Vrije Universiteit Amsterdam
Centrum Wiskunde en Informatica Amsterdam

Capacity Planning of Ambulance Services:

Statistical Analysis, Forecasting and Staffing

Author: G. M. Zuidhof

Supervisor: R. D. van der Mei

Supervisor: S. Bhulai

Preface

The program for the master of science in Mathematics is concluded by a research project or an external project to be carried out within a business, industry or research facility other than the department of Mathematics. I chose for an external project at the research facility CWI, the Centrum Wiskunde en Informatica.

The research area of my Master Mathematics is Stochastics. The head of the Probability and Stochastic Networks Group at CWI, prof.dr. Rob van der Mei, proposed several subjects for my master project. One of them was about the efficient planning of ambulance services, which I decided upon because of its social relevance.

I would like to thank dr. Sandjai Bhulai and prof.dr. Rob van der Mei for being my supervisors at the VU and CWI respectively, and dr. René Bekker for being the second reader. Because of their suggestions and remarks I was able to finish this long term project. Also I would like to thank my friends for their interest, support and understanding of my common absence. Above all, for their patience and supporting words I would like to thank my family, which of course includes Bas.

Geertje Zuidhof
July 2010

Executive Summary

The current staffing method for ambulances of one of the ambulance service providers of Amsterdam is static. During the weekend and on special occasions such as Queen's Day and the first day of January an extra number of ambulances is scheduled. In this thesis we investigate the data of the number of ambulance requests, the corresponding occupancy time, and the priority of the request.

Data analysis on the number of ambulance requests per day indicate correlations between the priorities, successive days and between seven days. Also patterns for the month of the year, days of the week, and hours of the day are visible. Based on these patterns we composed for each of the ambulance priorities a non-homogeneous Poisson process describing the number of ambulance requests per day, which performs well. A multinomial distribution is developed to spread the number of requests over the hours of the day. The occupancy time of the ambulance requests differs significantly for the different priorities, days of the week and hours of the day. The empirical occupancy time distributions are categorized according to these distinctions.

For each of the priorities seperately we applied several forecasting models to obtain forecasts of a two-week horizon. The multiple linear regression model generates the best fit to the data, and the most accurate predictions, since this model considers month of the year, day of the week, and holiday effects.

By modifying a simple staffing rule, different staffing levels are obtained. The modification takes into account the patterns in ambulance requests, and the differences in occupancy time. We distinguish between staffing levels based on the number of necessary ambulances of the different priorities combined, and a combination of the staffing levels generated separately for each of the priorities. The first type outperforms the current staffing method, on costs and performance.

Contents

Chapter 1

Introduction

In this chapter we will start with some words on CWI where this internship took place. Then we will give a motivation of the research done during this internship, and provide our research questions. We finish with the approach, and the structure of this master thesis.

1.1 About CWI

Founded in 1946, CWI is the national research center for mathematics and computer science in the Netherlands. The vision of CWI is twofold: to perform cutting-edge fundamental research in Mathematics and Computer Science, and to transfer knowledge to academia and to Dutch and European industry. This results in importance for our economy, from payment systems and cryptography to telecommunication and the stock market, from public transport and internet to water management and meteorology.

Within CWI the research group Probability and Stochastic Networks has a long-standing tradition in the field of performance modeling and solution techniques for stochastic evaluation and optimization problems. Examples can be found in areas like communication and information systems, biology, economics and logistics. This group develops and studies stochastic and statistical models that yield fundamental understanding and enable control and optimization of such systems. Analysis of these models relies on techniques from fundamental probability theory, queueing theory, stochastic scheduling, spatial stochastics and stochastic geometry. Besides its focus on methodological aspects of stochastic models, the group also has a strong focus on the applicability of the results. The group has a broad national and international network of collaborations with industrial partners, governmental and academic institutions.

More information on CWI can be found at its web site `http://www.cwi.nl`.

1.2 Motivation

In serious life-threatening situations where every second counts, an effective Emergency Medical Service (EMS) can make the difference between life and death. Therefore, ambulance service providers must meet strict requirements in terms of response times, the time between an incoming emergency call and the moment of arrival of an ambulance.
The primary performance measure for an EMS system is typically the fraction of calls to which help is provided within some time standard, from the instant the call was made. Other performance measures could be: the probability of all ambulances being occupied, the capability of a certain ambulance configuration to cover future calls, cost effectiveness, the type of care medical personnel is trained for, and enough necessary medical equipment. To realize short response times of the ambulances and crew (at affordable cost) efficient planning of ambulance services is crucial. The Dutch government recognizes the public interest of realizing short response times: the "Miljoenennota 2009" [27] states that "In 2009 12 million will be invested extra to improve the availability of ambulances. The goal is that in 95 % of the cases the response time will be less than 15 minutes."

The requests for medical emergency care, and the time an ambulance generally is occupied, varies over time. For some of these variations a pattern can be discovered. When the ambulance planner is trained in observing and foreseeing these variations, a timely planning can be determined, which results in more efficient planning of training personnel or maintenance of the ambulances.

Scientific problem: A highly complicating but scientifically challenging factor that influences the EMS demand is randomness; emergency-call arrival patterns tend to be higly bursty and heavily time- and location-dependent, and the availability of ambulance vehicles and personnel when needed is largely random due to uncertainty in their occupancy times. This is why the classical planning techniques, which assume that the demand of ambulances is known upfront, are not applicable. Hence the offered service quality is much lower than it could (and should) be. By predicting the demand of ambulances, and taking into account the randomness of their occupancy times, planning can be much more efficient and result in cost reduction.

1.3 Research Questions

The realization of an effective planning of ambulances for the near future invites some questions. We will focus on the following research questions, each of which covers a range of subquestions:

Research Question 1: How to accurately predict the number of ambulance rides over time?
Nowadays the number of planned ambulances is static. In the weekends and on special occasions, such as Queen's Day and New Year's eve, a number of extra ambulances are scheduled. Predicting the demand of medical service, by for instance taking into account weekly patterns, is the starting point of generating a planning. The importance of good forecasts is evident; inaccurate predictions can lead to understaffing.

Typical sub-questions are:

- Does the data of the number of ambulance requests show correlations, typical yearly, weekly and daily arrival patterns?
- Does the data provide reason to distinguish between ambulance rides of different priorities?
- Which factors are of influence when modeling a non-homogeneous Poisson process to the number of requested ambulances?
- What forecasting methods employ the observations of the number of ambulance requests?
- How can one measure the performance of a forecasting model?
- Which forecasting model is preferable?

Research Question 2: How to determine the occupancy time of ambulances?
The durations of ambulance rides can differ significantly, for instance because of: the type of medical emergency, the time medical care is needed, the distance between starting point of the ride and its destination, and busy traffic. During some rides, only after a few seconds, the ambulance driver can be requested to abort the ride because a more urgent medical situation needs attention. Even such a short ride contributes to the total number of rides per day. Hence when an estimate of the number of rides per day, or even per hour is provided, estimating when, what number of ambulances is occupied, requests for insights in the time that ambulances are occupied.

Typical sub-questions are:

- Does the data of the occupancy time of the ambulances show correlations, typical yearly, weekly and daily occupancy time patterns?

- Does the data provide reason to distinguish between ambulance rides of different priorities?

Research Question 3: How to properly staff ambulance vehicles?
The time-dependent number of requests for medical emergency care, and the time it takes for the ambulance to be called free again, can provide information of how many ambulances are at least required on what hour of each day. To be sufficient enough more ambulances could be scheduled, but each with its costs.

Typical sub-questions are:

- How to determine the quality of a certain staffing method?
- What is the performance of the current staffing method?
- Which staffing models are applicable to our data?
- Should the ambulance rides be staffed for each of the priorities separately?
- Which of the considered staffing models provides the best staffing levels?

1.4 Approach

Step 1: To establish answers to the proposed research questions we start by discussing literature concerning alike forecasting and staffing problems. Since studies also applied to other areas around emergency medical systems are of interest, literature regarding other topics are also discussed in Chapter 2.

Step 2: The data provided by one of the ambulance service providers in Amsterdam will be statistically analyzed in Chapter 3. The number of ambulance requests will be observed for each of the different priorities. By checking for correlations and seasonal patterns, a characterization of the requests for ambulances is acquired. Based on the observations a non-homogeneous Poisson process describing the number of ambulance rides per day and an intra-day pattern of the ambulance requests are provided, which will be validated.

