h^{j_1}) \prod_{j_2 \in \bar{U}^i} \rho_h^{j_2},$$

where $\bar{U}^i = \{1, 2, \ldots, b_h\} \backslash U^i$. This equals the probability mass function of the Poisson binomial distribution given in (2.2), but here we condition on $\omega_h^3(s)$. Therefore, $\sum_{i=1}^{I} \pi(U^i) < 1$, so we need to normalize. Let $\pi'(U^i)$ denote the normalized probabilities. That is,

$$\pi'(U^i) = \frac{\pi(U^i)}{\sum_{i=1}^{I} \pi(U^i)}$$

for each outcome i, $1 \leq i \leq I$. Now, we obtain a probability distribution on the set of outcomes and we sample an option from this distribution. Assume the sampled outcome is i. Then we define $Z^*(h)$ as follows:

$$Z^*(h, j) = \begin{cases} -1 & \text{if } j \in \bar{U}^i \text{ or } b_h < j, \\ Z(h, j) + 1 & \text{if } j \in U^i. \end{cases} \tag{2.5}$$

If $j \leq b_h$ and $j \in \bar{U}^i$, the j-th ambulance completes its service and is no longer busy; its elapsed service time is discarded. In the second case in (2.5), the j-th ambulance does not finish its treatment and thus its elapsed service time is increased by 1 time unit. Then, we sort $Z^*(h)$ in non-increasing order to make sure that there are no -1's in the first $b_h - \omega_h^3(s)$ entries.

Up to now, we only considered the completions of busy ambulances. However, occupied ambulances can arrive at hospital h as well. Note that during the transition from s to s', $D(h, 1)$ ambulances arrive at h. Then, $b'_h = b_h - \omega_h^3(s) + D(h, 1)$ and

$$Z'(h, j) = \begin{cases} Z^*(h, j) & \text{if } j \leq b_h - \omega_h^3(s), \\ 0 & \text{if } b_h - \omega_h^3(s) < j \leq b'_h, \\ -1 & b'_h < j. \end{cases}$$

Recall that we conditioned on $\omega_h^3(s)$. Alternatively, we could have chosen to consider all 2^{b_h} options. That is, we do not condition on $\omega_h^3(s)$. If we define a probability distribution on all these 2^{b_h} options, the probabilities sum up to 1. Sampling from this distribution, we immediately obtain a new configuration *and* the number of completions defined as $\omega_h^3(s)$. For $i \in \bar{\mathcal{H}}$, the evolution is similar, with $\omega_i^2(s)$ instead of $\omega_i^3(s)$ and no $D(h, 1)$-term.

5. Destinations and remaining driving times of ambulances in phase 3. Let $H(h)$ describe the set of demand locations for which hospital h is nearest among all hospitals. Formally,

$$H(h) = \{i \in \mathcal{N} \mid l(i, h) \leq l(i, h') \; \forall h' \in \mathcal{H}, \; h \neq h\},$$

where $l(i, h)$ denotes the required number of time units to travel from i to h. Remember that we assume that a patient, who needs transportation, is always transported to the nearest hospital. However, it is possible that there exist two hospitals h_1 and h_2 for which $H(h_1) \cap H(h_2) \neq \emptyset$, i.e., there is a demand location for which these two hospitals are both closest. We aim to send all occupied ambulances from this location to only one hospital, so we create a partition of the node set by using the recursion

$$H^*(h) = H(h) \backslash \bigcup_{j=1}^{h-1} H^*(j).$$

That is, if multiple hospitals are nearest, we send all occupied ambulances to the first hospital according to the enumeration of the nodes. Then, $D(h,t)$ evolves as follows:

$$D'(h,t) = D(h,t+1) + \sum_{i \in \bar{\mathcal{H}}} \omega_i^4 \left(s, \omega^2(s)\right) \mathbb{1}_{\{i \in H^*(h)\}} \mathbb{1}_{\{l(i,h)=t\}},$$

where $\omega_i^4 \left(s, \omega^2(s)\right)$ denotes the number of occupied ambulances departing for a hospital from location i. The term $\mathbb{1}_{\{i \in H^*(h)\}} \mathbb{1}_{\{l(i,h)=t\}}$ equals 1 if and only if h is the nearest hospital to location i and the travel time from i to h is t.

2.2.4 Objectives

In practice, each country, possibly even each ambulance service provider within a country, uses its own performance measure. In this section we demonstrate how to incorporate different objectives in our MDP formulation. We do this by introducing a non-negative continuous penalty (or cost) function Φ, which is a function of the response time solely, with domain $\mathbb{R}_{\geq 0}$. Several examples of cost functions are displayed in Figure 2.3.

Denote the cost in state $s = (x, y, z, Z, D)$ by $c(s)$. Let $F = \sum_{i=1}^N f_i$ and $X = \sum_{i=1}^N (x_i - y_i)^+$ denote the number of idle ambulances and the number of requests that are not served by an ambulance yet, respectively. We solve an assignment problem as described in Section 2.2.2 to obtain an assignment of idle ambulances to the X waiting patients. Unless $F < X$, each waiting request is assigned and a certain (remaining) response time to each of these requests is obtained. Denote these (remaining) response times by $R_1^s, R_2^s, \ldots, R_X^s$ for an enumeration of waiting requests. Note that $R_i^s > 0$, $i = 1, \ldots, X$. Now we define

$$c(s) = \sum_{i=1}^{X} (\Phi(R_i^s) - \Phi(R_i^s - 1)). \tag{2.6}$$

Note that the total penalty generated by request i equals

$$\sum_{t=1}^{\hat{R}_i^s} (\Phi(t) - \Phi(t-1)) = \Phi(\hat{R}_i^s).$$

This is the case if the ambulance assigned to it is not reassigned to a different request, where $\hat{R}_i^s$ denotes the total response time to request i. If the ambulance is reassigned, the penalty is slightly different. However, this hardly occurs in practice. If $F < X$, that is, there are not enough idle ambulances to respond to each of the waiting patients, we set the response time to an unassigned request

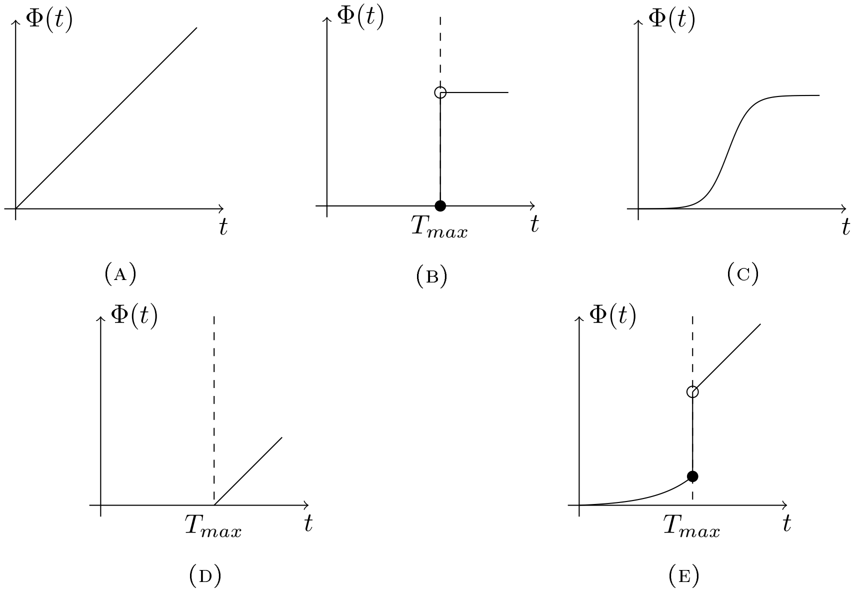


Figure 2.3: Examples of penalty functions $\Phi(t)$.

equal to a large number. After some time steps this request will get assigned, but before that it generates costs as well. The objective is to minimize average costs over an infinite horizon.

An obvious performance measure is the average response time to a request. The objective of minimizing the average response time corresponds to a linear cost function:

$$\Phi(t) = t,\ t \geq 0, \tag{2.7}$$

which is displayed in Figure 2.3a. Each additional time unit of delay generates the same penalty, since the derivative of this function is constant. Using this cost function results in a small average response time, but the variance may be large. Another commonly used type of performance measure is the percentage of emergency requests responded to within a certain maximum allowed response time threshold T_{max}, given by

$$\Phi(t) = \begin{cases} 0 & t \leq T_{max}, \\ 1 & t > T_{max}. \end{cases} \tag{2.8}$$

The penalty function corresponding to this performance measure is displayed in Figure 2.3b, and using it will relocate the ambulances in such a way that the coverage of the EMS region is maximized. However, in using this penalty function, there is no difference in penalty between a really short response time and a response time that is slightly below the maximum allowed one. To overcome this problem, one could use the penalty function

$$\Phi(t) = \frac{1}{1 + e^{-\beta(t - T_{max} - 0.5)}}, \ t \geq 0, \tag{2.9}$$

where $\beta \geq 0$ is a scaling parameter. This function is displayed in Figure 2.3c. This function is a smooth version of the function of Equation (2.8). The penalty function in Figure 2.3d has the interpretation of minimizing average lateness, and is given by

$$\Phi(t) = \begin{cases} 0 & t \leq T_{max}, \\ t - T_{max} & t > T_{max}. \end{cases} \tag{2.10}$$

The function in Figure 2.3e, which is suggested by practitioners in the field, combines the cost functions in Figures 2.3a–2.3d and is given by

$$\Phi(t) = \begin{cases} \frac{1}{\gamma}(e^t - 1) & 0 \leq t \leq T_{max}, \\ t - (T_{max} - 1) & t > T_{max}, \end{cases} \tag{2.11}$$

where $\gamma \geq 0$ is a scaling parameter. At T_{max}, the function makes a jump to ensure that not meeting the maximum allowed response time is much worse than a response time that does. For $t > T_{max}$, we use the performance measure of minimizing average lateness. To differentiate between response times before T_{max}, we use an exponential cost function. Moreover, other penalty functions, for instance, penalty functions related to survival of a patient as considered in Erkut et al. (2008) and Chapter 5, can be incorporated.

2.2.5 Tractability

The state space is high-dimensional; in theory, we have infinitely many states, since there is no upper bound on the number of requests per location. However, we can introduce such an upper bound to obtain a finite number of states, but even for small-size instances solving the problem, i.e., finding the optimal policy, becomes intractable. This is not only a consequence of the high-dimensional state space: the large number of actions plays a role as well. This number can be very large for states with few requests, since we allow ambulances to move to each neighbouring node, not only to designated nodes such as base stations. As a consequence, solving this problem by modeling it as an MDP and applying methods described by Puterman (1994) is not tractable for realistic settings, although we were able to compute the optimal policy for a simplified example, c.f., Section 2.4.2. Therefore, we resort to a heuristic solution, which is the topic of Section 2.3.

2.3 Heuristic Solution

In this section we propose a heuristic that computes an action, given the state of the system, in order to overcome the tractibility issues mentioned above. The general idea of this heuristic is to take the feasible action that minimizes the expected penalty generated by an arriving request during the next time step, given the current state of the system. It is a one step look-ahead method that generates several scenarios that may occur one step later. All of these scenarios are possible outcomes of the evolution of the system, described in Section 2.2, with the action in which each idle ambulance keeps its position. However, we only generate scenarios in which *at most one* request arrives. The reason behind this is twofold. First, we aim to bound the number of possible scenarios, since this facilitates the computations. Second, if Δt is small, it is not very likely that two or more requests arrive in the same time period. The probability that such a scenario in rural regions occurs is relatively small, and we do not consider this.

Consider state $s = (x, b, f, z, Z, D)$. We generate all possible outcomes of the evolutionary process described in Section 2.2.3, with the restrictions that $\sum_{i \in \bar{\mathcal{H}}} \omega_i^1 \leq 1$ and $a_i^s = 0$, $i \in \mathcal{N}$. That is, the number of arriving calls is bounded by 1 and each ambulance keeps its position. Let this set of possible scenarios when sampling from state s under these restrictions be denoted by $\mathcal{S}(s)$ and

$$s^n = (x^n, b^n, f^n, z^n, Z^n, D^n) \in \mathcal{S}(s)$$

denote the n-th scenario, where $1 \leq n \leq |\mathcal{S}(s)|$. Moreover, $\mathbb{P}\{s' = s^n | s\}$ denotes the probability that scenario n occurs. Due to the restriction on the number of requests that can happen at the same time, it holds that

$$\sum_{n=1}^{|\mathcal{S}(s)|} \mathbb{P}\{s' = s^n | s\} < 1.$$

For the calculation of $\mathbb{P}\{s' = s^n | s\}$, we use a slightly different arrival process of requests, since we know that at most one request arrives. Before, at demand location i exactly one request occurred with probability $p_i e^{-p_i}$, due to the fact that the number of arriving requests is Poisson distributed. However, for the calculation of the scenario probabilities, we assume that at location i exactly one request occurs with probability $1 - e^{-p_i}$. That is, we add the probability of *more* than one incoming request to the probability of *exactly* one incoming request. In the next section, we calculate the probability that scenario n occurs for each of the five state components, step by step.

2.3.1 Scenario Probabilities

1. The number of patients per demand location. We first consider $\mathbb{P}\{x' = x^n | s\}$ for scenario n. Since the arrival and completion process of requests is nodewise independent, it holds that

$$\mathbb{P}\{x' = x^n | s\} = \prod_{i=1}^{N} \mathbb{P}\{x_i' = x_i^n | s\}. \tag{2.12}$$

As in Section 2.2.3, we distinguish between $h \in \mathcal{H}$ and $i \in \bar{\mathcal{H}}$. We consider the case $h \in \mathcal{H}$ first. The arrival process is defined by the occupied ambulances arriving to h. This number is given by $D(h,1)$. The number of patients in scenario n for which the treatment is *not* completed, denoted by G_h^n at h, is Poisson binomially distributed. That is,

$$\mathbb{P}\{G_h^n = g_h^n|s\} = \sum_{U\in\mathcal{U}_{b_h}(g_h^n)} \prod_{j_1\in U}(1-\rho_h^{j_1}) \prod_{j_2\in\bar{U}} \rho_h^{j_2},$$

where $\mathcal{U}_{b_h}(g_h^n)$ is the set of subsets of $\{1,2,\dots,b_h\}$ with exactly g_h^n elements. Moreover, $\bar{U}$ is the complement of U^i in $\{1,2,\dots,b_h\}$ and ρ_h^j is the probability that the j-th patient at h will have been treated at the next time step. If $D(h,1)=j$ patients arrive in one time step at h, then the total number of patients at h in scenario n is in $\{j, j+1,\dots,j+x_h\}$. Moreover, given that j patients arrive and in scenario n we have x_h^n patients at h, we observe that for $x_h^n - j$ patients treatment is not completed. Now, we find

$$\mathbb{P}\{x'_h = x_h^n|s\} = \sum_{j=0}^{A} \mathbb{1}_{\{D(h,1)=j\}}\mathbb{1}_{\{j\le x_h^n\le x_h+j\}}\mathbb{P}\{G_h^n = x_h^n - j|s\}.$$

For $i\in\bar{\mathcal{H}}$, the arrival process of requests is not deterministic: a request arrives with probability $1-e^{-p_i}$. The total number of patients for which the service on scene is finished at i is again Poisson binomially distributed. Therefore,

$$\mathbb{P}\{x'_i = x_i^n|s\} = \begin{cases} (1-e^{-p_i})\times & \\ \mathbb{P}\{G_i^n = b_i-(x_i-x_i^n+1)|s\}+ & \\ e^{-p_i}\mathbb{P}\{G_i^n = b_i-(x_i-x_i^n)|s\} & \text{if } x_i-b_i\le x_i^n,\ x_i^n\le x_i+1, \\ e^{-p_i}\mathbb{P}\{G_i^n = 0|s\} & \text{if } x_i^n = x_i-b_i, \\ (1-e^{-p_i})\mathbb{P}\{G_i^n = b_i|s\} & \text{if } x_i^n = x_i+1, \\ 0 & \text{else,} \end{cases}$$

where b_i denotes the number of ambulances that are busy serving a patient, i.e., the number of patients that are treated by an ambulance. The first part of the sum above considers the situation in which a request arrives at i, with probability $1-e^{-p_i}$. Mind that this arriving request cannot be served immediately.

2. The number of ambulances either in phase 1, 2, or 4 per demand location. Now, we consider the transition probabilities for the second component. Similar to Equation (2.12), this probability can be written in product-form:

$$\mathbb{P}\{y' = y^n|s,x^n\} = \prod_{i=1}^{N}\mathbb{P}\{y'_i = y_i^n|s,x_i^n\}$$

for scenario n. Note that y_i^n depends on x_i^n. If in scenario n, $x_i - x_i^n$ treatments on scene at location i are finished, and these ambulances all leave for a hospital, we find that $y_i^n = y_i - (x_i - x_i^n)$. Moreover, there can be multiple possibilities for y_i^n that correspond to x_i^n. This is the case if for a particular location $i\in\bar{\mathcal{H}}$, we have multiple busy ambulances and at least one but not all of these need to go to a hospital. If, in scenario n, no requests arrive at i and $0 < x_i - x_i^n < x_i$ ambulances finish service on scene, we do not know how many of these $x_i - x_i^n$ ambulances need to transport a patient. Thus,

$$y_i - \min\{z_i, x_i - x_i^n\} \leq y_i^n \leq y_i - \max\{0, (x_i - x_i^n) - (b_i - z_i)\}.$$

Assume that x_i^n is given, and that in node i no hospital is located, i.e., $i \in \bar{\mathcal{H}}$. We make a distinction whether *no* or *one* extra request is considered at i in scenario n. If no extra request is considered, then for $x_i - x_i^n$ patients the treatment on scene ends. If an additional request is considered, then $x_i - x_i^n + 1$ ambulances finish their service at location i. Let $\mathbb{P}\{0|s, x_i^n\}$ denote the probability that x_i^n does *not* include an extra request and let $\mathbb{P}\{1|s, x_i^n\}$ denote the probability that it does. Because we assume that no more than one request can arrive per time period, it holds that $\mathbb{P}\{0|s, x_i^n\} + \mathbb{P}\{1|s, x_i^n\} = 1$. We distinguish three cases:

1. If $x_i^n = x_i - b_i$, all busy ambulances complete their treatment on scene and no request arrives. Hence, $\mathbb{P}\{0|s, x_i^n\} = 1$.

2. If $x_i^n = x_i + 1$, no ambulance completes its treatment on scene and one request arrives. Therefore, $\mathbb{P}\{1|s, x_i^n\} = 1$.

3. If $x_i - b_i < x_i^n < x_i + 1$, either no or one additional request is considered. Thus, $\mathbb{P}\{0|s, x_i^n\} = e^{-p_i}$ and $\mathbb{P}\{1|s, x_i^n\} = 1 - e^{-p_i}$.

Let $\mathbb{P}\{y_i' = y_i^n|s, x_i^n, 0\}$ and $\mathbb{P}\{y_i' = y_i^n|s, x_i^n, 1\}$ denote the probability that no and one extra request at i is considered in scenario n, respectively. Then,

$$\begin{aligned} &\mathbb{P}\{y_i' = y_i^n|s, x_i^n\} = \\ &\quad \mathbb{P}\{y_i' = y_i^n|s, x_i^n, 0\}\mathbb{P}\{0|s, x_i^n\} + \mathbb{P}\{y_i' = y_i^n|s, x_i^n, 1\}\mathbb{P}\{1|s, x_i^n\}. \end{aligned} \tag{2.13}$$

First, we determine $\mathbb{P}\{y_i' = y_i^n|s, x_i^n, 0\}$ in order to compute the left-hand side of Equation (2.13). Of the b_i busy ambulances at location i, $x_i - x_i^n$ finish their service on scene. If y_i^n ambulances remain at i, then $y_i - y_i^n$ of the z_i ambulances that have to transport a patient to a hospital leave location i. The remainder of the $x_i - x_i^n$ ambulances that complete their treatment on scene (that is, $(x_i - x_i^n) - (y_i - y_i^n)$ ambulances) finished serving patients that do not need transportation, of which there are $b_i - z_i$. Hence,

$$\mathbb{P}\{y_i' = y_i^n|s, x_i^n, 0\} = \frac{\dbinom{b_i - z_i}{(x_i - x_i^n) - (y_i - y_i^n)}\dbinom{z_i}{y_i - y_i^n}}{\dbinom{b_i}{x_i - x_i^n}},$$

where we define $\binom{K}{\kappa} = 0$ if $\kappa < 0$ or $\kappa > K$. Note that $x_i - x_i^n \leq b_i$, so the denominator is always positive.

If we consider one extra request, then $x_i - x_i^n + 1$ ambulances finish their service on scene. Then, if $x_i - x_i^n + 1 \leq b_i$, it holds that

$$\mathbb{P}\{y_i' = y_i^n | s, x_i^n, 1\} = \frac{\begin{pmatrix} b_i - z_i \\ (x_i - x_i^n + 1) - (y_i - y_i^n) \end{pmatrix} \begin{pmatrix} z_i \\ y_i - y_i^n \end{pmatrix}}{\begin{pmatrix} b_i \\ x_i - x_i^n + 1 \end{pmatrix}}.$$

If $x_i - x_i^n + 1 > b_i$, we define $\mathbb{P}\{y_i' = y_i^n | s, x_i^n, 1\}$ to be 0. However, if this is the case, $\mathbb{P}\{1 | s, x_i^n\} = 0$, so the second term in Equation (2.13) vanishes.

Note that for $h \in \mathcal{H}$, it holds that $y_h^n \geq y_h$, since we restrict ourselves to the action in which none of the idle ambulances leave for a neighbour. However, $D(h, 1)$ occupied ambulances arrive at h in the next time step. Therefore, $y_h^n = y_h + D(h, 1)$ for each scenario $s^n \in \mathcal{S}(s)$. Hence,

$$\mathbb{P}\{y_h' = y_h^n | s\} = \begin{cases} 1 & \text{if } y_h^n = y_h + D(h, 1), \\ 0 & \text{else.} \end{cases}$$

Using Equation (2.12), we can now compute the transition probabilities corresponding to the second component.

3. The number of busy ambulances per demand location required to transport patients. We now compute

$$\mathbb{P}\{z' = z^n | s, x^n, y^n\} = \prod_{i=1}^{N} \mathbb{P}\{z_i' = z^n | s, x_i^n, y_i^n\}.$$

We know that $\mathbb{P}\{z_h' = 0 | s, x^n, y^n\} = 1$, so we consider the case $i \in \bar{\mathcal{H}}$. Remember that d_i^n denotes the number of ambulances that start a new treatment on scene at i in scenario n. As before, we make a distinction whether *no* or *one* extra request is considered at i in scenario n. Let $d_i^n(u)$, $u = 0, 1$, denote the same quantity, but conditioned on the number of additional requests considered. Then, similar to what was done in the previous section, we find that

$$d_i^n(0) = \min\{(x_i - x_i^n) - (y_i - y_i^n) + f_i, x_i - b_i\},$$

using Equation (2.4). The first part corresponds to the number of ambulances that possibly can start a new treatment, while the second part equals the number of requests not treated by an ambulance at the moment. The number of ambulances that complete their treatment on scene is $x_i - x_i^n$, of which $y_i - y_i^n$ leave for a hospital, occupied by a patient. Besides, all idle ambulances at i can start a new service. Moreover, there are no incoming requests, so the treatment of $x_i - b_i$ patients could be started if there were enough ambulances. These $d_i^n(0)$ ambulances all make a diagnosis whether the patients they are serving need transportation. Therefore,

$$z_i - (y_i - y_i^n) \leq z_i^n \leq z_i - (y_i - y_i^n) + d_i^n.$$

The number of patients for which it is decided that they need transportation is binomially distributed on $d_i^n(0)$ trials. Remember that the probability of transportation is r. Note that

$$d_i^n(1) = \min\{(x_i - x_i^n + 1) - (y_i - y_i^n) + f_i, 1 + x_i - b_i\} = d_i^n(0) + 1,$$

again by using Equation (2.4). Then, for $u = 0, 1$:

$$\mathbb{P}\{z_i' = z_i^n | s, x_i^n, y_i^n, u\} = \begin{cases} \binom{d_i^n(u)}{j} r^j (1-r)^{d_i^n(u)-j} & \text{if } z_i^n = z_i - (y_i - y_i^n) + j, \\ & 0 \leq j \leq d_i^n(u), \\ 0 & \text{else,} \end{cases}$$

and

$$\mathbb{P}\{z_i' = z_i^n | s, x_i^n, y_i^n\} = \sum_{u=0}^{1} \mathbb{P}\{z_i' = z_i^n | s, x_i^n, y_i^n, u\} \mathbb{P}\{u | s, x_i^n\}.$$

4. The elapsed service time of ambulances in phases 2 and 4. Computing $\mathbb{P}\{Z' = Z^n | s, x^n, y^n\}$ requires more work. Let $h \in \mathcal{H}$ and denote the h-th row of Z by $Z(h)$. Then

$$\mathbb{P}\{Z' = Z^n | s, x^n, y^n\} = \prod_{h \in \mathcal{H}} \mathbb{P}\big\{Z'(h) = Z^n(h) | s, x_h^n, y_h^n\big\}.$$

We assume that $Z(h)$ is always sorted in non-increasing order. That is, the first b_h entries of $Z(h)$ denote the elapsed service times at h, and the remainder of the row equals -1. However, if an ambulance ends the treatment of a patient, its past service time is excluded from $Z(h)$. In other words, there is an extra -1. But since we assume Z' is sorted in non-increasing order, this -1 is placed among the last entries of $Z'(h)$. Thus, $Z'(h,j)$ does possibly not correspond to the same ambulance to which $Z(h,j)$ corresponds.

Let $\hat{Z}(h) \sim Z'(h)$, where the notation '$\sim$' means that if we sort $\hat{Z}(h)$ in non-increasing order, it equals $Z'(h)$. One can check that '$\sim$' indeed defines an equivalence relation. Moreover, if $\hat{Z}(h) \sim Z'(h)$, it holds that

$$\mathbb{P}\big\{\hat{Z}(h) = Z^n(h) | s, x_h^n, y_h^n\big\} = \mathbb{P}\big\{Z'(h) = Z^n(h) | s, x_h^n, y_h^n\big\}.$$

We divide $Z^n(h)$ in three parts. The first part consists of the first b_h entries corresponding to the ambulances that are treating a patient at h in s. The probability that ambulance j finishes its treatment is ρ_h^j defined in Equation (2.3), where $1 \leq j \leq b_h$. Then, we find that

$$\mathbb{P}\big\{\hat{Z}(h,j) = Z^n(h,j) | s\big\} = \begin{cases} \rho_h^j & \text{if } Z^n(h,j) = -1, \\ 1 - \rho_h^j & \text{if } Z^n(h,j) = Z(h,j) + 1, \\ 0 & \text{else.} \end{cases}$$

Note that we do not condition on x^n and y^n here since these determine how many ambulances end their treatments. Moreover, $y_h^n - y_h$ occupied ambulances arrive at h. Hence, $x_h^n - (y_h^n - y_h)$ of the b_h ambulances remain busy, while the rest finishes its treatment. Let $(Z(h,j))_{j \leq b_h}$ denote the first b_h entries of $Z(h)$. We denote the number of patients at h for which the treatment is *not* completed in scenario n by G_h^n. If $\sum_{j=1}^{b_h} \mathbb{1}_{\{Z^n(h,j)>0\}} = x_h^n - (y_h^n - y_h)$, then

$$\mathbb{P}\left\{ \left(\hat{Z}\left(h,j\right)\right)_{j\leq b_h} = \left(Z^n\left(h,j\right)\right)_{j\leq b_h} \middle| s, x_h^n, y_h^n \right\} = \frac{\prod_{j=1}^{b_h} \mathbb{P}\{\hat{Z}(h,j) = Z^n(h,j)|s\}}{\mathbb{P}\{G_h^n = x_h^n - (y_h^n - y_h)\}},$$

and 0 if this is not the case. The second part corresponds to the $D(h,1)$ arriving ambulances at h. Therefore, $\hat{Z}(h,j) = 0$ for $b_h + 1 \leq j \leq b_h + D(h,1)$. Hence,

$$\mathbb{P}\{\hat{Z}(h,j) = Z^n(h,j)|s, x_h^n, y_h^n\} = \begin{cases} 1 & \text{if } Z^n(h,j) = 0, \\ 0 & \text{else.} \end{cases} \tag{2.14}$$

In the last part, $b_h + D(h,1) + 1 \leq j \leq A$, and thus $\hat{Z}(h,j) = -1$. Therefore,

$$\mathbb{P}\{\hat{Z}(h,j) = Z^n(h,j)|s, x_h^n, y_h^n\} = \begin{cases} 1 & \text{if } Z^n(h,j) = -1, \\ 0 & \text{else.} \end{cases} \tag{2.15}$$

For $i \in \bar{\mathcal{H}}$, computing $\mathbb{P}\{\hat{Z}(i,j) = Z^n(i,j)|s, x_i^n, y_i^n\}$ differs slightly. The first part, for $j \leq b_i$, is similar. For the second part, Equation (2.14) holds for $b_i + 1 \leq j \leq b_i + d_i^n$, since d_i^n ambulances start a new treatment. Consequently, Equation (2.15) holds for $b_i + d_i^n + 1 \leq j \leq A$.

5. Destinations and remaining driving times of ambulances in phase 3. To compute $\mathbb{P}\{D' = D^n|s, x^n, y^n\}$, we again consider $h \in \mathcal{H}$. All ambulances that were already driving to h have progressed one unit distance at the next time, which is the length of one edge. Hence,

$$\mathbb{P}\{D'(h) = D^n(h)|s, x_h^n, y_h^n\} = \prod_{h\in\mathcal{H}} \prod_{t=1}^{L} \mathbb{P}\{D'(h,t) = D^n(h,t)|s, x_h^n, y_h^n\}.$$

Remember that for $i \in \bar{\mathcal{H}}$, $y_i - y_i^n$ ambulances leave for a hospital, all to hospital h for which $i \in H^*(h)$. Therefore,

$$D^n(h,t) = D(h,t+1) + \sum_{i\in\bar{\mathcal{H}}} (y_i - y_i^n) \mathbb{1}_{\{i\in H^*(h)\}} \mathbb{1}_{\{l(i,h)=t\}},$$

implies that

$$\mathbb{P}\{D'(h,t) = D^n(h,t)|s, x_h^n, y_h^n\} = 1,$$

and 0 if this is not the case. Now, we have described all ingredients to compute $\mathbb{P}\{s' = s^n|s\}$, the probability that scenario $s^n = (x^n, y^n, z^n, Z^n, D^n)$ occurs.

2.3.2 Response Time Expectations

Up to know, we assumed that the action we take is the action in which none of the ambulances moves, except for those transporting a patient to a hospital. In this section, we drop this assumption and we combine each scenario with each feasible action, which results in a potential new state. However, the expected response time to the additional request in this potential state is of more interest to us than the potential state itself. We consider the following ambulances as eligible for responding to this patient:

(I) The nearest idle unassigned ambulance,

(II) The nearest busy ambulance(s) transferring a patient at a hospital,

(III) The nearest busy ambulance(s) treating on scene, of which it is known that it is not required to transport a patient.

Note that we do not consider ambulances that are transporting patients. These ambulances will be busy for a deterministic remaining driving time and a stochastic treatment time at the hospital. Therefore, it probably takes ample time before they can be assigned to another incident. For the same reason, we assume ambulances treating on scene that know that they have to transport the patient they are serving, are not eligible. Formally, let $s = (x, y, z, Z, D)$ be our current state, where $y = f + b$, and let $a^s \in \mathcal{F}(s)$ be the action we take. Consider scenario s^n. We define $\tilde{s}(s^n, a^s)$ as the state in which action a^s has been carried out in scenario s^n. Moreover,

$$\tilde{s}(s^n, a^s) = \big(\tilde{x}(s^n, a^s), \tilde{y}(s^n, a^s), \tilde{z}(s^n, a^s), \tilde{Z}(s^n, a^s), \tilde{D}(s^n, a^s)\big),$$

and we omit the dependence on s^n and a^s in the remainder. The state componentes are defined as follows:

$$\tilde{x}_i = b_i^n + \mathbb{1}_{\{x_i^n = x_i + 1\}}, \;\; i \in \mathcal{N},$$

i.e., in $\tilde{x}$ only the patients that are being treated on scene and the additional waiting patient are considered. That is, we do not consider the waiting patients that were already present in s. Moreover,

$$\tilde{y}_i = y_i^n + a_i^s - \max\{0, \min_{\alpha^s \in \mathcal{F}(s)} \alpha_i^s\}, \;\; i \in \mathcal{N},$$

in other words, only the eligible ambulances mentioned are considered. Note that if $\min_{\alpha^s \in \mathcal{F}(s)} \alpha_i^s > 0$, each feasible action dispatches at least one ambulance to location i. Hence, location i is on the shortest path to a node where a patient waits. We do not consider ambulances traveling to waiting patients as eligible ones and therefore exclude them from $\tilde{y}$. Furthermore, $\tilde{z} = z^n$, $\tilde{Z} = Z^n$, and $\tilde{D} = D^n$. These state components do not depend on the action taken. We now compute the expected response time for the additional patient in $\tilde{s}$ from the eligible ambulances. All eligible ambulances are observed in $\tilde{y}$. Denote these response times from the ambulances defined in (I), (II), and (III) by $R^{(\iota)}(\tilde{s})$, where $\iota \in \{\text{I, II, III}\}$. We compute $\mathbb{E}\{R^{(\iota)}(\tilde{s})\}$ for the additional patient in $\tilde{s}$ as follows:

(I). The determination of $\mathbb{E}\{R^{(I)}(\tilde{s})\}$ is easy, since there is no randomness involved in its computation. The response time for the patient waiting from the nearest idle unassigned ambulance is just the travel time from the current location of the ambulance to the waiting patient. Assume that the additional patient in scenario n is at location i. Then,

$$\mathbb{E}\big\{R^{(I)}\,(\tilde{s})\,\big\} = \min_{j:\tilde{y}_j > \tilde{x}_j} l(j,i),$$

using that if for location j it holds that $\tilde{y}_j > \tilde{x}_j$, we have an idle unassigned ambulance at j.

(II). Now we compute $\mathbb{E}\{R^{(II)}(\tilde{s})\}$. Of all hospitals, we consider the nearest hospital with at least one busy ambulance. Possibly, there are more busy ambulances at this hospital. The expected response time from one of these ambulances consists of two parts: the expected time until at least one ambulance finishes its treatment and a deterministic travel time from the hospital to the additional patient. Assume that the patient waiting in $\tilde{s}$ is at location i. Moreover, suppose that hospital h is the nearest hospital with at least one busy ambulance. Assume that $\tilde{b}_h$ ambulances are busy at h. The elapsed service time of ambulance j at hospital h is given by $\tilde{Z}(h,j)$. For each of these $\tilde{b}_h$ ambulances, we can compute the probabilities that they finish their treatment in *exactly* t time units from now. That is, we compute

$$\rho^j_h(t) = \mathbb{P}\big\{B^j_h = \tilde{Z}(h,j) + t | B^j_h > \tilde{Z}(h,j)\big\},\ t \geq 1.$$

Now, define $T(h)$ to be the number of time steps it takes for *at least* one busy ambulance at h to complete its service. Then,

$$\mathbb{P}\big\{T(h) = 1|\tilde{Z}(h)\big\} = 1 - \prod_{j=1}^{b_h}\left(1 - \rho^j_h\,(1)\right),$$

which is the probability that at least one ambulance ends its treatment after exactly one time unit from now. We can generalize this to t time units as follows:

$$\mathbb{P}\big\{T(h) = t|\tilde{Z}(h)\big\} = \left(1 - \prod_{j=1}^{b_h}\left(1 - \rho^j_h\,(t)\right)\right)\left(1 - \sum_{\tau=1}^{t-1}\mathbb{P}\{T(h) = \tau|\tilde{Z}(h)\}\right), \tag{2.16}$$

where $t \geq 1$ and the last part corresponds to the probability that none of the busy ambulances at h finished its treatment before time t. Now, we compute

$$\mathbb{E}\{T(h)|\tilde{Z}(h)\} = \sum_{t=1}^{\infty} t\ \mathbb{P}\{T(h) = t|\tilde{Z}(h)\}, \tag{2.17}$$

which is the expected time until an ambulance at h ends its service if the system is in state $\tilde{s}$. The expected response time to the additional patient from ambulance(s) at the nearest hospital with at least one busy ambulance is given by

$$\mathbb{E}\{R^{(II)}(\tilde{s})\} = \mathbb{E}\{T(h)|\tilde{Z}(h)\} + l(h,i), \tag{2.18}$$

where we assume that the additional patient in scenario n is at location i and $h = \arg\min\{l(i,h) : \tilde{b}_h > 0, h \in \mathcal{H}\}$, i.e., the nearest hospital with at least one busy ambulance. If no such h exists, we define $\mathbb{E}\{R^{(II)}(\tilde{s})\} = \infty$.

(III). The term $\mathbb{E}\{R^{(III)}(\tilde{s})\}$ consists of two parts as well: the expected time until an ambulance finishes its treatment and a deterministic travel time. The computation of $\mathbb{E}\{R^{(III)}(\tilde{s})\}$ is similar to $\mathbb{E}\{R^{(II)}(\tilde{s})\}$, and we assume $\mathbb{E}\{R^{(III)}(\tilde{s})\} = \infty$ if there is no ambulance *not* required to transport, while treating a patient on scene.

Given $\tilde{s}$, we compute the shortest expected response time to the additional patient in s^n that is possible from the eligible ambulances. Let this quantity be defined by $\mathbb{E}\{R(\tilde{s})\}$: it reassembles the shortest response time possible if each busy ambulance would finish its treatment, either on scene or at the hospital, now. This quantity is given by

$$\mathbb{E}\{R(\tilde{s})\} = \min_{\iota \in \{\mathrm{I,II,III}\}} \mathbb{E}\{R^{(\iota)}(\tilde{s})\}$$

and it equals zero if and only if at the location of the additional patient there is an idle unassigned ambulance as well.

2.3.3 Action Selection

Consider state s and $\mathcal{F}(s)$, which is the set of feasible actions as computed by solving one of the two assignment problems mentioned in Section 2.2.2. For each feasible action, we compute the penalty of the weighted average shortest expected response time to an arriving request, using the penalty function Φ introduced in Section 2.2.4. This quantity serves as the measure for the effect the action has on the EMS system. For action a^s, we denote this quantity by $V(a^s)$ and we compute it by

$$V(a^s) = \sum_{n=1}^{|\mathcal{S}(s)|} \Phi\Big(\mathbb{E}\Big\{R\big(\tilde{s}(s^n, a^s)\big)\Big\}\Big)\mathbb{P}\{s' = s^n|s\}, \tag{2.19}$$

where $\mathbb{E}\{R(\tilde{s})\}$ and $\mathbb{P}\{s' = s^n|s\}$ are described in Sections 2.3.2 and 2.3.1, respectively. Note that $\mathbb{E}\{R(\tilde{s})\}$ is not necessarily integer, so for this reason Φ needs to be continuous. We compute Equation (2.19) for each action in $\mathcal{F}(s)$. Then, we select the action $a^s \in \mathcal{F}(s)$ for which

$$a^s = \arg\min_{\alpha^s \in \mathcal{F}(s)} V(\alpha^s).$$

That is, the action that minimizes the weighted average shortest expected response time to an arriving request in the upcoming time period, is taken.

2.3.4 Theoretical Weaknesses

We end this section with the discussion of two theoretical weaknesses of the heuristic described above: situations in which the heuristic might perform poorly. A first limitation of the method is that it considers only the nearest of the eligible ambulances: an ambulance driving from one town to another is not observed by the heuristic if in both towns an ambulance is present. A solution in which this ambulance is observed might result in a better policy. However, this only plays a role if the probability of having multiple busy ambulances per town is large, which is typically not the case in the rural regions we observe.

