Logistics of emergency response vehicles
Facility location, routing, and shift scheduling

van den Berg, Pieter

DOI
10.4233/uuid:f87b8985-b856-42d9-bca4-8c5cf3e20d45

Publication date
2016

Document Version
Final published version

Citation (APA)
van den Berg, P. (2016). *Logistics of emergency response vehicles: Facility location, routing, and shift scheduling*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:f87b8985-b856-42d9-bca4-8c5cf3e20d45

Logistics of emergency response vehicles

Facility location, routing, and shift scheduling

Logistics of emergency response vehicles

Facility location, routing, and shift scheduling

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus prof. ir. K.C.A.M. Luyben,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op
maandag 6 juni 2016 om 15:00 uur

door

PIETER LOURENS-JAN VAN DEN BERG

Master of Science in Econometrics and Operational Research,
geboren te Ermelo, Nederland.

This dissertation has been approved by the promotors:

Prof. dr. ir. K.I. Aardal and prof. dr. R.D. van der Mei

Composition of doctoral committee:

Rector Magnificus,	Chairman
Prof. dr. ir. K.I. Aardal,	Delft University of Technology
Prof. dr. R.D. van der Mei,	VU University Amsterdam and CWI

Independent members:

Prof. dr. L.G. Kroon,	Erasmus University Rotterdam
Prof. dr. G.T. Timmer,	VU University Amsterdam
Prof. dr. ir. A.W. Heemink,	Delft University of Technology
Dr. ir. H.N. Post,	Connexxion
Prof. dr. ir. G. Jongbloed,	Delft University of Technology, reserve member

Other member:

Prof. dr. S.G. Henderson,	Cornell University, United States

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)

Printed by: Ipskamp Printing, Enschede, the Netherlands

Cover design: Remco Wetzels | remcowetzels.nl

ISBN 978-94-6186-655-4

An electronic version of this dissertation is available at
`http://repository.tudelft.nl/`.

Contents

Summary

This thesis discusses different aspects of the logistics of emergency response vehicles. In most parts, we consider providers of ambulance care in the Netherlands. However, also firefighters and air ambulance providers in both Canada and Norway are considered. Even though significant differences exist between the considered systems, they share the task of providing adequate service in emergency situations. In these situations, a prompt response is important and this importance is typically expressed by a response time target set by law. For most emergency services, providing the appropriate care within this target in an efficient way is the main objective. This thesis uses optimization techniques to handle three aspects of the logistical process: facility location, routing, and shift scheduling. All three can have a significant impact on the performance of the system.

The first part of this thesis deals with the location of facilities. In the case of emergency responders, this concerns the location of the bases and the distribution of the vehicles over the selected bases. As the travel time is the most important component of the response time, the location of the vehicles is of utmost importance. Before introducing new models in Chapters 3-5, two experiments that help the modeling process are discussed in Chapter 2. First, six basic location models are compared on a wide range of criteria arising from practice in a computational experiment. The results give an indication of the suitability of the different models as a building-block for more advanced models. Second, the impact of different levels of data aggregation is evaluated. The results of a commonly used aggregation level are compared with a significantly more detailed level. The results show that for high coverage levels, the aggregation of data strongly influences the results. A method to partly overcome this problem without the need for detailed travel time data is proposed and evaluated.

In Chapters 3-5, new models for the location of base stations are introduced. The first two focus on ambulance providers, whereas the third is specifically designed for firefighter systems. Chapter 3 extends on the classical Maximum Expected Covering Location Problem (Daskin, 1983) by incorporating fluctuation in the system characteristics throughout the day. By penalizing the number

of base locations and ambulance relocations, we obtain more practical solutions. Chapter 4 considers the case where the coverage provided by a given base station is fractional rather than 0-1 valued. An Integer Linear Programming formulation is presented that can solve significantly larger instances than the nonlinear formulation by Ingolfsson et al. (2008). In Chapter 5, a location model for a firefighter department is introduced. The model incorporates firefighter-specific features, such as location-dependent response time targets, multiple vehicle types, and voluntary crews. The model is applied to the region of Amsterdam and the results show significant potential for improvement. Finally, Chapter 9 considers the location of emergency helicopters in Norway, where a very rural area must be covered. The basic Maximum Covering Location Problem (Church and ReVelle, 1974) is applied to propose changes to the current set of base locations.

The routing of vehicles is discussed in Chapters 6 and 8. Chapter 6 deals with the non-urgent transportation of patients, where patients must be transported between health care facilities. For these transportations, specific BLS ambulances are available. However, their capacity does typically not suffice to serve all calls, in which case a regular ALS ambulance is used. As these ambulances provide coverage for emergency calls, this can lead to a decrease in emergency coverage. A model is presented that finds routes for the BLS ambulances that minimize the impact on the emergency calls. In Chapter 8, a simulation model for the air ambulance service in Ontario, Canada is used to evaluate the impact of different routing policies. This simulation tool is later also used to evaluate alternative shift schedules.

In Chapters 7 and 8, we analyze the shift schedules of the vehicles. In Chapter 7, the model of Chapter 6 is used to evaluate the current schedule for the BLS ambulances. Based on these results, new schedules are defined that are again evaluated with the model. The results show that by changing the schedule, the performance can be increased without any additional capacity. In Chapter 8, a model is introduced to find a schedule that deviates for the 24 hour, flat schedule currently in use by the air ambulance service in Ontario. A limited number of shifts can be removed without a significant impact on the performance. However, due to the enormous area that must be covered, most of the capacity required during daytime is also necessary during the night.

The fact that many of the models presented in this thesis have (in slightly adapted form) been used to give concrete advice to emergency providers shows the potential of applying optimization techniques to emergency response services. However, this can only results in a real impact if the practitioners are closely involved in the process. Their involvement has therefore played a vital role in the realization of this PhD thesis.

Samenvatting

Dit proefschrift behandelt verschillende aspecten van de logistiek van aanbieders van spoedeisende hulp. Het grootste deel van het proefschrift beschouwt de ambulancezorg in Nederland. In andere delen worden ook de brandweer in Nederland en de luchtambulance in Noorwegen en Canada besproken. Ondanks dat er duidelijke verschillen tussen de beschouwde systemen zijn, delen ze de taak om snelle hulp te bieden in geval van ongevallen. In deze spoedeisende situaties is het van groot belang dat de hulpdiensten snel ter plaatse zijn. Dit belang wordt onderstreept door een wettelijke norm op de responstijd. Het op een efficiënte manier leveren van de juiste zorg binnen de gestelde norm is voor de meeste aanbieders de belangrijkste doelstelling. In dit proefschrift gebruiken we optimalisatietechnieken om drie aspecten van het logistieke proces te optimaliseren: (1) de locatie van standplaatsen, (2) de routering van voertuigen, en (3) de dienstroosters. Alle drie de aspecten kunnen invloed hebben op de geleverde kwaliteit van zorg.

Het eerste deel houdt zich bezig met het bepalen van standplaatsen en de verdeling van voertuigen over de geselecteerde standplaatsen. Gezien de rijtijd de belangrijkste component van de responstijd is, heeft de locatie van de voertuigen een significante impact op de dekking. Alvorens nieuwe modellen te introduceren in Hoofdstukken 3-5, worden in Hoofdstuk 2 twee experimenten behandeld waarvan de resultaten helpen in het modelleringsproces. Allereerst vergelijken we de uitkomsten van zes standaard locatiemodellen op verschillende criteria vanuit de praktijk. De resultaten hiervan laten zien welke van deze modellen geschikt zijn als basis voor meer geavanceerde modellen. Vervolgens analyseren we de impact van verschillende niveaus van data aggregatie. De resultaten op basis van het meest gebruikte aggregatieniveau worden vergeleken met de resultaten op basis van aanzienlijk gedetailleerdere data. Zeker in het geval dat hoge dekkingspercentages behaald kunnen worden zijn de verschillen groot. Om dit probleem gedeeltelijk te ondervangen stellen we een alternatieve aanpak voor waarvoor geen gedetailleerdere rijtijdendata nodig is.

In Hoofdstukken 3-5 introduceren we verschillende nieuwe modellen voor het bepalen van standplaatsen en de verdeling van de voertuigen over de standplaat-

sen. De eerste twee hoofdstukken richten zich op aanbieders van ambulancezorg, terwijl Hoofdstuk 5 speciaal ontwikkeld is voor de brandweer. Hoofdstuk 3 biedt een uitbreiding op het veel-bestudeerde MEXCLP (Daskin, 1983). Hierbij wordt rekening gehouden met de wijzigende eigenschappen van het systeem gedurende dag. Door een boete in te stellen op het aantal standplaatsen en het aantal relocaties van ambulance krijgen we realistischere resultaten. Hoofdstuk 4 beschouwt het geval dat de dekking geboden door een ambulance op een bepaalde standplaats fractioneel is, in plaats van louter 0 of 1. We introduceren een Geheeltallig Lineair Programmeringsformulering voor dit probleem waarmee we in staat zijn aanzienlijk grotere instanties op te lossen dan de niet-lineaire formulering van Ingolfsson et al. (2008). Het in Hoofstuk 5 geïntroduceerde model is specifiek ontwikkeld voor de brandweer. Het model bevat verschillende brandweer-specifieke eigenschappen zoals vraagpuntafhankelijke normtijden, verschillende voertuigtypen, en vrijwillige crew. Ten slotte wordt in Hoofdstuk 9 het MCLP model (Church and ReVelle, 1974) toegepast om de helikopterstandplaatsen in Noorwegen te optimaliseren.

De routering van voertuigen wordt besproken in Hoofdstuk 6 en Hoofdstuk 8. Hoofdstuk 6 behandelt het niet-spoedeisende en geplande vervoer van patiënten tussen zorginstellingen. Voor deze ritten zijn speciale zorgambulances beschikbaar. Over het algemeen hebben deze echter niet voldoende capaciteit om al het geplande vervoer uit te voeren. De resterende ritten moeten worden uitgevoerd met de reguliere ambulances, waardoor deze niet beschikbaar zijn voor spoedvervoer. Dit hoofdstuk bespreekt een model dat routes genereert voor de zorgambulances zodanig dat de impact op het spoedvervoer geminimaliseerd wordt. In Hoofdstuk 8 gebruiken we een simulatiemodel voor de luchtambulance in Ontario, Canada om verschillende routeringskeuzes te analyseren.

In Hoofdstuk 7 en Hoofdstuk 8 analyseren we de roosters voor de beschikbare voertuigen. Hoofdstuk 7 gebruikt het model van Hoofdstuk 6 om het huidige rooster voor de zorgambulances te evalueren. Op basis van de resultaten worden verschillende alternatieve roosters voorgesteld en vervolgens geëvalueerd met behulp van het model. Met kleine wijzigingen aan het rooster blijkt het mogelijk om betere kwaliteit te leveren zonder capaciteit toe te voegen. Hoofdstuk 8 bespreekt een model om roosters voor de Canadese luchtambulance aanbieder te vinden die afwijken van het huidige rooster waarin gedurende de hele dag dezelfde capaciteit beschikbaar is. Het aantal diensten per dag kan licht verlaagd worden zonder dat het serviceniveau significant verslechtert. Door de grootte van het gebied is echter een groot deel van de capaciteit die overdag nodig is ook nodig gedurende de nacht.

Het feit dat veel van de in dit proefschrift besproken modellen zijn gebruikt om tot concrete adviezen voor aanbieders van spoedvervoer te komen geeft het potentieel van optimalisatietechnieken aan. Het is echter onmogelijk om een werkelijke toegevoegde waarde te hebben zonder de nadrukkelijke betrokkenheid vanuit de sector. De samenwerking met de verschillende eindgebruikers is dan ook van cruciale waarde voor de totstandkoming van de resultaten in dit proefschrift gebleken.

1

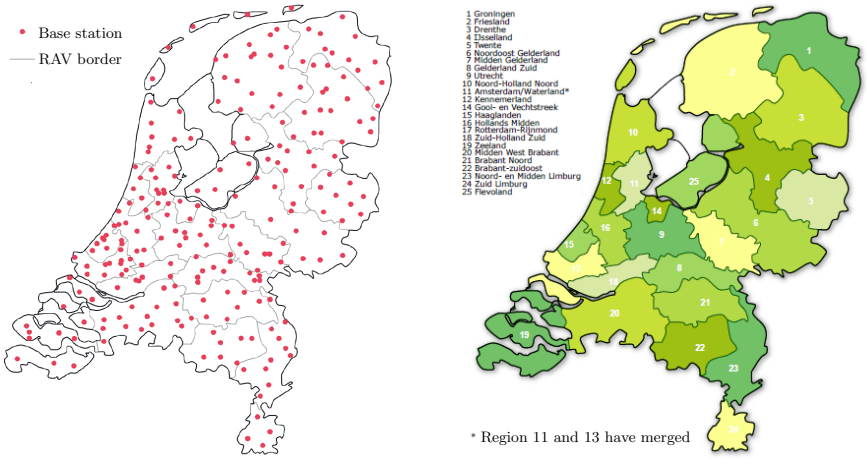
Introduction

Emergency Medical Service (EMS) providers are responsible for providing adequate care for emergency patients outside health care facilities. When a call arrives at the dispatch center, an ambulance is assigned to the call. The probability of survival of a patient that is involved in an accident is highly dependent on the time it takes for the ambulance to arrive at the scene (Larsen et al., 1993). As a consequence, prompt responses by EMS vehicles are of utmost importance. In most countries, this importance is stressed by a legal response time target in which a minimum fraction of calls must be reached. For example, in England, 75 percent of the most urgent calls must be reached within 8 minutes. Reaching this target requires all involved parties to efficiently allocate their resources. The research that resulted in this thesis was conducted within a larger research project called REPRO (from REactive to PROactive planning of ambulance services). In close cooperation with ambulance providers in the Netherlands, mathematical models were developed to improve the service provided with the available resources.

As in this thesis we mainly consider the Dutch EMS system, we will start by introducing the Dutch EMS system in Section 1.1. Even though most EMS systems have similar characteristics, there can be some differences between countries. For a survey among ten European countries regarding the structure of ambulance care, we refer to Hoogeveen (2010). After describing the structure of the Dutch EMS system, we will introduce a typical response process, where we define the different stages of the process. The definitions introduced here will be used throughout this thesis. In Section 1.2, we briefly introduce some of the Operations Research methodology that is used in this thesis. For a more comprehensive overview, we will refer to existing literature. Section 1.3 gives an overview of the planning problems that arise at EMS providers and gives an overview of the available literature for each of the planning problems. For planning problems that are addressed in this thesis, a more detailed overview of the relevant literature is given in the corresponding chapters. We conclude this chapter by giving an outline of the remainder of this thesis.

1.1 Ambulance care in the Netherlands

As in most countries, the ambulance providers in the Netherlands cover two main types of calls - emergency calls and patient transportations. Emergency calls are unscheduled calls for which an ambulance has to be sent as quickly as possible. In the Netherlands, two types of emergency calls are distinguished: A1 and A2 calls. A1 calls are the most urgent, life-threatening calls. In this case, an ambulance is required to be at the scene within 15 minutes in 95 percent of the cases. For urgent, but not life-threatening calls, an ambulance should be there within 30 minutes. Patient transportations, which are called B calls in the Netherlands, encompass the non-urgent and scheduled transportation of patients between health care facilities. Within the patient transportations, sometimes two categories are distinguished: B1 and B2. Here, B1 calls involve the transportation of patients in critical conditions, for which a fully equipped ambulance is required. For B2 calls, a less equipped ambulance suffices. In 2014, approximately 1.2 million calls were served in the Netherlands, of which 580,000 were A1, 290,000 were A2, and 320,000 were patient transportations. Besides the emergency calls and the patient transportations, some ambulance providers further assist in acute home care during the nights. In total, 755 ambulances were in use in the Netherlands in 2014. These were divided over 231 base stations (Ambulancezorg Nederland, 2014). The current locations of the base stations are shown in Figure 1.1a.



(a) *The current location of the base stations (Ambulancezorg Nederland, 2014).*

(b) *The 24 ambulance regions in the Netherlands.*

Fig. 1.1: *Current location of the base stations and ambulance regions.*

The ambulance care in the Netherlands is divided into 24 more or less independently operating regions (see Figure 1.1b), which in Dutch are called Regionale Ambulancevoorzieningen (RAVs). Each of them is operated by a single organization. The Dutch institute of public health and the environment (RIVM) computes the required capacity for each of the RAVs. These computations form the basis for the budgets of the RAVs. Although the budget is based on a number of ambulances and a number of bases, the RAVs are free to choose how to spend their budget. Every year, the branch organization of ambulance care in the Netherlands (AZN) publishes the performance of the different regions. The AZN further organizes the training for the ambulance crew.

Originally, every ambulance region had its own call center from which calls were taken and ambulances were dispatched. However, in the past years, multiple call centers have merged in order to improve efficiency. Currently, there are 19 call centers in the Netherlands. This will further reduce to ten in the coming years. Most call centers distinguish call takers and dispatchers. Call takers are responsible for triage, while dispatchers assign ambulances and instruct the ambulance crew. In principle, only the call taker requires medical training. However, in most call centers, call takers and dispatchers switch roles during the day, in which case both call takers and dispatchers are required to have a medical education.

In the Netherlands, all emergency calls that require transportation of a patient are served by an Advanced Life Support (ALS) ambulance. These ambulances are fully equipped and staffed by a paramedic and a driver. The paramedic is required to have completed a full nursing education and at least one follow-up course in acute care. Additionally, specific training is given by the AZN at the moment of hiring. The driver, on the other hand, does not need to have a medical background. The driver is there to assist the paramedic at the scene, for this the driver gets training in providing medical assistance. Furthermore, training in driving an ambulance is required.

In patient transportations, called B calls, two categories are distinguished depending on the medical conditions of the patient. In case life-threatening situations might occur during the transportation, an ALS ambulance is required. These calls are called B1. All other transportations, B2 calls, may also be executed by a Basic Life Support (BLS) ambulance. This is a less equipped ambulance staffed by two regular nurses.

In particular cases, other vehicles may be used. For patients that need transportation between the intensive care units of two hospitals, Mobile Intensive Care Units (MICU) are used. For children and newborns that require intensive care during transportation, a Pediatric Intensive Care Unit (PICU) or a Neonatal Intensive Care Unit (NICU) is available. Some regions additionally use rapid responders. This is a single paramedic that can provide care at the scene, but cannot transfer a patient to a hospital. The paramedic uses, for example, a car, motorbike, or even a normal bike to get to the scene. In the most severe cases, an additional Mobile Medical Team (MMT) is sent to the scene. Such a team consists of a medical specialist, a specialized nurse and a driver or pilot.

In the Netherlands, there are four MMTs located in Amsterdam, Rotterdam, Groningen and Nijmegen. The teams can use a car or a helicopter to get to the scene.

As in most countries, an ambulance can serve only one patient at a time. For both A1 and A2 calls, regular practice is to always send the closest available ambulance. Since all ambulances are equipped with GPS trackers, call center software shows the closest available ambulance to the dispatcher. If necessary, ambulance relocations are performed in order to maintain good coverage throughout the region. Often, these relocation decisions are based on so-called look-up tables. Some regions have specific locations where they can temporarily locate an ambulance as part of a relocation, for example, in the middle between two regular base stations.

1.1.1 Response process

Although there are small differences between countries, the structure of the response process is similar. However, often different terminology is used, which can result in misunderstanding. To avoid this confusion, we introduce a typical response process with the terminology that we use in this thesis. When a call arrives at the dispatch center, it takes some time for the dispatcher to assess the urgency of the call and assign an ambulance. This process is called triage and dispatch. The time between the assignment of an ambulance and the moment it starts driving is called the chute time. Together, the triage and dispatch, and the chute time accumulate to the pre-trip delay. Adding the travel time to the pre-trip delay gives the response time, which is the main performance measure for EMS providers. In some countries, the definition of response time is slightly different. In England, for example, the clock starts ticking up to 60 seconds later for serious but less immediately time critical incidents, than for cases where patients are in immediately life-threatening conditions. After spending some time on the scene with the patient, the patient might require transportation to a hospital. In that case, the ambulance becomes idle after dropping off the patient at the hospital. In many countries, congested Emergency Departments (EDs) result in long turnaround times (the time it takes the ambulance crew to hand over the patient and restock the vehicle so it is ready to attend another call), which can have an enormous impact on the performance of EMS systems (Channouf et al., 2007). Whenever an ambulance finishes a call, it is available for new calls. If no new calls are waiting, the ambulance returns to its base, or any other location where it waits for a new call. Figure 1.2 shows the different stages of the response process.

Even though there can be significant differences between countries, or even between EMS regions, we give some statistics to get insight into the duration of the components of the response time. In the Netherlands, the average time for dispatch and triage for the 24 different regions ranges from 1:07 minutes up to 2:36 minutes, with an average of 1:48 minutes. The average chute time is 0:56 minutes. This implies that the average pre-trip delay is almost three minutes,

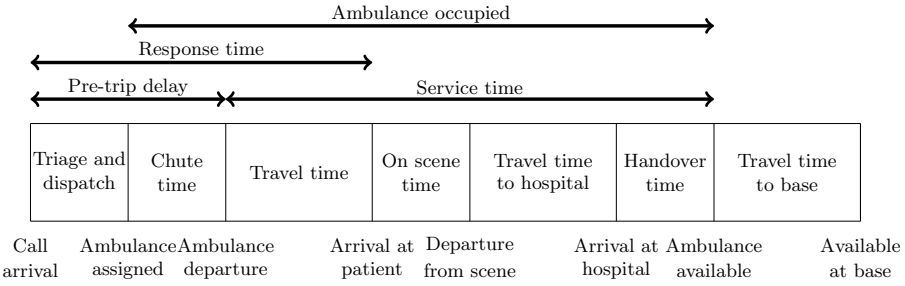


Fig. 1.2: *Overview of the different stages of the response process.*

which allows for a travel time of at most 12 minutes. The average response time in 2014 was 9:29 minutes and in 93.4% of the cases an ambulance was available within the target response time of 15 minutes. The target that is set by law to reach 95 percent of the life-threatening calls within 15 minutes has been reached by 7 of the 24 EMS regions (Ambulancezorg Nederland, 2014).

1.2 Brief introduction to Operations Research

Most of the techniques used in this thesis come from the area of Operations Research. This research field uses analytical models to assist in decision making. It is often considered a sub-discipline of mathematics. Two aspects of Operations Research that are used in this thesis are the modeling of problems arising from practice and finding the solutions to the resulting optimization problems. As the best formulation of a problem depends on the solution method used, we will introduce some solution techniques for commonly used optimization problems. Before that, we will give a very brief introduction to complexity theory. This field classifies optimization problem based on the complexity of solving the problem. Finally, we will discuss techniques for formulating problems in a framework that can be dealt with by the solution approaches. This section is not intended to give a complete introduction to Operations Research. Instead, it introduces the main ideas in an informal way. For a more precise and complete introduction, we refer to Papadimitriou and Steiglitz (1982) and Hillier and Lieberman (2014), for example.

1.2.1 Complexity theory

Complexity theory studies the complexity of solving optimization problems. When studying solution methods, a strict distinction between an optimization problem and an instance of the problem is made. The problem is the general formulation of the question under study without considering a particular case of the problem. One could, for example, study the problem of finding the shortest

path between two nodes in a road network. Once we include the data of a particular example of this problem, we obtain an instance of the problem. In case of finding the shortest path, an instance could be to find the shortest distance to get from Delft to Amsterdam using the Dutch road network. The concept of a solution is defined differently for problems and instances. Solving a problem requires a solution approach that can be used for every instance of the considered problem, whereas solving an instance just requires one solution. Getting back to the problem of finding a shortest path, a solution to the instance of shortest path is a set of directions to guide you from Delft to Amsterdam in the shortest way. Finding a solution to the shortest path problem, however, requires an algorithm that can find the shortest path between any two nodes in any road network. Even though we, from a practical point of view, are mainly interested in solving instances of problems, the general solution approach for solving a problem is typically used for solving instances. For that reason, the field of complexity theory studies the complexity of solving optimization problems, rather than instances of optimization problems.

In complexity theory, we distinguish optimization problems and decision problems. An optimization problem is the problem of finding the best solution within a set of feasible solutions. In a decision problem, the goal is to answer a "yes"/"no" question. In the example of the shortest path problem, the optimization problem searches the shortest path between two nodes. In the decision problem, the questions is whether there exists a path between two nodes of length at most L. It is easy to see that solving an optimization problem is at least as hard as solving the corresponding decision problem.

A solution procedure for an optimization or decision problem is defined as a procedure to solve any instance of the considered problem. The efficiency of a solution procedure is defined as the worst-case running time of the procedure as a function of the size of the input. To avoid excessive running times, an algorithm is considered efficient if the worst-case running time is bounded by a polynomial function of the input size. Algorithms with a running time that cannot be bounded by any polynomial function are considered inefficient. To classify problems based on their complexity, the problem classes P and NP are defined. The class P, standing for polynomial time, contains all problems for which a polynomial time algorithm exists. The class NP, which stands for nondeterministic polynomial time, is the class of all decision problems for which for every "yes"-instance there exists a certificate that can be verified in polynomial time. For example, the decision version of the shortest path problem belongs to the class NP, as a "yes"-answer can be verified in polynomial time by giving the path as a certificate. As Dijkstra's algorithm (Dijkstra, 1959) solves the shortest path problem in $O(|V|^2)$, where $|V|$ is the number of nodes in the network, the problem also belongs to the class P. Since solving a problem requires the verification of the solution, we have that $\mathrm{P} \subseteq \mathrm{NP}$. Even though it is widely believed that $\mathrm{P} \neq \mathrm{NP}$, this is still an open question. The Clay Mathematical Institute of Cambridge included this problem in the list of seven of the most significant open problems in mathematics at the beginning of the third millennium.

The set of decision problems that is at least as hard as any problem in NP is called NP-complete. More precisely, a problem Π is NP-complete if it belongs to the class P and every problem in the class NP can in polynomial time be reduced to Π. A polynomial time algorithm for any NP-complete problem would imply that P = NP. As all problems in NP can be reduced to any NP-complete problem, one could use the polynomial time algorithm to solve any problem in NP, implying P = NP. As we are typically interested in optimization problems rather than decision problems, we introduce the complexity class NP-hard that contains all optimization problems for which the decision version is NP-complete. As solving an optimization problem is at least as hard as solving the corresponding decision problem, we have that no polynomial time algorithm exists for any NP-hard problem, unless P = NP.

As there is limited hope for a polynomial time algorithm for any NP-hard problem, one should consider alternative solution approaches. The first option is to relax the constraint on the computation time. In that case, one would settle for a solution procedure that provides the optimal solution, but has exponential running time in the worst-case. For instances of limited size, this might still be acceptable. Alternatively, one could relax the optimality condition and settle for a reasonably good solution that can be found efficiently. Here, two classes of solution approaches can be distinguished: approximation algorithms and heuristics. In the first case, the algorithm must run in polynomial time and the worst-case gap between the value of the optimal solution and the provided solution must be bounded. For example, a 2-approximation algorithm for a minimization problem runs in polynomial time and returns a solution with an objective value within a factor 2 of the optimal value. A heuristic is an algorithm that typically gives a solution quickly, but does not provide any performance guarantee. Despite the fact that heuristics do not provide a performance guarantee, they are often used in practice. Even though some general frameworks are available, the design of approximation algorithms and heuristics is highly problem-specific.

In this thesis, we often formulate problems as an Integer Linear Programming Problem. This is a very general optimization problem that can be used to formulate a wide variety of optimization problems. The general form of Integer Linear Programming is NP-hard and thus no polynomial time algorithm is known. As this thesis mainly deals with strategic and tactical decision problems, we typically use the first approach for dealing with NP-hard problems where the running time is not polynomial, but the solution is guaranteed to be optimal with respect to the selected model. For this, commercial solvers like CPLEX (ILOG, 2013) and Gurobi (Gurobi Optimization, 2015) are available to solve reasonably large instances. As most problems can be formulated in different ways and the computation time is highly dependent on the formulation, we discuss some of the methods that are used by these solvers.

1.2.2 Integer Linear Programming

One of the most general NP-hard optimization problems is Integer Linear Programming (ILP). Here, the objective is to minimize (or maximize) a given linear function over a feasible region defined by a set of linear constraints and integrality constraints. The standard form ILP problem is defined as follows:

$$\begin{aligned} \min \quad & z_{IP} = c^T x, \\ \text{s.t.} \quad & Ax = b, \\ & x \geq 0, x \text{ integer}. \end{aligned}$$

Here, the input consists of an n-dimensional vector c, an m-dimensional vector b, and an $m \times n$ matrix A. The vector x consists of the n decision variables. As this problem is NP-hard, we use a non-polynomial time solution method to solve ILP instances. Many of these solution methods make use of the LP relaxation of the problem. This is the problem obtained by removing the integrality constraint. For this class of problems, called Linear Programming (LP) problems, Dantzig (1951) developed the simplex method. Despite the fact that this method is not known to run in polynomial time, very large instances can be solved in reasonable time by this method. Later, Khachiyan (1979) showed that LP is in the class P by showing that the ellipsoid method solves LP in polynomial time. As every feasible solution for ILP is feasible for the corresponding LP relaxation, we have that any solution to the LP relaxation provides a lower bound on the optimal solution of the ILP problem. In particular, we have that $z_{LP} \leq z_{IP}$, where z_{LP} is the value of the optimal solution of the LP relaxation. To measure the quality of the lower bound provided by the LP relaxation, we define the integrality gap (IG) as $\sup_I \frac{z_{IP}(I)}{z_{LP}(I)}$, for minimization problems. Here, $z_{IP}(I)$ and $z_{LP}(I)$ correspond to a particular instance I. If the objective is to maximize the objective function, the ratio is reversed. Consequently, we have that IG ≥ 1. If the optimal solution of the LP relaxation is integral, then $z_{IP} = z_{LP}$ and thus IG $= 1$.

One solution method for ILP that highly depends on the fact that the LP relaxation can be solved efficiently is Branch-and-Bound. In this solution approach, first the LP relaxation is solved. Then, if at least one decision variable has fractional value, the problem is divided into two subproblems, each with an additional constraint on the value of the fractional variable. Suppose the variable x_k has value $f \notin \mathbb{Z}$ in the LP relaxation. Then, the first subproblem is defined as the original problem with additional constraint $x_k \leq \lfloor f \rfloor$. The second subproblem has $x_k \geq \lceil f \rceil$ as additional constraint (see Figure 1.3). As no integer solutions are excluded, the best solution to any of the two subproblems gives the optimal solution to the original problem. Now, the same procedure is applied to the two subproblems. To avoid complete enumeration, Branch-and-Bound has three ways of pruning subproblems: (1) prune by infeasibility, (2) prune by integrality, and (3) prune by bound. We prune by infeasibility if the LP relaxation is infeasible, in which case no integer feasible solution exists either. Second, if the optimal solution to the LP relaxation of the subproblem is integral, there is no

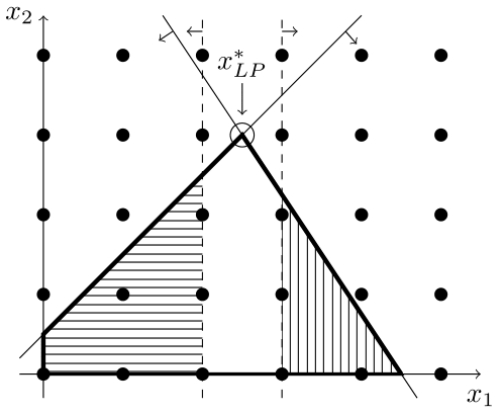


Fig. 1.3: *Example of branching in the Branch-and-Bound algorithm. Thick lines indicate the feasible region of the LP relaxation. x^*_{LP} is the optimal solution of the LP relaxation. As x_1 has fractional value, the algorithm introduces two subproblems with an additional constraint on the value of x_1. The horizontally hatched area is the feasible region of the first subproblem. The vertically hatched area is the feasible region of the second subproblem.*

reason to branch any further. The solution to the LP relaxation is the optimal solution for this subproblem and is feasible for the original problem and thus provides an upper bound on the value of the optimal solution of the original problem. The final reason for pruning a subproblem is that the value of the optimal solution of the LP relaxation guarantees that no improving solution for the original problem exists in this subproblem. For this, the algorithm records bounds on the optimal solution. Every feasible solution for the original problem provides an upper bound. As stated before, the optimal solution of the LP relaxation gives a lower bound for the considered subproblem. If the upper bound is smaller than or equal to the lower bound, the subproblem will not result in a better solution than the best solution found so far. And thus, the subproblem can be pruned. This final way of pruning subproblems can highly influence the running time of the Branch-and-Bound algorithm. For this reason, it is important to have a strong LP relaxation, i.e., an integrality gap close to one. This can, for example, be achieved by means of valid inequalities.

A valid inequality is an inequality that is satisfied for all feasible solutions of the ILP. Hence, including this inequality as a constraint in the ILP does not change the optimal solution. However, since the inequality might not be satisfied by all feasible solutions of the LP relaxation, the solution to this problem might change. Consequently, we get a stronger formulation and a smaller integrality gap by adding this constraint. Adding valid inequalities can therefore speed-up

the Branch-and-Bound procedure. There are roughly two types of methods to find valid inequalities. The first one, introduced independently by Gomory (1958, 1960, 1963) and Chvátal (1973), iteratively finds cuts that separate the optimal solution of the LP relaxation from the true feasible region. By adding this cut to the problem, we obtain a stronger LP relaxation. Alternatively, one could consider the problem at hand and search for inequalities that must be satisfied by any feasible solution. Typically, there are many formulations that result in the same feasible set, but that result in different LP relaxations. As this can influence the running time of solution procedures as Branch-and-Bound, it is important to consider the strength of the formulation in the modeling process.

1.2.3 Formulating Integer Linear Programming problems

Besides the fact that problems can be formulated in many different ways, it can also occur that it is not immediately clear how to model certain constraints in the ILP framework. In this section, we describe some modeling tricks to formulate certain constraints arising from practice. It is often useful to use binary variables that can only take value zero or one. Suppose we have binary variables x_i and x_j indicating whether product i and j are produced in a certain production plan and we want to enforce that product i can only be produced if product j is produced as well. Then, we can add the constraint

$$x_i \leq x_j,$$

which forces x_i to be zero if $x_j = 0$. To model integer variables that are restricted to a discrete set of values, we can use binary variables in the formulation. We include a binary variable for each possible value, indicating whether the variable takes that particular value. For example, if x_i can either take value 5, 8, 11, or 13, we add four binary variables y_1, y_2, y_3, and y_4 and add the following constraints:

$$x_i = 5y_1 + 8y_2 + 11y_3 + 13y_4,$$
$$y_1 + y_2 + y_3 + y_4 = 1.$$

In some applications, one might want to use the product of two binary variables, say x_i and x_j. In general, multiplication of variables does not fit the ILP framework, but for the product of binary variables it is possible to formulate it linearly. Let y be a binary variable that replaces the product of x_i and x_j. By adding the following three constraints, y is forced to take the value $x_i x_j$:

$$y \leq x_i,$$
$$y \leq x_j,$$
$$y \geq x_i + x_j - 1.$$

In case we have a discontinuous variable x that can either take value zero or any value between a lower bound l and an upper bound u, we can use a binary

variable y to indicate whether x takes a value larger than zero. Then, by adding the constraints

$$x \geq ly,$$
$$x \leq uy,$$

we ensure the right value for x. In many applications of Integer Linear Programming, decisions are made on whether a certain product is produced or not. Typically, such a decision involves a high start-up cost and a cost term that is proportional to the production. The cost function $c(x)$ is of the form

$$c(x) = \begin{cases} 0 & \text{if } x = 0, \\ k + cx & \text{if } x > 0, \end{cases}$$

where k is the start-up cost and c the per-unit cost. The standard approach is to use so-called big-M constraints. Here, a binary variable y indicates whether $x > 0$. The objective function then becomes

$$c(x) = ky + cx,$$

and the constraint

$$x \leq My,$$

is added, where M is a sufficiently large constant. In this context, sufficiently large means that M should be larger than the largest value that x can take in any feasible solution. Even though these big-M constraints are very useful in many applications, there is a clear downside to their usage. As M typically has to take values larger than the optimal value of x, the big-M constraints lead to a very weak LP relaxation. Since this can have a strong impact on the computation time, it is wise to carefully select the appropriate value of M. A too large value results in a weak LP relaxation, whereas a too small value restricts the feasible region and can lead to suboptimal solutions.

1.3 Literature review

This section gives an overview of the planning problems that arise in EMS systems. The different problems can be characterized by three different time horizons. At the strategic planning level, decisions are made for a time horizon of several years or even decades. Decisions include, for example, the construction or purchase of buildings for the dispatch centers or ambulance base stations. Hiring crew is part of this planning level as well, as contracts are typically envisioned for several years. Additionally, also the design of an EMS system and the division into EMS regions can be seen as a strategic decision. Decisions at the tactical level typically have an impact of one month up to one year. Problems at this

level include the number of ambulances at each base and the staffing levels of the crew. Problems at the tactical level are often solved simultaneously with some problems at the strategic level. At the operational level, short-term or even real-time decisions are made. This includes, for example, the dynamic relocation of ambulances and handling of unavailability of staff and vehicles. In the remainder of this section, we introduce the different planning problems and give a brief overview of the available literature. For the problems that are addressed in this thesis, additional literature is reviewed in the corresponding chapter.

1.3.1 Demand forecasting

In order to develop effective EMS deployment strategies, it is essential to first assimilate accurate predictions of demand per time period. However, despite the potential of advanced statistical models to offer accurate demand forecasts, most ambulance service providers still use rudimentary prediction methods. Typically, these methods involve dividing the week into 168 one hour increments, accumulating historical records of service requests and evaluating the number of calls received during each hour of the week (Matteson et al., 2011). In Wales, for example, the highest value for each hour of each day in each 10 week period in the previous 50 weeks is observed and the average of these is selected as the 'average peak demand value'. Then, the number of ambulances deployed for this hour in future weeks is based on the concept that there must be a sufficient number to cope with such demand. In Germany, however, a more complex risk-based method by Behrendt and Schmiedel (2002) is often used to determine the maximum number of ambulances needed. A Poisson distribution is used to determine a minimum number of ambulances that is required to have a probability of more simultaneous calls than the number of available ambulances below a certain threshold.

The demand for EMS has the characteristics of the essentially random occurrence of individual calls with seasonal patterns (Vile et al., 2015) and an underlying increase over the past 20 years (Lowthian et al., 2011). Several different methods to account for such fluctuations have been suggested, including linear, sinusoidal and support vector regression (Chen et al., 2015). Furthermore, simple moving averages and more complex time series approaches that allow inclusion of neighboring hours in the forecast are introduced by Matteson et al. (2011) and Baker and Fitzpatrick (1986). Integrated solutions have also been presented that both estimate ambulance demand and recommend deployment plans using queuing theory, simulation models and theoretical distributions (Bell and Allen, 1969; Larson, 1974; Rajagopalan, 2011).

Since the late 1980's, classical time series models such as Autoregressive Integrated Moving Average (ARIMA) and Holt-Winters methods have been used extensively to forecast call volumes (Bianci et al., 1993; Andrews and Cunningham, 1995; Holcomb and Sharpe, 2007) and specifically applied to ambulance demand in Channouf et al. (2007). These models, however, require restrictive data assumptions. Vile et al. (2012) have considered the potential of Singular

Spectrum Analysis (SSA) to produce accurate forecasts whilst adequately accounting for non-stationarity. They show that it considerably outperforms traditional methods for long-term forecasts and offers at least comparable forecasts for a short-term planning horizon. Artificial Neural Networks (ANNs) have also been demonstrated to be capable of producing accurate forecasts for small areas by Setzler et al. (2009).

1.3.2 Workload and service time forecasting

Whilst a large number of models have been developed to better predict demand for EMS, the most comprehensive mathematical models of EMS systems also take into account how response times and workload are expected to fluctuate over time (Ingolfsson, 2013). The relationship between these components is extremely complex, but the detailed call logs now standardly collected by most modern day EMS providers have supplied researchers with a wealth of historical data to analyze.

Each EMS call has an associated response and service time (see Figure 1.2), which are important for different reasons: the response time is the most common indicator of the quality of service provided by an EMS provider and the service times determine the workload on the EMS system. The travel time is usually the largest component of the response time (Ingolfsson, 2013) and has thus not surprisingly been found to be one of the main factors to influence overall system quality. Hence, by prudently distributing ambulances to bases, ambulance planners are able to improve their performance (Takeda et al., 2007). Most statistical analysis of EMS travel times has focused on either predicting travel times based on road types and travel conditions encountered when traveling from the base location to the scene of the incident (Henderson and Mason, 2004; Harewood, 2002; Kommer and Zwakhals, 2011), or based simply on the birds eye distance between both the two points, scaled by correction factors (Fujiwara et al., 1987; Aringhieri et al., 2007). Other techniques use graphical analysis, factor scaling (comparing travel time data to Google Maps travel times and distances), cluster analysis to group demand locations and find factors, and cluster analysis to group demand locations so that a significant distributional fit might be found to individual groups.

Knight and Harper (2012) have studied the effect of individual components of the ambulance service time using Coxian phase-type distributions. By fitting distributions to both the overall service times for different classes of patient priorities, as well as for the separate components of the service times, they were able to identify expected gains from adjusting specific components of the response process on the overall efficiency of an ambulance service. Ultimately, the insight they offered on the benefit of reducing turnaround times pointed towards the need for an entire systems approach given that the congestion in the hospital impacts on the ED and in turn on EMS turnaround times.

Further work in this area could involve studying how load-dependent average service times could be incorporated into mathematical models on EMS systems.

Ingolfsson (2013) has already shown that the chute time appears to decrease with load, whilst hospital time increases, but such observations have not yet been incorporated into time-varying models of the system.

1.3.3 Dispatch center

When optimizing the ambulance distribution, most research focuses on adjusting the travel time component with a fixed or 'known' pre-trip delay. However, the pre-trip delay time can greatly influence the overall response time, and thus, can the adequate management of the EMS call center lead to a reduction of the response time. Typically, incoming calls first receive complete triage before an ambulance is dispatched. However, another approach that could lead to quicker response times could be to dispatch an ambulance even before triage is completed. Since this can potentially reduce the pre-trip delay, quicker response times can be achieved. On the other hand, inappropriate ambulance assignment as a result of the incomplete triage could lead to a higher workload. It would be interesting to investigate the overall impact on the performance.

Apart from the potential response time reduction, an efficiency gain could be obtained by improving the call center staffing. For other applications, significant research has been done on the optimal staffing of call centers (c.f. Koole and Mandelbaum (2002)). Research specific for EMS call centers is limited. One of the few peer-reviewed papers (Kozan and Mesken, 2005) introduces a simulation tool that can be used for what-if scenarios to improve the staffing levels in EMS call centers. Dwars (2013) introduces a simulation tool that contains both specialized call takers and dispatchers and generalists that can do both. The tool is designed to find good configurations of call center crews. The main application is to investigate the potential gain of merging call centers. On the one hand, economies of scale can lead to significant savings, whereas the loss of region-specific knowledge can result in longer call durations and lower efficiency. Despite this lower efficiency as a result of the lack of this regional knowledge, Dwars (2013) shows that significant efficiency gains can be obtained by merging call centers. In the Netherlands, for example, one can observe a trend of merging call centers. From the 24 call centers that where there some years ago, only ten will remain in the upcoming years.

Despite the mentioned results, it is fair to say that the call center domain of EMS systems is not as well-studied as other domains and contains some good areas for future research.

1.3.4 Staff scheduling

For EMS planning, mainly two different types of staff are distinguished: the staff working in the dispatching center and the staff working in the ambulances. In almost all countries, the two types of staff are completely disjoint and thus can be scheduled separately. However, in some German regions the providers prefer that dispatchers also work at an ambulance from time to time to not loose contact

with practice. The staffing of dispatching centers has already been discussed in the previous section.

In general, staff scheduling problems have been extensively discussed in the literature. A review over existing approaches, methods and application areas can, for example, be found in Van den Bergh et al. (2013). While some of the general approaches might be used for fixing shifts for paramedics and assigning them to ambulances, several papers have already been published that explicitly consider the ambulance rostering problem. Bradbeer et al. (2000) discuss the ambulance roster problem and present three approaches that build upon each other and are based on genetic algorithms. They assume that the number and locations of ambulances is given. Li and Kozan (2009) define two stages for solving the problem. First, shift start times and the necessary number of ambulance staff to be assigned to each shift are determined using a deterministic model. Then, an allocation model assigns all ambulance staff to shifts resulting in a schedule for four weeks. In contrast, Erdoğan et al. (2010) combine the crew rostering and the ambulance location problem. Their objective is to schedule ambulance crews in order to maximize the coverage throughout a planning horizon. In order to do that, they first run tabu search to locate ambulances and use the output to solve the crew rostering problem. For that, they present two Integer Programming models. Also Rajagopalan (2011) present a two-stage approach for crew rostering and ambulance location planning. In the first stage, they solve a dynamic expected coverage model using tabu search. For the second stage, an Integer Programming model is presented. Jasim (2002) presents a set partitioning approach to solve the staff scheduling problem for the New Zealand EMS provider St John. In addition, a "fatigue model" is applied to the optimal solution to incorporate the personal needs of the employees.

1.3.5 Planning of emergency calls

Next, we describe different models that can be used for an efficient planning of emergency calls. First, we describe the problems that arise at a strategic and tactical level. This include models to determine good base locations and a good distribution of ambulances over the bases. Often, these two problems are solved simultaneously. Most models assume a fixed capacity and try to maximize the performance with the available resources. However, it is also interesting to consider the problem of deciding on an optimal capacity level so as to obtain a minimum performance. Second, we give an overview of some models that consider problems at the operational level. This includes, for example, dispatch rules and real-time relocation. The decisions made on the strategic and tactical level are typically considered as input at the operational level. Finally, we highlight some related problems that might have a significant impact on the performance of an EMS system.

Strategic and tactical level

The most important problems on the strategic and tactical level are to determine good base locations and a good distribution of ambulances over these bases. Although the first problem is more a strategic decision and the second more a tactical decision, these problems are often solved simultaneously. In Van Essen et al. (2013), approaches for solving the two problems simultaneously or subsequently are presented and compared. When fixing the set of bases, we can use the same models to solve the problem of distributing the ambulances separately. The models are typically formulated in a way to maximize the performance given a fixed set of resources. However, with slight modifications, most of the models can also be used to determine the required capacity to satisfy a minimum performance requirement. The vast majority of models use coverage-based performance measures. These models maximize the fraction of calls that can be reached within a given target response time. This is mainly due to the fact that in almost all countries, EMS providers are assessed on these kind of measures. Nevertheless, there are models that use different objectives. For example, Dzator and Dzator (2013) minimize the average response time by applying the p-median model (ReVelle and Swain, 1970) to ambulance location.

Two of the first ambulance location models did not incorporate the ambulance distribution. Toregas et al. (1971) introduced the Location Set Covering Model (LSCM) to determine the minimum required number of bases to cover the entire region within a fixed time threshold. The Maximal Covering Location Problem (MCLP) (Church and ReVelle, 1974) was introduced to maximize the coverage given a limited number of bases. Inspired by these two models, much research was done to include the ambulance distribution in the models. At first, it was assumed that a fixed number of ambulances was required to obtain full coverage. Examples of models of this type are DSM (Gendreau et al., 1997), BACOP (Hogan and ReVelle, 1986), and MALP (ReVelle and Hogan, 1989). After that, the concept of marginal coverage was introduced by Daskin (1983). Here, each additional ambulance covering some area provides some coverage to that region. This model uses expected coverage as opposed to the all-or-nothing coverage of the previous models. Many models were introduced that extend on Daskin's MEXCLP by incorporating time-dependent demand (Repede and Bernardo, 1994; Van den Berg et al., 2016), stochastic response times (Ingolfsson et al., 2008; Van den Berg and Aardal, 2015), or survival probabilities (Erkut et al., 2008; Knight et al., 2012). Recent approaches use stochastic programming as, for example, done by Nickel et al. (2015). A more extensive overview of the literature on ambulance location models can be found in Brotcorne et al. (2003) and Li et al. (2011). In Section 2.1, some of these basic models are compared in an empirical study.

Operational level

At the operational level of the planning, real-time decisions should be made, such as which ambulance to send to a call and how to relocate the remaining vehicles.

Gendreau et al. (2001) were one of the first to address the real-time ambulance location problem. They propose a dynamic version of the static Double Standard Model. It incorporates the current state of the system in finding good relocations. Whenever a redeployment decision must be made, the adapted version of DSM is solved. A similar approach is used by Gendreau et al. (2006), where MEXCLP is solved instead of DSM. Over the last ten years, many models were introduced that are specifically designed to capture the dynamics of an EMS system. Zhang (2012) solves the real-time relocation problem for a small number of ambulances by Dynamic Programming. For larger instances, their model suffers from the curse of dimensionality. Bjarnason et al. (2009) evaluate policies using a simulation tool. Based on the results of the simulation, an optimization tool is used to find better policies. As apposed to most models that significantly simplify the system, Maxwell et al. (2010) include as many details of the real system as possible and apply approximate dynamic programming (ADP) to find good relocation policies. ADP is further used by Schmid (2012) to find dispatch and relocation policies in case travel times and call rates fluctuate over time. However, redeployment decision are limited to the moment an ambulance becomes available after finishing a call. Alanis et al. (2013) pose a two-dimensional Markov chain to evaluate the system given a compliance table. Jagtenberg et al. (2015) introduce a heuristic in which an ambulance is sent to the base where it provides the highest marginal coverage according to the MEXCLP objective function. A heuristic approach to compute redeployment actions in an equidistant graph with a limited number of ambulances is presented by Van Barneveld et al. (2015).

Besides redeployment decisions, real-time decisions must be made on which ambulance to dispatch to a call. Even though Carter et al. (1972) already showed that it is not always optimal to dispatch the closest idle ambulance, it is still by far the most common dispatch rule. This assumes knowledge about the locations of the available ambulances. As observed by Dean (2008), this information is not always present. One notable exception of the closest-idle dispatch rule is Andersson and Värbrand (2007), who adopt alternative dispatch rules for low priority calls. However, they do not try to find optimal dispatch rules. Schmid (2012) uses approximate dynamic programming to find dispatch policies, and find that deviating from the closest-idle dispatch rule for non life-threatening calls can improve the overall performance.

Emergency doctors

In some countries, including for example Germany, the Franco-German system rather then the Anglo-American system is used for emergency medical services (Dick, 2003). In this system, besides the paramedic also an emergency doctor is sent to the scene. Typically, these medical doctors are working in a hospital or a private practice. In case of an emergency, the doctor is picked up by the ambulance if the ambulance is stationed at the same hospital as the doctor, or the doctor uses separate transportation to get to the scene. The system is

called a "rendez-vous" system, as the doctor and the ambulance are meeting at the scene. In some regions, the target response time does not only hold for the ambulance, but also for the emergency doctor. Therefore, the location of the emergency doctors can be crucial. However, to the best of our knowledge, there are no publications on explicitly locating emergency doctors in the literature. A main reason is probably that emergency doctors usually work in hospitals or practices while being on duty and these locations cannot be changed. In addition, it is not possible to just assign additional emergency doctors as a specific time-consuming and expensive training is a prerequisite for working as an emergency doctor. Nevertheless, the assignment of shifts is an important task, especially if in total more doctors and locations are available to choose from than necessary for each shift. This can result in a combination of a simple maximum coverage problem to make sure that the considered region is covered (as good as possible) with a shift scheduling problem. If, for example, there are only two emergency doctors in an area, these two should not have overlapping shifts.