Step 3: In Chapter 3 we also investigate whether patterns exist in the occupancy times of ambulances, the time an ambulance is deployed to an emergency or transportation ride. Based on significant differences between the occupancy time of the ambulance rides starting on different hours, empirical densities of the occupancy time distribution will be categorized.

Step 4: Different forecasting models (e.g., Holt-Winters, Multiple Linear Regression, Auto-Regressive Integrated Moving Average), based on time-series or queueing theory, will be investigated in Chapter 4. For each of the ride priorities, the models will be used to obtain predictions of the number of ambulance rides per day for a two week forecast horizon. By checking their goodness of fit to our data, and investigate the accuracy of the obtained forecasts the models will be compared.

Step 5: Based on the Erlang loss model, simple staffing methods will be presented in Chapter 5. We will implement time-dependent arrivals of the ambulance requests, and the categorized empirical densities of the occupancy time of the ambulances. The staffing levels are based on the requests priority combined and seperately, and we investigate the difference. The obtained staffing levels will be compared to the current staffing by costs, and a measure of performance which depends on the probability of all ambulances being occupied.

In Chapter 6 the prominent conclusions and topics for further research will be provided.

Chapter 2

Literature Review

In this chapter we will review some of the literature on ambulance logistics; the planning, implementation, and control of resources and information used to facilitate an efficient way of serving a person in need of out-of-hospital medical care including possible transportation. We start by providing a number of research questions, of which two can serve as a base to our research, and discuss the literature subject to these questions.

Many studies exist on improving the quality of service of emergency medical systems. For a survey we refer to [12]. Typical research questions one can ask are:

1. How to predict the number of incoming ambulance calls?
2. Where to locate ambulance bases?
3. What number and type of vehicles should be deployed at each base?
4. How to schedule the ambulances?
5. How to decide when to dispatch which ambulance to an emergency?
6. How to decide when to redeploy (allocate) vehicles as a function of the system state?

Any solution to these type of questions requires careful balancing of political, economic and medical objectives.

Since we focus on efficient planning of ambulances in this thesis, we will first address research questions 1 and 3.

How to predict the number of incoming ambulance calls?
According to [12], little work has been done on long term demand forecasting of ambulance demand, but "the ability to predict demand is of paramount importance". Most models use

deterministic data or the average of a sample since there are few good estimating procedures to obtain distributions.

The literature discussing predicting ambulance demand can be divided into two categories, models predicting the spatial distribution, and models predicting the demand over time. Our research focuses on the forecasting of the number of ambulances based on information of the past demand. The research discussed in [7] can be of help since forecasts are provided for each hour of the day, based on past data containing the time of occurence of ambulance rides. Time-series models are developed and evaluated to the emergency medical service of the Canadian city Calgary. The estimated models are compared in terms of goodness-of-fit and forecasting accuracy. For their data, an autoregressive model of daily volumes, obtained after eliminating trend, seasonality, and special-day effects, and a multinomial distribution for the vector of number of calls in each hour conditional on the total volume of calls during the day, are superior.

Ambulance rides can be categorized into different priorities, each possibly with typical traffic characteristics. To obtain accurate forecasts of the daily volume of daily emergency and non-emergency EMS calls (using data from four South Carolina counties) Winters exponential smoothing models are discussed in [3]. To choose the exponential smoothing parameters, goal and quadratic programming is used. The resulting forecasts were compared to those obtained by using a multiple linear-regression model and a single-objective Winters exponential smoothing model. The smoothing method yielded more accurate forecasts; with smaller MSE (4.4) for three of the four counties.

An example where predictions are based on information of space and time is discussed in [31]. An artificial neural network (ANN) is designed to forecast demand volume of specific areas during different times of the day, and compared to a method used by the EMS agency in Mecklenburg County in North Carolina (MEDIC). Three hour forecasts for a 4×4 square mile region are considered. The forecasts used by the EMS agency are determined by averaging the call volume of the previous four time periods over the past 5 years. This method is quite common in the industry although some agencies use slightly more or fewer datapoints. They conclude that for a 2×2 mile one and three hour granularity, the ANN does not perform better than MEDIC. (And that the MSE (4.4) of an all zeros forecast even performs better, from which they conclude that any method-derived forecast for these small levels of specificity have little or no practical value.) At both the 4×4 mile one and three hour granularities the ANN performs better.

A lot of research has been done on forecasting arrivals to call centers from a variety of industries. Results can be applicable to forecasting ambulance demand since call arrivals in

a call centre follow a Poisson process [22], which can also be assumed for emergency medical calls [7], [26] and [34]. Univariate time series methods for forecasting intra-day arrivals for lead times from half an hour ahead to two weeks ahead, were evaluated in [36]. A notable feature of the data used is the presence of both an intra-week and an intra-day seasonal cycle. Of the five methods considered, a strong potential is indicated for the use of seasonal ARIMA modeling and the extension of Holt Winters method for predicting up to about two to three days ahead and that, for longer lead times a simplistic historical average is difficult to beat. Call center arrival data can be analyzed and modeled by making use of single value decomposition (SVD) [32]. The data is described as the number of incoming calls during the jth time period of the ith day. The SVD method can be used for preliminary data analysis to, for instance, detect days with typical arrival patterns, and to obtain a forecast based on results taking into account intra-day and inter-day patterns. An extension to these models is described in [33]. The intra-day call volume profiles are treated as high-dimensional vector time-series. First the dimensionality of the matrix of historical intra-day profiles is reduced by SVD, then time-series and regression techniques are applied. Assuming that the intra-day properties stay the same, forecasting intra-day profiles can be reduced to inter-day forecasting. Their methods are easy to implement and appear to be robust against model assumptions in their simulation studies.

How to schedule the ambulances?

Erdoğan et al. [9] developed a solution method for the combined problem of scheduling the working hours of ambulance crews for a given planning horizon and allocating the ambulances at stations. Since the shift scheduling problem is complex, they solve the location problem first by using a tabu search algorithm (which was empirically shown to outperform the previous approaches in the literature) and use the results as input to the scheduling problem.

As underlined in survey [12], not much research has been devoted to the staffing of ambulances. Hence we take a look at the research done in other areas, for instance scheduling beds for hospital wards, or agents to a call center.

For the Erlang delay models applicable to call centers, the square root safety staffing rule is an excellent rule to determine the number of neccessary agents under the constraint that the delay probability stays below a certain value [37]. For different values of the load of the system, the delay probability can be held more or less constant. When the rule is applied to Erlang loss models, this property is lost but the square root safety staffing is still of great interest [4]. Since the ambulance requests can be described by such an Erlang loss model [30], we can make use of the staffing rule.

When the arrivals of ambulance calls fluctuate over time by a predictable pattern, modified offered load approximations based on Erlang delay models can be used to apply the square

root safety staffing rule [4].

Where to locate ambulance bases?
To guarantee required service levels, such as arriving at the emergency location within 15 minutes, bases should be placed at optimal locations.

The performance of the St. Johns Ambulance Service (Auckland region) in New Zealand is investigated in [18] and [19]. A preliminary study using queueing theory established that more ambulances were needed to achieve performance targets. However, the required assumptions in the queueing model were such that a more realistic modeling was necessary. To this end, they developed a simulation tool, BartSim for St. Johns, that observes performance. Based on plots, generated by BartSim, areas with both a poor overall on-time performance and a large number of calls can be located. Such areas are good candidates for extra ambulance resources.

Spatial Poisson processes were used for a problem of facility location and dividing a city into districts [23]. The travel distance was used as a measure of performance. An upper bound is generated by assuming that facilities are distributed as a homogeneous spatial Poisson process. To achieve a lower bound, implying minimal travel distance, the facilities should be positioned in a regular lattice.

In [1] an extension to models for the Maximal Covering Location Problem (MCLP) is applied to the Saudi Arabian Red Crescent Society (SARCS), Riyadh City, Saudi Arabia. The purpose is to identify the optimal locations of emergency medical service (EMS) stations. This is achieved by first locating these stations such that the maximum expected demand may be reached within a pre-specified target time. Then, ensuring that any demand located within the target time will find at least one ambulance available. The demand rates are identified when it is necessary to add an ambulance in order to maintain the performance level for the availability of ambulances.

What number and type of vehicles should be deployed at each base?
Static ambulance deployment problems have received a great deal of attention in the literature. We present a brief survey to give a feel for the primary approaches. For more detailed surveys, we recommend the excellent reviews [6], [12], [13] and [35].