Moreover, another possible weakness of this heuristic is that only an ambulance configuration in the 'neighborhood' of the current configuration can be attained in the next time step. This is a consequence of the fact that ambulances cannot traverse more than one edge per time unit. Therefore, the best action selected might not lead closer to the global optimal ambulance configuration. This is illustrated in the following small example.

Consider a chain with five equidistant nodes, where nodes 1 and 5 represent points of relatively high demand. The demand in the middle nodes 2, 3 and 4 is very low, as is typically the case in rural regions. Moreover, assume that the demand in node 5 is significantly higher than the demand in node 1. We use the penalty function of Figure 2.3a with $T_{max} = 1$ and we assume we have only one ambulance. The global optimal solution is to locate the ambulance at node 4, since it covers nodes 3, 4 and the high demand of node 5. However, if the ambulance is at node 1, the ambulance ends up in node 2. This is a consequence of the fact that the action of traversing the edge between nodes 2 and 3 is classified as a bad action, because if the ambulance is in node 3, then neither node 1 nor node 5 is covered. That is, instead of the global optimal configuration, a local optimal configuration is attained. However, instead of a weakness one can also interpret this as a strength, because attaining a local optimal configuration involves less driving. To investigate this, we compare the heuristic to a policy that focuses on attaining the global optimal configuration: the compliance table policy.

Two other assumptions that could impact the performance of the algorithm are the limitation of the scenarios with only one additional incident and the one-step lookahead. Relaxing these assumptions seriously increases the computation time and the question arises whether this is beneficial. This is probably not the case, since we focus on rural regions and as a consequence, the probability that two consecutive requests arrive in a short period of time, is relatively small. Results in Section 2.4.2 below, in which we compare the heuristic with the optimal policy for a small example, show indeed that the heuristic performs near-optimal for the performance indicator related to the chosen penalty function.

2.4 Numerical Results

The heuristic described in the previous section computes for each state an action in which the expected penalty is minimized. We call the policy obtained by performing the heuristic the *heuristic policy.* We compare it to a different policy: the compliance table policy, which we will explain in the next subsection.

2.4.1 Compliance Tables

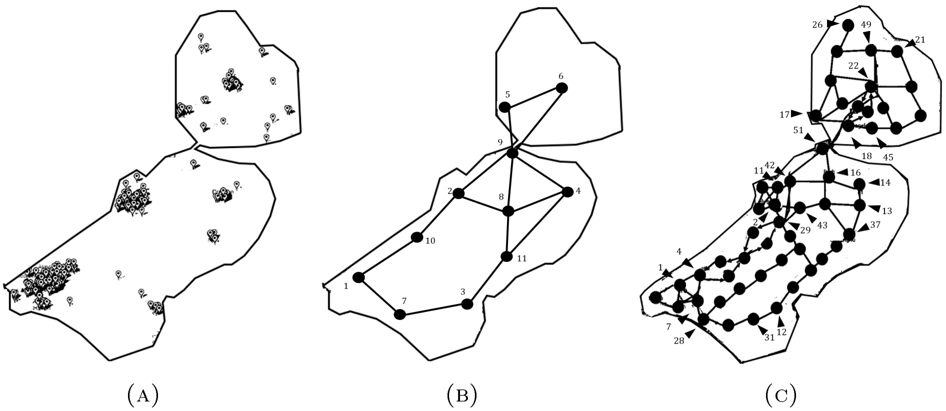
Compliance table policies are commonly used in practice for dynamic ambulance management (see Alanis et al. (2013), Gendreau et al. (2006)). Each row in a compliance table shows, for a given number of idle ambulances, the desired locations for these ambulances. If these ambulances are at their desired location, the system is *in compliance.* The number of idle ambulances changes when a request arrives or when an ambulance becomes idle again. Then, each idle ambulance is assigned to a possible new location. That is, in state s we first solve an assignment problem to assign ambulances to requests. After that, we solve a second assignment problem for the unassigned ambulances and desired locations. In our computations, we used LBAP for both assignment problems. Moreover, we assume that each ambulance *immediately* starts driving to its desired location.

We want to compare the heuristic described in Section 2.3 to a good compliance table with respect to the chosen penalty function. After all, for different penalty functions, compliance tables may differ. We assume that no more than one ambulance is deployed at a single location, because our setting is a rural region with a small number of ambulances. The arrival rate of incidents is low, and thus it is very unlikely that a second incident occurs just after the first in a certain area. We generate compliance tables by solving a static optimization problem for each level independently: the *p-median* problem.

In the p-median problem, which was formulated as an integer linear program by ReVelle and Swain (1970), one aims to find the location of a fixed number of facilities so as to minimize the weighted average distance. In the context of dynamic ambulance management, this translates to finding the location of the idle ambulances in such a way that the weighted sum over each node of the distance from the node to the nearest ambulance is minimized. Remember that $l(i, j)$ is the length of the shortest path between nodes i and j, and p is the parameter of the Poisson distribution that models the number of arriving requests. However, we do not use the shortest-path lengths itself, but the penalties corresponding to these to incorporate the penalty function of interest. The objective function is as follows:

$$\text{Minimize} \sum_{i \in \mathcal{N}} \sum_{j \in \mathcal{N}} p_i \Phi\Big(l(i, j)\Big) Y_{ij}, \tag{2.20}$$

where Y_{ij} is a binary decision variable: $Y_{ij} = 1$ if and only if a request at node i is served by an ambulance at node j, i.e., if the ambulance at j is the closest ambulance to node i. In addition, we introduce a binary decision variable X_j which equals one if an ambulance is placed at location j. Assume that there are F idle ambulances. Thus, we compute the F-th row of the compliance table. We



(A) (B) (C)

Figure 2.4: EMS region of Flevoland: spatial distribution of requests (A), simplified graph model (B), and extended graph model (C).

minimize Equation (2.20) under the following constraints:

$$
\begin{aligned}
&\sum_{j\in\mathcal{N}} Y_{ij} = 1 && i \in \mathcal{N}\\
&\sum_{j\in\mathcal{N}} X_j = F && \\
&Y_{ij} \leq X_j && i, j \in \mathcal{N}\\
&Y_{ij}, X_j \in \{0, 1\}. &&
\end{aligned}
$$

The first constraint states that each request has exactly one ambulance that is nearest. We need to find the desired locations of F ambulances, which is given by the second constraint. The third constraint induces that an ambulance at j can serve a request at node i only if j is a desired location. For each value of F, $1 \leq F \leq A$, we solve this p-median problem.

Note that there is no cohesion between the compliance table levels by applying this procedure. After all, the problem is solved for each level independently. Therefore, compliance table k aims for the optimal global configuration with k available ambulances, not a near-optimal one that can be attained faster in order to be in compliance earlier. However, by the same reasoning as before, the time between consecutive incidents is supposed to be large. As a consequence, it is justified to assume that there is enough time to attain the optimal configuration for this number of available ambulances before a next incident occurs. In short, given that no more than one ambulance is placed at a single location and each level is computed independent of each other, this procedure computes the optimal compliance table.

Level	Compliance Table
1	1
2	1-2
3	1-2-9
4	1-2-4-6

Table 2.4: Compliance table of the simplified example.

2.4.2 Case Study 1

To gain insight into the performance of both the heuristic and the compliance table policy, we first compare them to the optimal policy in an illustrative example. We apply these three policies to an EMS system belonging to a rural region in the Netherlands: Flevoland. A map of this region, as well as the spatial distribution of requests, is displayed in Figure 2.4a. We set Δt equal to 15 minutes and model the region by the graph in Figure 2.4b with 11 nodes and 15 edges. On average, there are 28.6 requests per day in our problem instance. There are six nodes with a non-zero arrival parameter, which varies between 1.2 and 15.1 requests per day. We consider an instance with four ambulances and we assume that none of the patients has to be transported to a hospital in our example. The treatment time on scene follows a geometric distribution with parameter 0.3. This results in a mean treatment time on scene of 50 minutes. These two simplifications greatly reduce the size of the state space, as now a state is described by the first two components only: (x, y). In order to compute the optimal policy, we truncate the state space by assuming that the maximum number of requests, denoted by $\bar{X}$, is five: $\sum_{i=1}^{N} x_i \leq \bar{X} = 5$. This results in a state space of 630,630 18-dimensional states, computed by

$$\sum_{i=0}^{\bar{X}} \binom{N' + i - 1}{N' - 1} \binom{A + N - 1}{N - 1},$$

in which there are $N' \leq N$ nodes with a non-zero arrival parameter, A ambulances and N nodes in total. Here, $\bar{X} = 5$, $N' = 6$, $A = 4$ and $N = 12$.

We model the problem as an MDP for the linear penalty function $\Phi(t) = t$, and solve it using Value Iteration (c.f. Puterman (1994)). We use LBAP to compute the set of feasible actions. The average size of the set of feasible actions is 1.9 actions. There are many states in which we only allow one action, namely the states with 4 or 5 requests in total. The maximum number of feasible actions in a state is 321, which obviously was a state without any request. The compliance table, as computed by solving the p-median problem, is displayed in Table 2.4. We simulate this table, the optimal policy and the heuristic policy for one million time steps. Results on late arrivals, response times and driving ambulances, as well as their 95% confidence bounds, are displayed in Table 2.5. The fraction of late arrivals represents the fraction of requests for which a maximum allowed response time of 15 minutes (1 time unit) is exceeded. The mean response time is expressed in time units. In the computation of the mean number of driving

Policy	Performance Statistics	Mean	95%-CI
Optimal	Fraction late arrivals	1.68%	[1.55%, 1.65%]
	Mean response time (time units)	0.0587	[0.0568, 0.0607]
	Mean no. driving ambulances	0.6280	[0.6259, 0.6301]
Heuristic	Fraction late arrivals	1.95%	[1.90%, 2.00%]
	Mean response time (time units)	0.0590	[0.0571, 0.0609]
	Mean no. driving ambulances	0.7232	[0.7213, 0.7251]
Compliance Table	Fraction late arrivals	2.22%	[2.16%, 2.27%]
	Mean response time (time units)	0.0630	[0.0611, 0.0649]
	Mean no. driving ambulances	0.9305	[0.9281, 0.9328]

Table 2.5: Results for the simplified example.

ambulances, ambulances traveling to a call, transporting a patient to a hospital and ambulances relocating themselves are included.

As expected, the optimal policy outperforms the other two policies on the performance measure related to the penalty function, although the differences between the mean response time induced by the optimal and heuristic policy are really close and their 95% confidence bounds overlap almost entirely. As a consequence, on this performance criterion the heuristic policy is a near-optimal policy. This shows that the two main assumptions stated at the end of Section 2.3.4, namely the limitation to scenarios with only one additional request and the one-step lookahead, have a very small impact on the performance. Relaxing these assumptions will seriously increase the computation time while there is little room for improvement. The optimal policy performs better on the two other performance indicators as well. It is also worth noting that the performance gap between the optimal and heuristic policy is smaller than the gap between the heuristic and the compliance table policy for all performance measures.

If we compare the results of the heuristic and the compliance table policy in Table 2.5, we observe that the heuristic policy outperforms the compliance table on any of the three performance criteria. The difference on mean number of driving ambulances is explained by the fact that there is a drift to node 1 in the compliance table, because node 1 has the highest call arrival rate. Together with the fact that in this node a hospital is present, many ambulances become idle again here. The heuristic takes this into account by considering ambulances transferring a patient at a hospital as eligible ones as well. In contrast, the compliance table of Table 2.4 sends an ambulance from elsewhere to node 1 each time the ambulance present in node 1 is dispatched, which happens relatively much due to the high arrival rate. This results in a relatively large amount of driving.

2.4.3 Case Study 2

We apply both the heuristic policy and the compliance table policy to a more realistic model of the EMS region of Flevoland. We set Δt equal to 5 minutes, and we model the region by the graph in Figure 2.4c, with 57 nodes and 74 edges. This

Level	Compliance Tables		
	Equation (2.7)	Equation (2.8)	Equation (2.9)
1	29	51	51
2	1-42	28-51	28-51
3	1-11-22	16-22-28	22-28-43
4	1-2-14-22	2-16-21-28	1-22-31-43
5	1-2-12-14-22	1-2-12-16-21	1-12-14-22-43
6	1-2-12-14-17-22	1-2-4-12-18-24	1-2-12-13-22-26
7	1-2-12-13-14-17-22	1-2-4-7-12-18-24	1-2-12-13-18-45-49

	Equation (2.10)	Equation (2.11)
1	29	29
2	28-51	28-51
3	16-21-28	22-28-43
4	2-22-28-51	1-12-22-43
5	1-2-12-16-22	1-2-12-14-22
6	1-2-6-12-22-37	1-2-12-14-17-22
7	1-2-6-7-12-16-22	1-2-12-13-14-17-22

Table 2.6: Compliance tables.

time of 5 minutes corresponds to a road distance of approximately 5 kilometers in the towns and to 8 kilometers in the rural areas. There are two hospitals in the region, one in the city in the Southwest and one in the western city in the middle. We use historical data to estimate the several distributions needed. The node-dependent arrival parameter of requests varies between 0.12 and 4.3 requests per day. On average, there are 24.2 requests per day. For the on-scene time we estimate a geometric distribution with a mean of approximately 10 minutes, and a standard deviation of 7 minutes. The hospital treatment time follows a Discrete Weibull distribution. The mass-function of the Discrete Weibull distribution with parameters μ and k is given by

$$\mathbb{P}\{X = x\} = (1-\mu)^{x^k} - (1-\mu)^{(x+1)^k}, \ x = 0, 1, 2, \ldots,$$

and is treated in detail by Rinne (2008). Here, $\mu = 0.1$ and $k = 2$, which results in a mean treatment time at the hospital of approximately 16 minutes and a standard deviation of 7.3 minutes. Moreover, 75% of the patients needs to visit a hospital, so $r = 0.75$. We consider cases with four ambulances and with seven ambulances, the latter being more realistic for this region.

We compute compliance tables for the five different penalty functions considered in Equations (2.7)–(2.11) in Section 2.2.4, where we take $\beta = 10$, $\gamma = 200$ and $T_{max} = 3$ time units (15 minutes). These functions are displayed in Figure 2.3 as well. The computed compliance tables are displayed in Table 2.6. Note that for Equation (2.7) and Equation (2.8), computing the compliance tables is equivalent to solving A classical p-median problems and A MCLP-problems, respectively.

We use LBAP to compute the set of feasible actions. As was stated at the end of Section 2.2.2, we incorporate the penalty function in this assignment problem. We simulate 100,000 time steps and observe from the simulations values for four different performance indicators, and their 95% confidence bounds. Results are displayed in Tables 2.7 and 2.8.

Observing these tables, one can make several interesting observations. In terms of penalty, the penalty function minimizing the number of late arrivals of Equation (2.8) combined with $A = 4$ is the only penalty function for which the heuristic policy performs worse than the compliance table policy. This is probably due to the fact that the heuristic policy only considers ambulance configurations that can be attained in one time step. As a consequence of the small differentiation in penalty for several response times, many actions are classified as equally good. This is also reflected in the fact that although Equation (2.8) focuses on minimizing the fraction of late arrivals, it is dominated on this criterion by three out of the four other penalty functions in the case with four ambulances. Specifically, the penalty function of Equation (2.9), that hardly differs from the one in Equation (2.8), performs much better on the fraction of late arrivals for the heuristic policy. These phenomena do not occur in the case with seven ambulances. After all, the action set is much larger in this case. Besides, with seven ambulances there are more opportunities to cover low demand points as well. Hence, there is more diversity in the classification of actions.

In general, the heuristic policy performs better than the compliance table policies on the mean response time for each of the considered penalty functions and cases, although the difference is not significant for the penalty function of Equation (2.7). This is explained by the fact that the heuristic focuses on the shortest expected response time from the eligible ambulances. In contrast to minimization of the fraction of late arrivals, the penalty function that focuses on minimizing the average response time, performs best on that criterion for both policies. The largest gap in terms of response times between the two policies is observed for the penalty function of Equation (2.8), in favour of the heuristic policy.

Comparing the fraction of late arrivals and the mean response times for each penalty function in the heuristic policy in Table 2.7, one may note that in the majority of the cases a shorter mean response time leads to an increase of the fraction of late arrivals. This is also the case for the compliance table policies both in Table 2.7 and Table 2.8. This negative correlation is in contrast to what one intuitively might expect. Note that this phenomenon is most clear in Table 2.7 in the compliance table policies for penalty functions (2.7) and (2.8). This is explained by the following reason. Equation (2.7) locates the ambulances close to the city centers of the two largest towns. As a consequence, some minor towns can not be reached within 15 minutes. Since approximately 56% of the incidents occurs in the two largest towns, especially in the city centers, this results in small response times to the areas of high demand. However, the response times to the areas of low demand are much larger, but this is only marginally noted in the mean.

Table 2.7: Main results for four ambulances.

Penalty Function	Performance Statistics	Heuristic Policy		Compliance Table	
		Mean	95%-CI	Mean	95%-CI
Equation (2.7)	Fraction late arrivals	12.44%	[12.19%, 12.70%]	14.08%	[13.82%, 14.33%]
	Mean response time (minutes)	7.7265	[7.6525, 7.8005]	7.8325	[7.7610, 7.9040]
	Mean no. of driving ambulances	0.5309	[0.5273, 0.5345]	0.8260	[0.8213, 0.8307]
	Mean penalty per time step	1.5453	[1.5305, 1.5601]	1.5665	[1.5522, 1.5808]
Equation (2.8)	Fraction late arrivals	10.64%	[10.47%, 10.81%]	7.78%	[7.52%, 7.98%]
	Mean response time (minutes)	8.0350	[7.9780, 8.0920]	10.707	[10.643, 10.772]
	Mean no. of driving ambulances	0.5851	[0.5806, 0.5896]	0.9132	[0.9073, 0.9190]
	Mean penalty per time step	0.1064	[0.1047, 0.1081]	0.0775	[0.0752, 0.0797]
Equation (2.9)	Fraction late arrivals	8.24%	[8.04%, 8.45%]	8.73%	[8.54%, 8.92%]
	Mean response time (minutes)	9.1170	[9.0525, 9.1815]	10.062	[9.9920, 10.132]
	Mean no. of driving ambulances	0.6537	[0.6495, 0.6578]	0.9521	[0.9461, 0.9581]
	Mean penalty per time step	0.0831	[0.0811, 0.0852]	0.0881	[0.0862, 0.0900]
Equation (2.10)	Fraction late arrivals	7.59%	[7.35%, 7.82%]	8.61%	[8.13%, 8.69%]
	Mean response time (minutes)	8.8175	[8.7430, 8.8925]	10.875	[10.816, 10.934]
	Mean no. of driving ambulances	0.7098	[0.7039, 0.7157]	0.9848	[0.9779, 0.9918]
	Mean penalty per time step	0.1659	[0.1589, 0.1729]	0.1948	[0.1877, 0.2018]
Equation (2.11)	Fraction late arrivals	7.70%	[7.55%, 7.84%]	8.67%	[8.49%, 8.85%]
	Mean response time (minutes)	8.8350	[8.7760, 8.8945]	9.2115	[9.1615, 9.2650]
	Mean no. of driving ambulances	0.7106	[0.7054, 0.7158]	0.9710	[0.9649, 0.9772]
	Mean penalty per time step	0.2758	[0.2690, 0.2825]	0.2908	[0.2838, 0.2978]

Table 2.8: Main results for seven ambulances.

Penalty Function	Performance Statistics	Heuristic Policy		Compliance Table	
		Mean	95%-CI	Mean	95%-CI
Equation (2.7)	Fraction late arrivals	1.96%	[1.46%, 2.46%]	2.51%	[2.05%, 2.96%]
	Mean response time (minutes)	3.5960	[3.3989, 3.7931]	3.7236	[3.5734, 3.8737]
	Mean no. of driving ambulances	0.3944	[0.3749, 0.4140]	0.8564	[0.8313, 0.8815]
	Mean penalty per time step	0.7192	[0.6798, 0.7586]	0.7447	[0.7147, 0.7747]
Equation (2.8)	Fraction late arrivals	1.18%	[0.82%, 1.55%]	1.91%	[1.65%, 2.18%]
	Mean response time (minutes)	3.6884	[3.5160, 3.8608]	6.642	[6.5310, 6.7530]
	Mean no. of driving ambulances	0.4147	[0.4012, 0.4283]	0.9325	[0.9093, 0.9557]
	Mean penalty per time step	0.0118	[0.0082, 0.0155]	0.0192	[0.0165, 0.0218]
Equation (2.9)	Fraction late arrivals	1.21%	[0.93%, 1.49%]	1.82%	[1.55%, 2.09%]
	Mean response time (minutes)	3.7744	[3.6364, 3.9125]	5.6210	[5.5400, 5.7020]
	Mean no. of driving ambulances	0.4508	[0.4373, 0.4644]	1.5192	[1.4949, 1.5436]
	Mean penalty per time step	0.0122	[0.0094, 0.0150]	0.0184	[0.0157, 0.0211]
Equation (2.10)	Fraction late arrivals	1.14%	[0.86%, 1.41%]	1.95%	[1.59%, 2.31%]
	Mean response time (minutes)	3.5967	[3.4657, 3.7278]	5.7025	[5.5840, 5.8210]
	Mean no. of driving ambulances	0.4259	[0.4140, 0.4378]	1.1518	[1.1294, 1.1741]
	Mean penalty per time step	0.0183	[0.0124, 0.0243]	0.0296	[0.0238, 0.0354]
Equation (2.11)	Fraction late arrivals	1.15%	[0.86%, 1.44%]	1.93%	[1.68%, 2.19%]
	Mean response time (minutes)	3.6441	[3.5264, 3.7617]	3.7811	[3.6812, 3.8811]
	Mean no. of driving ambulances	0.4348	[0.4213, 0.4482]	0.8972	[0.8777, 0.9167]
	Mean penalty per time step	0.0384	[0.0303, 0.0465]	0.0576	[0.0513, 0.0640]

In contrast, in the compliance table corresponding to penalty function (2.8) places ambulances in such a way that the demand that be reached within 15 minutes is maximized. Therefore, ambulances are further away from the areas of high demand, yielding a larger mean response time. An exception to this phenomenon is the heuristic policy in the case with seven ambulances. However, even in this case, the penalty function of Equation (2.7) induces the smallest response time, but the largest fraction of late arrivals.

Another interesting point is the number of driving ambulances. For each penalty function and each case, the heuristic policy greatly outperforms the compliance table policy on this performance indicator. This is caused by the fact that using compliance tables, one aims to attain a ambulance configuration only taking the number of available ambulances into account. In contrast, since ambulances can only traverse at most one edge per time unit, the heuristic computes a good local configuration. As a consequence, less driving is involved in using the heuristic policy. Moreover, comparing Tables 2.7 and 2.8, an increase in number of ambulances gives rise to a decrease of driving ambulances for the heuristic policy. In contrast, in the compliance table policy, more ambulances induce more driving in general, the penalty function of Equation (2.11) being the only exception.

If we compare Tables 2.7 and 2.8, we observe larger differences in patient-based results in the case with four ambulances. For instance, the fractions of late arrivals for the penalty functions of Equations (2.8)–(2.11) in the case with seven ambulances are very close to each other. In contrast, these differences in the case with four ambulances are much larger. Hence, a small change in setting (e.g., penalty function) may result in a large change in performance in such a case. This underlines what was stated in Section 2.1: if one has access to only a small number of ambulances, one has to be more careful about how to relocate them.

Apart from the first penalty function in the case with four ambulances, the heuristic outperforms the compliance table policy on each of the performance indicators. Therefore, it seems that attaining a good local ambulance configuration that can be reached quickly, performs better than attaining the desired configuration of ambulances supplied by the compliance table, which serves as a global configuration for this number of available ambulances.

2.5 Concluding Remarks

In this chapter, we proposed a DAM model for rural regions to solve the ambulance relocation problem. The model was formulated as a discrete-time Markov decision process. At each time step a relocation policy specifies, for each ambulance that is not busy, whether to move the ambulance to an adjacent node. A policy is sought that minimizes a general penalty function which is nondecreasing in the response time to a request. The function can be constructed to match the performance objectives of the EMS system being studied. Computation of the optimal policy in realistic settings is impractical, because the MDP has a high-dimensional state space. To address this, we developed a one-step look-ahead heuristic that, at each time step, relocates ambulances in order to minimize the expected response time for a possible call arriving in the next time step. We concluded this chapter with a numerical comparison of the performance of the heuristic policy to the optimal and to the compliance table policy. We observed that for the majority of the studied penalty functions, the heuristic policy outperformed the compliance table policy on most performance indicators.

3
The Penalty Heuristic and the Impact of Ambulance Relocations

Ambulance repositioning is generally believed to provide means to enhance the response time performance of emergency medical service providers. However, the implementation of DAM algorithms generally leads to additional movements of ambulance vehicles compared to the reactive paradigm. 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 in this chapter. We solve a linear bottleneck assignment problem to obtain the exact movements of ambulances from one configuration to a target configuration, 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.

This chapter is based on Van Barneveld et al. (2016a).

3.1 Introduction

In this chapter, we approach the ambulance relocation problem both from the patient's and from the crew's point of view. These perspectives are conflicting in the sense that patients desire the shortest possible response time at all costs. However, always striving for this objective may inconvenience the ambulance crew due to increased workloads and other reasons mentioned in Section 1.2, caused by additional relocations. The relationship between patient-based performance, i.e., response times and the number of ambulance relocations, which we regard as an appropriate measure for crew inconvenience, is complex. The consequences of moving an ambulance to a different base station are not known a priori, due to the 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 response time performance of the ambulance service provider. But even if this were the case, there is still a trade-off: would one carry out extra ambulance relocations for only a small gain in expected response times? 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 into the relationship between response times and the number of ambulance relocations are desirable.

To this end, we present an ambulance redeployment model, in which we incorporate different performance criteria by defining a suitable penalty function, like in Chapter 2. For this reason, we prefer to speak about the general notion of *(expected) performance* rather than about (expected) coverage or (expected) response times, specifically. We use a heuristic method, the so-called *penalty heuristic*, that computes an action concerning the relocation of ambulances in such a way that the expected performance is maximized. 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 were 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. In contrast to the previous chapter, relocation decision moments are not equally distributed over the time horizon. Instead, the dispatcher has the possibility to change the ambulance configuration at certain *events*. That is, the dispatcher may order ambulance crews to relocate when the number of available ambulances changes due to a vehicle dispatch or service completion.

Many authors of the papers in the EMS 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 expected performance by taking this action is very small. Possibly, this benefit does not outweigh the disadvantages regarding the number of additional ambulance relocations to achieve this gain. Therefore, we use the penalty heuristic to determine whether the redeployment action is really necessary, and we show results on several quantifications of 'really necessary'. Another important difference between the mainstream literature and this chapter is the way in which a redeployment action is carried out. We compute, using the penalty heuristic, 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 the origin to the destination, as assumed in most of the 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 idle ambulances that may be relocated simultaneously to obtain useful insights in the relationship between number of relocations and performance.

The remainder of this chapter is structured as follows. Section 3.2 describes the modified system dynamics (compared to the EMS process described in Chapter 2), introduces the main terminology used throughout this chapter, and states three problems dispatchers face in the emergency control center. Section 3.3 presents a heuristic method to solve these problems: the *penalty heuristic*. Section 3.4 describes the main experimental setup of our numerical study. Section 3.5 is concerned with the computational study, in which we use the penalty heuristic to study the trade-off between patient and crew based performance.

3.2 Model

In this section, we describe the EMS system dynamics and the modeling framework used in this chapter. The model of this chapter heavily differs from the one studied in the previous chapter. This is mainly caused by the difference in time scale: in Section 2.4 we considered time steps of 5 and 15 minutes. However, the model considered in this chapter is a continuous-time model. That is, at each moment in time, an emergency call can arrive. In that sense, the model described in this section is closer to practice compared to the MDP model of the previous chapter. Moreover, the locations at which an ambulance can idle is another difference: in the previous chapter, we assumed that it can stay at every node. In contrast, in the model in this chapter we require ambulances to return to a base station when idle.

3.2.1 System Dynamics

To investigate the relationship between number of relocations and expected performance, we slightly adjust the general EMS process described in Section 1.1. We assume that all incoming emergency calls are of the highest urgency, by a similar justification as given in Section 2.2. A large difference with the EMS model studied in the previous chapter, due to the different time scale, is the assumption on the arrival of emergency calls. We assume that this process is Poisson, i.e., the interarrival times are exponentially distributed. It is generally believed that this fits the EMS call arrival process quite well (Matteson et al., 2011).

Since all incidents are classified as A1-calls, the closest idle ambulance is always dispatched to a call. By 'closest', we mean closest in time. Note that this ambulance is not necessarily the closest one in space as well. In contrast to the previous chapter, in which ambulances could be reassigned to a different request during their response, this is not the case in the model in this chapter. When an ambulance is dispatched to an incident, we suppose that it starts driving immediately. This is a simplification of reality (see Figure 1.1), since the chute time, which is the time between the moment the mission is received by the crew and the moment of departure from the base station, is neglected. We do so because only the travel time to an incident is affected by the location of the ambulances. Instead of a maximum allowed response time threshold, we consider a maximum allowed travel time by subtracting both the dispatch and chute time from the

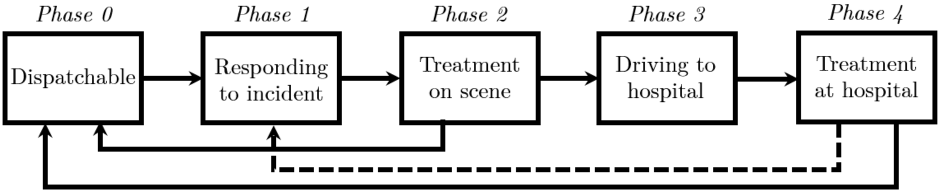


Figure 3.1: Ambulance phases.

response time threshold, which we consider both to be deterministic. However, chute times are typically smaller when ambulances are already on the road at the moment they are dispatched, so our assumption is a simplification of reality.

Due to the fact that ambulances can not be reassigned to respond to a different request during the response to a first one, we add an ambulance phase to the EMS process of Figure 2.1: we split phase 1 into phase 0 and phase 1. Figure 3.1 depicts a graphical representation of the new EMS process. Phase-0 ambulances are the ambulances currently not involved in the service of a patient, and thus, are the dispatchable ones. They are either at a base or executing a relocation. Moreover, an ambulance traveling to the emergency scene is supposed to be in phase 1. As before, phase-2 ambulances are the ones that are currently providing medical assistance to a patient on scene. If transportation is required, these ambulances enter phase 3, and we again assume that a patient is always transported to the nearest hospital. If this is not the case, phase 0 is entered. Phase-4 ambulances are currently involved in the drop-off of a patient.

Note that there is a dashed arrow from phase 4 to phase 1 in Figure 3.1. This is due to the following system characteristic: 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, i.e., a phase-4 ambulance. However, we only preempt if this ambulance is already more than a target time Δ busy with the transfer of this patient. That is, Δ 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 usually differs per ambulance service provider, but this is the case for the considered service provider in the numerical study in Section 3.5.

3.2.2 Ambulance Motions and Relocations

To ensure short response times to future incidents, dispatchers can proactively relocate ambulances to different base stations. We allow the dispatcher to make these decisions at the following moments, which we refer to as decision moments of the first and second type, respectively:

1. when an ambulance is dispatched to an incoming incident, and

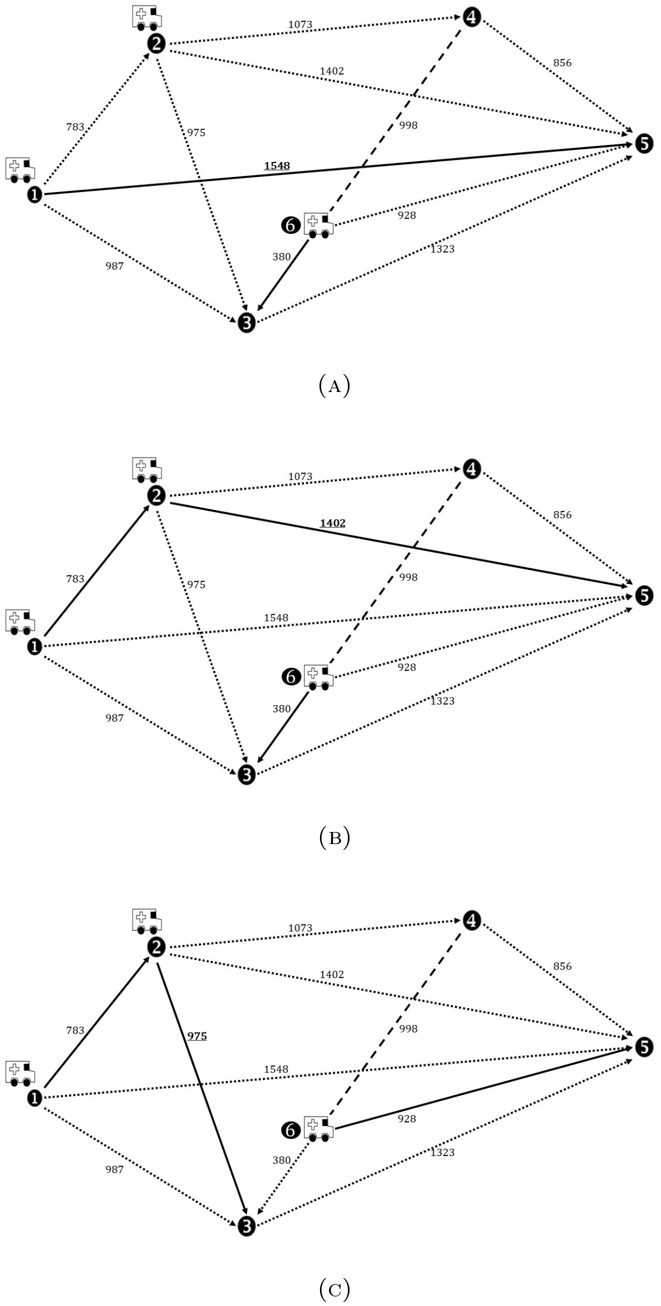
2. when an ambulance enters phase 0 again, either from phase 2 or phase 4.

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 location plan, 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 has 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 a single ambulance, it could be beneficial to break up the ambulance motion in two or more separate *ambulance relocations*, also called *chain relocations*. In this way, it is ensured that the new ambulance configuration is attained earlier. We refer to Figure 3.2 for an illustration of chain relocations.

In this illustration, full arcs denote the way in which ambulances are relocated. The numbers next to the arcs are the driving times in seconds. In all figures, 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 (Figure 3.2a). If one uses the ambulance at 2, this time can be reduced to 1,402 seconds, at the expense of one extra relocation (Figure 3.2b). In addition, if *redirection* is allowed, one can use the driving ambulance to decrease the time in which the new ambulance configuration is attained to 975 seconds (Figure 3.2c).

We assume that, like at the dispatch, the chute 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 Figure 3.2c,



(a)

(b)

(c)

Figure 3.2: Illustration of chain relocations.

neither counts as a relocation, as the crew is already en route. Note that there is a trade-off between the number of relocated ambulances and the time it takes to attain the new ambulance configuration.

3.2.3 Ambulance Relocation Problem

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 3.1. 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 has to 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.
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 Section 3.3 we present a heuristic method concerning these three problems.

3.2.4 Mathematical Model

In this section, we describe the mathematical model and introduce the notation. We model the region of interest as a weighted complete directed graph $G = (V \cup W, A, \tau)$. The region is discretized into geographical demand zones, e.g., municipalities, neighborhoods, postal codes or streets. We define V as the vertex set of these demand points. The fraction of demand occurring in node $i \in V$ is denoted by d_i, and we assume that incidents take place in a Poisson manner with rate λ. Hence, the arrival rate of incidents for node i equals λd_i. Let W be the set of potential waiting sites, $W \subseteq V$, and the number of ambulances is denoted by n. The road-network of the region is modelled by arcs $(i, j) \in A$, where $i, j \in V \cup W$. Moreover, τ_{ij} denotes the expected travel time (in seconds) between nodes i and j when driving with optical and sound signals turned on, typically used while responding to an emergency or the transportation of a patient to a hospital. These expected driving times are derived from a driving time table, estimated beforehand and thus assumed to be given. Table 3.1 depicts a brief overview of the used notation.

3.3 Penalty Heuristic

For the evaluation of the usefulness of ambulance motions and relocations, we present the penalty heuristic that can easily handle several types of restrictions

V	Set of demand points.
W	Set of potential waiting sites.
τ_{ij}	Expected travel time between nodes i and node j.
n	Number of ambulances.
d_i	Fraction of demand occuring in node $i \in V$.
λ	Incident arrival rate.
Δ	Target for hospital drop-off time.
Φ	Penalty function.
$\mathcal{A}^k(s)$	Number of ambulances in phase k in state s.
$des(j,s)$	Destination of ambulance j in state s.
$loc(j,s)$	Current location of ambulance j in state s.
$U(s)$	Unpreparedness level of state s.
Q	Motion threshold.
M	Bound on the number of ambulances in the relocation chain.

Table 3.1: Notation.

on the decisions of the dispatcher. First, we describe the heuristic method. Then, we 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 Section 3.3.1.

3.3.1 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. In the state, information about the current location of ambulances and the phases they are in, is captured. 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. Let $des(j,s)$ and $loc(j,s)$ denote the *destination* and *current location* of ambulance j

if the state of the system is s, respectively. We define

$$t_i^0(s) = \min_{j \in \mathcal{A}^0(s)} \tau_{des(j,s),i}, \tag{3.1}$$

as the driving time between the destination of the closest phase-0 ambulance and node i, $i \in V \cup W$. The destination equals the current location of the ambulance if the ambulance is not on the road, i.e., in this case, $loc(j,s) = des(j,s)$. The reason that we use the destination instead of the current location for dispatchable ambulances, is twofold. First, 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. Second, 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 already an ambulance is moving towards that area.

As explained in Section 3.2.1, phase-4 ambulances can respond to incoming incidents if their service has lasted for already at least Δ seconds. Similarly to $t_i^0(s)$, we define $t_i^4(s)$ to be the expected time until the closest phase-4 ambulance is able to be present at node i:

$$t_i^4(s) = \min_{j \in \mathcal{A}^4(s)} \left\{ \left[\Delta - t^{elapsed}(j,s)\right]^+ + \tau_{loc(j,s),i} \right\},$$

where $t^{elapsed}(j,s)$ denotes the transfer time at the hospital already elapsed of ambulance j in state s and $[\cdot]^+$ denotes the positive part. 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:

$$U(s) := \sum_{i=1}^{|V|} d_i \Phi(\min\{t_i^0(s), t_i^4(s)\}),$$

where Φ is the penalty function of interest, c.f., Section 2.2.4. To illustrate the computation of the unpreparedness, consider the system in Figure 3.2a. Assume each node has the same demand probability: $d_i = \frac{1}{5}, i = 1, \ldots, 5$. Moreover, suppose we use the penalty function corresponding to the minimization of the average response time: $\Phi(t) = t, t \geq 0$. 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 $t_i^4(s) = 0, i = 1, \ldots, 5$. We compute $t_1^0(s) = t_2^0(s) = 0$, since ambulances are present at nodes 1 and 2. Moreover, $t_3^0(s) = 0$ 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, $t_4^0(s) = 1073$, and $t_5^0(s) = 1323$. 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 phases 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 Δ seconds, we assume that Δ 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 ambulances 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.

3.3.2 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 when m_0 is carried out by $U(s_0)$. Subsequently, we evaluate ambulance motions. For all motions, we consider the state of the system as if the motion was carried out instantly and all driving phase-0 ambulances would be at their destinations. For all these states, we compute the unpreparedness, using Equation (3.1), to obtain a classification of the ambulance motions. The best motion is the ambulance motion that minimizes the unpreparedness.

For decision moments of the second type, we do something similar. However, the set of possible motions is usually smaller due to the fact that 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.