Helicopters

In many countries, helicopters are used in the most severe cases. Typically, the helicopter is not the first responder, but is used to provide more specialized medical assistance. This is mainly due to the high start-up times of helicopters. A land ambulance is used for the first response and if necessary, a specialized doctor arrives by helicopter. The helicopter can then also be used for the transportation to a hospital or trauma center. In many cases, the patient is transported by the land ambulance to a suitable place for the helicopter to land. One notable exception is the region of Ontario, Canada, where aircrafts and helicopters are also used for non-urgent patient transportations. For this region, Carnes et al. (2013) developed a model to better schedule the aircrafts that are used for these transportations. Another exception is Norway, where helicopters are used instead of land ambulance in rural areas. In Chapters 8 and 9, we further discuss these two systems. For the case where helicopters are only used for trauma patients, Erdemir et al. (2010) introduce a MCLP-based model for the simultaneous optimization of land and air ambulances. Here, a patient can be served by a land ambulance, a helicopter or both. Cho et al. (2014) optimize the location of the helicopters as well as the location of the trauma centers. The trauma centers can only be located at specific existing hospitals. Furuta and Tanaka (2014) consider the case where land transportation is also necessary and the ambulance and helicopter meet at a rendez-vous point. The goal is to reduce the access time for specialized care compared to the case where only land ambulances are used. Given the land ambulance distribution, the best location for helicopters and rendez-vous points is determined.

Drop-off at hospitals

It is in the interest of ambulance providers, patients and health care workers to experience a swift handover of care at the hospital. Long handover does not

only waste valuable resources but can also be harmful to patients, whose condition might deteriorate while waiting in the ambulance bay. This is, however, not always possible and conflicting targets for the ambulance service and the emergency departments (EDs) sometimes lead to long turnaround times. In order to promote swift handovers, some European countries issue target times for ambulance personnel to transfer patient care to the ED (for example, this target is fifteen minutes in the United Kingdom).

In the majority of European countries, it seams that the turnaround targets are attained at a reasonable level and therefore, patient handover is currently not a major area of concern. However, it is a notorious problem in the UK, as well as for some areas of the United States and Canada. In fact, the number of 'lost' ambulance hours due to long handovers has been estimated to have cost the UK National Health Service (NHS) millions of pounds a year. In Wales alone, there has been a five-fold increase in 'lost' ambulance hours in the last few years (from around 8,000 in 2008 to 40,000 in 2014). Hence, it is not surprising that this has been closely monitored by the media (Hughes, 2009; Jones, 2011; Clarke, 2015). Not only is this money wasted that could be better used within the service, distressed patients spend long periods of time waiting for transfer of care, resulting in potential deterioration in their condition and effectiveness of subsequent treatment. Furthermore, whilst waiting to handover patients, the crew's vehicles are blocked, which results in decreasing coverage (Lowthian et al., 2011).

The effect of reducing turnaround time on performance has been investigated in several simulation studies, which have focused on its impact on response time as this is the common measure for EMS systems and comparable across the countries (Knight et al., 2012). However, since some ambulance services are moving to clinical outcome based measures (e.g., the Welsh Ambulance Service Trust, WAST), it is also of interest to see if survival between different scenarios alters. Investigations into the effect of reducing the turnaround time have been undertaken in several studies; notably with Knight and Harper (2012) showing that if turnaround times were reduced to the extent that the government targets were met in Wales, this would be equivalent to 15% extra capacity on the ground.

As mentioned before, patient handover is also a critical issue in Canada. Therefore, Carter et al. (2015) propose the introduction of offload zones in hospitals to shorten the drop-off time while also controlling the workload in the ED. They describe it as an additional area next to the ED where a nurse looks after patients that were taken to hospital by an ambulance.

1.3.6 Patient transportation

When patients need to be transported to, from or between hospitals this is often organized by the EMS provider. In the Netherlands, for example, these transportations are called B calls. Depending on the medical condition of the patient, it is decided what type of vehicle is sent. In the Netherlands, this can either be an ALS or a BLS ambulance. In Germany, it is even possible that a taxi

or a private transport company fulfills the task. Calls for which no ambulance is required are called unqualified patient transportations. As these are not part of the logistics of EMS provider, they are excluded from the remainder of this section.

A transportation task involves picking up a patient at one location and dropping him off at a second location. Often, for one of the two actions a time window is given. Depending on the regulations in the country, these time windows are hard and must be fulfilled or soft and may be violated. In the latter case, minimizing the violations is often (part of) the objective function. In general, a distinct set of BLS ambulances is reserved to fulfill the transportation tasks. In that case, there are mainly two decisions to be made: (1) tasks must be assigned to the ambulances and (2) the routes for the ambulances must be constructed. If not all patients can be served by the set of ambulances, some tasks need to be assigned to ALS ambulances. This results in higher costs and coverage reduction and should therefore be prevented, if possible. The underlying problem can be expressed with a Dial-a-Ride (DARP) formulation.

The DARP itself is already well-studied (see, for example, Cordeau and Laporte (2007) and Parragh (2009)). There are only a few publications that study models and approaches for planning the patient transportation problem. Parragh et al. (2009), Schilde et al. (2011), and Ritzinger et al. (2012) study patient transportations in the Austrian EMS system. They provide different DARP formulation and solution approaches. Parragh et al. (2009) include two different types of vehicles having different capacities for transportation tasks and they additionally assign drivers to vehicles. In the system presented by Schilde et al. (2011), the transportations are requested by the patients, instead of the hospital. This system only considers transportations between patients' home locations and hospitals. Ritzinger et al. (2012) assume about 1,000 transportations per day. Therefore, an approach is needed that is fast in practice. Unfortunately, this is usually going along with some compromises on solution quality.

If not all patient transportations are known in advance, reoptimization of the schedule is required throughout the day. In Chapter 6, we present an online model to schedule the patient transportations on the BLS ambulance, while minimizing the impact on emergency calls.

1.3.7 Simulation

Due to the complexity of EMS systems, it is necessary to highly simplify the system in order to obtain tractable models. Computer simulation can help to get realistic estimates on how the decisions would influence the real system. Hence, it can serve as a playground for researchers and decision makers to evaluate system changes. Numerous simulation studies have demonstrated the potential of simulation in EMS settings. Simulation is used in two main ways in the decision process. The most common use is as a stand-alone evaluation tool. Here, different scenarios are compared by means of simulation. The simulation is not used to

generate these scenarios. Alternatively, simulation can also be used as a subroutine in the optimization. In that case, simulation is used to highlight directions for search in the optimization procedure. In this section, we will discuss some papers in both categories. For an extensive overview of simulation models, we refer to Aboueljinane et al. (2013).

As a consequence of the high level of detail that can be incorporated in simulation models, most simulation studies focus on one specific ambulance region. The simulation tools are often not easily transferable to other regions. Two notable exceptions are Henderson and Mason (2004) and Kergosien et al. (2014). The first paper introduces BartSim, a simulation tool that was originally developed for the region of Auckland, New Zealand. Later, BartSim formed the basis for a more general simulation tool commercialized by The Optima Corporation, which is now the market leader in EMS simulation software. Their software is used in many different countries. Kergosien et al. (2014) have also proposed a generic discrete event simulation-based analysis model that can be adapted to a wide range of EMS facilities. In particular, it considers how to optimally serve emergency requests in addition to patient transportations between their homes and other medical facilities. Other papers where simulation is used to evaluate scenarios are typically more region specific. The evaluated scenarios are proposed by decision makers (Aboueljinane et al., 2014), ILP models (Aringhieri et al., 2016; De la Mota et al., 2015), or heuristics (Jain and McLean, 2003).

A rather different approach is to use simulation as a subroutine of an optimization procedure. Lee et al. (2012) iteratively use simulation to estimate busy fractions in a static ambulance location model. With the new busy fraction, a new solution is found for which the busy fraction is estimated by the simulation. Yue et al. (2012) use simulation to obtain the objective value of solutions. In the optimization, the simulation is called every time the value of a solution is requested. Finally, in McCormack and Coates (2015) simulation gives the fitness of the current solution in a genetic algorithm.

For both uses of simulation (i.e., as a stand-alone tool and as a subroutine in the optimization process) it is crucial to have a realistic representation of the EMS system. For example, travel times should be incorporated in a realistic way. For some regions, Euclidean or Manhattan distances might yield good estimates, whereas for more irregular networks, other travel time models should be used. Another important step is the generation of calls. As mentioned in the 'Demand forecasting' section, it can be a challenging task to estimate demand distributions. An alternative could be to use trace-driven simulation where call streams are extracted from historical data. In this way, a particular period of time can be evaluated with the new configuration. A final example of modeling choices in simulation studies is the relocation policy. Since most EMS systems use at least some form of dynamic ambulance management, it is important to incorporate this in the simulation. However, often it is not clear under what circumstances relocations are executed. In order to realistically simulate the EMS system, some relocation rule should be implemented.

1.4 Thesis outline

The remainder of this thesis consists of three parts. In Part I, we discuss models for the location of ambulances and firefighter vehicles. This part consists of four chapters. In Chapter 2, two experiments are conducted that form the basis for the succeeding chapters. First, we perform a computational comparison of six ambulance location models from the literature. The results of the different models for the 24 ambulance regions in the Netherlands are compared on criteria arising from practice and by simulation. Second, we measure the impact of using a coarser grid of data aggregation. This is done by comparing the results of the commonly used level of aggregation with the results of a finer aggregation level. The results of this chapter give guidelines for the following chapters. Chapter 3 presents an extension to the well-known Maximal Covering Location Problem to incorporate fluctuating characteristics throughout the day. By including a penalty on the number of locations and ambulance relocations, we avoid completely different ambulance configurations at different times of the day. Chapter 4 introduces an Integer Linear Programming formulation for the version of MEXCLP with fractional coverage probabilities. This allows for the inclusion of stochastic travel times and survival probabilities into the model. In the literature, nonlinear formulations for this problem already exist. We show that the presented linear formulation is equivalent and results in a significant reduction of the computation time. This allows for solving larger instances with a larger number of potential base locations. In Chapter 5, we introduce a location model for firefighter vehicles. Here, firefighter-specific characteristics are taken into account, which significantly changes the models.

Part II focuses on the scheduling of non-urgent patient transportations. As typically some of these calls are served by an ALS ambulance, there is a link with the emergency coverage. In Chapter 6, we introduce a model to determine routes for the BLS ambulances that result in the smallest coverage reduction for emergency calls. In Chapter 7, this model is used to evaluate the current shift schedule for BLS ambulances in the region of Utrecht. Additionally, new shift schedules are proposed and evaluated with the model.

In Part III, the results of two studies with air ambulance providers are discussed. Chapter 8 presents a simulation model for the air ambulance provider in Ontario, Canada. Here, both helicopters and fixed wing aircrafts are used to provide care and transportation of the patients in this mainly rural area. With the simulation model, we evaluate different dispatch policies and the current shift schedule. Furthermore, an optimization model based on MEXCLP that allows for multiple vehicle types is introduced. This model is used to find alternative shift schedules which are again evaluated in the simulation model. Finally, Chapter 9 describes an application of the Maximal Covering Location Problem to the Norwegian air ambulance provider. A limited number of base changes is proposed that can already significantly increase the coverage throughout the country.

Part I

Facility Location Models

2

Preceeding computations

For emergency medical service providers, it is important to locate ambulances in such a way that patients can be reached as quickly as possible. As reviewed in Section 1.3, there already exist a large number of models that locate bases and ambulances such that the fraction of the demand that is reached within a specified target response time is maximized. Before we extend existing literature with some new models and apply these models to some regions in the Netherlands in Chapter 3-5, we conduct two experiments. First, we compare six of the basic models in order to get insight in the suitability of the models in practice. We evaluate the provided solutions for the 24 ambulance regions in the Netherlands on 11 criteria, and by means of simulation. The results of this experiment will help in selecting good models as a basis for more complicated models in the following chapters. Second, we evaluate the impact of different levels of data aggregation by applying the Maximal Covering Location Model (Church and ReVelle, 1974) to two regions for which we have more detailed travel time data. For most computations in this thesis, we will use the four digit postal codes[1] as demand points. This experiment quantifies the potential loss by this approach and proposes an alternative method to overcome this issue.

2.1 Computational comparison of static ambulance location models

In this section, several existing ambulance location models are compared. This comparison is of importance for two reasons. First of all, the considered models have been applied in numerous case studies. Most researchers only evaluate their models on a few criteria, which typically suit their model well. For application

[1] Postal codes in the Netherlands consist of four digits and two letters. The four digits correspond to a neighborhood, whereas the full postal code results in a part of a street. The combination of a postal code and a house number gives a unique address.

in real-life case studies, it is important to assess the performance of the models according to different criteria arising from practice. Second, the models have often been used as a basis for future research. As many of the characteristics of the basic models are preserved in the extended versions, understanding of the behavior of the basic version of the model is important.

Clearly, we cannot include all models in the comparison, and therefore we make a selection of six models. We select the models in such a way that different concepts underlying the models are included. As it is one of the first, and most easy to solve models, we include MCLP (Church and ReVelle, 1974). This model maximizes the single coverage. The concept of backup coverage is included by the Double Standard Model (Gendreau et al., 1997). Other models that use this concept are, for example, BACOP1 and BACOP2, introduced by Hogan and ReVelle (1986). MALP (ReVelle and Hogan, 1989) and MEXCLP (Daskin, 1983) are included as two models that incorporate busy fractions to determine the required coverage or the expected coverage, respectively. Despite the fact that almost all ambulance providers are assessed by a coverage-based performance measure, it is of interest to consider models that have response time-based objectives. Even though a patient that is reached within five minutes receives better care than a patient that is reached in nine minutes, this is not reflected in the coverage with respect to a ten minute response time target. To capture this, we also include two models that consider the average response time.

We compare the chosen models according to several criteria arising from practice. One of the most important requirement in practice is the achieved coverage. As for the realized fraction of calls that is reached in time, it makes a difference by how many ambulances a demand location is covered, we compare coverage by one, two, and three ambulances. In addition, we compare the expected coverage, which is the objective of MEXCLP. As indicated before, the achieved response time is at least as important as the target response time. Therefore, we also determine the achieved coverage for target response times other than the one used to obtain the solution. Finally, we consider the average and maximum response time.

The remainder of this section is structured as follows. In Section 2.1.1, we introduce and discuss the six considered models and the adjustments made to be able to compare the models fairly. The criteria on which the models are compared are discussed in Section 2.1.2. In Section 2.1.3, we present the results of our experiments and draw conclusions on the performance of the models. Section 2.1.4 presents conclusions and gives recommendations for further research.

2.1.1 Description of considered models

In this section, we describe the six models that we compare in Section 2.1.3. For each of the models, the set of potential base locations is given by the set I and the set of demand location is given by the set J. The travel times, including a fixed pre-trip delay, from all potential base locations $i \in I$ to demand locations $j \in J$ are given by t_{ij}. Most of the models from literature use a target response

time denoted by r that must be met for a demand location to be covered. To be more specific, each demand location $j \in J$ for which a base $i \in I$ is opened with $t_{ij} \leq r$, is covered by this base location. From a modeling perspective, it is useful to introduce the sets $I_j = \{i \in I | t_{ij} \leq r\}$ which represent the potential base locations that cover demand location $j \in J$. If at least one of these bases is opened, demand location $j \in J$ is covered.

All introduced models use integer variables x_i to indicate how many ambulances are located at base location $i \in I$. When x_i takes value 0, this means that base location $i \in I$ is not opened. The binary variables y_{jk} indicate whether demand location $j \in J$ is covered by at least k ambulances. The total number of ambulances is fixed and given by p.

To denote the importance of each demand location $j \in J$, the models use weights d_j. These weights can, for example, represent the average number of calls per year or the population at this demand location. The weights are used to give preference to more important demand locations when placing the ambulances.

Note that we present the models in the same way as originally published. For a fair comparison, some of the models are later adjusted. These adjustments are described at the end of this section.

Maximal Covering Location Problem

The first model we discuss is the Maximal Covering Location Problem (MCLP) that was introduced by Church and ReVelle (1974). This model maximizes the weighted number of demand locations that are covered by at least one ambulance. The variable x_i can only take values 0 or 1. This means that at most one ambulance can be placed at each base location.

The values of x_i are used to determine which demand locations $j \in J$ are covered by at least one ambulance which is represented by binary variables y_{j1}.

$$\max \quad \sum_{j \in J} d_j y_{j1}$$

$$\text{s.t.} \quad \sum_{i \in I_j} x_i \geq y_{j1} \qquad \forall j \in J \tag{2.1}$$

$$\sum_{i \in I} x_i = p \tag{2.2}$$

$$x_i \in \{0, 1\} \qquad \forall i \in I \tag{2.3}$$

$$y_{j1} \in \{0, 1\} \qquad \forall j \in J \tag{2.4}$$

The MCLP can be used to determine the optimal base locations. In addition, by solving the model for different values of p, we can determine how many base locations are needed to guarantee a certain coverage level. However, the model assumes that each ambulance is always available. Clearly, in practice, this is not the case. Therefore, the coverage that is indicated by the objective function can typically not be guaranteed in practice.

Double Standard Model

As a demand location might not be covered anymore when an ambulance is occupied, the Double Standard Model (DSM) (Gendreau et al., 1997) focuses on covering each demand location by two ambulances. DSM uses two target response times, namely r_1 and r_2. The target response time r_1 is the same as r for MCLP. Different from MCLP, the single coverage with respect to this target is not included in the objective function. A constraint is included to ensure that at least a fraction α of the demand is covered. In addition, all demand locations must be covered within r_2 which is ensured by constraints (2.5). Naturally, the value of r_2 must be larger than the value of r_1. As defined before, binary variables y_{j1} and y_{j2} indicate whether demand location $j \in J$ is covered within time r_1 by at least one or two ambulances, respectively. Similar to I_j, we introduce $I_j^{r_2}$ as the set of potential base locations that can cover demand point j within r_2 minutes. The objective of DSM is to maximize the weighted demand that is covered twice within target response time r_1.

As a second ambulance at a given base can now improve the solution, the number of ambulances placed at a base location is no longer limited to one. We now limit the number of ambulances at base location i by p_i.

$$\max \quad \sum_{j \in J} d_j y_{j2}$$

$$\text{s.t.} \quad \sum_{i \in I_j^{r_2}} x_i \geq 1 \qquad \forall j \in J \tag{2.5}$$

$$\sum_{j \in J} d_j y_{j1} \geq \alpha \sum_{j \in J} d_j \tag{2.6}$$

$$y_{j2} \leq y_{j1} \qquad \forall j \in J \tag{2.7}$$

$$\sum_{i \in I_j} x_i \geq y_{j1} + y_{j2} \qquad \forall j \in J \tag{2.8}$$

$$\sum_{i \in I} x_i = p \tag{2.9}$$

$$x_i \leq p_i \qquad \forall i \in I \tag{2.10}$$

$$y_{j1}, y_{j2} \in \{0, 1\} \qquad \forall j \in J \tag{2.11}$$

Average Response Time Model

One of the ambulance location models that considers the response time rather than the coverage is the Average Response Time Model (ARTM). This model is equivalent to the p-median model introduced by ReVelle and Swain (1970) and is applied to ambulance location by Dzator and Dzator (2013). We will refer to this model as ARTM, because this better fits our application of the model. The model minimizes the average response time from the nearest base. To that end,

we have a binary variable z_{ij} that takes value 1 if the opened base $i \in I$ is the closest base to demand location $j \in J$. As placing more than one ambulance per base does not improve the objective function, the number of ambulances per base is limited to one.

$$\begin{aligned}
\min \quad & \sum_{i \in I} \sum_{j \in J} d_j t_{ij} z_{ij} \\
\text{s.t.} \quad & \sum_{i \in I} z_{ij} = 1 && \forall j \in J && (2.12) \\
& x_i \geq z_{ij} && \forall i \in I, j \in J && (2.13) \\
& \sum_{i \in I} x_i = p && && (2.14) \\
& x_i \in \{0,1\} && \forall i \in I && (2.15) \\
& z_{ij} \in \{0,1\} && \forall i \in I, j \in J && (2.16)
\end{aligned}$$

Maximum Expected Covering Location Problem

The Maximum Expected Covering Location Problem (MEXCLP) introduced by Daskin (1983) is one of the first models that takes the probability that an ambulance is unavailable, called the busy fraction, into account. The model maximizes the weighted expected coverage of all demand locations while considering the probability that an ambulance is available within the target response time r. Binary variable y_{jk} indicates whether at least k ambulances can cover demand location $j \in J$. In the objective function, the probability that one of the ambulances is available is determined. Given that a demand point is covered by k ambulance and the busy fraction is denoted by q, this probability is $E_k = 1 - q^k$. Here, we assume that ambulance availabilities are independent. The marginal coverage of the k-th ambulance is then $E_k - E_{k-1} = (1-q)q^{k-1}$. Note that Constraint (2.18) states that the number of ambulances is smaller than or equal to p, whereas in the other models this has to hold with equality. This is, however, no restriction, as every optimal solution will use all p ambulances.

$$\begin{aligned}
\max \quad & \sum_{j \in J} \sum_{k=1}^{p} d_j (1-q) q^{k-1} y_{jk} \\
\text{s.t.} \quad & \sum_{i \in I_j} x_i \geq \sum_{k=1}^{p} y_{jk} && \forall j \in J && (2.17) \\
& \sum_{I \in I} x_i \leq p && && (2.18) \\
& x_i \in \mathbb{N} && \forall i \in I && (2.19) \\
& y_{jk} \in \{0,1\} && \forall j \in J, k \in \{1, \ldots, p\} && (2.20)
\end{aligned}$$

Maximum Availability Location Problem

Another model that takes the busy fraction q into account is the Maximum Availability Location Problem (MALP) introduced by ReVelle and Hogan (1989). Prior to formulating an instance of the model, the minimum number of ambulances b needed to guarantee a coverage level α is determined with the use of busy fraction q. The value of b is given by $\lceil \frac{\log(1-\alpha)}{\log q} \rceil$ as we must have that $1 - q^b \geq \alpha$.

$$\max \quad \sum_{j \in J} d_j y_{jb}$$

$$\text{s.t.} \quad \sum_{i \in I_j} x_i \geq \sum_{k=1}^{b} y_{jk} \qquad \forall j \in J \tag{2.21}$$

$$y_{jk} \leq y_{j(k-1)} \qquad \forall j \in J, k \in \{2, \ldots, p\} \tag{2.22}$$

$$\sum_{i \in I} x_i = p \tag{2.23}$$

$$x_i \in \{0, 1\} \qquad \forall i \in I \tag{2.24}$$

$$y_{jk} \in \{0, 1\} \qquad \forall j \in J, k \in \{1, \ldots, p\} \tag{2.25}$$

Expected Response Time Model

ARTM only considers the drive time from the nearest base location for each of the demand locations. Thus, this model does not take into account the fact that the nearest ambulance might not be available. Therefore, we also consider the Expected Response Time Model (ERTM) which minimizes the expected response time for all demand locations. This model is similar to the Reliability p-Median Problem (RPMP) introduced by Snyder and Daskin (2005). They applied this model to the facility location problem and included a penalty when all facilities are unavailable. Our model differs in this case, as we assume that all emergency calls are served, no matter if all ambulances are occupied. This is a realistic assumption, as in practice, there is always an ad-hoc decision possible that allows all emergency calls to be served. Therefore, we assume that when all ambulances are occupied in theory, the call is served by the farthest ambulance.

The ERTM determines the expected response in a similar way as MEXCLP determines the expected coverage. For each demand point, the location of the nearest ambulance, the second nearest ambulance, up to the p^{th}-nearest is determined with the use of binary variables z_{ijk}. These binary variables take value 1 if an ambulance at base $i \in I$ is the k-th-nearest ambulance for demand location $j \in J$. With this information, the expected response time for demand location $j \in J$ can be determined. The probability that demand location $j \in J$ is served by the nearest ambulance is given by $1 - q$, the probability that demand location $j \in J$ is served by the second nearest ambulance is given by $q(1 - q)$, etc. The probability that demand location $j \in J$ is served by the farthest or p^{th}-nearest

ambulance is slightly different to ensure that the probabilities sum up to one. Therefore, this probability is given by

$$1 - \sum_{k=1}^{p-1} (1-q)q^{k-1} = q^{p-1}.$$

ERTM then minimizes the weighted expected response time.

$$\begin{aligned}
\min \quad & \sum_{i\in I}\sum_{j\in J}\sum_{k=1}^{p-1} d_j t_{ij}(1-q)q^{k-1} z_{ijk} + \sum_{i\in I}\sum_{j\in J} d_j t_{ij} q^{p-1} z_{ijp} \\
\text{s.t.} \quad & \sum_{i\in I} z_{ijk} = 1 && \forall j \in J, k \in \{1,\dots,p\} && (2.26)\\
& x_i \geq \sum_{k=1}^{p} z_{ijk} && \forall i \in I, j \in J && (2.27)\\
& \sum_{i\in I} x_i = p && && (2.28)\\
& x_i \in \mathbb{N} && \forall i \in I && (2.29)\\
& z_{ijk} \in \{0,1\} && \forall i \in I, j \in J, k \in \{1,\dots,p\} && (2.30)
\end{aligned}$$

Adjustments

To make a fair comparison among the models, we have to adjust some of them. One of the adjustments is that we limit the number of bases that can be opened. The reason for doing this is twofold. First, in practice, opening bases is costly and therefore the number of bases is limited. Second, some of the models do not benefit from opening more bases, while other models do. Hence, by not limiting the number of bases, we favor some models which might result in a biased comparison. The following constraints are added to the models:

$$\begin{aligned}
& p \cdot f_i \geq x_i && \forall i \in I, && (2.31)\\
& \sum_{i\in I} f_i \leq f_{\max}, && && (2.32)\\
& f_i \in \{0,1\} && \forall i \in I. && (2.33)
\end{aligned}$$

The binary variable f_i takes value 1 when base location $i \in I$ is opened and 0 otherwise.

The maximum number of base locations to be opened $f_{\max}$ might be conflicting with the total number of ambulances p for the models MCLP, MALP and ARTM. For these models, the variable x_i is a binary variable that may prohibit the mentioned models from placing exactly p ambulances. Therefore, we change

binary variables x_i to integer variables to make sure that a feasible solution exists for models MCLP, MALP and ARTM.

The above mentioned adjustments cause a new problem for MCLP and ARTM. These models only focus on placing one ambulance at a base and distribute the remaining ambulances randomly over the opened bases. To make the comparison more fair, we limit the number of ambulances per base to make sure that the remaining ambulances are spread equally over the opened bases instead of locating them all at only one base. To model this, we add the following constraint to both MCLP and ARTM,

$$x_i \leq p_i \qquad \forall i \in I, \tag{2.34}$$

where $p_i = \lceil \frac{p}{f_{\max}} \rceil$.

2.1.2 Experimental setup

In the section, we describe how the models presented in Section 2.1.1 are compared. We apply the models to a set of test instances and evaluate the outcomes on 11 criteria. These criteria are based on the objectives of the different models and some other performance indicators that are important in practice. Note that the considered models do not focus on optimizing all these criteria, but we determine how well the models perform on important criteria even though these criteria are not taken into account in the model. We distinguish three categories of criteria: (1) coverage, (2) alternative response time targets, and (3) average response times. Additionally, we track the computation time of the models. One might argue that this criterion is not important, since we are dealing with strategic decisions. However, the models are often used as a basis for more complicated models, where computation times can explode. Additionally, we run a simulation with the results of the different models to evaluate how the solutions perform in a more realistic setting. Next, we introduce the criteria and the way they are computed.

Coverage criteria

We define four criteria based on the coverage within the time threshold r. The first three are the fraction of calls that is covered by one, two, or three ambulances. The first criterion is equivalent to the objective value of MCLP. Hence, this model will always perform best on this criterion. The second criterion is the objective value of DSM. However, since additional constraints are added to DSM, other models might outperform DSM on this criterion. In case $b = 3$, the third criterion corresponds to the objective of MALP. The fourth criterion is the expected coverage and is equivalent to the MEXCLP objective.

The four criteria are computed as follows:

$$\text{Crit. 1: } \frac{\sum_{j\in J} d_j y_{j1}}{\sum_{j\in J} d_j} \times 100\%, \qquad \text{Crit. 2: } \frac{\sum_{j\in J} d_j y_{j2}}{\sum_{j\in J} d_j} \times 100\%,$$

$$\text{Crit. 3: } \frac{\sum_{j\in J} d_j y_{j3}}{\sum_{j\in J} d_j} \times 100\%, \qquad \text{Crit. 4: } \frac{\sum_{j\in J} \sum_{k=1}^{p} d_j (1-q) q^{k-1} y_{jk}}{\sum_{j\in J} d_j} \times 100\%.$$

Here, y_{jk} is 1 if demand point $j \in J$ is covered by at least k ambulances within the time threshold r, and 0 otherwise.

Target response times

Since the response time target set by the regulator is not based on medical needs and is therefore rather arbitrary, we evaluate the outcomes of the models on different response time targets. In many countries, a target of 8 minutes is used. For that reason, we include the coverage within 8 minutes as the fifth criterion. As covering models do not penalize excessive response times, we add some criteria to incorporate this in the model evaluation as follows. First, we include the coverage within 20 minutes as the sixth criterion. For DSM, a constraint is added to ensure full coverage within this threshold. Additionally, we consider the maximum response time to a demand point as the seventh criterion. Finally, to avoid that small areas dominate this measure, we also consider the worst-case response time after deletion of the 5% calls with highest response time as the eighth criterion.

To determine whether a response time of 8 or 20 minutes is achieved, we introduce binary variables y_j^8 and y_j^{20}. These variables take value 1 if a response time of 8 respectively 20 minutes is achieved for demand location $j \in J$. To determine the correct values for y_j^8 and y_j^{20}, we introduce the sets $I_j^8 := \{i | t_{ij} \leq 8\}$ and $I_j^{20} := \{i | t_{ij} \leq 20\}$. We set y_j^8 to 1 when $\sum_{i\in I_j^8} x_i > 0$ and, similarly, we set y_j^{20} to 1 when $\sum_{i\in I_j^{20}} x_i > 0$. Then, criteria 5 and 6 can be computed in a similar way to criterion 1.

$$\text{Crit. 5: } \frac{\sum_{j\in J} d_j y_j^8}{\sum_{j\in J} d_j} \times 100\% \qquad \text{Crit. 6: } \frac{\sum_{j\in J} d_j y_j^{20}}{\sum_{j\in J} d_j} \times 100\%$$

To compute criteria 7 and 8, we introduce τ_j as the distance to demand point $j \in J$ from its closest open base. We have that $\tau_j = \min_{i\in F} t_{ij}$, where $F = \{i \in I | x_i > 0\}$. We get

$$\text{Crit. 7: } \max_{j\in J} \tau_j.$$

For criterion 8, we determine the minimum response time R such that 95 percent of the calls can be reached within R minutes.

Average response time

Although most ambulance service providers have coverage related targets, it is also important to provide short average response times. In our analysis, we incorporate two measures for average response times. The first one is the average response time from the closest opened base. This corresponds to the objective of the ARTM. The second measure is the expected response time, which corresponds to the objective of the ERTM. Given the variables z_{ij} and z_{ijk} as introduced in Section 2.1.1, we compute these criteria by:

$$\text{Crit 9: } \frac{\sum_{j \in J} d_j \tau_j}{\sum_{j \in J} d_j},$$

$$\text{Crit 10: } \frac{\sum_{i \in I} \sum_{j \in J} \sum_{k=1}^{p-1} d_j t_{ij} (1-q) q^{k-1} z_{ijk} + \sum_{i \in I} \sum_{j \in J} d_j t_{ij} q^{p-1} z_{ijp}}{\sum_{j \in J} d_j}.$$

Computation time

The last criterion we consider is the computation time. Even though the models are used at the strategic and tactical level, the computation time can still be important when the models are used as a basis for more complicated models. The models are implemented in AIMMS 3.14 (AIMMS BV, 2013) and solved with CPLEX 12.5.1 (ILOG, 2009) on an Intel Core i5-4300 CPU @ 1.90 GHz 2.50 GHz with 8 GB RAM. We simply implemented the ILP formulations given in this chapter without considering clever ways to improve the computation time, as this is out of the scope of the study. The computation time of ARTM, for example, could be reduced tremendously by using the optimization-based Lagrangian relaxation developed by Daskin (1995).

Overview of criteria

We conclude this section by giving an overview of the criteria.

1. Fraction of calls covered at least once within time threshold r.
2. Fraction of calls covered at least twice within time threshold r.
3. Fraction of calls covered at least three times within time threshold r.
4. Expected coverage.
5. Fraction of calls covered within 8 minutes.
6. Fraction of calls covered within 20 minutes.
7. Maximum response time.
8. Maximum response time within 95% of the calls with lowest response time.
9. Average response time of closest ambulance.
10. Average expected travel time.
11. Computation time.

2.1.3 Experimental results

In this section, the six ambulance location models discussed in Section 2.1.1 are compared based on the criteria described in Section 2.1.2. First, the data on which the results are based is described in full detail. Next, the models are compared with respect to the output of the models and the results of a simulation study. We end this section with a conclusion on the performance of the models.

Data

We apply the models to the 24 ambulance regions (RAVs) in the Netherlands. These regions differ in number of demand points, population size, surface area and density as shown in Table 2.1. Therefore, the considered regions represent a wide variety of geographical characteristics.

As demand points, we take the four digit postal code areas. We only exclude the demand points on the islands in the northern part of the Netherlands, because these islands are small and are operated separately. This gives a total of 3,990 postal codes, where the smallest region has 40 postal codes and the largest region has 456 postal codes. All these postal codes are available as a potential base station, i.e., $I = J$.

The population size of the regions varies between 243,540 and 1,247,858, the surface area between 273 and 5,748 km^2, and the density between 111 and 2,510 people per km^2. The region with the smallest population size is also the smallest region in terms of number of demand points and total surface area. The region with the largest population is, however, one of the smaller regions in terms of surface area. The region with the largest surface area is also the region with the smallest density, namely 111 people per km^2. The region with the highest density (2,510 people per km^2) is one of the regions with a higher population size.

Table 2.1 also indicates whether the regions are urban or rural. Regions with a population density of less than 750 inh/km^2 are considered rural. A density of more than 1000 inh/km^2 is considered urban, and other cases are considered mixed. Of the 24 regions, 16 are rural, four regions are urban and four regions have both rural and urban parts. A map of the different regions is given in Figure 1.1 in Chapter 1.

The travel time between two postal codes is given by a travel time model developed by Kommer and Zwakhals (2011). These travel times are based on observed travel speeds of ambulances on different road types. For the relative importance of covering demand point $j \in J$, denoted by d_j, we use the population of a demand point.

As stated before, the response time target in the Netherlands is 15 minutes. The response time consists of three parts: (1) triage and dispatch, (2) chute time, and (3) travel time, where the chute time is the time between the moment the crew is dispatched to a call and the moment the ambulance starts driving, see Figure 1.2. The average time for triage and dispatch in 2012 was 1.58 minutes, while the average chute time was 1.01 minute (Ambulancezorg Nederland, 2012).

Table 2.1: *Geographical characteristics of the 24 ambulance regions.*

Region	# Postal codes	Population	Area (km^2)	Density	Rurality
1	250	576,615	2,960	195	Rural
2	456	635,700	5,748	111	Rural
3	255	489,610	2,626	186	Rural
4	170	506,845	1,900	267	Rural
5	120	623,050	1,500	415	Rural
6	201	809,865	2,740	296	Rural
7	134	655,725	1,184	554	Rural
8	158	526,835	1,040	507	Rural
9	217	1,220,125	1,449	842	Mix
10	160	628,025	1,350	465	Rural
11	161	1,261,997	813	1,552	Urban
12	98	519,757	420	1,238	Urban
13	40	243,540	273	892	Mix
14	141	1,016,400	405	2,510	Urban
15	124	760,930	875	870	Mix
16	185	1,247,858	856	1,458	Urban
17	98	479,435	836	573	Rural
18	153	381,395	1,788	213	Rural
19	217	1,070,885	2,258	474	Rural
20	146	636,870	1,396	456	Rural
21	137	734,841	1,458	504	Rural
22	137	513,855	1,521	338	Rural
23	141	607,540	661	919	Mix
24	91	386,184	2,412	160	Rural

For this research, we assume a fixed pre-trip delay of three minutes. There are two ways to incorporate this in the model: we can add three minutes to the travel time or subtract three minutes from the response time target. We choose the first option, which implies that $r = r_1 = 15$. For DSM, we need another time threshold, within which all demand should be covered. As 20 minutes seems reasonable, we set $r_2 = 20$.

The presented models aim at finding the best distribution of a fixed number of ambulances. This fixed number is based on the required capacity according to a study by Kommer and Zwakhals (2008) given in Table 2.2. This study is used by the government to determine the budget for each RAV. We add a constraint to the models on the maximum number of bases that can be opened in a region. For this number, $f_{\max}$, we use the current number of bases, which are shown in Table 2.2.

It remains to find the appropriate values for the busy fraction q. We use the average busy fraction throughout the day. This is calculated by dividing the total workload in minutes by the total ambulance capacity in minutes. In the computation of the total workload, we distinguish two types of emergency

calls, A1 and A2. Here, A1 calls are life-threatening and have an average call duration of 42.9 minutes. A2 calls are not life-threatening and take on average 50.1 minutes (Zuidhof, 2010).

$$\begin{aligned}\textbf{Total workload} = {} & \#\text{ A1 calls a year} \times \text{average call duration A1} \\ & + \#\text{ A2 calls a year} \times \text{average call duration A2}\end{aligned}$$

The number of calls per region is extracted from the yearly report on the performance of the Dutch ambulance services (Ambulancezorg Nederland, 2012). For the total ambulance capacity, we multiply the average number of ambulances by the number of minutes in a year.

$$\textbf{Total capacity} = p \times 60 \times 24 \times 365$$

By dividing the total workload by the total capacity for each region, we get busy fractions between 0.089 and 0.447. The busy fraction of each region is given in Table 2.2.

Table 2.2: *Ambulance characteristics of the 24 ambulance regions.*

Region	Busy fraction	# Ambulances	# Bases
1	0.18	15	13
2	0.10	18	15
3	0.17	13	13
4	0.16	12	10
5	0.20	11	9
6	0.21	14	13
7	0.27	8	7
8	0.20	10	11
9	0.30	15	11
10	0.22	9	8
11	0.38	16	9
12	0.31	8	7
13	0.23	4	3
14	0.39	12	8
15	0.30	10	10
16	0.45	12	10
17	0.22	8	6
18	0.09	18	11
19	0.27	16	13
20	0.26	9	7
21	0.28	9	7
22	0.21	10	7
23	0.36	7	4
24	0.20	8	6

Both DSM and MALP require a fixed reliability α. For our computations, we take $\alpha = 95\%$. As a direct consequence of q and α, we get the value for b, which is given by $\lceil \frac{\log(1-\alpha)}{\log q} \rceil$. For the different regions, b varies between 2 and 4.

Computational results

Table 2.3 gives an overview of the performance of the different models on the 11 criteria. The table shows the average value over the 24 regions. Note that this average can be misleading when one region gives excessive values. For example, the average computation time of ERTM is highly influenced by the largest region, which has a computation time of 3.5 days. For all other regions, the computation time is more tractable, i.e., not more than four hours. In the remainder of this section, we evaluate the models on the four main groups of criteria: coverage, target response times, average response times, and computation time.

Table 2.3: *Average performance over the 24 regions for the considered criteria. Best and worst performing model highlighted in bold and italic, respectively.*

Criterion	MCLP	DSM	ARTM	MEXCLP	MALP	ERTM
Single coverage	**100.0%**	97.6%	97.6%	99.4%	*93.9%*	96.6%
Double coverage	*55.3%*	**95.4%**	73.7%	91.3%	92.6%	83.5%
Triple coverage	*25.1%*	29.5%	49.3%	**59.1%**	54.4%	52.6%
Expected coverage	*88.0%*	93.2%	90.4%	**95.5%**	91.4%	91.8%
8 min threshold	22.3%	25.3%	**55.8%**	34.5%	*21.3%*	55.3%
20 min threshold	**100.0%**	**100.0%**	99.9%	99.9%	*98.4%*	99.8%
Max. response time (min)	**15.3**	18.1	19.8	18.1	*22.8*	20.4
Avg. response time (min)	9.9	10.0	**7.9**	9.2	*10.6*	8.0
95% threshold (min)	13.9	14.4	**13.6**	13.7	*16.4*	14.0
Avg. ERT (min)	*11.2*	10.6	9.2	10.0	11.0	**9.0**
Computation time (sec)	**0.15**	21.30	69.92	3.59	0.81	*14,263.60*

Coverage criteria

By definition, MCLP outperforms all other models on single coverage. However, when backup coverage is required, MCLP performs very badly. For double, triple, and expected coverage, this model performs the worst. DSM has good performance on both single and double coverage. Since for regions with a relatively low busy fraction, double coverage suffices, also the expected coverage is reasonable. The two average response time models, ARTM and ERTM, provide rather good coverage, although the double coverage is significantly lower than for DSM, MEXCLP, and MALP. MALP provides the worst single coverage, while scoring rather well on the other coverage objectives. Overall, MEXCLP clearly outperforms all other models on coverage. For all four covering criteria, MEXCLP is among the three best models. For triple coverage and expected coverage, MEXCLP is even the best model.

Target response times

Except for DSM, no model considers other response time targets than the 15 minutes target. However, almost all models provide close-to-complete coverage within 20 minutes. Only MALP leaves more than 1% of the demand uncovered within this threshold. Also when considering the excessive response times, we see that MALP performs badly. Although ARTM and ERTM have rather high maximum response times, this effect vanishes when only the 95% best-covered demand is considered. This implies that only demand points with low population experience excessive response times. MALP provides the worst performance for the maximum response time, since for high values of b, MALP tends to ignore many demand points in order to focus on a small part of the region.

Average response time

As expected, ARTM and ERTM provide the best average response times. Here, ARTM has slightly better average response times, whereas ERTM gives better expected response times. Even though average response times are not considered by MEXCLP, the model still performs relatively well on these criteria. Especially the expected response times are better than for MCLP, DSM, and MALP.

Computation time

Although we are able to solve all instances to optimality, huge differences between the models are observed. The coverage based models can all be solved within a couple seconds. The average response time models, on the other hand, take significantly longer. ARTM takes on average a bit more than a minute to solve, while ERTM has an average computation time of four hours. Note that this average is highly dominated by one large instance, which had a computation time of 3.5 days. For all other instances, the computation time was less than four hours. Figure 2.1 shows the relationship between the size of the instance and the computation time for the different models. An exponential trendline is added based on the 23 smaller instances. The largest instance, with 456 demand points, has in most cases computation times according to that trend. For ERTM, this figure shows that the computation time for the largest instance is not disproportionately large. It is even below the trend based on the 23 other instances.

Simulation study

To get a better sense of the performance of the models in practice, we perform a simulation study. As input, we have the data described and the locations of the ambulances as determined by the models in Section 2.1.3. The arrival rate per region is determined with the use of the number of emergency calls served in 2012 as given in Ambulancezorg Nederland (2012). From this document, also the fraction of A1 and A2 calls is determined.

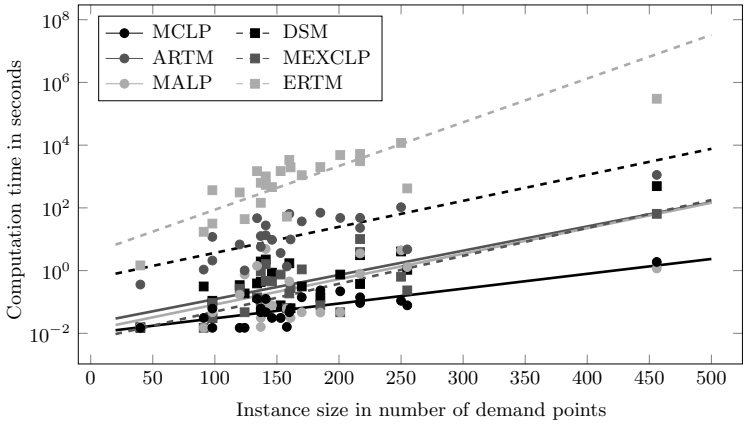


Fig. 2.1: *Computation time for different models and instance sizes. Exponential trend line is added based on 23 smaller instances.*

For each ambulance, we assume that the ambulance returns to its assigned base when a call has been fully served. However, this ambulance is already available to serve a new call when it drives from the hospital back to the base. It is important to note that we use two different travel speeds. The first is the travel speed of an ambulance going to a call location which corresponds to the travel times used as input. However, when an ambulance drives to a hospital or returns to its base location, the ambulance usually drives at a normal speed. In these situations, we assume that the travel time is approximately 10% longer than the travel times used as input.

We assume that the calls arrive as a Poisson process with the rate based on the data of 2012. For each call, we also have to generate the location of the call, the time spent at the accident scene, whether the patient needs transportation to the hospital, and if so, how long the transfer at the hospital will take. The spatial distribution of the calls is determined by means of the fraction of people living at a certain postal code. The time spent at the incident location is exponentially distributed as in Maxwell et al. (2010) with a mean of 18 minutes. The probability of a patient needing hospital treatment is 0.8034 and the time spent at the hospital has a Weibull distribution as in Maxwell et al. (2010) with a mean of 12 minutes. These numbers are all determined from Dutch data.

When a new call enters the system, the nearest ambulance is located and assigned to this call. This can be an ambulance waiting at a base location or an ambulance driving back from the hospital to its base location. In the latter case, the distance between the call location and the ambulance location is determined as follows. The time the ambulance needs to drive from the hospital to its base location is determined from the travel times used as input based on the postal codes. As we also know the current time in the simulation, we can calculate which fraction of the journey is already completed. Then, we draw a straight

line between the hospital and the ambulance base location and with the use of the (x, y)-coordinates of these two locations, we determine the (x, y)-coordinate of the current location of the ambulance. Then, the closest postal code to this (x, y)-coordinate is determined for which we know the actual drive time to the call location.

It might occur that all ambulances are occupied when a new call arrives. When this is the case, the call is put into a queue. A1 calls in the queue are prioritized over A2 calls, and for calls with the same priority, we use the first-come first-serve policy.

The order in which the calls in this queue are served depends both on the time and the priority of the call, which can be A1 or A2.

As output, we only consider the response times of the A1 calls, as this is also the focus of the considered models. Recall that in the Netherlands 95% of the A1 calls should be served within 15 minutes. In Table 2.4, we compare the results of the simulation study for the considered models. Note that criteria such as single, double, triple and expected coverage, and expected response times are not relevant in this case. The coverage criteria are replaced by the fraction of calls with a response time less than or equal to 15 minutes. The expected response time is omitted, as this is replaced by the average response time. Note further that these results are obtained from the simulation and thus incorporate ambulance unavailability. Most of the results in Table 2.3 are only based on the closest located ambulance. Hence, the behavior of the models cannot directly be compared with the results in Table 2.3.

Table 2.4: *Average performance over the 24 regions based on simulation.*

Description	MCLP	DSM	ARTM	MEXCLP	MALP	ERTM
15 min threshold	82.6%	89.0%	88.8%	93.6%	88.6%	90.7%
8 min threshold	19.3%	22.4%	44.6%	28.3%	20.2%	46.7%
20 min threshold	93.9%	96.3%	97.0%	98.1%	96.0%	97.9%
Max. response time (min)	43.3	43.1	41.0	39.9	41.2	39.0
Avg. response time (min)	11.7	11.0	9.5	10.1	11.2	9.2
95% threshold (min)	20.4	18.3	18.0	16.2	19.0	16.9

The 15 minutes threshold in Table 2.4 can best be compared with the expected coverage in Table 2.3. We see that for all models the realized coverage in the simulation is less than the expected coverage. This can be explained by the fact that for the simulation study the load of the system fluctuates over the day as a result of the randomized arrivals while in determining the expected coverage a stable situation is assumed. However, the overall conclusions still hold. MEXCLP gives the best coverage within 15 minute and MCLP the worst. When we look at the 8 minute threshold, we see that ARTM and ERTM still perform the best and MALP and MCLP perform the worst. Even though for each demand point there is a base location within 15 minutes (see Table 2.3), an ambulance

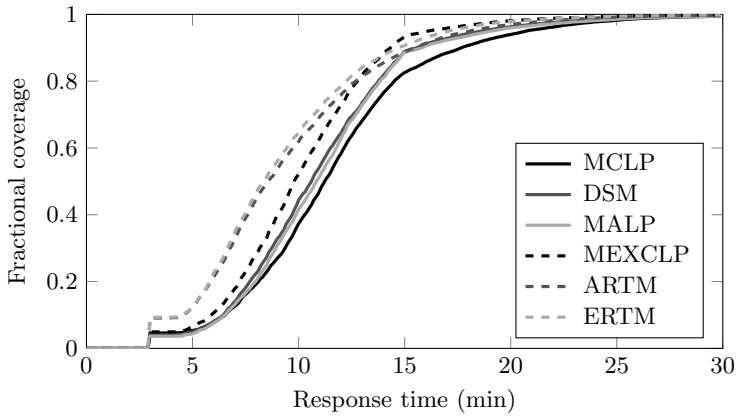


Fig. 2.2: *Response time distribution based on simulation.*

arrives at the patient within 20 minutes in only 93.9% of the cases. This is due to the lack of attention for backup coverage in MCLP. MEXCLP and ERTM perform best on the 20 minute threshold with a coverage of approximately 98%.

The values for the average response time, the maximum response time and the response time in which 95% of the calls are served are all higher for the simulation study compared to the results in Table 2.3. For the average response time, this can again be explained by the fact that for the simulation study the load of the system fluctuates over the day while in determining the expected coverage a stable situation is assumed. For the other two criteria, the reason is ambulance unavailability, which especially influences the maximum response time. If all ambulances in the system are occupied, calls are queued which leads to very high maximum response times. Since the maximum response time is dominated by queuing effects, no huge differences between the models are observed. For the 95% threshold criterion, this effect vanishes and we see that MEXCLP performs best. MCLP again performs worst, as it focuses only on single coverage. As expected, the average response time is smallest for ERTM and ARTM. Given the MEXCLP set-up, an ambulance would on average arrive almost one minute later at the scene, compared to ERTM.

When we have a look at the response time distributions of the different models in Figure 2.2, we see that ERTM and ARTM serve a large fraction of calls within a short response time. For response times closer to the target response time, MEXCLP gets closer and eventually outperforms the two models. Furthermore, it can be seen that only for higher response times ERTM scores better than ARTM. DSM and MALP show very similar behavior, while MCLP clearly shows the worst performance.