In [6] a review is given of ambulance location and relocation models proposed over the last 30 years. The three main approaches adopted use queueing models, mathematical programming, and simulation. The models are classified in two main categories. Deterministic models are used at the planning stage and ignore stochastic effects on the availability of ambulances. While probabilistic models take into account that ambulances operate as servers

in a queueing system. And in addition more recent dynamic models are mentioned, which repeatedly relocate ambulances throughout the day.

The emergency ambulance deployment on the Caribbean island Barbados is investigated in [16]. A multi-objective facility location model is used; maximizing the population covered within some standard of response and with some desired level of reliability, while minimizing cost of covering the population. The solutions obtained from the optimization were further analyzed by simulation.

An ambulance location optimization model is presented to maximize system-wide expected coverage in [20]. To determine an expected response time, in addition to modeling the uncertainty in the delay and in the travel time, uncertainty in the ambulance availability (pre-travel delay) is incorporated. Based on data of three real world ambulance location projects, one for the city of Edmonton, a second for the city of Calgary and a third for the city of St. Albert, all in Alberta, Canada, the latter is important and highly variable.

The call center assigning ambulances to medical emergencies of the area of Amsterdam, the Netherlands, requested an investigation of their targets [24]. To investigate the availability of ambulances a queueing model with different priorities is used. One of the conclusions was that it is quite possible to schedule the transportation rides of patients. This resulted in the recommendation of scheduling a small number of ambulances with less medical equipment instead of regular ones, to decrease costs. Another observation was that in the late afternoon much transportation rides for patients heading home are demanded, hence during those hours, instead of turning back to the base after a ride, ambulances should stay at one of the hospitals.

The ambulance deployment of an emergency medical system on a Brazilian highway, connecting the cities Sao Paulo and Rio de Janeiro is analyzed in [26]. To evaluate the systems performance, they focus on the mean response time of the system to an emergency call, the so-called hypercube model is used. The hypercube model is an effective tool for planning server-to-customer systems. The model expands the state space description of a simple multi-server queueing system; it also considers geographical and temporal complexities of the region.

The simulation tool BartSim for St. Johns, can assist in determining how to deploy its ambulances and staff to the various stations around Auckland [18]. The travel times predicted by their model are deterministic. It is stated that randomness in travel times can have a material effect on the predictions of a model, so this is an area they are beginning to investigate. BartSim has evolved into a more powerful system, Siren [19]. Its enhancements include among other things call generation using non-homogeneous Poisson processes and stochastic travel times.

How to decide when to dispatch which ambulance to an emergency?
The ambulance dispatch problem is to choose which ambulance to send to a patient. The dynamic ambulance relocation problem occurs in the operational control of ambulances. The objective is to find new locations for some of the ambulances, to increase the preparedness in the area of responsibility.

The complexities associated with dispatch decisions and strategies used to continually perform the dispatch task effectively are discussed in [17]. The complexities of dispatching were established after observing and interviewing fourteen ambulance command and control dispatchers located in two communications centers in New Zealand.

During the development of the London Ambulance Service Computer Aided Despatch system failures occurred at various stages. Whether responsibility models can be applied to prevent such failures is discussed in [8]. Questions as what types of responsibilities are considered to be important, where responsibilities within socio-technical systems should be located and when and where responsibility modeling should be applied.

The Ambulance Service St. Johns in New Zealand was considering the use of a dispatching strategy which would better match the necessary skills of the staff at an accident scene, and would result in more cases to be classified as priority 2 calls, based on an improved data collection [18]. These changes were built in the simulation tool, BartSim. The analysis played a large role in determining whether the proposed system would be adopted.

In [2] ambulance logistics which uses the Swedish public service enterprise SOS Alarm as a basis, is discussed. Described in detail is work on dispatching, deciding which ambulance to assign to each call, and relocating, deciding to relocate an ambulance if the dispatcher believes that there exists a location where the ambulance is more likely to be close to a new call. The main improvement lies in the development of decision support tools for the ambulance dispatchers. The tools developed are a preparedness calculator, an ambulance dispatch tool, a relocation tool and a simulation tool.

How to decide when to redeploy vehicles as a function of the system state?
Dynamic redeployment concerns the real-time relocation of idle ambulances so as to ensure better preparedness. The ambulance relocation problem is solved at each instant a call is registered [6].

Approximate dynamic programming on a dynamic program with a high-dimensional and uncountable state space is discussed in [29]. By computational experiments is shown that

the relocation policies obtained from this approach, relative to standard static-relocation policies, significantly improve performance.

A dynamic ambulance dispatching and redeployment system to assist realtime decision-making is developed and described in [11]. The main feature of this system lies in the precomputation of redeployment scenarios that allow immediate decision making when calls are received. It includes, for instance, constraints on the number of ambulances on each site, moving the same ambulances repeatedly and avoiding round and long trips. This system includes a parallel tabu search heuristic to precompute redeployment scenarios. By use of data of the emergency medical system of Montreal, computational results show that the proposed system can effectively solve real-life instances.

Chapter 3

Emergency Call Volumes and Traffic Patterns: Data Analysis

In this chapter we will investigate data provided to us by an Emergency Medical Service of Amsterdam, the capital of the Netherlands. We will start by explaining the dispatching process of an ambulance, and the status of the ambulance during a ride in Section 3.1. The data used for our research will be discussed in Section 3.2, and some preliminary numbers and observations of the number of executed rides and their duration time. In Section 3.3 we focus on the number of ambulance rides per day and per hour, and check for yearly, weekly and daily patterns in the ambulance requests. The provided results are used in Section 3.4 to obtain a non-homogeneous Poisson process describing the number of emergency call request per day, and a multinomial distribution describing the intra-day pattern of ambulance rides. The typical behavior of ride durations of the ambulances are investigated in Section 3.5, and we categorize the significantly alike empirical densities by weekdays and hours.

Questions we will address are: 'Does the data of the number of ambulance requests show correlations, typical yearly, weekly and daily arrival patterns?', 'Does the data provide reason to distinguish between ambulance rides of different priority?', 'Which factors are of influence when modeling a non-homogeneous Poisson process to the number of requested ambulances?', 'Does the data of the occupancy time of the ambulances show correlations, typical yearly, weekly and daily occupancy time patterns?' and 'Does the data provide reason to distinguish between ambulance rides of different priority?'.

3.1 Dispatching Process

In Amsterdam different ambulance service providers operate, yet all medical emergency calls in the area are handled by the same call center. The call center operator provides the first assistence to the caller, records the medical emergency, the address of the emergency, and

determines the priority of the emergency request:

A1	high priority, a life threatening situation is assumed. The ambulance responds as quickly as possible, and is allowed to make use of lights and sirens. As a standard, the ambulance should arrive at the emergency location within 15 minutes.
A2	high priority, no life threatening situation is assumed. The ambulance is not allowed to use light signals and should abide the standard speed limits. As a standard, the ambulance should arrive at the emergency location within 30 minutes.
B	low priority, most rides are scheduled, for instance, to provide first aid during big events, or to transport patients to or from hospitals.

When the call is denoted as a A1 or A2 call, the operator makes contact with the closest free ambulance, and dispatches this ambulance to the emergency. During every ride six different status values are logged by the ambulance crew: The departure of the ambulance after the instruction, its arrival at the patient, the time the ambulance leaves with the patient, the arrival at the patient's destination, the time the ambulance is called free, and the time it arrives at its base. At status two the ambulance can be at the scene of an emergency, a hospital (to transport a patient to home or another hospital), or a patient's home (to transporting a patient to the hospital for a procedure). When an ambulance only provides first aid, and no patient is transported, the third and fourth status values are of no interest. A graphic representation of the different status values of an ambulance ride is given in Figure 3.1.