Note that the number of possible motions is $\mathcal{O}(n|W|)$. For decision moments of the second type, the number of ambulance motions is even $\mathcal{O}(|W|)$, 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 $\mathcal{O}(n|V|)$ time, since for $|V|$ demand points we have to determine which of the n ambulances is the closest phase-0 and phase-4 ambulance. Therefore, the total complexity of the algorithm is $\mathcal{O}(n^2|V||W|)$, which is polynomial in the number of demand points, fleet size and number of base locations.

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

3.3.3 From Motions to Chain Relocations

Following the computation of the ambulance motion, the dispatcher has to make a decision concerning the exact execution of this motion, i.e., whether and how to set up a chain relocation. 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; see also Section 2.2.2). The LBAP can be solved to optimality in polynomial time, for instance by methods presented by Burkhard et al. (2009). In our setting, solving the LBAP 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.

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 benefits of too many ambulances moving in a chain relocation may be small. Therefore, in Section 3.5, we 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 a chain relocation.

3.3.4 Constraints on Decisions

To get a feeling about the necessity of the best motion, denoted by m_*, we compare it to the static motion m_0 defined above. To be more specific, we compute

$$q = \frac{U(s_0) - U(s_*)}{U(s_0)},$$

where $U(s_0)$ and $U(s_*)$ denote the unpreparedness of the state of the system when, respectively, the static and best motion are carried out instantly. We define Q to be the *motion threshold*: the dispatcher may carry out the best motion only if $q > Q$. Note that $0 \leq q \leq 1$, since $U(s_*) \leq U(s_0)$. 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 3.2.3: the way in which an ambulance motion is carried out, i.e., the number of ambulances used to perform an ambulance motion. We restrict the dispatcher to relocate no more than M phase-0 ambulances in a motion. The abovementioned M is a hard constraint that holds for both types of decision moments and $1 \leq M \leq n$. Recall 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.

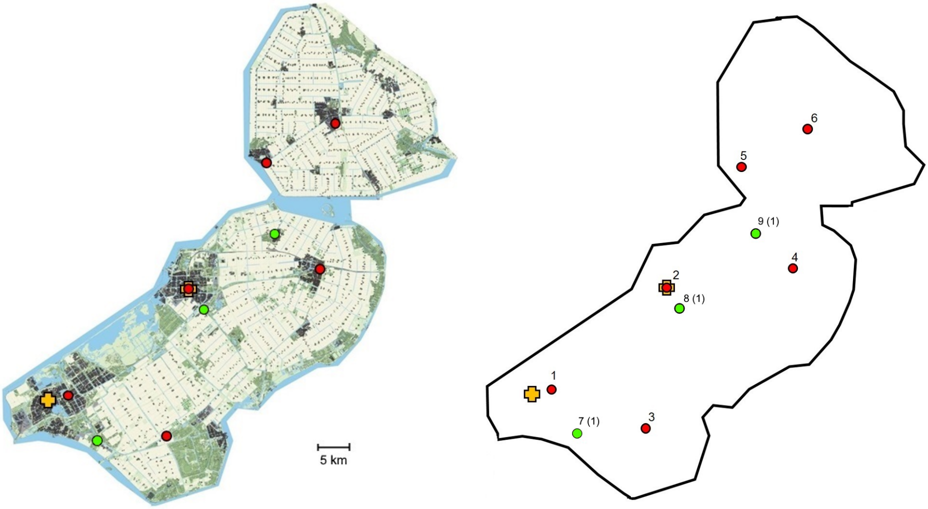
In short, the restrictions are given by (Q, M). Section 3.5 is concerned with results on the performance of the system and the number of relocations as a function of Q and M.

3.4 Experimental Setup

In this section, we describe the experimental setup for the numerical case study in this chapter and the following ones. We base our computations on two different EMS regions in the Netherlands: the EMS regions of Flevoland and Amsterdam. These regions are opposites of each other in terms of size and population. Flevoland is a large yet sparsely populated region, according to Dutch standards. On the other hand, Amsterdam is small but urban. Next, we will describe the regions in more detail. We refer to Figures 3.3 and 3.4 for a geographical representation of Flevoland and Amsterdam, respectively.

3.4.1 Flevoland

Flevoland covers approximately 1,400 km^2 and is home to nearly 400,000 people. Being raised from the sea in the 20th century, it is a very young region. With 285 inhabitants per squared kilometer, this region is quite rural for Dutch standards. Almost half of the total population of Flevoland lives in the city indicated with a ‘1’ in Figure 3.3b. The remaining population is mainly concentrated in one of the five other towns, although a couple of small villages exist as well, especially in the north-east. An ambulance base station, indicated by a dot in Figure 3.3, is located in or near each of the six major towns. These base stations are marked by a red dot in Figure 3.3. There are three additional waiting sites located at strategic places in the region, indicated by the green dots. The number between brackets



(A) (B)

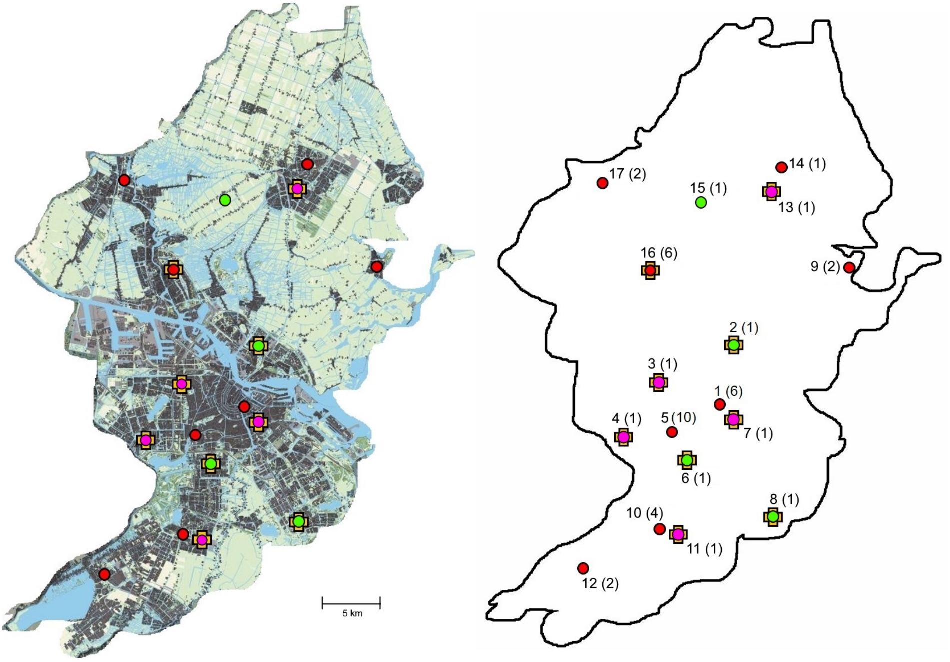
Figure 3.3: EMS region of Flevoland.

shows the ambulance capacity of the waiting site. The crosses in Figure 3.3 mark the two hospitals in Flevoland. We aggregate the region into 93 demand nodes, based on 4-digit postal codes. Actually, Flevoland is divided in 91 postal codes, but we cut two large areas in half. Note that the postal code corresponding to the dot indicated by a '2' contains both a waiting site and a hospital. The ambulance service provider of Flevoland is *GGD Flevoland*.

3.4.2 Amsterdam

The EMS region of Amsterdam and the surrounding areas is an amalgamation of two former EMS regions: the semi-rural Zaanstreek-Waterland (North) and the urban Amsterdam-Amstelland (South). The region is displayed in Figure 3.4. This region covers approximately 630 km^2 and is home to 1.2 million inhabitants, of which 68% live in Amsterdam itself. With 1,905 inhabitants per squared kilometer, it is very densely populated compared to Flevoland.

Ambulance waiting sites and hospital are present at the dots and crosses in Figure 3.4, respectively. The base stations and additional waiting sites are highlighted by red and green dots, respectively. The pink dots mark added, unofficial, waiting sites at hospitals. Many waiting sites coincide with one of the eight hospitals, marked by the crosses. The numbers in brackets denote the actual waiting site capacities. We aggregate the region into 162 demand points based on 4-digit postal codes. *Ambulance Amsterdam* runs the EMS operations in this region.



(A)

(B)

Figure 3.4: EMS region of Amsterdam.

3.4.3 Travel Times

Having access to accurate ambulance travel times is essential for testing the penalty heuristic. To that end, the RIVM[1] supplied us with an ambulance travel time table. They estimated deterministic travel times τ_{ij} for each 4-digit postal code-pair (i, j), from centroid to centroid. This was done as follows. 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 travel time for each pair of postal codes. We refer to Kommer and Zwakhals (2008) for a more detailed description of the computation of these emergency travel times.

However, estimating the actual location of an ambulance moving between postal code v_1 and v_m in a simulation of the relocation policy is difficult as the travel time table induces a *complete* graph $G = (V \cup W, A, \tau)$. To be able to keep track of the actual location of moving ambulances, we need the *route* between each pair of postal codes. We address this issue as follows: we construct a modified graph $\tilde{G} = (V \cup W, \tilde{A}, \tau)$ on the same vertex set as G, but with a different arc set. This subgraph is a node-incidence graph in which nodes are only connected by an arc if the corresponding postal codes are adjacent. Hence, $\tilde{A} \subseteq A$. Figure 2.4c in

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

the previous chapter shows an example of such a node-incidence graph, although the graph in this figure is equidistant and contains fewer nodes.

For an ambulance traveling between nodes v_1 and v_m, both the route and the travel time (induced by $\tilde{G}$) can be computed by a shortest-path algorithm. However, the computed shortest path length usually exceeds the estimated travel time τ_{v_1,v_m} as a consequence of the triangle inequality. Formally, let the shortest route between nodes v_1 and v_m be given by the sequence $v_1, v_2, \dots, v_m$. Moreover, we define $t_k(v_1, v_m)$ as the enter time of node k on the path $v_1, \dots, v_m$. That is, $t_k(v_1, v_m) = \sum_{l=1}^{k} \tau_{v_l,v_{l+1}}$. Since $t_m(v_1, v_m) \geq \tau_{v_1,v_m}$, we rescale the enter times $t_k(v_1, v_m)$, $1 \leq k \leq m$, to obtain modified enter times $\tilde{t}_k(v_1, v_m)$, as follows:

$$\tilde{t}_k(v_1, v_m) = \frac{\tau_{v_1,v_m} t_k(v_1, v_m)}{t_m(v_1, v_m)}.$$

Note that $\tilde{t}_m(v_1, v_m) = \tau_{v_1,v_m}$ and these modified enter times are dependent on the start and end node. We store the shortest path between any pair of nodes and the corresponding modified enter times $\tilde{t}$ in the memory in advance of the simulation. The actual location of a traveling ambulance on the route can then easily be estimated by considering the elapsed travel time and comparing it to the rescaled enter times: we assume that the ambulance is at the node for which the absolute difference between modified enter time and elapsed travel time is smallest.

3.5 Numerical Results

The purpose of this section is to show computational results on both the performance and the number of relocations using the penalty heuristic under different (Q, M)-regimes. We mainly focus on the EMS region of Flevoland, while we also provide a short numerical study on the Amsterdam region. For both regions, we only included the actual base stations, marked by the red dots in Figures 3.3 and 3.4, in our computations. The additional waiting sites are neglected. We generate computational results by a discrete-event simulation using historical data of the period January 2008 – September 2012, which was provided by GGD Flevoland. We have access to the following information of incidents: time and place (based on 4-digit postal code level) of occurrence, the on-scene time of the ambulance, whether the patient needed transportation to a hospital and the hospital transfer time.

In the simulations, 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 data set, 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). Moreover, 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. In addition, 71% 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, and 75% needs transportation.

We also use the supplied historical data for the computation of the demand probabilities d_i, where $i \in V$, by dividing the number of requests at node i by the total number of requests, for day and night separately. No randomness is involved in the simulation, since we use the actual historical data (trace-driven). The simulation evolves according to the system dynamics described in Section 3.2.1. If no ambulances are available at the time of the occurence of an incident, this request is placed in a first-come first-served queue. As soon as an ambulance becomes available, it will immediately respond to the first request in the queue.

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 3.3, 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%. 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%. 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.

As objective in the penalty heuristic, 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 chute time is 3 minutes, which induces 12 minutes (720 seconds) as maximum allowed travel time. The penalty function we use is

$$\Phi(t) = \begin{cases} \frac{1}{\beta(1+e^{-\alpha(t-T)})} & 0 \le t \le T, \\ \frac{\beta-1}{\beta} + \frac{1}{\beta(1+e^{-\alpha(t-T)})} & t > T, \end{cases} \tag{3.2}$$

and is displayed in Figure 3.5 for $\alpha = 0.008$, $\beta = 5$, and $T = 720$. We composed this function in consultation with practitioners from several ambulance service providers. Note that the focus in this penalty function is on minimizing the number of late arrivals rather than on minimizing the average response time. 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 dispatchable. Therefore, we set $\Delta = 600$.

3.5.1 Critical Night Situation

In the critical night situation, we assume there are $n = 4$ ambulances. In Figure 3.6a, we display the penalty per night as a function of the motion threshold

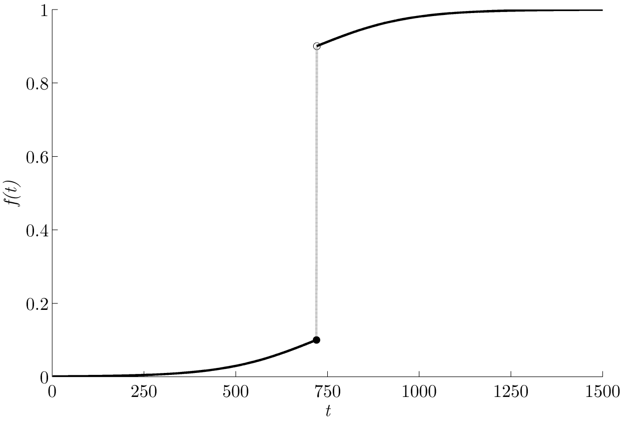
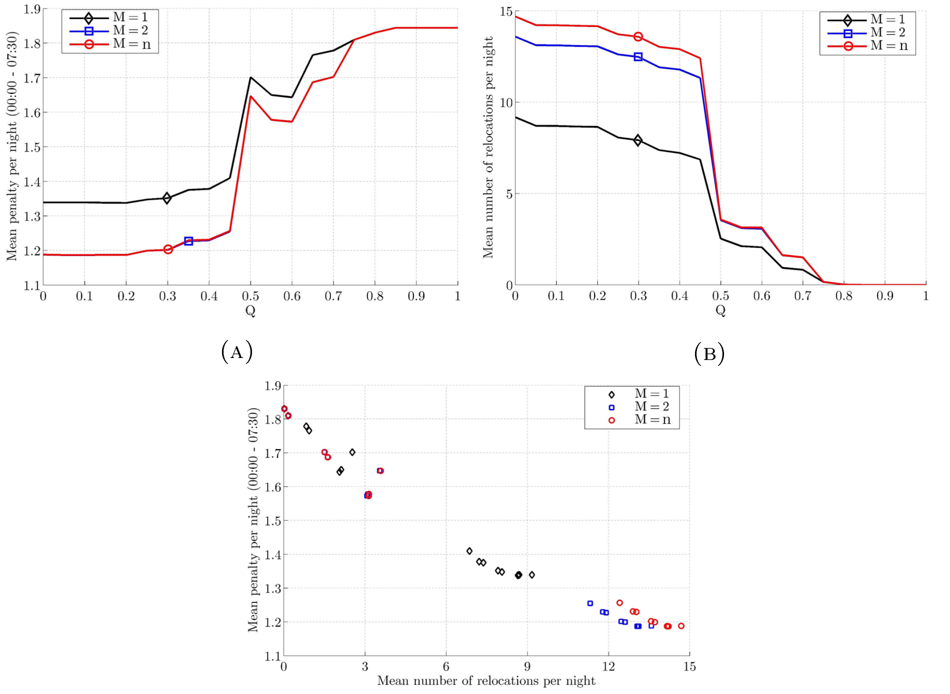


Figure 3.5: Penalty function of Equation (3.2).

Q, for $M = 1, 2, n$. However, since our system only contains four ambulances, the graphs for $M = 1$ and $M = n$ hardly differ. Note that the largest gap between $M = 1$ and $M = n$ is at $Q = 0$, i.e., if the dispatcher is always allowed to perform the motion. This gap is approximately 11.3%, as observed in Table 3.2. Thus, there is a beneficial effect on the 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 a function of Q for $M = 1, 2, n$ is displayed in Figure 3.6b. We observe approximately six additional ambulance relocations per night.

Column III of Table 3.2 shows a beneficial effect on the performance if one compares the $Q = 1$ and $Q = 0$ regime, i.e., if we always carry out the best motion instead of the static one. Furthermore, it is worth noting that although the graphs for $M = 2$ and $M = n$ coincide in Figure 3.6a, 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.

The increase just before $Q = 0.5$ in Figure 3.6a and the corresponding steep decrease in Figure 3.6b is explained through geographical reasons. Remember that there are two hospitals in the two largest cities (see Figure 3.3). These two cities together are inhabited by 68% 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 base station 6, 16% 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 city 6, is too small and thus the static motion is always performed. Since the majority (71%) of the ambulances finishes the treatment of a patient at a hospital, it hardly occurs that an ambulance becomes dispatchable again at one of the four other towns.



(A)

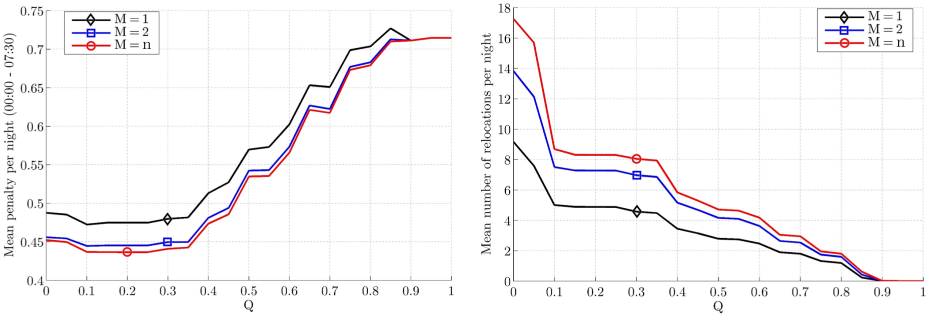
(B)

(C)

Figure 3.6: The mean penalty (Figure 3.6a) and number of relocations per night (Figure 3.6b) as a function of the motion-threshold Q for the critical night situation with $n = 4$. The width of the 95%-confidence intervals is approximately 0.16 for Figure 3.6a and 0.4 for Figure 3.6b. Figure 3.6c displays the relation between penalty and number of relocations per night.

	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%
$M = n$	11.3%	37.5%	35.6%	8.0%	41.2%	38.9%

Table 3.2: 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 Figures 3.6a and 3.7a is attained. Column III represents the gain in performance for Q_{min} with respect to $Q = 1$. Note that n equals 4 and 7 for the critical and realistic situation, respectively.



(A)

(B)

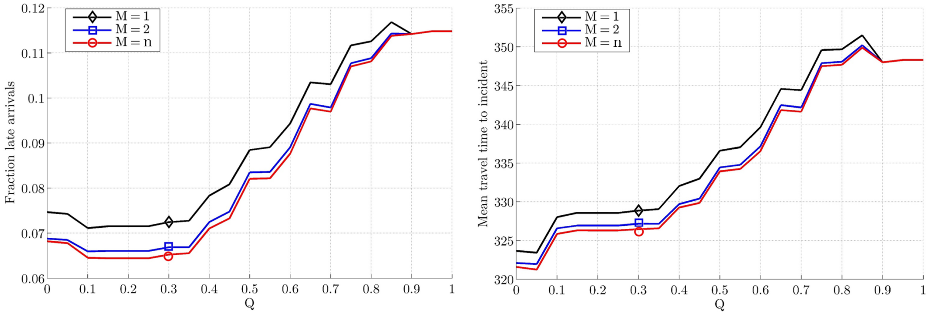
Figure 3.7: The mean penalty (Figure 3.7a) and number of relocations per night (Figure 3.7b) as a function of the motion-threshold Q for the realistic night situation with $n = 7$. The width of the 95%-confidence intervals is approximately 0.08 for Figure 3.7a and 0.4 for Figure 3.7b.

Therefore, an ambulance that just finished service of a patient in city 1 or 2 is seldom relocated to city 6, because it is forced to carry out the static motion. This results in a decrease in both performance and number of ambulance relocations.

The peak at $Q = 0.5$ in Figure 3.6a can be explained by a similar reasoning. For $Q > 0.5$, the static motion is performed if an incident occurs in city 1 and an ambulance is present in city 6. That is, no ambulance is redeployed from city 6 to city 1. However, at $Q = 0.5$, the best motion is performed in this situation, in which city 6 is the origin and city 1 the destination. The time to perform this motion for a single ambulance is approximately 32 minutes, while it takes at least 20 minutes when multiple ambulances participate in the motion. Since many ambulances finish their service in city 1, it often occurs that an ambulance finishes service before a relocated ambulance arrives there. For city 1, this is beneficial and a better performance is achieved there. However, these benefits not outweigh the performance loss in city 6. 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 reasons described above. In short, this illustration shows that performing the best motion does not always result in a better performance compared to the static motion.

3.5.2 Realistic Night Situation

In the realistic night situation, seven ambulances are on duty. The graphs for the mean penalty and the number of relocations as a function of Q are displayed in Figure 3.7. In Figure 3.7a, 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 $M = n$, unlike in the critical night situation. At $Q = 0.2$, this gap is approximately 6.2% and 8.0% for $M = 2$ and $M = n$ with respect to $M = 1$, as column I of Table 3.2 shows. Thus, by



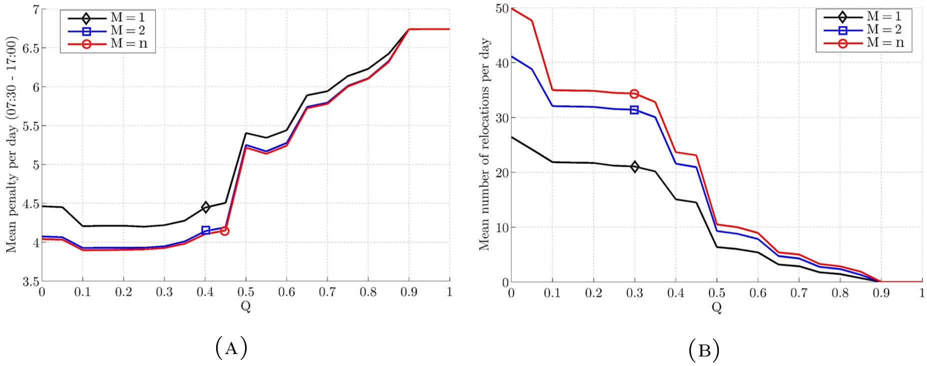
(A)

(B)

Figure 3.8: The fraction of incidents for which the maximum allowed travel time of 720 seconds is exceeded, and the mean response time as a function of the motion-threshold Q for the realistic night situation with $n = 7$. The width of the 95%-confidence intervals is approximately 0.08 for Figure 3.8a and 4 for Figure 3.8b.

allowing the dispatcher to use more than two ambulances in performing a motion, the performance improves. However, this improvement is small compared to the beneficial effect if one allows two ambulances to participate in the motion instead of one.

If we compare $Q = 0$ and $Q = 0.1$, we observe a tremendous decrease in number of relocations in Figure 3.7, and the penalty decreases as well, albeit to a lesser extent. This behaviour is explained by the choice of the penalty function. Results for the two objective functions compromised in the penalty function separately are displayed in Figure 3.8. Intuitively, one would expect that the criteria of minimization the fraction of late arrivals and minimization of the average response time are not 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: city 5 can be reached within 720 seconds from city 6 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 city 5. For $Q = 0$, the best motion is always performed, which sends an ambulance to city 5. 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 city 5, 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 Figure 3.8, especially for $M = 1$. It takes much time to perform the best motion, as an ambulance has to move from city 5 to city 1. If we compare the critical and realistic situation in Table 3.2, 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.



(a)

(b)

Figure 3.9: The mean penalty (Figure 3.9a) and number of relocations per day (Figure 3.9b) as a function of the motion-threshold Q for the critical day situation with $n = 6$. The width of the 95%-confidence intervals is approximately 0.4 for Figure 3.9a and 0.6 for Figure 3.9b.

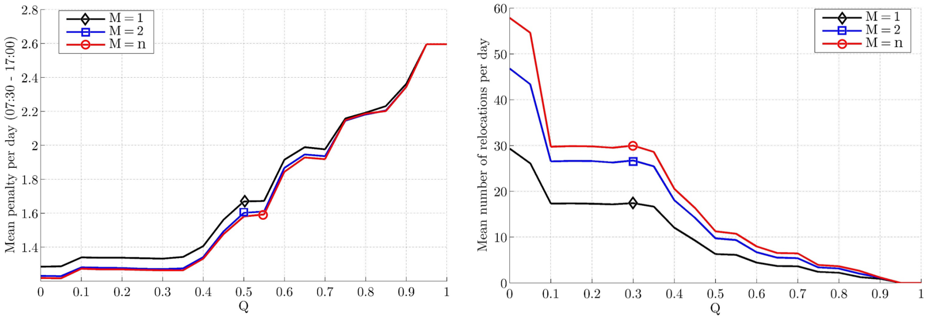
	Critical day situation			Realistic day situation		
	I	II	III	I	II	III
$M = 1$	-	-	37.5%	-	-	50.5%
$M = 2$	6.7%	32.0%	41.7%	4.2%	37.3%	52.6%
$M = n$	7.3%	37.7%	42.1%	5.2%	49.3%	53.1%

Table 3.3: 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 graphs in Figures 3.9a and 3.10a is attained. Column III represents the gain in performance for Q_{min} with respect to $Q = 1$. Note that n equals 6 and 12 for the critical and realistic situation, respectively.

3.5.3 Critical Day Situation

During daytime, the maximum number of ambulances needed to ensure that we are always in the critical situation is six. Compared to the night situation, slightly more patients need transportation to a hospital: this percentage is 75%. Results for this situation are displayed in Figure 3.9 and Table 3.3.

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 reasons as in the realistic night situation. We conclude from Table 3.3 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.



(A)

(B)

FIGURE 3.10: The mean penalty (Figure 3.10a) and number of relocations per day (Figure 3.10b) as a function of the motion-threshold Q for the realistic day situation with $n = 12$. The width of the 95%-confidence intervals is approximately 0.14 for Figure 3.10a and 0.8 for Figure 3.10b.

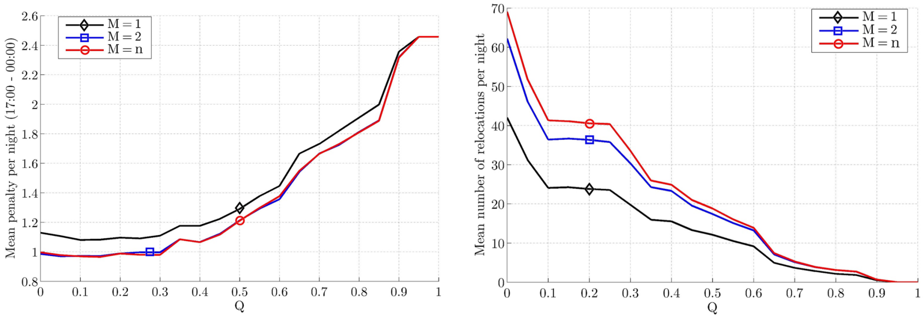
3.5.4 Realistic Day Situation

In the realistic day situation, twelve ambulances are present in the system. Results for this case are listed in Table 3.3 and Figure 3.10. There are some differences compared to the situations before. For instance, there is now an increase in penalty between $Q = 0$ and $Q = 0.1$, as observed in Figure 3.10a. 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 city 5. 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.

3.5.5 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. Historical data of the year 2011, provided by Ambulance Amsterdam, serves as the basis for our computations, and we again distinguish a day and a night situation. The total number of incidents in 2011 was 12,362 and 38,784 during night (00:00 – 07:30) and day (07:30 – 17:00), respectively. This results in 34 and 106 incidents on average per night and day. We again use the penalty function of Equation (3.2) with $\alpha = 0.008$, $\beta = 5$, and $T = 720$ and we retain the parameters corresponding to the response time threshold and the dispatch and chute time as in the Flevoland case.

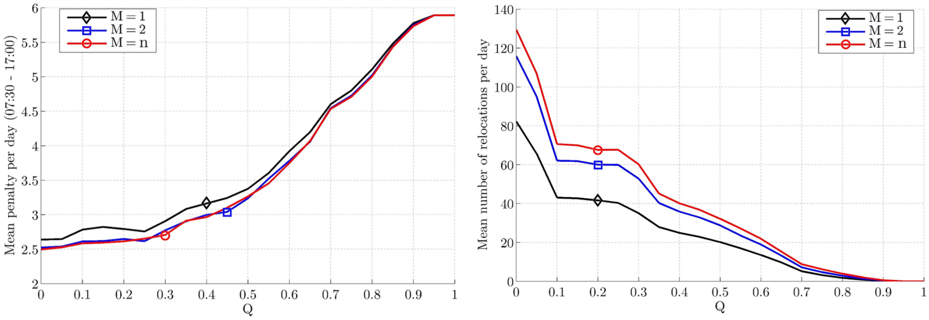
We consider both the realistic night and day situation with 15 and 24 ambulances, respectively. Results are displayed in Figures 3.11 and 3.12, and Table 3.4. Since Amsterdam is a smaller region than Flevoland and there are more base locations in Amsterdam, the driving times between base locations are smaller. Moreover, a lot more incidents occur in Amsterdam. However, these differences



(A)

(B)

Figure 3.11: The mean penalty (Figure 3.11a) and number of relocations per night (Figure 3.11b) as a function of the motion-threshold Q for the realistic night situation with $n = 15$. The width of the 95%-confidence intervals is approximately 0.24 for Figure 3.11a and 3 for Figure 3.11b.



(A)

(B)

Figure 3.12: The mean penalty (Figure 3.12a) and number of relocations per night (Figure 3.12b) as a function of the motion-threshold Q for the realistic day situation with $n = 24$. The width of the 95%-confidence intervals is approximately 0.3 for Figure 3.12a and 4.6 for Figure 3.12b.

	Critical day 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%
$M = n$	11.5%	71.6%	60.7%	5.7%	57.4%	57.6%

Table 3.4: 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 Figures 3.11a and 3.12a is attained. Column III represents the gain in performance for Q_{min} with respect to $Q = 1$. Note that n equals 15 and 24 for the critical and realistic situation, respectively.

are not reflected in the results: many of the phenomena observed in the results for Flevoland carry over to Amsterdam. We highlight one difference: in the Flevoland cases, $M = 2$ results in a higher penalty than $M = n$ in general. However, for Amsterdam, these two graphs are intertwined, as can be observed in Figures 3.11 and 3.12. 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 the shorter travel times between base stations, 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.

3.6 Concluding Remarks

In this chapter, we analyzed the effect of ambulance relocations on the performance of the ambulance service provider. For that purpose, 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 the penalty heuristic for computing ambulance motions and relocations. 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 (Q), and for the second we imposed a bound (M) on the number of ambulance relocations within a motion. We used historical data of two regions in the Netherlands to simulate the EMS system in which the penalty heuristic policy is carried out, under different (Q, M) regimes. Moreover, we showed results for one particular penalty function suggested by an ambulance service provider. 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 behaviour 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. This phenomenon could be caused by 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 into the relationship between performance and number of relocations.

Together with the DMEXCLP by Jagtenberg et al. (2015), the penalty heuristic developed in this chapter was adjusted for operational use by GGD Flevoland, the ambulance provider of the EMS region of Flevoland. It was implemented in a software tool for real-time decision support. The policy induced by the penalty heuristic was used in an actual dispatch center for a period of six weeks in the second half of the year 2015. To the best of our knowledge, this constituted the first non-commercial implementation of real-time ambulance repositioning policies in practice.Van Buuren et al. (2016) evaluate this pilot and provide statistics for the efficiency improvements. Moreover, they discuss the experiences of ambulance dispatchers and management.

4

Practical Insights into the Implementation of a Relocation Policy

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

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

4.1 Introduction

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

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

The inclusion of busy ambulances in the state description of the system. We investigate whether ambulance repositioning methods can benefit from taking into account vehicles that are currently dropping of a patient at a hospital. It is clear that these vehicles will become idle in the near future, but it is not trivial how one should model this, nor is it evident that this will have a positive effect on the performance. We show that taking ambulances at hospitals into account has hardly any effect on the response times, but it does slightly diminish relocation times, and thereby workload, for the crew.

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

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

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

Note that decision makers in practice may come to different conclusions based on the characteristics of their EMS region. For example, the size of the demand, as well as how it is spatially distributed, distances, and overall workload have a great effect on the dynamics in the EMS system. These characteristics may affect the performance of a relocation policy, and a policy that performs well in one region, does not necessarily give the same result elsewhere. Since we aim to construct a robust algorithm with respect to region characteristics, we include case studies for two different types of regions: the rural region of Flevoland, and the urban region of Amsterdam, both in the Netherlands. All our results are obtained from trace-driven simulations. While our primary focus is on minimizing the fraction of late arrivals, other metrics, such as crew related performance indicators, are also reported.

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

4.2 Model

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

As in the previous chapters, we consider a single type of ambulance and a single type of demand priority, inducing a single threshold or target, denoted by T, for the response time. We model the region as a doubly weighted complete directed graph $G = (V \cup W, A, (\tau^{(1)}, \tau^{(2)}))$, in which V, W and A are as defined in Section 3.2.4. Two different travel times are associated to each arc: $\tau_{ij}^{(1)}$ denotes the expected travel time between nodes i and j when driving with optical and sound signals turned on, typically used while responding to an emergency or the transportation of a patient to a hospital. If the ambulance is not performing patient-related duties, such as the return to a waiting site, then the optical and sound signals are not turned on. This yields a longer travel time, denoted by $\tau_{ij}^{(2)}$. Obviously, it holds that $\tau_{ij}^{(2)} \geq \tau_{ij}^{(1)}$. For an overview of notation we refer to Tables 3.1 and 4.1, the latter summarizing the newly introduced notation in this chapter.

4.3 Algorithms and Features

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

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

TABLE 4.1: Notation

in Chapter 3. Both methods have in common that it is only allowed to relocate vehicles to existing waiting sites. Such a relocation decision may only be taken at discrete *decision moments* in time, which we will define later. The decision is then computed by brute force in real time. Moreover, both methods incorporate the location of idle ambulances in the same way: for a traveling idle ambulance they pretend that it is already at its destination instead of at its current location. This choice has two advantages. First of all, for a real-life system it is typically easier to keep track of destinations since they change less often than current locations. Second, there is a methodological advantage: for a moving ambulance, its current location is only relevant for a very short time, while our relocation decision should be beneficial to the system for a longer time. In Section 4.3.3 we describe the incorporation of several aspects considered by Van Barneveld et al. (2016a) into the DMEXCLP method and into the simulation used for obtaining results.

4.3.1 Outline of DMEXCLP

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

MEXCLP defines the coverage of a region in terms of a so-called *busy fraction* p. This busy fraction is predetermined, and assumed to be the same for all vehicles. It can be estimated by dividing the expected load of the system by the total number of available ambulances. Furthermore, ambulances are assumed to operate independently. Consider a demand point $i \in V$ that is within the time threshold T of k ambulances. We can straightforwardly determine this number k using the expected travel times $\tau_{ij}^{(1)}$, $i, j \in V$. The probability that at least one of these k ambulances is available at any point in time, is then given by $1 - p^k$. If we let d_i be the demand at node i, the expected covered demand of this vertex is $E_k = d_i(1 - p^k)$. The MEXCLP positions the ambulances in such a way that the total maximum expected covered demand, summed over all demand points, is obtained.

DMEXCLP, or dynamic MEXCLP, reuses this definition of probabilistic coverage, but computes it for relocation purposes each time when an ambulance becomes available. At such a decision moment, the current state of the system is observed. DMEXCLP disregards all information about ambulances that are busy, and focuses purely on the set of idle vehicles. As mentioned, we only consider the destination of idle ambulances. If an ambulance is standing at a waiting site, we define its destination to be its current location. Information regarding the destination of each ambulance is captured by variables n_j: the number of idle ambulances that have waiting site j as destination, $j \in W$. In addition, DMEXCLP requires information on $(d_i)_{i \in V}$ and $(\tau_{ji}^{(1)})_{j \in W, i \in V}$.

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

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

where

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

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

4.3.2 Comparison to Penalty Heuristic

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

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

Number of decision moments: As we have seen, Jagtenberg et al. (2015) propose a relocation only when an ambulance becomes available. This choice has to do with the fact that DMEXCLP was originally designed for busy regions, in which vehicles often become idle. Although the authors state that the method can be easily adjusted for usage at other types of decision moments, it is not clear which ambulance should be relocated. In addition, Van Barneveld et al. (2016a) allow a relocation to be executed immediately after the dispatch of an ambulance to an incident.

Busy ambulances: As mentioned in Section 4.3.1, busy ambulances do not contribute to the coverage in the work by Jagtenberg et al. (2015). In contrast, Van Barneveld et al. (2016a) consider an ambulance as dispatchable if its transfer time at a hospital exceeds a predefined standard Δ. That is, after some time, the transfer may be interrupted if necessary. This influences the coverage of the region, as now a busy ambulance covers the direct neighborhood of the hospital.

Chain relocations: Jagtenberg et al. (2015) do not consider chain relocations, in contrast to Van Barneveld et al. (2016a). To attain the desired ambulance configuration in less time, the, otherwise possibly long, trip may be split into two or more trips, in which multiple ambulances are involved. Note that this extension does not influence the calculation of which waiting site should receive one additional vehicle: it can be regarded as a second step, executed after the computation of the new ambulance configuration.

Objective: The focus is on minimization of late arrivals solely in Jagtenberg et al. (2015): one incurs a penalty of 1 each time the response time to an incident exceeds T. In contrast, this objective can be generalized by the definition of a *penalty function*. This is a non-negative non-decreasing function on $\mathbb{R}_{\geq 0}$ relating a certain penalty to each possible response time. Note that the objective of DMEXCLP can be easily modeled by the penalty function $\Phi(t) = \mathbb{1}_{\{t>T\}}$. However, Van Barneveld et al. (2016a) question the dichotomous nature of this objective, as medical outcomes are completely ignored. Instead, they use a different penalty function, in which the primary goal is to maximize coverage as before, but there is more distinction between different response times. This function is given by Equation (3.2) and displayed in Figure 3.5 for $\alpha = 0.008$, $\beta = 5$, and $T = 720$, in the previous chapter.

We conclude that in one way DMEXCLP is richer than the penalty heuristic, as the multiple and non-integer MEXCLP coverage is a generalization of the penalty heuristic's single coverage. On the other points, the assumptions made by Jagtenberg et al. (2015) are generalized by Van Barneveld et al. (2016a). In the next section, we explain how we modify the original DMEXCLP method by incorporating a number of features related to the five aspects described above.

4.3.3 Modification of DMEXCLP

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

Decision moments

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

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

in which w_1 and w_2 denote the old origin and new destination of the vehicle to relocate, and $k(i, w, n_1, \ldots, n_{|W|})$ is as defined in Equation (4.2). In Equation (4.3) each possible waiting site pair with at least one ambulance at the origin is evaluated. Since the number of waiting sites is typically small, the maximization in Equation (4.3) can be computed by brute force.