To conclude, the simulation study confirms the conclusions based on the 11 criteria. MEXCLP and ERTM are still the two models that perform best. From the other models, ARTM would be the best alternative.

Conclusion

Based on the observed results, Table 2.5 gives a score to all models on the 11 criteria. Scores are expressed from very bad to very good: $--$, $-$, $+/-$, $+$, $++$. In the right column for each model, the score obtained in the static setting is given, and (if applicable) in the left column, the simulation score is depicted. We can easily conclude that MCLP and MALP are outperformed by the other models. MALP scores badly on the computational results and MCLP scores badly on the simulation results. The objective function of MALP focuses too much on one criterion to provide good scores on different measures. MCLP is easy to compute and provides good single coverage and relatively good response times in the computational results. In the simulation results, however, MCLP scores badly on all criteria. DSM scores a bit better on most criteria when compared to MCLP, but is still outperformed by ARTM, MECXLP and ERTM. These three models provide good results for most criteria, both in the computational and the simulation results, although ARTM is outperformed on backup coverage. MEXCLP performs slightly better on coverage criteria, while ERTM beats MEXCLP on average response times. Both models seem to provide solutions that consider most of the performance measures.

The only criterion for which the difference between the two models is larger is computation time. ERTM takes on average almost four hours, while MEXCLP can be solved within seconds. One can argue that this is not an important criterion, since we are dealing with strategic decisions and most instances could still be solved within a couple of hours. However, when the model is used as a basis for further, more complex, computations or the instance size increases, it might be of importance. In that case, MEXCLP seems more appropriate.

Table 2.5: *Rating of models on different criteria. On the left, the performance based on the simulation study, and on the right the performance based on the computational study. Scores range from very bad to very good:* $--$, $-$, $+/-$, $+$, $++$.

Criterion	MCLP		DSM		ARTM		MEXCLP		MALP		ERTM	
Single coverage		$++$		$+$		$+$		$++$		$-$		$+$
Double coverage		$--$		$++$		$-$		$+$		$+$		$+/-$
Triple coverage		$--$		$--$		$+/-$		$++$		$+$		$+$
Expected coverage	$-$	$-$	$+$	$+$	$+$	$+/-$	$++$	$++$	$+$	$+$	$+$	$+$
8 min threshold	$-$	$-$	$-$	$-$	$++$	$++$	$+/-$	$+/-$	$-$	$-$	$++$	$++$
20 min threshold	$--$	$++$	$+/-$	$++$	$+$	$+$	$++$	$+$	$+/-$	$-$	$++$	$+$
Max. response time	$-$	$++$	$-$	$+$	$+/-$	$+/-$	$+$	$+$	$+/-$	$--$	$++$	$+/-$
Avg. response time		$+/-$		$+/-$		$++$		$+$		$-$		$++$
95% threshold	$--$	$+$	$+/-$	$+/-$	$+/-$	$++$	$++$	$++$	$-$	$--$	$+$	$+$
Avg. ERT	$--$	$-$	$-$	$+/-$	$++$	$++$	$+$	$+$	$-$	$-$	$++$	$++$
Computation time		$++$		$+/-$		$-$		$+$		$+$		$--$

2.1.4 Conclusions and recommendations

In this section, we have compared several ambulance location models. Four of these models focus on maximizing the coverage while the other two focus on minimizing the response time. The models that focus on maximizing coverage are the Maximal Covering Location Problem (MCLP) (Church and ReVelle, 1974), the Double Standard Model (DSM) (Gendreau et al., 1997), the Maximum Expected Covering Location Problem (MEXCLP) (Daskin, 1983), and the Maximum Availability Location Problem (MALP) (ReVelle and Hogan, 1989). The two models that focus on minimizing the response time are the Average Response Time Model (ARTM) and the Expected Response Time Model (ERTM). ARTM is the p-median problem (ReVelle and Swain, 1970) which has also been applied to the ambulance location problem (Dzator and Dzator, 2013). A modified version of ERTM has already been applied to the facility location problem (Snyder and Daskin, 2005), but has not yet been applied to the ambulance location problem.

The purpose of this experiment is to investigate which of the six models performs the best on 11 criteria arising from practice. These criteria include coverage criteria, target and average response time criteria, and computation time. The results show that both MEXCLP and ERTM overall perform well on the 11 criteria. MEXCLP scores the best on the coverage criteria and ERTM is one of the best models when we consider the response time criteria. However, ERTM has the longest computation time which for the largest region amounts to approximately 3.5 days. Therefore, MEXCLP is the best option when coverage and computation times are important. When the response times are most important we would advice ERTM. When response times are most important, but computation time is limited, one could also choose ARTM. However, ARTM scores slightly worse on the coverage criteria, compared to ERTM.

An interesting option to investigate in future research is a combination of MEXCLP and ARTM. ARTM scores the best on the response time criteria, but not so good on the coverage criteria. In addition, the computation time for ARTM is still reasonable. MEXCLP scores the best on the coverage criteria and also relatively good on the response time criteria. By adding ARTM to MEXCLP, the scores on the response time criteria will likely improve.

2.2 Computational analysis of data aggregation error

All static ambulance location models require a discrete set of demand points. However, calls can in principle arise from every location in the area. Consequently, aggregation of demand points is required. Even though Francis et al. (2009) state that "the best aggregation is no aggregation at all", there are multiple reasons for further aggregation of the demand sites. Holmes et al. (2014) give three main reasons: data availability, privacy, and model solvability. In order to apply the location models on a certain aggregation level, at least the travel

times between demand points are required. This data is not always available, especially since ambulances' travel speeds differ from other road users when using optical and auditory signals. Privacy might become an issue when individual residents can be extracted from the data. A computational issue that might demand data aggregation is that computation times typically increase in the number of demand points. To guarantee tractability, data aggregation is required. Francis et al. (2009) add avoiding statistical uncertainty as a reason for aggregation. Larger demand points result in larger sample sizes, which reduces standard deviations in statistical analysis. Despite these reasons for aggregation, an error is made in doing so. This error is caused by demand points that are covered with respect to the aggregated demand points, but are uncovered in the unaggregated setting. Conversely, it can occur that demand points seem to be uncovered, while being covered.

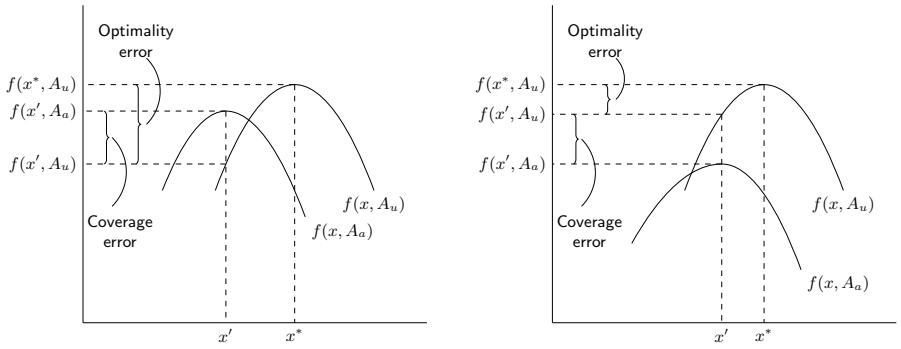
In the Netherlands, the RIVM uses the four digits of a postal code as aggregation level for all their computations that form the basis for the budgets of the different RAVs. The four digits of a postal code roughly correspond to neighborhoods, whereas the complete postal codes, consisting of the four digits and two letters, differ at least for every street. Typically, a street contains multiple postal codes. The combination of a complete postal code and a house number corresponds to a unique address. Travel times based on ambulances driving with warning lights and sirens are available for each pair of four digit postal codes. In this thesis, we will use the four digits of postal codes as our level of aggregation. The first reason to do so is the availability of the data. We do not have reliable travel time data for ambulances on any other level. Second, by using the same data as RIVM, our results can be compared to, and used for, the RIVM computations. Finally, also the tractability of the models benefits significantly from more aggregated demand points. We will see that for some models, we already reach the computational limits when using the four digits only.

In this section, we try to gain insight into the potential error we make by this level of aggregation. We evaluate this error by comparing the results for the most basic static ambulance location model, MCLP (Church and ReVelle, 1974), for six and four position postal codes for two regions for which we have travel time data on six position postal code level.

In the literature, many studies exist that analyze the impact of data aggregation for location models. In particular, the p-median problem (ReVelle and Swain, 1970) appears frequently in these analyses (see, for example, Hillsman and Rhoda (1978); Ballou (1994); Erkut and Bozkaya (1999)). Francis et al. (2009) present an extensive survey on demand point aggregation for location models. They conclude that most literature focuses on median-type models. Coverage models are underrepresented in these studies. However, as a consequence of the all-or-nothing objective inherent to coverage models, these models might suffer more severely than median-type models.

Typically, three types of aggregation errors are distinguished: cost errors, optimality errors, and location errors (Casillas, 1987). The cost error is the difference between the estimated performance of the obtained solution based on

aggregated data and the performance based on unaggregated data. Since for coverage models the performance is measured by the coverage, Daskin et al. (1989) call this the coverage error. The optimality error is the loss in objective value as a result over-aggregation of data. To compute this, the optimal objective value of the unaggregated model is compared with the objective value of the aggregated solution in the unaggregated model. Note that the optimality error always depicts a loss in coverage, whereas the coverage error can both be an over- or underestimation of the coverage. Figure 2.3 (Daskin et al., 1989) shows the two errors in the two cases. The location error is defined as the inadequate spacial distribution of the bases as a result of the aggregation and is harder to measure. Often, conclusions are drawn from geographical display of solutions rather than objective measures (Daskin et al., 1989). Erkut and Bozkaya (1999) further observe that this error is highly dependent on the selected solution, as location models have often a large set of (near-)optimal solutions. In line with these observations, we only include the first two error types in this analysis.



(a) *Negative coverage error* **(b)** *Positive coverage error*

Fig. 2.3: *(Daskin et al., 1989) Aggregation errors. x' and x^* denote the solution obtained with aggregated and unaggregated demand points, respectively. A_a denotes the high level of aggregation, whereas A_u denotes the unaggregated level. $f(x, A)$ gives the objective value of solution x based on aggregation level A.*

Besides the mentioned coverage and optimality error, another error occurs if the set of demand points coincides with the set of potential base locations. In that case, by aggregating demand sites, the feasible set is reduced (Hodgson et al., 1997). Although this error is caused by the aggregation of demand points, this should be measured separately (Erkut and Bozkaya, 1999).

Even though the main focus in the literature is on p-median related models, some studies do address the demand point aggregation error for coverage-based models. Daskin et al. (1989) consider MCLP and conclude that high levels of aggregation do not drastically affect the solution quality. Current and Schilling

(1990) show that coverage models are more sensitive for aggregation errors than median-type models. They introduce rules for demand point aggregation to limit the error. Holmes et al. (2014) also conclude that models with more gradual objective functions are less sensitive for aggregation errors. In their experiments, a version of MCLP with probabilistic travel times appears to be highly insensitive to significant aggregation. For the classical MCLP with deterministic travel times, serious optimality errors are observed.

2.2.1 Experimental setup

To quantify the error introduced by aggregating demand points on four digit postal codes rather than six position postal codes, we apply the well-known MCLP to two ambulance regions for which we have travel time data on complete postal code level: North-East-Gelderland and Zeeland. Table 2.6 shows the number of postal codes, the average size, and the average population for both cases. We see that the six position data is extremely detailed. Note that the travel time data contains the travel time between each pair of demand points. Hence, for North-East-Gelderland, this has $24,089^2 \approx 580,000,000$ entries.

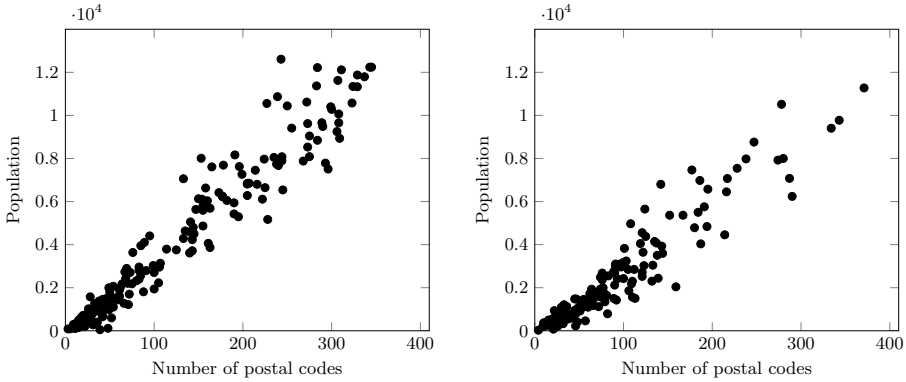
Table 2.6: *Postal code characteristics for the two considered regions.*

Region	Four digit postal codes			Six position postal codes		
	# Points	Avg. size	Avg. population	# Points	Avg. size	Avg. population
N-E Gelderland	200	13.7 km^2	4,049	24,089	0.11 km^2	34
Zeeland	153	11.7 km^2	2,493	14,165	0.13 km^2	27

For each aggregated demand point, we select a centroid that is used in the case with four digit postal codes. We select this centroid by taking the weighted average of the x and y coordinates of the six position postal codes in a four digit postal code area. This gives a point in the center of the four digit postal code area. Based on this center, we select the six position postal code that is closest to the obtained point, according to the Euclidean distance, as centroid. In principle, we could include all postal codes as potential base locations, but then the feasible set of the aggregated and unaggregated version would differ. To avoid this effect, we first limit the bases to the selected centroids.

In MCLP, we can assign weights to different demand points. These weights should indicate the importance of covering a particular demand point. Often, the population of an area is used. Since the population of a four digit postal code area is approximately proportional to the number of postal codes in that area (see Figure 2.4), we take the number of postal codes as weights for the aggregated demand points. We run the models for all values of p, which is the maximum number of base locations that can be selected, up to the number required for full coverage.

By adopting notation from Francis et al. (2009), we denote the optimal set of base locations for the aggregated and unaggregated case by x' and x^*, re-



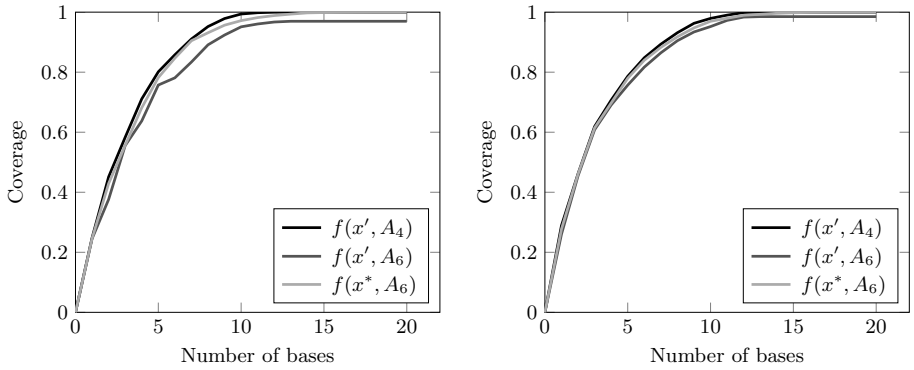
(a) *N-E Gelderland* **(b)** *Zeeland*

Fig. 2.4: *Correlation between population and number of six position postal codes in a four digit postal code for the two considered regions.*

spectively. We distinguish two levels of aggregation: A_4 and A_6, where A_4 is the four digit postal code aggregation level and A_6 the six position postal code level. Let f be a function that takes a solution x and an aggregation level A as input and returns the coverage of solution x in aggregation regime A. For example, $f(x^*, A_6)$ denotes the unaggregated coverage of the optimal solution in the unaggregated case. Now, we define the coverage error as $f(x', A_6) - f(x', A_4)$, which is the coverage estimation error as a result of demand point aggregation. The optimality error, sometimes called coverage loss, is defined as $f(x^*, A_6) - f(x', A_6)$. See Figure 2.3 for a graphical illustration of the two errors.

As in the aggregated case only the centroids have to be covered, this case is typically too optimistic in calculating the required number of bases for high levels of coverage. Whenever complete coverage is obtained, increasing the number of bases does not have any value. However, in the unaggregated case, more bases are required for full coverage. To overcome this, we introduce two more pessimistic versions of the aggregated model. First, we add the maximum travel time from a centroid to any points in the same four digit postal code area to all travel times to this centroid. This approach guarantees that points covered in the aggregated model are covered in the unaggregated model as well. In other words, the coverage error is non-negative. Since this approach might be overly pessimistic, we also include a second alternative where we add the average travel time from the centroid, instead of the maximum. The second approach does not guarantee the coverage error to be non-negative.

Finally, to measure the impact of a smaller feasible set by limiting the number of potential base locations to the centroids of four digit postal code areas, we compare the solution of the unaggregated problem with limited base set to the case with all potential bases included.



(a) *N-E Gelderland* **(b)** *Zeeland*

Fig. 2.5: *Coverage for aggregated and unaggregated solutions in different aggregation regimes for the two considered regions.*

2.2.2 Experimental results

In the first experiment, we compare the solution obtained by the four digit postal code aggregation with six position postal code aggregation to quantify the coverage error and the optimality error (or coverage loss). Table 2.7 and Figure 2.5 show $f(x', A_4)$, $f(x', A_6)$, and $f(x^*, A_6)$, as defined before. When comparing $f(x', A_4)$ and $f(x^*, A_6)$ it seems that the two versions of the model report similar coverage. However, if we evaluate the solution x' in the unaggregated regime A_6, we see that the coverage is typically overestimated. In other words, the coverage error, $f(x', A_6) - f(x', A_4)$ is positive. In particular for cases with large number of bases, this difference is significant. For both regions, full coverage is provided with 13 bases in the aggregated case. However, with unaggregated demand points, these solutions only yield a coverage of 0.970 and 0.985. Since, according to aggregated data, full coverage is provided, the model cannot be used to find solutions with higher coverage. This is particularly unsatisfactory since the single coverage model provides an upper bound on the potential coverage and EMS providers typically aim for very high single coverage. If selections of base locations with close to complete coverage are required, the aggregation of demand points does not seem appropriate. In the extreme case where we require complete coverage, the aggregated version suggest a minimum of 13 bases, whereas the unaggregated version requires at least 20 bases. Note that in the region of Zeeland, complete coverage cannot be obtained, since two postal codes cannot be reached within the time threshold from any of the selected centroids.

These results suggest that for large number of bases, the aggregation of demand points results in too optimistic results. To be more conservative, we provide two alternatives. We add the maximum and average travel time from the centroid to points in the same four digit postal code to the travel time to the centroid. Let x^m and x^a denote the solutions obtained in the two cases. Furthermore,

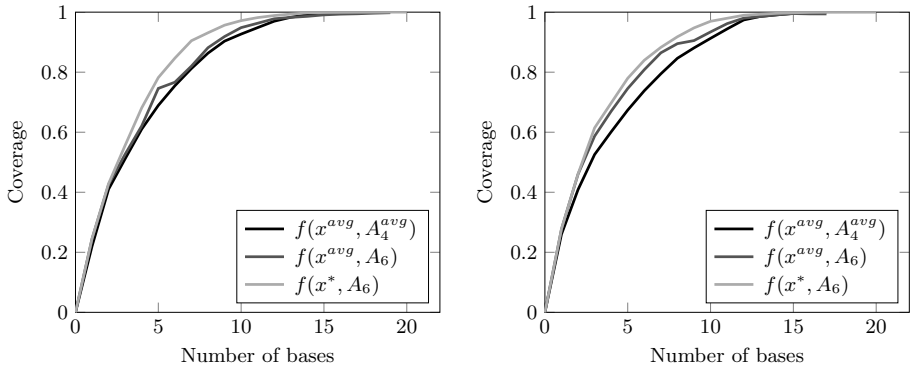
Table 2.7: *Coverage for different levels of demand point aggregation. x' and x^* denote the aggregated and unaggregated solution, respectively. A_4 and A_6 represent the aggregation level used for the coverage computation.*

# Bases	N-E Gelderland			Zeeland		
	$f(x', A_4)$	$f(x', A_6)$	$f(x^*, A_6)$	$f(x', A_4)$	$f(x', A_6)$	$f(x^*, A_6)$
1	0.248	0.245	0.246	0.289	0.260	0.278
2	0.451	0.376	0.429	0.462	0.454	0.461
3	0.583	0.555	0.557	0.619	0.608	0.614
4	0.711	0.638	0.681	0.707	0.691	0.697
5	0.801	0.757	0.782	0.786	0.757	0.780
6	0.858	0.781	0.847	0.848	0.817	0.840
7	0.911	0.834	0.905	0.893	0.865	0.883
8	0.952	0.891	0.931	0.932	0.905	0.918
9	0.979	0.924	0.957	0.963	0.934	0.948
10	0.994	0.951	0.972	0.979	0.952	0.970
11	0.998	0.961	0.982	0.990	0.973	0.980
12	0.9998	0.967	0.989	0.999	0.984	0.990
13	1	0.970	0.993	1	0.985	0.993
14			0.998			0.996
15			0.999			0.998
16			0.999			0.999
17			0.9996			0.9996
18			0.9999			0.9997
19			0.99996			0.9998
20			1			0.9999

let A_4^m and A_4^a denote the corresponding models. By looking at Table 2.8 and Figure 2.6, we first observe that adding the maximum travel time from the centroid is too pessimistic. According to this data, 20 bases provide only about 86% coverage. However, the provided solutions yield significantly higher coverage in the unaggregated version. Adding the average travel time within a postal code seems to give more realistic solutions. Especially for large number of bases, the optimality error is small. Also the required number of bases for extremely high coverage is more in line with the unaggregated case. This suggests that by adding a correction to the travel times that depends on the size of an aggregated demand point, aggregated demand points might be sufficient to obtain reasonable solutions. Holmes et al. (2014) already note that, for a high number of bases, a more stringent coverage standard results in a smaller coverage loss. Note that for this approach, we do require travel time data to compute the average travel time within a four digit postal code area. If this data is not available, as is the case for most regions in the Netherlands, we should find other ways to compute this number in order to apply this alternative approach.

Table 2.8: *Coverage for alternative travel times. x^m and x^a denote the solutions for the case in which the maximum and average travel time within postal code is added to travel time to the centroid, respectively. A_4^m and A_4^a denote the corresponding coverage regimes.*

# Bases	N-E Gelderland					Zeeland				
	$f(x^m, A_4^m)$	$f(x^m, A_6)$	$f(x^a, A_4^a)$	$f(x^a, A_6)$	$f(x^*, A_6)$	$f(x^m, A_4^m)$	$f(x^m, A_6)$	$f(x^a, A_4^a)$	$f(x^a, A_6)$	$f(x^*, A_6)$
1	0.166	0.235	0.221	0.245	0.246	0.219	0.277	0.260	0.277	0.278
2	0.286	0.416	0.410	0.426	0.429	0.329	0.455	0.409	0.460	0.461
3	0.370	0.528	0.512	0.527	0.557	0.413	0.582	0.526	0.587	0.614
4	0.444	0.603	0.611	0.621	0.681	0.462	0.647	0.601	0.670	0.697
5	0.498	0.634	0.689	0.746	0.782	0.507	0.709	0.674	0.746	0.780
6	0.539	0.694	0.755	0.767	0.847	0.549	0.737	0.738	0.808	0.840
7	0.576	0.805	0.812	0.820	0.905	0.587	0.803	0.794	0.865	0.883
8	0.612	0.830	0.863	0.882	0.931	0.624	0.813	0.846	0.895	0.918
9	0.641	0.842	0.903	0.919	0.957	0.654	0.873	0.881	0.905	0.948
10	0.673	0.852	0.927	0.948	0.972	0.681	0.912	0.913	0.934	0.970
11	0.703	0.862	0.949	0.963	0.982	0.704	0.925	0.944	0.962	0.980
12	0.731	0.896	0.970	0.978	0.989	0.726	0.925	0.974	0.980	0.990
13	0.756	0.904	0.983	0.983	0.993	0.746	0.949	0.986	0.985	0.993
14	0.776	0.918	0.991	0.986	0.998	0.766	0.949	0.991	0.991	0.996
15	0.793	0.925	0.994	0.991	0.999	0.785	0.954	0.996	0.995	0.998
16	0.812	0.923	0.996	0.993	0.999	0.804	0.975	0.998	0.994	0.999
17	0.828	0.922	0.998	0.995	0.9996	0.819	0.977	1	0.994	0.9996
18	0.842	0.926	0.999	0.997	0.9999	0.830	0.977			0.9997
19	0.857	0.932	1	0.998	0.99996	0.841	0.982			0.9998
20	0.869	0.932			1	0.852	0.991			0.9999



(a) *N-E Gelderland* **(b)** *Zeeland*

Fig. 2.6: *Coverage for alternative approach based on aggregated demand points for the two considered regions.*

Besides the loss in coverage by aggregating demand points, another effect plays a role if all demands points are included as potential base locations. Namely, the set of feasible solutions is restricted by going from six position postal codes to four position postal codes. To quantify this effect, we apply MCLP to the case where all postal codes are included as demand points and as potential base locations. Note that computationally this is very demanding, as the number of postal codes is significantly larger than the number of aggregated postal codes (see Table 2.6). Let x^{**} denote the optimal solution to the case with the larger set of potential base locations. Table 2.9 and Figure 2.7 compare the results of the two sets of potential bases. By definition, x^{**} gives a coverage that is at least as high as the coverage provided by x^*, since its feasible set is strictly larger. Even though the results show that this limitation of the solution space has an impact on the coverage, it should be noted that the larger set of potential bases might be too detailed. Especially if the location is as detailed as a part of a street, it is unrealistic to assume that bases will be stationed at these exact locations. As Daskin et al. (1989) state, "it is unlikely that any decision maker will accept exactly that location". Thus, it might be more realistic to consider neighborhoods, or four digit postal codes, as potential bases.

2.2.3 Conclusions

As all ambulance location models require a discrete set of demand points, it is important to consider what level of aggregation should be used. Even though aggregation in principle leads to aggregation errors, such as coverage errors and optimality errors, some higher level of demand point aggregation might be necessary or preferable. In the Netherlands, the standard level of aggregation for ambulance location models is the four digits of a postal code. These roughly

Table 2.9: *Coverage for restricted and unrestricted set of potential base locations for a different maximum number of base stations. x^* denotes solution with restricted set of bases, x^{**} corresponds with the unrestricted set of bases.*

# Bases	N-E Gelderland		Zeeland	
	$f(x^*, A_6)$	$f(x^{**}, A_6)$	$f(x^*, A_6)$	$f(x^{**}, A_6)$
1	0.246	0.274	0.278	0.312
2	0.429	0.461	0.461	0.487
3	0.557	0.613	0.614	0.643
4	0.681	0.723	0.697	0.736
5	0.782	0.830	0.780	0.819
6	0.847	0.885	0.840	0.881
7	0.905	0.939	0.883	0.928
8	0.931	0.961	0.918	0.957
9	0.957	0.979	0.948	0.974
10	0.972	0.987	0.970	0.987
11	0.982	0.994	0.980	0.992
12	0.989	0.997	0.990	0.996
13	0.993	0.998	0.993	0.998
14	0.998	0.9997	0.996	0.999
15	0.999	0.99996	0.998	0.9996
16	0.999	1	0.999	0.9999
17	0.9996		0.9996	1
18	0.9999		0.9997	
19	0.99996		0.9998	
20	1		0.9999	

correspond to neighborhoods. If travel time data would be available, an alternative would be to use the complete postal code, consisting of four digits and two letters. We evaluated the impact of using a higher level of aggregation by comparing the results of the well-studied MCLP for four and six position postal codes.

We see that for a small number of bases, the impact is limited. However, for a number of bases close to the number required for full coverage, a significant error is made at this level of aggregation. Also, the required number of bases for a given high level of coverage is highly underestimated.

As an alternative to avoid these errors, we propose to add a correcting factor to the travel times to compensate for the travel time within an aggregated demand point. Since travel times within an area increase with the area size, we let this correction depend on the size of a demand point. A first, extremely conservative, option is to add the maximum travel time from the centroid of an area to any postal code in the area to all travel times to the centroid. In this case, demand points covered according to the aggregated data are guaranteed to be covered for the unaggregated data. This approach turns out to be too pes-

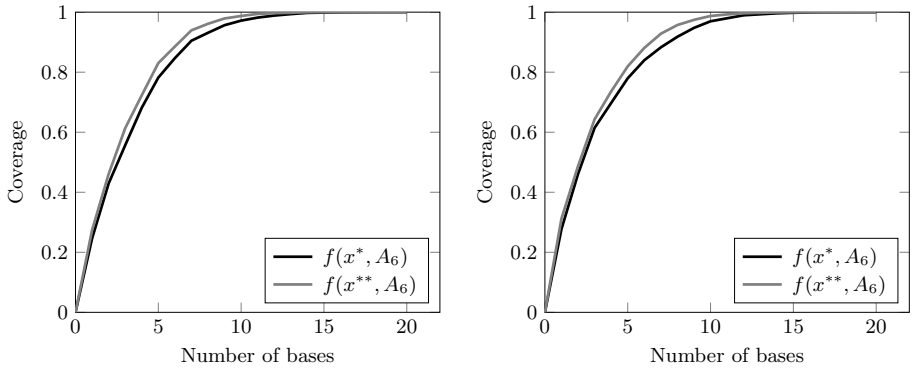


Fig. 2.7: *Coverage for solutions with restricted and unrestricted set of potential bases for the two considered regions.*

simistic to yield reasonable results. Alternatively, we find that adding the average rather than the maximum gives solutions very close to the unaggregated results. Both the coverage error and the optimality error are small for this approach. Furthermore, the number of bases required for a given level of coverage is more realistic.

The aggregation of data can also lead to a smaller set of potential bases, which reduces the solution space. Consequently, lower levels of coverage are obtained for each given number of bases. However, one might argue that allowing bases in each postal code is overly optimistic. There is only a small probability that decision makers can indeed select the exact location provided by the model. It is more likely that the neighborhood of the proposed solution is considered for the base location. So, it might be better to consider a smaller set of potential base locations.

In the remainder of this thesis, we will use the four digits of a postal code as aggregation level for the computations. One reason is the availability of data. For most regions, no travel time data on six position postal code level is available. Additionally, by using the same data as the RIVM, our results can be compared. In Chapter 3, we do perform experiments with the correction factor as proposed in this section.

3

Time-dependent MEXCLP with start-up and relocation cost

3.1 Introduction

Adequate location of the ambulance base stations has a significant impact on the response times of EMS providers. As reviewed in Section 1.3, much research is conducted on the optimal locations for base stations. Despite significant fluctuation in the characteristics of EMS systems in practice, most models assume that the system characteristics used as input are static throughout the day. For example, call volumes, travel speeds, and ambulance availability are typically far from stable. Figure 3.1 shows the call volume fluctuation during a 24 hour period for A1 calls in the Netherlands.

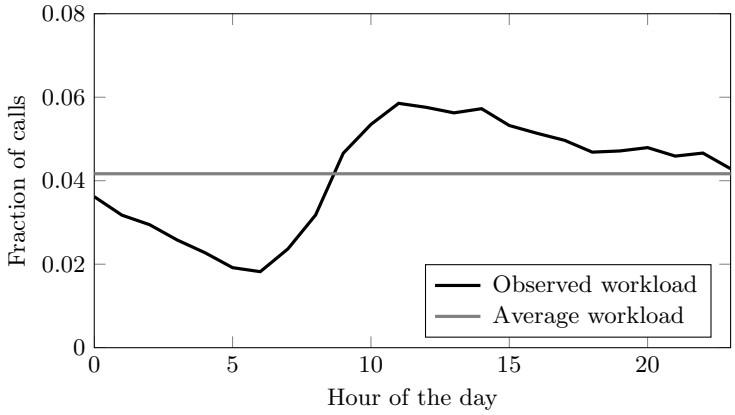


Fig. 3.1: *Distribution of life-threatening emergency calls (A1) over the day (Ambulancezorg Nederland, 2013).*

We see that during the night, the call intensity is only half the daily average, whereas about 50% more calls arise during the afternoon. At least three papers

take these fluctuations into account in finding good base locations. Repede and Bernardo (1994) extend MEXCLP to incorporate time-dependent inputs. They introduce multiple time periods and maximize the expected coverage over the day. Since there is no connection between the different time periods in their model, the problem can be solved independently for each time period. As a consequence, large differences in the location of ambulances can occur between time periods, which can result in high costs for the EMS provider. Second, Rajagopalan et al. (2008) compute the number of ambulances needed to satisfy a certain coverage requirement in a multiperiod setting. The coverage requirement is modeled by the hypercube model by Larson (1974). Finally, Schmid and Doerner (2010) introduce an extension of the Double Standard Model (Gendreau et al., 1997) to include time-dependent travel times. To partly overcome the above-mentioned problem, a penalty is added for each relocation of an ambulance between two time periods. A relocation means that an ambulance is located at a different location in two consecutive time periods. This penalty does capture the inconvenience for the crew to move between bases during a shift. However, the costs to arrange facilities at the selected locations are not captured.

In this chapter, we combine ideas from these papers and add a penalty for the selected number of base locations. Since Section 2.1 indicates that MEXCLP is an appropriate building block for more advanced models, we consider a time-dependent version of MEXCLP. In contrast to Repede and Bernardo (1994), we do incorporate dependencies between time periods. As in Schmid and Doerner (2010), we include a penalty for ambulance relocations. Since the construction, or rental, of a base location incurs high costs, we further add a penalty for each location that is used as a base station during at least one time period.

3.2 Model formulation

We consider a time-dependent version of MEXCLP in which we define a set $T = \{t_1, t_2, ..., t_{|T|}\}$ of time periods. As in previous descriptions of ambulance location models, we denote the set of potential base locations by I and the set of demand points by J. Compared to Schmid and Doerner (2010), we do not only assume time-dependent travel times, but also time-dependent demand and ambulance availability. As a consequence, we no longer have one demand d_j for each demand point $j \in J$, but we have d_{jt} which is the demand generated from demand point j during time period t. The available number of ambulances during time period t is denoted by p_t. Since the ambulance distribution and the resulting coverage may differ among time periods, we have to adjust the decision variables x and y. Variable x_{it} will now denote the number of ambulances located at base location i during time period t, while y_{jkt} will indicate whether demand point j is covered by at least k ambulances during time period t. As a result of the time-dependent demand, service time, and ambulance availability, we also have a time-dependent busy fraction, q_t. The set of base locations that can cover demand point j in time period t is denoted by I_{jt}.

In the model, we take the relationship between the different time periods into account by adding a penalty for the number of relocations between the time periods. Therefore, we introduce the binary variable $r_{ii't}$ which indicates whether an ambulance is relocated from location i to location i' at the end of time period t. The penalty for a relocation is γ. In practice, there are costs involved in making use of a base location. These costs occur when a base location is used during at least one time period. We add a penalty, β, for each location that is used. Note that these cost can vary per base location. For example, it might be cheaper to make use of an existing base location than to build a new one. This can be modeled by a location-dependent penalty β_i. To keep track of the opened base locations we introduce the binary variable f_i indicating whether base location i is used during at least one time period. To ensure a correct value for these variables, we need so-called big-M constraints. These constraints make use of a sufficiently large constant M. The model can be formulated as follows:

$$\max \quad \sum_{t\in T}\sum_{j\in J}\sum_{k=1}^{p_t} d_{jt}(1-q_t){q_t}^{k-1}y_{jkt} - \beta\sum_{i\in I} f_i - \gamma\sum_{i\in I}\sum_{i'\in I}\sum_{t\in T} r_{ii't},$$

$$\text{s.t.} \quad \sum_{i\in I_{jt}} x_{it} \geq \sum_{k=1}^{p_t} y_{jkt} \qquad \forall j\in J, t\in T, \tag{3.1}$$

$$\sum_{i\in I} x_{it} \leq p_t \qquad \forall t\in T, \tag{3.2}$$

$$\sum_{t\in T} x_{it} \leq M f_i \qquad \forall j\in J, \tag{3.3}$$

$$x_{it} + \sum_{i'\in I} r_{i'it} - \sum_{i'\in I} r_{ii't} = x_{i(t+1)} \qquad \forall t\in T\backslash t_{|T|}, i\in I, \tag{3.4}$$

$$x_{i|T|} + \sum_{i'\in I} r_{i'i|T|} - \sum_{i'\in I} r_{ii'|T|} = x_{i1} \qquad \forall i\in I, \tag{3.5}$$

$$x_{it}, r_{ii't} \in \mathbb{N} \qquad \forall i,i'\in I, t\in T, \tag{3.6}$$

$$y_{jkt}, f_i \in \{0,1\} \qquad \forall i\in I, j\in J, \quad k\in\{1,...,p_t\}, t\in T. \tag{3.7}$$

The objective function of this model consists of three terms. The first term calculates the expected coverage over all demand points and all time periods. Recall that the marginal coverage of the k-th ambulance, given independent ambulance availability with busy fraction q, is $(1-q)q^{k-1}$. The second term penalizes the number of opened locations. Finally, the third term subtracts a penalty for the number of relocations between time periods. Constraints (3.1) ensure that a demand point is only covered by at least k ambulances, if at least k ambulances can reach this demand point within the target response time. Constraints (3.2) limit the number of ambulances in each time period. Constraints (3.3) state that ambulances can only be located at locations that are opened, i.e., $f_i = 1$. For this constraint to be valid, M should have at least the value

of the left hand side. However, as discussed in Section 1.2, a too high value for M will result in a weak LP relaxation and an increase in computation time. In Section 3.4, we address how to determine appropriate values for M. Constraints (3.4) and (3.5) ensure that the $r_{ii't}$'s have the correct value. In case the number of ambulances is not constant over the day, we need a dummy base location to which off-duty ambulances can be assigned.

3.3 Computational results

In order to test the model, we apply it to the region of Amsterdam-Waterland in the Netherlands. After a description of the data, we compare the optimal solution according to the model with the current set of base locations in the region. Next, we compare the results with a time-independent version of MEXCLP. The impact of the penalties is evaluated by analyzing the outcomes for different values of β and γ. Finally, we incorporate the data aggregation error correction strategy as proposed in Section 2.2.

3.3.1 Data description

The considered region for the case study has a population of approximately 1.1 million. We divide the region into 161 demand points based on four digit postal codes. All postal codes are considered as a potential location for a base station. The set T consists of 12 time periods of two hours. For the average travel times between the different nodes, we use a travel time model developed by Kommer and Zwakhals (2011). In the model, the driving speeds on different road types is estimated base on historical driving speeds of ambulances. Based on these estimations, the average travel time between nodes is computed. Since the model is time-independent, it does not distinguish different times of the day. In order to get some indication of the fluctuation in the travel times, we use the average drive time to a call. For each time period, we compute the average drive time to a patient. This average is compared with the average over the entire day. By this procedure, we get that the average travel time varies between 97% and 108% of the average travel speed. Note that this might be a conservative estimation, since Schmid and Doerner (2010) report a travel speed fluctuation up to 25% of the daily average. Further research is required to obtain better estimates of the travel times throughout the day. As response time target, we consider 15 minutes, as is stated in Dutch law for the most urgent calls. To account for the average pre-trip delay in the region of Amsterdam, we add four minutes to the travel times. Note that in the previous chapter, we used a pre-trip delay of three minutes. The three minutes corresponds to the average pre-trip delay in the Netherlands, whereas the four minutes corresponds to the average in the region of Amsterdam.

For the weights, d_{jt}, we use the historical data regarding the number of calls per part of the day. For each two-hour time interval, we have the number of

emergency calls per demand point in the years 2008-2011. The busy fraction is based on data regarding the average call duration, the available number of ambulances, and the number of calls. Since all these components are considered time-dependent, we also have a time-dependent busy fraction. The characteristics of the different time periods are shown in Table 3.1.

Table 3.1: *Call volumes, travel times, ambulance availability, and busy fraction for 12 two-hour time periods.*

Interval	# Calls	Travel time	# Ambulances	Busy fraction
00:00-02:00	27,382	1.0054	13	0.2678
02:00-04:00	22,214	0.9837	13	0.1973
04:00-06:00	17,932	1.0365	13	0.1625
06:00-08:00	19,138	1.0817	13	0.2275
08:00-10:00	37,438	1.0084	18	0.3811
10:00-12:00	45,516	0.9707	18	0.4912
12:00-14:00	48,526	0.9799	18	0.5483
14:00-16:00	49,526	0.9812	18	0.5151
16:00-18:00	36,342	1.0030	17	0.4866
18:00-20:00	42,808	0.9877	17	0.4200
20:00-22:00	40,506	0.9625	17	0.3639
22:00-24:00	35,114	0.9993	17	0.2837

3.3.2 Base case

All computations are executed on a 2.9 GHz Intel(R) Core(TM) i7-3520M laptop with 8 GB of RAM. CPLEX 12.5 is used as our solver (ILOG, 2009). The penalty for opening a new facility is set to 0.5% of the total number of calls and the penalty for relocating an ambulance is set to 0.0005% of the total number of calls. These values imply that a new base location is only opened if it results in a coverage increase of at least 0.5 percentage points. Section 3.3.5 gives results for different values.

Given these settings, which we will call the base case, the model gives a solution with an expected coverage of 0.9763 within 15 minutes. In this solution, five base stations are used and three ambulances are relocated between time periods. Even though coverage within 15 minutes is the main target for ambulance providers in the Netherlands, it is interesting to evaluate the provided coverage with respect to different targets. Although this is not reflected in the objective function, better service is provided to patients that are reached in less than the required 15 minutes. On the other hand, uncovered patients should not be completely neglected. Figure 3.2 shows the coverage of the provided solution for different response time targets. We see that more than 99% of the calls is covered within a time threshold of 18 minutes, whereas 50% of the calls is reached in less than 10 minutes.

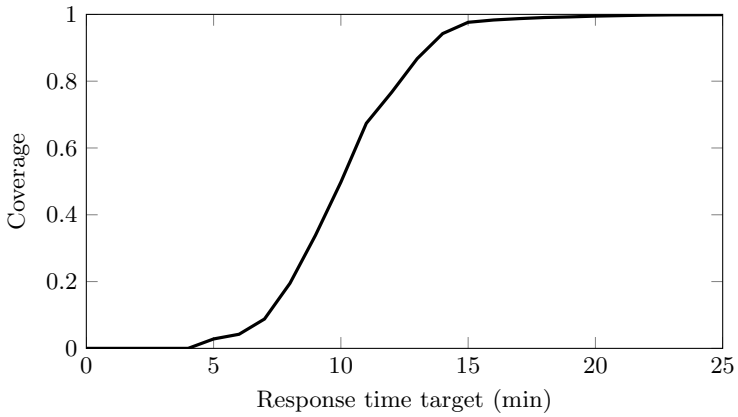


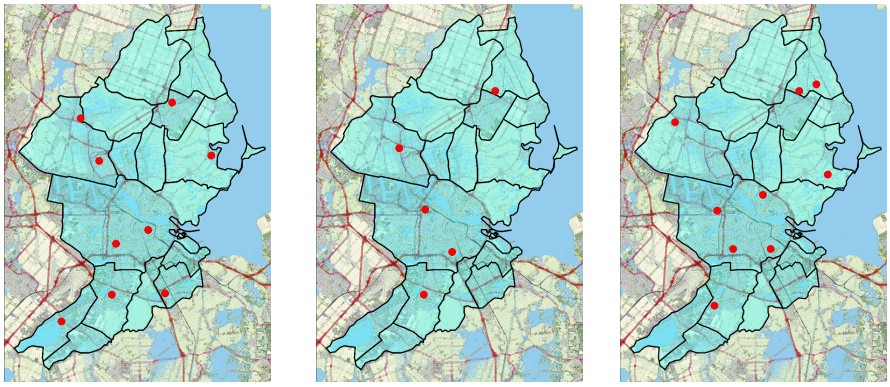
Fig. 3.2: *Coverage with respect to different response time targets of the solution obtained with a threshold of 15 minutes.*

To get insight into the potential improvement that can be achieved, we compare the solution of the model with the current set of base stations. To do so, we apply the model with a fixed set of base locations. The model determines the optimal ambulance distribution over the current bases. This gives a coverage of 0.9477 with six relocations. The optimal solution, with respect to the model, thus provides better coverage with fewer locations and relocations.

For further analysis of the solution, we consider the geographical distribution of the base locations. The selected locations are plotted on a map of the region in Figure 3.3. Figure 3.3a shows the locations of the current bases, while the optimal locations are shown in Figure 3.3b. Our first observation based on this map is that the base locations tend to be located far away from the border of the region. In fact, ambulances are typically located as far away from the border as possible, while still covering the demand points at the border. This is a direct consequence of a coverage-based objective and the fact that no value is given to coverage of demand points in neighboring regions. As a result, the response times will typically be larger for calls at the border of the region than for calls at the center of the region. Another observation that can be made is that not all demand points can be reached from a base within the response time target. It appears that a better objective value can be obtained if some demand points are left uncovered in order to provide backup coverage to others. In the next section, we have a closer look at this issue.

3.3.3 Uncovered demand points

One of the characteristics of a probabilistic model is that a coverage of 95% does not mean that 5% of the demand points is not covered. In fact, a part of the 5% is uncovered because of ambulance unavailability. For example, given a busy



(a) *Current locations* **(b)** *Optimal locations* **(c)** *Cover at least once*

Fig. 3.3: *Maps of different sets of base locations. 3.3a shows the current set of base locations. 3.3b shows the optimal set of base locations with respect to the base case. 3.3c gives the optimal base locations if each demand point should be covered by at least one ambulance in each time period.*

fraction of 0.5, a demand point that is covered by two ambulances is covered for 75%. From a fairness perspective, it is better to have 100% of the demand covered for 95% than 95% of the demand covered for 100%. For that reason, it is interesting to investigate what fraction of the demand is not covered at all. In the solution provided by the model, 0.62 percent of the calls is completely uncovered. The remaining 1.76 percent is uncovered as a result of ambulance unavailability. A total of 18 demand points are not covered in each time period.

It may be considered as unfair that some demand points are uncovered in order to provide backup coverage to other demand points. From this point of view, it might be interesting to see what overall coverage can still be obtained if we require all demand points to be covered by at least one ambulance in each time period. Therefore, we conduct an experiment in which we add constraints (3.8) to the model.

$$\sum_{i \in I_{jt}} x_{it} \geq 1 \qquad j \in J, t \in T \tag{3.8}$$

We see (Table 3.2) that by requiring each demand point to be covered at least once in each time period, we lose about 0.5% of coverage and we need 4 additional base locations. Furthermore, we see that the number of relocations increases significantly. The selected locations in this solution are plotted in Figure 3.3c.

Table 3.2: *Objective value, expected coverage, number of locations, and number of relocations of optimal solution with or without the constraint of coverage of all demand points in each time period.*

	Base case	**Cover at least once**
Objective value	411,368	400,393
Expected coverage	0.9763	0.9710
# Locations	5	9
# Relocations	3	12

3.3.4 Time-dependent versus time-independent MEXCLP

To analyze the importance of the time-dependent aspects of the model, we conduct the following experiment. We first solve the model assuming that the travel speed, call volume, and busy fraction are constant throughout the day. For all these characteristics, we take the average over all time periods in the time-dependent case. Since, for comparison, we need a feasible solution for the time-dependent case, we keep the fluctuation in the number of available ambulances. The optimal solution for this problem is analyzed in the time-dependent environment so that we obtain the solution value and the expected coverage. We compare this with the original solution. Table 3.3 shows that by not taking into account the time-dependency, we lose some coverage. Two out of the five bases are located differently. Figure 3.4 further shows that the coverage predicted by the time-independent version does not correspond to the more realistic, time-dependent, variant. The coverage is over- or underestimated up to two percentage points.

Table 3.3: *Results of the time-dependent and time-independent model evaluated in the time-dependent environment.*

	Time-dependent	**Time-independent**
Objective value	411,368	410,762
Expected coverage	0.9763	0.9749
# Locations	5	5
# Relocations	3	0

3.3.5 Impact of penalties

The main difference between our model and the TIMEXCLP model introduced by Repede and Bernardo (1994) is that we take the relationship between time periods into account. We do this by adding a penalty for opened locations and relocations. To examine the impact of these penalties, we compare the results of

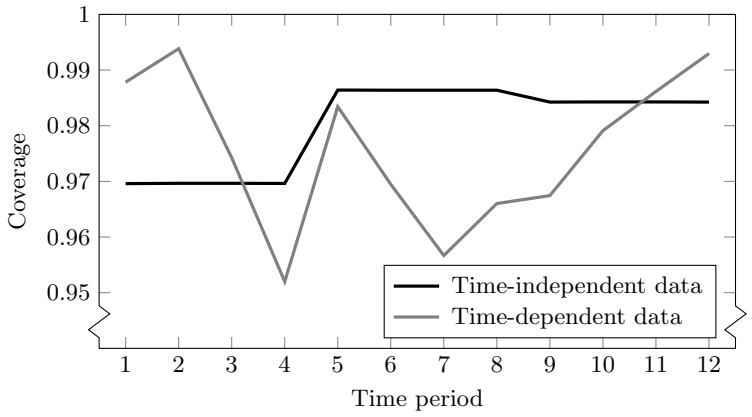


Fig. 3.4: *Coverage for each time period of the optimal solution for time-independent data with respect to time-dependent and time-independent data.*

our model with the result obtained without the penalties. In Table 3.4, we see that without the penalties we can increase coverage by 0.69 percentage points. However, we need 20 additional locations and 51 additional relocations. Clearly, this is not a good solution in practice.

Table 3.4: *Results of solutions of the base case and the case with* $\beta = \gamma = 0$.

	Base case	**No penalties**
Expected coverage	0.9763	0.9832
# Locations	5	25
# Relocations	3	54

As defining the right values for these penalties might be difficult, we run the model for different values of β. Clearly the lower the value of β, the more bases will be opened and the higher the coverage. Figure 3.5 presents the coverage as a function of β. We obtain a step function where each jump represents opening an additional base. We see that opening the sixth base yields a coverage increase of 0.35 percentage point. Decision makers should make the trade-off between this better performance and the higher cost that a new base will incur.

3.3.6 Data aggregation error correction

In Section 2.2, we showed that using four position postal codes as demand point aggregation level might result in suboptimal solutions. Furthermore, the coverage

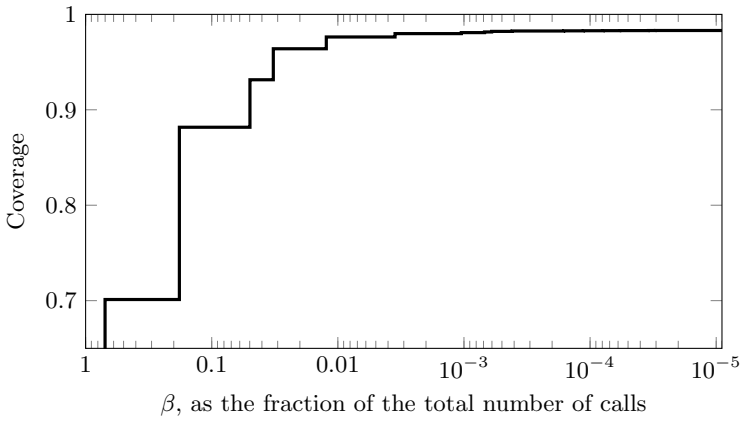


Fig. 3.5: *Coverage provided solutions for different values of β. γ is fixed at 0.0005% of total number of calls.*

provided by the solution is typically overestimated. We proposed a correction based on the average travel time within a demand point. In this section, we evaluate the impact of this approach on the new model. Since for the region of Amsterdam, we do not have travel time data on six position postal code level, we have to compute the average travel times differently. As x and y coordinates of postal codes are available, we use the Euclidean distance as an approximation for the travel time within a postal code. We assume a constant speed of 50 kilometers per hour. Using these estimated travel times, we apply the same method as described in Section 2.2.1.