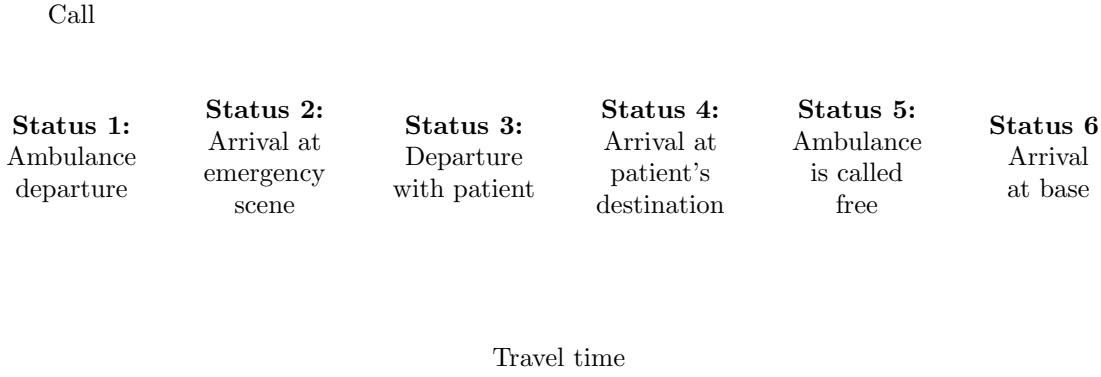

Figure 3.1: *Chart flow of the status values registered by the ambulance crew during a ride.*

When planning the number of ambulances one needs to take into account the time an ambulance is occupied. Throughout this thesis we will use the term travel time to refer to the time an ambulance is occupied during a ride. In Figure 3.1 the travel time is explained by the status values kept by ambulance drivers during a ride.

3.2 Data Set

In this thesis we focus on the planning of the ambulance rides dispatched to the GGD (Geneeskundige en Gezondheidsdienst) Amsterdam, one of the ambulance service providers in Amsterdam. In Amsterdam different emergency medical services provide their services, differences between these ambulance services are investigated in [24]. The GGD uses one station where all its ambulances are based, and takes emergency calls in the Amsterdam area, but also transports people to a remote location, for instance, because of a medical procedure.

The data we use from each call are the priority of the call, the starting time of the ride (status 1) and the time the ambulance is called free (status 5), since in between these two times the ambulance cannot respond to incoming emergency calls.

We will use a part of the data set, the first of January of 2008 until 31st of August in 2008, as a test set to validate models and forecasts.

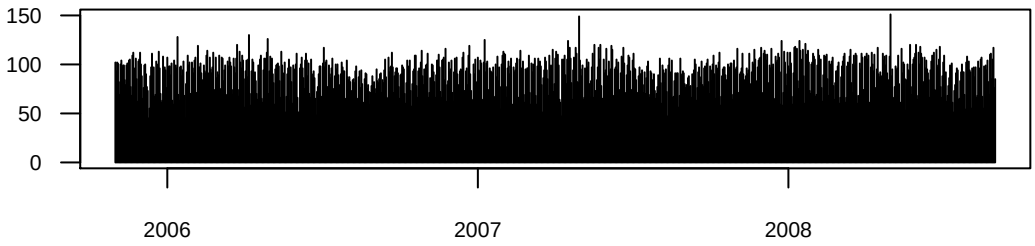

Figure 3.2: *Number of daily ambulance rides throughout our data set.*

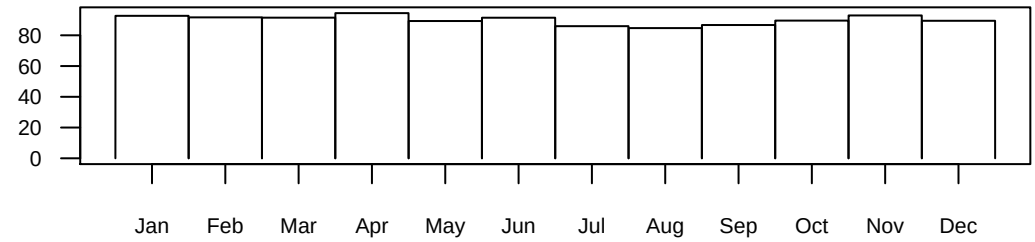

Figure 3.3: *Mean number of ambulance rides per day, for each month of the year.*

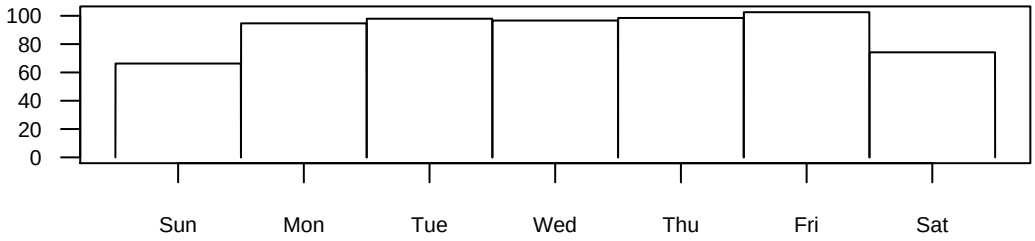

Figure 3.4: *Mean number of ambulance rides per day, for each day of the week.*

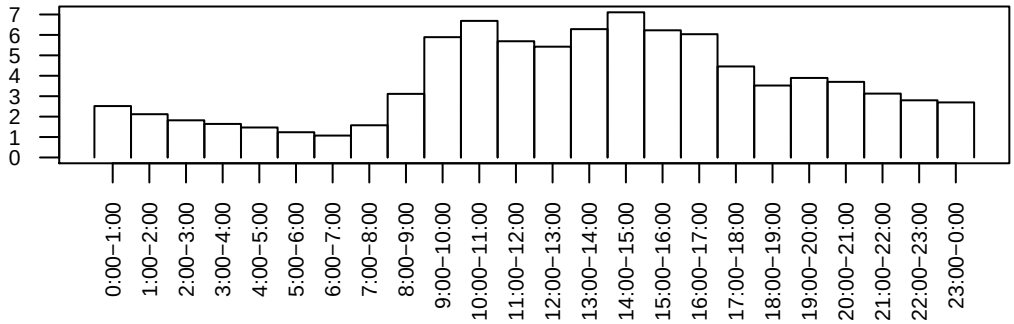

Figure 3.5: *Mean number of ambulance rides per day, for each hour of the day.*

In 2006 and 2007 approximately 32350 and 32900 ambulance rides were dispatched to the GGD respectively, about 90 per day. In Figure 3.2 we can see that *for the total number of rides per day there is no clear trend visible over time.* On April the 30th in 2007 and 2008, two clear peaks are visible. On those days the Dutch celebrate Queen's Day and the capital becomes crowded with party people. As can be expected also on the first of January a lot of emergency calls have been made. A closer look at Figure 3.3 shows that *larger daily volume can be seen in the months April and November, while July, Augustus and September are the most quiet months. On Sunday and Saturday are the least number of ambulance rides dispatched to the GGD. And on average the most rides occur between 9am to 5pm.* As explained in Section 3.1, the medical emergency calls are categorized in three levels of priorities. About 51% of the dispatched rides are categorized as A1 priority, about 10% are priority A2 and 38% were categorized as B rides.

In Figure 3.6 box plots are shown of the number of rides per day for the A1, A2 and B rides seperately. The mean of the number A1 of rides per day is 46.25, for the A2 rides we calculated a mean of 9.08 and for the B calls we found 34.77. The standard deviation of the number of A1 rides per day is 8.53, which indicates that *the number of A1 rides per day does not fluctuate that much. The standard deviation for the number of daily A2 rides is also low, but relative to the mean it is high*, that is about 3.29. This can result in high relative measures of the error of forecasts of the number of A2 rides per day. For the number

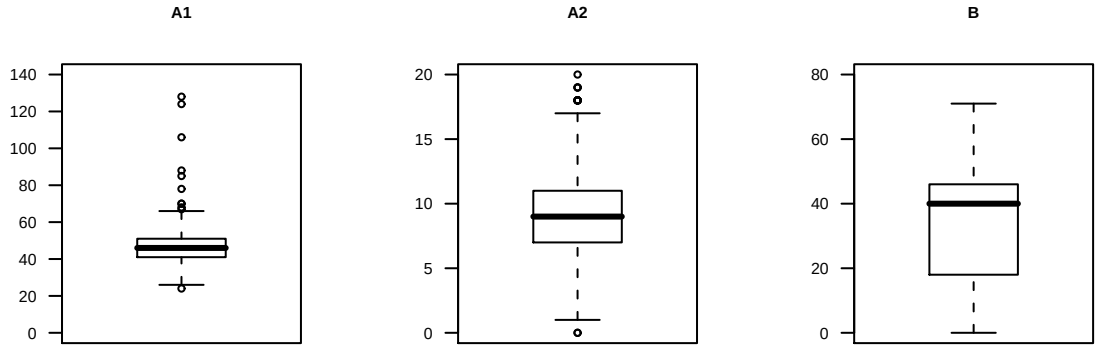

Figure 3.6: *Box plots of the number of daily ambulance rides for the different types of ambulance rides.*

of B rides per day we calculated a relatively high standard deviation of 15.13, which can be explained by the distinction in the daily number of B rides during the weekend versus weekdays. Since hospitals plan the most surgeries during weekdays and hence admit less patients during the weekend, *the number of B rides per day shows a clear difference between weekdays and weekend, while the A1 rides occur more frequently during the weekend.* This can explain the negative correlation of -0.19 between the daily A1 and daily B calls. While the A2 calls have a correlation of 0.18 with the B rides.