Busy ambulances

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

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

denote the expected remaining transfer time of ambulance a if its transfer already lasted for $\tau(a)$ units of time. Moreover, let $h(a)\in V$ denote the demand zone in which the hospital where ambulance a is busy, is located. Let $\mathcal{A}$ be the set of ambulances currently dropping off a patient at a hospital. We adjust Equation (4.2) as follows:

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

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

Chain relocations

As stated before, the use of chain relocations is not a modification of the DMEXCLP method, but the calculation of this chain is a subsequent step: the expression of Equation (4.1) is not modified. As mentioned in the previous chapter, the linear bottleneck assignment problem is considered for this computation. In Chapter 3, we concluded that the benefit to the patient-based performance of a chain relocation consisting of more than two links is very small. We observed a large performance gain, however, if chains consisting of exactly two links are used, instead of a regime in which no chain relocations are allowed. The crew-based performance decreases if chains consist of more than two links, as a consequence of an inflation in number of relocations. As the regions considered in the numerical study of this paper are the same as in Chapter 3, we follow this conclusion and restrict that at most two ambulances may take part in a chain relocation.

Relocation time bounds

At a decision moment, the DMEXCLP method searches for the waiting site for which the expected coverage is maximized, without taking into account the current location of the ambulance. However, from both patient and crew perspective, it might be beneficial to steer the system towards a good, but not necessarily the best, configuration that can be attained quickly. After all, driving to a waiting site, although best classified by DMEXCLP, may take long. To study the behaviour of the performance if the focus is on good local configurations, we impose an upper bound B on the relocation time of an ambulance. That is, we do not allow the relocation of an ambulance to a waiting site for which the driving time between its current location and destination exceeds B time units. Let c be the current location of the ambulance under consideration. Then, we modify Equation (4.1) as follows:

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

That is, we evaluate only the waiting sites that can be reached within B time units from the current location of the ambulance in the maximization. In Section 4.4.6 we analyze the behaviour of the system on both patient and crew-based performance for different values of B.

Performance criteria

The incorporation of a different performance criterion, such as the one considered in Equation (3.2) and Figure 3.5, requires more effort than the previous features: one can no longer simply count the number of ambulances within range of demand node i. After all, each idle ambulance contributes to the coverage of i, no matter how far away. Due to the notion of probabilistic coverage, this contribution levels off the farther away an ambulance: with probability $1-p$ the closest one to i is available and responds to an incident occuring there, inducing a penalty of $\Phi(\tau^{(1)}_{ji})$ if the closest ambulance to i is located at waiting site j. With probability $(1-p)p$ the second closest responds, generating $\Phi(\tau^{(1)}_{j'i})$ penalty if this ambulance is at j', and so on.

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

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

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

4.4 Numerical Results

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

4.4.1 Experimental Setup

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

Historical data on emergency requests in the year 2011 was used for our analyses. We built two traces based on this data and simulate them in a discrete-event simulation. The trace is constructed as follows. We consider all emergency requests occuring between 7 AM and 6 PM, generally the busiest time of the day. In the trace, we include the following incident related information:

- Time of occurence, i.e., the time of the emergency call;
- Location of occurence (4-digit postal code);
- Time spent on scene by the ambulance;
- Hospital transfer time.

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

7,632 resp. 41,996 incidents occur in Flevoland and Amsterdam, respectively. This yields an hourly arrival rate of 1.97 resp. 10.84 emergency requests. Moreover, around 87% resp. 73% of the patients needs transportation to a hospital. The average busy time of an ambulance is 0.74 resp. 0.73 hours, excluding relocation time after the transfer. To ensure an out-of-sample validation, we estimate the demand probabilities per postal code based on the year 2010, and not 2011.

In our simulations, the closest idle ambulance always responds to the incident. If no ambulance is available, the call enters a queue. Once an ambulance becomes available from service again, it is immediately dispatched to the longest waiting request. Moreover, if a patient needs transportation to a hospital, the closest hospital is selected. In the simulation model, we use travel times estimated by the RIVM, which provided us tables containing travel times between each pair of postal codes in the regions of consideration. We refer to Section 3.4.3 and Kommer and Zwakhals (2008) for a more detailed description on the travel time model used for the estimation of these travel times. We interpret the travel times in these tables as the arc lengths $\tau^{(1)}$. The travel times $\tau^{(2)}$ are obtained by multiplying $\tau^{(1)}$ with a factor of $\frac{10}{9}$. Moreover, we use the framework described in Section 3.4.3 for the computation of the travel routes, in order to keep track of the actual location of a moving vehicle. We do not simulate a dispatch time or pre-trip delay.

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

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

2. Mean response time.

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

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

5. Total relocation time.

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

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

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

4.4.2 Original DMEXCLP method

In this section, we report results for both regions of interest, Flevoland and Amsterdam, of the original DMEXCLP method, as explained in Section 4.3.1. Moreover, we compare these results to the static policy according to the MEXCLP solution: each ambulance returns to its home base station when newly idle. Results are listed in Table 4.2.

A large performance improvement in terms of late arrivals can be observed in Table 4.2 for the Amsterdam region. This quantity decreases from on average 6.19% to 4.10%, a difference of 2.09 percentage point and a decrease of 33.76%, even outperforming the performance gain reported in the original article (Jagtenberg et al. (2015), for the region of Utrecht). However, the performance gain regarding this criterion is small for Flevoland: a difference of 0.11 percentage point, which is a decrease of only 2.1%. Moreover, the confidence bounds for this region overlap almost entirely. In addition, the gaps in mean single coverage and mean MEXCLP coverage between the static and DMEXCLP policy are much smaller for Flevoland. This was already foreseen by Jagtenberg et al. (2015), and a possible explanation for this phenomenon is given: the DMEXCLP method is designed for busy areas in particular. The hourly arrival rate of incidents in Flevoland is much smaller compared to the urban Amsterdam region. As a consequence, there are fewer relocation moments, inducing a smaller performance improvement. In the next subsection, we allow additional decision moments.

In contrast to Flevoland, the number of ambulance relocations in Amsterdam does not equal the number of incidents. This is explained by the fact that in Amsterdam sometimes the situation occurs that none of the ambulances is available for a reported incident. As soon as an ambulance finishes service of a patient, it is immediately dispatched to a waiting call. This is not recorded as a relocation and hence, the number of relocations does not necessarily equal the number of incidents. Based on Table 4.2 one can compute that the total number of incidents for which no ambulance was immediately available, equals 655 and 575 for the static and DMEXCLP policy, respectively.

Note that both the mean single and MEXCLP coverage performance indicators serve as an estimate of the number of calls for which the response time treshold is achieved. As observed in Table 4.2, the mean single coverage is an optimistic ap-

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

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

proximation of this quantity for both policies, as expected. After all, ambulance unavailability is not taken into account in the concept of single coverage. The relative gap between mean single coverage and percentage on time is smaller for Flevoland, compared to Amsterdam, for both policies. This is not very surprising, since in Flevoland the overlap in coverage of multiple ambulances is very small: the distances between the 6 large towns generally exceed the time threshold. Furthermore, the busy fraction in Flevoland is relatively low. Therefore, the error made when ignoring ambulance unavailability will also be small.

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

4.4.3 Decision Moments

As explained in Section 4.3.3, we allow the dispatcher to make an ambulance relocation decision if the number of available ambulances decreases, just after the dispatch. As a consequence, the number of opportunities to steer the EMS system is multiplied by two. Results are displayed in Table 4.3. In this table and the forthcoming ones, the default policy is the DMEXCLP policy explained in Section 4.3.1, without any additional features. This policy outperforms the static policy, commonly used as benchmark policy in ambulance literature, on the most important performance indicators, as Table 4.2 underlines.

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

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

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

Surprisingly, the results on mean response times do not concur with those on the late arrivals criterion: in Flevoland, a performance gain of only 1.64% is achieved. In contrast, the mean response time in Amsterdam decreases with 7.44%. A possible explanation for this behaviour is the following: since Flevoland is a rural region, an ambulance traveling between two waiting sites provides no or very little coverage. After all, few people live in the areas between the cities, c.f., Figure 3.3. In contrast, a large part of the Amsterdam region is urban, c.f., Figure 3.4. In an urban area, an ambulance performing a relocation drives through a densely populated area, being able to respond to an incoming call in that area quickly. As the number of ambulance relocations almost doubles for both regions, this effect will be largest in Amsterdam, resulting in a relative large decrease in mean response time.

In the crew-related performance indicators, we observe both an increase in number of relocations and average relocation time. As a consequence, the total relocation time is more than doubled. A trade-off between patient- and crew-based performance, which is the subject of Chapter 3, is clearly visible here as well. The question arises whether this large increase outweighs the gain in patient-based performance. It is up to the ambulance service provider to decide on this, but we suspect that the answer depends on the daily workload of the crew. As this is typically lower in rural regions, we expect those EMS providers to be more open to additional relocation moments.

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

4.4.4 Hospitals

In this section, we explore the differences in performance if ambulances transferring patients at hospitals are taken into account. We do this in two ways. First, we consider the data obtained via the ambulance service providers and fit a distribution on the busy times of an ambulance at a hospital. As mentioned in Section 4.3.3, we plug in the expected remaining service time in the formula, given the hospital time already elapsed. As an alternative approach, we simulate the system in which we have ‘perfect information’ regarding the hospital transfer time. We assume that we know this time when an ambulance arrives at the hospital, which results in a deterministic remaining service time. This approach clearly is a rather optimistic approach, and it can be interpreted as a bound on the knowledge that one can have on the remaining service time. However, this approach is more realistic than one might expect at first glance, as ambulance crews and dispatchers in the Netherlands are able to estimate the hospital transfer time rather accurately, as we have learned from discussions with dispatchers and management. In particular, hospitals in the Netherlands do not suffer from queues building up at an emergency department, in contrast to North America where the average transfer time can be very large and highly variable, c.f., Carter et al. (2015).

We estimate the service time at a hospital by a Weibull distribution, for both regions. In our experience, this distribution provides a rather accurate approximation. Moreover, a Weibull distribution for this quantity was also used in both Maxwell et al. (2010). The means of the fitted distributions are 966 seconds and 1,160 seconds for Flevoland and Amsterdam, respectively. The differences in mean are probably explained by the fact that the hospitals in Amsterdam are typically larger, and thus the ambulance personnel spends more time on the transport of the patient to the appropriate department within the hospital. Based on the Weibull distributions, we calculate the expected remaining transfer time for each possible value of service time already elapsed.

In Table 4.4, we list simulated results on the assumption of Weibull distributed transfer times and perfect information, and we compare those to the default policy explained above. We observe neither an increase nor a decrease in the patient-related performance indicators in the Weibull case. A small decrease in average relocation time can be observed, which has a small effect on the total relocation time as well. Based on these observations, one might conclude that the inclusion of ambulances busy at a hospital in the algorithm in the way described in Section 4.3.3

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

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

does not significantly influence the performance.

Alternatively, the Weibull distribution used for the estimation of the transfer time may perhaps be a poor approximation. To test whether this indeed is the case, we simulate the system in which we have perfect information about the transfer time to exclude this source of randomness. However, we do not observe an improvement in the patient-related performance indicators. Based on these results, we claim that taking into account ambulances busy at a hospital in the way we did (as explained in Section 4.3.3), has no effect on the patient-related performance, regardless the distribution used.

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

4.4.5 Chain Relocations

In Chaper 3, it is stated that it is beneficial to use chain relocations: the break-up of a certain long lasting relocation into multiple short relocations by different ambulances. Moreover, their computational results, based on the same regions considered in this paper, show substantial benefit when using two links instead of one, but using more than two links appears to be redundant. We simulate the system according to this regime: a relocation is decomposed into a chain relocation of length two if this reduces the time until the new configuration is attained. Results are displayed in Table 4.5.

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

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

Although the time until the desired configuration is attained is decreased, we do not observe a gain on the patient-related performance criteria. Instead, even a slight deterioration can be seen in Table 4.5. This contradicts the findings of Van Barneveld et al. (2016a). This is probably due to the fact that they allow extra decision moments, as considered in Sections 4.3.3 and 4.4.3. In Section 4.4.7, we study the effect of the combination of extra decision moments and chain relocations.

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

4.4.6 Relocation Time Bounds

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

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

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

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

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

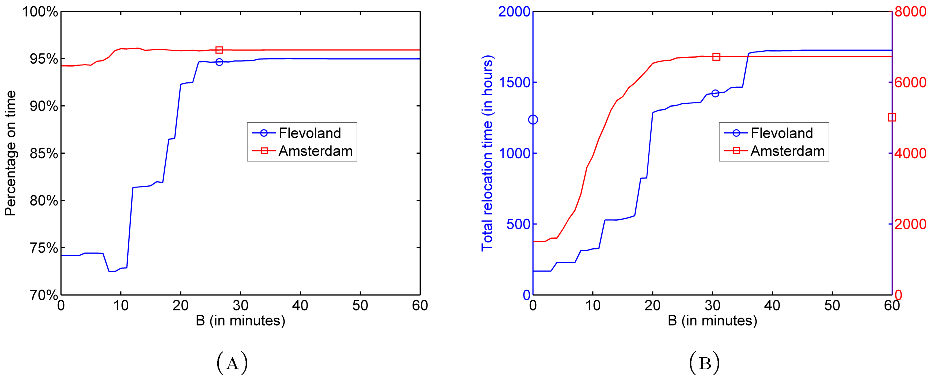


FIGURE 4.1: Percentage on time and total relocation time as a function of B.

In Figure 4.1a we observe a large difference in the system's behaviour. For Amsterdam, the bound B is of little influence only: the percentage of calls reached within the time threshold is close to 95% for all levels of B. In contrast, we see a huge improvement in performance for larger values of B in Flevoland: for $B < 12$ the percentage on time is below 75% and this increases up to approximately 95%. This phenomenon has a simple explanation: it is a consequence of both the size and the number of waiting sites and hospitals in Flevoland. The mean distances between two waiting sites are much larger, so for small values of B there are few possibilities for the destination of an ambulance after a service completion. Moreover, since there are only two hospitals in the region and the vast majority of the ambulances becomes available there, relocations to waiting sites 3, 4, 5 and 6 (in the enumeration of Figure 3.3) do not take place.

Another interesting point is the drop between $B = 7$ and $B = 8$ for Flevoland. This behaviour is due to one relocation in particular: the relocation time for an ambulance between the hospital in city 1 and waiting site 7 is exactly 7.5 minutes. Thus, for $B = 7$, an ambulance becoming free at this hospital moves to waiting site 1, regardless of the number of ambulances already present there. In contrast, for $B = 8$, this ambulance travels to waiting site 7, if unoccupied. The benefit of covering the southeastern part is outweighed by the performance loss in city 1. This aspect can be observed in the coverages displayed in Table 4.6 as well.

All large jumps are easily explained as well: the jump at $B = 12$ is due to the allowance of a relocation from 2 to 9; the one at $B = 18$ is due to the relocation from 1 to 3. If $B = 20$, it is now allowed to relocate an ambulance from 2 to both 4 and 5 as well. Finally, waiting site 6 can be reached from 2 if B exceeds 23 minutes. These jumps are largely visible in Figure 4.1b as well. Moreover, the large increase in total relocation time at $B = 36$ is due the fact that relocations from 1 to 4 and 6 both are acceptable now.

The pattern for Amsterdam is of different shape: the best performance is achieved for $10 \leq B \leq 13$, although the differences are minor. Apparently, it is beneficial to the performance if one chooses a relatively close waiting site if an

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

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

ambulance is newly free. That is, a local optimum that can be reached quickly performs better than a global one for which it takes long until that configuration is attained. A possible explanation for this phenomenon is the large number of events and thus decision moments in Amsterdam. This behaviour is also reflected in Table 4.7: the coverage levels corresponding to $B = 30$ are higher than for $B = 10$, although $B = 10$ yields a larger percentage on time. Note that there is also a reduction in mean response time of approximately 2.1% for $B = 10$ compared to $B = 30$.

4.4.7 Combinations

In this section, we combine different promising features and test the resulting methods for both regions. Moreover, we compare the performance to the penalty heuristic as presented in Chapter 3. We test the following combinations and methods:

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

2. DMEXCLP with extra decision moments, with chain relocations; busy time at the hospital follows the Weibull distribution considered in Section 4.4.4.

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

4. Penalty heuristic (see Chapter 3).

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

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

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

with the first columns in Tables 4.8 and 4.9, we see that performance improvements of 0.64 and 0.88 percentage points are achieved for Flevoland and Amsterdam, respectively. That is, the number of late arrivals decreases with 14.55% and 24.11%. This behaviour is probably explained by the following observation: it is more likely that a poor ambulance configuration arises just after the dispatch than when an ambulance becomes available. Therefore, at that decision moment, it is more important to attain the desired configuration quickly. This is achieved by using chain relocations, explaining the difference in performance.

If we compare columns 1, 2 and 3 in Tables 4.8 and 4.9, we barely see any differences in patient-based performance. This underlines the observations in Section 4.4.4. Results on crew-based performance are similar to those obtained in Section 4.4.4 as well.

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

Note that the penalty heuristic does not focus on coverage solely, but it uses the penalty function of Equation (3.2). One can observe in Tables 4.8 and 4.9 that minimizing the average response time is included in this penalty function

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

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

as well, as this method yields the shortest mean response time for both regions. In addition, the single coverage concept is used in the penalty heuristic. As a consequence, the mean single coverage levels are highest for the penalty heuristic, at the expense of a lower mean MEXCLP coverage.

If we modify the DMEXCLP method of Jagtenberg et al. (2015) in such a way that extra decision moments and chain relocations are allowed, we observe an improvement over other policies on most performance indicators if the coverage penalty function is used. In the next section, we consider different penalty functions and explore the performance of the DMEXCLP method with additional features.

4.4.8 Different Performance Criteria

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

- $\Phi_1(t) = \mathbb{1}_{\{t>720\}}$: the coverage penalty function, with a time threshold of 720 seconds.
- $\Phi_2(t) = t$: this penalty function focuses on minimization of the average response time.
- $\Phi_3(t)$: the penalty function of Equation (3.2), which is a compromise between minimizing late arrivals and minimizing average response times.

Results are displayed in Table 4.10. One might expect that the number of late arrivals and average response time are positively correlated. However, the results contradict this hypothesis: an increase of 6.00% resp. 9.42% in late arrivals is observed if one uses Φ_2 instead of Φ_1, for Flevoland and Amsterdam, respectively. In contrast, the average response time is reduced with 5.82% and 11.59%, respectively. Similar behaviour was also observed in Chapter 2.

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

4.5 Concluding Remarks

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

Based on the results in Table 4.10, we would suggest to use $\Phi_3(t)$ in a DMEXCLP environment. However, we want to note that $\Phi_1(t)$ makes for a fine alternative, as the results only differ slightly (7 to 20 seconds for the average response time). A reason to choose $\Phi_1(t)$ could be to make it easier to explain the behaviour of the system to EMS management and/or crew.

Adding extra decision moments (i.e., also relocating when a vehicle is dispatched to an incoming incident) is something we highly recommend in rural regions. We draw this conclusion based on the results in Table 4.3. For urban regions, we consider this an optional addition, that may be implemented if the region is willing to increase the crew's workload. Moreover, we recommend the use of chain relocations only if these extra decision moments are added. After all, Table 4.5 shows that no performance gain is achieved, while the workload on the crew is much higher. In contrast, if extra decision moments are added, the effect of chain relocations on the performance is much larger, c.f., Tables 4.8 and 4.9.

When it comes to ambulances involved in a drop-off at a hospital, our initial recommendation is to ignore them (in terms of coverage provided). The reason for this, is that including them makes the relocation strategy somewhat harder to implement (and explain), while it does not benefit the patients. An exception to this rule could be, when an ambulance service providers struggles with the workload of EMS crews: in that case, including the ambulances at hospital could be worthwhile, because it slightly reduces the relocation times (as seen in Table 4.4).

Part II

Offline Approaches

5

The Minimum Expected Penalty Relocation Problem for Ambulance Compliance Tables

The previous chapters were concerned with the online approach to the ambulance relocation problem: the majority of the computational effort is done at the decision moment itself. From this chapter onwards, we shift our focus to the offline approach, in which most computations are done a priori and the solutions are stored. When a certain situation occurs, the solution is consulted and applied, possibly proceeded by an additional short computation. Whereas online policies can take into account many characteristics of the state of the EMS system, this is impractical for offline methods as this would induce many system states for which a solution have to be computed in advance. Therefore, offline methods use little information about the state of the system.

A commonly used offline strategy is the *compliance table* policy. The system state in a compliance table is purely given by the number of available vehicles: each compliance table level indicates the desired waiting site locations for the available ambulances. If these ambulances are at their desired waiting sites, the system is said to be *in compliance*. The number of available ambulances changes when a request arrives or when an ambulance becomes available again. Then, each idle ambulance may be assigned to a different waiting site. As the number of units is bounded by the fleet size, computation of efficient compliance tables can be done offline.

To this end, we introduce the minimum expected penalty relocation problem (MEXPREP) in this chapter. In this problem, which we formulate as an integer linear program, one has the ability to control the number of waiting site relocations. Moreover, different performance measures related to response times, such as survival probabilities, can be incorporated. We show by simulation that the MEXPREP compliance tables outperform both the static policy and compliance tables obtained by the maximum covering relocation problem (MECRP), which both serve as benchmarks. Besides, we perform a study on different relocation thresholds and on two different methods to assign available ambulances to desired waiting sites.

This chapter is based on Van Barneveld (2016).

5.1 Introduction

In addition to the ability to calculate compliance tables offline, another important strength of the compliance table policy is that it is simple to explain to and to use by dispatchers, since the state of the EMS system is only described by the number of available ambulances. However, surprisingly little has been published about ambulance compliance tables, despite some practical advantages over online methods, as mentioned above. To the best of our knowledge, Gendreau et al. (2006) were the first to propose a methodology for computing compliance tables, by formulating the MECRP. Although the MECRP, which we will in summarize in Section 5.2.1 below, is a good and easily applicable model to compute compliance tables, it has some major limitations:

1. An area is covered if an idle ambulance is present within the coverage radius: multiple idle ambulances within the coverage radius do not contribute to the coverage of the area. Especially in an EMS system with a high call arrival rate, it may happen that another incident occurs before the idle ambulances reach the locations to which they are assigned, according to the compliance table. The MECRP does not take this into account: it only focuses on the next future emergency request. In other words, the MECRP utilizes the notion of single coverage, whereas probabilistic coverage may benefit the compliance table.
2. There are at least as many waiting site locations as ambulances. This is a rather strong assumption and not generally true in practice, although it tends to be more and more common in the US and Canada to park up (temporarily) at a street corner or other strategic hotspot. However, it may be dictated by law that ambulances are allowed to idle at designated ambulance base stations only.
3. The capacity of each waiting site location equals one. This may be true for designated ambulance parking spaces, but in general not for base stations.
4. Only a performance measure related to coverage can be incorporated.

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

The last limitation is related to coverage. As pointed out by De Maio et al. (2003), the most common EMS standard is to respond to 90% of all urgent calls within 8 minutes. Many EMS systems use the percentage of calls covered as performance measure. However, as stated by Erkut et al. (2008), the black-and-white nature of the coverage concept is an important limitation, and standard coverage models should not be used for ambulance location. First, coverage can result in large measurement errors because of their limited ability to discriminate between different response times. Second, these measurement errors are likely to result in large optimality errors when one uses covering models to locate ambulances instead of a model that takes survival probabilities into account. The difference between 'coverage' and 'survival' is demonstrated by an artifical example by Erkut et al. (2008), and it is shown that covering models can result in arbitrarily poor location decisions for ambulances.

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

The remainder of this paper is organized as follows. In Section 5.2.1 we explain the MECRP of Gendreau et al. (2006). In Sections 5.2.2 and 5.2.3 we treat the limitations mentioned above, resulting in the formulation of the MEXPREP in Section 5.2.4. In Section 5.3, we consider two models for the assignment problem, which needs to be solved to obtain an assignment of available ambulances to the waiting sites corresponding to the compliance table level. We conclude the paper by a numerical study in Section 5.4.

5.2 Model

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

5.2.1 Maximal Expected Coverage Relocation Problem

The MECRP is defined on a directed graph $G = (V \cup W, A)$ representing the region of interest. The region is discretized into demand zones, e.g., postal codes, in which V is the vertex set of these demand points. Moreover, W is the vertex set of potential waiting sites for n emergency vehicles and A is a set of arcs defined on $(V \cup W)^2$. A travel time is associated to each arc $(i, j) \in A$ and d_i denotes the demand at vertex $i \in V$. This d_i may, for instance, correspond to the population of demand zone i, or to the probability that an incoming emergency call occurs in demand zone i, which can be estimated by analyzing historical data. A vertex i is said to be covered by a vertex $j \in W$ if the expected travel time from j to i, denoted by τ_{ji}, is less than a given coverage radius T, expressed in time. We denote by W_i the subset of vertices of W covering i. We refer to Table 3.1 for an overview of the used notation.

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

$$\pi_k = \binom{n}{k}(1-p)^k p^{n-k}, \quad k = 0, 1, \ldots, n. \tag{5.1}$$

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

$$\begin{aligned}
\text{MECRP: Maximize} \sum_{k=1}^{n} \sum_{i \in V} d_i \pi_k y_{ik} & & (5.2)\\
\text{Subject to: } \sum_{j \in W_i} x_{jk} \geq y_{ik} & \qquad i \in V,\ k = 0, 1, \ldots, n & (5.3)\\
\sum_{j \in W} x_{jk} = k & \qquad k = 0, 1, \ldots, n & (5.4)\\
x_{jk} - x_{j,k+1} \leq a_{jk} & \qquad j \in W,\ k = 1, \ldots, n-1 & (5.5)\\
\sum_{j \in W} a_{jk} \leq \alpha_k & \qquad k = 1, \ldots, n-1 & (5.6)\\
x_{jk} \in \{0, 1\} & \qquad j \in W,\ k = 0, 1, \ldots, n & (5.7)\\
y_{ik} \in \{0, 1\} & \qquad i \in V,\ k = 0, 1, \ldots, n & (5.8)\\
a_{jk} \in \{0, 1\} & \qquad j \in W,\ k = 1, \ldots, n-1. & (5.9)
\end{aligned}$$

In this model, the objective function (5.2) maximizes the expected coverage. Constraints (5.3) induce that vertex $i \in V$ is covered only if at least one ambulance is located at at least one of the waiting sites in W_i, in compliance table level

k. Constraints (5.4) ensure that exactly k waiting sites are occupied in compliance table level k. Constraints (5.5) and (5.6) control the number of waiting site changes between compliance table levels k and $k+1$. The designated waiting sites at compliance table level k are given by decision variables x_{jk}. Although $k=0$ is included in the original MECRP by Gendreau et al. (2006), it is not necessary to include this case.

5.2.2 Expected Covered Demand

In the MECRP, the objective function for a given compliance table level k is to maximize the demand covered within the response time threshold. Then, each level is weighted according to π_k, the probability of being in a situation with k available ambulances, which can be computed using Equation (5.1). As stated by Gendreau et al. (2006), the MECRP reduces to the MCLP with k ambulances if $\pi_k = 1$. After all, always k ambulances are available, since $\pi_i = 0$ for $i \neq k$.

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

In the MEXCLP, the busy fraction is incorporated as follows: if vertex $i \in V$ is covered by k ambulances, the expected covered demand is $d_i(1-p^k)$. Moreover, the marginal contribution of the k^{th} ambulance equals $d_i(1-p)p^{k-1}$ (see also Section 4.3.1). This expression is incorporated in the objective value of the MEXCLP:

$$\text{MEXCLP: Maximize} \sum_{i\in V}\sum_{k=1}^{n} d_i(1-p)p^{k-1}z_{ik}, \tag{5.10}$$

$$\text{Subject to:} \sum_{j\in W_i} x_j \geq \sum_{k=1}^{n} z_{ik} \qquad i \in V \tag{5.11}$$

$$\sum_{j\in W} x_j \leq n \tag{5.12}$$

$$x_j \in \{0,1,\ldots,n\} \qquad j \in W \tag{5.13}$$

$$z_{ik} \in \{0,1\} \qquad i \in V,\ k=1,\ldots,n. \tag{5.14}$$

Here, $z_{ik}=1$ if and only if vertex i is covered by at least k ambulances. Note that constraint (5.12) is an inequality, while its MCLP counterpart is an equality. This is due to the concavity of the objective function in k for each i, which implies that if $z_{ik}=1$, then $z_{i1}=z_{i2}=\ldots=z_{ik}=1$ and if $z_{il}=0$, then $z_{i,l+1}=z_{i,l+2}=\ldots=z_{in}=0$. Moreover, the objective is to be maximized. Hence, constraint (5.12) will be satisfied at equality.

Analogous to the extension of the MCLP to the MEXCLP, we extend the MECRP, to address the first three shortcomings of the MECRP mentioned in Section 5.1. This is done by replacing the objective function of the MECRP, expression (5.2), by the following objective function:

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

where $z_{ikl} = 1$ if and only if in compliance table level k, vertex i is covered by at least l ambulances. Otherwise, $z_{ikl} = 0$. Moreover, constraint (5.3) is replaced by

$$\sum_{j \in W_i} x_{jk} \geq \sum_{l=1}^{k} z_{ikl}, \qquad i \in V, \ k = 1, \ldots, n. \tag{5.15}$$

This constraint is satisfied at equality by the same reasons as before. None of the other constraints of the MECRP change, except for constraints (5.7) and (5.8), which become $x_{jk} \in \{0, 1, \ldots, n\}$ and $z_{ikl} \in \{0, 1\}$, where $j \in W$, $i \in V$, $k = 1, \ldots, n$ and $l = 1, \ldots, k$. Moreover, constraint (5.9) is changed into $a_{jk} \in \{0, 1, \ldots, n\}$, where $j \in W$ and $k = 1, \ldots, n-1$.

5.2.3 General Performance Measures

As stated in Section 5.1, another limitation of the MECRP is the incapability to incorporate other EMS performance measures than coverage, such as patient survivability. This is a limitation of the MCLP and the MEXCLP as well. In this section we demonstrate how to incorporate different objectives in the MECRP. Similar to the previous chapters, we do this by introducing a non-negative non-decreasing penalty or cost function Φ, which is a function of the response time solely, with domain $\mathbb{R}_{\geq 0}$. A penalty function assigns to each different response time a penalty, and thus several performance measures related to response times can be incorporated. The commonly used EMS performance measure of coverage can be translated into the penalty function $\Phi(t) = \mathbb{1}_{\{t > T\}}$, where t denotes the response time and T the coverage radius, expressed in time. Other examples of objectives could be minimizing the average response time or minimizing the average lateness, modeled by penalty functions $\Phi(t) = t$ and $\Phi(t) = \max\{0, t - T\}$, respectively (see also Section 2.2.4). In addition, Erkut et al. (2008) consider survival functions, which we can use as penalty function as well (see Section 5.4).

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

As a consequence, constraint (5.3) of the MECRP needs to be replaced by a different constraint, which is able to take all available ambulances for each vertex into account. That is, for each vertex i, we need an ordering of ambulances according to their expected travel time to i, because we incorporated ambulance unavailability in our model: with probability $(1-p)$ the closest ambulance will respond to the request, generating a certain penalty $\Phi(t_1)$, with probability $(1-p)p$ the second closest ambulance will respond, generating penalty $\Phi(t_2) \geq \Phi(t_1)$, and so on, up to the k^{th} ambulance for compliance table level k. Moreover, to specify $\Phi(t_1), \Phi(t_2), \ldots, \Phi(t_k)$ for compliance table level k, we need to incorporate the expected travel times $t_1, t_2, \ldots, t_k$ in our model, since the penalty function relies on these.

As previously stated, the expected travel time from waiting site $j \in W$ to demand point $i \in V$ is denoted by τ_{ji}. If $\tau_{ji} \leq \tau_{j'i}$ then it holds that $\Phi(\tau_{ji}) \leq \Phi(\tau_{j'i})$ from the definition of the penalty function. Moreover, for the ordering of ambulances, we define $z_{ijkl} = 1$ if and only if for compliance table level k, the l^{th} closest ambulance to vertex i is at waiting site j. We need to introduce the constraint $\sum_{j \in W} z_{ijkl} = 1$ to ensure that at compliance table level k, there is exactly one ambulance that is the l^{th} closest to i. For an overview of the decision variables, we refer to Table 5.1. Now we have all the ingredients to formulate the minimal expected penalty relocation problem (MEXPREP).

5.2.4 Minimal Expected Penalty Relocation Problem

The MEXPREP is formulated as follows:

$$\text{Minimize} \sum_{k=1}^{n} \sum_{l=1}^{k} \sum_{i \in V} \sum_{j \in W} \pi_k d_i (1-p) p^{l-1} \Phi(\tau_{ji}) z_{ijkl} \tag{5.16}$$

$$\begin{aligned}
\text{Subject to:} \sum_{l=1}^{k} z_{ijkl} &= x_{jk} && i \in V, j \in W, k = 1, \ldots, n && (5.17)\\
\sum_{j \in W} z_{ijkl} &= 1 && i \in V,\ k = 1, \ldots, n,\ l = 1, \ldots, k && (5.18)\\
\sum_{j \in W} x_{jk} &= k && k = 1, \ldots, n && (5.19)\\
x_{jk} - x_{j,k+1} &\leq a_{jk} && j \in W,\ k = 1, \ldots, n-1 && (5.20)\\
\sum_{j \in W} a_{jk} &\leq \alpha_k && k = 1, \ldots, n-1 && (5.21)\\
x_{jk} &\in \{0, 1, \ldots, n\} && j \in W,\ k = 1, \ldots, n && (5.22)\\
z_{ijkl} &\in \{0, 1\} && i \in V,\ j \in W,\ k = 1, \ldots, n\ l = 1, \ldots, k && (5.23)\\
a_{jk} &\in \{0, 1, \ldots, n\} && j \in W,\ k = 1, \ldots, n-1. && (5.24)
\end{aligned}$$

Note that there is only a contribution to the objective value if $z_{ijkl} = 1$, i.e., if for compliance table level k, the l^{th} closest ambulance to vertex i is at waiting

x_{jk}	Number of ambulances placed at waiting site $j \in W$ in compliance table level k.
a_{jk}	Difference in number of occurences of waiting site $j \in W$ in level k compared to level $k+1$.
z_{ijkl}	Equals 1 iff in compliance table level k, the l^{th} closest ambulance to vertex $i \in V$ is at waiting site $j \in W$, and 0 otherwise.

TABLE 5.1: Decision variables of the MEXPREP.

site j. The marginal contribution of this l^{th} closest ambulance to vertex i is $d_i(1-p)p^{l-1}\Phi(\tau_{ji})$ for given vertex i, waiting site j, and compliance table level k. That is, with probability $(1-p)p^{l-1}$, the l^{th} closest ambulance to vertex i is the closest available one, inducing a penalty of $d_i\Phi(\tau_{ji})$. Such as in the MECRP, each compliance table level k is weighted according to the probability that the system is in a situation with k available ambulances, as computed in Equation (5.1).

Constraints (5.17) and (5.18) take over the role of constraint (5.3) in the MECRP formulation. In constraint (5.17), both the left- and the right-hand side represent the number of ambulances at waiting site j for compliance table level k. Note that no i-index is present in the right-hand side. Since constraint (5.17) holds for each $i \in V$, it is immediately forced that

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

This should hold in a feasible solution to the problem, since for level k all the ambulances at waiting site j contribute to the penalty induced by each demand point in the objective function. As stated before, constraint (5.18) ensures that at compliance table level k, there is exactly one ambulance that is the l^{th} closest to i. All the other constraints are the same as the constraints in the MECRP formulation, except for the integer and binary constraints. Note that since the objective is to be minimized and the penalty function $\Phi(t)$ is non-decreasing in t, we do not require constraints related to the ordering of ambulances.

5.2.5 Adjusted MEXPREP

In the MEXCLP-formulation of Daskin (1983), some simplifying assumptions are made: (1) ambulances operate independently, (2) each ambulance has the same busy fraction, and (3) ambulance busy fractions are invariant with respect to the ambulance locations. Moreover, the MEXPREP formulation, like the formulations of MEXCLP and MECRP, assumes that the busy fraction is an input. However, in reality, the busy fraction p is an output as the service rate that is needed to calculate the busy fraction depends on the allocation of ambulances to waiting sites. The use of a universal busy fraction is a rough approximation of reality, since the actual busy fractions depend on both the compliance table itself and on the dispatch policy.

Batta et al. (1989) consider an adjustment of the objective function in the MEXCLP, relaxing the assumptions on busy fractions. In this problem, called the AMEXCLP, correction factors $Q(n, p, k)$, $k = 0, \ldots, n-1$, derived by Larson (1975), are incorporated in the objective function of the MEXCLP. We extend the MEXPREP to the AMEXPREP by incorporating the correction factors $Q(n, p, k-1)$ in Equation (5.16), where

$$Q(n,p,k) = \frac{\sum_{j=k}^{n-1} \frac{(n-k-1)!(n-j)}{(j-k)!} \frac{n^j}{n!} p^{j-k}}{(1-p)\sum_{i=0}^{n-1} \frac{n^i}{i!} p^i + \frac{n^n p^n}{n!}}, \quad k = 0, \ldots, n-1, \tag{5.25}$$

analogous to the work done by Batta et al. (1989). In Section 5.4.6, we explore the differences between the MEXPREP and the AMEXPREP.

5.3 Assignment Problem

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

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

In contrast to the computation of compliance tables, fast methods exist for solving the MWBM and the LBAP, e.g., the Hungarian Method of complexity $\mathcal{O}(n^3)$ for MWBM and the Threshold Algorithm of complexity $\mathcal{O}(n^{2.5}/\sqrt{\log n})$ for LBAP, both explained by Burkhard et al. (2009). Hence, this can be done in real-time and an offline solution is not necessary. After all, this would require a complex state dependent policy which shows relocation moves for every realized state of the system. Moreover, an online implementation of the assignment problem allows takes into account the actual locations of driving ambulances and hence a redirection of ambulances to different waiting sites. Therefore, we recommend to compute compliance tables offline, and the assignment problem online. In Section 5.4.4, we will explore the differences in the MWBM and the LBAP.

5.4 Computational Study

The MEXPREP computes compliance tables taking into account ambulance unavailability, general performance measures, and a restriction on the number of waiting site changes. We apply the MEXPREP to the Amsterdam EMS region, extensively described in Section 3.4.2. In this chapter, we assume that ambulances can idle at any of the 17 dots depicted in Figure 3.4. Results are generated by simulation using historical data.

5.4.1 Experimental Setup

Historical data on emergency requests in the year 2011 was provided by Ambulance Amsterdam, which runs the emergency medical services in this region. We only consider the time-period between 7 AM and 6 PM, like in Section 4.4.1. During the considered time-period, 33 ambulances are present in the system. However, of these ambulances, many are busy with ordered transport: taxi-type transport of patients not able to travel to the hospital themselves, usually scheduled in advance. Therefore, we assume a fleet size of 21 in our computations.

In 2011 between 7 AM and 6 PM, the total number of emergency requests was 44,966, yielding an hourly arrival rate of 11.2 requests. Only 44,520 of these requests are useful, because of the remainder historical data was not complete. We build a trace on this data and simulate it in a discrete-event simulation. We refer to Section 4.4.1 for an enumeration of the incident related information included in the trace. Unlike the mentioned section, we do not remove days, as the fleet capacity of 21 ambulances is satisfactory. We connect the 365 days in the trace such that 6 PM is followed directly by 7 AM, for the same reason as explained in Section 4.4.1.

We also use historical data to compute the busy fraction, by dividing the total patient-related work during these 4,015 hours by the total duty time of 21 ambulances, to obtain a busy fraction of $p = 0.43047$. The average busy time (excluding relocation time after transferring the patient at the hospital) of an ambulance is 0.82 hours. The annual number of emergency requests ranges between 2 (in a postal code somewhere between waiting sites 9 and 13) and 1,545 (in the city center of Amsterdam, near waiting site 1), with an average demand of 275 per node. We define d_i as the probability that an incoming request occurs in vertex i, computed by normalization of the number of emergency requests.