Table 3.5: *Results for solution of model with and without the data aggregation correction. x and x' represent the solution with and without the correction, respectively. Corr or No corr indicates the regime in which the objective value and expected coverage are evaluated.*

	(x , *No corr*)	(x', *Corr*)	(x', *No corr*)
Objective value	411,368	406,172	408,688
Expected coverage	0.9763	0.9693	0.9751
# Locations	5	6	6
# Relocations	3	6	6

Table 3.5 shows the solutions of the two versions. We see that we need one more base and three more relocations in the new situation. The expected coverage is reduced to 0.9693. The coverage loss with respect to the version without the correction is limited to 0.0012. Similar to the original solution, we have two bases

in the Northern part of the region, and one base in the Southern part. Now, we have three bases to cover the central part of the region, which contains the city of Amsterdam.

3.4 Computational aspects

In this section, we address some issues related to the computations. In the implementation, we made some adjustments to the mathematical formulation of the model in order to reduce the number of variables and decrease the upper bound obtained by the LP relaxation of the model. First, we replaced the variables $r_{ii't}$ by two variables r_{i1t} and r_{i2t} indicating the number of relocations to and from base location i after time period t, respectively. By doing so, we only obtain the number of relocations and not the specific relocations that are required. This can be done, because we assume the penalty for relocations to be independent of the base locations that are involved in the relocations. This implementation reduces the number of r-variables for the Amsterdam case from 311,052 to 3,864. The computation time of the base case of the model is reduced from 403 seconds to 68 seconds.

To model the variables f_i, we need some big-M constraints. This type of constraints can result in a large integrality gap, especially when M is required to have high values. In our model, M can be bounded by the maximum number of ambulances that can be located at one base location. In principle, this can be as big as the maximum number of available ambulances, but because it is clear that it is not optimal to locate all ambulances at one location, we can do with a smaller M. In our experiments, the number of ambulances at one location in one time period never exceeds five. To be on the safe side, we bound that number by seven. As a consequence, we can use $7 \times 12 = 84$ as the value for M. However, we can change the formulation in order to get an even tighter value for M. Therefore, we will replace constraint (3.3) by constraint (3.9).

$$x_{it} \leq M f_i \qquad i \in I, t \in T \tag{3.9}$$

In this formulation, it is sufficient to set M to 7. So, in the LP relaxation, we will see no smaller fractions for f_i than 1/7. If multiple ambulances are located at each opened base location, we get even better values for f_i. Replacing constraint (3.3) by (3.9) results in better LP relaxations when the number of ambulances at a base location varies over the day.

We might even consider to introduce a base location-dependent value for M, say M_i. This can be the result of practical limitations of specific base locations. For example, if some base locations does not have enough capacity for more than a certain number of ambulances, we can add this constraint and decrease the value of M. Alternatively, we can use the base location-dependent M_i for purely computational reasons. For base locations in rural areas, for example, it is known in advance that no optimal solution will locate more than two or three

ambulances there. We can improve the LP relaxation by using a smaller value M_i for these locations.

3.5 Conclusions and future research

In this chapter, we introduced a time-dependent probabilistic location model for EMS vehicles. We consider time-dependent travel times, demand and ambulance availability. In comparison with the original time-dependent MEXCLP, we penalize the number of base locations and the number of relocations. By doing so, we make sure that the relationship between different time periods is taken into account. Computational results show that this is of great importance in order to get practical solutions. By applying the model to the region of Amsterdam-Waterland, the Netherlands, we show that the current location of the ambulances is not optimal with respect to our model and we can obtain a higher coverage with even fewer base locations. From a fairness perspective, it might be important to ensure that each demand point is covered at least once in each time period. Incorporating this constraint would slightly reduce the expected coverage and increase the number of locations and relocations. The decision maker should make the trade-off between a fair solution and a solution with a higher total expected coverage. Finally, we applied the proposed data aggregation error correction method from Section 2.2 to this model and obtain that we might need one more base location in the city center.

For future research, we would like to take a closer look at what happens at the border of the considered region. We see that base locations tend to be located far away from the border, which result in higher response times to calls at the border of the region. This problem does not only occur in our model, but is typical for coverage-based location models. Furthermore, we would like to extend our case study with three aspects. First, it might be interesting to incorporate variations between days of the week or times of the year. These variations can be accounted for by the same model, where a time period has a different interpretation. Second, it would be interesting to examine the effect of a location-dependent penalty β_j to take into account the preferences of the EMS providers. For example, we could have a smaller penalty for using the current base locations than we have for new base locations. Finally, we would like to improve the travel time model so that we could work on a finer grid and give better estimations for the time-dependency in the travel times.

4

Linear formulation for MEXCLP with fractional coverage

4.1 Introduction

In the previous chapter, we introduced an adaption of MEXCLP for the case that the input varies over the day. Another assumption of MEXCLP, and most other ambulance location models, is that the obtained coverage by assigning a call to a particular base is either zero or one. In some applications, it would be suitable to relax this assumption and allow for fractional coverage. We discuss two well-studied examples: coverage probabilities and survival probabilities.

In most models, it is assumed that the response time from a particular base to a particular demand point is fixed. Typically, this is equal to the average travel time plus some fixed pre-trip delay, where the pre-trip delay is the time elapsed before the ambulance starts driving. In practice, however, these response times vary, due to traffic jams and weather conditions. At least two ways of handling this uncertainty are used in literature. Koç and Bostancioğlu (2011) introduce a required reliability α, and say that a base can cover a demand point only if the response time is within the time threshold with probability at least α. With this approach, they have again a zero or one coverage and then they apply DSM. Another, more common approach is to compute the coverage probability directly. This approach is applied to several of the basic ambulance location models. Both Daskin (1987) and Karasakal and Karasakal (2004) include the coverage probability in MCLP. Marianov and ReVelle (1996) adapt MALP so that it can handle coverage probabilities. A version for MEXCLP with coverage probabilities was introduced by Goldberg et al. (1990) and Goldberg and Paz (1991), for which they used heuristics to find approximate solutions. Based on these two papers, Ingolfsson et al. (2008) developed a nonlinear formulation of the variant of MEXCLP with coverage probabilities. For small instances, typically with a fixed set of bases, the model can be solved to optimality. However, the computation time increases rapidly when the instance size increases. To determine the coverage probabilities, they assume that both the pre-trip delay and the travel times are nondeterministic.

A second example of the usage of fractional coverage is the concept of survival probabilities. Erkut et al. (2008) argue that even though most EMS providers are assessed on coverage-related criteria, it is worthwhile to consider performance measures related to health outcomes. They introduce a version of MEXCLP that maximizes the survival probability of a patient rather than the expected coverage. By replacing the coverage probability by the survival probabilities, we get that this model is equivalent to Ingolfsson et al. (2008). Again, the presented model is a Nonlinear Integer Programming problem. Knight et al. (2012) extend the model to allow for different survival probabilities for different types of patients. Later, Mayorga et al. (2013) used survival probabilities to develop dispatch policies.

In this chapter, we present an Integer Linear Programming formulation for the version of MEXCLP with fractional coverage. Compared to the nonlinear formulation by Ingolfsson et al. (2008) and Erkut et al. (2008), this reduces the computation time and allows for solving larger instances. Erkut et al. (2009) solve the nonlinear model for 180 demand points and 16 bases, but note that finding optimal solution for instances with more bases would be problematic. To apply the model to determine optimal base locations rather than an optimal distribution of the ambulances given a fixed set of bases, we need to solve instances with more base locations. We will show that our linear model can be solved for larger instances. Note that the two models are equivalent and thus provide the same solutions.

In Section 4.2, we will first show why a straightforward formulation will result in a nonlinear model. Second, we will show how the problem can be reformulated as an Integer Linear Programming problem. Finally, we will prove the equivalence of the two models. Section 4.3 provides an empirical comparison of the computation times of the linear and nonlinear formulation. In Section 4.4, we apply the model to the region of Amsterdam to show the behavior of the model. Conclusions and possible extensions of this research are discussed in Section 4.5. Note that in the description of the model, we use the stochastic response times as our underlying application, but for the application to survival probabilities, the model is equivalent.

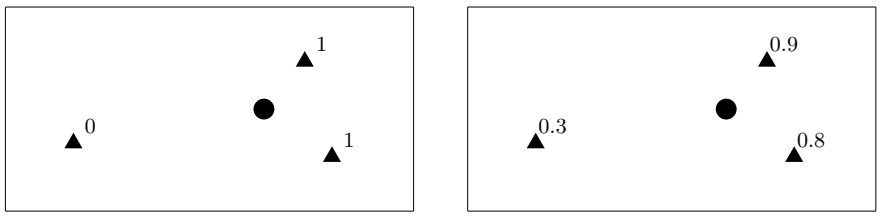
4.2 Model description

Even though ambulance location models typically use all-or-nothing coverage, multiple authors have noted that it might be more realistic to use fractional coverage probabilities. In this section, we introduce an Integer Linear Programming formulation for an adapted version of MEXCLP where a coverage w_{ij} is obtained when an available ambulance at base $i \in I$ responds to a call at demand point $j \in J$. Different from the classical MEXCLP this probability does not have to be 0-1 valued. This w_{ij} can, for example, be interpreted as the probability of reaching demand point j from base i within the time threshold, or as the probability that a patient at location j survives when served by an ambulance from

base i. A nonlinear formulation for this problem was previously introduced by Ingolfsson et al. (2008).

As in MEXCLP, we assume that each ambulance is unavailable for a fraction q of the time, called the busy fraction. Furthermore, we assume that the availability of an ambulance is independent of the availability of the other ambulances. The probability that at least one ambulance out of k is available is then $E_k = 1 - q^k$. The expected coverage of a demand point covered by k ambulances is thus E_k. In our model, we will use this concept, introduced by Daskin (1983), to determine the expected coverage.

We now give two examples to show the effect of fractional coverage probabilities on the expected coverage. In both examples, we consider a region with one demand point j and three base locations, each with one ambulance located. Each base i has a probability w_{ij} of covering the demand point. In the first example, these are 0.9, 0.8 and 0.3, respectively. In the second example, we have 0.7, 0.4, and 0.3. In the deterministic case, the coverage is 1 if $w_{ij} \geq 0.5$ and 0 otherwise. Figure 4.1 depicts Example 1.



(a) *Deterministic case* **(b)** *Stochastic case*

Fig. 4.1: *Representation of the first example of the difference between the deterministic and the stochastic case. A circle represents a demand point, a triangle represents a base station. The numbers next to the triangles show the probability that the demand point can be reached within the time limit from the particular base.*

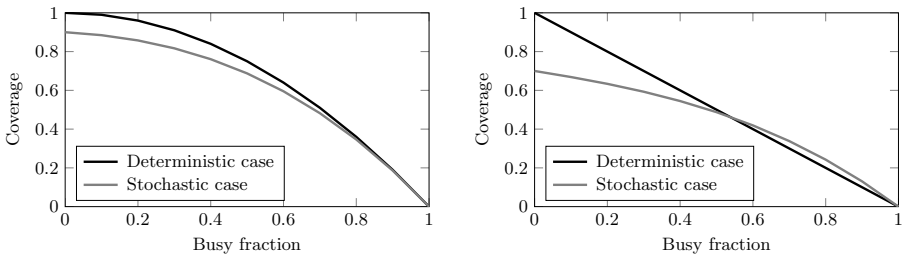
Figure 4.2 shows the expected coverage for the case with deterministic and stochastic coverage probabilities for both examples, varying the busy fraction. Note that since both models are based on MEXCLP, we do have stochastic ambulance availability in both cases. To show how the expected coverage is computed, we show the computation for Example 1 with a busy fraction of 0.4. In the deterministic case, we get

$$\begin{aligned}\text{Cov} = &\mathbb{P}(\text{1st available}) \times \mathbb{P}(\text{1st in time}) + \\ &\mathbb{P}(\text{1st unavailable}) \times \mathbb{P}(\text{2nd available}) \times \mathbb{P}(\text{2nd in time}) + \\ &\mathbb{P}(\text{1st and 2nd unavailable}) \times \mathbb{P}(\text{3th available}) \times \mathbb{P}(\text{3th in time}) = \\ &0.6 \times 1 + 0.4 \times 0.6 \times 1 + 0.4^2 \times 0.6 \times 0 = 0.84.\end{aligned}$$

For the stochastic case, we get

$$\begin{aligned}\text{Cov} = &\mathbb{P}(\text{1st available}) \times \mathbb{P}(\text{1st in time}) + \\ &\mathbb{P}(\text{1st unavailable}) \times \mathbb{P}(\text{2nd available}) \times \mathbb{P}(\text{2nd in time}) + \\ &\mathbb{P}(\text{1st and 2nd unavailable}) \times \mathbb{P}(\text{3th available}) \times \mathbb{P}(\text{3th in time}) = \\ &0.6 \times 0.9 + 0.4 \times 0.6 \times 0.8 + 0.4^2 \times 0.6 \times 0.3 \approx 0.76.\end{aligned}$$

Figure 4.2 shows that using 0-1 coverage results in different estimations of the expected coverage than using the fractional coverage. Typically, the deterministic case overestimates the expected coverage, even though Example 2 shows that for high busy fractions, it can also be the other way around. These examples stress the importance of including fractional coverage probabilities.



(a) *Example 1* **(b)** *Example 2*

Fig. 4.2: *Difference between the deterministic and the stochastic case for different parameters. In the first example, we have three bases with coverage probabilities 0.9, 0.8, and 0.3, respectively. In the second example, we have three bases with coverage probabilities 0.7, 0.4, and 0.3. In the deterministic case, we have a coverage probability of 1 if the stochastic coverage probability is at least 0.5, and 0 otherwise. The figures show the expected coverage for different busy fractions.*

4.2.1 Model formulation

As in the previous chapters, we are given a set of demand points J and a set of possible base locations I. For each demand point j, we have a given demand d_j. This d_j should be a measure for the number of calls within demand point j. See, for example, Channouf et al. (2007), and Setzler et al. (2009) for EMS call

volume forecasting methods. Each base location has a capacity p_i, which is the maximum number of ambulances that may be located at that station. In total, we are allowed to use at most $f_{\max}$ base locations. The total number of available ambulances is p. The busy fraction of an ambulance is denoted by q. For each combination of a base location i and a demand point j, we have a probability w_{ij} that an ambulance departing from base i will reach demand point j within the time threshold. For a fixed demand point j, given w_{ij}, we can order the base locations from the closest to the furthest for this demand point. Let a_{ij} denote the index of the base location that is in position i in this ordering for demand point j. Similarly, let $ranking(i, j)$ be the ranking of base i in the ordering of demand point j. So, by definition we have $ranking(a_{ij}, j) = i$.

The most straightforward way of modeling our problem is to introduce a decision variable x_i denoting the number of ambulances located at location i. The expected coverage of demand point j in terms of x_i is then

$$c_j(x) = \sum_{i \in I} q^{\sum_{k<ranking(i,j)} x_{a_{kj}}} (1 - q^{x_{a_{ij}}}) w_{a_{ij}j}. \tag{4.1}$$

Clearly, this formulation is not linear in the decision variables. When solving larger instances, this can result in longer computation times. To avoid this, we present a different formulation for which the objective is linear in the decision variables.

To formulate a linear model, we introduce a new binary decision variable z_{ijk} indicating whether the k-th preferred, with respect to w_{ij}, ambulance for demand point j is located at base location i. If, for example, base location 1 is the closest one for demand point 2 and we have three ambulances located at that base location, we get $z_{121} = z_{122} = z_{123} = 1$. Additionally, we introduce a binary variable f_i, which has value 1 if and only if at least one ambulance is located at base location i. This variable is needed to limit the number of base locations that is used. Based on the values of z, we can compute the coverage of demand point j, $c_j(z)$.

$$c_j(z) = \sum_{k=1}^{p} (1 - q) q^{k-1} \sum_{i \in I} z_{ijk} w_{ij} \tag{4.2}$$

The value $c_j(z)$ is calculated by conditioning on the number of unavailable ambulances. The probability that the k-th ambulance is the first available one equals $(1 - q)q^{k-1}$. If the k-th preferred ambulance is located at location i, we obtain an expected coverage of w_{ij}.

Now, we are able to formulate our model as follows:

$$\max \quad C^{MILP} = \sum_{j\in J} d_j c_j(z) = \sum_{j\in J} d_j \sum_{k=1}^{p} (1-q)q^{k-1} \sum_{i\in I} z_{ijk} w_{ij},$$

$$\text{s.t.} \quad \sum_{k=1}^{p} z_{ijk} \le x_i \qquad \forall i \in I, j \in J\,, \tag{4.3}$$

$$\sum_{i\in I} z_{ijk} = 1 \qquad \forall j \in J, k \le p\,, \tag{4.4}$$

$$\sum_{i\in I} f_i \le f_{\max}\,, \tag{4.5}$$

$$x_i \le p_i f_i \qquad \forall i \in I\,, \tag{4.6}$$

$$\sum_{i\in I} x_i \le p\,, \tag{4.7}$$

$$f_i, z_{ijk} \in \{0,1\} \qquad \forall i \in I, j \in J, k \le p\,, \tag{4.8}$$

$$x_i \in \mathbb{N} \qquad \forall i \in I\,. \tag{4.9}$$

The objective is to maximize the expected coverage over all demand points. This is defined as the sum of the coverages that can be provided to an individual node by the whole system, $c_j(z)$, multiplied by the total demand generated at this node, d_j. Constraints (4.3) state that no more than x_i ambulances may be assigned to base i. This makes sure that the z_{ijk}'s have the right value. Constraints (4.4) ensure that the k-th preferred ambulance of demand point j is located at no more than one base location. In order to design a realistic system, we add a limitation on the maximum number of base locations by Constraint (4.5). This constraint is not included in Ingolfsson et al. (2008). They assume that the set of bases is fixed. Constraints (4.6) guarantee that the number of vehicles located at each station does not exceed its capacity. Finally, Constraint (4.7) states that no more than p ambulances are used.

In Appendix A, the complete description of the nonlinear version is given. Now, we prove that the two formulations are equivalent.

Theorem 4.1. $C^{MINLP} = C^{MILP}$

Proof. Given a solution (f', x') for MINLP, we construct the following solution (f, x, z) for MILP. Let $f := f'$, $x := x'$, and

$$z_{ijk} := \begin{cases} 1 & \text{for } k = \{\sum_{l<ranking(i,j)} x_{a_{lj}} + 1, ..., \sum_{l\le ranking(i,j)} x_{a_{lj}}\}, \\ & i \in I, j \in J \\ 0 & \text{otherwise.} \end{cases}$$

We show that this solution is feasible and that it has the same objective value as (f', x').

Since Constraints (4.5), (4.6) and (4.7) are equivalent to (4.18), (4.19) and (4.20), the constructed solution satisfies these constraints. In the construction of z, we set exactly

$$\sum_{l<ranking(i,j)} x_{a_{lj}} - \sum_{l\leq ranking(i,j)} x_{a_{lj}} = x_i$$

variables to 1, given i and j. Hence, Constraint (4.3) is satisfied. Finally, Constraint (4.4) is satisfied because the order a_{ij} fully determines at which base the k-th ambulance for demand point j is located.

Now, we define $c_{ij}(x) := q^{\delta}(1-q^{\lambda})w_{a_{ij}j}$, where $\delta = \sum_{l<ranking(i,j)} x_{a_{lj}}$ and $\lambda = x_{a_{ij}}$. Then, we get

$$\begin{aligned} c_{ij}(x) = q^{\delta}(1-q^{\lambda})w_{a_{ij}j} &= q^{\delta}\sum_{k=1}^{\lambda} q^{k-1}(1-q)w_{a_{ij}j} = \sum_{k=1}^{\lambda} q^{\delta+k-1}(1-q)w_{a_{ij}j} \\ &= \sum_{k=\delta+1}^{\lambda+\delta} q^{k-1}(1-q)w_{a_{ij}j} = \sum_{k=1}^{p} q^{k-1}(1-q)w_{a_{ij}j}z_{a_{ij}jk}. \end{aligned}$$

For the first equality, we use the geometric sequence. This gives that $1-q^{\lambda} = \sum_{k=1}^{\lambda} q^{k-1}(1-q)$. The last equality is true by construction of z. All terms that are added to the sum have $z_{ijk} = 0$.

By summing over all demand points j and base stations i, we get

$$\begin{aligned} \sum_{j\in J} d_j c_j(x) = \sum_{j\in J} d_j \sum_{i\in I} c_{ij}(x) &= \sum_{j\in J} d_j \sum_{i\in I}\sum_{k=1}^{p} q^{k-1}(1-q)w_{a_{ij}j}z_{a_{ij}jk} \\ &= \sum_{j\in J} d_j \sum_{i\in I}\sum_{k=1}^{p} q^{k-1}(1-q)w_{ij}z_{ijk} = \sum_{j\in J} d_j c_j(z). \end{aligned}$$

Since for every solution (f', x') for MINLP we can find a solution (f, x, z) for MILP with the same objective value, we have that $C^{MINLP} \leq C^{MILP}$.

To show that $C^{MINLP} \geq C^{MILP}$, we prove that given an optimal solution (f^*, x^*, z^*) for MILP, we have that (f^*, x^*) is a feasible solution for MINLP with the same objective value. Clearly, (f^*, x^*) is feasible. Without loss of generality, we can assume that the optimal solution for MILP satisfies the relation between x and z as before. It is optimal to respect the order a_{ij} in filling z, because $q^{k-1}(1-q)$ is concave. As a result, by the same arguments as before, we have that $C^{MINLP} \geq C^{MILP}$. Hence, $C^{MINLP} = C^{MILP}$. □

4.3 Comparison of computation time

To analyze the difference in computation time between our formulation (MILP) and the nonlinear formulation (MINLP), used, for example, by Ingolfsson et al.

(2008), we solve both models for a set of 20 generated test instances. We implement both models in AIMMS 3.14 (AIMMS BV, 2013) and use the default solvers, which are CPLEX 12.6 (ILOG, 2013) and BARON 12 (The Optimization Firm, 2013), respectively.

We create two sets of ten instances differing in the number of demand points and potential bases. We randomly generate demand points and base locations in the unit square. The average travel time between two points is the Euclidean distance multiplied by 1,500 seconds. We assume that both the pre-trip delay and the travel times are stochastic. The travel times are assumed to be normally distributed with a coefficient of variation of 0.25, which corresponds to a standard deviation of 25% of the mean. The pre-trip delay is incorporated in the same way as in the case study, a lognormal distribution with mean 5.2967 and standard deviation 0.4574. The time threshold is set to 900 seconds, or 15 minutes. For each demand point, we generate a weight d_j uniformly between 10 and 30. Hence, the maximum difference in importance between two demand points is a factor of 3. We use a busy fraction of 42%, which corresponds to the observed busy fraction in the region of Amsterdam.

For the first set of instances, which we call 10-180, we take 10 potential bases and 180 demand points. We set $f_{\max}$ to 10, so that there is no limitation on the number of opened bases. Basically, the model only decides how to distribute the available ambulances over the given bases. The dimensions of these instances correspond to the test cases in Ingolfsson et al. (2008).

The second set of instances, called 100-100, is used to test how the formulations perform when not all bases can be opened. To that end, we take instances with 100 base locations and 100 demand points, while allowing to open only 10 bases, i.e., $f_{\max} = 10$. In both sets, we set the number of ambulances to 18 and the maximum number of ambulances per base to 5.

Recall from Theorem 4.1 that the two formulations, MILP and MINLP, are equivalent and thus have the same optimal objective value.

4.3.1 Results

As described above, we have a total of 20 test instances, which can be divided in two groups. To all instances, we apply both models with different time limits. For the easier set of instances, 10-180, we set the time limit to 5 minutes, 30 minutes, and 24 hours. Since the linear formulation already provided all optimal solutions within 5 minutes, we did not run it with longer time limits. For the second set of instances, we used time limits of 30 minutes and 24 hours. The results for all instances are summarized in Table 4.1. For the results per instance, see Table 4.5 and 4.6 in Appendix 4.B.

The table shows a significant difference in performance between the two formulations for both set of instances. For the first set, MILP was able to solve all instances to optimality in less than 3 seconds, while for MINLP, optimality could not be guaranteed for any of the instances within 30 minutes. However, in seven cases the best solution found after 30 minutes was the optimal one. For one

Table 4.1: *Results for comparison of computation time. The first three columns describe the instances. Column 4 shows the number of instances that are solved to optimality. In column 5, it is stated in how many of those cases the optimality could be verified by the solver. Column 6 shows the number of instances for which no solution is returned by the solver after the time limit has exceeded. The final two columns show the average gap and the average computation time. For the computation of the average gap, only the instances that returned a non-optimal, but feasible solution are included.*

Size	Formulation	Time limit	# Opt	# Verified	# No sol.	Avg. gap	Avg. time
10-180	**MINLP**	**300 sec.**	2	0	1	0.65%	300 sec.
10-180	**MINLP**	**1,800 sec.**	7	0	0	0.05%	1,800 sec.
10-180	**MINLP**	**86,400 sec.**	10	8	0	–	39,303 sec.
10-180	**MILP**	**300 sec.**	10	10	0	–	1.96 sec.
100-100	**MINLP**	**1,800 sec.**	0	0	3	30.95%	1,800 sec.
100-100	**MINLP**	**86,400 sec.**	0	0	2	28.76%	86,400 sec.
100-100	**MILP**	**1,800 sec.**	0	0	0	0.07%	1,800 sec.
100-100	**MILP**	**86,400 sec.**	9	9	0	0.02%	31,308 sec.

instance, the solver did not provide a feasible solution within 5 minutes. Even within a time limit of 24 hours, optimality could not be guaranteed for two of the instances, although the optimal solution was found.

For the second instance set, no optimal solutions were found within 30 minutes. For MILP, the average gap was only 0.07%, while for MINLP this was 30.95%. Note that the gap is defined as the value of the best found solution divided by the best found upper bound. The upper bound found in the linear formulation is also used to compute the gap for the nonlinear case. When the time limit is set to 24 hours, MILP was able to solve nine instances to optimality. For the remaining instance, the gap was only 0.02%. The nonlinear model gave no optimal solutions and the average gap was 28.76%. In two cases, no solution was returned by the solver.

4.4 Case study

In this section, we apply the presented model to the region of Amsterdam, the Netherlands. We define the set J of demand points as the set of all postal codes in this region. This gives a total of 161 points, which corresponds to an average size of 3.9 km^2 per postal code area. We assume that each demand point is also available as a potential base location, i.e., $I = J$. However, in the solution, we are allowed to use at most nine of these bases, which corresponds to the number of bases currently in use in this region. The number of available ambulances is set to 18.

4.4.1 Data Analysis

In order to apply the model, we have to determine the busy fraction q, the demand d_j for each $j \in J$, and the coverage probabilities w_{ij}. For the busy fraction, we take the average busy fraction during the day over the last four years, which is equal to 0.42. The expected demand for demand point j, d_j, is estimated by the average number of calls that have arisen from that demand point over the years 2008-2012. This data is provided by the ambulance provider in the region of Amsterdam. To compute w_{ij}, we have to estimate the pre-trip delay distribution and the travel time distribution. Below, we describe these estimations. Based on these two distributions, we compute w_{ij} by taking the convolution of the two distributions. Let R_{ij} be a random variable representing the response time for a call from demand point j served by base i. Furthermore, let $t_{ij}(x)$ be the travel time distribution for trips between i and j. Finally, let $h(x)$ be the distribution function of the pre-trip delay. Note that the pre-trip delay is independent of i and j. As in Ingolfsson et al. (2008), we compute w_{ij} as follows:

$$w_{ij} = \mathbb{P}(R_{ij} \leq r) = \int_0^r h(x)t_{ij}(r-x)dx, \tag{4.10}$$

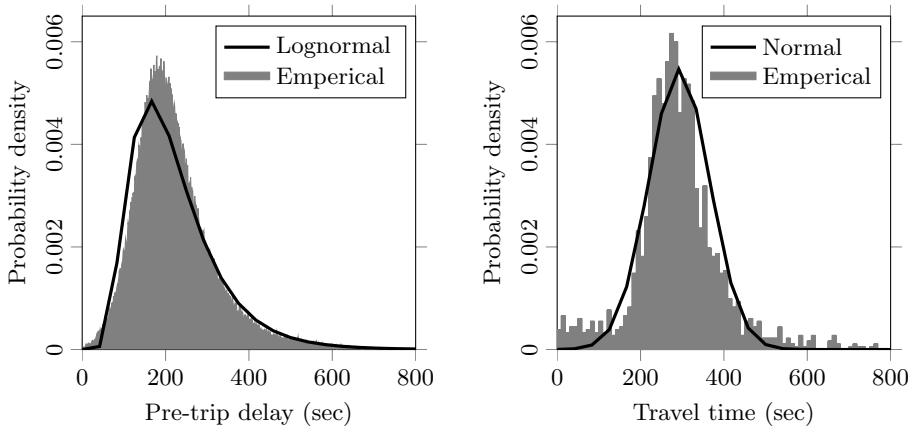
where r is the response time target, which is 15 minutes in the Netherlands.

Pre-trip delays

The pre-trip delay is the time spent before the ambulance leaves the station. Based on 446,290 calls of high urgency, we find that a lognormal distribution gives a reasonable fit. This is the same result as obtained by Ingolfsson et al. (2008). For our data, the pre-trip delay is best approximated by a lognormal distribution with mean 5.2967 and standard deviation 0.4574. This corresponds to an average pre-trip delay of 222 seconds and a standard deviation of 107 seconds. The average is similar to the numbers reported in the annual EMS-reports in the Netherlands (Ambulancezorg Nederland, 2012). Figure 4.3a shows the empirical and fitted lognormal distribution of the pre-trip delay.

Travel times

The calculation of the travel time distribution is more complicated, since we have a different travel time for each pair of base location and demand point. In order to estimate these distributions, we analyzed 10 pairs with more than 750 samples in our database. Based on these, we conclude that the travel times are well-approximated with a normal distribution with a coefficient of variation of 0.25. One problem with a normal distribution is that it could generate negative values, which cannot occur in practice. However, since the coefficient of variation of 0.25, this happens only for values smaller than $\mu - 4\sigma$, which happens in only 0.3% of the cases. For the mean travel time between two points, we use the travel



(a) *Empirical and fitted distribution of pre-trip delay.*

(b) *Empirical and fitted distribution of travel times for one particular pair of postal codes.*

Fig. 4.3: *Empirical and fitted distribution of pre-trip delay and travel time.*

time model introduced by Kommer and Zwakhals (2008), which is specifically developed for ambulances in the Netherlands. This model estimates the driving speed on each road type and uses that to compute the travel times. The estimated driving speeds that we use are based on rush hours, workdays from 06:30 till 09:30 and from 15:00 till 19:00. Figure 4.3b shows for one pair of points that this can give a reasonable fit, although we can also see that the fit is worse than for the pre-trip delay. To account for the potential misfit, we will evaluate the sensitivity of the model with respect to the travel time distribution in Section 4.4.2.

4.4.2 Results

To evaluate the performance of the model, we conduct multiple experiments. First, we compare the optimal solution according to the model with the current set of base locations. This shows the potential performance increase. Second, we compare the solution with the cases where we do not take into account randomness in either the pre-trip delay or the travel times. This provides insight into the importance of modeling the uncertainty. Furthermore, by plotting the selected bases, we get insight into the structure of the provided solutions. Third, we investigate the impact of the restriction on the number of bases. In our base case, we limit the number of bases by the current number, which is nine. These results may provide a trade-off between the number of bases to open and the coverage that can be obtained. Finally, we evaluate the sensitivity to the chosen travel time distribution. We show how the solutions change, when a different distribution or a different coefficient of variation is used. All computations were

executed on a 2.9 GHz Intel(R) Core(TM) i7-3520M laptop with 8 GB of RAM. We used CPLEX 12.5 as our solver (ILOG, 2009).

Current versus optimal

In the current situation, there are nine base locations in the region of Amsterdam. To investigate whether these nine bases are located in an optimal way, we compare the optimal solution according to the model with the best solution given that the bases are fixed. Note that the number of ambulances remains fixed at 18. In the optimal solution, we are only allowed to open the same number of bases as in the current situation. We will refer to this case as the base case. Comparing the results, we see that without changing the base stations, we can obtain a coverage of 0.9203. By changing the bases, however, we can obtain an increase of 2.92 percentage points, to 0.9495. This corresponds to reaching 37% of the previously uncovered calls. Note that the actual coverage in 2012 was 0.933 for this region (Ambulancezorg Nederland, 2012). This coverage is higher than expected by the model, which can be explained by some of the simplifying assumptions of the model. The model ignores that dynamic ambulance management is used to improve the real-time performance. Additionally, in practice there is a link with non-urgent patient transportations that are partly executed with the same ambulances. We did not incorporate this link in the case study.

Impact of randomness

To investigate the impact of the randomness in both the pre-trip delay and the travel times, we create four test instances. The first assumes stochastic pre-trip delays and travel times and is the same as the base case defined earlier. Then, we define two instances in which the randomness of one of the two response time components is ignored. The last instance has both deterministic delays and travel times and corresponds to the classical MEXCLP. In Table 4.2, we show the coverage according to the w_{ij}'s used in the optimization and the coverage according to the w_{ij}'s in the base case. Clearly, the coverage in the base case is the highest, since this gives the optimal solution with respect to the w_{ij}'s in the base case.

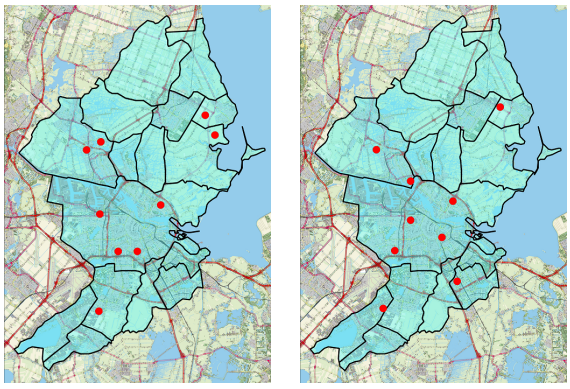
We observe that in order to get the optimal solution, it is important to take the randomness in both the components into account. In particular, when both random components are ignored, we obtain far from optimal solutions with respect to the input of the stochastic case. Note that this case corresponds to the classical MEXCLP. Furthermore, we see that the coverage is consistently overestimated when the randomness is not incorporated. In the fully deterministic case, this overestimation is almost 5.5 percentage point.

Since the nonlinear formulation was not able to solve the model for many potential bases, it is interesting to see the impact of the fractional coverage probabilities on the selected set of bases. Figure 4.4 shows the selected bases for the fully deterministic case, which corresponds to the classical MEXCLP, and the

Table 4.2: *Importance of taking into account randomness in pre-trip delay and travel times. Estimated coverage is the coverage with respect to the w_{ij}'s used in the optimization. Real coverage is the coverage with respect to w_{ij} where both pre-trip delay and travel times are stochastic.*

Pre-trip delay	Travel times	Estimated coverage	Real coverage
Deterministic	**Deterministic**	0.9852	0.9304
Deterministic	**Stochastic**	0.9656	0.9487
Stochastic	**Deterministic**	0.9623	0.9490
Stochastic	**Stochastic**	—	0.9495

fully stochastic case. We see that in the stochastic case, bases are evenly spread out over the city center, so as to provide good coverage to these regions with high call volume. This is not necessary in MEXCLP, because a coverage within 15 minutes suffices. In three cases, two bases are located close to each other. This is necessary to avoid some demand points to be completely uncovered. This is a direct consequence of the strict 0-1 coverage.



(a) *MEXCLP locations* **(b)** *Stochastic MEXCLP locations*

Fig. 4.4: *Maps of base locations selected by deterministic MEXCLP (4.4a) and stochastic MEXCLP (4.4b).*

Limited number of bases

In this part, we investigate the impact of the number of bases on the expected coverage. We run the model for different values of $f_{\max}$ and compare the coverage. Note that the total number of ambulances is fixed.

Table 4.3: *Coverage for different number of bases.*

# Bases	Coverage	# Bases	Coverage
1	0.6680	**10**	0.9508
6	0.9381	**11**	0.9517
7	0.9434	**12**	0.9524
8	0.9481	**13**	0.9533
9	0.9495	**18**	0.9553

Table 4.3 shows that reducing the number of bases from nine to seven does not have a large impact on the expected coverage. Similarly, adding one or two bases hardly increases the coverage. When no limit is set on the number of bases, which corresponds to a different base for each ambulance, the coverage increases by only 0.58 percentage point compared to the base case. For this coverage to be reached, we need twice as many bases. Ambulance providers should make the trade-off between the cost of an additional base and the increase in coverage.

Sensitivity to travel time distribution

Since the distribution of the travel times might influence the optimal ambulance locations, we compare the outcome of the model for different travel time distributions. In the base case, we used a normal distribution with a coefficient of variation of 0.25, corresponding to a standard deviation of 0.25 times the mean. We vary this value from 0.1 to 0.5. Additionally, we evaluate the results for a lognormal travel time distribution with the same coefficients of variation. In all cases, the pre-trip delay remains the same as in the base case. The expected coverage and the number of changes in the ambulance distribution are given in Table 4.4. The fifth column gives the number of bases that are located differently, while the sixth column lists the number of ambulances that are assigned to a different base. Note that if a base is located differently, the ambulances assigned to that base are counted as assigned to a different base.

We see that the coverage decreases when the variability in the travel times increases. Due to the relatively high coverage percentage, the loss of coverage as a consequence of a more negative worst-case is higher than the benefit from a better best-case travel time realization. Furthermore, we can conclude that the optimal locations of the ambulances is rather robust against different travel time distributions. For example, changing from a normal distribution to a lognormal distribution with the same coefficient of variation does not change the optimal solution. The solution provided by the base case gives close to optimal coverage with respect the different distributions.

Sensitivity to busy fraction

The model, as presented, takes the busy fraction of an ambulance as an input. Typically, this busy fraction is hard to estimate and might depend on the se-

Table 4.4: *Solution for different travel time distributions. Column two gives the coverage of the optimal solution with respect to a particular distribution. Column three gives the coverage of the solution provided by the base case with respect to the different distributions. Column four evaluates the different solutions with respect to the base case. Column five and six give the number of bases and ambulances that are located differently.*

Var	Coverage	Coverage of solution for base case with respect to different distributions	Coverage of different solutions with respect to base case	Changed bases	Changed assignment
Normal distribution					
0.1	0.9600	0.9590	0.9494	1	1
0.2	0.9539	0.9535	0.9494	1	1
0.3	0.9448	0.9448	0.9495	0	0
0.4	0.9342	0.9342	0.9495	0	0
0.5	0.9231	0.9226	0.9462	1	3
Lognormal distribution					
0.1	0.9600	0.9590	0.9490	2	3
0.2	0.9536	0.9533	0.9494	1	1
0.25	0.9493	0.9493	0.9495	0	0
0.3	0.9449	0.9449	0.9495	0	0
0.4	0.9354	0.9354	0.9485	1	2
0.5	0.9269	0.9263	0.9475	2	4

lected bases and ambulance distribution. To overcome this, one could use an iterative method where, based on the outcomes of the model, the busy fraction is estimated. With the updated value, the model is solved until some convergence criterion is met (Ingolfsson et al., 2008). To gain insight into the sensitivity of the model to the busy fraction, we run the model for different values of q. Furthermore, the solution obtained with $q = 0.42$ is evaluated for different busy fractions. The results are shown in Figure 4.5. For values of q between 0.3 and 0.5 the solution does not change. Only when very high or very low busy fractions are used in the optimization, we obtain suboptimal solution with respect to a busy fraction of 0.42. Similarly, if we use 0.42 in the optimization, the obtained solution is also optimal for some cases with different busy fractions. Even if the busy fraction is significantly different, the coverage loss as a result of the incorrect estimation is limited. This shows that the model is rather robust against busy fraction estimation errors.

4.5 Conclusions and future work

In this chapter, we presented an ambulance location model based on the maximum expected coverage model, introduced by Daskin (1983). In contrast to the classical MEXCLP, we allow the coverage provided by base i to demand point j to be fractional. This allows to include stochastic travel times and survival

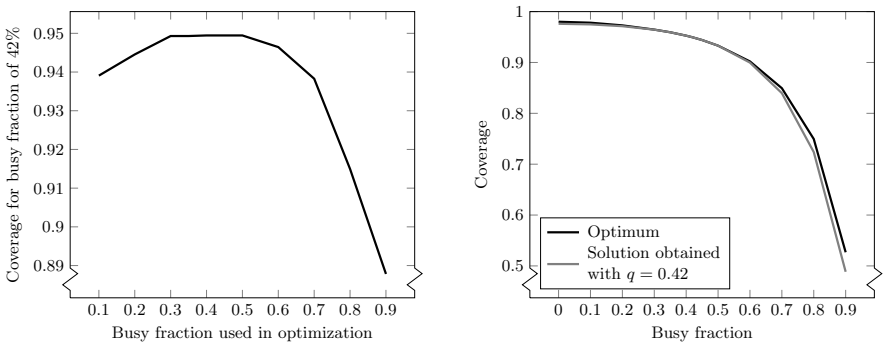


Fig. 4.5: *Impact of busy fraction on obtained solution. On the left, the coverage, with respect to busy fraction of 0.42, of solution obtained with different busy fractions is given. On the right, the solution obtained with busy fraction of 0.42 is compared with optimum for different busy fractions.*

probabilities. These applications were already studied by Ingolfsson et al. (2008) and Erkut et al. (2008). They used a nonlinear formulation to model fractional coverage probabilities. We presented a linear formulation for this problem, which is proved to be equivalent to their formulation. We compared the computation time of our linear formulation with the nonlinear formulation and observed that significant improvement can be obtained. Instances of the nonlinear model that take more than 30 minutes to solve can now be solved within a few seconds. We further applied the model to the region of Amsterdam and observe that higher coverage can be obtained according to our model. Furthermore, we saw that including the randomness in pre-trip delay and travel times has an important impact on the obtained solution. Since travel time distributions are hard to estimate, we evaluated the impact of different travel time distributions. The results show that small changes in the distribution do not have a high impact on the optimal solution. Nevertheless, it would be useful for future research to investigate potential improvements in the estimation of the travel time distributions.

An interesting extension of this research would be to incorporate busy fractions that depend on the base station. This would allow to incorporate workload variations within a region. In the current formulation, this would result in a nonlinear model. The results of Section 4.3 show that tractability benefits significantly from a linear formulation. Hence, investigating potential linear formulations might be worthwhile for future research.

Finally, we highlight that most proposed extensions of the Maximum Expected Coverage Location Model can be included in this model as well. For example, although the model is formulated to maximize the coverage given fixed resources, it can also be used to determine the required number of ambulances to reach a fixed coverage level by applying the model for different values of p.

4.A Model formulation

In this appendix, we state both the MILP and MINLP formulation. Both models use the variables x_i and y_i. Here, x_i is the number of ambulances located at base i and y_i takes value 1 if base i is opened and 0 otherwise. Additionally, MILP uses the variables z_{ijk} indicating whether the k-th preferred, with respect to w_{ij}, ambulance for demand point j is located at base location i. The two formulations are then as follows.

MILP

$$\max \quad C^{MILP} = \sum_{j\in J} d_j c_j(z) = \sum_{j\in J} d_j \sum_{k=1}^{p} (1-q)q^{k-1} \sum_{i\in I} z_{ijk} w_{ij}$$

$$\text{s.t.} \quad \sum_{k=1}^{p} z_{ijk} \leq x_i \qquad \forall i \in I, j \in J \tag{4.11}$$

$$\sum_{i\in I} z_{ijk} = 1 \qquad \forall j \in J, k \leq p \tag{4.12}$$

$$\sum_{i\in I} f_i \leq f_{\max} \tag{4.13}$$

$$x_i \leq p_i f_i \qquad \forall i \in I \tag{4.14}$$

$$\sum_{i\in I} x_i \leq p \tag{4.15}$$

$$f_i, z_{ijk} \in \{0,1\} \qquad \forall i \in I, j \in J, k \leq p \tag{4.16}$$

$$x_i \in \mathbb{N} \qquad \forall i \in I \tag{4.17}$$

MINLP

$$\begin{aligned} \max \quad C^{MINLP} &= \sum_{j\in J} d_j c_j(x) \\ &= \sum_{j\in J} d_j \sum_{i\in I} q^{\sum_{k<ranking(i,j)} x_{a_{kj}}} (1 - q^{x_{a_{ij}}}) w_{a_{ij}j} \end{aligned}$$

$$\text{s.t.} \quad \sum_{i\in I} f_i \leq f_{\max} \tag{4.18}$$

$$x_i \leq p_i y_i \qquad \forall i \in I \tag{4.19}$$

$$\sum_{i\in I} x_i \leq p \tag{4.20}$$

$$f_i \in \{0,1\} \qquad \forall i \in I \tag{4.21}$$

$$x_i \in \mathbb{N} \qquad \forall i \in I \tag{4.22}$$

4.B Results computational comparison

Table 4.5: *Results of computational comparison.*

Instance	Formulation	Time limit (sec.)	Outcome	Gap	Comp. time (sec.)
10-180-01	**MINLP**	**300**	Feasible	0.04%	300
10-180-02	**MINLP**	**300**	Optimal	–	300
10-180-03	**MINLP**	**300**	Feasible	0.63%	300
10-180-04	**MINLP**	**300**	Optimal	–	300
10-180-05	**MINLP**	**300**	Feasible	2.83%	300
10-180-06	**MINLP**	**300**	Feasible	0.29%	300
10-180-07	**MINLP**	**300**	Feasible	0.09%	300
10-180-08	**MINLP**	**300**	Feasible	0.43%	300
10-180-09	**MINLP**	**300**	No solution	–	300
10-180-10	**MINLP**	**300**	Feasible	0.20%	300
10-180-01	**MINLP**	**1,800**	Optimal	–	1,800
10-180-02	**MINLP**	**1,800**	Optimal	–	1,800
10-180-03	**MINLP**	**1,800**	Optimal	–	1,800
10-180-04	**MINLP**	**1,800**	Optimal	–	1,800
10-180-05	**MINLP**	**1,800**	Optimal	–	1,800
10-180-06	**MINLP**	**1,800**	Optimal	–	1,800
10-180-07	**MINLP**	**1,800**	Feasible	0.03%	1,800
10-180-08	**MINLP**	**1,800**	Feasible	0.07%	1,800
10-180-09	**MINLP**	**1,800**	Feasible	0.07%	1,800
10-180-10	**MINLP**	**1,800**	Optimal	–	1,800
10-180-01	**MINLP**	**86,400**	Verified	–	29,262
10-180-02	**MINLP**	**86,400**	Verified	–	29,759
10-180-03	**MINLP**	**86,400**	Verified	–	23,448
10-180-04	**MINLP**	**86,400**	Optimal	–	86,400
10-180-05	**MINLP**	**86,400**	Verified	–	21,927
10-180-06	**MINLP**	**86,400**	Optimal	–	86,400
10-180-07	**MINLP**	**86,400**	Verified	–	32,383
10-180-08	**MINLP**	**86,400**	Verified	–	26,848
10-180-09	**MINLP**	**86,400**	Verified	–	24,260
10-180-10	**MINLP**	**86,400**	Verified	–	32,314
10-180-01	**MILP**	**300**	Verified	–	1.65
10-180-02	**MILP**	**300**	Verified	–	1.69
10-180-03	**MILP**	**300**	Verified	–	2.12
10-180-04	**MILP**	**300**	Verified	–	2.09
10-180-05	**MILP**	**300**	Verified	–	1.83
10-180-06	**MILP**	**300**	Verified	–	1.79
10-180-07	**MILP**	**300**	Verified	–	2.31
10-180-08	**MILP**	**300**	Verified	–	2.31
10-180-09	**MILP**	**300**	Verified	–	2.18
10-180-10	**MILP**	**300**	Verified	–	1.59

Table 4.6: *Results of computational comparison.*

Instance	Formulation	Time limit (sec.)	Outcome	Gap	Comp. time (sec.)
100-100-01	**MINLP**	**1,800**	Feasible	16.92%	1,800
100-100-02	**MINLP**	**1,800**	Feasible	28.56%	1,800
100-100-03	**MINLP**	**1,800**	Feasible	14.05%	1,800
100-100-04	**MINLP**	**1,800**	Feasible	14.44%	1,800
100-100-05	**MINLP**	**1,800**	No solution	–	1,800
100-100-06	**MINLP**	**1,800**	Feasible	41.53%	1,800
100-100-07	**MINLP**	**1,800**	No solution	–	1,800
100-100-08	**MINLP**	**1,800**	Feasible	66.97%	1,800
100-100-09	**MINLP**	**1,800**	No solution	–	1,800
100-100-10	**MINLP**	**1,800**	Feasible	34.18%	1,800
100-100-01	**MINLP**	**86,400**	Feasible	16.92%	86,400
100-100-02	**MINLP**	**86,400**	Feasible	28.56%	86,400
100-100-03	**MINLP**	**86,400**	Feasible	14.05%	86,400
100-100-04	**MINLP**	**86,400**	Feasible	14.44%	86,400
100-100-05	**MINLP**	**86,400**	No solution	–	86,400
100-100-06	**MINLP**	**86,400**	Feasible	41.53%	86,400
100-100-07	**MINLP**	**86,400**	No solution	–	86,400
100-100-08	**MINLP**	**86,400**	Feasible	66.97%	86,400
100-100-09	**MINLP**	**86,400**	Feasible	13.46%	86,400
100-100-10	**MINLP**	**86,400**	Feasible	34.18%	86,400
100-100-01	**MILP**	**1,800**	Feasible	0.13%	1,800
100-100-02	**MILP**	**1,800**	Feasible	0.08%	1,800
100-100-03	**MILP**	**1,800**	Feasible	0.02%	1,800
100-100-04	**MILP**	**1,800**	Feasible	0.05%	1,800
100-100-05	**MILP**	**1,800**	Feasible	0.12%	1,800
100-100-06	**MILP**	**1,800**	Feasible	0.01%	1,800
100-100-07	**MILP**	**1,800**	Feasible	0.03%	1,800
100-100-08	**MILP**	**1,800**	Feasible	0.13%	1,800
100-100-09	**MILP**	**1,800**	Feasible	0.09%	1,800
100-100-10	**MILP**	**1,800**	Feasible	0.002%	1,800
100-100-01	**MILP**	**86,400**	Optimal	–	65,961
100-100-02	**MILP**	**86,400**	Optimal	–	66,116
100-100-03	**MILP**	**86,400**	Optimal	–	3,568
100-100-04	**MILP**	**86,400**	Optimal	–	9,863
100-100-05	**MILP**	**86,400**	Feasible	0.02%	86,400
100-100-06	**MILP**	**86,400**	Optimal	–	5,455
100-100-07	**MILP**	**86,400**	Optimal	–	12,737
100-100-08	**MILP**	**86,400**	Optimal	–	24,695
100-100-09	**MILP**	**86,400**	Optimal	–	35,927
100-100-10	**MILP**	**86,400**	Optimal	–	2,360

5

Location model for firefighters

5.1 Introduction

In the previous chapters, we focused on the location of bases for EMS providers. Additionally, we considered the distribution of the ambulances over the selected bases. In this chapter, we develop a model to determine good locations for fire stations. Clearly, this problem is strongly related to the ambulance location problem. However, some characteristics of firefighter systems differ from EMS systems, calling for different models. Some features of the models are no longer necessary, whereas other features need to be added to make the model suitable for firefighters.