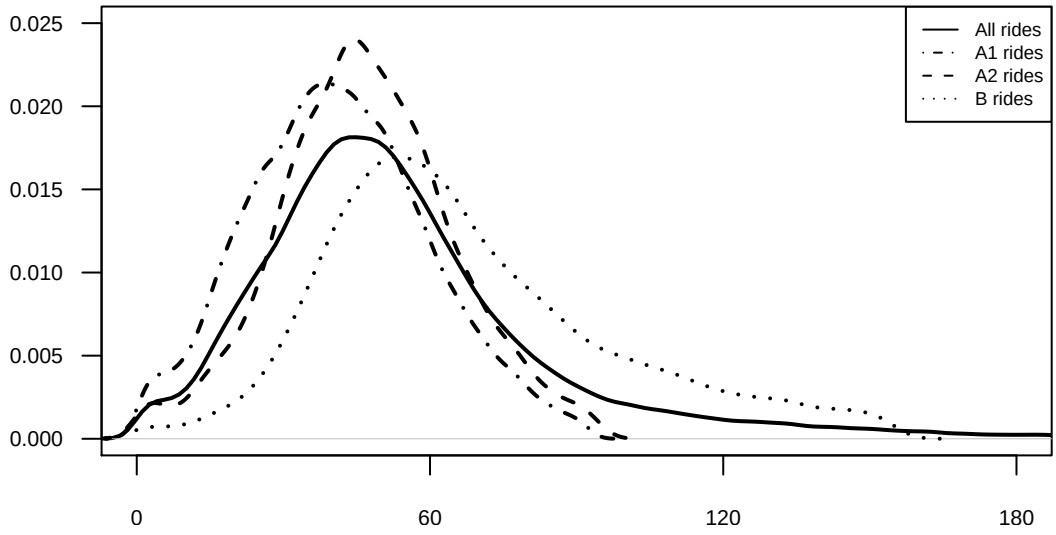

Figure 3.7: *Empirical density of the travel time of an ambulance ride.*

A first look at the density of the travel times of the ambulance rides dispatched to the GGD indicates some high travel times, see the right tail in Figure 3.7. Examples of those rides are scheduled ambulances for events which take on one day, or transportation rides to Groningen or Sittard. The peak of the density of the travel time of A1 and A2 rides is higher than that of the B rides, which indicates that *the length of the travel time of a B ride is*

more spread. The centre of the peak for the A1 rides is positioned before the peak of the A2 rides, which is positioned before the peak of the B rides. This indicates on average shorter travel times for A1 than A2 rides, and shorter travel times of the A2 rides compared to the B rides. *The travel times of ambulance rides during the weekend, and during hours of 16pm to 6am appear to be of shorter length*, see Figures 3.4 and 3.5. This can be explained by the occurance of less B rides during those days and hours, which have a longer travel time duration, and that during the weekend more ambulances are scheduled.

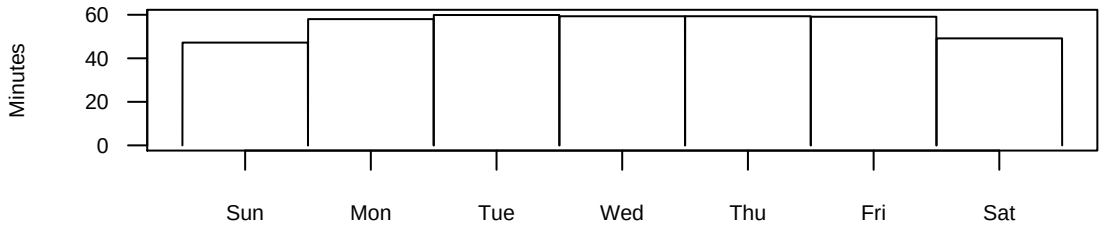

Figure 3.8: *Mean travel time of ambulance rides for each day of the week.*

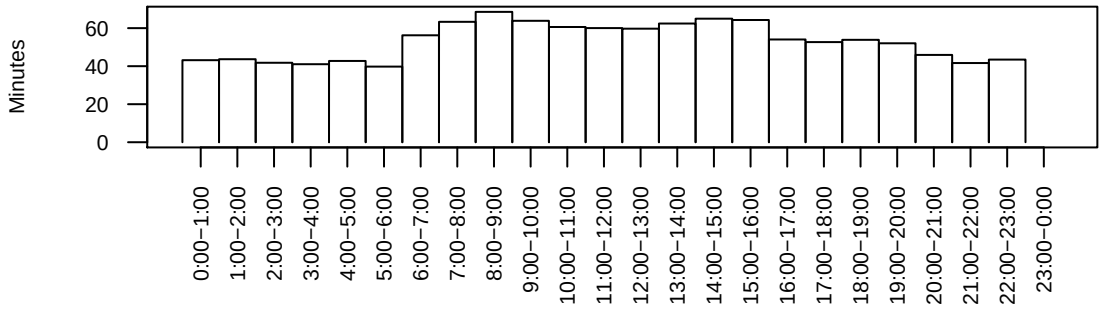

Figure 3.9: *Mean travel time of ambulance rides for each hour of the day.*

The length of the travel time of the ambulances differs for the three types of priorities. In Figure 3.10 box plots are shown of the travel times of A1, A2 and B rides seperately. The high priority calls, A1, have a mean of 42.9 minutes and a standard deviation of 21.6 minutes. The A2 rides, have a mean of 50.1 minutes and a standard deviation of 29.5 minutes. The ambulance needs to be at the emergency scene within fifteen minutes during an A1 ride, and within 30 minutes during an A2 ride. But as explained in Figure 3.1, the continuation of the ride can take different forms, which results in the relatively high standard deviations. The duration of B rides is a lot higher, their mean is 76.8 minutes with a standard deviation of 62.4 minutes, yet the median of the B calls is 63.7 minutes, more than 10 minutes lower than the mean. This can be explained by the fact that B rides can be aborted when a high priority request comes in.

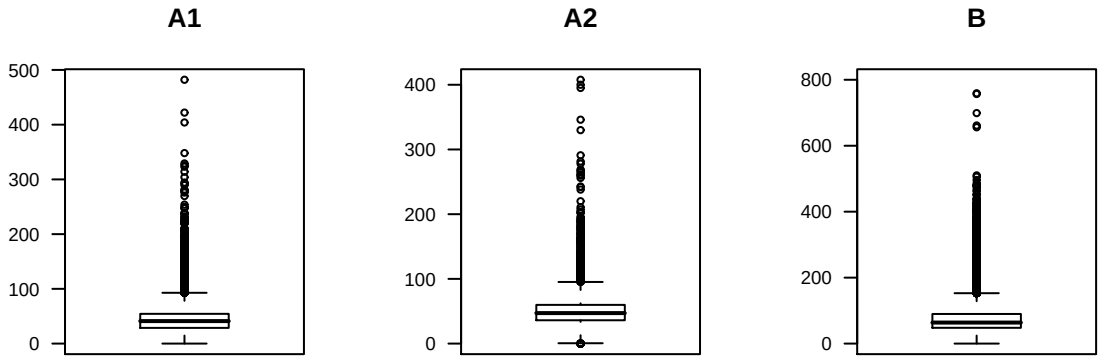

Figure 3.10: *Box plots of the travel times in minutes of the ambulance rides for the different call types.*

3.3 Daily and Hourly Numbers of Dispatched Rides

As already mentioned in Section 3.2, a different behavior of the number of daily rides can be seen over the year. In this section we will take a closer look at the data. By looking for extreme values, yearly and weekly patterns we are able to categorize the number of daily calls by their distinct behavior for the three types of priorities separately. Investigating the different daily patterns of the found categories will provide a daily pattern of the incoming calls per hour. We work with the number of daily and hourly calls instead of their time of occurence since we want to make use of time series in Chapter 4 to forecast the ambulance rides.