We assume a deterministic dispatch time of 120 seconds and a deterministic chute time of 60 seconds for ambulances at a waiting site. There is no chute time if the dispatched ambulance is already on the road. Moreover, the pre-trip delay for moving an ambulance from a waiting site to another one is assumed to be 180 seconds. The ambulance that can be present fastest at the emergency scene is always dispatched to the request.

We perform simulations using the computed compliance tables and the actual emergency requests in the region during the daytime of the year 2011. To keep track of the actual locations of ambulances, we use the travel routes as computed in Section 3.4.3, which use the travel time table estimated by the RIVM (Kommer and Zwakhals, 2008) as input. The relocation travel times were computed by multiplying the emergency travel times by a factor $\frac{10}{9}$. Computation of the assignment of ambulances to waiting sites is done online by solving either the MWBM or the LBAP during the simulation. We test performance according to six statistics:

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

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

$$\Phi_1(t) = \mathbb{1}_{\{t>720\}}, \tag{5.26}$$

$$\Phi_2(t) = t, \tag{5.27}$$

$$\Phi_3(t) = 1 - (1 + e^{0.679+0.0044t})^{-1}, \tag{5.28}$$

$$\Phi_4(t) = 1 - (1 + e^{0.113+0.0041t})^{-1}, \tag{5.29}$$

$$\Phi_5(t) = 1 - (1 + e^{0.04+0.005t})^{-1}. \tag{5.30}$$

Function Φ_1 is based on coverage, in which we consider a response time threshold of 720 seconds (12 minutes). Φ_2 represents the penalty function focusing on the

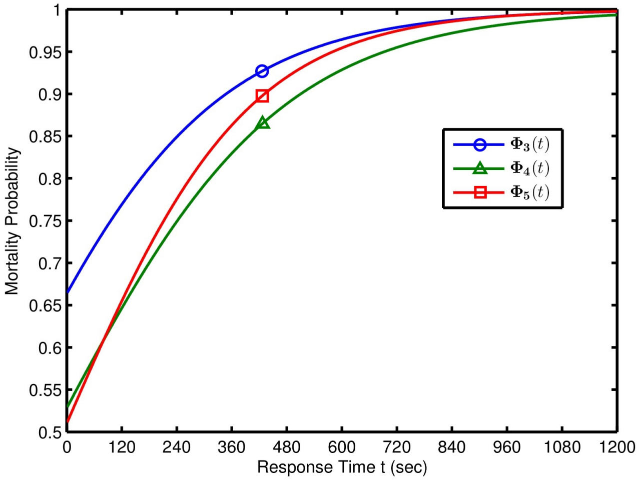


Figure 5.1: Mortality probabilities as a function of the response time.

objective of minimizing the average response time. Functions Φ_3, Φ_4, and Φ_5 represent the survival functions of De Maio et al. (2003), Valenzuela et al. (1997), and Waaelwijn et al. (2001), respectively, in a penalty function (mortality) setting. A graphical representation of Φ_3, Φ_4, and Φ_5 is given in Figure 5.1.

5.4.2 Comparison of MEXPREP with MECRP

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

We only display the compliance tables for the $\alpha_k = 0$ case, since these compliance tables are nested and thus can be represented efficiently. We represent such a nested compliance table by a one-dimensional vector, where compliance table level k is given by entries 1 up to k. The computed MECRP and MEXPREP compliance tables for $\alpha_k = 0$, for the two different settings are displayed in Equations (5.31) and (5.32), respectively. Note that none of these four compliance tables equals an-

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

TABLE 5.2: Simulation results for 17 ambulances and penalty function $\Phi(t) = \mathbb{1}_{\{t>720\}}$, based on 44,520 requests in 2011.

other, although the two MECRP-tables are very similar. Simulation results, using MWBM as assignment policy, for these compliances tables are listed in Tables 5.2 and 5.3, respectively. These tables include 95% confidence intervals around the percentage of requests responded to within 720 seconds.

$$\begin{aligned} \text{MECRP:} \quad & (1,16,12,14,5,9,17,8,11,15,3,10,4,13,2,6,7) \\ \text{MEXPREP:} \quad & (1,1,6,16,6,15,2,10,16,14,1,10,15,9,6,17,12) \end{aligned} \tag{5.31}$$

$$\begin{aligned} \text{MECRP:} \quad & (1,16,12,14,5,9,17,8,10,15,11,3,4,13,2,6,7) \\ \text{MEXPREP:} \quad & (1,6,16,1,15,10,2,14,6,16,10,9,17,2,12,14,5) \end{aligned} \tag{5.32}$$

Note that in Table 5.2 as well as in Table 5.3, the MEXPREP significantly outperforms the MECRP on the most important performance indicator: the percentage of requests responded to within the response time threshold of 720 seconds. We observe improvements on this criterion between 0.7% (second setting, $\alpha_k = 0$) and 1.89% (first setting, $\alpha_k = 1$). Moreover, this performance gain is achieved with fewer relocations, although the average relocation time is longer for MEXPREP. A small disadvantage of the MEXPREP compared to the MECRP is the computation time. However, as stated before, the computation time of the MEXPREP compliance tables is of less importance, since the problem can be solved in an offline fashion.

Observing the results listed in Tables 5.2 and 5.3, we note that the benefit of allowing non-nested compliance tables is very marginal with respect to the percentage of requests for which the response time threshold is achieved, and to the average response time. In some cases it is even disadvantageous to allow more

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

TABLE 5.3: Simulation results for 21 ambulances, compliance tables up to level 17 and penalty function $\Phi(t) = \mathbb{1}_{\{t>720\}}$, based on 44,520 requests in 2011.

than zero waiting site changes. Besides that, in the second setting the MEXPREP computes the same compliance tables for $\alpha_k = 0$, $\alpha_k = 1$ and $\alpha_k = \lceil \frac{k}{2} \rceil$. However, the effect on the number of relocations is large if one uses the compliance tables with no restrictions on waiting site changes rather than compliance tables with restrictions. The question arises whether this marginal performance improvement outweighs this increase in number of relocations. In line with Gendreau et al. (2006), the average relocation time decreases if more waiting site changes are allowed, as expected.

5.4.3 Relocation Thresholds

The number of relocations in Table 5.3 is quite large. For instance, for the MEXPREP with $\alpha_k = 0$, the average number of relocations per day is 32. This is due to the large number of changes in availability of ambulances. After all, each time an ambulance is dispatched or finishes service, relocations may be performed. However, one could argue the effect of ambulance relocations if enough ambulances are still available. As example, it probably makes no sense to relocate ambulances if $n - 1$ instead of n ambulances are available, since frequent movements may inconvenience ambulance crews. A way to address this is the introduction of a *relocation threshold*, denoted by K. If the number of available ambulances is below this threshold, we use the compliance table policy. However, if this is not the case, we carry out the *static policy*: we perform no relocations if an ambulance is dispatched, and we send a newly finished ambulance back to its home waiting site. If a transition from level K to $K + 1$ occurs, each ambulance is sent back to its home waiting site. Note that these ambulance movements do *not* contribute to the number of relocations, as it is beneficial from the crew's perspective to be present at the home waiting site.

The determination of the ideal level of this relocation threshold K is an interesting topic. If K is too high, it is possible that too many relocations are performed. On the other hand, a low value of K may result in a worse performance of an ambulance service provider. To investigate the behavior of different relocation thresholds K, we compute compliance tables by the MEXPREP for $K = 7$, $K = 14$, and $K = 21$, for the five different penalty functions of (5.26)–(5.30), where $\alpha_k = 0$. That is, we change n in the MEXPREP-formulation to K and compute K compliance table levels. Except for the fact we do not change the π_k-values in the objective function, we compute the MEXPREP as if there were K ambulances instead of n.

In addition, we compute an initial configuration of the $n = 21$ ambulances by an ordinary location problem, which is a modification of the MEXPREP, as follows. In the MEXPREP, we set $k = 21$ in all constraints and in the objective function. Moreover, we discard constraints (5.20), (5.21) and (5.24), as well as π_k in the objective function. Note that for penalty function Φ_1 this modification of the MEXPREP is equivalent to the MEXCLP.

Then, we simulate our system for $K = 0$ (the static policy), $K = 7$, $K = 14$ and $K = 21$, starting in the initial configuration. This initial configuration also determines the home waiting site of each ambulance. In the simulation, we solve the MWBM to obtain a solution to the assignment problem. The results are listed in Table 5.4.

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

Interestingly, fewer ambulance relocations are performed when a relocation threshold $K = 21$ is used instead of $K = 14$. This behavior is easily explained by the following observation: the majority of the ambulance relocations are done when a transition from level $K + 1$ to K occurs. If $K = n = 21$, there are no transitions from level $K+1$ to level K. Due to the nesting of the compliance table, relatively few ambulance relocations are performed. However, for $K = 14$, there are many transitions from level 15 to level 14. Together with the fact that level 14 is generally not nested in the ambulance configuration with 15 ambulances, many ambulance relocations are carried out. This behavior is also reflected in Figure 5.2, where the total number of relocations and mean penalty as a function of K is displayed. It is not a surprise that the peak of the number of relocations is at $K = 12$. After all, the mean number of available ambulances is between 12 and 13, so many transitions from a situation with 13 to a situation with 12 available ambulances take place.

Note that for the static policy $K = 0$, the performance indicators differ for the considered penalty functions in general, although no compliance table policy is carried out. This is a direct consequence of the differences in the initial con-

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

Table 5.4: Simulation results for several levels of K, $n = 21$ ambulances and $\alpha_k = 0$, based on 44,520 requests in 2011.

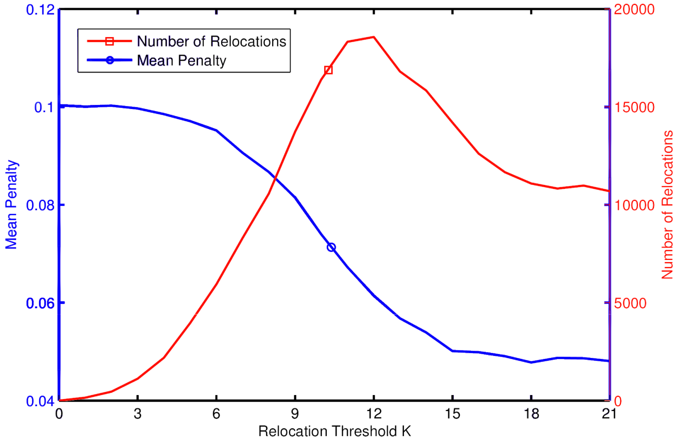


FIGURE 5.2: Total number of relocations and mean penalty as a function of the relocation threshold K, for 21 ambulances, $\alpha_k = 0$ and penalty function $\Phi(t) = \mathbb{1}_{\{t>720\}}$.

figurations. Moreover, it is worth noting that the coverage penalty function Φ_1 is outperformed by the average response time penalty function Φ_2 on the percentage on time criterion, despite the fact that Φ_1 focuses on maximizing this percentage. This underlines the conclusion made by Erkut et al. (2008) about the weakness of models based on coverage.

5.4.4 Assignments

We proceed this numerical study with a comparison of the two models for solving the assignment problem, mentioned in Section 5.3, namely the MWBM and the LBAP.

The results in Table 5.5 show that using the LBAP for the assignment problem results in a slightly better performance regarding the patient-based performance indicators. This small increase is explained by the observation that the LBAP minimizes the maximum travel time of a relocated ambulance. As a consequence, the ambulance configuration corresponding to the new compliance table level is attained faster. Hence, as expected, the average relocation time per ambulance decreases drastically. After all, using the LBAP, a long trip of one ambulance is split into multiple shorter trips, thus reducing the average relocation time per ambulance. However, the total number of relocations is approximately quadrupled with respect to the usage of the MWBM as assignment problem. This is probably not acceptable from the crew perspective. It is up to the ambulance service provider to decide whether this tremendous increase of number of relocations outweighs the benefits of the increase in patient-based performance.

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

Table 5.5: Simulation results for $n = K = 21$ and $\alpha_k = 0$, based on 44,520 requests in 2011.

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

Table 5.6: Expected number of survivors for $n = 21$ and $\alpha_k = 0$, based on 44,520 requests in 2011.

5.4.5 Expected Number of Survivors

Another interesting indicator that provides insight into the performance of the compliance tables, is the expected number of survivors. This expected number is easily computed by the summation of the 44,520 penalties for the survival functions Φ_3, Φ_4 and Φ_5. Moreover, we perform cross-comparisons of these functions: we evaluate the compliance table corresponding to the solution of the MEXPREP for one specific penalty function (rows) using the other ones (columns), for both the MWBM and the LBAP. The results are listed in Table 5.6.

If one considers the rows corresponding to Φ_3, Φ_4 and Φ_5 in Table 5.6, one may observe that the differences within these columns are small: the numbers differ at most by 0.5%. We conclude that the chosen survival function is not of influence on the maximization of survivors. In contrast, the number of survivors differs for the compliance tables induced by the penalty functions based on coverage and average response times, Φ_1 and Φ_2, respectively. Especially for Φ_1, this difference

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

Table 5.7: MEXPREP objective values and simulated penalties for $n = 21$, based on 44,520 requests in 2011.

is around 5% compared to the survival functions. However, the difference between the survival functions and Φ_2 is relatively minor. As a consequence, it seems that the average response time is a better approximation for survival than coverage.

As can be observed in Table 5.6, there are differences between the MWBM and the LBAP. For instance, the expected number of survivors using the LBAP increases with approximately 2.6% with respect to the case in which the MWBM is used as assignment problem, for Φ_3. This was to be expected due to the increase in performance of the LBAP with respect to the MWBM, as can be observed in Table 5.5. The expected number of survivors is smallest when the compliance tables are evaluated using penalty function Φ_3. This is explained by the fact that Φ_3 is the most pessimistic survival function (see Figure 5.1).

5.4.6 AMEXPREP

In Section 5.2.5, we discussed some limitations and assumptions on busy fractions. These assumptions may result in a objective value of the MEXPREP that differs from the values computed through simulation. In Table 5.7, objective and simulated values are listed for the two extremes $\alpha_k = 0$ and $\alpha_k = k$, for both the MWBM and the LBAP.

From Table 5.7, we conclude that MEXPREP's estimation of the system performance is somewhat too pessimistic. This is most evident in Φ_1, in which the relative gap between objective value and simulated values is largest. Moreover, we observe a difference only in the fourth digit in the objective values for $\alpha_k = 0$ and $\alpha_k = k$ for Φ_1. From this observation, one could draw the conclusion that nested compliance tables are already close to optimal. This is also underlined by the simulated values. In all cases, the simulated values using the MWBM are closer to the objective values than in the simulation that uses the LBAP as the assignment problem. This is as expected, since the use of the LBAP results in better patient-based performance (see Table 5.5).

As opposed to the objective values of the MEXPREP, the AMEXPREP presented in Section 5.2.5 provides an optimistic estimation of the system performance, as can be observed in Table 5.8. For the penalty functions based on survival, Φ_3, Φ_4 and Φ_5, the objective value of the AMEXPREP differs more from

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

TABLE 5.8: AMEXPREP objective values and simulated penalties for $n = 21$ and $\alpha_k = 0$, based on 44,520 requests in 2011.

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

TABLE 5.9: Deviations of unrestricted MEXPREP compliance tables with respect to actual capacities and restricted MEXPREP for $\alpha_k = 0$.

the simulated values than is the case for the MEXPREP. Surprisingly, for Φ_1 and Φ_2 it is the opposite. At last, it is worth noting that the AMEXPREP performs slightly better than the MEXPREP on the penalty criterion in general.

5.4.7 Base Station Capacities

In this section, we solve the MEXPREP, taking into account the actual waiting site capacities depicted in Figure 3.4. These restrictions can easily be incorporated in the MEXPREP by introducing constraints of the type

$$x_{jk} \leq c_j \qquad\qquad j \in W,\ k = 1, \ldots, n-1, \qquad (5.33)$$

where c_j denotes the capacity of waiting site $j \in W$. We compute the restricted version of MEXPREP for $\alpha_k = 0$. We compare the obtained compliance table to the actual capacities. The number of deviations is reported in the columns *c1* in Table 5.9. For instance, for Φ_1, the number of capacity violations is 13 for the whole compliance table, and these violations occur in levels nine up to 21. In addition, columns *c2* report the numbers for the restricted compliance table compared to the unrestricted one. Note that the compliance tables consist of 231 numbers in total.

Only for Φ_1 the computation of the restricted MEXPREP results in a different objective value compared to the unrestricted MEXPREP: 0.0576. For the other penalty functions, the objective values do not differ in the first four digits, although different compliance tables were generated, as can be observed in Table 5.9. From this observation, one could draw the conclusion that minor differences in compliance tables are hardly noticed in the objective value: there are many compliance

$\lvert V\rvert$	100	200	300	400	500	600
Number of variables	3.2×10^5	6.3×10^5	9.5×10^5	1.3×10^6	1.6×10^6	1.9×10^6
Number of constraints	3.1×10^5	6.2×10^5	9.2×10^5	1.2×10^6	1.5×10^6	1.8×10^6
CPU time $\Phi_2,\ \alpha_k = 0$	53 s	168 s	348 s	695 s	1119 s	1770 s
CPU time $\Phi_5,\ \alpha_k = 0$	38 s	196 s	387 s	808 s	1182 s	1689 s
CPU time $\Phi_2,\ \alpha_k = k$	63 s	197 s	459 s	595 s	1049 s	1594 s
CPU time $\Phi_5,\ \alpha_k = k$	47 s	189 s	349 s	576 s	997 s	1650 s

Table 5.10: Computation times for the artificial problem instance.

tables that are near-optimal. It is also interesting to see that there are deviations in lower levels for penalty functions Φ_3, Φ_4 and Φ_5 with respect to the restricted compliance table, while these are not present in the middle levels.

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

5.4.8 Computation Times

We conclude this section with an investigation on computation times of the MEXPREP. Unfortunately, we are not able to investigate the increase in computation time by choosing a different demand aggregation for the considered case, since we only have access to travel times between 4-digit postal codes. As an alternative, we create an artificial problem instance: we pick $|V|$ demand nodes out of a grid of size 100×100, for different values of $|V|$, and assign demand probabilities to them. Travel times between nodes are calculated by the Manhattan metric. For the base locations, we select $|W| = 15$ points, and we consider $n = 20$ ambulances. Then, we solve the MEXPREP for the extremes $\alpha_k = 0$ and $\alpha_k = k$, and for Φ_2 and Φ_5, since in Table 5.5 the computation time of these penalty functions is shortest and longest, respectively. Results on computation times, as well as number of variables and constraints (namely, Equations (5.17)–(5.21)) are listed in Table 5.10.

For large values of $|V|$, it takes more time to obtain a solution for $\alpha_k = 0$ compared to $\alpha_k = k$, as can be observed in Table 5.10. The explanation of this phenomenon is probably in the method CPLEX uses to compute a solution. From Tables 5.4 and 5.5 one may conclude that the use of Φ_2 and Φ_5 induce the shortest and longest computation times, respectively. However, Table 5.10 shows that Φ_2 did not consistently result in shorter computation times than Φ_5.

5.5 Concluding Remarks

In this paper, we presented the minimum expected penalty relocation problem (MEXPREP) to compute compliance tables. The MEXPREP is an extension of the maximal covering relocation problem (MECRP) formulated by Gendreau et al. (2006) in two directions. First, we incorporated the objective function of the

MEXCLP into the objective function of the MECRP, to anticipate multiple future emergency requests beyond a first request. Then, we introduced penalty functions in order to focus on performance measures other than coverage, including survival probabilities. Moreover, based on the assumptions and limitations of busy fractions, we introduced an adjusted version of the MEXPREP. In this adjusted version, called the AMEXPREP, correction factors proposed by Batta et al. (1989) were incorporated. Additionally, we considered both the minimum weighted bipartite matching problem (MWBM) and the linear bottleneck assignment problem (LBAP) as assignment problem for the assignment of available ambulances to the waiting sites indicated by the compliance table level.

We concluded this paper with a numerical study, based on 44,520 emergency requests in 2011 in the region of Amsterdam and its surroundings. In this study, we compared the MEXPREP compliance tables to both the MECRP compliance tables and the static policy, and we observed that the MEXPREP outperforms both of them on most performance indicators. We also carried out a comparison between several restrictions on waiting site changes. Moreover, we considered several relocation thresholds, and compared both the resulting performance when using the LBAP and the MWBM as assignment problems. In addition, we compared the objective values with the simulated values for both the MEXPREP and the AMEXPREP. Studies regarding computation times of the MEXPREP and the effect of base station capacities were conducted as well.

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

6

Compliance Tables for an EMS System with Two Types of Medical Response Units

Like Chapter 5, compliance tables are the topic of this chapter as well. However, a key difference between this chapter and the previous one is the fleet homogeniety. Before, we assumed that only one type of ambulance is used. However, in the Netherlands, several types of medical response units are used. In addition to the regular ambulances, there are for instance mobile intensive care units and trauma helicopters. Additionally, the use of a new type of response unit is emerging: so-called *rapid responder ambulances* (RRAs). Recently, the Dutch Minister of Public Health was questioned by the parliament regarding the deployment of these RRAs (Schippers, 2014). These units are typically motor cycles, used for fast first response to an emergency request. They are staffed by highly educated persons equipped with the same gear regular ambulance personnel takes inside a patient's house in order to provide Advanced Life Support (ALS). Basically, there are two differences between RRAs and regular transport ambulances (RTAs): RRAs are faster, but they lack the ability of RTAs to transport a patient to a hospital.

To maintain the ability to respond to emergency requests timely when ambulances get busy, we consider so-called *two-dimensional* compliance tables for proactive relocation purposes. In this chapter, we propose an integer linear program (ILP) formulation for the computation of compliance tables for an EMS system with two types of vehicles. We use outcomes of a Hypercube model (see Larson (1975)) as input parameters in the ILP. Moreover, we include nestedness constraints and we set bounds on the relocation times. To obtain more credible results than the objective function of the ILP solely, we simulate the computed compliance tables for different input parameters. Results show that bounding the time a relocation may last seems beneficial. Besides, including the nestedness constraints ensures that the number of relocations and the relocation time can be bounded, while the performance stays unaffected.

This chapter is based on Van Barneveld et al. (2017).

No. of RRAs	No. of RTAs	Base stations					
		1	2	3	4	5	6
1	0	R	0	0	0	0	0
2	0	R	R	0	0	0	0
0	1	0	0	0	T	0	0
1	1	R	0	0	T	0	0
2	1	R	R	0	T	0	0
0	2	0	T	0	0	T	0
1	2	R	T	0	T	0	0
2	2	R	R,T	0	T	0	0

TABLE 6.1: The two-dimensional compliance table indicates the desired locations for the available RRAs and RTAs.

6.1 Introduction

This chapter is concerned with the problem of computing compliance tables in an EMS system with multiple ambulance types. We observe that almost all models with multiple vehicle types make a distinction in the level of care an ambulance can provide: either Advanced (ALS) or Basic Life Support (BLS), and ambulances are classified as such, the MEXPREP2 model of McLay (2009) being an exception. In this model, ALS and BLS ambulances are considered, but the author introduces another distinction as well: ALS ambulances are non-transport Quick Response Vehicles (QRVs), comparable to the RRAs studied in this chapter. The regular transport ambulances are limited to provide BLS care, being a difference with this chapter in which transport ambulances are also able to provide ALS care.

Computing compliance tables for an EMS system with two vehicle types brings forth additional complexity to the usual approach in which only one type of ambulance is considered. After all, the state of the system in our model is described by the number of available units of both types, making it two-dimensional. For each of these states an ambulance configuration for both types of units needs to be computed in a so-called two-dimensional compliance table. We refer to Table 6.1 for an example of a two-dimensional compliance table.

To the best of our knowledge, the problem of computing compliance tables for an EMS system with two types of medical response units has not been studied before. We extend both the MECRP model by Gendreau et al. (2006) and the MEXCLP2 model proposed by McLay (2009), of which we also use the modification of the Hypercube model by Jarvis (1985) for the estimation of the input parameters (e.g., busy fractions). In the ILP formulation of our problem, we incorporate cohesion between the different compliance table levels in two different ways. First, we restrict the number of ambulances that is instructed to relocate at the same decision moment, per vehicle type. There are several reasons why this restriction on a compliance table would be incorporated. For instance, the budget an ambulance service provider may spend is limited and costs, (e.g., fuel and redemption) are involved with each relocation. Moreover, as stated before, relocations are not popular among the ambulance personnel. This restriction on simultaneous moves is also present in the MEXPREP presented in Chapter 5, and in the MECRP of Gendreau et al. (2006).

In addition to the *nestedness* constraints mentioned above, we also impose bounds on the time a relocation may take in the compliance table (see Section 4.4.6). Without these restrictions, it is possible that a long trip of an ambulance is needed to attain the ambulance configuration indicated by the compliance table. However, another event may occur during this relocation with high probability, e.g., a busy ambulance becomes available, or another incident occurs. In case of the latter, the system may possibly not be able to respond to the new incident timely, as the system is 'out of compliance' due to the fact that the relocated ambulance has not arrived at its new location. Therefore, it is desirable to be in compliance, according to the compliance table, as quickly as possible. Moreover, such bounds are desirable from the crew's perspective since these limit the time medical personnel spends on the road.

Moreover, to get a more realistic idea about the effect of applying relocation policies, such as compliance tables, it is useful to perform simulation experiments, as stated at the end of Chapter 4. Although objective values in a mathematical model serve as approximations of the performance of the EMS system, ambulance service providers are far more interested in the relocation policy itself rather than in theoretically computed numbers. It is not impossible that policies yielding good theoretical results perform worse in practice compared to ones with inferior theoretical results, and vice versa. Analyzing the simulation results of these two-dimensional compliance tables, we obtain several interesting insights.

The remainder of this chapter is organized as follows. In Section 6.2 we describe the EMS process. The dispatch process differs from the process explained in Chapter 1 as we now have two types of vehicles. Section 6.3 is concerned with the presentation of the ILP model for the computation of two-dimensional compliance tables. We also describe how we estimate the input parameters (e.g., busy fractions) and we provide the formulation of the mentioned constraints. The chapter is concluded with a numerical study based on the EMS region of Flevoland in Section 6.4.

6.2 System Dynamics

In this section, we describe the EMS system dynamics studied in this chapter. This process differs from the one described in Chapter 3 and onwards, as a different dispatching policy is used due to multiple ambulance types.

When idle, both RRA and RTA crews spend their shift at base stations. In our setting it is assumed that there are more medical units than base stations, resulting in multiple occupancy of one or more base stations. This is common in the Netherlands and this assumption differs from the one done in the compliance table model by Sudtachat et al. (2016), in which each base station can be occupied by at most one vehicle. If the situation requires, medical units may be asked to

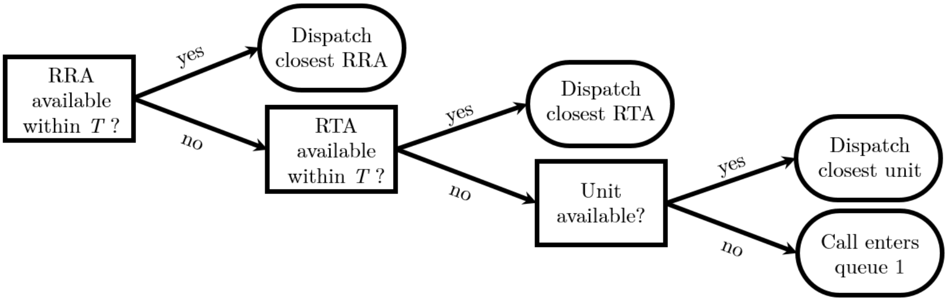


FIGURE 6.1: Dispatch policy of the first response.

relocate to other base stations. These decisions are made when the number of available ambulances changes, e.g., when an ambulance is instructed to respond to a call or when a unit finishes service.

In case of the first event type, a medical unit needs to be dispatched to the patient. As we do not distinguish between ALS and BLS type of care, we assume a single type of call: a patient always needs ALS care as soon as possible. The dispatch policy is as follows: if there is at least one RRA available that can reach the patient within the time threshold, the closest (in time) RRA is dispatched. Otherwise, an available RTA present within the time threshold is selected to respond to this call. In the situation in which neither an RRA nor an RTA can respond to the patient timely, the nearest medical unit is assigned, regardless of the type. Such a response counts as a *late arrival.* If no unit at all is available for the response, the call enters a first-come first-served queue: the first unit that becomes available is dispatched. Figure 6.1 shows a graphical representation regarding the first reponse dispatch policy.

We assume that it is not known beforehand whether the patient needs transportation to a hospital. This information becomes available at the emergency control center when a unit arrives at the emergency scene. After all, it is typically difficult to determine the severity of the incident based on the descriptions of the caller: he/she is usually upset and may give an inadequate description of the status of the patient. If an RRA responds to the incident and the patient needs transportation, the closest RTA is sent to the emergency scene as well. If no RTA is available, this call enters another first-come first-served queue with less priority than the one mentioned above. Meanwhile, the RRA paramedic provides care to the patient. This on-scene care can take either longer or shorter than the response time of the RTA. In the first case, the RTA leaves with this patient for the hospital as soon as the on-scene treatment finishes. If the response time of the RTA exceeds the time of the care needed on scene, the RRA waits until the RTA arrives. In either case, the RRA paramedic does not accompany the patient to the hospital but he/she becomes available when the RTA leaves the emergency scene. We assume that a patient is always transported to the closest hospital. Having arrived there, it takes some time for the RTA to drop off the patient. When this task is

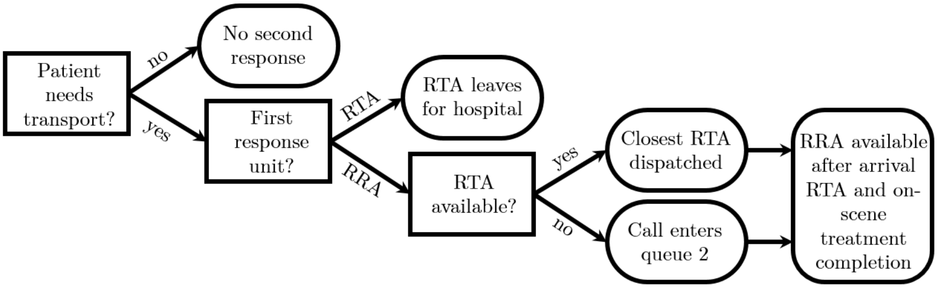


Figure 6.2: Dispatch policy of the second response.

finished, the ambulance becomes idle again. If an RRA responds to a patient not requiring transportation, no subsequent dispatch of a transport unit takes place (see Figure 6.2).

The dispatch process described above is assumed to be fixed. Moments at which the number of available units changes are the dispatch of a response unit (either the first or the second response), the finish of the service of a patient who does not require transportation, the departure time of the transport ambulance from the emergency scene and the service completion of a patient at a hospital. At these events relocation decisions are taken, according to a two-dimensional compliance table. Our goal is to compute a two-dimensional compliance table that minimizes the fraction of calls for which the response time exceeds the time threshold: the fraction of late arrivals.

6.3 Model

In this section we formulate the abovementioned problem as an ILP. First, we introduce the framework and some notation. We define V as the set of locations at which demand for care can occur. Calls arrive according to a Poisson process with rate λ and d_i denotes the fraction of demand occuring at demand node $i \in V$. We denote the set of base stations by W. We assume that both RRAs and RTAs use the same base stations, although this is not a limiting assumption in general. The total number of RRAs and RTAs is denoted by N_R and N_T, respectively. We assume that both the on-scene treatment and the hospital drop-off time are exponentially distributed, with rates μ_1 and μ_2, respectively.

Deterministic driving times are given: $\tau^R(i,j)$ and $\tau^T(i,j)$ denote the driving time between nodes i and j, $i,j \in V \cup W$ of an RRA and an RTA, respectively. As RRAs are faster, we assume $\tau^R(i,j) < \tau^T(i,j)$. The abovementioned driving times are based on the emergency speeds, which are used when an ambulance is carrying out patient-related tasks, e.g., response or transport. An ambulance performing a relocation is not allowed to turn on optical and sound signals, and so these driving times are longer. We denote these relocation driving times by $\tau^2(i,j)$ for $i,j \in V \cup W$ and both vehicle types. The time threshold is denoted by

λ	Call arrival rate.
μ_1	On-scene treatment rate.
μ_2	Hospital transfer rate.
$\tau^R(i,j)$ $(\tau^T(i,j))$	Emergency driving time from i to j for an RRA (RTA), $i, j \in V \cup W$.
$\tau^2(i,j)$	Relocation time between i and j, $i, j \in V \cup W$.
T	Time threshold on the response time.
V	Set of demand nodes.
W	Set of waiting sites.
N_R (N_T)	Total number of RRAs (RTAs).
$\mathcal{S}$	State space.
d_i	Fraction of demand occuring at node $i \in V$.
p_R (p_T)	Busy fraction RRA (RTA).
J_i^R (J_i^T)	Subset of base stations from which an RRA (RTA) can respond to node $i \in V$ within time threshold T.
K_s^R (K_s^T)	Number of available RRAs (RTAs) in state $s \in \mathcal{S}$.

TABLE 6.2: Notation.

T. We define J_i^R as the subset of base stations from which an RRA can respond to an incident at node $i \in V$ within the time threshold, according to τ^R:

$$J_i^R = \{j \in W : \tau^R(j,i) \leq T\}.$$

The RTA counterpart J_i^T is defined similarly. Note that $J_i^T \subseteq J_i^R \subseteq W$ due to the fact that RRAs are faster than RTAs.

We denote the *busy fractions* of RRAs and RTAs by p_R and p_T. These fractions correspond to the probability that a specific unit is unavailable due to the service of a patient, at an arbitrary moment in time. Note that these fractions heavily rely on λ, μ_1 and μ_2, but also on the response time and the transportation time of a patient to a hospital. The state of our system is described by the number of available vehicles of both types. We denote the state space by $\mathcal{S}$ and a state $s \in \mathcal{S}$ is given by $s = (s_R, s_T)$ with $0 \leq s_R \leq N_R$ and $0 \leq s_T \leq N_T$. In the remainder, we denote the number of available RRAs and RTAs in state s by K_s^R and K_s^T, respectively. For each state, except the state $(0,0)$, a desired configuration of available ambulances is computed in order to produce a two-dimensional compliance table. Table 6.2 provides an overview of the introduced notation.

The first step in the formulation of our model is to extend the MEXCLP2 model by McLay (2009) to fit into the compliance table framework. The objective of the MEXCLP2 is to optimally deploy two types of vehicles in a geographic area; optimally in the sense that the expected number of highest urgency calls responded to within T is maximized. That is, it computes the optimal configuration for the state (N_R, N_T). We extend this model to compute these configuration for any state, resulting in a two-dimensional compliance table.

6.3.1 Hypercube Model

An important model used to obtain input parameters for the MEXCLP2 is the Hypercube model proposed by Larson (1974) and its approximation by the same author (Larson, 1975). This model was extended by Jarvis (1985) to include multiple customer types and two types of servers. This extension considers a loss system with distinguishable servers and multiple customer types, each arriving according to a Poisson process with a customer-type dependent arrival rate. Exactly one server is assigned to each customer. If no servers are available, the customer is lost. Moreover, servers are assigned to customers according to a fixed preference assignment rule for that customer type. If all servers of the most preferred type are busy, the customer is assigned to a server of the less preferred type. The assignment is made at the moment of the arrival of the customer. The expected service times for each server-customer pair are known in advance.

The approach taken by McLay (2009) is similar to the one by Jarvis (1985), except for the fact that an infinite queue system is used instead of the loss system. The motivation for this model is that patients generally wait for a medical unit to become available. Moreover, the Hypercube model by Jarvis (1985) assumes that exactly one unit is assigned to each call, which does not hold in the MEXCLP2 model. Therefore, McLay (2009) considers calls existing of multiple customers. In our model, this translates to the arrival of one customer when the emergency call is made (first response) and the arrival of one customer when the RRA informs the emergency control center about the necessity of an RTA (second response). Note that in our model the preference assignment rule is to first assign an RRA and if none of these are available within range, an RTA is dispatched.

An approximation procedure to estimate performance measures for the Hypercube model assuming exponential service times is presented by Jarvis (1985), based on the one given by Larson (1975). This procedure was used by McLay (2009) to estimate busy fractions for the MEXCLP2 model. In our framework, we need the following ingredients for this approximation procedure.

In the remainder, we replace the R of RRA and the T of RTA by $* \in \{R, T\}$ if statements hold for both vehicle types. We denote by P_0^* the steady-state probability that all units of type $*$, $* \in \{R, T\}$, are busy, which corresponds to the fraction of time none of the ambulances of type $*$ is available. This quantity is computed by

$$P_0^* = \left(\frac{N_*^{N_*} p_*^{N_*}}{N_*!(1-p_*)} + \sum_{j=0}^{N_*-1} \frac{N_*^j p_*^j}{j!} \right)^{-1}, \tag{6.1}$$

as in an $M/M/N_*$-queue. Moreover, we define 'correction factors' $Q_*(N_*, p_*, j)$. These factors correct for computing the probability that the $(j+1)^{st}$ selected ambulance of type $*$ is the first available one, assumed that ambulances operate independently, given a total of N_* servers and a busy fraction p_*. Therefore, j in bounded from above by $N_* - 1$. The correction factors are computed by Larson

(1975) via

$$Q_*(N_*, p_*, j) = \sum_{k=j}^{N_*-1} \frac{(N_* - j - 1)!(N_* - k)N_*^k p_*^{k-j}}{(k-j)!N_*!(1-p_*)} P_0^*,$$

where $j = 1, 2, \ldots, N_* - 1$, and with $Q_*(N_*, p_*, 0) = 1$. We define customer type 1 and 2 to correspond to the request for first and second response, respectively. We denote the corresponding arrival rates by λ_1 and λ_2, and the service rates by μ_1^R, μ_1^T and μ_2^T. Note that μ_2^R is not defined since a customer of type 2 is solely served by an RTA. Recall that all type 1 customers prefer to be served by an RRA. We denote the fraction of type 1 customers responded to by an RRA by f. We can compute this quantity by

$$f = \sum_{j=0}^{N_R-1} Q_R(N_R, p_R, j)(1-p_R)p_R^j.$$

An update on the busy fractions p_R and p_T can now be computed by

$$p_R = \frac{f\lambda_1}{\mu_1 N_R}, \tag{6.2}$$

and

$$p_T = \frac{1}{N_T}\left(\frac{\lambda_2}{\mu_2} + (1-f)\frac{\lambda_1}{\mu_1}\right). \tag{6.3}$$

The procedure used to estimate busy fractions is to initialize

$$p_R = \frac{\lambda_1}{\mu_1 N_R} \quad \text{and} \quad p_T = \frac{\lambda_2}{\mu_2 N_T}$$

and then to iteratively compute Equations (6.1)-(6.3) until a certain stopping criterion is met, e.g., when the differences in busy fractions between subsequent iterations have become small enough. This procedure is similar to the ones by Jarvis (1985) and McLay (2009). The one by the latter seems more comprehensive as multiple call priorities are taken into account. However, we have chosen not to do so because this complicates our simulation. In addition, the MEXCLP2 model itself focuses on the response to a single type of call, like our model.

Note that the approximations of the busy fractions computed by the above procedure are rough estimates on the true values. This has several causes. First, the Hypercube model assumes that servers operate independently. However, this is not the case as an RRA periodically summons an RTA. Therefore, the call arrival process for RTAs depends on that for RRAs. The reason that we make this assumption is for tractability reasons. Moreover, it does not capture the actual locations of the ambulances. As a consequence, the Hypercube model assumes that an RRA is dispatched to each customer of type 1, regardless of the location of the incident. However, if the ambulance configuration is such that no RRA is present within range while an RTA is, this is not the case. Therefore, the Hypercube model overestimates p_R, while p_T is underestimated especially if the number of RRAs is small compared to the number of RTAs. Besides, the busy fractions depend on the response and transportation time as well, since response and transportation is part of the busy time of an ambulance. The mean transportation time can be estimated rather accurately since the location of hospitals and the demand of each node are known, so this can be taken into account in the computation of μ_2. However, it is not possible to estimate the mean response time as we need the locations of the ambulances as well. Therefore, we assume that the response times in the Hypercube model are zero, which underestimates the busy fractions. In addition, the Hypercube model assumes exponentially distributed busy times, which is generally not true in practice. At last, in using the Hypercube model we make the assumption that an RTA arrives at the emergency scene before the on scene treatment time has finished, in case of an RRA response to a patient requiring transportation. In short, the computed approximation of the busy fractions should be viewed with some caution.