The first main difference is the variety of vehicles that is used to handle the calls. For emergency responses of ambulances, almost all calls are served by an Advanced Life Support (ALS) ambulance. There are other vehicles, such as MICU and PICU, but their usage is very limited and is typically not included in location models. Fire departments typically use a wide variety of vehicles to serve different calls and this cannot be neglected in determining base locations.

Second, ambulance services in the Netherlands have only two different response time targets, which only depend on the medical condition of the patient. This target is independent of the location of the patient. For firefighters, many different targets are set by law, depending on, for example, the type of building, the function of the building, and the type of call. As a consequence, the response time target is highly location-dependent and fluctuates significantly throughout the region. Additionally, the targets are typically stricter than the 15 minutes for the most urgent ambulance calls. For example, in case of a fire in the old city center of Amsterdam, a fire apparatus vehicle must be present within 6 minutes.

Third, the call volumes for fire departments are lower than for EMS providers. On average, the fire department in the region of Amsterdam-Amstelland handles approximately 30 calls per day, whereas the EMS provider in the region Amsterdam-Waterland serves approximately 170 A1 calls per day. This, in combination with the very stringent response time targets, leads to a very low utilization of the vehicles.

Finally, for fire departments, it is common to staff some vehicles with a voluntary, or on-call, crew. In this case, the crew is not located at the base, but is called in case of emergency. Clearly, this leads to an increase in chute time, which is part of the pre-trip delay. On the other hand, staffing vehicles with voluntary brigades significantly reduces the cost. Note that typically more personnel is called than required, so that the increase in chute time is limited.

All these differences call for a different model for determining the optimal configuration of fire stations. In this chapter, we describe how these differences can be incorporated and what the impact is on the location of fire stations in the region of Amsterdam-Amstelland.

This region contains the city of Amsterdam that was the first city in the Netherlands to start a professional fire service in 1874. With 144 crew members and nine fire stations covering 30 km^2, it ensured the fire protection safety for approximately 285,000 inhabitants. Today, the regionally organized fire department Amsterdam-Amstelland is responsible for over 1,000,000 inhabitants in an area of 354 km^2. For this, 34 vehicles, distributed over 19 bases, and 1,150 crew members are available. Obviously, over time the questions of how many fire stations were needed and where to locate them had to be answered numerous times as new needs and means for fire protection safety emerged. These decisions were made with the information and technology available at the time, and it is interesting to see how the resulting configuration compares to an ideal configuration and how small changes can improve the coverage.

As we have seen in Section 1.3, the location of ambulance bases is widely studied, and some of the models are also suitable for the placement of fire stations. Only some scholars focused on developing models specifically designated to firefighters. Much of the research in the seventies was done within the Rand Fire Project in the city of New York. The research within this project is documented in Rand Fire Project (1979). Both Swersey (1994) and Green and Kolesar (2004) give overviews of the successes of this project. Hogg (1968) determined the set of locations of fire stations that minimizes the sum of losses from fire and the cost of providing the service. Building further on this and the detailed study of Toregas et al. (1971), Plane and Hendrick (1977) used the response time as a standard for coverage and applied a location set covering problem (LSCP) optimization model. The amount of needed resources (i.e., firemen or fire apparatus) was included as a decision rule in Swersey (1982), expanding the work on square root laws for fire engine response distances by Kolesar and Blum (1973). Later, Batta and Mannur (1990) also optimized the number of resources sent to an emergency, but only considered one type of fire apparatus. Andersson and Särdqvist (2007) developed a model which allows for combinations of multiple resource types as well as multiple event types. An approach for fire protection locational decisions based on the maximal coverage location problem (MCLP) was described by Schilling et al. (1980). Variants of the MCLP were then used by Murray and Tong (2009) and Chevalier et al. (2012), further improving the model by including a risk-modeling approach to estimate demand. For the case where many fire stations become empty as a result of large fires, Kolesar and

Walker (1974) present an algorithm to dynamically relocate fire trucks to empty fire stations.

In the remainder of this chapter, we introduce a model to determine good locations for fire stations, good distribution of vehicles over the bases, and appropriate crew configurations. The model allows to incorporate firefighter-specific characteristics such as multiple response time targets, different vehicle types, and different crew types. In Section 5.2, the model is described in detail. In Section 5.3, we describe the data used in a case study for fire department Amsterdam-Amstelland. The data is collected in close cooperation with the fire department to ensure a good representation of their system. In Section 5.4, we discuss the results for both a greenfield scenario as well as scenarios with only limited base changes. We further evaluate the impact of different crew configurations. Finally, Section 5.5 restates the main conclusions and discusses some applications of the model that we have in mind.

5.2 Model description

In this section, we introduce a model to determine good locations for fire stations. The model differs from typical ambulance location models by including some firefighter-specific features. First, we include multiple types of vehicles in the optimization, which are used to serve different types of calls. To that end, we define a set V of vehicle types. The demand for a particular vehicle $v \in V$ from demand point $j \in J$ is given by d_{jv}, where, as in previous chapters, J denotes the set of demand points. For each vehicle type $v \in V$, we have a fixed number of p_v available units. Since opening base locations is costly, we limit the total number of opened bases by a parameter f_{max}. Note that if $f_{max} \geq \sum_{v \in V} p_v$, this constraint will not impose any limitation.

We further consider multiple types of crews to distinguish between professional and voluntary brigades. In general notation, we define a set L of different crew types, where in the computations we typically use $|L| = 2$. As professional crews are more costly than voluntary crews, we have a limit on the number of professional crews. There are multiple ways of implementing this limitation. For example, one could introduce a crew budget and costs for crews of different types. Then, the model could decide on the best way to spend the budget. However, since we currently fix the total number of vehicles and crews are only useful if assigned to a vehicle, the total number of crews is fixed. We fix the number of each type by some value c_l. The model decides to what vehicle types the different crews are assigned.

The objective of the model is to maximize the provided coverage, and thereto we define a set of base locations that can cover a certain demand point. This set depends on the pre-trip delay, the travel time, and the response time target. We assume a fixed time for triage and dispatch and a chute time that depends on the type of crew. Let τ_l denote the resulting pre-trip delay in case a vehicle with crew type $l \in L$ is sent. The travel time between base location $i \in I$ and demand

point $j \in J$ is assumed to be fixed and independent of the type of vehicle, and is denoted by t_{ij}. Here, I denotes the set of potential locations for a fire station. The response time target is set for each type of vehicle and each demand point separately. We denote this by r_{iv}. Combining all this, we get that the set of bases that can cover demand point $j \in J$ by vehicle type $v \in V$ if staffed by crew of type $l \in L$ is equal to $I_{jlv} = \{i \in I | t_{ij} + \tau_l \leq r_{iv}\}$.

In the formulation of the model, we use three types of variables. First, we have a variable x_{ilv} denoting the number of type v vehicles that are located at potential base location i and staffed with crew of type l. Since we do not take into account backup coverage, these variables are 0-1 valued. Second, we introduce a binary variable y_{jv} indicating whether demand point j is covered by a vehicle of type v. Finally, binary variable f_i indicates whether base location i is used by at least one vehicle type, and thus must be opened.

We formulate the model as an Integer Linear Programming (ILP) problem. This type of problem can in general not be solved in polynomial time. However, commercial solvers like CPLEX (ILOG, 2009) can provide an optimal solution in reasonable time for realistic instance sizes. The model is defined as follows:

$$\max \quad \sum_{j \in J} \sum_{v \in V} d_{jv} y_{jv},$$

$$\sum_{i \in I_{jlv}} x_{ilv} \geq y_{jv} \qquad \forall j \in J, l \in L, v \in V, \tag{5.1}$$

$$\sum_{i \in I} \sum_{l \in L} x_{ilv} \leq p_v \qquad \forall v \in V, \tag{5.2}$$

$$\sum_{i \in I} \sum_{v \in V} x_{ilv} \leq c_l \qquad \forall l \in L, \tag{5.3}$$

$$x_{ilv} \leq f_i \qquad \forall i \in I, l \in L, v \in V, \tag{5.4}$$

$$\sum_{i \in I} f_i \leq f_{max}, \tag{5.5}$$

$$x_{ilv}, y_{jv}, f_i \in \{0, 1\} \qquad \forall i \in I, j \in J, l \in L, v \in V. \tag{5.6}$$

The objective of the model is to maximize the number of calls that is covered by a vehicle of the appropriate type. Constraints 5.1 state that demand point j is only covered by vehicle type v if there is at least one vehicle of type v at a base close enough to demand point j. Whether a base location is close enough depends on the crew that is assigned to a vehicle at that base. This is ensured by the set I_{jlv}. Constraints (5.2) and (5.3) ensure that the limitations on the number of vehicles and crew of each type are respected. Constraints (5.4) and (5.5) limit the total number of bases that is used.

Note that we did not include backup coverage in the models. This is in contrast with many other studies in the literature, especially those dealing with ambulance modeling. There, the backup coverage is often included by taking

into account the ambulance unavailability (see, for example, Daskin (1983), Hogan and ReVelle (1986), and Gendreau et al. (1997)). As already mentioned by Swersey (1994), the call volumes for fire departments are significantly lower, and simultaneous calls are thus not as common. However, they do occur, but not at an alarming rate. Data analysis we conducted showed that fire department Amsterdam-Amstelland handles on average 32 calls per day with 19 base locations and 34 vehicles. Furthermore, in the computations, we will not include all vehicles that are currently in use, so that the other vehicles can be used to provide backup coverage.

5.3 Data description

To apply the model to the region of Amsterdam-Amstelland, we have to determine the appropriate data. In close cooperation with the fire department Amsterdam-Amstelland, we defined a set of 2,643 demand points, which corresponds to the sections currently in use. This includes some areas of neighboring regions at the borders. We assume that in most demand points that are part of the region, a base can be located. We exclude demand points in neighboring regions and demand points that only contain highways. This gives 2,223 potential base locations. Travel times between potential base locations and demand points are provided by the fire department and are based on estimated travel times on the road network between each location.

In our analysis, we include the four most common types of vehicles used at Dutch fire departments: fire apparatus (FA), aerial apparatus (AA), rescue apparatus (RA) and marine rescue units (MR). The number of available vehicles of each type is 22, 9, 3 and 2, respectively. In the current configuration, we have 19 bases. Since the objective does not benefit from backup coverage, three of the 22 FA vehicles do not contribute to the coverage in the current situation. In the computations, we will not include these three vehicles. The number of FA vehicles in the computation is thus 19 and the other three can be used as backup for the rare occurrence of simultaneous calls. For each demand point and for each vehicle type, we have to define the weight. We take the absolute number of calls per vehicle that occurred in 2011 to model the recurring risk. In this period, 39,516 calls were registered. The number of calls per vehicle type is 29,016, 9,182, 615 and 703, respectively. Since for many pairs of demand point and vehicle type we have zero calls, we add one call to each pair. By doing so, we avoid that we completely ignore sections that did not have any calls yet, enabling for sporadic risks (Chevalier et al., 2012). Another benefit of this approach is that we increase the relative importance of vehicle types with low call volume. Since call volumes for RA and MR are so low, we risk too high focus on FA and AA in the optimization. By adding calls artificially, we limit this uneven focus. We end up with a total weight of 50,088.

Since different vehicle types are used for different kind of calls, different response time targets might be used for each vehicle type. Dutch law states a

response time target for both FA and AA. These targets depend on the type and function of a building, and vary between 6 and 10 minutes for FA and between 6 and 15 minutes for AA. Based on these requirements, we set a response time target for each demand point for FA and AA. For RA and MR, no requirements are set by law. For these vehicle types, we set the response time requirement to 15 minutes for all demand points. A summary of the data is given in Table 5.1.

Table 5.1: *Summary of the data for the different vehicle types.*

Vehicle type	**FA**	**AA**	**RA**	**MR**
# Vehicles	19	9	3	2
# Calls	29,016	9,182	615	703
Total weight	31,659	11,825	3,258	3,346
Minimum target	6	6	15	15
Average target	7.98	14.68	15	15
Maximum target	10	15	15	15

For each crew type, we have to determine the number of available crews and the pre-trip delay. In the current execution, there are 27 professional crews and 6 voluntary crews. All six voluntary brigades operate on an FA vehicle. For the professional staff where the crew is present at the fire station, the average pre-trip delay is 3 minutes. For voluntary staff, where the crew is on-call, this is 6 minutes. Note that often more volunteer personnel is called than necessary. In that case, the crew can depart whenever sufficient crew members have arrived.

5.4 Computational results

We apply the model to the data of the region Amsterdam-Amstelland as described before. We consider three different cases regarding the set of bases. First, we analyze the performance of the current set of 19 bases. Here, we distinguish cases based on the freedom we allow in relocating vehicles and reassigning crew. We compare the current assignment of vehicles and crew with the optimal distribution over the current bases. We further evaluate the impact of replacing some voluntary crews by professionals and vice versa. Second, we consider the case where we allow for a small number of base changes. We allow for a free distribution of the vehicles and crew over the selected bases. Finally, we study a greenfield scenario where the current bases are not incorporated. This provides us with insight into what the optimal distribution of the fire stations would be.

5.4.1 Fixed set of bases

When we fix the current set of bases, we can obtain the current coverage and evaluate the potential improvement by redistributing vehicles and crew over the

current set of bases. Note that for FA vehicles, we cannot improve coverage by redistributing vehicles, as we already have an FA at every base. Table 5.2 shows the coverage for the different vehicle types for four different cases regarding the freedom in redistributing vehicles and reassigning voluntary and professional crews. The total number of each vehicle type and each crew type, as well as the set of bases, remains the same.

Table 5.2: *Coverage with fixed set of bases for different vehicle types in different cases regarding freedom in redistribution of vehicles and reassignment of crew.*

Crew assignment	Vehicle distr.	Coverage				
		FA	**AA**	**RA**	**MR**	**Total**
Fixed	**Fixed**	0.8016	0.9291	0.8938	0.7989	0.8375
Free	**Fixed**	0.8208	0.9274	0.8938	0.7989	0.8492
Fixed	**Free**	0.8016	0.9634	0.9497	0.8619	0.8535
Free	**Free**	0.8208	0.9578	0.9497	0.8619	0.8643

We see that the current configuration provides a coverage of 0.8375. If we allow for reassignment of the crew, we can increase coverage by 1.17 percentage point. In the corresponding solution, four AA vehicles are staffed with voluntary crew instead of professionals. Conversely, four more FA vehicles are staffed with a professional crew. This decreases coverage for AA by 0.0017, but increases FA coverage by 0.0192. With the current crew configuration, but a free distribution of the vehicles over the bases, we can increase coverage by 0.0160, compared to the current configuration. Finally, if we combine crew reassignment and vehicle redistribution, a coverage of 0.8643 can be obtained. Recall that this does not require any base changes, nor does it require any additional resources.

Next, we evaluate the impact of replacing voluntary crews by professionals and vice versa. In this experiment, the set of bases is fixed and vehicles can be freely distributed over the current bases. As the total number of vehicles remains fixed, also the total number of shifts remains the same.

Table 5.3 shows that the impact of replacing some professional crews by volunteers is limited. Increasing the number of voluntary crews from six to nine leads to a coverage reduction of only 0.0028. An all professional crew yields a coverage increase of 0.18 percentage point.

5.4.2 Limited base changes

Now, we evaluate the impact of a small number of base changes. We allow for a free distribution of the vehicles and reassignment of the crew. We consider the case of adding or replacing up to four of the 19 bases. Additionally, we consider the case with unlimited additions or changes. The last two cases correspond to the greenfield scenario that we discuss in Section 5.4.3, with $f_{max} = 33$ and

Table 5.3: *Coverage with fixed set of bases for different configuration of the crew. In the current configurations 27 professional crews and 6 voluntary crews are available.*

Crew changes	Coverage				
	FA	**AA**	**RA**	**MR**	**Total**
Add 3 voluntary crews	0.8198	0.9548	0.9282	0.8619	0.8615
Add 2 voluntary crews	0.8198	0.9548	0.9497	0.8619	0.8629
Add 1 voluntary crew	0.8198	0.9578	0.9497	0.8619	0.8636
Current	0.8208	0.9578	0.9497	0.8619	0.8643
Add 1 professional crew	0.8208	0.9597	0.9497	0.8619	0.8647
Add 2 professional crews	0.8208	0.9621	0.9497	0.8619	0.8653
Add 3 professional crews	0.8213	0.9621	0.9497	0.8619	0.8656
All professional crews	0.8216	0.9634	0.9497	0.8619	0.8661

$f_{max} = 19$. Since the number of available vehicles is equal to 33, $f_{max} = 33$ implies that there is no limitation on the number of fire stations.

Table 5.4: *Coverage given a limited number of changes to the current set of bases, where vehicles and crews can be freely distributed over the selected bases.*

Base changes	Coverage				
	FA	**AA**	**RA**	**MR**	**Total**
Add 1	0.8506	0.9849	0.9497	0.8619	0.8895
Add 2	0.8879	0.9818	0.9497	0.8619	0.9123
Add 3	0.9087	0.9822	0.9626	0.8751	0.9273
Add 4	0.9185	0.9822	0.9626	0.8751	0.9335
Change 1	0.8506	0.9849	0.9497	0.8619	0.8895
Change 2	0.8879	0.9818	0.9497	0.8619	0.9123
Change 3	0.9087	0.9818	0.9626	0.8751	0.9272
Change 4	0.9185	0.9818	0.9626	0.8751	0.9334
No changes	0.8208	0.9578	0.9497	0.8619	0.8643
Unlimited changes	0.9753	0.9914	0.9770	0.9032	0.9744
Unlimited additions	0.9753	0.9941	0.9834	0.9097	0.9759

Table 5.4 shows that the difference between adding a base and replacing a base is very limited. Up to two changes, there is no difference in coverage between adding and replacing a fire station. For unlimited changes, the difference in coverage is only 0.0015. This suggests that the current number of fire stations is appropriate, but also that some of the current stations are not adequately positioned. We further see that only a few changes already lead to a significant coverage improvement. For example, three changes yield a 6.29 percentage point

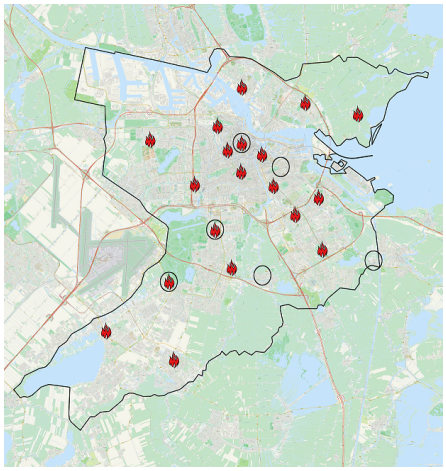


Fig. 5.1: *Distribution of fire stations in case three stations are moved. The empty circles represent the closed fire station and filled circles represent the new fire stations.*

coverage increase. Figure 5.1 shows the movements of bases in the case where the location of three fire stations may be changed.

5.4.3 Greenfield

As a final experiment, we consider a greenfield scenario where the current fire stations are not incorporated. Again, we allow for a free distribution of the vehicles and assignment of the crew over the selected bases. Both the number of vehicles and crews of each type remains fixed. We consider values for f_{max}, which is the maximum number of opened fire stations, from nine up to 19. Furthermore, we consider the case with an unlimited number of bases.

Table 5.5 shows enormous potential to improve the coverage by an optimal distribution of fire stations throughout the region. Even with only nine optimally located bases, a higher coverage can be obtained than with the current 19 bases. One particularly surprising observation is that the coverage of RA and MA vehicles can be significantly improved. Since only three or two vehicles are available, one would expect that the the current 19 bases would contain a good tuple, or pair, of bases for those types. However, as a coverage increase of 3.37 and 4.78 percentage point can be obtained, this does not seem to be the case.

5.5 Conclusions

In this chapter, we presented a model specifically designated to determine good locations for fire stations. Other than the typical ambulance location models,

Table 5.5: *Coverage in greenfield scenario for different maximum number of bases.*

# Bases	Coverage				
	FA	**AA**	**RA**	**MR**	**Total**
9	0.8499	0.9678	0.9638	0.8906	0.8879
10	0.8797	0.9816	0.9742	0.8748	0.9095
11	0.9035	0.9822	0.9742	0.8748	0.9247
12	0.9156	0.9937	0.9748	0.8861	0.9359
13	0.9302	0.9940	0.9748	0.8840	0.9451
14	0.9422	0.9939	0.9748	0.8840	0.9526
15	0.9549	0.9930	0.9718	0.8924	0.9608
16	0.9613	0.9933	0.9748	0.8978	0.9655
17	0.9687	0.9933	0.9748	0.8897	0.9696
18	0.9714	0.9913	0.9770	0.9032	0.9719
19	0.9753	0.9914	0.9770	0.9032	0.9744
Unlimited	0.9753	0.9941	0.9834	0.9097	0.9759
Current	0.8208	0.9578	0.9497	0.8619	0.8643

the model incorporates multiple vehicle types for different types of calls. For each call type, but also for each demand point, the model can have different response time targets. Furthermore, the model distinguishes multiple crew types to incorporate both professional and voluntary crews. As a consequence of the very low call volumes for fire stations, it is common to have some crews on-call rather than present at the fire station.

One concept that is wide-spread in ambulance location models that is not incorporated in the model is backup coverage. As a result of the low call volumes and the relatively high number of vehicles, simultaneous calls for the same vehicle are not common for firefighters. For that reason, the model presented in this chapter uses single coverage only. However, for FA vehicles we only included 19 vehicles, while currently there are 22 FA vehicles. The other three vehicles could be used to provide backup coverage.

The results of this study show that even without changing the bases, a significant coverage improvement can be obtained. By changing the distribution of the vehicles and reassigning the crew, the coverage can be increased by 2.68 percentage point. If up to three base changes are allowed, another 2.5%, 4.8% or 6.3% of coverage can be gained. For the first two changes, the removal of one of the current bases does not affect the coverage. This shows that some of the current bases are not adequately located. This observation made the fire department Amsterdam-Amstelland decide to close one of the current bases. This particular base is staffed with voluntary crew only, but a professionally staffed FA vehicle at another base reaches all demand points quicker. The results of the greenfield computation further show that there is an enormous potential for improvement by optimally locating the fire stations. With fewer stations, the same or even better coverage levels can be reached.

In one aspect the model differs from the current practice at fire departments in the Netherlands. Currently, coverage targets are set for each building rather than each area. Therefore, it can occur that one building has a 5 minute response time target while its neighbor has a 10 minute target. In the model, we define the response time target for each demand point, and thus for some buildings this will deviate from the real response time target. However, the Dutch government is about to change the coverage definition to adopt an area-based target. Thus, the presented model is already prepared for this new approach.

Besides the usage of the model on the strategic and tactical level, there is also a possible application of the model in real-time. Even though simultaneous calls are rare, there is some risk in case of a very large fire. In this case, five or more trucks are sent. As typically the five closest FA vehicles are sent, it might be beneficial to relocate some vehicles to restore coverage. For future research, we intend to use the model to compute relocation policies for different scenarios. Depending on the location of the fire, we compute which trucks are sent and which relocations are necessary to ensure the remaining coverage.

Part II

Patient Transportation

6

Incorporating emergency calls in scheduling patient transportations

6.1 Introduction

In Part I of this thesis, we focused on the location of bases in order to provide good coverage for emergency calls. Apart from these A1 and A2 calls, EMS providers are also responsible for the transportation of patients from and to hospitals. In the Netherlands, these are called B calls. In 2013, 28.7% of all calls were patient transportations (Ambulancezorg Nederland, 2013).

Many regions in the Netherlands have ambulances that are specifically designated to these patient transportations. These Basic Life Support (BLS) ambulances are less equipped than regular Advanced Life Support (ALS) ambulances. Furthermore, BLS ambulances are only staffed by two regular nurses, whereas each ALS ambulances is staffed by at least one paramedic. Consequently, BLS ambulances cannot be used for emergency calls. ALS ambulances, on the other hand, can be used for patient transportations. As the ALS vehicles are primarily used to provide coverage for emergency calls, assigning them to patient transportation requests will reduce emergency coverage and should be done with great care.

Despite the non-urgent classification of patient transportations, each call does have a requested execution time. Even though this requested time is not a hard constraint, for the hospital's and patient's convenience, it is important to respect this requested time as often as possible. A significant fraction of the requests are known in advance, so that scheduling of these calls is possible. For the other part of the requests that arises on the day of execution, the schedule should be adapted in real-time.

The scheduling of non-urgent transportation requests is related to the Dial-A-Ride Problem (DARP), which is a special case of the Vehicle Routing Problem with Pickup and Delivery. DARP consists of designing vehicle routes to fulfill pickup and delivery requests between origins and destinations. The scheduling of BLS ambulances is a special case of DARP, as the capacity of BLS ambulances is limited to one patient. For DARP, Cordeau and Laporte (2007) make a distinction between the static DARP and the dynamic DARP. In the static case,

all transportation requests are known in advance and the schedule can thus be made with all necessary input. In contrast, in the dynamic case, the transportation requests arrive throughout the day, and thus, the schedule must be updated every time a request arrives. For the considered situation of scheduling BLS ambulances, we have a combination of the two cases. Some of the transportation requests are known in advance, but most requests arrive throughout the day.

Chen and Xu (2006) make a distinction between two classes of methods for dealing with the dynamic aspect. The first class uses local approaches, which means that the routes are solely based on the currently known information without considering the future. The second class uses look-ahead approaches, which try to incorporate a forecast of the future. For our case, we use a local approach as it is hard to predict when and where future transportation requests will occur.

There are several papers that apply DARP in the context of patient transportation. Most of them consider either an efficiency-based objective function (such as transportation cost or travel distance) or an objective function based on patients' inconvenience (such as lateness or excess drive time). Ritzinger et al. (2014) consider the static DARP with travel time minimization as objective and constraints on patients' inconvenience. Multiple Dynamic Programming (DP) based algorithms are used to provide heuristic solutions. Carnes et al. (2013) introduce a set partitioning formulation for the scheduling of patient transportations by fixed wing aircrafts in Ontario, Canada.

Different from Ritzinger et al. (2014), Melachrinoudis et al. (2007) and Parragh et al. (2009) include patients' inconvenience in the objective, which results in a static multi-objective DARP. Melachrinoudis et al. (2007) solve the problem as an Integer Linear Programming (ILP) problem and compare this to a Tabu Search (TS) heuristic for solving larger instances. Parragh et al. (2009) use Variable Neighborhood Search (VNS) to obtain an initial set of solutions which is used to generate additional efficient solutions by a path relinking module. Efficient solutions are solutions that are Pareto optimal with respect to the trade-off between efficiency and patients' inconvenience.

As opposed to the static DARP, Beaudry et al. (2010) allow requests to arrive throughout the day. They focus on the efficient and timely transport of patients between several locations on a hospital campus. This means that only short distances are considered. To find solutions to this problem, they use an insertion approach followed by a TS heuristic. Ritzinger et al. (2012) also consider the dynamic DARP where the objective is to balance the total travel time and patients' inconvenience. A heuristic DP algorithm is used to find an initial solution for requests that are known in advance. Requests that arrive throughout the day are included by an insertion heuristic. For the special case where vehicle capacity is limited to one patient, Kergosien et al. (2011) introduce a TS heuristic to obtain solutions. In case the number of vehicles does not suffice, they have the possibility of subcontracting a private company.

These three dynamic models all use a local approach where no information about future requests is used. Schilde et al. (2011), on the other hand, explicitly use the stochastic information about future requests to find better solutions.

Many of the patients that are transported from home to a hospital require transportation back home the same day. Using this information in the optimization results in a significant improvement. This method can be considered a look-ahead approach.

The main difference between the described papers and the situation that we consider is that the transportation requests can only be fulfilled by BLS ambulances, whereas in our situation also ALS ambulances can be used if needed. As the ALS ambulances are primarily used for emergency calls, using one of these ambulances reduces the available capacity for emergency calls. Therefore, assigning a non-urgent transportation request to an ALS ambulance must be done with great care such that the effect on the coverage by the ALS ambulances is minimized.

As reviewed in Section 1.3, there exist several measures for the coverage of emergency calls. For example, Church and ReVelle (1974) aim at maximizing the weighted number of demand locations within a given travel time from a base location. Daskin (1983) uses the weighted expected coverage as a measure of coverage which takes into account the probability that at least one ambulance is available within a given time limit. The Maximum Availability Location Problem of ReVelle and Hogan (1989) views coverage as the weighted number of demand locations that can be reached within a given time limit by a predefined number of ambulances. The model developed in this chapter is set up such that these and other coverage measures can be used.

This chapter is structured as follows. In Section 6.2, we first introduce an ILP formulation to determine the optimal routes for the BLS ambulances in the offline case where all non-urgent transportation requests are known beforehand. This ILP formulation is used as the basis for the online scheduling approach introduced in Section 6.3. In Section 6.4, we present the results for both the offline and online scheduling approaches and perform extensive sensitivity analysis. Section 6.5 presents conclusions and gives recommendations for further research. In Chapter 7, an application of the model to evaluate shift schedules for the region of Utrecht is discussed.

6.2 Offline model

As stated in the introduction, we consider the situation where some transportation requests are known beforehand, but most requests arrive throughout the day. In this section, we introduce an ILP formulation that can be used to determine the routes of BLS ambulances when the information of all requests is available. This formulation cannot be used in practice as most requests arrive throughout the day, which means that not all information is available, but the solution to this ILP yields an upper bound on the performance that can be obtained in practice. We call the situation, where the information of all requests is available, the offline case and we call the case where the information arrives throughout the day the online case. The solution approach for the offline case,

which is introduced in this section, is used as a basis for the solution method for the online case, which is discussed in Section 6.3.

One of our contributions is to include the coverage for emergency calls by ALS ambulances in scheduling BLS ambulances. Since ALS ambulances are used to serve non-urgent patient transportation requests when the capacity of the BLS ambulances is not sufficient, inadequate planning of BLS ambulances decreases the coverage for emergency calls. Therefore, we present a model that determines routes for the BLS ambulances such that the remaining coverage for emergency calls is maximized. To determine the remaining coverage, we assign patient transportation requests that are not executed by a BLS ambulance to a base station where one or more ALS ambulances are stationed. The number of available ambulances at that station is then reduced for a given amount of time. By doing so, we reserve capacity for the execution of the non-urgent transportation requests. The coverage is calculated based on the remaining capacity at the ambulance bases.

The following sets are considered as input to the model:

C	set of non-urgent transportation requests,
I	set of base locations,
J	set of demand points for emergency calls,
S	set of BLS shifts,
T	set of time periods.

The sets I, J, and T are specifically used to determine the coverage given a schedule for the BLS ambulances. As in previous chapters, we derive a related set I_j which is the set of all bases that can cover demand point $j \in J$ within the given time threshold.

As a tool for building feasible routes for BLS ambulances, we define a network that consists of nodes and arcs between the nodes. The nodes correspond to tasks for an ambulance. An arc between node n and n' corresponds to the execution of task n directly before task n' by a BLS ambulance. We define three types of nodes that correspond to tasks for an ambulance: the start of a shift, the execution of a transportation request and the end of a shift. Since the transportation requests are non-urgent, there is flexibility in the execution time of these requests. For example, we can have requests that can be served at any time between an hour before and after the requested time. We model this by including multiple copies of a request in the model, where the different copies have different start times. For each copy, we have a node in our network. Let M_c denote the set of nodes corresponding to request $c \in C$. Note that in general, we can schedule the transportation requests at every time in the interval between the earliest and latest execution time, but the model requires a discrete number of options. If the available computation time increases, we could add more nodes to the sets M_c. We further define $M := \cup_{c \in C} M_c$, which is the set of all nodes that correspond to requests. Additionally, we have nodes for the start and end of a BLS shift. We denote the set of nodes corresponding to origins and destinations

of a shift by O and D, respectively. Adding all this together, we get the set of nodes $N := O \cup M \cup D$.

Note that in the formulation, we strictly distinguish between nodes and requests. If, for example, we say that ambulance $s \in S$ executes node $n \in M_c$, this means that the copy of request c that corresponds to node n is executed by this ambulance. We will use the index c for requests and n for nodes.

Based on the start time, the end time, the start location and the end location of a node $n \in N$, we can derive the sets B_n and A_n that contain all nodes that can be visited directly before or after n in a feasible tour, respectively. A node n' is in the set B_n if the difference between the end time of n' and the start time of n is sufficient to travel from the end location of n' to the start location of n. The set A_n is constructed similarly. For $n' \in O, n \in D$, we have that $n' \in B_n$ if and only if n and n' correspond to the same BLS shift. In that case, we also have that $n \in A_{n'}$. Note that also nodes in O and D have a time and a location. In this way, we ensure that tours start and end at the right location, and it implies that we do not allow for overtime. Figure 6.1 gives a graphical representation of a simplified network with only two requests and two BLS shifts.

To estimate the remaining coverage for a given solution, we can use any of the static ambulance location models from the literature, see Brotcorne et al. (2003) or Section 1.3 for an overview. Regardless of this choice, we need the number of ambulances at each base, the demand of the demand points, and the time that an ambulance from base i is occupied when assigned to request c, as input. These inputs are assumed to be known and are denoted by:

- d_{jt} demand at demand point $j \in J$ during time period $t \in T$,
- a_{it} number of available ALS ambulances at base location $i \in I$ during time period $t \in T$,
- b_{int} binary parameter that indicates whether node $n \in M$ would occupy an ambulance at base location $i \in I$ in time period $t \in T$ if it were to be assigned to an ALS ambulance at this base.

To compute b_{int}, we subtract the travel time from base location i to the start location of n from the start time of n and add the travel time from the end location back to i to the end time of n. If this time interval intersects time period t, we set $b_{int} = 1$.

We formulate the problem as an ILP problem. To that end, we define the following variables:

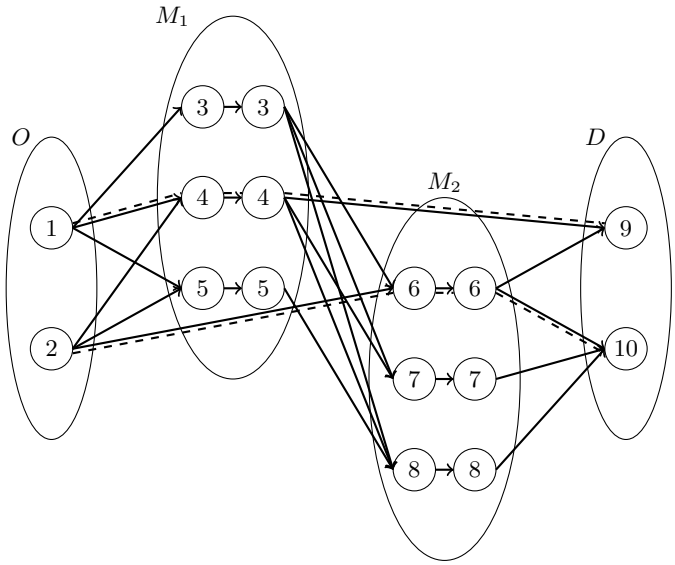


Fig. 6.1: ***Example of a network.*** *This figure represents the network of a problem with two BLS shifts and two requests. Each request can be executed at three different points in time. Nodes 1 and 2 represent the start of the two shifts. Nodes 9 and 10 represent the end of the shifts. Nodes 3, 4, and 5 and nodes 6, 7, and 8 correspond to request 1 and 2, respectively. Each node in the sets M_1 and M_2 is divided into two parts to highlight that "within the node" the ambulance travels from the origin of the patient to its destination. In the formulation this is represented by one node with a start location and an end location. Nodes are connected if they can be executed directly after each other. For example, if an ambulance executes request 1 at its latest possible time, node 5, then this ambulance can execute request 2 at its latest possible time, node 8, only. Another example would be that ambulance 2 can execute request 1, but in that case, request 1 cannot be executed at its earliest time. Arcs that are implied by transitivity are not shown, but are included in the network. For example, even though the arc from node 1 to 6 is not shown, it does exist in the network. This arc is implied by the arcs from 1 to 3 and from 3 to 6. The dashed lines in the network represent a feasible solution in which both requests are executed. Shift 1 executes request 1 at its second possible time, node 4, and after that returns to its base. Request 2 is executed by shift 2 at its earliest possible time and shift 2 returns to base after executing the request.*

u_{in} binary variable, which takes value 1 when node $n \in M$ is assigned to an ALS ambulance stationed at base $i \in I$, and 0 otherwise,

v_n binary variable, which takes value 1 when node $n \in M$ is assigned to a BLS ambulance, and 0 otherwise,

$w_{nn's}$ binary variable, which takes value 1 when BLS ambulance $s \in S$ executes node $n \in N$ directly before node $n' \in N$, and 0 otherwise,

x_{it} the number of ALS ambulances at base $i \in I$ that remain available for emergency calls during time period $t \in T$,

y_{jt} number of ALS ambulances that can cover demand point $j \in J$ during time period $t \in T$ within the given time threshold.

The objective of the model is to maximize the coverage that can be obtained by the remaining capacity of the ALS ambulances, i.e.,

$$\max \quad \sum_{j \in J} \sum_{t \in T} d_{jt} \, \text{coverage}(y_{jt}),$$

where coverage(y_{jt}) is a function that gives the coverage given a fixed number of ambulances that are stationed within the time threshold. This function depends on the coverage model that is used. For example, if the model MCLP (Church and ReVelle, 1974) is used, we get coverage(y_{jt}) is equal to one if and only if $y_{jt} \geq 1$.

As a straightforward constraint, we have that every transportation request should be executed, either by a BLS ambulance ($v_n = 1$) or by an ALS ambulance ($u_{in} = 1$) at one of the bases.

$$\sum_{n \in M_c} \left(\sum_{i \in I} u_{in} + v_n \right) = 1 \qquad \forall c \in C$$

Furthermore, we require that transportation requests that are assigned to BLS ambulances, i.e., $v_n = 1$, appear in one of the routes of a BLS ambulance.

$$\sum_{s \in S} \sum_{n' \in A_n} w_{nn's} = v_n \qquad \forall n \in M$$

The assignment of nodes to ambulances should satisfy some standard routing constraints (see, for example, Cordeau and Laporte (2007)). Recall that nodes in O correspond to the start of a BLS shift, nodes in M with executing a transportation request, and nodes in D with the end of a BLS shift. Note that even though the network does not restrict connections between the origin node of an ambulance and the destination node of another ambulance, these constraints enforce that each ambulance ends at its own base.

$$\sum_{n' \in B_n} w_{n'ns} - \sum_{n \in A_n} w_{nn's} = -1 \qquad \forall n \in O, s \in S$$
$$\sum_{n' \in B_n} w_{n'ns} - \sum_{n \in A_n} w_{nn's} = 0 \qquad \forall n \in M, s \in S$$
$$\sum_{n' \in B_n} w_{n'ns} - \sum_{n \in A_n} w_{nn's} = 1 \qquad \forall n \in D, s \in S$$

The relation between the variables u, x, and y is ensured by the following two constraints:

$$x_{it} + \sum_{n \in M} b_{int} u_{in} = a_{it} \qquad \forall i \in I, t \in T,$$
$$\sum_{i \in I_j} x_{it} \geq y_{jt} \qquad \forall j \in J, t \in T.$$

Finally, we have bounds on the variables.

$$u_{in}, v_n, w_{nn's} \in \{0, 1\} \qquad \forall n, n' \in N, i \in I, s \in S$$
$$x_{it}, y_{jt} \in \mathbb{N} \qquad \forall i \in I, j \in J, t \in T$$

For a complete overview of the model, see Appendix 6.A.

6.2.1 Coverage function

As stated before, we can choose numerous coverage functions to use in the model. We choose to use an adapted version of the well-known MEXCLP that was introduced by Daskin (1983). In MEXCLP, the expected coverage is determined by conditioning on the number of unavailable ambulances. The unavailability of the ambulances is denoted by the busy fraction of an ambulance, which is defined as the average fraction of time an ambulance is occupied. In the original MEXCLP, this busy fraction is the same for every part of the region. In practice, we typically see that the workload of ambulances varies over the region. In our model, we use a different busy fraction for each demand point. This busy fraction is given by the busy fraction of the nearest base location.

Another adaptation of the model compared to MEXCLP is that we do not re-optimize the distribution of the ambulances over the bases. We only consider the changes in capacity due to non-urgent transportation requests that are scheduled on an ALS ambulance.

Note that the demand, busy fractions, and number of ambulances at each base change over time. Consequently, we have different input values for the coverage model for each time period. To incorporate this coverage function, we introduce the following variables:

y_{jkt} binary variable that takes value 1 when demand point $j \in J$ is covered by at least k ambulances within the time threshold during time period $t \in T$.

Let q_{jt} denote the busy fraction of ambulances covering demand point $j \in J$ during time period $t \in T$. Then the function coverage(y_{jt}) is defined as

$$\text{coverage}(y_{jt}) = \sum_{k=1}^{\sum_{i \in I_j} a_{it}} (1 - q_{jt}) q_{jt}^{k-1} y_{jkt}.$$

To ensure that y_{jkt} has the right value, we add the following constraint:

$$\sum_{k=1}^{\sum_{i \in I_j} a_{it}} y_{jkt} \leq y_{jt} \qquad \forall j \in J, t \in T.$$

6.2.2 Further remarks

In the model description, we incorporated time flexibility in the execution of transportation requests. For each request $c \in C$, we have a set M_c of nodes, differing in execution time. In the construction of M_c, we can distinguish different types of requests. If we have a request without flexibility, we would have $|M_c| = 1$. If a patient has to be picked up after surgery, M_c would only contain start times after the requested time, which typically is the earliest possible pick-up time for this kind of request.

Even though we assume in the offline case that all information is known in advance, we cannot schedule a request before it is requested at the call center. We call this moment the release date of a request. If, for example, a request arrives at the call center at 14:00 that could be executed at any time between 13:00 and 15:00 if the request was made earlier, we do not allow for the request to be scheduled before 14:00. This is done, because the potential loss of efficiency as a result of the late request cannot be avoided by better planning. In the result section we do, however, evaluate the case where we ignore release dates. We do this to quantify the potential gain that can be obtained if hospitals send out a request earlier.

Another comment that should be made is that, up to now, we assumed that all transportation requests can be executed by the less equipped BLS ambulance. In practice, however, some transportation requests require an ALS ambulance. We can easily incorporate this in the model by adding the constraint $v_n = 0$ for all nodes $n \in M_c$ corresponding to request c of this type. Those requests will be assigned to an ALS ambulance at a particular base.

6.3 Online model

In the previous section, we introduced a model to solve the patient transportation problem if all the requests are known in advance. In practice, however, this is often not the case. Typically, a large fraction of the requests is released on the day

of execution. It even frequently happens that requests are made for immediate transportation. To incorporate this, we model the online version of the problem as an iterative Integer Linear Programming problem.

Since the patient transportation requests do not show too much structure, it is hard to predict future requests. Therefore, we introduce a local approach, in the terms of Chen and Xu (2006). We iteratively solve the offline version of the problem with the information that is available at that moment. Every time new information becomes available, i.e., a new request is released, we solve an instance of the offline model. This release date of a request can be as early as a day before the requested time or as late as the requested time.

When reoptimizing the schedule, we fix the assignments of ambulances to requests that have already started. For example, if a BLS ambulance is already with the patient, we cannot assign it to a different request. Even stronger, we do not allow for redirecting an ambulance that is on its way to a patient. The constraint that we cannot "change the past" does also apply to idle time of an ambulance.

When a request is completed, we remove it from the list of requests and do not include it in the following offline instances. The BLS shifts are adjusted accordingly. The new start location of the BLS shift is the drop-off location of the patient. Since we do not incorporate finished requests nor requests that are not yet released, the different offline instances that are solved in the online case are typically rather small. However, since for every release date of a request we have to solve an instance, we have many instances.

In the offline version of the model, we allow some flexibility in the execution time of a request. We do not incorporate an incentive to stimulate early execution of a request. However, in the online case this means that BLS ambulances might be left idle even when there are requests that can be executed. If a new request arises, it would have been better if we had scheduled the request earlier. To overcome this undesirable behavior, we implement a small reward for scheduling a request earlier. The reward is small enough so that it works as a tie-breaking rule only. Hence, the coverage in the offline version would not be affected by this modification. This might be considered a look-ahead approach in Chen and Xu's classification (Chen and Xu, 2006). Section 6.4.5 highlights the impact of this minor modification of the model.

6.4 Computational results

In this section, we discuss our computational results. First, we evaluate the solution provided by the model and compare it to the current execution. Then, we compare the offline and online cases and perform an extensive sensitivity analysis. In addition, we compare the effects of some modifications of the introduced model.

6.4.1 Data

We apply the models to the region of Utrecht, which is one of the largest ambulance regions (RAVs) in the Netherlands. As non-urgent transportation requests, we have the requests from the first three quarters of the year 2014. For all these requests, we know the start location, end location, release time, preferred start time, and realized duration. Note that the realized duration is in practice not known beforehand, but we assume that good estimations of this duration will be available. In Section 6.4.4, we investigate the effect of uncertainty in this duration. In addition, we know for each request whether or not it can be fulfilled by a BLS ambulance. A distinction between B1 and B2 calls is made. B1 calls are non-urgent patient transportations for which the medical condition of the patient is such that an ALS ambulance is required. For B2 calls, a BLS ambulance suffices, even though an ALS ambulance can also be used. So, in the model, we want to schedule B2 calls on BLS ambulances and assign the other B2 calls and all B1 calls to ALS bases such that the coverage provided by the remaining ALS capacity is maximized. To determine this coverage, we need some additional input data. We need the demand locations for emergency calls, the demand for each demand location, the number and location of the ALS ambulances, the busy fraction, and a time threshold in which the emergency calls should be served. As demand locations for emergency calls, we take the four digit postal codes, which gives a total of 217 demand points. The demand is time-dependent and is based on historical data provided by the ambulance provider. For the base locations, we take the current 12 base locations, and the number of available ALS ambulances per time period is obtained from the current shift schedule. The busy fraction is calculated by dividing the total workload of emergency calls by the total available ALS capacity. This capacity is obtained from the current shift schedule and the total workload is obtained by multiplying the total number of emergency calls with the average duration of the calls. As time threshold for determining the coverage, we take 15 minutes, which is the standard in the Netherlands. Since the pre-trip delay is assumed to be equal to 3 minutes, this gives a maximum drive time of 12 minutes.

We apply the model to all days in the first nine months of 2014 separately. Since the workload during the night is very low, we do not see the need to run the model for nine months consecutively and we can solve the model for different days separately. We use the current BLS shift schedule as input for the model. The schedule contains 10 shifts on weekdays, 7 shifts on Saturdays, and 5 shifts on Sundays. Since the schedule includes one shift that runs over multiple days, we split this shift into two parts: one that runs from 00:00 till 08:00 and one that starts at 23:00. The shift that starts at 23:00 is not required to return to its base before midnight. For weekdays, this gives the shift schedule as depicted in Figure 6.2.

We include all patient transportation requests in this study. We distinguish two categories: B1 and B2, where B1 requests are the patient transportation requests that require an ALS ambulance and B2 requests can be executed by

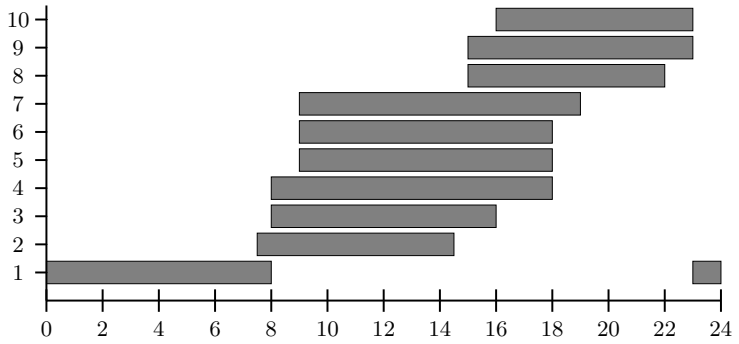


Fig. 6.2: *Current shift schedule. Shift one runs over two different days, and is therefore split in two parts.*

a BLS ambulance. In the considered period, we have a total of 20,278 patient transportation requests, of which 10,044 are B2. A workday has on average twice the number of B calls as a day in the weekend.

For each request, we have a given release date. For approximately 50% of the B2 requests, this release date equals the requested execution time. For B2 requests, we allow for flexibility in scheduling the request by scheduling the request between one hour before and one hour after the requested time. However, we do take the release date into consideration. So, if, for example, the release date equals the requested execution time, we do not allow the request to be executed before its requested time.

Since the model requires a discrete number of possible start times for a request, we use a time step of 15 minutes, which gives a maximum of 9 start times for each request, i.e., $|M_i| \leq 9$. We do not allow for time flexibility for B1 requests, these requests are assigned to a base at their requested time.

In the following sections, we refer to the instances with the settings described here as the base case.

6.4.2 Results base case

If we consider the results of the base case, we see that 87.8% of all B2 calls can be executed with a BLS ambulance. In the current execution, this is only 80.8%. Since the base case does not use any information about the future, this represents a solution that is in principle feasible in practice. In the model, we even allow for less flexibility than in practice. In the current execution, 13.5% of all calls are executed more than 60 minutes after the requested time. This is not allowed in the model, where each call is scheduled within 1 hour from the requested time. Figure 6.3 shows the number of B2 calls that is executed by a BLS ambulance per day of the week.