Figure 3.2 indicates that *on a number of days a large number of rides occurred*. To make a good analysis of the data we will not take these outliers, and extremely small numbers of daily rides, into account. We define a data point as an outlier when its value is lower than the first quantile minus 1.5 times the interquantile range, or when its value is higher than the third quantile plus 1.5 times the interquantile range. In Figure 3.6 the whiskers extend to 1.5 times the interquartile range from the box, and the outliers are given by dots. For type A1 rides, the outliers represent the dates: 2006-01-01, 2006-04-29, 2006-09-18, 2007-01-01, 2007-04-30, 2007-05-18, 2007-07-17, 2007-11-24, 2008-01-01. And for the A2 priority the dates with an extreme low or high number of calls are 2005-11-03, 2005-12-10, 2005-12-11, 2006-01-13, 2006-02-13, 2006-05-27, 2006-12-01, 2006-12-19, 2006-12-22, 2007-01-24, 2007-03-09, 2007-09-21, 2008-01-11, 2008-01-14, 2008-01-31. *The number of daily rides on some of these dates can be explained to be outliers*. For instance, in the early night of the dates 2006-01-01, 2007-01-01 and 2008-01-01 New Year's eve is celebrated. And on 2006-04-29 and 2007-04-30 people from everywhere in the Netherlands head of to the capital to celebrate Queen's Day.

3.3.1 Tests for Correlation

In this subsection, the numbers of ambulance rides per day which appeared to be outliers (see Page 21) were omitted from the analysis for correlations. Spearman's rho or Kendall's tau statistic [15] is used to estimate a rank-based measure of association between bivariate data defined as $(X_1, Y_1), \ldots, (X_n, Y_n)$. Both these tests are non-parametric and can be used if the data does not necessarily come from a bivariate normal distribution. For notation let $S_1, \ldots, S_n$ be the rank numbers of $X_1, \ldots, X_n$ of the ordered row $X_{(1)}, \ldots, X_{(n)}$ and $R_1, \ldots, R_n$ the rank numbers of $Y_1, \ldots, Y_n$ of the ordered row $Y_{(1)}, \ldots, Y_{(n)}$. The null and the alternative hypothesis for the two sided problem are given by:

$$H_0: \quad X_i \text{ and } Y_i \text{ are independent, for } i = 1, \ldots, n, \tag{3.1}$$

$$H_1: \quad X_i \text{ and } Y_i \text{ are dependent, for } i = 1, \ldots, n.$$

The Spearman's rho statistic is based on the (sample) coefficient of correlation of the two groups of rank numbers r_i and s_i. We reject the nullhypothesis for values of ρ close to -1 or 1; close to one indicates a positive correlation and a value close to -1 a negative.

$$\rho = \frac{n(\sum_{i=1}^n r_i s_i) - (\sum_{i=1}^n s_i)(\sum_{i=1}^n r_i)}{\sqrt{n(\sum_{i=1}^n r_i^2) - (\sum_{i=1}^n r_i)^2} - \sqrt{n(\sum_{i=1}^n s_i^2) - (\sum_{i=1}^n s_i)^2}}.$$

The Kendall's tau statistic measures the degree of correspondence between the two rankings and assesses the significance of this correspondence. The nullhypothesis is rejected for values of τ close to -1 or 1 again, close to one indicates a positive correlation and a value close to -1 a negative. Since we observed ties in the data, Kendall's tau statistic is defined as:

$$\tau = \frac{n_c - n_d}{\sqrt{\left(\frac{1}{2}n(n-1) - \sum_{i=1}^n \frac{1}{2}r_i(r_i-1)\right)\left(\frac{1}{2}n(n-1) - \sum_{i=1}^n \frac{1}{2}s_i(s_i-1)\right)}},$$

where n_c denotes the number of concordant pairs of the ranked data, and n_d the number of discordant pairs.

After performing tests based on Spearman's rho or Kendall's tau statistic we can conclude that *the number of type A1 calls per day and the number of A2 calls per day are uncorrelated but the type B calls are negatively correlated to the number of A1 rides per day* (both p-values smaller than 0.001) *but positively to the A2 rides* (both p-values smaller than 0.0001). This can be explained by the difference in number of rides during the weekend and weekdays; during the weekend more A1 rides occur compared to the weekdays, on the contrary less A2 and less B rides occur during the weekend. Even though there appears to be a correlation we will investigate the three types of priorities separately since the travel times of B rides are much higher than those of the A1 and A2 calls, see Figures 3.7 and 3.10.

The number of rides per day is for all three types of priorities positively correlated with the number of rides during the previous day . The Spearman's rank correlation test generates p-values smaller than 0.0001 for A1 rides, 0.036 for the A2 priority and a p-value less than 0.0001 for the ordered rides. The Kendall's rank correlation test returns p-values smaller than 0.0001, 0.037 and smaller than 0.0001 for the B calls. *The number of type A1 rides per day appears to be positively correlated to the number of rides two days earlier*, the rank correlation test of Spearman provides a p-value of 0.021 and Kendall's rank correlation test returns a p-value of 0.020. *The number A2 rides per day appear not to be correlated to the number of rides two days earlier*, the tests returns a rho of $\rho = 0.022$ and a tau of $\tau = 0.016$. We can state that *the number of B rides per day are significantly negatively correlated to the number of ordered rides two days earlier* since the both tests return p-values smaller than 0.0001.

Box plots of the number of rides per day of the week, for each of the ambulance priorities seperately, are shown in Figure 3.11. The plots suggest the existence of a weekly pattern, especially for the B rides. A question one can ask is, are the number of calls per day correlated to the number of calls seven days before? Again we used the rank correlation tests of Spearman and Kendall to verify this observation. *A positive correlation between the number of rides per day with the number of rides seven days ago, is determined for the A1 and A2 rides.* The corresponding p-values are given by 0.0002 and 0.0002, respectively 0.002 and 0.002. *For the B rides the positive correlation is also significant* since the p-values for both tests are lower than 0.0001.

In the box plots of Figure 3.12 some outliers can be detected, skewness [25], since some of the boxes are asymmetrical, and for some months the variances are unequal since the boxes have a different size.

The rank correlation tests of Spearman and Kendall can only be used on paired data. To check whether the number of calls per day during weekend versus weekdays is of significant difference, and for which type of calls, we use the Wilcoxon-two-sample or Mann-Whitney test [15]. Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ denote the daily number of calls during weekdays, respectively, during weekends. Suppose that X_i and X_j have distributions F and G, respectively. The null and alternative hypothesis are given by:

$$H_0 : \quad F = G, \tag{3.2}$$

$$H_1 : \quad F \neq G.$$

Let $R_1, \ldots, R_n$ denote the ranks of the X_i in the combined sample. The Mann-Whitney