6.3.2 MEXCLP2 for Compliance Tables

In this section we explain the ILP used to compute compliance tables for an EMS system with two vehicle types. That is, for each state this ILP computes the desired waiting sites for the available RRAs and RTAs. Although we focus on RRAs and RTAs, this ILP can be applied to any type of vehicle mix with predescribed dispatch process and preference assignment lists.

To define the objective function, we need some additional definitions. We denote the fraction of time the system is in state $s = (s_R, s_T)$ by π_s. These steady-state probabilities can be estimated using the steady-state probabilities of an $M/M/N_*$-queue with a load equal to the busy fraction p_*, $* \in \{R, T\}$, as done by Larson (1975). Let $\pi^*_{s_*}$ denote the steady-state probability that exactly s_* units of type $* \in \{R, T\}$ are available. We know that $\pi^*_{N_*} = P^*_0$, defined in Equation (6.1). We compute

$$\pi^*_{s_*} = \frac{N_*^{N_*-s_*} p_*^{N_*-s_*} \pi^*_{N_*}}{(N_* - s_*)!}, \tag{6.4}$$

for $s_* = 1, 2, \ldots, N_* - 1$, $* \in \{R, T\}$. Moreover,

$$\pi^*_0 = \frac{N_*^{N_*} p_*^{N_*} \pi^*_{N_*}}{(1 - p_*) N_*!}, \tag{6.5}$$

and assuming that RRAs and RTAs operate independently (which we assume for tractability reasons), we compute

$$\pi_s = \pi^R_{s_R} \pi^T_{s_T}.$$

We also define $\pi^R_0\{k_R\}$ to respresent the probability that no RRAs are available in an $M/M/k_R$-queue. This quantity can be estimated by replacing N_R by k_R in Equation (6.1) and Equation (6.5).

Now, we have all ingredients to formulate the ILP model. The ILP is based on the decision variables listed in Table 6.3. The objective of this ILP, as the

$x^*_{s,j}$	Number of units of type $*$ placed at waiting site $j \in W$ in state $s \in \mathcal{S}$, $* \in \{R, T\}$.
y^R_{s,i,k_R}	Equals 1 if in state $s \in \mathcal{S}$, demand point $i \in V$ is covered by at least k_R RRAs, and 0 otherwise.
y^T_{s,i,k_T,k_R}	Equals 1 if in state $s \in \mathcal{S}$, demand point $i \in V$ is covered by at least k_T RTAs and *exactly* k_R RRAs, and 0 otherwise.

TABLE 6.3: Decision variables.

one by McLay (2009), is to maximize the demand covered within time threshold T. A call is covered if either an RRA or an RTA responds timely, but an RRA is preferred. An RTA is only dispatched if none of the RRAs can arrive at the emergency scene within the specified amount of time. The objective function is given by

$$\begin{aligned}\text{Max} \sum_{s \in \mathcal{S}} \sum_{i \in V} \pi_s d_i \Bigg(& \sum_{k_R=1}^{K_s^R} Q(K_s^R, p_R, k_R - 1)(1 - p_R) p_R^{k_R - 1} y^R_{s,i,k_R} + \\ & \sum_{k_T=1}^{K_s^T} \sum_{k_R=0}^{K_s^R} Q(K_s^T, p_T, k_T - 1)(1 - p_T) p_T^{k_T - 1} \pi_0^R\{k_R\} y^T_{s,i,k_T,k_R} \Bigg).\end{aligned} \tag{6.6}$$

Given a state $s \in \mathcal{S}$ and a node $i \in V$, the expected coverage consists of two parts: the first part (the upper line in Equation (6.6)) corresponds to the expected coverage induced by RRAs. This term is similar to the objective function in the AMEXCLP model by Batta et al. (1989). In the second part (the lower line in Equation (6.6)) the expected coverage induced by RTAs is added, weighted by a factor $\pi_0^R\{k_R\}$ corresponding to the approximated probability of having no available RRA within range, assuming that demand node i is covered by exactly k_R RRAs. Both parts are concave in k_R and k_T, respectively, for each state $s \in \mathcal{S}$ and each demand node $i \in V$. This is due to the same reason as the objective function of the MEXCLP model is concave, and implies that both sequences $(y^R_{s,i,k_R})_{k_R=1}^{K_s^R}$ and $(y^T_{s,i,k_R,k_T})_{k_T=1}^{K_s^T}$ are non-increasing in an optimal solution.

As in the original MEXCLP and MEXCLP2 model of Daskin (1983) and McLay (2009), respectively, we need to limit the number of units to be placed. In state s, we are allowed to locate no more than K_s^* vehicles of type $*$:

$$\sum_{j \in W} x^*_{s,j} \leq K_s^* \qquad s \in \mathcal{S},\ * \in \{R, T\}. \tag{6.7}$$

In addition, we need constraints that link the x- and y-variables. For RRAs, these constraints are given by

$$\sum_{k_r=1}^{K_s^R} y^R_{s,i,k_R} \leq \sum_{j \in J_i^R} x^R_{s,j} \qquad s \in \mathcal{S},\ i \in V. \tag{6.8}$$

These constraints enforce that a demand point $i \in V$ is only covered by at least k_R vehicles if the base stations within range of i contain at least k_R vehicles together. Connecting the x^T- and y^T-variables is harder as indices belonging to the number of RRAs are involved as well in y^T_{s,i,k_T,k_R}. To ensure the above condition for RTAs, we include the constraint

$$\sum_{k_T=1}^{K_s^T} \sum_{k_R=0}^{K_s^R} y^T_{s,i,k_T,k_R} \leq \sum_{j \in J_i^T} x^T_{s,j} \qquad s \in \mathcal{S},\ i \in V \qquad (6.9)$$

in our model. Note that if for $s \in \mathcal{S}$, $i \in V$, $k_T = 1, \ldots, K_s^T$ and $k_R = 0, \ldots, K_s^R$ it holds that $y^T_{s,i,k_T,k_R} = 1$, then $y^T_{s,i,k_T,k'_R} = 0$ for $k'_R \neq k_R$, which makes constraint (6.9) similar to constraint (6.8). To link the y^R_{s,i,k_R} and y^T_{s,i,k_R,k_T} we introduce variables z_{s,i,k_T} similar to McLay (2009), as follows:

$$z_{s,i,k_T} = \begin{cases} 0 & \text{if } y^R_{s,i,k_T,k_R} = 0,\ s \in \mathcal{S},\ i \in V,\ k_T = 1, \ldots, K_s^T,\ k_R = 1, \ldots, K_s^R, \\ 1 & \text{otherwise.} \end{cases}$$

Moreover, the following constraints are introduced:

$$\sum_{k_R=1}^{K_s^R} (k_R y^T_{s,i,k_T,k_R}) + K_s^R z_{s,i,k_T} \geq \sum_{j \in J_i^R} x^R_{s,j} \qquad s \in \mathcal{S}, i \in V, k_T = 1, \ldots, K_s^T \qquad (6.10)$$

$$\sum_{k_R=0}^{K_s^R} (y^T_{s,i,k_T,k_R}) + z_{s,i,k_T} \leq 1 \qquad s \in \mathcal{S}, i \in V, k_T = 1, \ldots, K_s^T. \qquad (6.11)$$

If demand node i is covered by exactly k_R RRAs and at least k_T RTAs in state $s \in \mathcal{S}$, then constraint (6.11) forces z_{s,i,k_T} to be 0, $i \in V$, $k_T = 1, \ldots, K_s^T$. In addition, constraint (6.10), which will be satisfied as equality if $z_{s,i,k_T} = 0$, has a similar interpretation as constraints (6.8) and (6.9). However, if $\sum_{k_R=0}^{K_s^R} y^T_{s,i,k_T,k_R} = 0$, it can still be the case that demand node i is covered by exactly k_R RRAs in state s, but not by at least k_T RTAs. In order to maintain proper linking, z_{s,i,k_T} must be 1, which is assured by constraint (6.10).

Now, the ILP is given by the objective function of Equation (6.6) subject to constraints (6.7)–(6.11), and the following integer and binary constraints:

$$x^*_{s,j} \in \{0, 1, \ldots, K_s^*\} \qquad s \in \mathcal{S},\ j \in W \qquad (6.12)$$

$$y^R_{s,i,k_R} \in \{0, 1\} \qquad s \in \mathcal{S},\ i \in V,\ k_R = 1, \ldots, K_s^R \qquad (6.13)$$

$$y^T_{s,i,k_T,k_R} \in \{0, 1\} \qquad s \in \mathcal{S},\ i \in V,\ k_T = 1, \ldots, K_s^T,\ k_R = 0, 1, \ldots, K_s^R \qquad (6.14)$$

$$z_{s,i,k_T} \in \{0, 1\} \qquad s \in \mathcal{S},\ i \in V,\ k_T = 1, \ldots, K_s^T. \qquad (6.15)$$

Note that there is no cohesion between the configurations in different states. That is, if the steady-state probabilities π_s were to be removed from the objective function, the same solution would be computed. In the next two subsections, we will incorporate dependence between desired configurations in different states.

6.3.3 Nestedness

A first way to incorporate cohesion between different compliance table levels is the introduction of nestedness constraints as done in the MEXPREP presented in Chapter 5. These constraints limit the number of units instructed to relocate if a state transition occurs. Recall that in a nested compliance table, the set of desired locations of a lower state is a subset of each higher state, where lower and higher correspond to the number of units available and a station at which multiple units are positioned counts as multiple elements. By using nested compliance tables, at most one ambulance is instructed to move at each decision moment, which avoids unnecessary movement of other ambulances, as stated by Sudtachat et al. (2016).

As we consider two-dimensional compliance tables, we can have nestedness in both the RRA- and RTA-direction. In addition to the above described condition for a compliance table to be nested, we require that the desired configurations are the same if the number of available units does not change. For instance, the two-dimensional compliance table displayed in Table 6.1 is nested in the RRA-direction: the configuration belonging to each state with one available RRA (base station 1) is a subset of each state with two available RRAs (base stations 1 and 2). As a consequence, if in a state with two RRAs available the one from station 1 is dispatched, the other RRA travels from station 2 to 1. If the one from 2 is dispatched, no relocation is necessary. Moreover, if an RTA is dispatched, no relocation of an RRA is required.

However, in the RTA-direction the two-dimensional compliance table of Table 6.1 is *not* nested as the set of desired locations for the RTAs in state $(0,1)$ is not a subset of the one of state $(0,2)$. Moreover, the RTA-configurations of states $(1,2)$ and state $(0,2)$ do not coincide. We define

$$\mathcal{S}_0^* = \{s' \in \mathcal{S} : K_{s'}^* = 0\}$$

as the subset of states without an available unit of type $* \in \{R,T\}$. Moreover, we define

$$\mathcal{S}_s^R = \{s' \in \mathcal{S} : K_{s'}^R = K_s^R - 1,\ K_{s'}^T = K_s^T\}$$

as the subset with one RRA fewer available and the same number of RTAs available, $s \in \mathcal{S} \backslash \mathcal{S}_0^R$. The set $\mathcal{S}_s^T$ is defined similar. Note that both sets contain precisely one element. Similar to the a-variables of the previous chapter, we define $a_{s,s',j}^*$ as the number of units that is added to base station $j \in W$ if a transition from state $s \in \mathcal{S} \backslash \mathcal{S}_0^*$ to state $s' \in \mathcal{S}_s^R \cup \mathcal{S}_s^T$ occurs, i.e., at the dispatch of either an RRA or an RTA. It is this number that we want to restrict. We do this by defining $\alpha_{s,s'}^*$ as the bound on base station changes for a vehicle of type $*$ if an state transition from s to s' takes place. We introduce the constraints

$$x_{s',j}^* - x_{s,j}^* \leq a_{s,s',j}^* \qquad s \in \mathcal{S} \backslash \mathcal{S}_0^*,\ s' \in \mathcal{S}_s^R \cup \mathcal{S}_s^T,\ j \in W,\ * \in \{R,T\}, \tag{6.16}$$

$$\sum_{j \in W} a_{s,s',j}^* \leq \alpha_{s,s'}^* \qquad s \in \mathcal{S} \backslash \mathcal{S}_0^*,\ s' \in \mathcal{S}_s^R \cup \mathcal{S}_s^T,\ j \in W,\ * \in \{R,T\}, \tag{6.17}$$

which serve the same purposes as constraints (5.21) and (5.22) in the MEXPREP model, respectively: constraint (6.16) ensures that $a_{s,s',j}^*$ takes a non-negative value if more ambulances of type $*$ are located at base station $j \in W$ in state $s \in \mathcal{S}\backslash\mathcal{S}_0^*$ compared to state $s' \in \mathcal{S}_s^R \cup \mathcal{S}_s^T$. Note that if this number is nonnegative, the compliance table in this direction is not nested: in a state with fewer available units, a certain base station contains more ambulances than in the higher state. This implies that at least one ambulance needs to relocate.

In constraint (6.17) we bound the number of these base station changes. Note that if we set $\alpha_{s,s'}^* \equiv 0$ for each (s,s')-pair with $s \in \mathcal{S}\backslash\mathcal{S}_0^*$, $s' \in \mathcal{S}_s^R \cup \mathcal{S}_s^T$, and $* \in \{R,T\}$, a nested compliance table in both directions is obtained. The other extreme value is $\alpha_{s,s'}^* \equiv K_s^*$. If this value is implemented, no nestedness restrictions are present. At last, we include the integer constraints

$$a_{s,s',j}^* \in \{0,1,\dots,K_s^*\} \quad s \in \mathcal{S}\backslash\mathcal{S}_0^*,\ s' \in \mathcal{S}_s^R \cup \mathcal{S}_s^T,\ j \in W,\ * \in \{R,T\} \quad (6.18)$$

in our ILP formulation.

6.3.4 Bounds on Relocation Times

In practice, it may take a while before the desired configuration according to the two-dimensional compliance table is attained, since the new destinations of relocated ambulances may not be close to their origins. For the preparedness of the EMS system this may be disadvantegeous. After all, the model assumes that each ambulance is at its new location just after the state transition and it bases its decision on that assumption. However, in practice this is far from reality. Possibly, there may be much to be gained if relocation times are kept short. In addition, from a crew-perspective this is also desirable as ambulance personnel does not have to spend that much time on the road.

We extend the ILP formulation of Section 6.3.2 to take into account bounds on relocation times. Therefore, we introduce binary variables $v_{s,j}^*$, $s \in \mathcal{S}$, $j \in W$, $* \in \{R,T\}$:

$$v_{s,j}^* \in \{0,1\},\ s \in \mathcal{S},\ j \in W.$$

A variable $v_{s,j}^*$ equals 1 if base station j is occupied by at least one ambulance of type $*$ in state s, and 0 otherwise. This can easily be ensured by incorporation of the following two constraints:

$$v_{s,j}^* \leq x_{s,j}^* \qquad s \in \mathcal{S},\ j \in W,\ * \in \{R,T\}, \quad (6.19)$$

$$x_{s,j}^* - K_s^* v_{s,j} \leq 0 \qquad s \in \mathcal{S},\ j \in W,\ * \in \{R,T\}. \quad (6.20)$$

These constraints enforce that $v_{s,j}^* = 1$ if and only if $x_{s,j}^* > 0$. A relocation between base stations j and j' if a state transition from $s \in \mathcal{S}\backslash\mathcal{S}_0^*$ to state $s' \in \mathcal{S}_s^R \cup \mathcal{S}_s^T$ occurs can be prevented by forbidding that both $v_{s,j}^*$ and $v_{s',j'}^*$ equal 1 in a solution. Let $M_{s,s'}^*$ be a bound on the time any relocation may take if a transition from state s to state s' occurs. To model this restriction in our ILP, we include the constraint

$$v_{s,j}^* + v_{s',j'}^* \leq 1 \qquad s \in \mathcal{S}\backslash\mathcal{S}_0^*,\ s' \in \mathcal{S}_s^R \cup \mathcal{S}_s^T,\ j,j' \in W,\ * \in \{R,T\}, \quad (6.21)$$

for the base station pairs (j,j') for which it holds that $\tau^2(j,j') > M_{s,s'}^*$. Note that this constraint also bounds the relocation time of idle ambulances if a state transition in the other direction occurs, i.e., when an ambulance becomes available. This bound is only imposed on idle ambulances and not on a unit that just finished service. After all, it is very uncertain where this vehicle becomes available. Therefore, it might still happen that this unit performs an overly long relocation.

The ILP formulation to compute a two-dimensional compliance table with nestedness constraints and bounds on the relocation time is now given by objective function (6.6), subject to constraints (6.7)-(6.21).

6.4 Computational Study

In this section, we compute two-dimensional compliance tables for the EMS region of Flevoland (see Section 3.4.1), in which units can only idle at the six actual base stations (the red dots in Figure 3.3). Some outskirts of this region can not be reached by an RTA, departing from a base station, within the time threshold. However, RRAs can reach these areas timely, as they are faster than RTAs. In addition to the computation of the compliance tables, we generate results by a discrete-event simulation of the obtained two-dimensional compliance tables based on the description of the process described in Section 6.2.

For each postal code-pair deterministic emergency driving times τ^T for RTAs are estimated by and provided by the RIVM (Kommer and Zwakhals, 2008). We also need emergency driving times of RRAs (τ^R) and relocation times for both types of vehicles (τ^2). We compute these by division resp. multiplication of the driving times τ^T by a factor $\frac{10}{9}$. To keep track of the actual location of units in our simulation, we use the travel routes as computed in Section 3.4.3.

In our study, we consider three different fleet mixes. We assume that always ten units are on duty in total and we base our computations on fleet mixes $(N_R, N_T) = (2, 8)$, $(5, 5)$ and $(8, 2)$. The number of ambulances, as well as the vehicle mix, is kept constant throughout the day: we do not model ambulance shifts. This results in 26, 35, and 26 states, respectively. Note that the 'state' $(0, 0)$ is not classified as such as no computation of an ambulance configuration is required for $(0, 0)$. The response time threshold T is 12 minutes, although the statutory threshold time is 15 minutes in the Netherlands. However, we do not take into account answering the emergency call and pre-trip delay, which together last for 3 minutes on average.

6.4.1 Application of the Hypercube Model

To apply both the Hypercube model as described in Section 6.3.1 and the ILP of Sections 6.3.2–6.3.4, we need to estimate the input parameters regarding the demand probabilities, the arrival and service rates, and the hospital probabilities. To this end, GGD Flevoland provided us historical data on emergency requests occurred in the year 2011. This data includes the time and location of occurrence, as well as the on-scene treatment time and hospital drop-off time. We focused on the time interval 7AM to 6PM, which are the hours with the highest arrival intensity.

(N_R, N_T)	(2,8)	(5,5)	(8,2)
p_R	0.4123	0.2158	0.1356
p_T	0.2005	0.2699	0.6719

TABLE 6.4: Busy fractions estimated by the Hypercube model for different fleet mixes (N_R, N_T), with ten vehicles.

In the year 2011, 7,632 emergency requests were reported in the considered time interval, which corresponds to an hourly arrival rate of 1.97 incidents. This corresponds to $\lambda = 0.0328$ incidents per minute. Note that in order to apply the described Hypercube model, we need to distinguish two different arrival rates: $\lambda_1 = \lambda$ corresponds to the request for an ambulance for first response, and λ_2 is the arrival rate of the request for an RTA by an RRA. This quantity is computed by multiplication of the probability that a patient needs transportation to a hospital and λ. Around 87% of the patients require transportation in our data set, so $\lambda_2 = 0.0286$. The demand probabilities d_i, $i \in V = \{1, \ldots, 93\}$ are easily estimated by division of the number of occurred incidents in node i by the total number of incidents.

The estimation of the quantities μ_1^R, μ_1^T, and μ_2^T requires more work. These factors correspond to the on-scene treatment rate of an RRA and RTA, and to the hospital transfer rate, obviously by an RTA, respectively. However, we have no information on μ_1^R in our data set, as this system was not implemented in the year 2011. Therefore, we assume $\mu_1^R = \mu_1^T$, i.e., the on-scene treatment is independent of the type of first response unit. We compute a mean on-scene treatment time of 17.7 minutes, which corresponds to $\mu_1^R = \mu_1^T = 0.0567$.

To obtain accurate estimates of the busy time of an RTA transporting a patient, we also consider the expected transportation time, in addition to the actual drop-off time at the hospital. This expected transportation time is computed, under the assumption that each patient is transported to the closest hospital, as follows: for each postal code i the travel time to the closest hospital is considered, based on the emergency travel times provided. Then, we weight this time by d_i for postal code i, and add the results to obtain an estimate on the mean transportation time. This results in an average transportation time of 8.55 minutes. Based on the historical data, we estimate an actual mean drop-off time of 16.5 minutes. Hence, $\mu_2^T = 0.0400$.

Now, the Hypercube model can be applied in order to estimate the busy fraction p_R and p_T, and consequentely, all factors that depend on these: the correction factors and steady-state probabilities. Busy fractions generated by the procedure explained in Section 6.3.1 for the three fleet mixes of consideration are listed in Table 6.4.

6.4.2 Two-dimensional Compliance Tables

In this section, we solve the ILP given by objective function (6.6) and subject to constraints (6.7)-(6.21). Based on τ^R and τ^T, the sets J_i^R and J_i^T can be computed for demand node $i \in V$. These are the subsets of base stations from which an RRA and RTA can respond to node i within 12 minutes, respectively. Without loss of generality, we can further aggregate the demand nodes in the region, as follows: if for two demand nodes u and v it holds that $J_u^R = J_v^R$ and $J_u^T = J_v^T$, then we replace these nodes by a new node w with $d_w = d_u + d_v$. This results in 20 demand nodes in our region, which we again will denote by V for the sake of simplicity. This reduces the number of variables in the ILP. For each fleet mix, we consider four regimes related to nestedness. We refer to these by R1–R4.

R1. $\alpha^*_{s,s'} \equiv 0$ for each $s \in \mathcal{S}\backslash\mathcal{S}_0^*$, $s' \in \mathcal{S}_s^R \cup \mathcal{S}_s^T$, $* \in \{R, T\}$.

R2. $\alpha^R_{s,s'} \equiv K_s^R$ for each $s \in \mathcal{S}\backslash\mathcal{S}_0^R$, $s' \in \mathcal{S}_s^R \cup \mathcal{S}_s^T$.
$\alpha^T_{s,s'} \equiv 0$ for each $s \in \mathcal{S}\backslash\mathcal{S}_0^T$, $s' \in \mathcal{S}_s^R \cup \mathcal{S}_s^T$.

R3. $\alpha^R_{s,s'} \equiv 0$ for each $s \in \mathcal{S}\backslash\mathcal{S}_0^R$, $s' \in \mathcal{S}_s^R \cup \mathcal{S}_s^T$.
$\alpha^T_{s,s'} \equiv K_s^T$ for each $s \in \mathcal{S}\backslash\mathcal{S}_0^T$, $s' \in \mathcal{S}_s^R \cup \mathcal{S}_s^T$.

R4. $\alpha^*_{s,s'} \equiv K_s^*$ for each $s \in \mathcal{S}\backslash\mathcal{S}_0^*$, $s' \in \mathcal{S}_s^R \cup \mathcal{S}_s^T$, $* \in \{R, T\}$.

Note that R1 forces the compliance table to be nested in both directions, while no nestedness conditions are present in R4. Moreover, we study five different bounds on the relocation time: $M^*_{s,s'} \equiv \frac{1}{2\lambda}, \frac{3}{4\lambda}, \frac{1}{\lambda}, \frac{5}{4\lambda}, \frac{3}{2\lambda}$ for each $s \in \mathcal{S}\backslash\mathcal{S}_0^*$, $s' \in \mathcal{S}_s^R \cup \mathcal{S}_s^T$, $* \in \{R, T\}$. We let the bounds depend on λ because the expected time until the next incident occurs is $\frac{1}{\lambda}$, assuming Poisson arrivals. After all, we aim to be well positioned before the next incident happens. In addition, we study deviations from this bound by 25% and 50% to both sides. Incorporating the bound $\frac{3}{2\lambda}$ is equivalent to the unbounded program, as there is no relocation time between any pair of base stations that exceeds $\frac{3}{2\lambda}$.

We solve the $3 \times 4 \times 5 = 60$ instances of the ILP using CPLEX 12.6 on a 2.2 GHz Intel(R) Core(TM) i7-3632QM laptop with 8 GB of RAM. The optimal solution for each instance was found in approximately 1 second for fleet mixes $(2, 8)$ and $(8, 2)$, and within 10 seconds for fleet mix $(5, 5)$. Note that this last one has substantially more variables due to the larger number of states. However, the computation time is not an issue as compliance tables are usually computed offline.

The objective values for R1 are displayed in Figure 6.3. The values for R2–R4 are within the 1% range, and therefore we do not show them in the figure. Table 6.5 shows the two-dimensional compliance tables for regime R1. We only display these compliance tables, as these are fully nested, and thus they can be represented efficiently. We represent such compliance tables by two one-dimensional vectors of length N_R and length N_T, respectively. The desired ambulance configuration belonging to state s is then given by the first K_s^R entries of the first, and the first K_s^T entries of the second vector. The computed compliance tables are displayed in Table 6.5, based on the enumeration of the base stations of Figure 3.3b. The

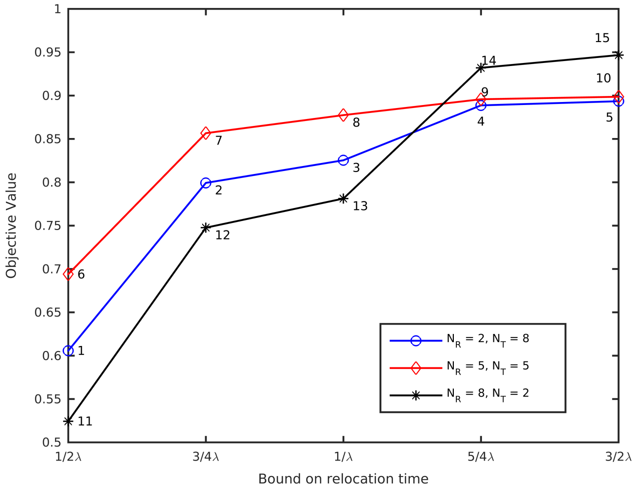


Figure 6.3: Objective function values for R1 as a function of the relocation time bound.

numbers before the compliance tables correspond to the numbers displayed in Figure 6.3.

Figure 6.3 and Table 6.5 lead to several interesting observations. One would expect that fleet mix $(5, 5)$ would have its objective values between those of $(2, 8)$ and $(8, 2)$, but for bounds up to $\frac{1}{\lambda}$ this is not the case. This is probably caused by the following reason. In, for instance, solutions 2, 7 and 12 in Table 6.5 there is a clear division visible in the compliance tables: all vehicles of one specific type are located in the northern part of the region, while all units of the other type are positioned in the south, which is given priority due to the large cities located there. As a consequence, only two units are placed in the north in solutions 2 and 12, while there is overcapacity in the southern part because the relocation time bound does not allow relocations from north to south or vice versa. In solution 7 one also observes a north-south division, but now 5 ambulances are positioned in both parts. Hence, the objective function value for fleet size $(5, 5)$ is higher.

The intersection of the line corresponding to fleet mix $(8, 2)$ with the other two is also an observation that requires discussion. It is closely related to the above explanation. For relocation time bounds up to $\frac{1}{\lambda}$ the northern part is covered very sparsely. However, in solution 14, another partition of the region is induced: the town near base station 3 is isolated from the rest as relocations from base stations 6 to 1 are now allowed, while relocations from 6 to 3 are not. Therefore, a very large part of the region is covered by RRAs. Together with the small busy fraction p_R, this explains the large improvement of the objective function for fleet mix $(8, 2)$ between bounds $\frac{1}{\lambda}$ and $\frac{5}{4\lambda}$.

Solution	Compliance Tables	
	RRAs	RTAs
1	$(2, 2)$	$(1, 1, 1, 1, 1, 1, 1, 1)$
2	$(6, 4)$	$(1, 2, 1, 2, 1, 2, 1, 2)$
3	$(6, 4)$	$(1, 2, 1, 2, 3, 1, 2, 3)$
4	$(3, 4)$	$(1, 2, 6, 1, 4, 2, 6, 1)$
5	$(6, 4)$	$(1, 2, 6, 1, 2, 3, 4, 6)$
6	$(1, 1, 1, 1, 1)$	$(2, 2, 2, 2, 2)$
7	$(2, 5, 4, 2, 5)$	$(1, 1, 3, 1, 1)$
8	$(2, 6, 4, 2, 6)$	$(1, 1, 3, 1, 2)$
9	$(1, 4, 2, 3, 1)$	$(6, 1, 2, 6, 1)$
10	$(1, 4, 2, 6, 1)$	$(1, 2, 6, 3, 1)$
11	$(1, 1, 1, 1, 1, 1, 1, 1)$	$(2, 2)$
12	$(1, 2, 1, 2, 1, 2, 1, 2)$	$(6, 4)$
13	$(1, 2, 1, 3, 2, 1, 3, 2)$	$(6, 4)$
14	$(1, 2, 6, 4, 1, 2, 6, 4)$	$(3, 3)$
15	$(1, 2, 6, 4, 1, 3, 2, 6)$	$(2, 1)$

Table 6.5: Nested compliance tables computed by the ILP, for different relocation time bounds and fleet mixes.

6.4.3 Sensitivity Analysis

This section studies the sensitivity of the computed compliance tables with respect to the estimated inputs. To this end, we consider a variation in treatment rates (μ_1^R, μ_1^T, and μ_2^T), and we multiply the mean treatment times by a factor γ, for different values of γ. Based on these modified treatment rates, we compute new busy fractions p_R and p_T. Then, we compute nested two-dimensional compliance tables under regime R1. We do not impose a bound on the relocation time.

Table 6.6 displays the computed compliance tables and busy fractions for different values of γ. In this table, we observe small changes if treatment rates are larger or smaller. For fleet mixes $(2, 8)$ and $(8, 2)$ the eight RTAs and eight RRAs, respectively, occupy the same base stations if they are all available, for each value of γ. However, there are some minor changes in the order. For instance, for fleet mix $(2, 8)$, base station 2 and 3 are switched between $\gamma = 0.75$ and $\gamma = 1$. As the load of the system increases, it is more important to have an RTA positioned in the city where base station 2 is located. This behavior is also reflected in fleet mix $(8, 2)$: as the load increases, base stations 1 (in the largest city) and 2 are preferred over base station 3. Moreover, base station 1 also appears in the RRA- and RTA-part of the compliance tables for $(2, 8)$ and $(8, 2)$, respectively, if the busy fractions become large enough. The fact that the busier base stations are occupied longer in the states with fewer units is also reflected in the compliance tables for fleet mix $(5, 5)$: both station 3 and 4 move further to the right if γ increases in the RTA- and RRA-part, respectively. Especially base station 1 is an important one,

γ	p_R	p_T	Compliance Tables	
			RRAs	RTAs
0.50	0.2451	0.1193	$(4,6)$	$(1,2,6,1,3,2,4,6)$
0.75	0.3377	0.1576	$(4,6)$	$(1,2,6,1,3,2,4,6)$
1.00	0.4123	0.2005	$(4,6)$	$(1,2,6,1,2,3,4,6)$
1.25	0.4729	0.2469	$(4,1)$	$(1,2,6,1,2,3,4,6)$
1.50	0.5226	0.2960	$(4,1)$	$(1,2,6,1,2,3,4,6)$
0.50	0.1085	0.1804	$(4,1,6,2,1)$	$(2,3,1,6,2)$
0.75	0.1625	0.2248	$(1,4,6,2,1)$	$(2,1,3,6,2)$
1.00	0.2159	0.2698	$(1,4,2,6,1)$	$(1,2,6,3,1)$
1.25	0.2678	0.3164	$(1,4,2,6,1)$	$(1,2,6,3,1)$
1.50	0.3174	0.3652	$(1,2,4,6,1)$	$(1,6,2,1,3)$
0.50	0.0678	0.4509	$(1,2,6,4,3,1,2,6)$	$(6,4)$
0.75	0.1017	0.5614	$(1,2,6,4,3,1,2,6)$	$(6,1)$
1.00	0.1356	0.6719	$(1,2,6,4,1,3,2,6)$	$(2,1)$
1.25	0.1695	0.7824	$(1,2,6,4,1,3,2,6)$	$(2,1)$
1.50	0.2034	0.8930	$(1,2,6,4,1,2,3,6)$	$(1,6)$

TABLE 6.6: Nested compliance tables computed by the ILP for different treatment rates.

as a second occurence replaces station 2 between $\gamma = 0.75$ and $\gamma = 1$. Besides, the first occurence of station 2 shifts to the right in favor of station 1. Station 2 shifts to the left in the RRA-part in order to compensate for this.

We also study the impact of the demand variation throughout the considered time interval (7 AM – 6 PM). To this end, we divide the mentioned interval into eleven time blocks of one hour, and we consider the arrival rate per block. We select the minimum and maximum hourly arrival rate: 0.93 incidents (7 AM – 8 AM) and 2.34 incidents (1 PM – 2 PM), respectively. These correspond to $\lambda = 0.0154$ and $\lambda = 0.0390$ incidents per minute. We compute busy fractions and nested compliance tables based on these values for λ. All the other inputs in the Hypercube model are held constant. No relocation time bound is imposed. Table 6.7 displays the results.

As in the case of larger mean treatment times, we observe that it becomes more important to occupy the base stations located in the largest cities (1 and 2) in states with a few number of units available. This is not surprising since longer treatments and an increased arrival intensity both have the same consequence: larger busy fractions. Another interesting question is whether the proposed ILP for the computation of two-dimensional compliance tables scales to city-sized networks. To this end, we have run a variety of experiments based on the EMS region of Amsterdam and its surroundings. We tested a variety of fleet mixes to assess the computation times. The results show that the computation times are short for small- and medium-sized cities (up to, say, 18-20 ambulances), but tend to become

Interval	p_R	p_T	Compliance Tables	
			RRAs	RTAs
7AM-8AM	0.2327	0.0847	$(4, 6)$	$(1, 2, 6, 3, 1, 2, 4, 6)$
1PM-2PM	0.4389	0.2262	$(4, 1)$	$(1, 2, 6, 1, 2, 3, 4, 6)$
7AM-8AM	0.1021	0.1265	$(1, 4, 6, 2, 1)$	$(2, 3, 1, 6, 4)$
1PM-2PM	0.2382	0.2992	$(1, 4, 2, 6, 1)$	$(1, 2, 6, 3, 1)$
7AM-8AM	0.0638	0.3162	$(1, 2, 4, 6, 3, 1, 2, 4)$	$(6, 1)$
1PM-2PM	0.1500	0.7434	$(1, 2, 6, 4, 1, 3, 2, 6)$	$(2, 1)$

Table 6.7: Nested compliance tables computed by the ILP for different demand arrival rates.

significant for larger cities.

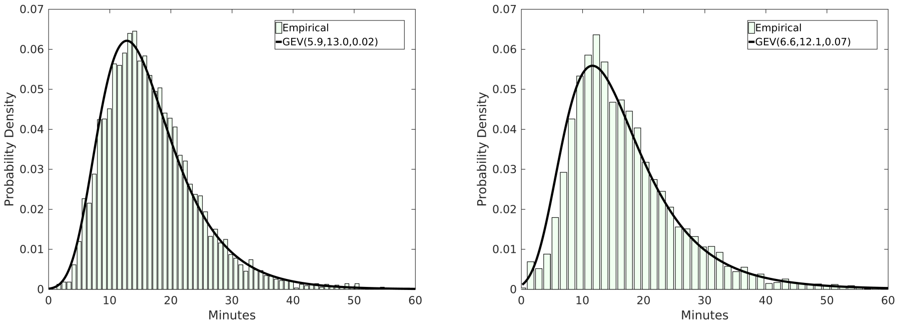
6.4.4 Simulation

To obtain a more realistic estimate of the system performance, we simulate the process described in Section 6.2 according to the parameters estimated in Section 6.4.1 with one exception: by performing a data analysis on the historical data provided, it turned out that the treatment and transfer times are not exponentially distributed, as assumed by the Hypercube model. We fitted several distributions and the *generalized extreme value (GEV) distribution* was the best, according to the Bayesian Information Criterion (Schwarz, 1978). The probability density function of this distribution is given by

$$f(x) = \frac{1}{a}\big(1 - \frac{b}{a}(x-c)\big)^{\frac{1-b}{b}} exp\Big(- 1\big(1 - \frac{b}{a}(x-c)\big)\Big)^{\frac{1}{b}},$$

where $a > 0$ and c are the scale and location parameter, and b is the shape parameter. We refer to Singh (2010) for an extensive description of this probability distribution. See Figure 6.4 for a graphical illustration of GEV(a, c, b). We simulate the on-scene treatment time and hospital transfer time according to this distribution to stay as close to reality as possible. For the same reason, we use the actual postal codes as the demand points, and not the aggregated version. In estimating the on-scene treatment time, we distinguish between patients that need transportation and those who do not, since the on scene treatment time for the last category is substantially longer: 25.7 minutes vs. 16.5 minutes. Note that if one weighs these numbers with the probability that transportation is required, one obtains the mean treatment time of 17.7 minutes mentioned before.

Our simulation length is ten years for each of the 60 compliance tables. That is, we consider the system to be in continuous operation with the fleet size fixed, deterministic driving times τ^R, τ^T and τ^2 and the estimated parameters. This avoids that the system becomes empty over night, and thereby our approach allows us to obtain measurements that are close to ‘steady-state’. We test the performance through simulation on the following performance measures:



(A)

(B)

Figure 6.4: Histogram of the on-scene treatment time if transportation is required (6.4a), and of the hospital transfer time (6.4b), and the fitted probability distribution.

1. Percentage on time: the fraction of requests responded to within $T = 12$ minutes, as well as 95%-confidence intervals.

2. Mean response time of first response unit (in seconds).

3. Number of relocations.

4. Total relocation time (in hours).

Results on these performance indicators are displayed in Tables 6.8, 6.9, and 6.10, and Figure 6.5.

Note that for fleet mix $(8, 2)$ the shape of the simulated performance is similar to the corresponding objective values in Figure 6.3. Surprisingly, this is not the case for the other two fleet mixes as one would expect on basis of the objective function values, underlining the necessity of performing simulations. The maximum for both is attained at $\frac{1}{\lambda}$, which is the expected time until the next incident occurs.

If one compares the fully nested two-dimensional compliance tables of fleet mix $(2, 8)$ for $\frac{1}{\lambda}$ and $\frac{3}{2\lambda}$ (Solutions 3 and 5 in Table 6.5), one observes that in both solutions the RRAs are located in the north of the region, which is a sparsely populated area. The difference in both solutions is that in solution 3, RTAs are positioned only in the south. As a consequence, there are relatively many late arrivals in the north of the region in the simulation. In solution 5, RTAs are located across the whole region, which causes many late arrivals in the south: the city in which base station 1 is located and the town near base station 3, in particular. As the call arrival rate in the south is larger, solution 3 outperforms solution 5. Moreover, the results on number of relocations and total relocation time indicate that the system corresponding to solution 3 is in compliance faster than the one of solution 5, which also has an effect on the patient-based performance. The non-nested cases are explained by a similar reasoning.