We see that both the number of served as well as the number of unserved calls increases as the number of requests increases. So, more calls allow for more

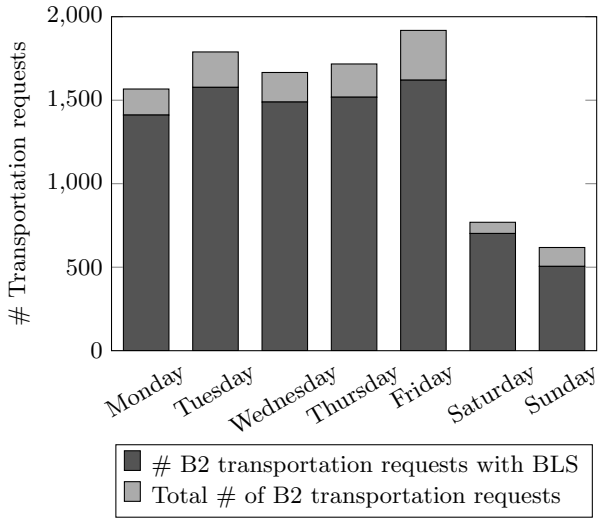


Fig. 6.3: *Number of B2 transportation request that can be served by a BLS ambulance per day of the week in the base case.*

efficient scheduling of calls on the BLS ambulances, but this efficiency gain is not sufficient to fully compensate for the higher workload. Figure 6.4 shows the average utilization of the different shifts on weekdays. The shift numbers correspond with the numbering of the shifts in Figure 6.2. We see that the afternoon shifts can obtain a utilization of almost 80%, whereas the evening shifts have a utilization of less than 60%. The night shift has very low utilization, but this shift is also used to provide acute home care which is not included in this utilization. The figure further shows that approximately 70% of the busy time of an ambulance is spent with a patient. The remaining 30% of the time, the ambulance is on its way to a patient. The figure indicates that it might be worthwhile to move an evening shift towards the afternoon. Chapter 7 discusses the use of the model to evaluate the impact of such changes.

6.4.3 Offline versus online scheduling

In this section, we compare three different cases of dealing with the dynamic aspects of the data. In the first case, release dates are not included in the model. This corresponds to the case where all calls are known at the start of the day. This deviates from practice in two ways: first, we have more flexibility in B2 calls with release date within 1 hour of the requested time; second, since all information is known in advance, more efficient schedules can be made. In the second case, we assume all information is known in advance, but the release dates have to be respected. This gives a feasible solution for the base case, as all constraints of the model are respected. However, since, in practice, calls are not

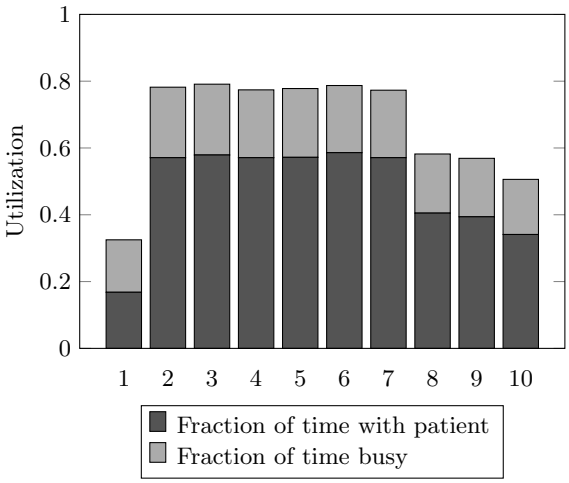


Fig. 6.4: *Average utilization of the different shifts in the base case. In light, the actual time with a patient. In dark, the total utilization, including the travel time to the patient.*

known before their release date, this schedule could not be derived in real-time. This does give an upper bound on the performance of the base case. In the third case, which corresponds to the base case, calls become available at their release date. In this setting, this is called the online case. The difference in performance between the second and third case gives us the loss in efficiency as a result of making the wrong decision as a result of not knowing the future. The difference between the first and second case measures the impact of the loss of flexibility as a result of late notification of the hospital. Together, they give the loss in performance as a result of not knowing all requests at the start of the day.

Figure 6.5 shows that the impact of flexibility is smaller than the impact of knowing future requests. This is because 49.7% of the B2 requests is already known an hour before its requested time. For these requests, there is no difference between the first and second case. In the case where we do not consider release dates, we can execute 95.9% of the B2 requests with a BLS ambulance. For the offline case, this is 94.0% and for the online case, this is 87.8%. Figure 6.6 shows the remaining coverage of the ALS ambulances for the emergency calls. The same behavior can be seen here, i.e., flexibility has less impact than having information of the future.

6.4.4 Sensitivity analysis

In this section, we perform a sensitivity analysis on the effect of flexibility and discretization in the execution time of requests. Furthermore, we analyze the impact of uncertain request duration.

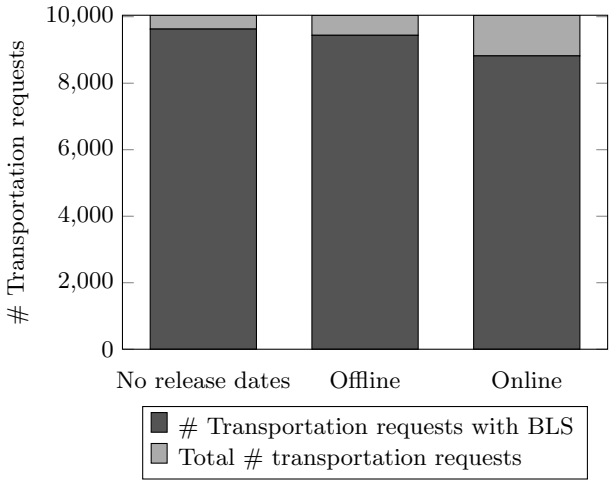


Fig. 6.5: *Number of B2 transportation requests served by a BLS ambulance for online and offline version of the model.*

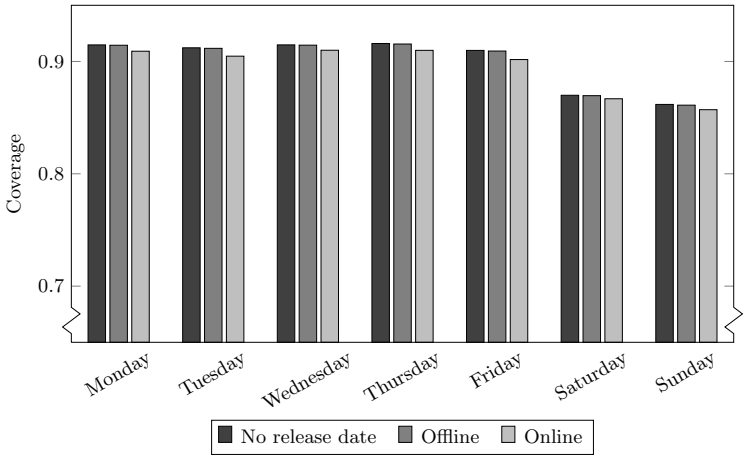


Fig. 6.6: *Remaining coverage for emergency calls for online and offline version of the model.*

Effect of flexibility

In the base case, we allow for a flexibility of 1 hour around the requested time for B2 calls. Here, we will evaluate the impact of reducing this flexibility to 15 minutes or 30 minutes. If a flexibility of 15 minutes is used, calls can either be executed 15 minutes before the requested time, at the requested time, or 15

minutes after the requested time. Clearly, reducing the flexibility will reduce the performance.

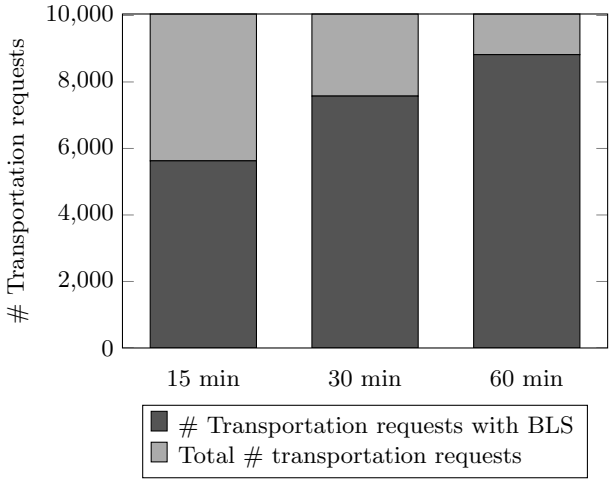


Fig. 6.7: *Number of B2 transportation requests served by a BLS ambulance for different levels of flexibility in execution time.*

Figure 6.7 shows that with a flexibility of 15 minutes, we can execute only 56.0% of the B2 requests with a BLS ambulance. By increasing the flexibility to 30 minutes, this percentage increases to 75.4%. For a flexibility of 60 minutes, which corresponds to the base case, this is 87.8%.

From the input data, we know that 48.4% of the B2 requests is released at their requested time. With the time step set to 15 minutes, we have only two moments to schedule the requests if we set the time window to $\pm$ 15 minutes. Namely, either at the requested time or 15 minutes after the requested time. Picking up a patient immediately is only possible if an ambulance would already be at the patient's origin location. As this rarely happens, the patient can only be scheduled 15 minutes after the requested pick-up time. However, this is also hard to realize as the travel time is usually more than 15 minutes. Therefore, only a small portion of the requests released at its requested time can be served by a BLS ambulance within a 15 minute time frame.

By increasing the time window to $\pm$ 30 minutes, we obtain a huge improvement as most patients can be reached within 30 minutes by a BLS ambulance. Increasing the time window further to $\pm$ 60 minutes increases the flexibility of scheduling requests enormously. This can also be seen in Figure 6.8 which shows the remaining coverage for the three considered cases.

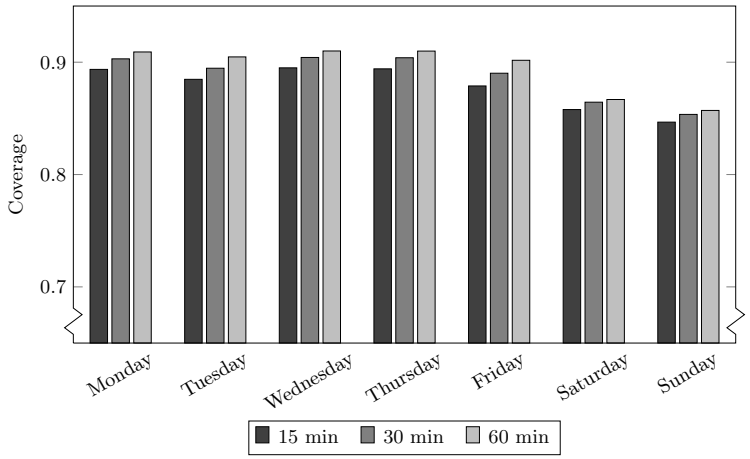


Fig. 6.8: *Remaining coverage for emergency calls for different levels of flexibility in execution time.*

Effect of discretization

Since the model only allows for a discrete number of nodes, we have to discretize the start time of requests. In the base case, we use a 15 minute discretization, meaning that a request can be executed at a maximum of nine different times. Namely, 1 hour, 45 minutes, 30 minutes and 15 minutes before the requested time, at the requested time, and 15 minutes, 30 minutes, 45 minutes and 1 hour after the requested time. For each allowed start time, we have a node in the network. Ideally, we would not discretize, and therefore, we evaluate the impact of discretizing. We compare the base case, with a discretization of 15 minutes, to the case with 5 minute and 30 minute discretization. Note that the number of nodes increases significantly if a time step of 5 minutes is used. In that case, we have a maximum of 25 nodes for each request compared to 9 nodes in the base case. By setting the time step to 30 minutes, we have a maximum of 5 nodes per request.

Figure 6.9 shows that by setting the time step to 5 minutes, we can schedule 89.5% of the B2 requests on the BLS ambulances. This decreases to 87.8% for the base case with a time step of 15 minutes and it decreases further to 85.0% when we set the time step to 30 minutes. This shows an improvement of 2.8% if we go from 30 to 15 minutes and an increase of 1.7% if we go from 15 minutes to 5 minutes. This means that we can obtain a huge improvement by decreasing the time step from 30 to 15 minutes whereas the computation time increases from approximately 3 to 4 seconds per instance. The latter is because we only go from 5 to 9 scheduling options, whereas if we decrease the time step from 15 to 5 minutes, we go from 9 to 25 scheduling options. Therefore, decreasing the time step from 15 to 5 minutes does not imply much gain in the quality of

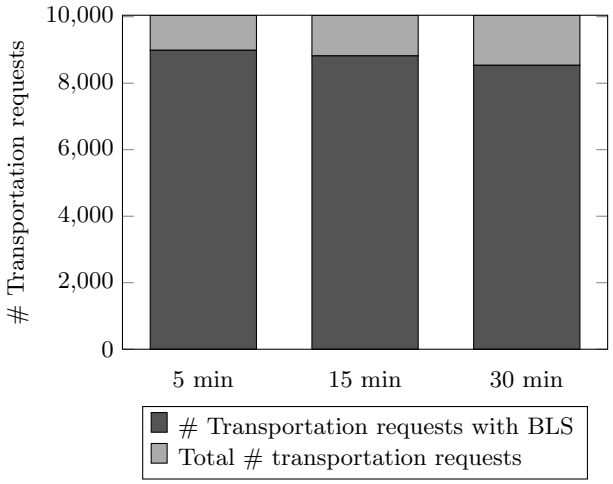


Fig. 6.9: *Number of B2 transportation requests served by a BLS ambulance for different levels of discretization.*

the solution but does increase the computation time from approximately 4 to 15 seconds per instance. Note that we have to solve an instance every time new information becomes available. The same effect can be seen in Figure 6.10, which presents the remaining coverage for the emergency calls.

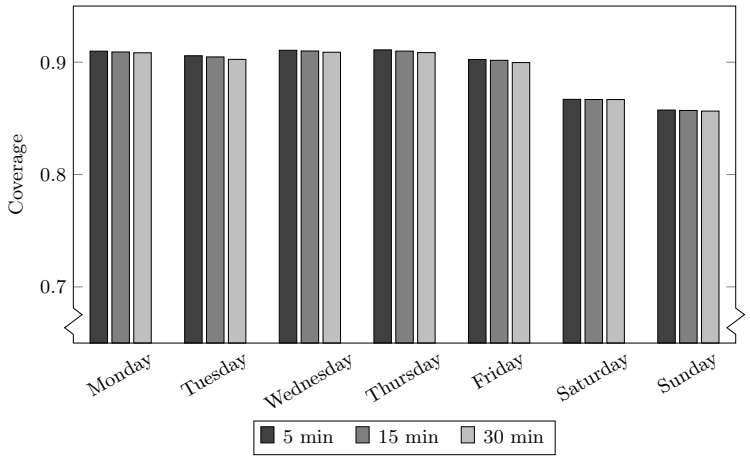


Fig. 6.10: *Remaining coverage for emergency calls for different levels of discretization.*

Effect of uncertain request duration

Up to now, we assumed that the duration of a request is known at the release date. In the base case, we take the realized duration in practice as the duration of a request. However, the exact duration of a request is typically not known at the moment the request arrives at the call center. In this section, we evaluate the impact of uncertainty in the request duration.

We assume that we know an expected, minimum, and maximum duration for each request. Based on some distribution, we generate the real duration of the request. We handle the uncertainty in the following way. For each request, we initially assume that the expected duration is its real duration. If the request finishes earlier than expected, the ambulance is available at this earlier time and we reoptimize the schedule given this new information. If a request is not yet finished at its expected end time, we reoptimize the schedule assuming that the duration of the request is its maximum duration. Since the request has already been started, it is not possible to change its assignment. Again, the request might finish earlier than expected, in which case we follow the previously described procedure. Note that the delay in the execution of a request can cause a shift to run in overtime. In the original version of the model, we do not allow for this to happen, but given the uncertain duration, this is unavoidable. The overtime can, however, never be more than the difference between the expected and maximum duration of the last request scheduled on a shift. Similarly, it can happen that, as a result of the longer duration of a call that is assigned to an ALS ambulance, the capacity at the selected base does not suffice. As we again cannot change the assignment, this would lead to overtime of an ALS shift.

To evaluate the impact of the uncertainty in the request duration, we apply the new version of the model for varying minimum and maximum duration. We compare the base case to the case with a maximum deviation of 5%, 10%, and 20% of the expected duration. To generate the real duration, we use the triangular distribution. Generating from this distribution can be done by

$$X = \begin{cases} \min + \sqrt{U(\exp - \min)(\max - \min)} & 0 \leq U \leq 0.5 \\ \max - \sqrt{(1-U)(\exp - \min)(\max - \min)} & 0.5 \leq U \leq 1, \end{cases}$$

where U is uniformly distributed in the interval $(0, 1)$. One advantage of the triangular distribution is that it has a continuous density function, whereas, for example, a truncated normal distribution has jumps at the minimum and maximum call duration.

Figure 6.11 and 6.12 show the results for the base case (0%) and for a flexibility of 5%, 10%, and 20%. As expected, the number of executed BLS requests decreases with increasing uncertainty. The same pattern is observed in the remaining coverage for emergency calls. In the 20% case, we have to execute an additional 0.8% of the B2 requests by an ALS ambulance.

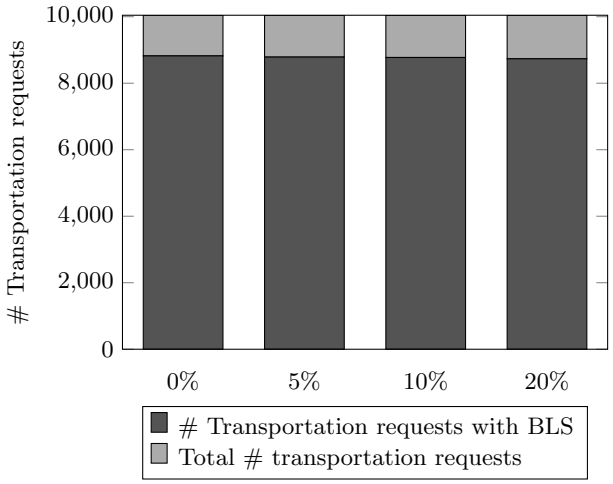


Fig. 6.11: *Number of B2 transportation requests served by a BLS ambulance for different levels of call duration uncertainty.*

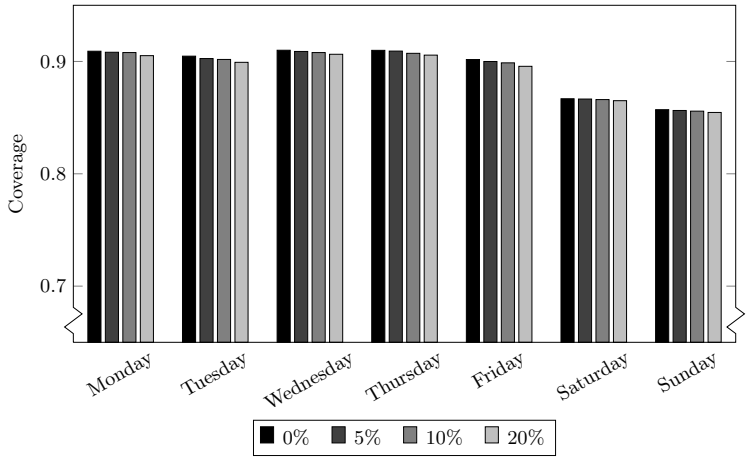


Fig. 6.12: *Remaining coverage for emergency calls for different levels of call duration uncertainty.*

6.4.5 Effect of online scheduling rule

In Section 6.3, we discussed a tie-breaking rule to stimulate the early execution of requests. The main reason for including this rule is to avoid unnecessary idle time for BLS ambulances. Without this online scheduling rule, it can occur that BLS ambulances are idle, even though requests are available for execution. Note that since it is only a tie-breaking rule, adding the rule does not change the coverage of

the offline version. In Figure 6.13, we see that by including the online scheduling rule, 87.8%, instead of 83.6%, of the requests are executed by a BLS ambulance. This corresponds to an increase of 4.2 percentage point. Figure 6.14 shows that also the coverage increases by adding this simple rule.

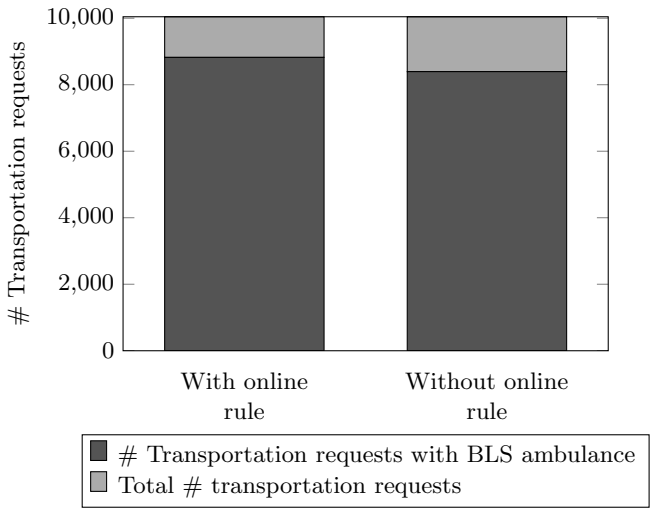


Fig. 6.13: *Number of B2 transportation requests served by a BLS ambulance for the online version of the model with and without the online scheduling rule.*

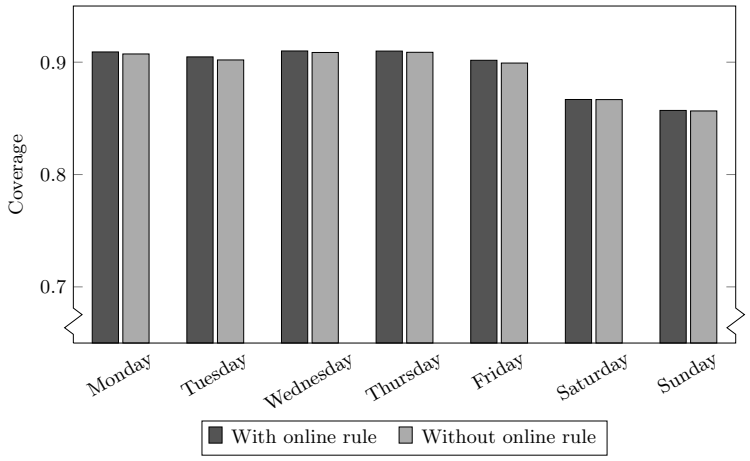


Fig. 6.14: *Remaining coverage for emergency calls for the online version of the model with and without the online scheduling rule.*

6.4.6 Effect of maximizing number of executed requests

One novelty of our model is the use of the coverage for emergency calls as the objective in scheduling patient transportation requests. Another, more common, approach is to exclude the coverage and simply focus on the number of requests executed by a BLS ambulance. One might expect that by maximizing the number of requests executed by BLS ambulances, and thus minimizing the workload on the ALS ambulances, the coverage for emergency calls is maximized as well. However, Figure 6.15 and Figure 6.16 show that this is not the case. The two figures compare the results of the model in the base case with a version where the objective function is changed such that the number of requests executed by a BLS ambulance is maximized. The objective is then

$$\max \quad \sum_{n \in M} Z_n.$$

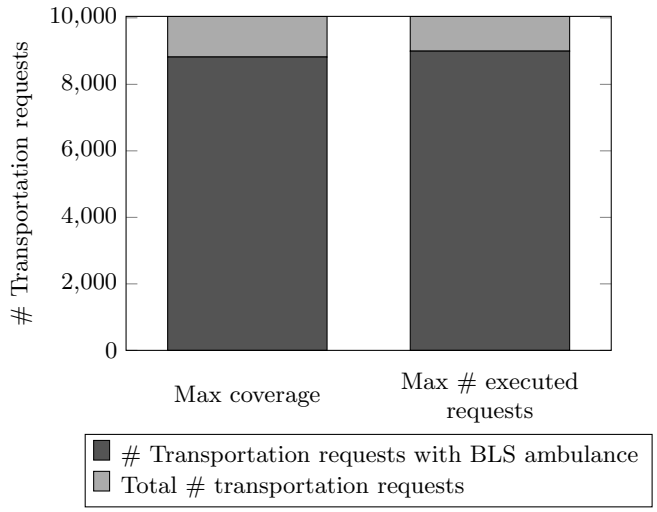


Fig. 6.15: *Number of B2 transportation requests served by a BLS ambulance for maximizing number of calls served by BLS ambulance and for maximizing the emergency coverage.*

We see that even though 1.8% more requests are executed by BLS ambulances, the coverage still slightly decreases. Apparently, it is important to carefully select which requests are not assigned to a BLS ambulance. The model ensures that ALS ambulances are only used for patient transportation requests in time periods with sufficient capacity for emergency calls. In Chapter 7, we will see that in evaluating different schedules, this behavior is even more apparent.

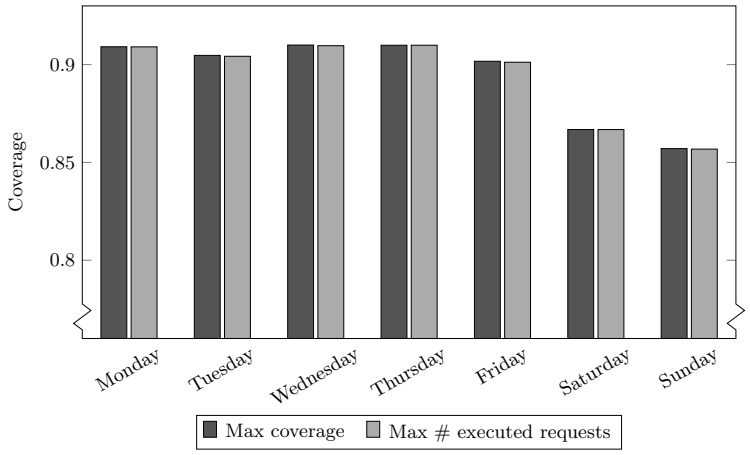


Fig. 6.16: *Remaining coverage for emergency calls for maximizing the number of calls served by BLS and for maximizing the emergency coverage.*

6.5 Conclusions

We have introduced a method to optimize the routes of Basic Life Support (BLS) ambulances for non-urgent patient transportation while maximizing the remaining Advanced Life Support (ALS) capacity for emergency calls. We consider the situation where part of the non-urgent transportation requests are known at the start of the day and the remainder of the requests arrives throughout the day. Most of these transportation requests can be executed by BLS ambulances, but due to the limited capacity of BLS ambulances and the basic level of care provided by the BLS ambulances, several of the non-urgent transportation requests have to be executed by ALS ambulances. As the primary task of ALS ambulances is to respond to emergency calls, we have to make sure that the non-urgent transportation requests are assigned to the ALS ambulances in such a way that the remaining coverage for emergency calls is maximized. We include this by setting our objective function such that expected coverage, as defined by MEXCLP (Daskin, 1983), is maximized.

One of our contributions is taking the coverage of ALS ambulances for emergency calls into account. Most papers make a strict distinction between non-urgent and urgent transportation requests. By also allowing ALS ambulances to respond to non-urgent transportation requests, we are able to use fewer BLS ambulances, and thus, improve the utilization of the BLS ambulances. This means that both the ALS and BLS ambulances are used more efficiently and we are better able to meet the targets. When we compare our approach to the standard approach of maximizing the number of requests executed by BLS ambulances, we see that we could execute more requests with a BLS ambulance, but that this

reduces the remaining coverage of ALS ambulances for emergency calls. Even though this reduction is small, we see that our objective function is needed to maximize the remaining coverage.

Another contribution is that we solve the problem as an Integer Linear Program (ILP) instead of using a heuristic approach. However, we cannot model the problem exactly as we have to discretize the time in order to make the ILP solvable within a reasonable amount of time. We choose to model the problem with a time step of 15 minutes. Our sensitivity analysis shows that with a time step of 15 minutes, we can assign 87.8% of the requests to a BLS ambulance whereas we could assign 89.5% of the requests to a BLS ambulance when we would use a time step of 5 minutes. However, the loss of coverage for the emergency calls is only 0.08% if we compare the 5 and 15 minutes case. In addition, the computation time increases from approximately 4 to 15 seconds per instance. Note that we have to solve an instance every time new information becomes available. As our intention is to implement the method for real-time planning of BLS ambulances, this increase in computation time is not preferred in practice.

One disadvantage of our approach is that we only take the expected request duration into account. Our sensitivity analysis shows that the number of requests served by a BLS ambulance decreases with 0.4% when we allow 10% deviation in the duration of the requests. This percentage increases to 0.8% if we allow 20% deviation. As this decrease is very moderate and we expect dispatchers to be able to make good predictions, we do not consider this uncertain call duration a significant problem.

Although most non-urgent patient transportation requests cannot be predicted, some can. For example, some of the patients that have to be transported from home to a hospital also need to be transported back home on the same day. For future research, it would be interesting to investigate the potential benefit of taking expected future requests into account. Schilde et al. (2011) already showed that using this information can improve the results significantly. This effect is also shown in Figures 6.5 and 6.6, where we compare our base case to the case where we would have all the information available beforehand.

As the idea for this research originated from one of the ambulance providers in the Netherlands, we aimed at developing a method that could be used in practice for the real-time planning of BLS ambulances. Despite the fact that the developed method is suitable to do this, implementing our approach in the system of the ambulance provider is a challenging task. It would be interesting to see how the results described in this chapter hold up in practice.

Even though the implementation of the model for the real-time scheduling of patient transportation requests requires more work, two other applications that are easier to implement come to mind. First, the model could be used to tune the shift schedule of the BLS ambulances. The developed method can already be used to compare several schedules. In the next chapter, we will discuss a project with the ambulance provider in the region of Utrecht, in which we used the model to propose changes in the current shift schedule to improve coverage. For future research, it would be interesting to develop a method that can optimize the shift

schedule such that a good balance between the efficiency of BLS ambulances and the remaining coverage of ALS ambulances can be obtained.

The second application of the model is to steer the incoming transportation requests of the hospitals such that the requests are spread more equally over the day. Currently, there is a peak load of transportation requests at 11:00 and 15:00 of patients that are admitted to or discharged from the hospital. This means that around these times, not enough BLS ambulances are available, whereas at other times there are BLS ambulances available. With the use of the obtained information in this study, the ambulance providers are able to set up a plan with the hospitals to spread the requests more evenly over the day. In this way, the BLS ambulances can be used more efficiently, the remaining coverage for emergency calls can be improved, and the requested pick-up times can be met more often.

6.A Model formulation

$$\max \quad \sum_{j\in J}\sum_{t\in T} d_{jt} \sum_{k=1}^{\sum_{i\in I_j} a_{it}} (1-q_{jt})q_{jt}^{k-1} y_{jkt}$$

$$\text{s.t.} \quad \sum_{n\in M_c}\left(\sum_{i\in I} u_{in} + v_n\right) = 1 \qquad \forall c \in C$$

$$\sum_{s\in S}\sum_{n'\in A_n} w_{nn's} = v_n \qquad \forall n \in M$$

$$\sum_{n'\in B_n} w_{n'ns} - \sum_{n\in A_n} w_{nn's} = -1 \qquad \forall n \in O, s \in S$$

$$\sum_{n'\in B_n} w_{n'ns} - \sum_{n\in A_n} w_{nn's} = 0 \qquad \forall n \in M, s \in S$$

$$\sum_{n'\in B_n} w_{n'ns} - \sum_{n\in A_n} w_{nn's} = 1 \qquad \forall n \in D, s \in S$$

$$x_{it} + \sum_{n\in M} b_{int} u_{in} = a_{it} \qquad \forall i \in I, t \in T$$

$$\sum_{i\in I_j} x_{it} \geq y_{jt} \qquad \forall j \in J, t \in T$$

$$\sum_{k=1}^{\sum_{i\in I_j} a_{it}} y_{jkt} \leq y_{jt} \qquad \forall j \in J, t \in T$$

$$u_{in}, v_n, w_{nn's} \in \{0,1\} \qquad \forall n, n' \in N, i \in I, s \in S$$

$$x_{it}, y_{jt} \in \mathbb{N} \qquad \forall i \in I, j \in J, t \in T$$

7

Application of non-urgent patient transportion model

7.1 Introduction

In Chapter 6, we introduced a model to determine routes for BLS ambulances such that many of the B2 calls can be executed, and the impact on the coverage for emergency calls is minimized. Even though the model is developed for the real-time scheduling of patient transportations, two other applications of the model were mentioned. The first one is to use the model to control the input of transportation requests. The model can provide insight into the remaining BLS capacity at particular times. Call takers at the dispatch center can use this information to give alternative times if a call arises with a requested time at which no capacity is left. Second, the model can be used to evaluate shift schedules for BLS ambulances. By applying the model with a given schedule and evaluating the outcomes, we can come up with alternative, better schedules. These new schedules can again be evaluated with the model to assess the impact. In this chapter, we analyze the current shift schedule for the region of Utrecht by means of the introduced model. Furthermore, based on the results, we suggest slight changes that could lead to better performance. These alternative schedules are again evaluated with the model.

7.2 Data analysis

As in the previous chapter, we consider the region of Utrecht. We evaluate the current shift schedule for workdays only, since call volumes are lower in the weekends. At first, all B calls from October 2013 till September 2014 are included. Table 7.1 shows the average number of calls per day for the four considered quarters. From the table, we see that the fourth quarter of 2013 shows lower call volumes than the three quarters in 2014. A possible explanation is that at December 23, 2013, a new hospital was opened in the region that changed the demand pattern for patient transportations. For this reason, we exclude the

fourth quarter of 2013 from the data and only consider the first three quarters of 2014.

Table 7.1: *Average number of B2 calls per day per quarter.*

	Q4 2013	**Q1 2014**	**Q2 2014**	**Q3 2014**
Monday	33	42	36	47
Tuesday	35	50	47	49
Wednesday	34	41	45	45
Thursday	31	45	43	48
Friday	24	51	50	51
Saturday	15	19	19	22
Sunday	14	16	16	16

The current shift schedule for BLS ambulances is given in Figure 7.1. The schedule consists of one night shift, six day shifts, and three evening shifts. The aim of this study is to evaluate whether this schedule is appropriate.

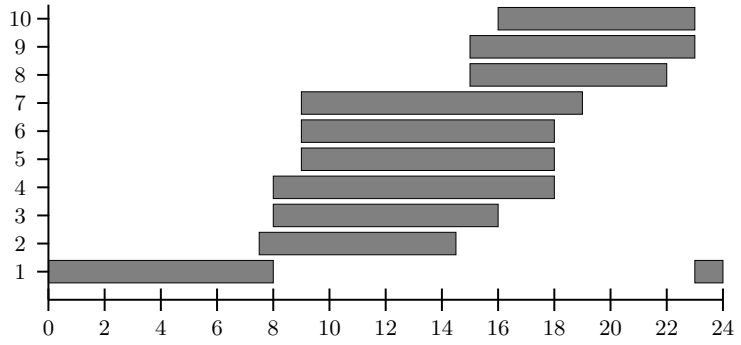
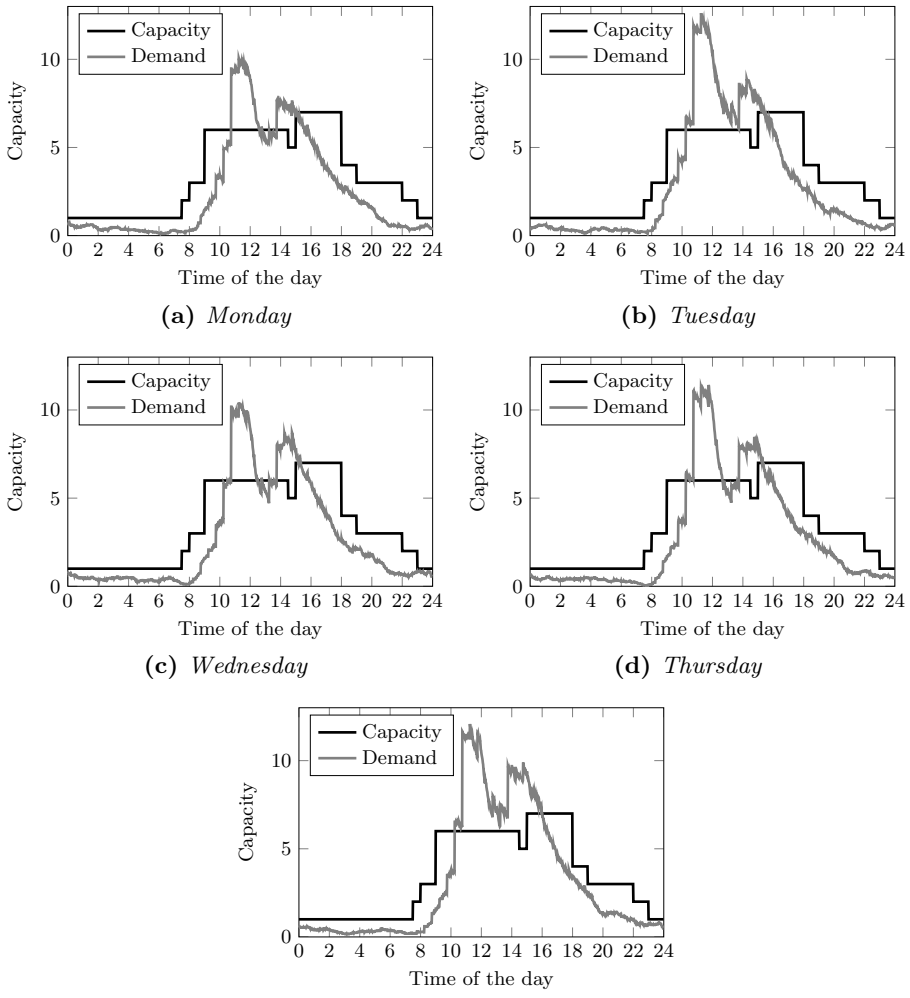


Fig. 7.1: *Current shift schedule.*

First, we compare the available capacity with the demand for BLS ambulances. Figure 7.2 shows for every workday the average capacity and average demand. To compute the demand, we include all B2 calls and increase the demand by one unit from the requested time for a period of time equal to the realized call duration, which is equal to the time from the ambulance assignment till the patient drop-off. To compensate for the drive time to and from the patient, we add 15 minutes on both sides. The figure clearly shows that there is a large peak in demand at 10:30 and 11:00 in the morning. This peak is caused by patients that have to stay overnight and can leave the next day after they have seen the physician, which typically happens between 10:00 and 11:00. There is another peak right after lunch. When comparing demand with capacity, we see

that there is undercapacity in the later morning and early afternoon, whereas there is overcapacity in the evening. This suggest that moving some evening shifts to the afternoon might improve the performance. Next, we will apply the model as described in Chapter 6 to further evaluate the current schedule.



(a) *Monday*

(b) *Tuesday*

(c) *Wednesday*

(d) *Thursday*

(e) *Friday*

Fig. 7.2: *Demand and capacity for BLS ambulances for different days of the week.*

7.3 Evaluation of current schedule

We apply the online version of the model, which corresponds to the base case, on the data from January till September 2014. Figure 7.4 gives the number of unserved calls at different times of the day. This approximately corresponds to the workload for ALS ambulances for patient transportations. For every day, we see a peak in unserved calls around 14:00. We further see a smaller peak in the late evening when the evening shifts end. Especially on Friday, the afternoon peak is very high. In the early evening, almost all B2 calls can be served with the available BLS ambulances.

Next, we consider the utilization of the different shifts. Figure 7.3 shows the utilization of the different shifts on the different days of the week. Clearly, the night shift has the lowest utilization, but this shift is also used for providing acute home care at night. However, also the three evening shifts have lower utilization than the day shifts. The day shifts have a utilization of 70-80%, whereas the evening shifts typically have a utilization of less than 60%. This again suggests that it might be beneficial to move some evening shifts to the afternoon.

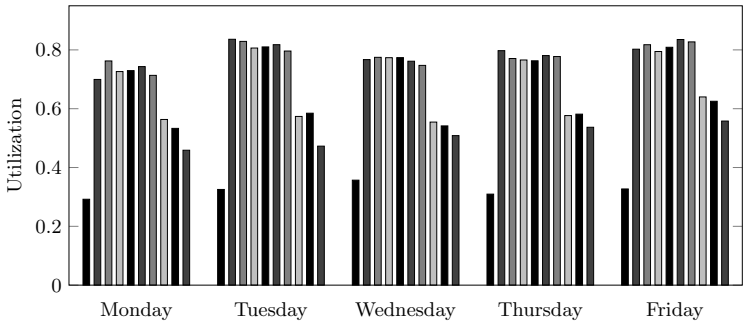
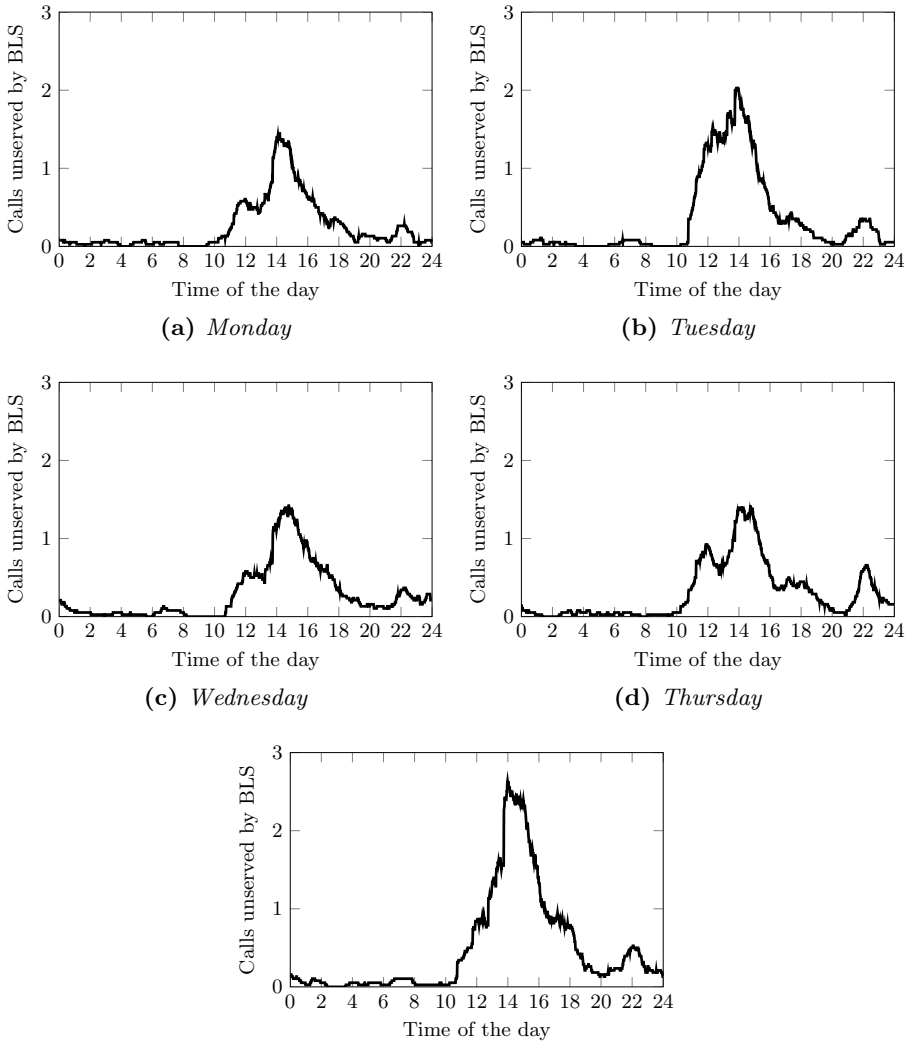


Fig. 7.3: *Utilization of the different shifts in the current schedule.*

7.4 Alternative schedules

Based on the previous observations, we propose two different schedules that are based on the current schedule, by moving only one or two shifts. Additionally, we define a schedule that matches the demand and capacity by completely redefining the schedule. This is done by the ILP formulation that is given in Appendix 7.A. For future research, it would be interesting to see how better schedules can be developed. For the first new schedule (Schedule 1), we let shift 10 start earlier so that it starts at 12:00 and ends at 19:00. The second new schedule (Schedule 2) is created by letting both shift 8 and shift 10 start earlier. The two new schedules, as well as the result of the ILP formulation (Schedule 3), are given in Figure 7.5. Note that the total capacity is not increased and that we have exactly the same shifts. Only the start times of the shifts have changed.

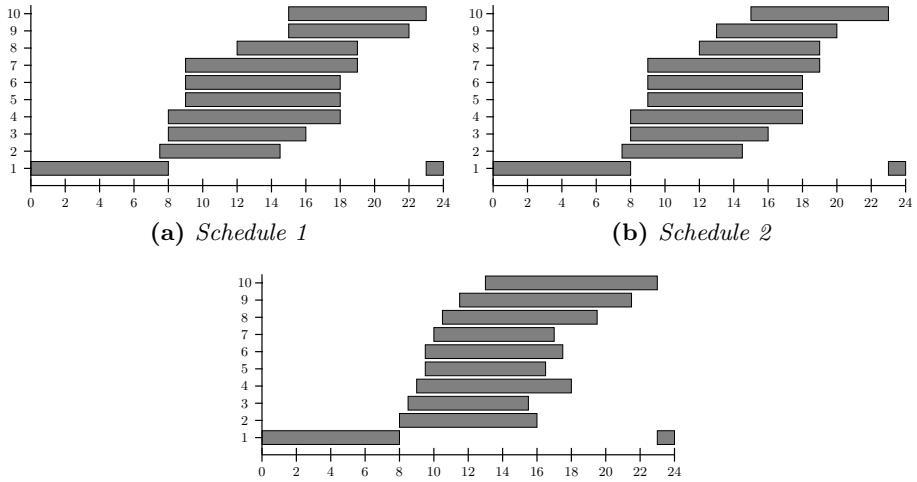


(a) *Monday* **(b)** *Tuesday*

(c) *Wednesday* **(d)** *Thursday*

(e) *Friday*

Fig. 7.4: *Distribution of unserved calls over the different days, given the current schedule.*



(a) *Schedule 1* **(b)** *Schedule 2*

(c) *Schedule 3*

Fig. 7.5: *Three alternative schedules.*

7.5 Evaluation of alternative schedules

First, we compare the four schedules on the number of B2 calls that are served by a BLS ambulance. Figure 7.6 shows that all three new schedules are able to serve more BLS calls. Schedule 3 clearly serves most calls, and Schedule 1 serves slightly more calls than Schedule 2. As unserved calls occupy an ALS ambulance for some time, it is important to evaluate the distribution of unserved calls over the day. Figure 7.7 shows this for the four schedules. We see that all three schedules are able to lower the peak in the early afternoon. On the other hand, the number of unserved calls in the evening is increased. Schedule 3 has its main peak in the late afternoon. In all cases, the unserved calls are spread more evenly over the day. When looking at the utilization of the shifts in Figure 7.8, we see that all new schedules result in a more balanced workload among the shifts. The highest utilization is achieved by Schedule 3. Finally, if we consider the remaining coverage for emergency calls, we see that Schedule 2 results in lower coverage than the current schedule. Schedule 1 and 3 provide a coverage improvement of 0.04% and 0.1%, respectively. So, even though Schedule 2 serves more calls by a BLS ambulance, the impact on emergency coverage is larger. The reason for this is that the ALS capacity in the evening is smaller, and the impact of using one ALS ambulance for a patient transportation is larger.

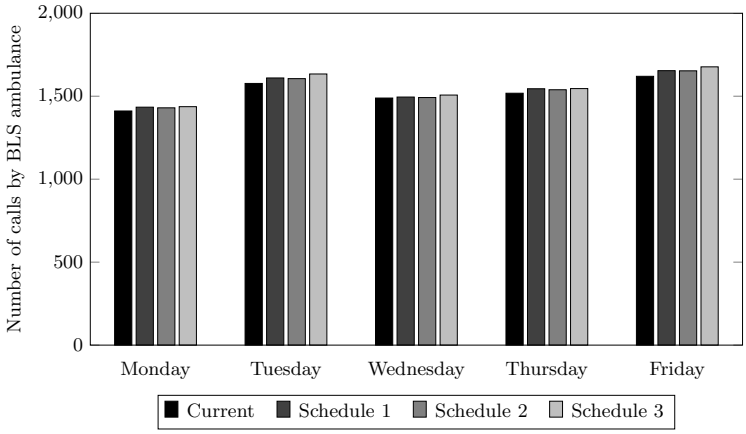


Fig. 7.6: *Total number of B2 calls served by a BLS ambulance for the four considered schedules.*

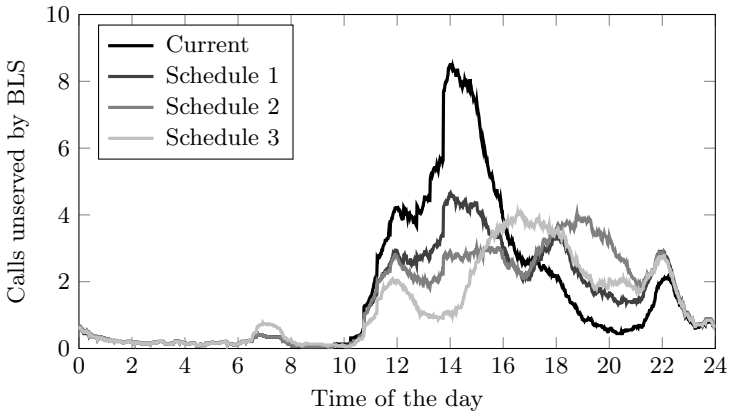
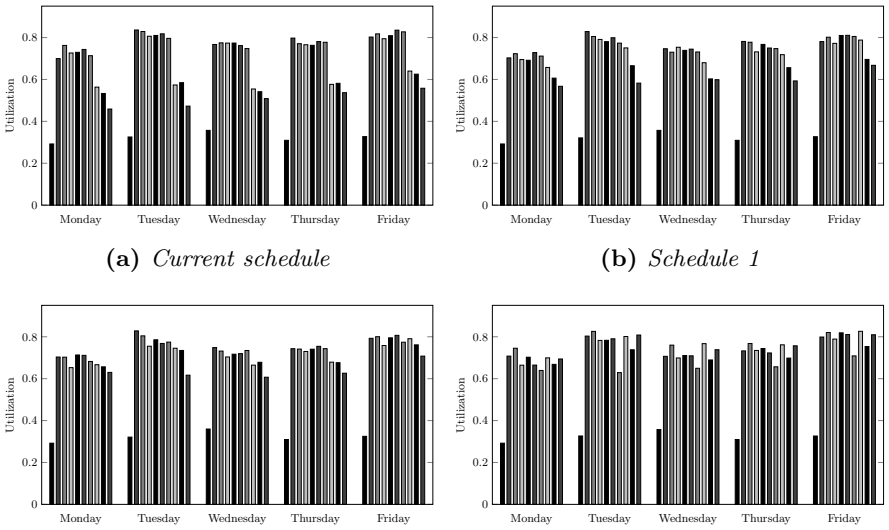


Fig. 7.7: *Distribution of unserved calls over the day for the four considered schedules.*

7.6 Conclusions

In this chapter, we have seen how the model as described in Chapter 6 can be applied to analyze shift schedules for BLS ambulances. We evaluated three different schedules, of which two were based on the evaluation of the current schedule. The results show that by moving one evening shift to the afternoon, we can obtain a 0.04% coverage improvement. In total, 123 more B2 calls can be served by a BLS ambulance and the utilization is more balanced among the different shifts as a result of this small change. Additionally, the unserved calls,



(a) *Current schedule* **(b)** *Schedule 1*

(c) *Schedule 2* **(d)** *Schedule 3*

Fig. 7.8: *Utilization of the different shifts for the different schedules.*

and thus the workload on ALS ambulances, are divided more equally over the day.