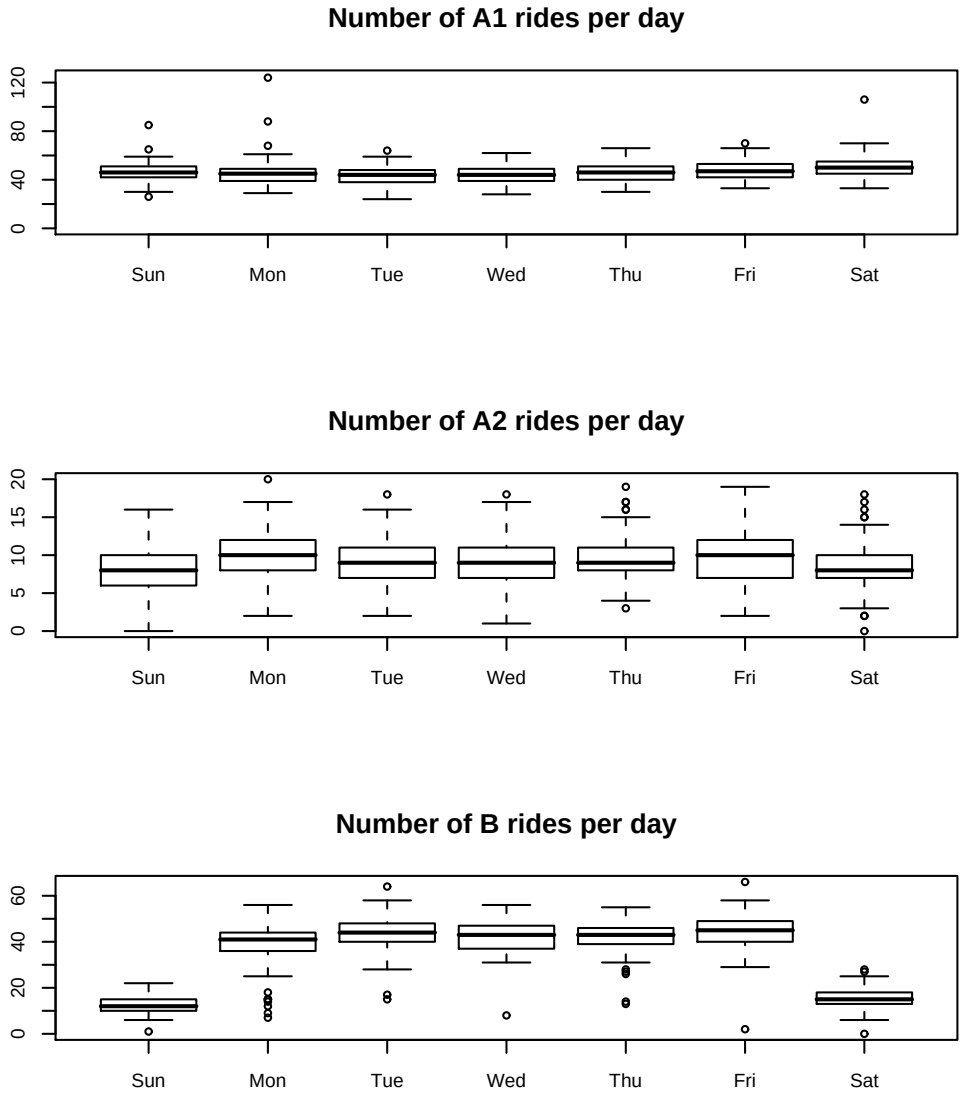

Figure 3.11: *Box plots of the number of A1, A2, and B rides per day for every day of the week separately.*

statistic is defined as:

$$U = \sum_{i=1}^{n} R_i - \min\{R_i\}.$$

The A1 calls seem to occur more often during the weekend than during weekdays, hence we have as alternative hypothesis that the number of weekend calls is larger. The p-value of the test is lower than 0.0001, hence *the number of A1 rides per day is significantly higher during the weekend, whereas priority A2 and B rides intend to appear less during weekends*, this difference is significant with a p-values smaller than 0.0001 and for the ordered rides smaller than 0.0001. We can conclude that for all types of priorities there is a significantly difference

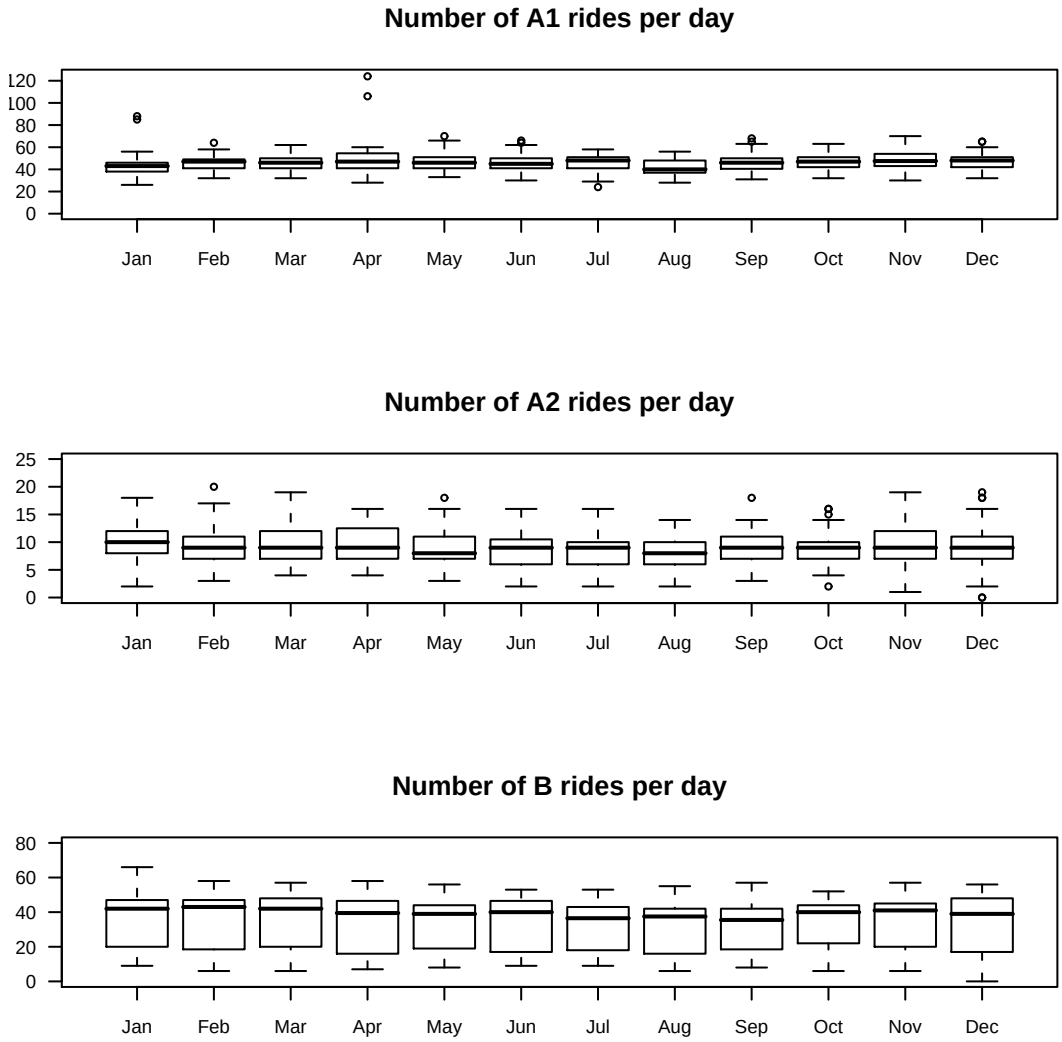

Figure 3.12: *Box plots of the number of A1, A2, and B rides per day for every month separately.*

between the number of incoming calls for weekdays and the weekend.

3.3.2 Pattern over the Day

Not only does a pattern over the week and over the year for the dispatched ambulance rides exist, Figure 3.13 suggests there exists also a pattern over the day, the vertical dotted lines denote the transition between successive days. For instance the A1 rides seem to appear more frequent from the hours 9 am until 6 pm and during the early night on Saturday and Sunday. During the night less B rides per hour occur than during daytime, and two clear peaks are visible around 10 am and 3 pm; around that time the hospitals transmit their patients.