		$\frac{1}{2\lambda}$	$\frac{3}{4\lambda}$	$\frac{1}{\lambda}$	$\frac{5}{4\lambda}$	$\frac{3}{2\lambda}$
R1	Percentage on time	71.46%	92.61%	95.72%	95.27%	95.08%
	Lower Bound 95%-CI	71.12%	92.36%	95.50%	95.13%	94.87%
	Upper Bound 95%-CI	71.80%	92.86%	95.94%	95.42%	95.29%
	Mean response time	520 s	323 s	303 s	340 s	307 s
	Number of relocations	0	33,968	43,166	45,336	48,972
	Total relocation time	0 h	10,087 h	13,701 h	19,128 h	19,066 h
R2	Percentage on time	71.54%	92.29%	95.68%	95.41%	94.82%
	Lower Bound 95%-CI	71.26%	92.13%	95.52%	95.21%	94.64%
	Upper Bound 95%-CI	71.83%	92.45%	95.85%	95.61%	95.00%
	Mean response time	519 s	325 s	302 s	340 s	331 s
	Number of relocations	0	34,419	43,473	39,060	58,664
	Total relocation time	0 h	10,169 h	13,791 h	20,892 h	25,765 h
R3	Percentage on time	71.80%	92.86%	95.93%	95.97%	95.84%
	Lower Bound 95%-CI	71.02%	92.43%	95.65%	95.62%	95.56%
	Upper Bound 95%-CI	72.52%	93.29%	96.26%	96.33%	96.38%
	Mean response time	519 s	322 s	302 s	339 s	304 s
	Number of relocations	0	33,823	42,753	52,888	55,430
	Total relocation time	0 h	10,045 h	13,551 h	21,516 h	20,603 h
R4	Percentage on time	71.33%	92.44%	95.83%	95.26%	95.10%
	Lower Bound 95%-CI	71.02%	92.22%	95.67%	95.09%	94.88%
	Upper Bound 95%-CI	71.64%	92.67%	95.99%	95.44%	95.32%
	Mean response time	519 s	324 s	302 s	342 s	327 s
	Number of relocations	0	34,301	42,911	57,152	64,882
	Total relocation time	0 h	10,137 h	13,611 h	23,841 h	27,525 h

Table 6.8: Simulation results for fleet mix $(2, 8)$.

		$\frac{1}{2\lambda}$	$\frac{3}{4\lambda}$	$\frac{1}{\lambda}$	$\frac{5}{4\lambda}$	$\frac{3}{2\lambda}$
R1	Percentage on time	71.88%	93.29%	95.32%	93.11%	92.64%
	Lower Bound 95%-CI	71.38%	93.00%	95.10%	92.84%	92.40%
	Upper Bound 95%-CI	72.38%	93.58%	95.54%	93.40%	92.87%
	Mean response time	503 s	335 s	310 s	316 s	318 s
	Number of relocations	0	27,609	31,788	50,697	57,393
	Total relocation time	0 h	8,018 h	10,123 h	19,791 h	22,928 h
R2	Percentage on time	71.97%	93.13%	95.05%	93.05%	93.02%
	Lower Bound 95%-CI	71.58%	92.81%	94.82%	92.81%	92.77%
	Upper Bound 95%-CI	72.37%	93.44%	95.29%	93.31%	93.27%
	Mean response time	503 s	337 s	312 s	322 s	323 s
	Number of relocations	0	27,440	44,539	58,445	81,327
	Total relocation time	0 h	7,975 h	14,370 h	23,595 h	31,199 h
R3	Percentage on time	72.00%	93.48%	94.99%	93.26%	92.52%
	Lower Bound 95%-CI	71.61%	93.28%	94.82%	93.03%	92.27%
	Upper Bound 95%-CI	72.38%	93.67%	95.16%	93.48%	92.77%
	Mean response time	503 s	335 s	312 s	316 s	318 s
	Number of relocations	0	28,014	39,120	59,692	71,237
	Total relocation time	0 h	8,142 h	12,821 h	24,292 h	30,340 h
R4	Percentage on time	71.77%	93.21%	95.11%	93.05%	92.47%
	Lower Bound 95%-CI	71.40%	93.00%	94.89%	92.75%	92.17%
	Upper Bound 95%-CI	72.13%	93.41%	95.34%	93.36%	92.76%
	Mean response time	503 s	337 s	312 s	319 s	319 s
	Number of relocations	0	30,813	42,897	76,110	74,926
	Total relocation time	0 h	9,034 h	14,220 h	30,160 h	31,681 h

Table 6.9: Simulation results for fleet mix $(5, 5)$.

		$\frac{1}{2\lambda}$	$\frac{3}{4\lambda}$	$\frac{1}{\lambda}$	$\frac{5}{4\lambda}$	$\frac{3}{2\lambda}$
R1	Percentage on time	63.36%	78.60%	81.60%	92.68%	94.60%
	Lower Bound 95%-CI	62.99%	78.24%	81.12%	92.21%	94.26%
	Upper Bound 95%-CI	63.74%	78.96%	82.08%	93.14%	94.95%
	Mean response time	629 s	418 s	414 s	323 s	298 s
	Number of relocations	0	17,844	26,937	28,638	37,603
	Total relocation time	0 h	5,480 h	8,440 h	12,380 h	14,094 h
R2	Percentage on time	63.67%	78.19%	81.74%	93.06%	94.31%
	Lower Bound 95%-CI	63.31%	77.63%	81.28%	92.83%	93.83%
	Upper Bound 95%-CI	64.02%	78.75%	82.18%	93.29%	94.79%
	Mean response time	622 s	431 s	404 s	307 s	308 s
	Number of relocations	0	15,847	26,387	29,281	35,252
	Total relocation time	0 h	4,826 h	8,294 h	12,602 h	13,370 h
R3	Percentage on time	63.32%	78.51%	81.61%	92.60%	94.64%
	Lower Bound 95%-CI	62.92%	77.98%	81.18%	92.08%	94.29%
	Upper Bound 95%-CI	63.72%	79.03%	82.04%	93.12%	95.00%
	Mean response time	630 s	435 s	399 s	319 s	302 s
	Number of relocations	0	17,210	27,103	29,067	52,800
	Total relocation time	0 h	5,284 h	8,507 h	12,504 h	21,133 h
R4	Percentage on time	63.03%	78.44%	81.68%	92.75%	94.69%
	Lower Bound 95%-CI	62.44%	78.02%	81.32%	92.36%	94.31%
	Upper Bound 95%-CI	63.61%	78.86%	82.04%	93.13%	95.07%
	Mean response time	655 s	421 s	406 s	314 s	302 s
	Number of relocations	0	17,089	27,204	29,348	52,282
	Total relocation time	0 h	5,246 h	8,554 h	12,619 h	20,889 h

TABLE 6.10: Simulation results for fleet mix $(8, 2)$.

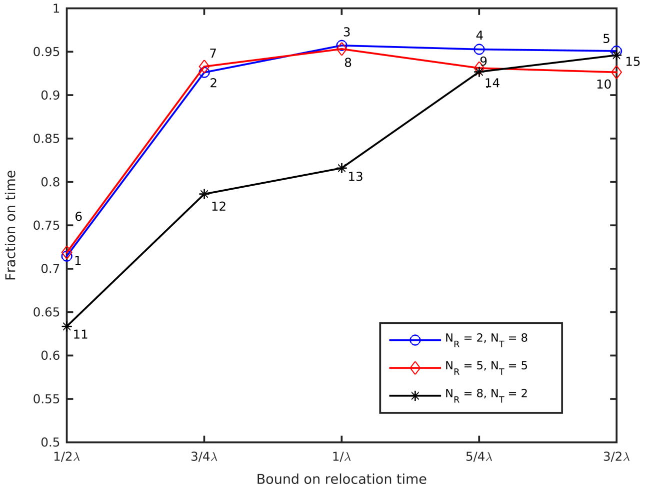


Figure 6.5: Simulated fractions on time for R1 as a function of the relocation time bound.

Whereas the gap in the percentage on time performance indicator between $\frac{1}{\lambda}$ and $\frac{3}{2\lambda}$ for fleet mix $(2,8)$ is relatively small (within 1 percent point), it is much larger for fleet mix $(5,5)$. Even compliance tables with a bound of $\frac{3}{4\lambda}$ outperform the unrestricted version (solutions 7 and 10 in Table 6.5), as observed in Figure 6.5 and Table 6.9, although no unit is assigned to the strategic base station 6 at all. Simulation of the compliance table of solution 7 results in a huge number of late arrivals in the far north and northeast as no unit is able to respond to some postal codes timely if base station 6 is not occupied. However, this reduction is offset by the performance improvement in the rest of the region due to the reduction in time before the system is in compliance again, compared to solution 10, as indicated by the crew-based performance indicators. The performance gap in the simulated on-time percentage between solutions 7 and 8 is explained by the fact that base station 6 is selected instead of 5, resulting in a large performance improvement due to the abovementioned postal codes that now can be reached within 12 minutes.

The performance of fleet mix $(8,2)$ behaves more as expected compared to the other mixes: it is increasing if the relocation time bound is relaxed, as observed in Figure 6.5. This is due to the decreased ambulance availability: in the compliance table belonging to solution 13, for instance, the RRAs are located in the south as this is the most populous part of the region and hence multiple coverage is necessary here. As a consequence, the RTAs are positioned in the north in order to cover this part of the region as well. Since the arrival rate in the south is much larger than in the north, the RTAs are very often instructed to head to the south for the transportation of a patient there. Hence, they are barely available for first response in the north. Moreover, this influences the availability of the RRAs as they need to wait until an RTA arrives at the emergency scene for transportation, which takes a relatively long time as in the majority of the cases the closest RTA is far away. This is the reason behind the increase in performance between relocation time bound $\frac{5}{4\lambda}$ and $\frac{3}{2\lambda}$: in the compliance table of solution 15, the RTAs are located far more strategically.

Another interesting observation is the strange behavior of the response time as a function of the relocation time bound for fleet mix $(2, 8)$, especially the relatively long mean response time of the bound $\frac{5}{4\lambda}$ compared to $\frac{1}{\lambda}$ and $\frac{3}{2\lambda}$. This phenomenon is explained by the fact that in the fully nested two-dimensional compliance table corresponding to bound $\frac{5}{4\lambda}$ (solution 4 in Table 6.5) no RRA is present at base station 6. As one can observe in Figure 3.3, there are many small villages around base station 6. Therefore, the response time from station 6 to one of these villages is quite long. The fact that in solution 9 the first response unit to an incident occuring in one of these villages is always a, relatively slow, RTA, results in a longer mean response time for this relocation bound. The same explanation holds for the non-nested cases, the compliance tables with bound $\frac{3}{2\lambda}$ in R2 and R4 in particular.

Regarding the nestedness, it is worth noting that fully nested compliance tables (R1) are not significantly performing worse on the patient-based performance indicators than non-nested ones (R2, R3 and R4). However, the gaps between the fully nested and fully non-nested regimes in number of relocations and total relocation time are large if one compares these quantities in Tables 6.8, 6.9 and 6.10, especially for the larger relocation time bounds.

6.4.5 Exponentially Distributed Treatment Times

We end this section with a study on the impact of the assumption of exponentially distributed treatment times instead of using the GEV distributions displayed in Figure 6.4. For that purpose, we simulate the nested compliance tables with a relocation time bound of $1/\lambda$ (solutions 3, 8, and 13 in Table 6.5). Only the treatment times are changed with respect to the simulations in Section 6.4.4; the time and place of demand requests are maintained, as well as whether transportation is required. We consider four settings: (1) both the on-scene treatment time and the hospital transfer time are exponentially distributed, (2) only the on-scene treatment time follows an exponential distribution, (3) only the hospital transfer time is exponentially distributed, and (4) both follow the GEV distribution as in Section 6.4.4. The used exponential distributions have the same means as their GEV distributed counterparts, but a larger variance. Results on the percentage on time criterion are listed in Table 6.11.

The results consistently show that especially the use of the exponential distribution instead of the GEV distribution for the on-scene treatment time results in a performance decrease, albeit a small one. This behavior is explained as follows: due to the relative large variance of the exponential distribution, there are many short treatment times, but also many long ones. The short treatment times do not influence the performance much as the RRA has to wait for an arriving RTA anyway (if transportation is required). However, if the on-scene treatment time takes long, the unit availability decreases as both the RRA and the RTA are busy for a

		(EXP,EXP)	(EXP,GEV)	(GEV,EXP)	(GEV,GEV)
$(2,8)$	Percentage on time	95.55%	95.58%	95.74%	95.72%
	Lower Bound 95%-CI	95.37%	95.36%	95.53%	95.50%
	Upper Bound 95%-CI	95.73%	95.80%	95.95%	95.94%
$(5,5)$	Percentage on time	95.13%	95.04%	95.31%	95.32%
	Lower Bound 95%-CI	94.68%	94.82%	95.08%	95.10%
	Upper Bound 95%-CI	95.39%	95.27%	95.55%	95.54%
$(8,2)$	Percentage on time	81.15%	81.23%	81.69%	81.60%
	Lower Bound 95%-CI	80.69%	80.79%	81.27%	81.12%
	Upper Bound 95%-CI	81.61%	81.68%	82.12%	82.08%

Table 6.11: Performance for different distributions of treatment times.

long time. Hence, the performance decreases if a distribution with large variance (e.g., the exponential distribution) is used for the on-scene treatment time. This phenomenon does not occur if a distribution with large variance is used for the hospital transfer time, since only RTAs are involved in the drop-off process.

6.5 Concluding Remarks

In this chapter, we studied an EMS system with two types of medical response units: RRAs and RTAs, and we proposed a mathematical model for the computation of compliance tables in such a system. To this end, we extended the MECRP model by Gendreau et al. (2006) and the MEXCLP2 model by McLay (2009), and formulated our problem as an ILP. To estimate the input parameters needed in this ILP, we used the Hypercube model and iterative procedure described in McLay (2009), which are closely related to the work done by Jarvis (1985). We forced cohesion between the desired configurations in the two-dimensional compliance tables in two ways: we included nestedness constraints and we set bounds on the time a relocation may take. The resulting ILP was applied to the EMS region of Flevoland, for different nestedness regimes, relocation time bounds and fleet mixes. We simulated the obtained two-dimensional compliance tables in a discrete-event simulation to obtain practically relevant results and insights.

Including the two mentioned types of constraints in the model yields some interesting results, most notable the performance improvement if one imposes bounds on the time a relocation may take for fleet mixes with several RRAs. Based on the corresponding objective values, this was not expected. The relocation time bound $\frac{1}{\lambda}$ plays here an important role, because imposing this bound induces the best patient-based performance for the mentioned fleet mixes. Hence, it seems that relating the relocation time bound to the call arrival rate is a good idea. After all, one aims to be in compliance before the next incident occurs, which is expected to happen in $\frac{1}{\lambda}$ time, assuming Poisson arrivals.

In addition, nestedness constraints are a valuable contribution to the two-dimensional compliance table model as well. Simulation shows that no significant performance gain is obtained on the patient-based performance measures if these constraints are dropped. However, the number of relocations and total relocation time are greatly reduced if this type of constraints is included. This reduction on the crew-based performance measures is beneficial for both ambulance crews and managers, as the same patient-based performance can be realized with less driving, and hence, less costs.

7

A Unified View on the Online and Offline Approach

In the previous chapters, we presented several online and offline approaches to solve the ambulance relocation problem. We did numerous experiments regarding different characteristics to analyze the proposed methods. This short chapter is concerned with the presentation of a unified view on the online and offline approach. To that end, we perform several experiments to compare representants of both the online and offline methods proposed in this thesis. That is, we simulate an online method and an offline method on the same setting and the same traces, for both EMS regions described in Section 3.4.1 and Section 3.4.2. We test the chosen representants on four different penalty functions considered throughout this thesis.

7.1 Introduction

In this thesis we made a separation between the online and offline approach to solving the ambulance relocation problem, in which we may proactively relocate ambulances to ensure a fast response to each emergency request. The main difference between both approaches is the way in which the computational work is done. Online methods, usually heuristics, base their decisions on a detailed state description of the system. Due to the large number of states, it is impossible to compute a relocation decision for each state beforehand. Therefore, the computations are done in an online fashion: a relocation decision, based on the current state of the EMS system, is computed from scratch when a decision moment occurs. In contrast, in the offline approach most computational work is done in advance. As the system is usually described by a less detailed state description, the state space is much smaller (although sophisticated techniques like ADP can handle large state spaces in an offline setting as well (Maxwell et al., 2010; Schmid, 2012). This allows for the computation of ambulance location plans for each state a priori. When a decision moment occurs, the corresponding location plan is applied, possibly preceeded or proceeded by a fast computation concerning the actual relocation decision.

In this chapter, we perform such a comparison between online and offline policies. This chapter shows similarities with all previous chapters in the sense that we use methods, insights and setups considered before. We use the insight of Chapter 3 that the restriction of two on the number of ambulances used in a chain relocation is a good choice. Moreover, Chapter 4 and Chapter 5 present the representant of the online and offline approach, respectively. Additionally, the determination of the busy fraction is done in a similar way as in Chapter 6. We also simulate the system according to insights obtained in this chapter. That is, we estimate the on-scene treatment time and the hospital time by means of a GEV distribution (see Section 6.4.4). However, unlike Chapter 6, we consider one type of unit. In the next section, we explain the chosen representants in more detail.

7.2 Methods

To perform the comparison between online and offline approaches, we have selected two relocation methods developed in this thesis that we regard as highly promising in their respective fields, underlined by insights obtained in the research that led to the previous chapters. These two methods serve as representant for their approach. Note that we cannot speak about 'the' methods as we can make many choices regarding the actual implementation of the representants, for instance, choices on the allowance of chain relocations, on the relocation time (see Section 4.4.6) or relocation thresholds (Section 5.4.3), or on the objective criterion (penalty function). In that sense, the representants cover actually a range of several methods. We continue with the explanation of the selected representants. Moreover, we describe which implementations we have picked for both the online and offline representant in our comparison. The representants and implementations are chosen in such a way that the expected results in both patient and crew based performance are similar, and hence, comparable in a fair way.

7.2.1 Online Representant

We choose the amalgamation of the DMEXCLP method and the penalty heuristic as representant for the online approach (see Chapter 4). This method is flexible in the sense that different penalty functions can be implemented easily. Moreover, simulation of this representant shows good results, as indicated by Table 4.10. We have chosen the following implementation:

- Decisions are made at each change in unit availability, i.e., just after the dispatch of an ambulance and when an ambulance becomes available.
- We do not take into account ambulances currently involved in the drop-off of a patient at a hospital. After all, the way in which we modelled this aspect did not lead to better performance, even not in a regime in which we have perfect information concerning the transfer times (see Table 4.4).

- Chain relocations may be carried out. However, the maximum number of vehicles participating in an ambulance motion is two, following the insight obtained in Chapter 3.
- We do not impose relocation time bounds: each ambulance can be relocated to each base station, without being restricted by maximum trip lengths.
- We implement four different objective criteria: maximization of coverage, minimization of the average response time, minimization of the penalty induced by the compromise between coverage and response time, and maximization of the expected number of survivors. The corresponding penalty function are described by

$$\Phi_1(t) = \mathbb{1}_{\{t>T\}}, \tag{7.1}$$

$$\Phi_2(t) = t, \tag{7.2}$$

$$\Phi_3(t) = \begin{cases} \frac{1}{\beta(1+e^{-\alpha(t-T)})} & 0 \leq t \leq T, \\ \frac{\beta-1}{\beta} + \frac{1}{\beta(1+e^{-\alpha(t-T)})} & t > T, \end{cases} \tag{7.3}$$

$$\Phi_4(t) = 1 - (1 + e^{0.679+0.0044t})^{-1}, \tag{7.4}$$

where we set $\alpha = 0.008$, $\beta = 5$, and $T = 720$. Note that $\Phi_4(t)$ represents the survival function of De Maio et al. (2003), in a mortality setting. Functions $\Phi_3(t)$ and $\Phi_4(t)$ are displayed in Figure 3.5 and Figure 5.1, respectively.

7.2.2 Offline Representant

For the offline representant we choose the compliance tables as computed by the AMEXPREP discussed in Section 5.2.5. Using a compliance table implicitly ensures that decisions are taken at the same moments as described above. We also use the same chain relocation policy. Additionally, we consider nested compliance tables to ensure a similar crew workload with respect to the online representant. After all, using such a regime, only one ambulance motion may be performed in both the online and offline representant. Hence, the patient based performance can be compared fairly in this way.

We adjust the computation of the nested AMEXPREP compliance tables on one point, with respect to the description in Chapter 5: we compute the probability of being in a situation with a certain number of ambulances by means of a multi-server queue, as done by Larson (1975). That is, we compute these steady-state probabilities by Equations (6.1), (6.4) and (6.5) as an alternative to the binomial distribution of Equation (5.1), since we think that using the queueing model is a more realistic approach to reflect practice.

7.3 Numerical Results

In this section, we perform various simulations to compare the representants of the online and offline approach. We describe our experimental setup, show some results and conclude this section with a discussion on these results.

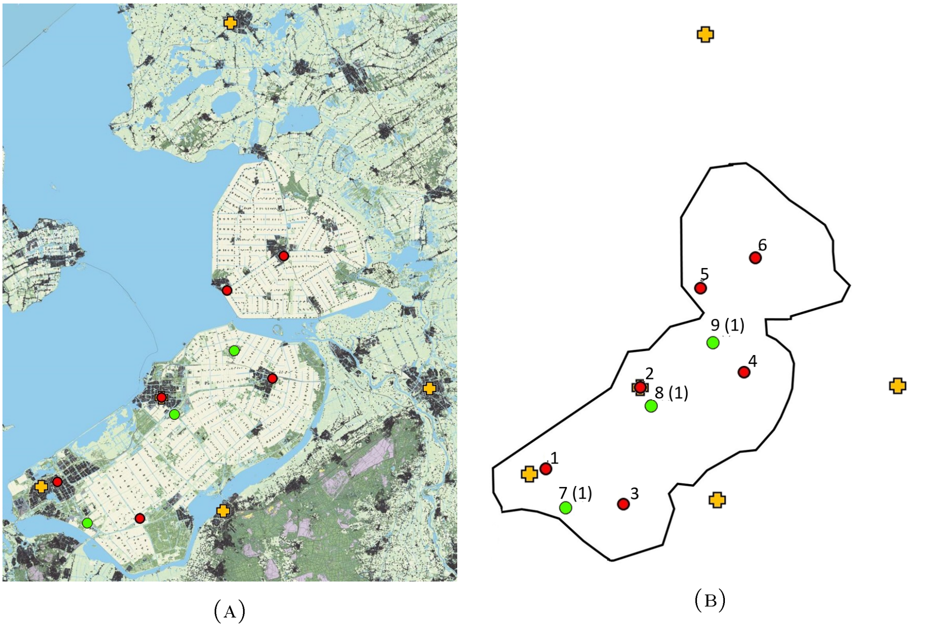


Figure 7.1: EMS region of Flevoland with out of region hospitals.

7.3.1 Experimental Setup

We use both the Flevoland (Section 3.4.1) and the Amsterdam (Section 3.4.2) EMS region as a test bed for the proposed representants. We apply one small modification to Flevoland: we add three hospitals to which patients can be transported outside the region. These are located east of waiting site 3, east of waiting site 4, and north of waiting site 6 (see Figure 7.1). For a proportion of demand points, i.e., 4-digit postal codes, one of these hospitals is nearest.

We test the performance on the following criteria:

1. The fraction of incidents responded to within 12 minutes, as well as 95% confidence intervals. Actually, the time threshold in the Netherlands is 15 minutes, but we do not simulate dispatch and chute time, which usually takes approximately 3 minutes.

2. Mean response time.

3. Total number of relocations. An ambulance travelling back to a base station after a task also counts as relocation.

4. Average relocation time.

5. Expected number of survivors, computed via Equation (7.4).

6. Realized busy fraction. We also compare this quantity with the busy fraction computed via Equation (7.5).

Quantity	Distribution	Flevoland		Amsterdam	
		Mean	Variance	Mean	Variance
Interarrival times	Exponential	1829.9	3.3×10^6	327.2	1.1×10^5
To hospital	Bernoulli	0.87	0.11	0.73	0.20
On scene time hosp.	GEV	990.7	2.2×10^5	1156.8	2.7×10^5
On scene time no hosp.	GEV	1539.5	9.5×10^5	1514.7	1.1×10^6
Hospital time	GEV	986.1	3.2×10^5	1167.4	3.6×10^5

Table 7.1: Overview of the distribution of the sources of randomness in the simulation, and their means and variances, in seconds and seconds^2, respectively.

Unlike Chapters 3–5, we do not simulate the EMS system based on the actual incident data. Instead, the approach in this chapter is similar to the one in Chapter 6 in which we estimate the quantities related to the incidents occurred in 2011 in the time interval 7 AM – 6 PM. In line with Section 6.4.4, we assume that the on-scene treatment time and the hospital transfer time follow a generalized extreme value distribution. For a graphical representation of this distribution, we refer to Figure 6.4. Table 7.1 provides an overview of the used distributions and their means and variances for the sources of randomness in the simulation. We sample one random trace consisting of 50,000 incidents according to the distributions displayed in this table. We test the performance of the online and offline representant on this trace in order to cancel out simulation noise.

The busy fraction p is input to both the online and offline representant. We compute it, similar to Chapter 6, through

$$p = \frac{\lambda}{n\mu}. \tag{7.5}$$

That is, we assume that the service process follows an exponential distribution, as in an $M/M/n$-queue. The mean interarrival times are displayed in Table 7.1. The computation of the service rates require somewhat more work. Note that the response time is part of the busy time of an ambulance, and hence, is of influence on the busy fraction. However, the response time is again influenced by the ambulance location plan. To avoid this problem, we assume a mean response time of 300 seconds, inspired by the findings in Tables 4.8 and 4.9. This allows us to compute expected busy times of 2,652 and 2,576 seconds for Flevoland and Amsterdam, respectively.

Responding to urgent demand requests is not the only duty of ambulances: carrying out ordered transport (B-calls) is part of their mission as well. As we do not consider this type of calls, it is not realistic to adopt the same fleet-size as used in practice. We compute realistic fleet sizes in two ways: we compute the busy fraction by Equation (7.5) for several values of n (fleet size). Then, we solve the AMEXCLP problem proposed by Batta et al. (1989) to compute a distribution of the n ambulances over the base stations and the corresponding objective value, that is, the expected coverage of the region. One of the fleet sizes we consider

	Objective	Location plan
Flevoland 10 units	95.94%	(1,1,2,3,4,6,6,7,8,9)
Flevoland 11 units	96.57%	(1,1,2,3,3,4,6,6,7,8,9)
Amsterdam 13 units	95.85%	(1,2,5,5,5,6,10,10,14,15,16,16,16)
Amsterdam 18 units	99.62%	(1,2,5,5,6,9,9,10,10,10,10,14,14,15,16,16,16,16)

TABLE 7.2: Configurations computed by the AMEXCLP.

Performance Indicators	Flevoland		Amsterdam	
	10 units	11 units	13 units	18 units
Percentage on time	94.93%	95.48%	83.78%	96.61%
Lower Bound 95%-CI	94.78%	95.34%	82.71%	95.40%
Upper Bound 95%-CI	95.08%	95.63%	84.85%	96.91%
Mean response time	304 s	300 s	476 s	345 s
Number of relocations	50,000	50,000	44,399	49,821
Average relocation time	456 s	454 s	397 s	356 s
Expected no. survivors	7,353	7,763	5,250	6,282
Realized busy fraction	0.1442	0.1310	0.6517	0.4483
Computed busy fraction	0.1450	0.1318	0.6055	0.4373

TABLE 7.3: Results by simulation of the static policy, as computed by the AMEXCLP.

in our simulation is the first n for which the objective value of the AMEXCLP exceeds the 0.95 mark, since in the Netherlands one aims to respond to emergency requests in 95% of the cases timely. This results in a fleet size for Flevoland of 10 ambulances, and 13 ambulances for Amsterdam. The corresponding ambulance location plans as well as the objective values of the AMEXCLP are displayed in Table 7.2.

Moreover, we simulate the static policy according to the AMEXCLP configuration for different fleet sizes and we use the first fleet size for which the simulated coverage exceeds 0.95 as second fleet size. This results in a fleet size of 11 and 18 ambulances, respectively. Table 7.3 displays the simulated results. Note that the number of relocations for Flevoland equals 50,000 (the number of generated incidents) due to the fact that an ambulance travelling back to its home station counts as relocation. However, the total number of relocations for Amsterdam is smaller than the total number of incidents, as in some cases an ambulance is inmediately dispatched to another call upon becoming available. One can compute that for 5,601 resp. 179 emergency requests no ambulance is available at the moment the call is made.

Table 7.3 shows that the busy fraction computed a priori is a rather accurate estimation on the actual realized busy fraction, especially for Flevoland. For the

	Function	Compliance tables
Flevoland 10 units	Φ_1	(7,8,9,6,1,3,2,4,6,1)
	Φ_2	(2,1,6,4,1,3,8,5,9,2)
	Φ_3	(1,8,9,6,1,3,2,4,6,7)
	Φ_4	(1,2,6,4,1,5,2,9,3,8)
Flevoland 11 units	Φ_1	(7,8,9,6,1,3,2,4,6,1,3)
	Φ_2	(2,1,6,4,1,3,8,5,9,6,2)
	Φ_3	(1,8,6,9,1,3,2,4,6,7,3)
	Φ_4	(1,2,6,4,5,1,2,9,3,8,6)
Amsterdam 13 units	Φ_1	(5,2,6,16,5,16,10,15,1,16,10,14,5)
	Φ_2	(1,5,16,1,6,2,15,1,10,16,5,12,1)
	Φ_3	(5,2,6,16,5,16,1,15,10,16,10,14,5)
	Φ_4	(1,5,1,16,6,2,1,14,10,5,16,3,1)
Amsterdam 18 units	Φ_1	(5,16,6,2,15,10,16,10,14,1,16,10,9,5,14,16,5,9)
	Φ_2	(1,5,16,1,2,6,14,10,16,3,8,17,1,14,12,4,13,5)
	Φ_3	(5,16,6,2,15,10,1,16,14,10,1,9,16,10,5,14,17,3)
	Φ_4	(1,5,16,1,2,6,14,10,3,17,8,1,16,13,4,12,5,9)

Table 7.4: Compliance tables computed by the AMEXPREP.

Amsterdam setting with 13 ambulances there is some deviation, probably caused by the relatively large mean response time due to the large fraction of calls that are queued. This has an impact on the service rate, and thus, on the busy fraction. In the setting with 18 ambulances, the mean response time is closer to the value of 300 seconds estimated beforehand. Hence, the error between computed and realized busy fraction is smaller than in the setting with 13 ambulances.

7.3.2 Results

Table 7.4 displays the nested compliance tables computed by the AMEXPREP. These are represented in a similar way as in Equation (5.31): the first k entries of the vectors displayed designate the k^{th} compliance table level. The numbers correspond to the enumeration of the base stations, as displayed in Figures 7.1 and 3.4. Note that none of the compliance tables in the settings with fewer ambulances is contained in those with more ambulances, except for Φ_1 in Flevoland. This is due to the difference in the computed steady-state probabilities (according to Equations (6.1), (6.4), and (6.5)), for different settings. Moreover, it appears that the compliance tables for Φ_1 and Φ_3 on the one hand and those for Φ_2 and Φ_4 are similar (but not the same). As the first and the third objective function heavily rely on the coverage criterion, there is an incentive to position ambulances at base stations from which many demand points can be reached in a timely manner: for Flevoland they are usually placed at the base stations between the towns first, and for Amsterdam at the edges of the city. In contrast, the other two penalty func-

Method	Performance Indicators	Φ_1	Φ_2	Φ_3	Φ_4
Online	Percentage on time	96.52%	96.34%	96.52%	94.35%
	Lower bound 95%-CI	96.41%	96.21%	96.40%	94.18%
	Upper bound 95%-CI	96.64%	96.47%	96.64%	94.53%
	Mean response time	292 s	274 s	285 s	281 s
	No. relocations	128,396	131,175	136,819	117,037
	Average relocation time	826 s	863 s	799 s	781 s
	Expected no. survivors	7,500	7,927	7,659	8,007
Offline	Percentage on time	96.52%	96.38%	96.52%	94.35%
	Lower bound 95%-CI	96.41%	96.25%	96.40%	94.18%
	Upper bound 95%-CI	96.64%	96.52%	96.64%	94.53%
	Mean response time	292 s	272 s	284 s	281 s
	No. relocations	128,396	124,337	136,457	117,037
	Average relocation time	826 s	764 s	798 s	781 s
	Expected no. survivors	7,500	7,983	7,670	8,007

Table 7.5: Simulation results of the Flevoland setting with 10 ambulances.

tions, in which the mean response time plays a major role, position ambulances inside the towns first (Flevoland), and for Amsterdam it is important to occupy the base station in the city center (around base station 1).

In Tables 7.5-7.8 we show the simulated results for the mentioned EMS systems. Note that we made the underlying assumption that any of the 50,000 patients suffers from a cardiac arrest in the computation of the expected number of survivors. We omitted the realized busy fractions, as these are very similar to the ones displayed in Table 7.3, albeit a little smaller due to the reduced mean response time in comparison to the static policy.

The results are most remarkable in the sense that the performance for both the online and offline representant is very similar: in none of the considered situation one approach convincingly outperforms the other. In some settings (Flevoland with 10 ambulances, first and fourth penalty function) even exactly the same decisions are made at each decision moment. Only the Amsterdam framework with 13 ambulances and Φ_2 and Φ_4 as penalty function show a substantial difference in the percentage on time criterion between the online and offline representant, in favor of the offline policy. However, the difference is not significant, as the confidence intervals overlap. Additionally, this setting is not a realistic one as the fleet size is far too limited to reach the 95% threshold. At last, the performance gain of the offline respresentant is associated with a substantial increase in crew workload, as both the number of relocations and the average relocation time are larger than for the online approach.

With respect to the crew-based performance indicators, we observe that the performance of both approaches is similar in general as well. However, an exception is the setting of Flevoland with a fleet size of 10 and Φ_2 as the penalty function

Method	Performance Indicators	Φ_1	Φ_2	Φ_3	Φ_4
Online	Percentage on time	97.06%	96.75%	97.07%	95.73%
	Lower bound 95%-CI	96.94%	96.63%	96.94%	95.59%
	Upper bound 95%-CI	97.19%	96.89%	97.20%	95.87%
	Mean response time	280 s	265 s	276 s	269 s
	No. relocations	130,615	130,398	141,976	121,549
	Average relocation time	750 s	837 s	725 s	799 s
	Expected no. survivors	7,721	8,106	7,813	8,141
Offline	Percentage on time	97.07%	96.77%	97.07%	95.74%
	Lower bound 95%-CI	96.94%	96.65%	96.93%	95.59%
	Upper bound 95%-CI	97.21%	96.90%	97.20%	95.88%
	Mean response time	280 s	265 s	276 s	269 s
	No. relocations	131,285	130,050	142,649	121,343
	Average relocation time	783 s	838 s	754 s	797 s
	Expected no. survivors	7,721	8,106	7,814	8,146

Table 7.6: Simulation results of the Flevoland setting with 11 ambulances.

Method	Performance Indicators	Φ_1	Φ_2	Φ_3	Φ_4
Online	Percentage on time	84.59%	83.38%	84.44%	82.56%
	Lower bound 95%-CI	83.61%	82.44%	83.50%	81.65%
	Upper bound 95%-CI	85.58%	84.32%	85.38%	83.48%
	Mean response time	450 s	430 s	446 s	435 s
	No. relocations	75,522	73,909	74,903	73,026
	Average relocation time	416 s	380 s	411 s	378 s
	Expected no. survivors	5,583	6,174	5,718	6,122
Offline	Percentage on time	84.55%	84.62%	84.38%	83.83%
	Lower bound 95%-CI	83.59%	83.63%	83.44%	82.92%
	Upper bound 95%-CI	85.52%	85.61%	85.31%	84.74%
	Mean response time	447 s	422 s	445 s	426 s
	No. relocations	75,747	76,589	74,797	77,199
	Average relocation time	414 s	390 s	411 s	391 s
	Expected no. survivors	5,653	6,264	5,750	6,251

Table 7.7: Simulation results of the Amsterdam setting with 13 ambulances.

Method	Performance Indicators	Φ_1	Φ_2	Φ_3	Φ_4
Online	Percentage on time	98.77%	98.43%	98.78%	98.32%
	Lower bound 95%-CI	98.57%	98.23%	98.59%	98.13%
	Upper bound 95%-CI	98.96%	98.63%	98.97%	98.51%
	Mean response time	278 s	239 s	256 s	238 s
	No. relocations	98,486	108,545	102,224	106,368
	Average relocation time	455 s	424 s	443 s	422 s
	Expected no. survivors	7,530	8,585	8,088	8,625
Offline	Percentage on time	98.73%	98.39%	98.80%	98.29%
	Lower bound 95%-CI	98.54%	98.20%	98.61%	98.10%
	Upper bound 95%-CI	98.93%	98.57%	98.99%	98.48%
	Mean response time	272 s	237 s	254 s	236 s
	No. relocations	99,747	106,813	101,927	106,553
	Average relocation time	444 s	432 s	437 s	426 s
	Expected no. survivors	7,691	8,625	8,139	8,687

Table 7.8: Simulation results of the Amsterdam setting with 18 ambulances.

of consideration: the offline policy yields a smaller number of relocations, while also a decrease in the average location time is achieved, compared to the online regime. This phenomenon can be explained as follows: if a state transition from availability level 9 to level 10 occurs, the offline policy sends a unit to base station 2 (see Table 7.4). As the majority of ambulances becomes available in either the hospital in city 1 or the one in city 2, this yields a short relocation time: no chain relocation is used. However, in the same situation, the online policy favors base station 6 instead of base station 2. Therefore, it frequently occurs that an ambulance from either city 1 or city 2 is sent to town 6. To decrease the time until the system is in compliance, a chain relocation is carried out, which induces an increase in the total number of relocations compared to the online regime.

One can explain the differences in crew based performance for the first and third penalty function on Flevoland with 11 units by a similar reasoning, although it is the other way around: the online policy performs better than the offline equivalent. This is due to the fact that the compliance table sends an ambulance to base station 6 when a transition from availability level 8 to level 9 occurs. In contrast, the online policy sends one to waiting site 9, which is usually closer to the location at which ambulances become available frequently.

On one performance measure there is a difference in both approaches: without any exception, the offline policies perform at least as good as the online equivalents on the performance measure of expected number of survivors. Although the differences for some settings are very minor, we emphasize that one should not think light-headed about this phenomenon. After all, every saved life counts. In that sense, the expected number of survivors is perhaps the best theoretical measure for the evaluation of EMS providers. However, unfortunately, in practice it is very hard to quantify the effects of the response time on the survival probability of a patient, as many other factors play a role. For instance, the type of disease, the speed at which the patient (or a bystander) can make the emergency call, whether first aid is applied by a bystander all have a large effect as well.

7.4 Concluding Remarks

In this last chapter, we compared representants of both the online and the offline approach to the ambulance relocation problem. The simulations show very similar results, both from a patient and a crew perspective. Due to the limited diversity in performance (at least, for the considered representants), other aspects are important in the consideration about whether one should implement an online or an offline method, as both approaches have their strenghts and shortcomings.

A benefit of the offline over the online approach is its simplicity, although there are some exceptions like the mentioned ADP policies. Offline policies are typically easy to implement: in case of a compliance table policy, one could print the whole relocation policy on one piece of paper (or even write it down in one line, see Table 7.4). There is no need to build a decision support system for implementing an offline policy, although one could do this. In contrast, the implementation of an online policy requires far more work, as such a system has to be designed. Moreover, it has to be connected to a computer-aided dispatch system (CAD), because it needs to know the location and status of each vehicle. Implementation of such a policy easily can take a couple of months, as reported by Van Buuren et al. (2016) who implemented the penalty heuristic of Chapter 3 and the DMEXCLP method explained in Chapter 4 in a real-time decision support system in the emergency control center of Flevoland.

