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

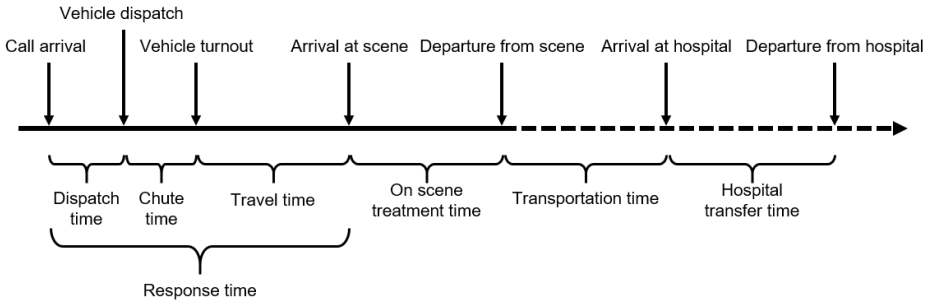


FIGURE 1.1: EMS process.

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.

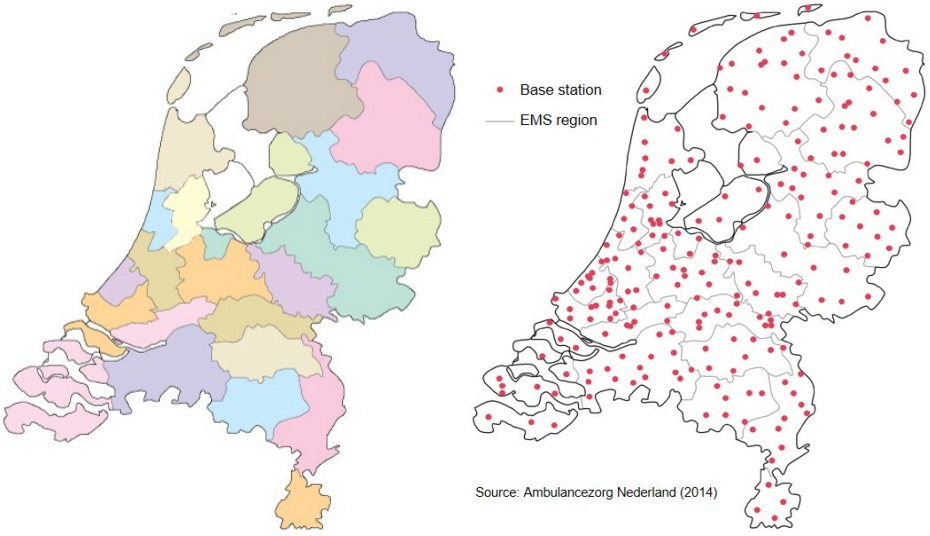
During the transit, one crew member usually continues to provide appropriate medical care, if necessary. Meanwhile, the driver travels to the selected hospital as fast as possible. At the hospital, the ambulance crew unloads the patient and takes her/him to a suitable department, usually the emergency 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).



(A) (B)

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 am-

bulance 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 equipment-

emplacement 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 Gen-

dreau 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.