We further observed that even though Schedule 2, where two shifts are moved, can serve more B2 calls with a BLS ambulance, the resulting coverage for emergency calls will decrease. Finally, we also evaluated a schedule that was the result of an ILP formulation that matches demand and capacity for BLS ambulances. Without adding any capacity, 186 more calls can be assigned to a BLS ambulance and a 0.1% coverage improvement can be obtained, compared to the current schedule. For future research, it would be interesting to develop better models to generate new schedules. The new schedules could then again be evaluated with the model.

7.A ILP formulation

In this appendix, the ILP formulation that is used to generate the new schedule is given.

7.A.1 Input

S	set of BLS shifts
T	set of time periods $= \{1, 2, ..., 48\}$
n_t	required capacity during time period $t \in T$
d_s	duration of BLS shift $s \in S$

7.A.2 Variables

a_t	available capacity during time period $t \in T$
c_t	shortage of capacity during time period $t \in T$
x_{st}	binary variable indicating whether shift $s \in S$ starts at the beginning of time period $t \in T$

7.A.3 Model

$$\begin{aligned}
\min \quad & \sum_{t \in T} c_t \\
\text{s.t.} \quad & \sum_{t \in T} x_{st} = 1 && \forall s \in S \\
& a_t = \sum_{s \in S} \left(\sum_{t' \geq t - d_s + 1}^{t} x_{st'} + \sum_{t' \geq t - d_s + 49}^{48} x_{st'} \right) && \forall t \in T \\
& c_t \geq n_t - a_t && \forall t \in T \\
& c_t, a_t \geq 0 && \forall t \in T \\
& x_{st} \in \{0, 1\} && \forall s \in S, t \in T
\end{aligned}$$

Part III

Air Ambulances

8

Simulation and optimization for air ambulance provider in Ontario

8.1 Introduction

In the previous chapters, we mainly considered the Dutch ambulance services. Now, we discuss some aspects of the logistics of an air ambulance provider in Ontario, Canada. Ontario is one of the ten provinces in Canada and covers an area of more than 1 million square kilometer. With a population of 13.8 million, this gives a population density of 12.8 inh/km^2 (Statistics Canada, 2015). In order to provide good health care to its inhabitants, Ontario uses land ambulances, as well as, aircrafts and helicopters. The not-for-profit organization Ornge provides health care and transportation of patients, mainly by aircrafts and helicopters. As the largest air ambulance provider in Canada, Ornge serves more than 18,000 patients a year. For those calls, Ornge uses two types of aircrafts: rotor wing (RW) aircrafts and fixed wing (FW) aircrafts. Besides the aircrafts that are owned by Ornge, additional FW aircrafts are available with subcontractors. These can only be used if the medical condition of the patient permits it, and are called Standing Agreement (SA) aircrafts.

The calls are categorized based on the location of the patient, the urgency of the call, and the medical condition of the patient. Three urgency classes are distinguished: (1) non-urgent, (2) urgent, and (3) emergent. The level of care that a patient needs can be classified as either primary, advanced, or critical. A significant part of the patients does not originate at a location that is reachable by a regular aircraft. This part of the calls is called on scene and must be served by a RW aircraft. In case the patient is already at a health care facility, also a FW aircraft can be used.

The aim of this study is to evaluate and potentially improve the shift schedule at Ornge. For that, two main techniques are used: simulation and optimization. In the simulation, Ornge's operations are modeled in great detail in order to evaluate different scenarios. This enables us to gain insight into the impact of changes in the shift schedules or the dispatch policies. However, simulation is not suitable for the design of new schedules, and for that, we use the optimization model. In the optimization procedure, shift schedules are optimized in a highly

simplified representation of the system. The resulting schedules from the optimization are then evaluated in the simulation tool to assess the impact in a more realistic setting.

With the simulation tool that is described in Section 8.2, we will first evaluate the impact of different dispatch policies. Then, we will evaluate the impact of the shift change times in the current shift schedule. Currently, all shifts change around the same times, which results in undercapacity around that time. Finally, the additional value of each of the current shifts is measured by running the simulation after the removal of one or two shifts.

With the optimization, introduced in Section 8.4, we construct new schedules for the Ornge aircrafts. First, we fix the current distribution of the aircrafts over the bases. We limit the number of shifts to move away from the 24 hour, flat schedule. Despite significant fluctuation in call volumes, Ornge uses a shift schedule that has the same number of aircrafts at every time of the day. Then, we allow for a redistribution of the aircrafts over the given bases. Again, the number of shifts are restricted to obtain schedules with fluctuating capacity over the day. For all experiments with the optimization model, the simulation model is used to assess the different solutions.

8.2 Simulation model

In this section, we introduce the simulation model that is used to evaluate the different scenarios. In the simulation, we aim at a realistic representation of the response process at Ornge. As the goal is to optimize the shift schedule for Ornge's aircrafts, and these are only used for emergent and urgent calls, we exclude non-urgent calls. We use discrete event simulation where the events include: the start and end of a shift, the arrival of a call, the cancellation of a call, and the arrival of an aircraft at the scene, the hospital or its base. In the simulation, every time a call arrives or an aircraft becomes available, we try to assign an aircraft to a call. Depending on the type of call, different dispatch policies are followed. In all cases, the appropriate aircraft with the shortest time to definite care is selected. The access time to definite care is defined as the response time, plus the time spent on scene, plus the travel time to the hospital. Thus, this includes two flight legs: (1) from the base to the the scene, and (2) from the scene to the hospital. Note the difference with the response time, where only the first flight leg is included. Since the different aircrafts have different speeds, the aircraft with the shortest time to definite care does not necessarily coincide with the aircraft with shortest response time. Besides the time to definite care, also the response time will be one of the key performance indicators in this study. For every aircraft type, we have a maximum range for each flight. We cannot assign an aircraft to a call if that would imply a flight leg that exceeds the maximum range. Next, we describe how the calls of different types are handled, and what dispatch policies are implemented.

8.2.1 Call handling

We distinguish three main types of calls: (1) onscene calls, (2) interfacility calls, and (3) river hops. The onscene and interfacility calls are further characterized by the level of care the patient requires, which can be critical, advanced, or primary. Only for interfacility calls, this characterization influences the way the calls are handled. The urgency level of a call, which in the simulation can either be emergent or urgent, only influences the redirection and waitlist policy, which we discuss later.

Onscene calls

Onscene calls are calls that originate at a location different than an airport. Consequently, fixed wing aircrafts cannot respond to these calls, and only helicopters can be assigned. Since all different levels of care are treated in the same way, no distinction is made in the simulation. For these calls, the RW aircraft that would give the shortest time to definite care is selected. If no RW aircraft is available, the call is added to the waitlist.

Interfacility calls

For interfacility calls, the set of aircrafts that can serve the call depends on the level of care. If the patient requires critical care, only Ornge aircrafts (RW or FW) can be used. For advanced care patients, an Ornge aircraft is preferred. However, Standing Agreement aircrafts can serve as backup. Primary care calls are generally served by SA aircrafts, but the data shows that a significant fraction is in fact served by an Ornge aircraft. In the simulation, we assume that with a fixed probability, the call is served by an SA aircraft. All other calls must be served by an Ornge aircraft and are thus treated in the same way as critical care interfacility calls. For all calls, if an Ornge aircraft is available, the aircraft that results in the shortest time to definite care is selected. If no aircraft is available, critical and primary care are directly added to the waitlist. For advanced care calls, we assume that with a fixed probability, an SA aircraft is available as backup. In that case, the call leaves the system, otherwise the call enters the waitlist.

River hops

The helicopter located at the Moosonee base is often used to transport a patient from the hospital at one side of the river to the airport at the other side of the river. From the airport, the patient is then transported by a FW or SA aircraft. Even though these calls are not Ornge's main responsibility, they have a significant impact on the workload of the RW at the Moosonee base. In the simulation, these calls are treated different from the other calls. If a river hop request arrives, we check whether a RW is available. If a RW is available, it

is assigned to the call and spends a fixed time with the patient, after which it becomes available. If no RW is available, it is assumed that the call is handled differently. These calls are not included in the standard performance indicators, but the performance is reported separately.

8.2.2 Call cancellation

In the current execution, a significant part of the transportation request is eventually canceled. The cancellation can have a wide range of reasons, which we categorize in two main groups. First, we have calls that are canceled for reasons beyond Ornge's control. These means that regardless of Ornge's logistics, these calls will get canceled. For example, cancellation as a result of weather conditions and cancellation by the local EMS provider available at the scene are considered as beyond Ornge's control. On the other hand, there are calls that get canceled because Ornge has no aircraft available in time. The patient will be served by other means of transportation. These patients could have been served if Ornge had more available capacity or had used its capacity in a more efficient way. Ornge's main concern is to minimize the second category of unserved calls by better allocating the available capacity.

In the simulation, both types of cancellations are included. The second category is included in the same way as regular calls and the goal of this study is to increase the number of calls that can be served. Calls that are canceled beyond Ornge's control are included in a different way. As the calls will get canceled, one might be tempted to remove these calls from the data. However, these calls might have an impact on the workload of the aircrafts, as an aircraft might already be on its way at the moment of cancellation. Up to the moment of cancellation, the calls are treated as regular calls. When the call gets canceled, the aircraft becomes available for other calls at its current location. If the aircraft reaches the patient before the cancellation time, we assume that the call is canceled at the aircraft's arrival.

8.2.3 Waitlist policy

As discussed in the call handling section, calls for which no aircraft is available are added to a waitlist. Whenever an aircraft becomes available, either because of the start of a shift or the completion of a call, the waitlist is checked for calls that can be served by the newly available aircraft. If multiple calls can be served by the aircraft, the following dispatch rules are followed:

1. First serve calls of higher priority. In other words, emergent calls are served before urgent calls.
2. First-come, first-serve. For calls of the same priority, the calls that has the longest wait time is served first.

Note that other dispatch rules might result in better performance. One could, for example, serve the calls that is closest to the available aircraft to decrease the

total mileage, and consequently the workload. For each priority level, we define a maximum wait time, before the call gets canceled (under Ornge's control). This time is set to 1 hour for emergent calls and 12 hours for urgent calls.

8.2.4 Aircraft redirection

Aircrafts that are assigned to a call but did not reach the patient yet can be redirected to serve a call of higher priority. This means that urgent calls can be interrupted for emergent calls. Multiple policies for these redirections can be used. First, one could decide to always redirect an aircraft if this decreases the time to definite care for the high priority call. However, this might lead to an unnecessarily high deterioration of the performance of low priority calls. Alternatively, one could decide to only redirect if no other aircraft is available. In Section 8.3, we evaluate these two redirection policies. Furthermore, we evaluate the performance of the case where redirection is not allowed at all. Note that a mixture of the two policies is also possible, but is not considered in this study.

8.2.5 Shift changes

Currently, Ornge operates a 24 hour, flat schedule, where the crew works in shifts of 12 hours. At the end of a shift, an aircraft has to return to its base to change crew. Since Ornge does typically not have more aircrafts available, the new shift cannot start before the preceding shift has finished. For calls that arise close to the end of a shift, it is allowed for an aircraft to run in overtime. However, the total duration of a shift can never exceed 13 hours and 45 minutes. The maximum allowed overtime is thus 1 hour and 45 minutes. To avoid excessive overtime, a policy for handling the end of a shift has to be defined. In Section 8.3, two main policies are evaluated. In both cases, urgent calls are not allowed to incur overtime. In the first policy, we only allow for overtime for emergent onscene calls. Emergent interfacility calls are only assigned to calls if sufficient time is available for the aircraft to return to its base in time. In the second policy, also emergent interfacility calls can cause overtime. In case an aircraft which results in overtime is selected, we always assign the aircraft with minimum overtime. So, we do not allow for additional overtime in order to reduce the time to definite care.

8.3 Results simulation

In this section, we perform different experiments with the simulation model to evaluate the impact of certain decisions. We change some settings and compare the results with a fixed base case. In this base case, the current shift schedule is used and overtime is allowed for all emergent calls. Redirection of aircrafts to calls of higher priority is only allowed if no other aircraft is available. First, we describe

the data that is used in the experiments. Second, the impact of the overtime and redirection policy and the cancellation of calls is evaluated. Then, we experiment with changing the time at which the crew changes shifts. Currently, almost all shifts change at the same time, which might have a negative impact on the performance. Finally, the importance of each of the shifts is measured by running the simulation after removal of each of the shifts separately.

8.3.1 Data description

In the simulation, we use trace-driven simulation where the calls from January 2011 till June 2014 are used. In this period, over 93,000 calls were recorded. Approximately, 700 calls are excluded as a result of some data recording error. Another 10,000 calls are excluded as they are served by a land ambulance. Of the remaining calls, 2,003 are classified as river hops around the Moosonee base. Almost 25,000 calls are non-urgent and are not included in the simulation. 55,000 regular calls remain, of which 18,000 are onscene and 37,000 are interfacility calls. For the interfacility calls, we distinguish three different levels of care: (1) critical, (2) advanced, and (3) primary. For the primary care calls, an SA aircraft is preferred, but in 27 percent of the cases an Ornge aircrafts is requested as no SA aircraft is available. In the simulation, each primary care interfacility calls is included with a probability of 0.27. These calls are handled in the same way as critical care calls. Of the remaining interfacility calls, 48 percent have critical level of care and 52 percent requires advanced care.

In total, more than 11,000 calls are canceled beyond Ornge's control. In most cases, this involves onscene calls for which 47% is canceled. Almost all onscene calls are considered emergent, whereas for interfacility calls this is only the case for approximately 35 percent of the calls. Table 8.1 gives an overview of the calls that are included in the simulation. For the computation of the expected total number of calls, the number of primary care interfacility calls is multiplied by 0.27.

Table 8.1: *Summary of included calls.*

	Total	**Canceled**	**Emer.**	**Urg.**
Onscene	18,274	8,502	18,076	198
Critical interfacility	12,365	1,550	9,599	2,766
Advanced interfacility	14,091	1,369	9,902	4,189
Primary interfacility	10,652	11	4,714	5,938
Riverhop	2,003	0	2,003	0
Total	57,385	11,432	44,294	13,091
Expected total	49,609	11,424	40,853	8,756

Figure 8.1 gives the distribution of the included calls over the day. Despite the fact that this shows enormous fluctuation over the day, Ornge currently uses

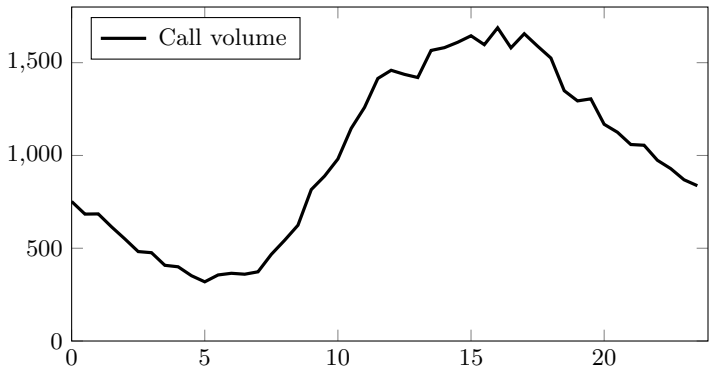


Fig. 8.1: *Call distribution over the day.*

a 24 hour, flat schedule. One part of this study focuses on the potential benefit of changing towards a schedule that better matches the available capacity with the demand.

Since the focus of this study is on the shift schedule of the Ornge aircrafts, we do not include the Standing Agreement aircrafts in the simulation. For the Ornge aircrafts, we consider the current shift schedule in the base case. Currently, Ornge deploys a 24 hour, flat schedule with eight rotor wing aircrafts and four fixed wing aircrafts distributed over nine different bases. Each aircraft is operated by two 12 hour shifts. All but two shifts change at 07:00 and 19:00. These two exceptions change at 06:00 and 18:00 and at 08:00 and 20:00, respectively. Figure 8.2 gives the geographical distribution of the aircrafts over the region. For each aircraft type, we have to specify the speed, the range, and the pre-flight preparation time. Rotor wing aircrafts travel with an average speed of 100 miles per hour (mph) and have a range of 150 miles. The pre-flight preparation time is 20 minutes. Fixed wing aircrafts are more suitable for longer distances as they have a range of 500 miles and an average speed of 150 mph. Their pre-flight preparation is slightly longer: 30 minutes. Note that the ranges hold for the different flight legs separately. There are guidelines for a maximum total distance for the two flight legs together, but these are regularly violated and not included in the simulation.

8.3.2 Results base case

First, we describe the results of our base case. In this scenario, overtime is allowed for all emergent calls and redirection of aircrafts of higher priority occurs only if no other aircraft is available for the high priority call. Table 8.2 shows that approximately 88% of all emergent onscene calls that are not canceled can be served. This corresponds to 47% of all emergent onscene calls, as a significant part of these calls is canceled. For interfacility calls, this is approximately 85%. Surprisingly, the coverage for urgent onscene calls is lower. Apparently, call volumes for emergent calls are too high to adequately serve urgent calls. For

Fig. 8.2: *Current distribution of the aircrafts over the region. Gray dots correspond to FW aircrafts, black dots to RW aircrafts.*

interfacility calls, we see that urgent calls have better service level than emergent calls. The average response time for onscene calls is less than 1 hour, which is not bad given the large distances and the relatively high start-up time of 20 minutes. For interfacility calls, the response time is higher, as distances involved are typically larger. In 73% of the cases is the rotor wing at the Moosonee base available when requested for a river hop.

Table 8.2: *Results of the base case with 95% confidence intervals. The first fraction of served calls is the fraction of the total calls that is served. In the second fraction, the canceled calls are not included. The response time and the time to definite care are in hours.*

Type of call	# Calls	% Served		Resp. time	Care time
Onscene Emer.	**18,076**	0.471 ± 0.030	0.883 ± 0.008	0.92 ± 0.01	1.72 ± 0.02
Onscene Urg.	**198**	0.535 ± 0.092	0.815 ± 0.080	3.03 ± 0.48	4.45 ± 0.50
Interfacility Emer.	**20,788**	0.760 ± 0.015	0.848 ± 0.011	1.33 ± 0.01	2.58 ± 0.02
Interfacility Urg.	**8,572**	0.848 ± 0.015	0.930 ± 0.009	3.18 ± 0.09	4.59 ± 0.10
Riverhop	**2,003**	0.732 ± 0.046	0.732 ± 0.046	-	-

When we consider the distribution of unserved calls over the day, we see a strong peak some time before the shift change times at 07:00 and 19:00 (see Figure 8.3). As all aircrafts have to return to their base around that time, limited capacity is available to serve calls. As a result of a waitlist time of 1 hour for emergent calls, this peak occurs slightly earlier. Section 8.3.4 evaluates the impact of changing the shift change times.

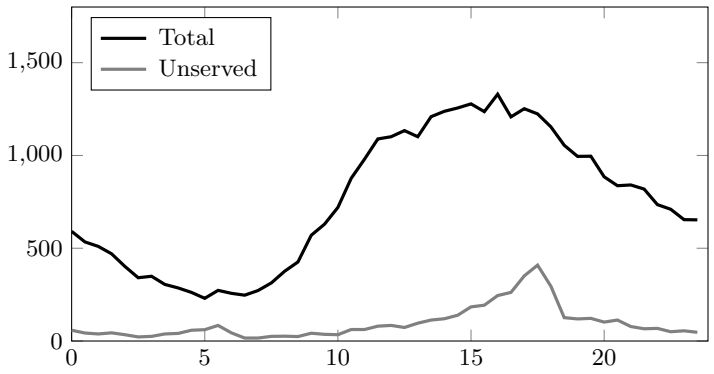


Fig. 8.3: *Number of unserved calls under Ornge's control per time of the day in the base case.*

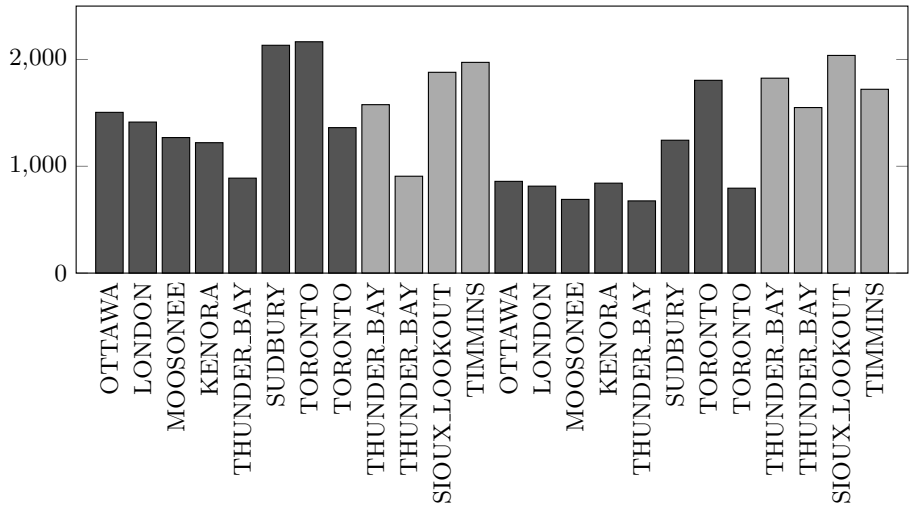


Fig. 8.4: *Number of calls per shift. The darker color corresponds to RW aircrafts, the lighter color to FW. On the left are the shifts starting in the morning, on the right the shifts that start in the evening.*

Figure 8.4 gives the number of calls executed by each of the shifts. It shows that the rotor wing shifts starting in the morning serve more calls than the shift starting in the evening. It further shows the importance of the helicopter at the Sudbury and the Toronto base. Section 8.3.5 further evaluates the importance of the different shifts by running the simulation after removal of one or two shifts.

8.3.3 Impact of implementation choices

Here, we evaluate the impact of some of the implementation choices. First, we experiment with different ways of handling the end of shifts. Then, different policies for the redirection of aircrafts are compared. Finally, the impact of call cancellations is measured by changing the cancellation times.

Overtime settings

In the base case, overtime is allowed in order to serve emergent calls of all types. By law, overtime can never exceed 1 hour and 45 minutes. We consider two alternative overtime policies: (1) no overtime is allowed at all, and (2) overtime is only allowed to serve emergent onscene calls. We measure the impact on the number of unserved calls, the response time, and the amount of overtime.

Table 8.3: *Results for different overtime policies. The second column gives the results for the case where no overtime is allowed. The third column overtime is allowed for emergent onscene calls. In the last column, all emergent calls can cause overtime. The first fraction of served calls is the fraction of the total calls that is served. In the second fraction, the canceled calls are not included.*

	No overtime	**Only onscene**	**All calls**
Onscene served	0.447 0.835	0.476 0.891	0.472 0.882
Interfacility served	0.720 0.800	0.718 0.798	0.786 0.873
Riverhop served	0.705	0.704	0.732
Resp. time onscene	0.94	0.92	0.92
Resp. time interfacility	1.31	1.31	1.33
Care time onscene	1.73	1.72	1.72
Care time interfacility	2.53	2.53	2.58
# Overtime RW	0	1,813	3,134
Total Overtime RW (h)	0	1,410	2,614
# Overtime FW	0	0	1,587
Total overtime FW (h)	0	0	1,425

Table 8.3 shows that by not allowing any overtime, significantly more calls remain unserved. The fraction of calls that can be served reduces by 4.7 and 7.3 percentage point, for onscene and interfacility calls, respectively. On the other hand, in the base case, a total of more than 4,500 shift run in overtime. That means that every day on average 3.5 of the 24 shifts run in overtime. The total overtime is a bit more than 3 hours per day. If only onscene calls can infer overtime, the average overtime is limited to 1 hour and 6 minutes per day. In

this case, an increase of the fraction of served onscene calls of 5.6 percentage point can be obtained. However, the service level of interfacility calls slightly decreases. The results show that allowing for some overtime can significantly increase the number of served calls, but it is for the decision maker to decide whether this improvement is worth the overtime.

Redirection of aircrafts

In the simulation, as in practice, it is allowed to redirect an aircraft assigned to a call of lower priority to serve a call of higher priority. An interesting question is when to make use of this possibility. On the one hand, the redirections lead to better service for the high priority calls. On the other hand, it reduces the service for low priority calls and it might even increase the total workload. In this experiment, we evaluate three redirection policies: (1) no redirections, (2) only redirect if no other aircraft is available, and (3) always redirect if the time to definite care can be reduced. Note that one can think of many more policies. For example, one could balance the last two policies or even include the impact on the low priority call in the decision.

Table 8.4: *Results for different redirection policies. In the first case, no aircrafts are redirected. In the second case, we only redirect an aircraft if no other aircraft is available. In the last case, an aircraft is redirected to a call of higher priority if the time to definite care of that call can be reduced.*

	No redirection	**If only option**	**Always**
Onscene served	0.465 0.870	0.472 0.882	0.472 0.883
Interfacility served	0.775 0.860	0.786 0.873	0.784 0.871
Riverhop served	0.732	0.732	0.733
Resp. time onscene (Emer.)	0.92	0.92	0.92
Resp. time interf. (Emer.)	1.34	1.33	1.31
Care time onscene (Emer.)	1.72	1.72	1.72
Care time interf. (Emer.)	2.58	2.58	2.55
Resp. time onscene (Urg.)	2.58	3.03	3.00
Resp. time interf. (Urg.)	2.75	3.18	3.33
Care time onscene (Urg.)	3.99	4.45	4.42
Care time interf. (Urg.)	4.17	4.59	4.74
# Redirections	0	1,470	2,156

Table 8.4 shows that redirection of aircrafts has an important impact on the number of calls that can be served. By allowing redirection, an increase of 1.2 and 1.3 percentage point can be obtained for onscene and interfacility calls,

respectively. The redirection can also have a slight positive effect no the response time and time to definite care for emergent calls. However, these times increase for urgent calls as a result of redirections. The difference between the last two redirection policies is limited. To limit the number of redirections, we recommend the policy where we only redirect an aircraft if no other aircraft is available.

Call cancellation

In Ornge's operations, it frequently occurs that calls get canceled. As an aircraft might already be assigned to the call, this can have an impact on the workload and consequently the performance of the system. In this section, we evaluate the impact of these calls. In the base case, we assume that the notification of cancellation is made 30 minutes after the call is requested. However, no reliable data is available regarding the cancellation times, so we consider multiple alternatives. We include two cases where this time is increased to 1 hour and 90 minutes. Additionally, we consider the case where all canceled calls are removed from the data.

Table 8.5: *Results for different cancellation times for canceled calls.*

	Cancellation			
	Base case	**Removed**	**30 min later**	**60 min later**
Onscene served	0.472	0.474	0.463	0.461
	0.882	0.886	0.866	0.862
Interfacility served	0.786	0.787	0.783	0.782
	0.873	0.874	0.869	0.869
Riverhop served	0.732	0.733	0.731	0.731

Table 8.5 shows the number of served calls in each of the scenarios. We see that in the base case, the impact of the cancellation is limited. If all canceled calls are removed, only slightly more calls can be served. However, if the time to cancellation is increased, a significant increase in unserved calls is observed. The number of unserved onscene calls can increase up to 2 percentage point.

8.3.4 Impact of shift change times

As we have seen in the analysis of the base case, there is a strong peak in the number of unserved calls around the time the shifts change. To verify that this peak is indeed caused by the shift changes, we run the simulation with a shift schedule with only 24 hour shifts. All these shifts change at 07:00. Additionally, we consider scenarios where all shifts change at a busy time of the day, 12:00 and 00:00. We expect that this worsens the effect, and leads to worse performance. Finally, we consider a potentially good schedule, where all shifts change at a

Table 8.6: *Results for different shift change times.*

	Shift changes			
	Current	**All at 12**	**All different**	**Only 7 am**
Onscene served	0.472 0.882	0.469 0.878	0.472 0.883	0.483 0.903
Interfacility served	0.786 0.873	0.786 0.873	0.807 0.896	0.824 0.915
Riverhop served	0.732	0.721	0.720	0.739
# Overtime RW	3,134	3,254	3,233	754
Total overtime RW (h)	2,614	2,640	2,640	607
# Overtime FW	1,587	1,349	1,819	390
Total overtime FW (h)	1,425	1,223	1,676	338

different time. In this way, the capacity reduction as a result of shift changes is equally distributed over the day.

Table 8.6 gives the results of the different schedules, and Figure 8.5 gives the distribution of the unserved calls over the day. We see that the peak in the evening disappears if we consider 24 hour shifts. This clearly indicates that the peak is caused by the shift changes. This is confirmed by the bad performance of the schedule with shift changes at 12:00 and 00:00 and the good performance of the schedule with equally distributed shift change times. For interfacility calls, the fraction of served calls be increased by 2.3 percentage point compared to the base case.

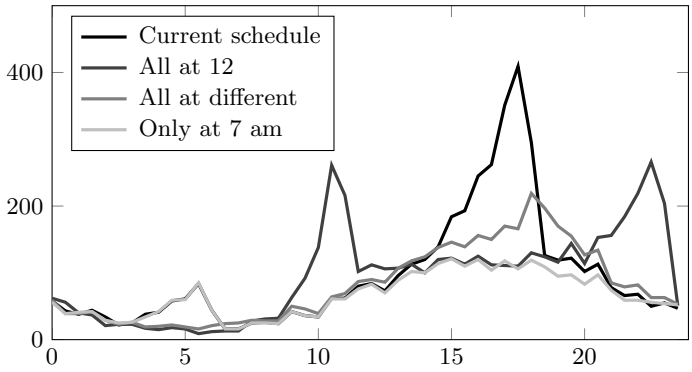


Fig. 8.5: *Number of unserved calls under Ornge's control per time of the day for different shift change times.*

8.3.5 Impact of shift removal

To gain further insight into the importance of a particular shift, we measure the impact of removing a shift from the schedule. In the analysis of the base case, we have already seen the number of calls served by each shift. However, this does not tell whether another shift was available to serve the call. Table 8.7 gives the number of additionally unserved calls if one particular shift is removed. The shifts are ordered by the number of additional unserved calls, not including the river hops.

Table 8.7: *Number of additional unserved calls after removal of one shift.*

Shifts			**Additional unserved calls**			
Base	**Time**	**Type**	**Onscene**	**Interfacility**	**Total**	**Riverhop**
MOOSONEE	**19:00**	**RW**	18	29	47	452
MOOSONEE	**07:00**	**RW**	26	75	101	1,014
TORONTO	**19:00**	**RW**	40	68	108	3
LONDON	**19:00**	**RW**	59	108	167	3
TORONTO	**06:00**	**RW**	90	88	178	3
TORONTO	**18:00**	**RW**	87	111	198	1
THUNDER_BAY	**07:00**	**FW**	22	275	297	1
THUNDER_BAY	**07:00**	**FW**	23	275	298	1
TORONTO	**07:00**	**RW**	150	180	330	3
OTTAWA	**19:00**	**RW**	122	235	357	2
LONDON	**07:00**	**RW**	136	266	402	5
THUNDER_BAY	**19:00**	**FW**	12	444	456	3
THUNDER_BAY	**19:00**	**RW**	461	20	481	0
THUNDER_BAY	**19:00**	**FW**	16	466	482	3
SIOUX_LOOKOUT	**07:00**	**FW**	26	459	485	1
SIOUX_LOOKOUT	**19:00**	**FW**	15	573	588	1
TIMMINS	**19:00**	**FW**	31	574	605	8
TIMMINS	**07:00**	**FW**	71	536	607	17
KENORA	**20:00**	**RW**	610	33	643	0
SUDBURY	**19:00**	**RW**	583	74	657	0
THUNDER_BAY	**07:00**	**RW**	741	8	749	1
OTTAWA	**07:00**	**RW**	459	347	806	2
KENORA	**08:00**	**RW**	936	44	980	0
SUDBURY	**07:00**	**RW**	1,217	145	1,362	2

The table shows that the removal of the helicopter at the Moosonee base has the lowest impact on onscene and interfacility calls. However, this shift is often used for the river hops. Despite the large number of calls that is served by the Toronto rotor wing aircrafts, we see that removing one of the two does not have a very high impact. Since there are two helicopters located at this base and there are some other helicopters nearby, the removal of one shift only slightly increases the number of unserved calls. For the RW at Sudbury, we do see that serving

many calls can be an indication of high importance. Apparently, this helicopter is the only available one for many onscene calls. For the FW aircrafts, we see that the second aircraft at the Thunder Bay has the lowest value. For RW aircrafts, we see that the day shifts are more important than the night shifts. This is due to the fact that especially for onscene calls, more calls occur during daytime. Table 8.10 in Appendix 8.B gives similar results for the case where two shifts are removed simultaneously.

8.4 Optimization model

In this section, we present the optimization model that is used to develop new shift schedules for the Ornge aircrafts. The main purpose of the models is to find good schedules for cases where we move away from a 24 hour, flat schedule. To that end, the model not only distributes aircrafts over the selected bases, but also assigns crew to the aircrafts. As the number of shifts might not suffice to staff all aircrafts 24 hours a day, the model decides which aircrafts to equip with crew at what time. To capture the difference between onscene and interfacility calls and between rotor wing and fixed wing aircrafts, the model includes two types of calls and two types of aircrafts. The model is based on the version of MEXCLP with two vehicle types introduced by Chong et al. (2015). For onscene calls, the expected coverage is determined in the same way as in the original MEXCLP by Daskin (1983). Here, only the availability of RW aircrafts is taken into account. For interfacility calls, both RW and FW aircrafts are included. Given a busy fraction q_1 for RW and q_2 for FW aircrafts, the coverage provided by a RW aircrafts and b FW aircrafts is equal to

$$\text{cov}(a, b) = 1 - q_1^a \times q_2^b. \tag{8.1}$$

As in most ambulance location models, we have a set I of potential base locations and a set J of demand points. In the experiments, we set I to be equal to the current set of bases. In this case, the model searches for the best distribution of aircrafts and crew over the current bases. To compute the coverage given a shift schedule, we introduce the sets I_{jRW} and I_{jFW} that contain all base locations that are within the range of the corresponding aircraft type from demand point j. For both types of calls, we have a given demand in each time period, denoted by d_{jt1} and d_{jt2} for onscene and interfacility calls, respectively. Here t is an element of the set of time periods T. To denote the set of time periods at which a 12 shift can start to be available during time period t, we introduce the set T_t. For example, if $T = \{1, 2, ..., 24\}$, then $T_{15} = \{4, 5, ..., 15\}$ and $T_8 = \{21, 22, 23, 24, 1, 2, ..., 8\}$. Let $V = \{RW, FW\}$ be the set of aircraft types.

In the model, we distinguish aircrafts and shifts. All shifts are assumed to have a duration of 12 hours. Each shift requires its own aircraft, except for the case of two consecutive shifts at the same base. The two shifts together form a

24 hour shift, for which only one aircraft is required. The number of shifts of each aircraft type is limited by s_{RW} and s_{FW} for fixed wing and rotor wing, respectively. Similarly, the number of aircrafts is limited by p_{RW} and p_{FW}.

In the model, we have two main sets of variables to denote the shift schedule and the corresponding coverage. The integer variables x_{iRW} and x_{iFW} give the number of 24 hour shifts at base i for RW and FW aircrafts, respectively. Similarly, x_{itRW} and x_{itFW} denote the number of 12 hour shifts starting at time t of the different aircraft types. For the coverage of onscene calls, we have a set of binary variables y_{jta} indicating whether demand point j is covered by exactly a RW aircrafts during time period t. Similarly, y_{jtab} indicates whether j is covered by exactly a RW and exactly b FW aircrafts. Note the difference with the other chapters, where y typically indicates that j is covered by at least a given number of ambulances. With this, we can formulate the model.

$$\max \quad \sum_{t\in T}\sum_{j\in J} d_{jt1}\sum_{a=0}^{p_1} y_{jta}\ \mathrm{cov}(a,0) + d_{jt2}\sum_{a=0}^{p_1}\sum_{b=0}^{p_2} y_{jtab}\ \mathrm{cov}(a,b)$$

$$\text{s.t.} \quad \sum_{a=0}^{p_1} a y_{jta} \le \sum_{i\in I_{jRW}} \left(x_{iRW} + \sum_{t'\in T_t} x_{it'RW} \right) \qquad \forall j \in J, t \in T \tag{8.2}$$

$$\sum_{a=0}^{p_1} a y_{jtab} \le \sum_{i\in I_{jFW}} \left(x_{iFW} + \sum_{t'\in T_t} x_{it'FW} \right) \qquad \forall j \in J, t \in T, b \le p_2 \tag{8.3}$$

$$\sum_{b=0}^{p_2} b y_{jtab} \le \sum_{i\in I_{jFW}} \left(x_{iFW} + \sum_{t'\in T_t} x_{it'FW} \right) \qquad \forall j \in J, t \in T, a \le p_1 \tag{8.4}$$

$$\sum_{a=0}^{p_1} y_{jta} = 1 \qquad \forall j \in J, \forall t \in T \tag{8.5}$$

$$\sum_{a=0}^{p_1}\sum_{b=0}^{p_2} y_{jtab} = 1 \qquad \forall j \in J, \forall t \in T \tag{8.6}$$

$$\sum_{i\in I} 2x_{iv} + \sum_{t\in T} x_{itv} \le s_v \qquad \forall v \in V \tag{8.7}$$

$$\sum_{i\in I} x_{iv} + \sum_{t\in T} x_{itv} \le p_v \qquad \forall v \in V \tag{8.8}$$

$$x_{iRW}, x_{itRW}, x_{iFW}, x_{itFW} \in \mathbb{N} \qquad \forall i, \in I, t \in T \tag{8.9}$$

$$y_{jta}, y_{jtab} \in \{0,1\} \qquad \forall i \in I, j \in J,\ a \le p_1, b \le p_2 \tag{8.10}$$

The objective of the model is to maximize the coverage provided to onscene and interfacility calls. Constraints 8.2-8.4 ensure that the coverage level depicted by

the y-variables is obtained by the shift schedule represented by the x-variables. Constraints 8.5 and 8.6 state that j can only be covered by exactly one number of aircrafts. Constraints 8.7 limit the number of shifts of each type. Here, 24 hour shifts correspond to two shifts. Similarly, Contraints 8.8 restrict the number of aircrafts.

8.5 Results optimization

In this section, the optimization model presented in Section 8.4 is applied to find better shift schedules for the Ornge aircrafts. We are particularly interested in schedules other than the current 24 hour, flat schedule. By limiting the number of shifts, we obtain schedules with fluctuating capacity over the day. First, we describe the data that is used in the optimization. Then, we use the model to find schedules given the current distribution of the aircrafts. Here, we find which combination of two shifts of 12 hours can best be replaced by one 12 hour shift and what the start time for this shift should be. Finally, we allow for a redistribution of the aircrafts over the given bases. Again, we restrict the number of shifts. In both experiments, the resulting shift schedules are evaluated in the simulation model.

8.5.1 Data description

As we consider alternative shift schedules rather than alternative base locations, we take I to be equal to the current set of nine bases. For the set of demand points J, we use a 10 mile by 10 mile grid for the region of Ontario. We exclude all grid cells for which no calls are recorded in the period from January 2011 till June 2014. A total of 1,029 grid cells remains. The model requires the demand for each demand point, each time period, and each call type as an input. For this, we take all calls that are included in the simulation, including canceled calls and primary care interfacility calls. Each call is assigned to a demand point based on the origin location of the patient. The time-dependent distribution of the demand is included by multiplying the total demand from a particular demand point by the fraction of calls in that time period. In this way, we avoid having a lot of pairs of demand points and time periods with zero demand. As the total number of calls is artificially increased by also including the calls that can potentially be served by an SA aircraft, we risk too high focus on interfacility calls. To overcome this, we normalize the total number of onscene calls and the total number of interfacility calls to 1. In this way, we have an equal focus on both onscene and interfacility coverage.

As a limit on the number of aircrafts, we take the current availability, which is equal to 8 rotor wing aircrafts and 4 fixed wing aircrafts. In the current schedule, each aircraft has two shifts of 12 hours, which gives a total number of shifts of 16 and 8, respectively. In the experiments, we reduce this number to evaluate the impact of moving away from the 24 hour, flat schedule. In the optimization,

we use a smaller range for the aircrafts than in the simulation to avoid many very long-distance flights. Additionally, this avoids high workloads as a result of these long distances. We use 100 miles for RW and 300 miles for FW. The busy fractions are obtained from the base case of the simulation. This gives busy fractions of 0.15 for RW and 0.35 for FW. The pre-flight delay and the travel speed are the same as in the simulation.

As the model does not specify the shift change time of the crew on an aircraft that is staffed 24 hours a day, but this does influence the results of the simulation, we have to specify this shift change time. In order to follow the current practice, we let these 24 hour shifts change at 07:00 and 19:00. The only exception to this rule is that if two aircrafts of the same type are located at the same base, we let the second shift change at 08:00 and 20:00.

8.5.2 Limiting the number of shifts

In the first experiment, we analyze the impact of reducing the number of shifts. The aircrafts remain located at their current base location. We consider the case where up to two shifts are removed and compare the results of the simulation with the current situation with 16 RW shifts and 8 FW shifts. To avoid many unserved river hops, we do not allow the removal of one of the shifts at the Moosonee base. Note that the performance of the different schedules is evaluated in the simulation model.

Table 8.8: *Results for different number of RW and FW shifts. Aircrafts remain located at their current base.*

	16 | 8	**16 | 7**	**16 | 6**	**15 | 8**	**14 | 8**	**15 | 7**
Onscene served	0.472	0.471	0.470	0.471	0.451	0.471
	0.883	0.881	0.879	0.881	0.843	0.880
Interfacility served	0.790	0.786	0.772	0.790	0.789	0.785
	0.878	0.873	0.857	0.877	0.876	0.872
Riverhop served	0.732	0.731	0.731	0.732	0.731	0.731
Resp. time onscene	0.92	0.92	0.92	0.93	0.91	0.93
Resp. time interfacility	1.33	1.35	1.35	1.35	1.35	1.36
Care time onscene	1.72	1.73	1.73	1.73	1.70	1.73
Care time interfacility	2.58	2.60	2.60	2.60	2.60	2.62

Table 8.8 gives the results for the different scenarios. We see that by removing up to one shift of each type, only a slight decrease in the number of served calls is observed. For removing one FW shift, this reduction is 0.2 percentage point for onscene calls and 0.5 percentage point for interfacility calls. In that case, two of the four shifts at the Thunder Bay base are replaced by one shift starting at 10:00. Removing one RW shift yields a decrease of 0.2 and 0.1 percentage point, respectively. Here, one shift is removed from the Toronto base. Removing

these two shifts simultaneously gives a reduction of 0.3 and 0.6 percentage point, respectively. Figure 8.6 gives the corresponding shift schedule.

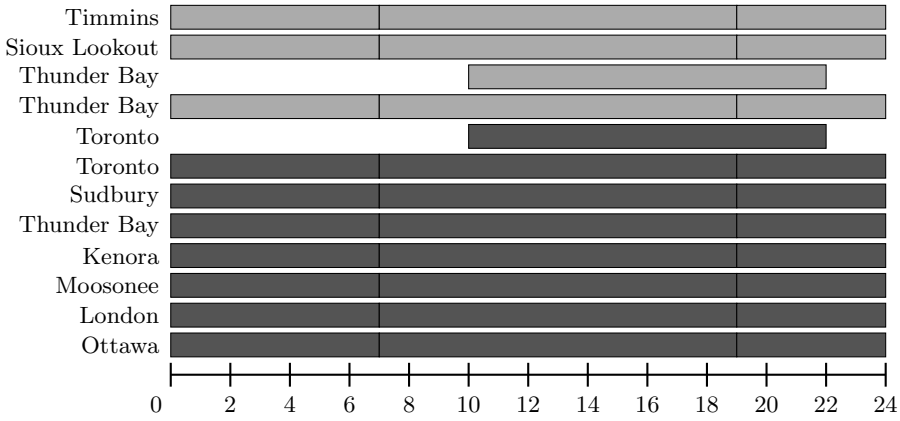


Fig. 8.6: *Shift schedule after removal of one FW shift and one RW shift. Aircrafts remain located at their current base location. The darker color corresponds to RW aircrafts, the lighter color corresponds to FW aircrafts.*

8.5.3 Aircraft redistribution

Besides limiting the number of shifts, it is also interesting to see how the schedule would change if we allow the redistribution of aircrafts. Here, we use the current bases as potential locations for the aircrafts and let the model decide how to distribute the aircrafts. The number of aircrafts of each type is not changed. Again, we do the computations for a different number of shifts.

Table 8.9 gives the performance of the obtained schedules. As in the previous experiment, we see that removing up to one shift of each type has only limited impact. Removing two shifts, however, leads to significantly more unserved calls. Figure 8.7 gives the shift schedule corresponding to the schedule with 15 RW shifts and 7 FW shifts. In this schedule, the second RW at Toronto is moved to Sudbury. In both Figure 8.4 and Table 8.7, we have already seen the importance of the Sudbury base. For the FW aircrafts, both aircrafts at the Thunder Bay base are moved to Sudbury and the aircraft at Timmins is moved to Sioux Lookout.

Comparing the results from Table 8.8 and Table 8.9, we see that by redistributing the aircrafts, we can obtain an increase in served calls of about two percentage point for onscene calls and 0.5 percentage point for interfacility calls. Additionally, the response time and time to definite care for interfacility calls is reduced.

Table 8.9: *Results for different number of RW and FW shifts. Aircrafts can freely be distributed over the current bases.*

	16 8	**16 7**	**16 6**	**15 8**	**14 8**	**15 7**
Onscene served	0.483	0.483	0.482	0.480	0.460	0.480
	0.903	0.904	0.902	0.898	0.861	0.897
Interfacility served	0.795	0.790	0.768	0.795	0.795	0.790
	0.883	0.878	0.853	0.883	0.883	0.878
Riverhop served	0.721	0.721	0.721	0.721	0.721	0.721
Resp. time onscene	0.92	0.92	0.92	0.92	0.91	0.92
Resp. time interfacility	1.28	1.30	1.31	1.28	1.28	1.30
Care time onscene	1.72	1.72	1.72	1.72	1.69	1.72
Care time interfacility	2.52	2.54	2.54	2.52	2.52	2.54

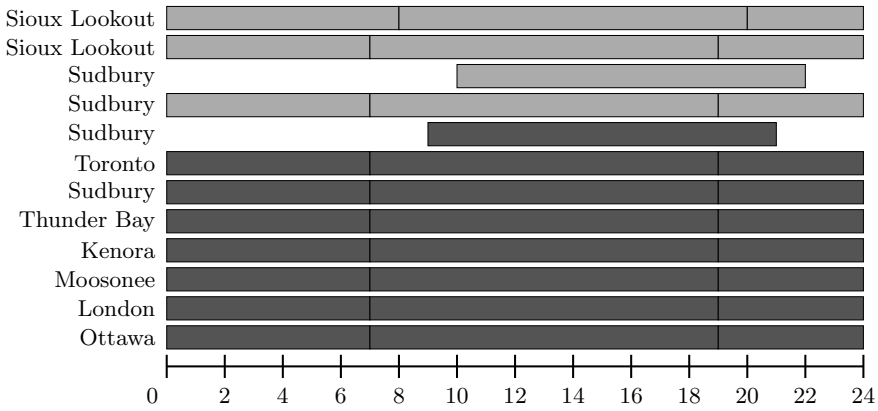


Fig. 8.7: *Shift schedule after removal of one FW shift and one RW shift, and the redistribution of aircrafts over the current bases. The darker color corresponds to RW aircrafts, the lighter color corresponds to FW aircrafts.*

8.6 Conclusion

In this chapter, we used both simulation and optimization techniques to evaluate and improve the current shift schedule at an air ambulance provider in Ontario, Canada. The simulation model gives a detailed description of the response process and allows for the evaluation of changes in the dispatch policy. Additionally, the simulation is used to indicate the importance of each of the current shifts. The introduced optimization model generates shift schedules that might result in better coverage. The model uses the expected coverage as defined in Daskin (1983) and extends that to include multiple vehicle types in a way similar to Chong et al. (2015). The shift schedules obtained with the optimization model

are evaluated in the simulation model to get more realistic insights into the change in performance.

The experiments with the simulation model show that the overtime and aircraft redirection policies have an impact on the number of served calls. By allowing shifts to run in overtime for emergent calls, up to 5 percent more onscene and 7 percent more interfacility calls can be served. The redirection of aircrafts assigned to calls of lower priority to calls of high priority can lead to 1 percentage point more served calls. The results of the simulation for different shift change times clearly show that these shift change times matter. In the current shift schedule, almost all shifts change at 07:00 and 19:00, which leads to a significant peak in unserved calls around those times. By equally distributing the shift changes over the day, significantly more calls can be served. Finally, the simulation model was applied after the removal of one shift, to indicate the importance of the different shifts.

The results of the optimization model are used to evaluate the impact of moving away from a 24 hour, flat schedule. As long as the number of removed shifts is limited to one of each type, only a minor decrease in performance is observed. However, if more shifts are removed, we start to see a stronger decrease in performance. When we allow for shift schedules in which the aircraft distribution can be changed, we can obtain a 2 percentage point increase in serving onscene calls. A smaller increase for interfacility calls is obtained as well.

For future research, we would like to use the optimization model to combine the selection of base locations, the distribution of aircrafts, and the generation of shift schedules. Currently, only the last two are included in the experiments. Additionally, we would like to include the observed coverage decrease around shift change times into the model. In that way, we can potentially find even better shift schedules.