Additionally, offline policies are usually simple to explain to and to use by dispatchers. Since this kind of policy does not need implementation in a decision support system, it is probably closer to the way dispatchers are used to work. Therefore, it might meet less resistance from dispatchers in comparison with online policies. As a consequence, an EMS provider exploring the use of proactive relocation methods might get an offline policy accepted easier than when an online policy would be introduced.

However, offline policies are rather adamant in the sense that a completely new policy has to be computed if there is just one small abnormality of the system, e.g., a deviation of the estimated travel times due to a road construction, a temporary change in expected busy fraction, or a slightly different demand distribution. From this point of view, online policies are more flexible: one can easily adjust the parameters in the decision support system. Moreover, due to the online character, it is relatively easy to incorporate additional features in those policies. This is perhaps best reflected in Sections 4.3.3 and 6.3.4: in these sections, bounds on the relocation time were incorporated, in an online and offline fashion, respectively. For the online variant incorporation such bounds benefit the computation time. After all, the set of decisions is reduced as some base stations are excluded. However, the incorporation of relocation bounds in the offline approach is much harder, as the size of the ILP increases due to the addition of extra variables and constraints.

Although the adoption of an offline policy is a good first step in the implementation of relocation policies in the emergency control center, we believe that in the end online policies have a large potential to outperform offline policies due to the flexibility to incorporate additional system state characteristics. Simulation is a highly helpful tool in the evaluation of ambulance relocation policies due the very complex and stochastic nature of the efficient planning of ambulances services.

List of Acronyms

ADP	Approximate Dynamic Programming
AED	Automated External Defibrillator
ALS	Advanced Life Support
AMEXCLP	Adjusted Maximum Expected Location Problem
AMEXPREP	Adjusted Minimal Expected Penalty Relocation Problem
BACOP	Backup Coverage Problem
BLS	Basic Life Support
CAD	Computed-Aided Dispatch
DAM	Dynamic Ambulance Management
DMEXCLP	Dynamic Maximum Expected Location Problem
DSM	Double Standard Model
EMS	Emergency Medical Services
GAAP	Generalized Ambulance Assignment Problem
GABAP	Generalized Ambulance Bottleneck Assignment Problem
GEV	Generalized Extreme Value
GPS	Global Positioning System
ILP	Integer Linear Programming
LBAP	Linear Bottleneck Assignment Problem
LSCP	Location Set Covering Problem
MALP	Maximum Availability Location Problem
MCLP	Maximum Coverage Location Problem
MDP	Markov Decision Process
MECRP	Maximal Expected Coverage Relocation Problem
MEXCLP	Maximum Expected Location Problem
MEXPREP	Minimal Expected Penalty Relocation Problem
MS	Management Science
MWBM	Minimum Weighted Bipartite Matching

OR	Operations Research
REPRO	From REactive to PROactive planning of ambulance services
RIVM	Rijksinstituut Volksgezondheid en Milieu
RP^t	Redeployment Problem at time t
RRA	Rapid Responder Ambulance
RTA	Regular Transport Ambulance
TIMEXCLP	Time-dependent MEXCLP

Bibliography

L. Aboueljinane, E. Sahin, and Z. Jemai. A Review on Simulation Models applied to Emergency Medical Service Operations. *Computers & Industrial Engineering*, 66(4):734–750, 2013.

R. Alanis, A. Ingolfsson, and B. Kolfal. A Markov Chain Model for an EMS System with Repositioning. *Production and Operations Management*, 22(1): 216–231, 2013.

Ambulancezorg Nederland. Ambulances in-zicht 2014. Technical report, 2014.

T. Andersson and P. Värbrand. Decision Support Tools for Ambulance Dispatch and Relocation. *Journal of the Operational Research Society*, 58(2):195–201, 2007.

S. Ansari, L. A. McLay, and M. E. Mayorga. A Maximum Expected Covering Problem for District Design. *Transportation Science*, 2015.

A. Başar, B. Çatay, and T. Ünlüyurt. A Multi-Period Double Coverage Approach for Locating the Emergency Medical Service Stations in Istanbul. *Journal of the Operational Research Society*, 62(4):627–637, 2011.

A. Başar, B. Çatay, and T. Ünlüyurt. A Taxonomy for Emergency Service Station Location Problem. *Optimization Letters*, 6:1147–1160, 2012.

D. Bandara, M. E. Mayorga, and L. A. McLay. Optimal Dispatching Strategies for Emergency Vehicles to Increase Patient Survivability. *International Journal of Operational Research*, 15(2):195–214, 2012.

D. Bandara, M. E. Mayorga, and L. A. McLay. Priority Dispatching Strategies for EMS Systems. *Journal of the Operational Research Society*, 65:572–587, 2014.

R. Batta, J. M. Dolan, and N. N. Krishnamurthy. The Maximal Expected Covering Problem: Revisited. *Transportation Science*, 23(4):277–287, 1989.

V. Bélanger, A. Ruiz, and P. Soriano. *Recent Advances in Emergency Medical Services Management.* Tech. Rep. CIRRELT-2015-28, CIRRELT, 2015.

O. Berman. Repositioning of Distinguishable Urban Service Units on Networks. *Computers & Operations Research*, 8(2):105–118, 1981a.

O. Berman. Dynamic Repositioning of Indistinguishable Service Units on Transportation Networks. *Transportation Science*, 15(2):115–136, 1981b.

O. Berman. Repositioning of Two Distinguishable Service Vehicles on Networks. *IEEE Transactions on Systems, Man and Cybernetics*, 11(3):187–193, 1981c.

L. Brotcorne, G. Laporte, and F. Semet. Ambulance Location and Relocation Models. *European Journal of Operational Research*, 147(3):451–463, 2003.

S. Budge, A. Ingolfsson, and E. Erkut. Approximating Vehicle Dispatch Probabilities for Emergency Service Systems with Location-Specific Service Times and Multiple Units per Location. *Operations Research*, 57(1):251–255, 2009.

S. Budge, A. Ingolfsson, and D. Zerom. Empirical Analysis of Ambulance Travel Times: The Case of Calgary Emergency Medical Services. *Management Science*, 56(4):716–723, 2010.

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

A. Carter, J. Gould, P. Vanberkel, J. Jensen, J. Cook, S. Carrigan, M. Wheatley, and A. Travers. Offload Zones to Mitigate Emergency Medical Services Offload Delay in the Emergency Department: a Process Map and Hazard Analysis. *Canadian Journal of Emergency Medicine*, pages 1–9, 2015.

N. Channouf, P. L'Ecuyer, A. Ingolfsson, and A. N. Avramidis. The Application of Forecasting Techniques to Modeling Emergency Medical System Calls in Calgary, Alberta. *Health Care Management Science*, 10(1):25–45, 2007.

A. Charnes and J. Storbeck. A Goal Programming Model for the Siting of Multilevel EMS Systems. *Socio-Economic Planning Sciences*, 14(4):155–161, 1980.

K. C. Chong, S. G. Henderson, and M. E. Lewis. The Vehicle Mix Decision in Emergency Medical Service Systems. *Manufacturing & Service Operations Management*, 2015.

R. Church and C. S. ReVelle. The Maximal Covering Location Problem. *Papers Regional Science Association*, 32(1):101–118, 1974.

M. S. Daskin. Application of an Expected Covering Model To Emergency Medical Service System Design. *Decision Sciences*, 13:416–439, 1982.

M. S. Daskin. A Maximum Expected Covering Location Model: Formulation, Properties and Heuristic Solution. *Transportation Science*, 17(1):48–70, 1983.

M. S. Daskin and E. H. Stern. A Hierarchical Objective Set Covering Model for Emergency Medical Service Vehicle Deployment. *Transportation Science*, 15: 137–152, 1981.

V. De Maio, I. Stiell, G. Wells, and D. Spaite. Optimal Defibrillation for Maximum Out-of-Hospital Cardiac Arrest Survival Rates. *Annals of Emergency Medicine*, 42(2):242–250, 2003.

D. Degel, L. Wiesche, S. Rachuba, and B. Werners. Time-dependent Ambulance Allocation Considering Data-driven Empirically Required Coverage. *Health Care Management Science*, 18(4):444–458, 2015.

K. F. Doerner, W. J. Gutjahr, R. F. Hartl, M. Karall, and M. Reimann. Heuristic Solution of an Extended Double-Coverage Ambulance Location Problem for Austria. *Central European Journal of Operations Research*, 13:325–340, 2005.

G. Erdoğan, E. Erkut, A. Ingolfsson, and G. Laporte. Scheduling Ambulance Crews for Maximum Coverage. *Journal of the Operational Research Society*, 61 (4):543–550, 2009.

E. Erkut, A. Ingolfsson, and G. Erdoğan. Ambulance Location for Maximum Survival. *Naval Research Logistics*, 55(1):42–58, 2008.

E. Erkut, A. Ingolfsson, T. Sim, and G. Erdoğan. Computational Comparison of Five Maximal Covering Models for Locating Ambulances. *Geographical Analysis*, 41(1):43–65, 2009.

R. D. Galvão and C. S. ReVelle. A Lagrangean Heuristic for the Maximal Covering Location Problem. *European Journal of Operational Research*, 88(1):114–123, 1996.

R. D. Galvão, F. Y. Chiyoshi, and R. Morabito. Towards Unified Formulations and Extensions of Two Classical Probabilistic Location Models. *Computers & Operations Research*, 32:15–33, 2005.

M. Gendreau, G. Laporte, and F. Semet. Solving an Ambulance Location Model by Tabu Search. *Location Science*, 5(2):75–88, 1997.

M. Gendreau, G. Laporte, and F. Semet. A Dynamic Model and Parallel Tabu Search Heuristic for Real-time Ambulance Relocation. *Parallel Computing*, 27 (12):1641–1653, 2001.

M. Gendreau, G. Laporte, and F. Semet. The Maximal Expected Coverage Relocation Problem for Emergency Vehicles. *Journal of the Operational Research Society*, 57:22–28, 2006.

J. B. Goldberg. Operations Research Models for the Deployment of Emergency Services Vehicles. *EMS Management Journal*, 1:20–39, 2004.

J. B. Goldberg, R. Dietrich, J. Ming Chen, M. G. Mitwasi, T. Valenzuela, and E. Criss. Validating and Applying a Model for Locating Emergency Medical Vehicles in Tuczon, AZ. *European Journal of Operational Research*, 49(3):308–324, 1990.

L. V. Green and P. J. Kolesar. Improving Emergency Responsiveness with Management Science. *Management Science*, 50(8):1001–1014, 2004.

K. Hogan and C. S. ReVelle. Concepts and Application of Backup Coverage. *Management Science*, 34:1434–1444, 1986.

A. Ingolfsson, S. Budge, and E. Erkut. Optimal Ambulance Location with Random Delays and Travel Times. *Health Care Management Science*, 11(3):262–274, 2008.

C. J. Jagtenberg, S. Bhulai, and R. D. van der Mei. An Efficient Heuristic for Real-time Ambulance Redeployment. *Operations Research for Health Care*, 4: 27–35, 2015.

C. J. Jagtenberg, S. Bhulai, and R. D. van der Mei. Optimality of the Closest-idle Policy in Advanced Ambulance Dispatching. *Health Care Management Science (to appear)*, 2016.

J. P. Jarvis. Approximating the Equilibrium Behavior of Multi-Server Loss Systems. *Management Science*, 31(2):235–239, 1985.

V. Jayaraman and R. Srivastava. A Service Logistics Model for Simultaneous Siting of Facilities and Multiple Levels of Equipment. *Computers & Operations Research*, 22(2):191–204, 1995.

O. Karasakal and E. K. Karasakal. A Maximal Covering Location Model in the Presence of Partial Coverage. *Computers & Operations Research*, 31(9):1515–1526, 2004.

R. B. O. Kerkkamp and K. I. Aardal. A Constructive Proof of Swap Local Search Worst-case Instances for the Maximum Coverage Problem. *Operations Research Letters*, 44(3):329–335, 2016.

V. A. Knight, P. R. Harper, and L. Smith. Ambulance Allocation for Maximal Survival with Heterogeneous Outcome Measures. *Omega*, 40(6):918–926, 2012.

P. J. Kolesar and W. E. Walker. An Algorithm for the Dynamic Relocation of Fire Companies. *Operations Research*, 22(2):249–274, 1974.

G. J. Kommer and S. Zwakhals. Referentiekader spreiding en beschikbaarheid ambulancezorg, 2008.

G. Laporte, F. V. Louveaux, F. Semet, and A. Thirion. Application of the double standard model for ambulance location. In *Innovations in Distribution Logistics*, pages 235–249. Springer, 2009.

M. Larsen, M. Eisenberg, R. Cummins, and A. Hallstrom. Predicting Survival from Out-of-Hospital Cardiac Arrest - a Graphic Model. *Annals of Emergency Medicine*, 22:1652–1658, 1993.

R. C. Larson. A Hypercube Queuing Model for Facility Location and Redistricting in Urban Emergency Services. *Computers & Operations Research*, 1(1):67–95, 1974.

R. C. Larson. Approximating the Performance of Urban Emergency Service Systems. *Operations Research*, 23(5):845–868, 1975.

S. Lee. The Role of Preparedness in Ambulance Dispatching. *Journal of the Operational Research Society*, 62:1888–1897, 2011.

X. Li, Z. Zhao, X. Zhu, and T. Wyatt. Covering Models and Optimization Techniques for Emergency Response Facility Location and Planning: a Review. *Mathematical Methods of Operations Research*, (74):281–310, 2011.

C. S. Lim, R. Mamat, and T. Bräunl. Impact of Ambulance Dispatch Policies on Performance of Emergency Medical Services. *IEEE Transactions on Intelligent Transportation Systems*, 12:624–632, 2011.

M. Maleki, N. Majlesinasab, and M. Mehdi Sepehri. Two New Models for Redeployment of Ambulances. *Computers & Industrial Engineering*, 78:271–284, 2014.

M. B. Mandell. Covering Models for Two-Tiered Emergency Medical Service Systems. *Location Science*, 6:355–368, 1998.

V. Marianov and C. S. ReVelle. The Capacitated Standard Response Fire Protection Siting Problem: Deterministic and Probabilistic Models. *Annals of Operations Research*, 40(1):303–322, 1992.

V. Marianov and C. S. ReVelle. The Queueing Maximal Availability Location Problem: a Model for the Siting of Emergency Vehicles. *European Journal of Operational Research*, 93(1):110–120, 1996.

V. Marianov and D. Serra. Hierarchical Location-allocation Models for Congested Systems. *European Journal of Operational Research*, 135(1):195–208, 2001.

A. J. Mason. Simulation and Real-Time Optimised Relocation for Improving Ambulance Operations. In B. Denton, editor, *Handbook of Healthcare Operations: Methods and Applications*, pages 289–317. Springer, New York, 2013.

D. S. Matteson, M. W. McLean, D. B. Woodard, and S. G. Henderson. Forecasting Emergency Medical Service Call Arrival Rates. *Annals of Applied Statistics*, 5 (2B):1379–1406, 2011.

M. S. Maxwell, M. Restrepo, S. G. Henderson, and H. Topaloglu. Approximate Dynamic Programming for Ambulance Redeployment. *INFORMS Journal on Computing*, 22(2):266–281, 2010.

M. S. Maxwell, S. G. Henderson, and H. Topaloglu. Tuning Approximate Dynamic Programming Policies for Ambulance Redeployment via Direct Search. *Stochastic Systems*, 3(2):322–361, 2013.

M. S. Maxwell, E. C. Ni, C. Tong, S. G. Henderson, H. Topaloglu, and S. R. Hunter. A Bound on the Performance of an Optimal Ambulance Redeployment Policy. *Operations Research*, 62(5):1014–1027, 2014.

M. E. Mayorga, D. Bandara, and L. A. McLay. Districting and Dispatching Policies for Emergency Medical Service Systems to Improve Patient Survival. *IIE Transactions on Healthcare Systems Engineering*, 3(1):39–56, 2013.

L. A. McLay. A Maximum Expected Covering Location Model with Two Types of Servers. *IIE Transactions*, 41(8):730–741, 2009.

L. A. McLay and M. E. Mayorga. Evaluating Emergency Medical Service Performance Measures. *Health Care Management Science*, 13(2):124–136, 2010.

L. A. McLay and M. E. Mayorga. A Dispatching Model for Server-to-Customer Systems that Balances Efficiency and Equity. *Manufacturing & Service Operations Management*, 15(2):205–220, 2013a.

L. A. McLay and M. E. Mayorga. A Model for Optimally Dispatching Ambulances to Emergency Calls with Classification Errors in Patient Priorities. *IIE Transactions*, 45(1):1–24, 2013b.

M. Moeini, Z. Jemai, and E. Sahin. Location and Relocation Problems in the Context of the Emergency Medical Service Systems: a Case Study. *Central European Journal of Operations Research*, 23:641–658, 2014.

R. Nair and E. Miller-Hooks. Evaluation of Relocation Strategies for Emergency Medical Service Vehicles. *Transportation Research Record*, 2137:63–73, 2009.

J. Naoum-Sawaya and S. Elhedhli. A Stochastic Optimization Model for Real-Time Ambulance Redeployment. *Computers & Operations Research*, 40:1972–1978, 2013.

S. H. Owen and M. S. Daskin. Strategic Facility Location: A Review. *European Journal of Operational Research*, 111:423–447, 1998.

M. A. Pereira, E. L. F. Senne, and L. A. N. Lorena. A Decomposition Heuristic for the Maximal Covering Location Problem. *Advances in Operations Research*, 2010.

M. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, USA, 1st edition, 1994.

H. K. Rajagopalan and C. Saydam. A Minimum Expected Response Model: Formulation, Heuristic Solution, and Application. *Socio-Economic Planning Sciences*, 43(4):253–262, 2009.

H. K. Rajagopalan, C. Saydam, and J. Xiao. A Multiperiod Set Covering Location Model for Dynamic Redeployment of Ambulances. *Computers & Operations Research*, 35(3):814–826, 2008.

J. F. Repede and J. J. Bernardo. Developing and Validating a Decision Support System for Locating Emergency Medical Vehicles in Louisville, Kentucky. *European Journal of Operational Research*, 75(3):567–581, 1994.

C. S. ReVelle. Review, Extension and Prediction in Emergency Service Siting Models. *European Journal of Operational Research*, 40(1):58–69, 1989.

C. S. ReVelle and K. Hogan. The Maximum Availability Location Problem. *Transportation Science*, 23:192–200, 1989.

C. S. ReVelle and R. W. Swain. Central Facilities Location. *Geographical Analysis*, 2(1):30–42, 1970.

D. P. Richards. *Optimised Ambulance Redeployment Strategies.* Master thesis, University of Auckland, New Zealand, 2007.

H. Rinne. *The Weibull Distribution: a Handbook.* Taylor & Francis, 2008.

E. S. Savas. Simulation and Cost-Effectiveness Analysis of New York's Emergency Ambulance Service. *Management Science*, 15(12):B608–B627, 1969.

C. Saydam and M. McKnew. A Separable Approach to Expected Coverage: an Application to Ambulance Location. *Decision Sciences*, 16:381–389, 1985.

C. Saydam, H. K. Rajagopalan, E. Sharer, and K. Lawrimore-Belanger. The Dynamic Redeployment Coverage Location Model. *Health Systems*, 2(2):103–119, 2013.

D. A. Schilling, D. J. Elzinga, J. Cohon, R. L. Church, and C. S. ReVelle. The Team/Fleet Models for Simultaneous Facility and Equipment Siting. *Transportation Science*, 13(2):163–175, 1979.

E. Schippers. Beantwoording kamervragen over de inzet van rapid responders en brambulances (answers to parliamentary questions regarding rapid responders and brambulances). 2014.

V. Schmid. Solving the Dynamic Ambulance Relocation and Dispatching Problem using Approximate Dynamic Programming. *European Journal of Operational Research*, 219(3):611–621, 2012.

V. Schmid and K. F. Doerner. Ambulance Location and Relocation Problems with Time-dependent Travel Times. *European Journal of Operational Research*, 207 (3):1293–1303, 2010.

A. Schrijver. *Combinatorial Optimization - Polyhedra and Efficiency*, volume A, chapter 17. Springer-Verlag Berlin Heidelberg, 2003.

G. Schwarz. Estimating the Dimension of a Model. *Annals of Statistics*, 6(2): 461–464, 1978.

H. Setzler, C. Saydam, and S. Park. EMS Call Volume Predictions: A Comparative Study. *Computers & Operations Research*, 36(6):1843–1851, 2009.

V. P. Singh. Generalized Extreme Value Distribution. In *Entropy-Based Parameter Estimation in Hydrology*, chapter 11, pages 169–183. 2010.

P. N. Skandalakis, P. Lainas, O. Zoras, J. E. Skandalakis, and P. Mirilas. "To Afford the Wounded Speedy Assistance": Dominique Jean Larrey and Napoleon. *World Journal of Surgery*, 30(8):1392–1399, 2006.

K. Sudtachat, M. E. Mayorga, and L. A. McLay. Recommendations for Dispatching Emergency Vehicles under Multitiered Response via Simulation. *International Transactions in Operational Research*, 21:581–617, 2014.

K. Sudtachat, M. E. Mayorga, and L. A. McLay. A Nested-Compliance Table Policy for Emergency Medical Service Systems under Relocation. *Omega*, 58: 154–168, 2016.

C. Toregas, R. W. Swain, C. S. ReVelle, and L. Bergman. The Location of Emergency Service Facilities. *Operations Research*, 19:1363–1373, 1971.

H. Toro-Díaz, M. E. Mayorga, S. Chanta, and L. A. McLay. Joint Location and Dispatching Decisions for Emergency Medical Services. *Computers and Industrial Engineering*, 64(4):917–928, 2013.

H. Toro-Díaz, M. E. Mayorga, L. A. McLay, H. K. Rajagopalan, and C. Saydam. Reducing Disparities in Large-Scale Emergency Medical Service Systems. *Journal of the Operational Research Society*, 66(7):1169–1181, 2014.

T. Valenzuela, D. Roe, S. Cretin, D. Spaite, and M. Larsen. Estimating Effectiveness of Cardiac Arrest Intervention - a logistic Regression Survival Model. *Circulation*, 96:3308–3313, 1997.

T. C. van Barneveld. The Minimum Expected Penalty Relocation Problem for the Computation of Compliance Tables for Ambulance Vehicles. *INFORMS Journal on Computing*, 28(2):370–384, 2016.

T. C. van Barneveld, S. Bhulai, and R. D. van der Mei. A Dynamic Ambulance Management Model for Rural Areas. *Health Care Management Science (to appear)*, 2015.

T. C. van Barneveld, S. Bhulai, and R. D. van der Mei. The Effect of Ambulance Relocations on the Performance of Ambulance Service Providers. *European Journal of Operational Research*, 252(1):257–269, 2016a.

T. C. van Barneveld, C. J. Jagtenberg, S. Bhulai, and R. D. van der Mei. Real-Time Ambulance Relocation: Assessing Real-Time Deployment Strategies for Ambulance Relocation. *Submitted for publication*, 2016b.

T. C. van Barneveld, R. D. van der Mei, and S. Bhulai. Compliance Tables for an EMS system with Two Types of Medical Response Units. *Computers & Operations Research*, 80:68–81, 2017.

M. van Buuren, C. J. Jagtenberg, T. C. van Barneveld, R. D. van der Mei, and S. Bhulai. Ambulance Dispatch Center Pilots Proactive Relocation Policies to Enhance Effectiveness. *Submitted for publication*, 2016.

P. L. van den Berg. *Logistics of Emergency Response Vehicles: Facility Location, Routing, and Shift Scheduling*. PhD thesis, Delft University of Technology, The Netherlands, 2016.

P. L. van den Berg and K. I. Aardal. Time-dependent MEXCLP with Start-up and Relocation Cost. *European Journal of Operational Research*, 242(2):383–389, 2015.

P. L. van den Berg, G. J. Kommer, and B. Zuzáková. Linear Formulation for the Maximum Expected Coverage Location Model with Fractional Coverage. *Operations Research for Health Care*, 8:33–41, 2014.

R. Waaelwijn, R. de Vos, J. Tijssen, and R. Koster. Survival Models for Out-of-Hospital Cardiopulmonary Resuscitation from the Perspectives of the By-stander, the First Responder, and the Paramedic. *Resuscitation*, 51:113–122, 2001.

Y. H. Wang. On the Number of Successes in Independent Trials. *Statistica Sinica*, 3(2):295–312, 1993.

B. S. Westgate, D. B. Woodard, D. S. Matteson, and S. G. Henderson. Travel Time Estimation for Ambulances using Bayesian Data Augmentation. *Annals of Applied Statistics*, 7(2):1139–1161, 2013.

B. S. Westgate, D. B. Woodard, D. S. Matteson, and S. G. Henderson. Large-Network Travel Time Distribution Estimation for Ambulances. *European Journal of Operational Research*, 252(1):322–333, 2016.

L. Zhang. *Simulation Optimisation and Markov Models for Dynamic Ambulance Redeployment.* PhD thesis, University of Auckland, New Zealand, 2012.

Summary

In life-threatening situations where every second counts, the ability of ambulance service providers to arrive at the emergency scene within a few minutes to provide medical aid can make the difference between survival or death. In the Netherlands the response-time target is 15 minutes for incidents of the highest urgency. To realize short response times at affordable costs, adequate planning of ambulance services is crucial. An essential element herein is an efficient distribution of ambulances over the region. After all, the location of emergency vehicles at the time an incident is reported at the emergency control center determines to a large extent whether the response-time target is achieved. A complicating factor in addressing these ambulance location problems is the omnipresence of uncertainty in the ambulance service-provisioning process. Especially the uncertainty regarding the availability of emergency vehicles is crucial. Ambulances are not always dispatchable to an incident during their duty time, for instance, due to the treatment of a patient at the emergency scene of an earlier incident or due to the transportation of such a patient. This unavailability can result in a temporarily inefficient distribution of ambulances over the region.

A way to resolve the problem mentioned above is a temporary redeployment of one or more ambulances. This is the core of Dynamic Ambulance Management (DAM). An important advantage of DAM is the ability to anticipate future incidents in a more flexible way. Response times, and hence, mortality and morbidity, can be reduced through the repositioning of ambulances in real time. This dissertation is concerned with several models and algorithms for the optimization of the coverage by means of performing so-called *proactive relocations*.

The optimization methods in this thesis can be roughly classified into two main categories: (1) *online* and (2) *offline* algorithms. The key difference between both is the moment at which the (majority of the) computational work is done. For online methods this happens at the moment at which a relocation decision needs to be taken. This class of methods can handle a very detailed state description, because just for one specific state of the EMS system at the time the relocation decision is computed. In the offline approach the majority of the computations is done a priori and for each possible state the corresponding relocation decision is stored. When a certain situation (i.e., state) occurs, the relocation decision is retrieved and applied. The computation time highly depends on the number of possible states. Therefore, the state description is typically less sophisticated than in the online approach in order to keep the computation time manageable.

The first part of this dissertation (Chapters 2-4) is concerned with the online approach to the ambulance relocation problem. Chapter 2 focuses on rural regions. These usually differ from their urban counterparts due to a smaller number of incidents, a smaller number of ambulances on duty, and the geographical spread of incidents over the region. The relocation problem is modeled as a Markov decision problem, in which the response-time dependent performance objective can be chosen arbitrarily through the selection of a suitable *penalty function*. This function assigns to each possible response time a certain penalty. Besides, this chapter describes a heuristic for the calculation of good relocation decisions, based on the presented model. Chapter 2 concludes with the illustration of this heuristic for several penalty functions using a realistic EMS system corresponding to a Dutch region as test bed.

A frequently mentioned drawback of DAM is the increase in the crew's workload, due to the additional number of relocations. Chapter 3 addresses this issue: the trade-off between the number of relocations and the response-time performance is studied. To this end the penalty heuristic is proposed in this chapter. This heuristic uses the concept of penalty function. Moreover, this chapter introduces the use of so-called chain relocations: a otherwise long trip can be split into multiple shorter ones to attain the desired ambulance configuration (as computed by the penalty heuristic) faster. The chapter is concluded with an extensive numerical study regarding the mentioned trade-off. The consequences of restrictions on the number of ambulance relocations on the performance are studied, based on a large number of realistic situations, using a specific penalty function provided by ambulance practitioners. The response-time performance increases significantly if only a few ambulance relocations are carried out. However, frequent repositioning can possibly result in a performance loss.

In Chapter 4 several insights concerning the implementation of relocation strategies in practice are presented. To this end, the proposed penalty heuristic of Chapter 3 is combined with the Dynamic MEXCLP algorithm of Jagtenberg et al. (2015). The following five aspects are discussed: (1) the frequency of redeployment decision moments, (2) the inclusion of busy ambulances in the state description of the system, (3) the performance criterion on the quality of the relocation strategy, (4) the use of chain relocations, and (5) time bounds on the relocation time. The chapter continues with an extensive simulation study regarding the practical implementation of these facets.

The offline approach for solving the ambulance relocation problem is the topic of the second part of this dissertation (Chapters 5 and 6). Chapter 5 describes a integer linear programming formulation for the computation of *ambulance compliance tables*, called MEXPREP. This model is an extension of the MECRP model by Gendreau et al. (2006) in two different directions: (1) it accounts for the fact that ambulances might become busy during the execution of the compliance table policy, and (2) a generic performance objective can be chosen through the definition of the corresponding penalty function. This chapter also introduces an adjusted version (called AMEXPREP) that relaxes the assumptions made on the busy fraction. A section with numerical results, based on the simulation of the compliance tables computed by MEXPREP, concludes the chapter.

Chapter 6 is devoted to the computation of optimal compliance tables as well. Two types of medical response units are considered in this chapter: vehicles with and without transport capability. The last class usually consists of motor cycles, so this type of unit is typically present faster at the emergency scene. This complicates the calculation of compliance tables, as an extra dimension is added to the state space of the EMS system. This chapter presents a integer linear programming formulation for the computation of so-called *two-dimensional compliance tables*. In this model the number of relocations required to be carried out at once is bounded. Moreover, there are also restrictions on the time a specific relocation may last. Subsequently, several regimes are tested for different fleet mixes in a rural region by simulation. This study shows that imposing a time bound that is equal to the expected interarrival times of incidents seems a good choice.

The last chapter of this dissertation, Chapter 7, presents a unified view on the online and offline approaches for solving the ambulance relocation problem. To this end, representants proposed in the previous chapters are chosen. These are the combination of the penalty heuristic and DMEXCLP (Chapter 4) and the compliance tables obtained through solving AMEXPREP (Chapter 5) for the online and offline approach, respectively. Both methods are simulated for different fleet sizes and performance objectives. In this study, the chosen representants show similar performance, both from a patient and a crew perspective.

Samenvatting

In levensbedreigende noodgevallen waarin elke seconde telt, kan het al dan niet op tijd ter plaatse zijn van een ambulance het verschil maken tussen leven en dood. In Nederland bedraagt deze normtijd 15 minuten voor ongevallen van de hoogste urgentieklasse. Om tegen betaalbare tarieven zulke korte aanrijtijden te realiseren is een uitgekiende planning van ambulancediensten noodzakelijk. Een wezenlijk onderdeel hierin is het bepalen van een efficiënte verdeling van ambulances over de regio. Immers, de locatie van de hulpvoertuigen op het moment dat een ongeval wordt gerapporteerd in de meldkamer is in hoge mate bepalend voor het al dan niet behalen van de normtijd. Een sterk complicerende factor bij deze locatievraagstukken is de grote mate van onzekerheid die inherent is aan vrijwel alle facetten van het ambulance service-proces. Met name de onzekerheid rond de beschikbaarheid van hulpvoertuigen is van cruciaal belang. Ambulances zijn niet te allen tijde tijdens hun dienst inzetbaar om te beantwoorden aan een oproep. Dit kan verscheidene oorzaken hebben, bijvoorbeeld de behandeling van de patiënt op de plaats van een eerder ongeval. Deze onbeschikbaarheid kan ertoe leiden dat er tijdelijk een inefficiënte spreiding van ambulances over de regio ontstaat.

Een manier om bovenstaand probleem te verhelpen is het tijdelijk herpositioneren van één of meerdere hulpvoertuigen. Dit is de kern van het zogenaamde Dynamisch Ambulance Management (DAM). Het grote voordeel van DAM is dat door het in real-time verplaatsen van voertuigen veel flexibeler kan worden geanticipeerd op toekomstige incidenten, waardoor de aanrijtijden, en daarmee ook de mortaliteit en morbiditeit, kunnen worden gereduceerd. Dit proefschrift behandelt verschillende modellen en algoritmen voor het optimaliseren van de gebiedsdekking door middel van het uitvoeren van deze zogenaamde proactieve relocaties.

De optimalisatiemethoden in dit proefschrift zijn grofweg in te delen in twee categorieën: (1) *online* en (2) *offline* algoritmen. Het kenmerkende verschil tussen beide is het tijdstip waarop het computationele werk gedaan wordt. Bij online methoden gebeurt dat op het moment waarop een relocatiebeslissing genomen dient te worden. Deze klasse van methoden kan dan ook omgaan met een zeer gedetailleerde toestandsbeschrijving, aangezien bij de berekening van een beslissing dit voor slechts één specifieke toestand gedaan wordt. Bij offline methoden wordt het merendeel van de berekeningen op voorhand gedaan en voor iedere mogelijke toestand van het systeem wordt de bijbehorende relocatiebeslissing opgeslagen. Als een bepaalde situatie (c.q., toestand) zich voordoet, wordt deze beslissing opgezocht en toegepast. Om de rekentijd, die sterk afhangt van het aantal toestanden, hanteerbaar te houden, is de toestandsbeschrijving bij offline methoden doorgaans minder gedetailleerd dan bij online tegenhangers.

Het eerste deel van dit proefschrift (Hoofdstukken 2-4) houdt zich bezig met de online benadering van het ambulance relocatieprobleem. Hoofdstuk 2 richt zich op landelijke regio's. Deze verschillen doorgaans van hun stedelijke tegenhangers, onder andere in het lagere aantal incidenten, in het kleinere aantal ambulances dat dienst heeft en in de spreiding van ongevallen over de regio. Het relocatieprobleem wordt gemodelleerd als Markov-beslissingsprobleem, waarbij het aanrijtijdsafhankelijke prestatiedoel generiek gekozen kan worden door het definiëren van een geschikte boetefunctie. Deze functie kent aan iedere mogelijke aanrijtijd een bepaalde boete toe. Daarnaast beschrijft dit hoofdstuk ook een heuristiek voor goede relocatiebeslissingen, gebaseerd op het beschreven model. Hoofdstuk 2 wordt afgesloten met het illustreren van deze heuristiek op een realistisch ambulance systeem behorend bij een Nederlandse regio, voor verschillende boetefuncties.

Een veelgenoemd nadeel van DAM is dat ambulances vaak proactief moeten verplaatsen, en dat de werkdruk daardoor verhoogt. Hoofdstuk 3 houdt zich bezig met deze problematiek: de afweging tussen het aantal relocaties en de prestatie met betrekking tot de aanrijtijden wordt onderzocht. Hiertoe wordt in dit hoofdstuk de zogenaamde *boeteheuristiek* ontwikkeld. Deze heuristiek gebruikt het concept van boetefunctie uit Hoofdstuk 2. Ook wordt in dit hoofdstuk het gebruik van zogeheten ketenrelocaties geïntroduceerd. Hierbij wordt een langdurende relocatie opgeknipt in meerdere korte relocaties, om de gewenste spreiding van ambulances (berekend door de boeteheuristiek) te verkrijgen. Het hoofdstuk sluit af met een numerieke studie van de bovengenoemde afweging. Op basis van een door praktijkexperts uit het veld aangedragen boetefunctie wordt voor een grote hoeveelheid scenario's onderzocht in welke mate restricties op het aantal ambulancerelocaties van invloed zijn op de prestatie. Deze verbetert al significant als slechts een paar relocaties worden uitgevoerd. Echter, veelvuldige herpositionering kan ertoe leiden dat de prestatie niet meer verbetert, of zelfs verslechtert.

In Hoofdstuk 4 worden enkele inzichten betreffende de implementatie van relocatiestrategieën in de praktijk gepresenteerd. Hiertoe wordt de in Hoofdstuk 3 beschreven boeteheuristiek gecombineerd met het Dynamisch MEXCLP (DMEXCLP) algoritme van Jagtenberg et al. (2015). De volgende vijf aspecten komen hierbij aan bod: (1) de frequentie van beslismomenten, (2) het rekening houden met bezette ambulances in de beschrijving van de toestand van het systeem, (3) het prestatiedoel, (4) het gebruik van ketenrelocaties en (5) tijdsgrenzen op de relocatietijd. Het hoofdstuk vervolgt met een uitgebreide simulatiestudie met betrekking tot de praktische implementatie van deze aspecten.

In het tweede deel van dit proefschrift (Hoofdstukken 5 en 6) staat de offline benadering voor het oplossen van het relocatieprobleem centraal. Hoofdstuk 5 beschrijft een geheeltallig lineair programmeringsprobleem voor de berekening van *schuifregeltabellen*, genaamd MEXPREP. Dit model is een uitbreiding op het MECRP model van Gendreau et al. (2006) in twee verschillende opzichten: (1) het houdt rekening met het feit dat ambulances bezet kunnen raken tijdens het relocatieproces en (2) een vrij te kiezen prestatiedoel kan in het model ingebouwd worden via de definitie van de corresponderende boetefunctie. Daarnaast introduceert dit hoofdstuk een aangepaste versie (genaamd AMEXPREP) die de gedane aannames omtrent de bezettingsgraad afzwakt. Een sectie met numerieke resultaten, gebaseerd op de simulatie van de met MEXPREP verkregen schuifregels, sluit het hoofdstuk af.

Ook Hoofdstuk 6 is gewijd aan de berekening van optimale schuifregeltabellen. In dit hoofdstuk worden twee types ambulances beschouwd: voertuigen met en voertuigen zonder vervoerscapaciteit. De laatste categorie bestaat doorgaans uit motoren of personenauto's, dus dit type eenheid is doorgaans sneller bij een ongeval ter plaatse. Dit bemoeilijkt de berekening van schuifregeltabellen, aangezien een extra dimensie wordt toegevoegd aan de toestandsruimte van het systeem. Dit hoofdstuk presenteert een geheeltallig lineair programmeringsprobleem dat zogeheten tweedimensionale schuifregeltabellen berekent. In dit model wordt het aantal relocaties dat per beslismoment dient te worden uitgevoerd, beperkt. Ook worden er in het model grenzen gesteld aan de tijd die een ambulance mag besteden aan een specifieke relocatie. Verschillende regimes worden vervolgens getest in een simulatie, voor verschillende samenstellingen van de ambulancevloot in een landelijke regio. Uit deze studie blijkt dat een tijdsrestrictie op de relocatieduur die gelijk is aan de verwachte tijdsduur tussen twee opeenvolgende ongevallen, een goede keus is.

Het laatste hoofdstuk van dit proefschrift, Hoofdstuk 7, kan worden beschouwd als het verbindende hoofdstuk tussen de online en offline benaderingen voor het oplossen van het ambulance relocatieprobleem. Van beide benaderingen wordt een representant gekozen uit de voorgaande hoofdstukken. Dit zijn de combinatie van de boeteheuristiek en DMEXCLP (Hoofdstuk 4) en de schuifregeltabellen verkregen door oplossen van AMEXPREP (Hoofdstuk 5) voor respectievelijk de online en offline benadering. Beide methoden worden gesimuleerd voor verschillende ambulanceaantallen en prestatiedoelen, uitgaande van twee Nederlandse regio's. Hieruit blijkt dat de gekozen representanten zeer vergelijkbaar presteren wat betreft zowel patiënt- als personeel-gerelateerde prestatiematen.