8.A Current shift schedule

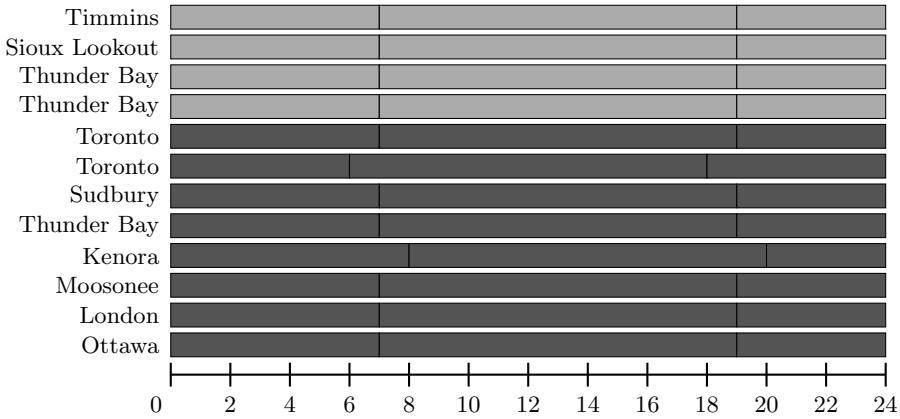


Fig. 8.8: *Current shift schedule. The darker color corresponds to RW aircrafts, the lighter color corresponds to FW aircrafts.*

8.B Impact of removal of two shifts

Table 8.10: *Number of additional unserved calls after removal of two shifts.*

Shift 1			Shift 2			Additional unserved calls			
Base	**Time**	**Type**	**Base**	**Time**	**Type**	**Onscene**	**Interfacility**	**Total**	**Riverhop**
TORONTO	**06:00**	**RW**	**TORONTO**	**19:00**	**RW**	143	169	312	6
TORONTO	**06:00**	**RW**	**LONDON**	**19:00**	**RW**	149	206	355	5
THUNDER_BAY	**07:00**	**FW**	**TORONTO**	**19:00**	**RW**	60	346	406	3
THUNDER_BAY	**07:00**	**FW**	**TORONTO**	**19:00**	**RW**	62	346	408	3
LONDON	**19:00**	**RW**	**TORONTO**	**19:00**	**RW**	169	271	440	4
THUNDER_BAY	**07:00**	**FW**	**LONDON**	**19:00**	**RW**	79	388	467	4
THUNDER_BAY	**07:00**	**FW**	**LONDON**	**19:00**	**RW**	80	391	471	4
TORONTO	**06:00**	**RW**	**THUNDER_BAY**	**07:00**	**FW**	113	365	478	5
TORONTO	**06:00**	**RW**	**THUNDER_BAY**	**07:00**	**FW**	114	367	481	5
TORONTO	**07:00**	**RW**	**LONDON**	**19:00**	**RW**	198	291	489	6
SUDBURY	**07:00**	**RW**	**THUNDER_BAY**	**19:00**	**RW**	1,679	167	1,846	2
SUDBURY	**07:00**	**RW**	**SIOUX_LOOKOUT**	**07:00**	**FW**	1,234	644	1,878	3
SUDBURY	**07:00**	**RW**	**TIMMINS**	**19:00**	**FW**	1,239	716	1,955	10
SUDBURY	**07:00**	**RW**	**SIOUX_LOOKOUT**	**19:00**	**FW**	1,227	729	1,956	4
SUDBURY	**07:00**	**RW**	**KENORA**	**20:00**	**RW**	1,827	182	2,009	2
SUDBURY	**07:00**	**RW**	**TIMMINS**	**07:00**	**FW**	1,236	818	2,054	18
THUNDER_BAY	**07:00**	**RW**	**SUDBURY**	**07:00**	**RW**	1,958	160	2,118	3
OTTAWA	**07:00**	**RW**	**SUDBURY**	**07:00**	**RW**	1,684	504	2,188	4
SUDBURY	**07:00**	**RW**	**SUDBURY**	**19:00**	**RW**	1,961	245	2,206	2
KENORA	**08:00**	**RW**	**SUDBURY**	**07:00**	**RW**	2,155	192	2,347	2

9

Application of MCLP to the Norwegian air ambulance

9.1 Introduction

In this thesis, we have developed multiple new models to improve the performance of EMS providers. We applied the models to different regions in the Netherlands. Furthermore, in Chapter 5, we developed a model for the fire department Amsterdam-Amstelland. Here, many new features had to be included to make the model applicable to firefighter systems. Chapter 8 showed that similar techniques can be applied to air ambulance services. Here, significant modeling effort was required to obtain an appropriate model. The expertise gained with these projects led to a project together with the Norwegian air ambulance provider. Their aim was to improve coverage provided by their medical helicopters in the long-stretched country of Norway. This chapter discusses the application of a basic ambulance location model from the literature and shows how mathematically very simple models can have an immediate impact. This research was initiated by nation-wide consternation as a result of a study showing that large parts were not covered within the targets set by law.

Norway is a long-stretched country with a wide-spread rural population. Despite large geographical distances and substantial uninhabited areas, the government requirements state that 90 percent of the population should be reached by a physician manned ambulance service within 45 minutes (NMoHaC Services, 2000). An effective Helicopter Emergency Medical Service (HEMS) is considered essential in order to achieve the desired equality in health care and the objective of the Norwegian air ambulance service to provide advanced emergency medicine to critically ill or severely injured patients.

In order to ensure optimal coverage, and homogeneity in health care throughout, the location of the air ambulance bases is important. Currently, there are 12 helicopter bases in Norway providing HEMS. The bases have been established from the late 1970s, at geographical locations that at the time led to a significant coverage improvement.

In this study, we explore the mathematically optimal locations of helicopter ambulance bases using the Maximal Covering Location Problem (MCLP), intro-

duced by Church and ReVelle (1974) and given in Chapter 2 of this thesis. Using fine detail population density data for the whole of Norway, we fit MCLP to explore optimal base structures given different parameter values. We perform both a greenfield analysis, assuming no existing bases, and optimization conditioned on the current bases, in order to explore whether improvements to the existing base structure could be achieved by moving a limited number of bases.

9.2 Data and model description

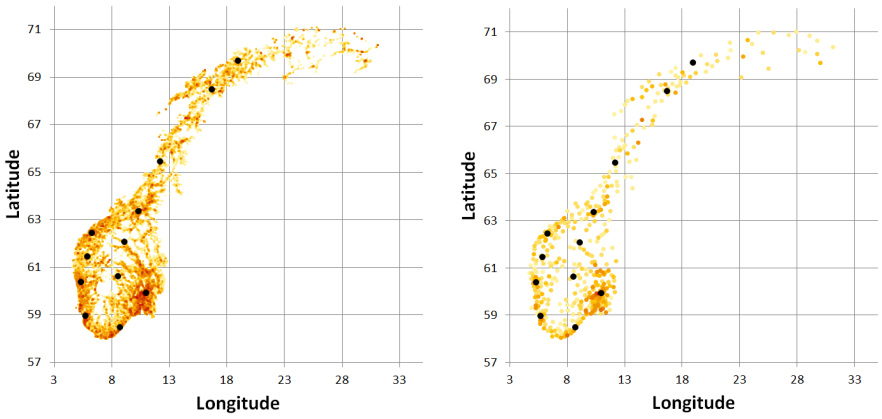
Mainland Norway covers 323,780 km^2 at the far North of Europe, stretching 1,790 kilometer from North to South. The country has a mainly rural population with an average population density of 16.1 inh/km^2, ranging from 1129.5 inh/km^2 in Oslo to 1.5 inh/km^2 in Finnmark. On January 1st, 2015, the population of Norway was 5.2 million (Statistics Norway, 2015). Official population statistics on a fine grid with cells of dimension 1km x 1km are freely available from Statistics Norway (2015). This gives fine detail information on the population density of Norway. In 2015, only 55,213 (10.3%) of the grid cells were inhabited. Median (IQR) inhabitants for the inhabited grid cells was 13 (5-36).

Official statistics are often collected and reported on municipality level. In order to explore whether this coarser information would lead to estimation bias or otherwise essential loss of precision, we also perform the analysis on this coarser data set. In 2015, Norway consisted of 428 municipalities. For each municipality, there is a population weighted centroid representing the population center of the municipality. The 428 municipalities have a median (IQR) of 4,697 (2,180-10,654) inhabitants. Figure 9.1 shows the geographical distribution of the population on both the fine grid and the municipality level.

The average pre-flight preparation time, or pre-trip delay, of HEMS in Norway is 5.5 minutes (Zakariassen et al., 2015), and the helicopter ground speed is about 220 km/h. For an air ambulance helicopter flying at this speed, a 1km x 1km grid cell is thus crossed in 15-20 seconds.

The optimal base locations for the emergency helicopters are determined by the Maximal Covering Location Problem (Church and ReVelle, 1974). The model maximizes the number of inhabitants that is covered by at least one helicopter within a pre-specified target response time. Given a fixed number of helicopters, the single coverage is maximized. By applying the model for a different number of bases, the least number of bases to cover a given fraction of the population can be computed.

As MCLP uses the concept of single coverage, it assumes that a helicopter is always available. Clearly, this is not realistic, however, in some sense, it represents a best-case scenario. If an area cannot be covered within the response target from any of the bases, it never can be reached in time. As the purpose of this study is to evaluate and potentially improve the fair access of care throughout the country, MCLP can give useful insights.



(a) *Fine grid* **(b)** *Municipality*

Fig. 9.1: *Geographical distribution of the population on both fine grid and municipality level.*

In the fine grid analysis, we use the 55,213 inhabited grid cells as demand locations. In order to keep computations tractable, we use a smaller, but still relatively large, set of potential base locations. For this, we use a coarser 10km x 10km grid. Grid cells with zero population are included as well, since uninhabited locations surrounded by several densely populated areas can still be good locations. The travel times between base locations and demand points, including a 5.5 minute pre-trip delay, are computed using a fixed speed of 220 km/h.

To explore the practical consequences of various target times, we calculate the number of bases needed to cover various percentages of the population for target response times of 30 and 45 minutes. We first compute the optimal base locations assuming no current bases existed, so-called greenfield analysis. This yields the truly optimal base locations, given the chosen model. Such an analysis is, however, not practically feasible, as this would imply tearing down all existing bases and start building anew. We thus also compute the optimal location of new helicopter bases. Given the existing 12 bases in Norway, we compute the possible gain of moving one or two bases.

In order to quantify potential information loss by using municipality level data, we compare the solutions of the fine grid data with the solutions of the municipality level data. The models are implemented in Java and solved with IBM ILOG CPLEX Optimization Studio 12.6 (ILOG, 2013).

9.3 Results

First, we evaluate the minimum number of optimally located bases that is required for a given level of coverage. This number is computed for a coverage

level of 90%, 95%, and 100% with a target response time of 30 minutes and 45 minutes. For this experiment, we use the fine grid data, a pre-trip delay of 5.5 minutes, and an average flight speed of 220 km/h. For the 45 minute threshold, which corresponds to the official target, we see that only four bases are required to reach a coverage of 90%. For a coverage of 95%, five bases are required and complete coverage can be obtained by nine bases. Note that the current number of bases is 12. When a stricter 30 minute target is used, eight, ten, or 21 bases are required to ensure coverage levels of 90%, 95%, and 100%, respectively. Figure 9.3 shows the geographical distribution of the bases and the coverage in the considered cases.

As official statistics are often reported at the coarser municipality level, we perform the calculation at this level as well. To evaluate the coverage loss as a result of the data aggregation, we compute the optimal locations according to MCLP with both demand points and potential base locations at municipality level. The coverage provided by the set of bases is evaluated at the fine grid. This coverage is compared to the coverage provided by the solution of the fine grid data. Note that the two cases use a different set of potential locations. On the fine grid, we use the 4,218 10km x 10km grid cells, whereas on the municipality level, we use the 428 population weighted centroids. Since these centroids are not a subset of the centroids of the 10km x 10km grid cells, it can occur that the solution of the municipality level gives a higher coverage. Figure 9.2 gives the coverage for different number of bases and different response time targets. We see that the fine grid solution only slightly outperforms the municipality solution, which indicates that the municipality data already gives a reasonable representation of the system for this model. However, if computation times allows for the computation of the fine grid solution, this is of course preferred.

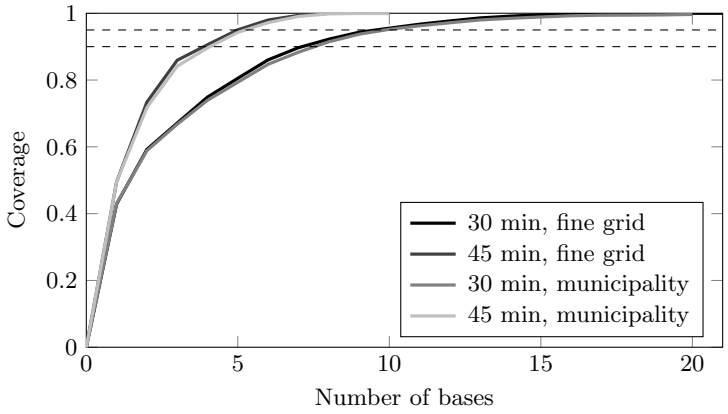
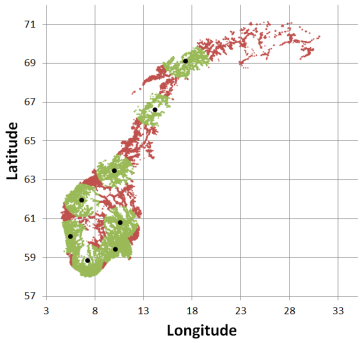
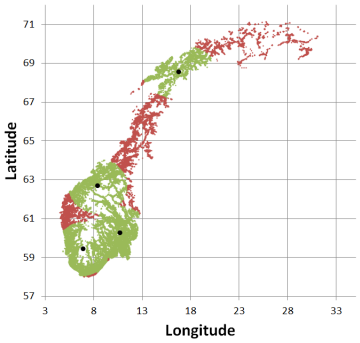


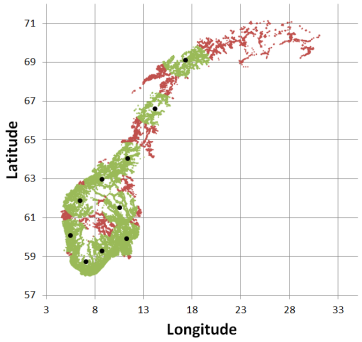
Fig. 9.2: *Fine grid versus municipality evaluated in fine grid.*



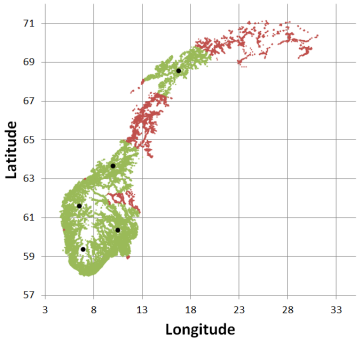
(a) *30 minutes, 90% coverage*



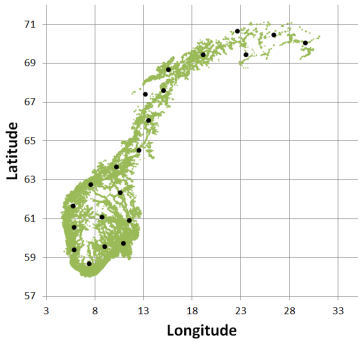
(d) *45 minutes, 90% coverage*



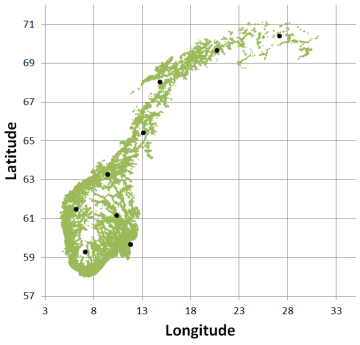
(b) *30 minutes, 95% coverage*



(e) *45 minutes, 95% coverage*



(c) *30 minutes, 100% coverage*



(f) *45 minutes, 100% coverage*

Fig. 9.3: *Distribution of the bases and corresponding coverage in the greenfield scenario for response time targets of 30 and 45 minutes. Solutions for a coverage of at least 90%, 95%, and 100% coverage are given.*

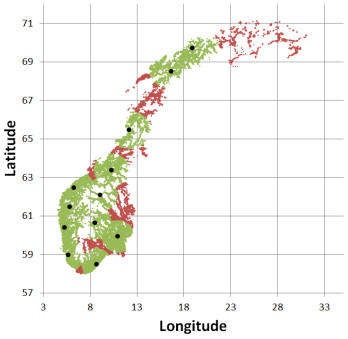
For practical purposes, it is of interest to find the coverage improvement that can be obtained with only a small number of base changes. For this, we use the fine grid data and the two different response time targets. First, we compute the coverage of the current set of bases. Then, the maximum coverage that can be obtained by replacing one or two of the current 12 bases is computed. Figure 9.4 gives the optimal base changes and the corresponding coverage. The current bases already cover 97.84% of the population within 45 minutes, and 90.40% within 30 minutes. By replacing two of the bases, this can be improved to almost complete coverage within 45 minutes, and 94.01% coverage within 30 minutes.

9.4 Discussion

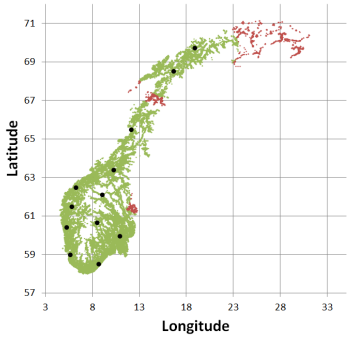
In this chapter, we applied the well-known MCLP to the air ambulance service in Norway. We considered a greenfield scenario where the optimal locations of the base stations were computed for different response time targets. We conclude that the required coverage can be achieved with considerably fewer bases than there are currently in use. A coverage of 95% of the population within 45 minutes can, for example, be achieved with only five bases. With nine optimally located bases, the entire population can be covered. The computations that include the current location of the bases show that almost complete coverage can be obtained by replacing two of the 12 existing bases. The location of the two new bases in Northern Norway are similar to the current location of two existing Sea King bases. Sea King helicopters are larger helicopters that are primarily used for offshore missions. These helicopters are not part of the official air ambulance system, but are sometimes used for air ambulance missions. As the coverage of the current bases would increase to 99.26% if regular helicopters would be located at the Sea King bases, these bases appear to fill the gap in the existing air ambulance base structure.

In determining the best locations for the emergency helicopters, we used MCLP, which maximizes the single coverage. As we have seen in previous chapters, single coverage does not always suffice. However, it gives an upper bound on the performance that can be obtained, as if no base is available within the time threshold, the patient will not be reached in time. In this way, the model has practical relevance. For future research, it would be interesting to evaluate the importance of incorporating backup coverage. If MEXCLP based models will be used for the air ambulance service in Norway, it is important to somehow incorporate region-dependent busy fractions, as significant differences in population density can be found in Norway.

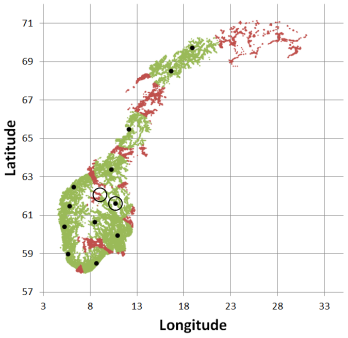
In the present model, ground services are not included. There is, however, a link with land ambulances in practice. Especially in the few densely populated areas in Norway, a land ambulance is the preferred means of transportation. For future research it would be interesting to see how the air ambulance configuration would change if the coverage by land ambulances is included. One could expect



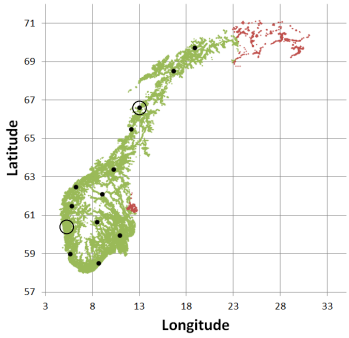
(a) *Current situation, 30 minute target (90.40%)*



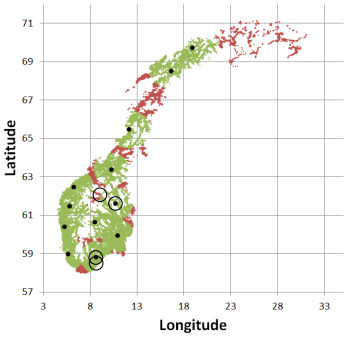
(b) *Current situation, 45 minute target (97.84%)*



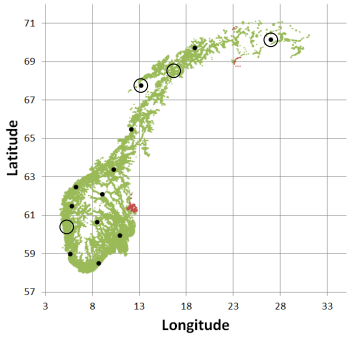
(c) *Change 1 base, 30 minute target (92.57%)*



(d) *Change 1 base, 45 minute target (98.89%)*



(e) *Change 2 bases, 30 minute target (94.01%)*



(f) *Change 2 bases, 45 minute target (99.88%)*

Fig. 9.4: *Distribution of the bases and corresponding coverage for a limited number of base changes with response time targets of 30 and 45 minutes.*

that the impact is limited, as Norway only has a few urban areas. When covering all rural areas, it is possible that most urban areas are automatically covered.

Currently, we used the population as a proxy for the number of calls in an area. Even though this is in line with the official requirements, it might not be optimal. Kristiansen et al. (2014) already showed that the incidence density and the population density do not necessarily overlap. Norway covers a large geographical area with diverse nature and strong seasonal weather effects. Consequently, people tend to flock to the coast in the summer, and to the mountains in the winter. This might call for different weights of the demand points. One problem for finding better estimates for call volumes might be that the location of calls is only recorded on municipality level. So, if weights based on historical demand are used, this higher level of data aggregation is required or data on a more detailed level must be collected. The experiments in this chapter suggest that data on municipality level already give very reasonable outcomes.

Given the increasing evidence that quick response times are essential in prehospital medical care (Lossius et al., 2002; Lossius and Lund, 2012; Østerås et al., 2015), decreasing the target response time is a topic for the political and medical debate. Our analysis quantifies possible practical consequences of reducing the target to 30 minutes. The results show that significantly more bases are required in order to fulfill such requirement. This fact should be incorporated if changing the response time target is considered.

List of acronyms

AA	Aerial Apparatus
ADP	Approximate Dynamic Programming
ALS	Advanced Life Support
ANN	Artificial Neural Network
ARIMA	Autoregressive Integrated Moving Averages
ARTM	Average Response Time Model
AZN	Ambulance Zorg Nederland
BACOP	Backup Coverage Problem
BLS	Basic Life Support
DAM	Dynamic Ambulance Management
DARP	Dial-A-Ride Problem
DP	Dynamic Programming
DSM	Double Standard Model
ED	Emergency Department
EMS	Emergency Medical Service
ERTM	Expected Response Time Model
FA	Fire Apparatus
FW	Fixed Wing
HEMS	Helicopter EMS
IG	Integrality Gap
ILP	Integer Linear Programming
IQR	Interquartile Range
LP	Linear Programming
LSCP	Location Set Covering Problem

MALP	Maximum Availability Location Problem
MCLP	Maximal Covering Location Problem
MEXCLP	Maximum Expected Coverage Location Model
MICU	Mobile Intensive Care Unit
MILP	Mixed Integer Linear Programming
MINLP	Mixed Integer Nonlinear Programming
MMT	Mobile Medical Team
MR	Marine Rescue Unit
NICU	Neonatal Intensive Care Unit
NP	Nondeterministic polynomial time
OR	Operations Research
P	Polynomial time
PICU	Pediatric Intensive Care Unit
RA	Rescue Apparatus
RAV	Regionale Ambulancevoorziening
REPRO	From REactive to PROactive planning of ambulance services
RIVM	Rijksinstituut voor Volksgezondheid en Milieu
RPMP	Reliability p-Median Problem
RW	Rotor Wing
SA	Standing Agreement
SSA	Singular Spectral Analysis
TIMEXCLP	Time-dependent MEXCLP
TS	Tabu Search
VNS	Variable Neighborhood Search

References

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.

L. Aboueljinane, E. Sahin, Z. Jemai, and J. Marty. A simulation study to improve the performance of an Emergency Medical Service: application to the French Val-de-Marne department. *Simulation Modelling Practice and Theory*, 47:46–59, 2014.

AIMMS BV. AIMMS, the user's guide. Technical report, 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 2012. Technical report, 2012.

Ambulancezorg Nederland. Ambulances in-zicht 2013. Technical report, 2013.

Ambulancezorg Nederland. Ambulances in-zicht 2014. Technical report, 2014.

T. Andersson and S. Särdqvist. Planning for effective use of fire and rescue service resources. In *Interflam 2007: 11th International fire science and engineering conference*, pages 1561–1566, 2007.

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

B. H. Andrews and S. M. Cunningham. L. L. Bean improves call-center forecasting. *Interfaces*, 25(6):1–13, 1995.

R. Aringhieri, G. Carello, and D. Morale. Ambulance location through optimization and simulation: the case of Milano urban area. In *The 38th annual conference of the Italian operation research society optimization and decision sciences*, pages 1–29, 2007.

R. Aringhieri, G. Carello, and D. Morale. Supporting decision making to improve the performance of an Italian Emergency Medical Service. *Annals of Operations Research*, 236(1):131–148, 2016.

J. R. Baker and K. E. Fitzpatrick. Determination of an optimal forecast model for ambulance demand using goal programming. *Journal of Operational Research Society*, 37(11):1047–1059, 1986.

R. H. Ballou. Measuring transport costing error in customer aggregation for facility location. *Transportation Journal*, 33(3):49–59, 1994.

T. C. van Barneveld, S. Bhulai, and R. D. van der Mei. A dynamic ambulance management model for rural areas: computing redeployment actions for relevant performance measures. *Health Care Management Science*, 2015.

R. Batta and N. R. Mannur. Covering-location models for emergency situations that require multiple response units. *Management Science*, 36(1):16–23, 1990.

A. Beaudry, G. Laporte, T. Melo, and S. Nickel. Dynamic transportation of patients in hospitals. *OR Spectrum*, 32(1):77–107, 2010.

H. Behrendt and R. Schmiedel. Ermittlung der bedarfsgerechten Fahrzeugvorhaltung im Rettungsdienst. *Notfall und Rettungsmedizin*, 5(3):190–203, 2002.

C. E. Bell and D. Allen. Optimal planning of an emergency ambulance service. *Socio-Economic Planning Sciences*, 3:95–101, 1969.

J. van den Bergh, J. Beliën, P. de Bruecker, E. Demeulemeester, and L. de Boeck. Personnel scheduling: a literature review. *European Journal of Operational Research*, 226(3):367–385, 2013.

P. L. van den Berg and K. 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, 2016.

L. Bianci, J. E. Jarrett, and C. R. Hanumara. Forecasting incoming calls to telemarketing centers. *Journal of Business Forecasting*, 12(2):3–11, 1993.

R. Bjarnason, P. Tadepalli, A. Fern, and C. Niedner. Simulation-based optimization of resource placement and emergency response. In *Proceedings of the Twenty-First Innovative Application of Artificial Intelligence Conference*, pages 47–53, 2009.

P. V. G. Bradbeer, C. Findlay, and T. C. Fogarty. An ambulance crew rostering system. In *Real-World Applications of Evolutionary Computing*, pages 267–279. Springer-Verlag Berlin Heidelberg, 2000.

L. Brotcorne, G. Laporte, and F. Semet. Ambulance location and relocation models. *European Journal of Operational Research*, 147(3):451–463, 2003.

T. A. Carnes, S. G. Henderson, D. B. Shmoys, M. Ahghari, and R. D. MacDonald. Mathematical Programming guides air-ambulance routing at Ornge. *Interfaces*, 43(3):232–239, 2013.

A. J. E. Carter, J. B. Gould, P. Vanberkel, J. L. Jensen, J. Cook, S. Carrigan, M. R. Wheatley, and A. H. Travers. Offload zones to mitigate Emergency

Medical Services (EMS) offload delay in the emergency department: a process map and hazard analysis. *Canadian Journal of Emergency Medicine*, 17(06): 670–678, 2015.

G. M. Carter, J. M. Chaiken, E. Ignall, and N. M. Jun. Response areas for two emergency units. *Operations Research*, 20(3):571–594, 1972.

P. A. Casillas. Data aggregation and the p-median problem in continuous space. In *Spatial analysis and location-allocation models.*, pages 227–244. Van Nostrand Reinhold Co., NY, 1987.

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. Chen, T. Lu, M. Ma, and W. Sun. Demand forecast using data analytics for the pre-allocation of ambulances. *IEEE Journal of Biomedical and Health Informatics*, 2015.

Z. Chen and H. Xu. Dynamic column generation for dynamic vehicle routing with time windows. *Transportation Science*, 40(1):74–88, 2006.

P. Chevalier, I. Thomas, D. Geraets, E. Goetghebeur, O. Janssens, D. Peeters, and F. Plastria. Locating fire stations: an integrated approach for Belgium. *Socio-Economic Planning Sciences*, 46(2):173–182, 2012.

S. Cho, H. Jang, T. Lee, and J. G. Turner. Simultaneous location of trauma centers and helicopters for Emergency Medical Service planning. *Operations Research*, 62(4):751–771, 2014.

K. C. Chong, S. G. Henderson, and M. E. Lewis. The vehicle mix decision in Emergency Medical Service systems. *Manufacturing & Service Operations Management*, pages 1–45, 2015.

R. L. Church and C. S. ReVelle. The maximal covering location problem. *Papers in Regional Science*, 32(1):101–118, 1974.

V. Chvátal. Edmonds polytopes and a hierarchy of combinatorial problems. *Discrete Mathematics*, 4:305–337, 1973.

O. Clarke. 40,000 hours of ambulance delays at Welsh hospital A&Es. *BBC News Wales*, 2015.

J. Cordeau and G. Laporte. The Dial-a-Ride Problem: models and algorithms. *Annals of Operations Research*, 153(1):29–46, 2007.

J. R. Current and D. A. Schilling. Analysis of errors due to demand data aggregation in the Set Covering and Maximal Covering Location Problems. *Geographical Analyis*, 22(2):116–126, 1990.

G. B. Dantzig. Maximization of a linear function of variables subject to linear inequalities. In *Activity Analysis of Production and Allocation*, pages 339–347. Wiley, NY, 1951.

M. S. Daskin. A Maximum Expected Covering Location Model: formulation, properties and heuristic solution. *Transportation Science*, 17(1):48–70, 1983.

M. S. Daskin. Location, dispatching, and routing model for emergency services with stochastic travel times. In *Spatial analysis and location-allocation models.*, pages 224–265. Van Nostrand Reinhold Co., NY, 1987.

M. S. Daskin. *Network and discrete Location: models, algorithms, and applications*. 1995.

M. S. Daskin, A. E. Haghani, M. Khanal, and C. Malandraki. Aggregation effects in maximum covering models. *Annals of Operations Research*, 18:115–140, 1989.

S. F. Dean. Why the closest ambulance cannot be dispatched in an urban Emergency Medical Services system. *Prehospital and Disaster Medicine*, 23(02): 161–165, 2008.

W. F. Dick. Anglo-American vs. Franco-German Emergency Medical Services system. *Prehospital and Disaster Medicine*, 18:29–37, 2003.

E. W. Dijkstra. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1(1):269–271, 1959.

R. P. Dwars. *Capacity planning of emergency call centers*. Master thesis, VU University Amsterdam, 2013.

M. Dzator and J. Dzator. An effective heuristic for the p-median problem with application to ambulance location. *Opsearch*, 50(1):60–74, 2013.

E. T. Erdemir, R. Batta, P. A. Rogerson, A. Blatt, and M. Flanigan. Joint ground and air emergency medical services coverage models: a greedy heuristic solution approach. *European Journal of Operational Research*, 207(2):736–749, 2010.

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, 2010.

E. Erkut and B. Bozkaya. Analysis of aggregation errors for the p-median problem. *Computers & Operations Research*, 26:1075–1096, 1999.

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 Analyis*, 41:43–65, 2009.

J. T. van Essen, J. L. Hurink, S. Nickel, and M. Reuter-Oppermann. Models for ambulance planning on the strategic and the tactical level. Technical Report WP-434, Beta Research School for Operations Management and Logistics, Eindhoven, 2013.

R. L. Francis, T. J. Lowe, M. B. Rayco, and A. Tamir. Aggregation error for location models: survey and analysis. *Annals of Operations Research*, 167(1): 171–208, 2009.

O. Fujiwara, T. Makjamroen, and K. K. Gupta. Ambulance deployment analysis: a case study of Bangkok. *European Journal of Operational Research*, 31(1): 9–18, 1987.

T. Furuta and K. Tanaka. Maximal Covering Location Model for doctor-helicopter systems with two types of coverage criteria. *Urban and Regional Planning Review*, 1:39–58, 2014.

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(1):22–28, 2006.

J. B. Goldberg and L. Paz. Locating emergency vehicle bases when service time depends on call location. *Transportation Science*, 25(4):264–280, 1991.

J. B. Goldberg, R. Dietrich, J. Ming Chen, M. Mitwasi, T. D. 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.

R. E. Gomory. Outline of an algorithm for integer solution to linear programs. *Bulletin of the American Mathematical Society*, 64(5):275–278, 1958.

R. E. Gomory. Solving Linear Programming Problems in Integers. *Combinatorial Analysis*, 10:211–215, 1960.

R. E. Gomory. An algorithm for integer solutions to linear programs. *Recent Advances in Mathematical Programming*, 64:260–302, 1963.

L. V. Green and P. J. Kolesar. Improving emergency responsiveness with Management Science. *Management Science*, 50(8):1001–1014, 2004.

I. Gurobi Optimization. Gurobi Optimizer Reference Manual. Technical report, 2015.

S. I. Harewood. Emergency ambulance deployment in Barbados: a multi-objective approach. *The Journal of the Operational Research Society*, 53(2): 185–192, 2002.

S. G. Henderson and A. J. Mason. Ambulance service planning: simulation and data Visualisation. In *Operations Research and Health Care: A Handbook of Methods and Applications*, chapter 4, pages 77–102. Springer US, Boston, 2004.

F. S. Hillier and G. J. Lieberman. *Introduction to Operations Research*. McGraw-Hill, 10th edition, 2014.

E. L. Hillsman and R. Rhoda. Errors in measuring distances from population to service centers. *The Annals of Regional Science*, 12(3):74–88, 1978.

M. J. Hodgson, F. Shmulevitz, and M. Körkel. Aggregation error effects on the discrete-space p-median model: the case of Edmonton, Canada. *The Canadian Geographer*, 41(4):415–428, 1997.

K. Hogan and C. S. ReVelle. Concepts and applications of backup coverage. *Management Science*, 32(11):1434–1444, 1986.

J. M. Hogg. The siting of fire stations. *Journal of the Operational Research Society*, 19(3):275–287, 1968.

J. P. Holcomb and N. R. Sharpe. Forecasting police calls during peak times for the city of Cleveland. *CS-BIGS*, 1(1):47–53, 2007.

G. Holmes, A. Ingolfsson, R. Patterson, and E. Rolland. Model specification and data aggregation for emergency services facility location. *Production and Operations Management (Submitted)*, 2014.

M. Hoogeveen. Ambulance care in Europe. Technical report, Ambulancezorg Nederland, 2010.

O. Hughes. Crews waste 30,000 hours at A&E. *Daily Post. Wales*, page 8, 2009.

ILOG. Ilog CPLEX 12.5 reference manual. Technical report, 2009.

ILOG. Ilog CPLEX 12.6 reference manual. Technical report, 2013.

A. Ingolfsson. EMS planning and management. In *Operations Research and Health Care Policy*, pages 105–128. Springer New York, 2013.

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.

S. Jain and C. McLean. A framework for modeling and simulation for emergency response. In *Proceedings of the 2003 Winter Simulation Conference*, pages 1068–1076, 2003.

H. Jasim. Relief staff rostering for the St. John ambulance service. 2002.

A. Jones. Slow ambulance turnarounds cost NHS more than 10m pounds. *BBC News Wales*, 2011.

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.

Y. Kergosien, C. Lenté, D. Piton, and J.-C. Billaut. A tabu search heuristic for the dynamic transportation of patients between care units. *European Journal of Operational Research*, 214(2):442–452, 2011.

Y. Kergosien, P. Soriano, M. Gendreau, Y. Kergosien, V. Bélanger, P. Soriano, and A. Ruiz. A generic and flexible simulation-based analysis tool for EMS management. 2014.

L. G. Khachiyan. A polynomial algorithm in linear programming. *Doklady Akademii Nauk SSSR*, 244:1093–1096, 1979.

V. A. Knight and P. R. Harper. Modelling Emergency Medical Services with phase-type distributions. *Health Systems*, 1(1):58–68, 2012.

V. A. Knight, P. R. Harper, and L. Smith. Ambulance allocation for maximal survival with heterogeneous outcome measures. *Omega*, 40(6):918–926, 2012.

M. L. Koç and M. Bostancioğlu. A reliability based solution to an ambulance location problem using fuzzy set theory. *International Journal of Natural and Engineering Sciences*, 5(1):13–17, 2011.

P. J. Kolesar and E. H. Blum. Square root laws for fire engine response distances. *Management Science*, 19(12):1368–1378, 1973.

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. L. N. Zwakhals. Referentiekader spreiding en beschikbaarheid ambulancezorg 2008. Technical report, 2008.

G. J. Kommer and S. L. N. Zwakhals. Modellen referentiekader ambulancezorg 2008. Technical report, 2011.

G. Koole and A. Mandelbaum. Queueing models of call centers: an introduction. *Annals of Operations Research*, 113:41–59, 2002.

E. Kozan and N. Mesken. A simulation model for emergency centres. In *Proceedings of the International Congress on Modelling and Simulation. Advances and Applications for Management and Decision Making*, pages 2602–2608, 2005.

T. Kristiansen, H. M. Lossius, M. Rehn, P. Kristensen, H. M. Gravseth, J. Røislien, and K. Søreide. Epidemiology of trauma: a population-based study of geographical risk factors for injury deaths in the working-age population of Norway. *Injury*, 45(1):23–30, 2014.

M. P. Larsen, M. S. Eisenberg, R. O. Cummins, and A. P. Hallstrom. Predicting survival from out-of-hospital cardiac arrest: a graphic model. *Annals of emergency medicine*, 22(11):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.

T. Lee, H. Jang, S. Cho, and J. G. Turner. A simulation-based iterative method for trauma center - air ambulance location problem. In *Proceedings of the 2012 Winter Simulation Conference*, pages 955–966, 2012.

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(3):281–310, 2011.

Y. Li and E. Kozan. Rostering ambulance services. In *Asia Pacific Industrial Engineering and Management Systems Conference*, pages 795–801, 2009.

H. M. Lossius and C. G. Lund. Prehospital assessment of stroke: time is brain. *Practice*, 132:1848–1849, 2012.

H. M. Lossius, E. Søreide, R. Hotvedt, S. A. Hapnes, O. V. Eielsen, O. H. Førde, and P. A. Steen. Prehospital advanced life support provided by specially trained physicians: is there a benefit in terms of life years gained? *Acta anaesthesiologica Scandinavica*, 46:771–778, 2002.

J. A. Lowthian, P. A. Cameron, J. U. Stoelwinder, A. Curtis, A. Currell, M. W. Cooke, and J. J. McNeil. Increasing utilisation of emergency ambulances. *Australian Health Review*, 35(1):63–69, 2011.

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.

D. S. Matteson, M. W. McLean, D. B. Woodard, and S. G. Henderson. Forecasting emergency medical service call arrival rates. *The Annals of Applied Statistics*, 5(2B), 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. 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.

R. McCormack and G. Coates. A simulation model to enable the optimization of ambulance fleet allocation and base station location for increased patient survival. *European Journal of Operational Research*, 2015.

E. Melachrinoudis, A. B. Ilhan, and H. Min. A Dial-a-Ride problem for client transportation in a health-care organization. *Computers and Operations Research*, 34(3):742–759, 2007.

I. F. de la Mota, A. V. Garduño, and E. S. Pérez. Simulation and optimization of the pre-hospital care system of the national university of Mexico. In *Applied Simulation and Optimization*, pages 233–276. Springer International Publishing, Cham, 2015.

A. T. Murray and D. Tong. GIS and spatial analysis in the media. *Applied Geography*, 29(2):250–259, 2009.

S. Nickel, M. Reuter-Oppermann, and F. Saldanha-da Gama. Ambulance location under stochastic demand: a sampling approach. *Operations Research for Health Care*, 2015.

NMoHaC Services. Stortingsmelding 43: Om Akuttmedisinsk Beredskap. Technical report, 2000.

Ø. Østerås, G. Brattebø, and J.-K. Heltne. Helicopter-based Emergency Medical Services for a sparsely populated region: A study of 42,500 dispatches. *Acta Anaesthesiologica Scandinavica*, 2015.

C. H. Papadimitriou and K. Steiglitz. *Combinatorial Optimization: Algorithms and Complexity*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1982.

S. N. Parragh. *Ambulance routing problems with rich constraints and multiple objectives*. PhD thesis, 2009.

S. N. Parragh, K. F. Doerner, and R. F. Hartl. A heuristic two-phase solution approach for the multi-objective Dial-a-Ride Problem. *Networks*, 54(4):227–242, 2009.

D. R. Plane and T. E. Hendrick. Mathematical Programming and the location of fire companies for the Denver Fire Department. *Operations Research*, 25 (4):563–578, 1977.

H. K. Rajagopalan. Ambulance deployment and shift scheduling: an integrated approach. *Journal of Service Science and Management*, 04(01):66–78, 2011.

H. K. Rajagopalan, C. Saydam, and J. Xiao. A multiperiod set covering location model for dynamic redeployment of ambulances. *Computers & Operations Research*, 2008.

Rand Fire Project. *Fire department deployment analysis*. Elsevier North-Holland, New York, 1979.

J. F. Repede and J. J. Bernardo. Case study 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 and K. Hogan. The Maximum Availability Location Problem. *Transportation Science*, 23(3):192–200, 1989.

C. S. ReVelle and R. W. Swain. Central facilities location. *Geographical Analysis*, 2(1):30–42, 1970.

U. Ritzinger, J. Puchinger, C. Rudloff, and R. F. Hartl. Real-world patient transportation. In *19th ITS World Congress*, pages 1–4, 2012.

U. Ritzinger, J. Puchinger, and R. F. Hartl. Dynamic programming based metaheuristics for the Dial-a-Ride Problem. *Annals of Operations Research*, pages 1–18, 2014.

M. Schilde, K. F. Doerner, and R. F. Hartl. Metaheuristics for the dynamic stochastic Dial-a-Ride Problem with expected return transports. *Computers & Operations Research*, 38(12):1719–1730, 2011.

D. A. Schilling, C. S. ReVelle, J. Cohon, and D. J. Elzinga. Some models for fire protection locational decisions. *European Journal of Operational Research*, 5 (1):1–7, 1980.

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.

H. Setzler, C. Saydam, and S. Park. EMS call volume predictions: a comparative study. *Computers & Operations Research*, 36(6):1843–1851, 2009.

L. V. Snyder and M. S. Daskin. Reliability models for facility location: the expected failure cost case. *Transportation Science*, 39(3):400–416, 2005.

Statistics Canada. Population by year, by province and territory. Technical report, 2015.

Statistics Norway. Population, 1 January 2015. Technical report, 2015.

A. J. Swersey. A Markovian decision model for deciding how many fire companies to dispatch. *Management Science*, 28(4):352–365, 1982.

A. J. Swersey. The Deployment of Police, Fire, and Emergency Medical Units. In *Handbook in OR & MS*, volume 6, chapter 6, pages 151–200. Elsevier Science BV, Amsterdam, 1994.

R. A. Takeda, J. A. Widmer, and R. Morabito. Analysis of ambulance decentralization in an urban emergency medical service using the hypercube queueing model. *Computers & Operations Research*, 34(3):727–741, 2007.

The Optimization Firm. BARON user manual v. 12.0.0. Technical report, 2013.

C. Toregas, R. W. Swain, C. S. ReVelle, and L. Bergman. The Location of emergency service facilities. *Operations Research*, 19(6):1363–1373, 1971.

J. L. Vile, J. W. Gillard, P. R. Harper, and V. A. Knight. Predicting ambulance demand using SSA. *Journal of Operational Research Society*, 63(11):1556–1565, 2012.

J. L. Vile, J. Gillard, P. R. Harper, and V. A. Knight. Time-dependent stochastic methods for managing and scheduling Emergency Medical Services. *Operations Research for Health Care*, 2015.

Y. Yue, L. Marla, and R. Krishnan. An efficient simulation-based approach to ambulance fleet allocation and dynamic redeployment. In *AAAI Conference on Artificial Intelligence*, pages 398–405, 2012.

E. Zakariassen, O. Uleberg, and J. Røislien. Helicopter Emergency Medical Services response times in Norway: do they matter? *Air Medical Journal*, 34 (2):98–103, 2015.

L. Zhang. *Simulation optimisation and Markov models for dynamic ambulance redeployment*. PhD thesis, University of Auckland, 2012.

G. M. Zuidhof. Capacity planning of ambulance services: statistical analysis, forecasting and staffing. Technical report, 2010.

Publications by the author

Journal publications

P. L. van den Berg and K. Aardal. Time-dependent MEXCLP with start-up and relocation cost. *European Journal of Operational Research*, 242(2):383-389, 2015.
(Basis for Chapter 3)

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, 2016.
(Basis for Chapter 4)

K. Aardal, P. L. van den Berg, D. Gijswijt, and S. Li. Approximation algorithms for hard capacitated k-facility location problems. *European Journal of Operational Research*, 242(2):358-368, 2015.

Conference proceedings

P. L. van den Berg, J. T. van Essen, and E. J. van Harderwijk. Comparison of static ambulance location models. To appear in *Proceedings of the 3th International IEEE conference on Logistics Operations Management*, 2016.
(Basis for Section 2.1)

Unpublished manuscripts

M. Reuter-Opperman, P. L. van den Berg, and J. Vile. Logistics for Emergency Medical Service systems. *Under review.*
(Basis for Section 1.3)

P. L. van den Berg, G. A. G. Legemaate, and R. D. van der Mei. Boosting the responsiveness of firefighter services by relocating few base stations in Amsterdam. *Under review.*
(Basis for Chapter 5)

P. L. van den Berg, J. T. van Essen. Scheduling non-urgent patient transportation while maximizing emergency coverage. *Working paper*.
(Basis for Chapter 6)

P. L. van den Berg, S. G. Henderson, M. Ahghari, and R. D. MacDonald. Simulation and optimization for air ambulance provider in Ontario. *Working paper*.
(Basis for Chapter 8)

J. Røislien, P. L. van den Berg, T. Lindner, E. Zakariassen, K. Aardal, and J. T. van Essen. Exploring optimal air ambulance base locations in Norway using advanced mathematical modelling. *Under review*.
(Basis for Chapter 9)

C. J. Jagtenberg, P. L. van den Berg, and R. D. van der Mei. Benchmarking online dispatch algorithms for Emergency Medical Services. *Under review*.

Other publications

P. L. van den Berg, and J. T. van Essen. Analyse dienstrooster zorgambulances voor RAV Utrecht. Technical report, 2015.
(Basis for Chapter 7)

P. L. van den Berg. Standplaatsanalyse voor RAV Flevoland. Technical report, 2014.

K. Aardal, T. C. van Barneveld, P. L. van den Berg, S. Bhulai, M. van Buuren, J. T. van Essen, C. J. Jagtenberg, G. J. Kommer, G. A. G. Legemaate, and R. D. van der Mei. Van reactieve naar proactieve planning van ambulancediensten. *Nieuw Archief voor Wiskunde*, 2015.

P. L. van den Berg, Guido Legemaate, and R. D. van der Mei. Boosting the responsiveness of firefighter services with Mathematical Programming. *ERCIM news*, 105:20-21, 2016.

Acknowledgments

Na ruim vier jaar is het dan zo ver dat mijn proefschift voltooid is. Dit resultaat is natuurlijk tot stand gekomen met hulp van velen, waarvan ik een deel op deze plek graag wil bedanken.

Mijn dank gaat bovenal uit naar mijn promotors. Karen, naast de inhoudelijke begeleiding die ik van jou ontvangen heb, wil ik je vooral danken voor jouw rol in mijn ontwikkeling tot kritische academicus. Jouw hoge standaard en kritische blik zullen mij helpen in het vervolg van mijn carrière. Verder wil ik je danken voor de kansen die je me hebt gegeven tot het geven van onderwijs. Met veel plezier ben ik betrokken geweest bij de verschillende optimaliseringsvakken. Rob, jouw enthousiasme heeft in grote mate bijgedragen aan het succes van het REPRO-project. Bedankt ook voor de talloze keren dat jij mee ging naar presentaties bij de verschillende eindgebruikers van onze resultaten.

De overige REPRO-leden wil ik graag bedanken voor de vele nuttige discussies die zeker hebben bijgedragen aan de totstandkoming van dit proefschrift. In het bijzonder wil ik Martin bedanken voor zijn hulp bij tal van IT-problemen. Naast de overige onderzoekers binnen REPRO gaat mijn dank ook uit naar de verschillende ambulancediensten die betrokken waren bij het REPRO-project. Gesprekken met hen garandeerden de praktische relevantie van mijn onderzoek en hebben geleid tot vele nieuwe onderzoeksrichtingen.

I would further like to thank Shane Henderson for traveling all the way from Ithaca to be part of my PhD committee and for being a wonderful host during my stay at Cornell University. From the very first day I arrived in Ithaca, I felt more than welcome. For that, I would also like to thank the PhD students at ORIE for including me in numerous activities.

Ook de overige leden van mijn promotiecommissie wil ik graag bedanken voor de tijd die zij gestoken hebben in het bestuderen van mijn proefschrift en hun aanwezigheid bij mijn verdediging.

Tijdens mijn promotietraject heb ik de kans gehad om met veel verschillende mensen samen te werken. De nieuwe inzichten die hieruit voortvloeiden hebben de kwaliteit van mijn proefschrift zeker goed gedaan. In het bijzonder dank ik de

organisatoren en deelnemers van het European Summer Institute OR in Health Care voor twee intensieve maar zeer nuttige weken.

De afgelopen jaren heb ik mijn tijd mogen verdelen tussen mijn werkplekken in Delft en Amsterdam. Ondanks dat ik steeds vaker koos om mijn tijd bij het CWI door te brengen, heb ik de vierde verdieping van de TU Delft altijd als een prettige werkplek ervaren. Met name wil ik mijn kamergenoten Fred, Matthijs, Theresia en Teun bedanken. Mijn CWI collega's bedank ik voor het creëren van een ontspannen werksfeer waar altijd tijd was voor een kop koffie, een potje tafeltennis, of een borrel om de week mee af te sluiten.

Verder wil ik graag mijn paranimfen Vivian en Jan-Pieter bedanken. Natuurlijk voor hun bereidheid om mij bij te staan tijdens mijn verdediging, maar bovenal voor de vele redenen die ze mij hebben gegeven om hen als paranimfen te vragen.

Tenslotte wil ik graag mijn vrienden en familie bedanken voor een meer indirecte bijdrage aan mijn promotie. Hun geïnteresseerde vragen naar mijn onderzoek hebben bijgedragen aan het blijven zien van de grote lijn. Maar meer nog heeft hun gezelschap geleid tot het behouden van het besef dat er meer is dan een promotietraject. Papa, ondanks dat je dit niet mee hebt mogen maken, weet ik dat je ongelofelijk trots zou zijn. Bedankt dat ik trots kan zijn iedere keer als mij gezegd wordt dat ik steeds meer op jou begin te lijken. Anne Frances, bedankt dat dankzij jou het laatste jaar van mijn promotie zo veel leuker is geweest.

Pieter van den Berg
Mei, 2016
Delft

About the author

Pieter van den Berg was born in Ermelo, the Netherlands on November 21th, 1988. He obtained his VWO diploma at the Christelijk College Groevenbeek in Ermelo in 2007. After that, he joined the Bachelor's program Econometrics and Operational Research at VU University Amsterdam, which he completed in 2011. In 2012, he completed the corresponding Master's program with a specialization in Operational Research cum laude at the same university. His Master's thesis was written during an internship at the innovation department of the Dutch railway company NS. Here, he focused on the real-time rolling stock management in case of disruptions.

In 2012, Pieter started as a PhD student at the Delft University of Technology under supervision of Karen Aardal and Rob van der Mei. His research was part of a larger research project on the logistics of emergency medical service providers. At the end of the third year of his PhD program, Pieter spent three months at the School of Operations Research and Information Engineering (ORIE) at Cornell University in the United States. After finishing his PhD, Pieter will continue to do research as an assistant professor at the Rotterdam School of Management at Erasmus University.