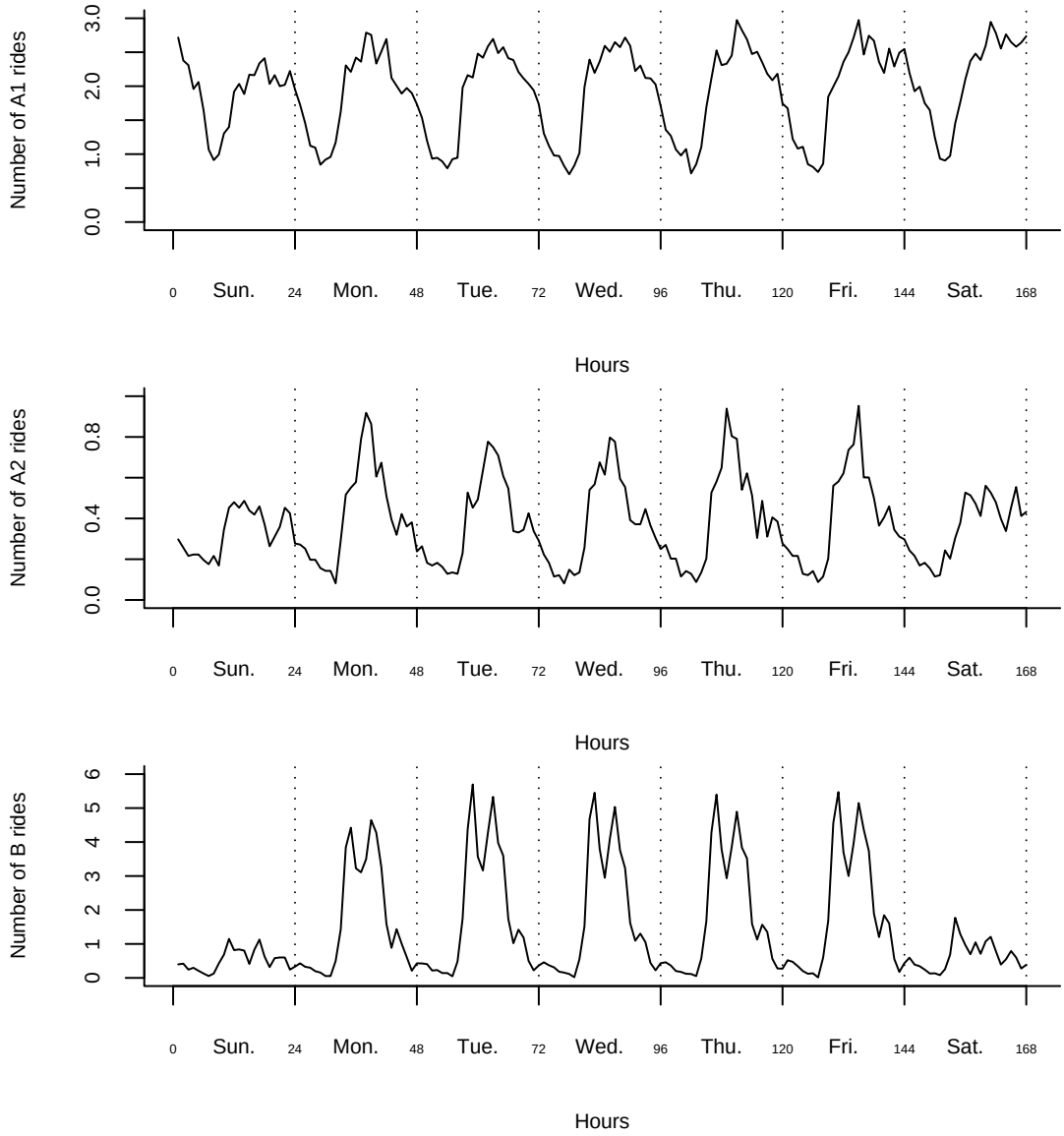

Figure 3.13: *Mean hourly volume on every day of the week for the priority A1, A2, and B rides separately.*

In Section 3.5 we will discuss modeling the number of daily rides over the day itself, by use of the multinomial distribution conditional on the total rides that day [7]. Since some days of the week show a different mean hourly volume (Figure 3.13), days of the week which have the same daily pattern $(p_1, p_2, \ldots, p_{24})$ will be grouped in one category. To analyze the relationship between the number of rides per hour and the day of the week, we will use contingency tables [15]. We counted the total number of executed rides of every day of the week separately, in every daily hour from the first of November 2005 until the 31st of December of 2007 (113 weeks) and denote them by N_{ij}, for $i \in \{1, \ldots, 7\}$ and $j \in \{1, \ldots, 24\}$. Hence we obtain a $k \times r$ contingency table with $k = 7$ independent samples from a $r = 24$

nomial distribution, where the ith sample has parameters $N_{i\cdot}, p_{i1}, \ldots, p_{ir}$. We assume

$$\sum_{i=1}^{k} N_{i\cdot} = \sum_{j=1}^{r} N_{\cdot j} = N_{\cdot\cdot},$$

where $N_{1\cdot}, \ldots, N_{k\cdot}$ are known. We demand that under this model we have

$$p_{i\cdot} = \sum_{j=1}^{r} p_{ij} = 1, \quad i = 1, \ldots, k.$$

Since we are interested in the intra-day pattern of the ambulance rides and whether these are distinct for different days of the week we test the following hypothesis:

$$H_0: \quad p_{1j} = p_{2j} = \cdots = p_{kj} \equiv p_j, \quad j = 1, \ldots, r = 24. \tag{3.3}$$

The test statistic we use to determine whether we can reject this hypothesis is given by

$$X^2 = \sum_{i=1}^{7} \sum_{j=1}^{24} \frac{(N_{ij} - \frac{1}{n} N_{i\cdot} N_{\cdot j})^2}{\frac{1}{n} N_{i\cdot} N_{\cdot j}}. \tag{3.4}$$

For large enough $N_{i\cdot}$, X^2 is distributed as a χ^2 distribution with $(k-1)(r-1)$ degrees of freedom. The application of this distribution is reliable when under the nullhypothesis the expected N_{ij} are larger than 1 for all (i, j), and at least 80% is larger than 5, which is the case.

First we will investigate the effect of the day of the week on the daily pattern of the A1 rides. As Figure 3.13 suggests the daily pattern for the A1 rides is not the same for every day of the week. The test statistic X^2 has value 1133.6, the number of degrees of freedom is $(7-1)(24-1) = 69$, resulting in a p-value smaller than 0.0001. *The number of A1 rides per hour over the day appear to be alike on Monday, Tuesday, Wednesday and Thursday.* The statistic in (3.4) now has value 80.9, the number of degrees of freedom is $(4-1)(24-1) = 69$, hence we obtain a p-value of 0.15 and cannot reject the hypothesis (3.3) of having the same distribution. Comparing the daily patterns of Sunday, Friday and Saturday separately with the daily pattern of the Monday-Thursday gives us p-values of 0.0001, 0.0001, and 0.0002 respectively. *Comparing the hourly patterns of A1 rides on Sunday, Friday and Saturday against each other leads to the conclusion that each of these days has a distinct daily pattern.* Hence we group the days of the week, according to the daily pattern of the A1 rides, as $g_1^{(\mathrm{A1})}$, denoting Sunday, $g_2^{(\mathrm{A1})}$ containing Monday-Thursday, $g_3^{(\mathrm{A1})}$ denotes Friday and $g_4^{(\mathrm{A1})}$ Saturday.

When we look at the daily patterns of the A2 rides, again we use the statistic mentioned in (3.4) to group the days of the week in categories. Taking Monday, Tuesday, Wednesday,

Thursday and Friday as g_1 we obtain $X^2 = 83.31$, having $(5-1)(24-1) = 138$ degrees of freedom we obtain a p-value of 0.730. Analyzing the behavior of the hourly rides on Saturday and Sunday we obtain a p-value of 0.321. *Hence the daily patterns of the A2 rides can be grouped in two categories,* $g_1^{(A2)}$ denoting the *Monday to Friday and* $g_2^{(A2)}$ indicating *Saturday and Sunday.*

The number of requested B rides per day, show a significantly difference in the daily patterns between Tuesday to Friday and the other days of the week separately. The daily pattern of Sunday behaves not like those of Monday and Saturday (we found both p-values to be smaller than 0.001) and Monday is also significantly different from Saturday. Hence we obtained four groups for the B rides. The first $g_1^{(B)}$, denotes the daily pattern on Sunday specifically, the second $g_2^{(B)}$ denotes Monday, then we get $g_3^{(B)}$ containing Tuesday to Friday and the daily pattern of the B rides on Saturday is denoted by $g_4^{(B)}$.

3.4 Modeling the Hourly and Daily Rides

Literature on modeling emergency call arrivals provide compelling theoretical reasons to assume that the number of emergency call arrivals per day follows a non-homogeneous Poisson process, see for instance [7], [26] and [34]. To investigate whether the number of ambulance rides on day t behaves as such a process we need to determine the corresponding piecewise constant parameter μ_t. Its value can be influenced by the day of the week, the month of the year and being it a certain holiday. After obtaining the estimation of μ_t we will investigate whether our data actually behaves like a non-homogeneous Poisson process.

As seen in the previous section the demand for ambulances varies at different hours of the day, for different days of the week. We make use of a multinomial distribution, conditional on the parameter μ_t, to describe the demand per hour [7].

3.4.1 A Non-homogeneous Poisson Process Describing the Number of Ambulance Rides per Day

In Figures 3.11 and 3.12 box plots are provided of the number of daily rides per day of the week and month of the year, respectively, per priority. The A1 rides seem to have a higher daily number during the weekend, whereas the A2 and B rides seem to be executed less during the weekends. For the A1 rides the most quiet month appears to be August, while April appears to be the busiest. In January and April more A2 rides have been performed per day than in the other months, and again August has been the most quiet month. The number of daily B rides appears to be the highest during the first three months of the year, whereas July, August and September are the months with the lowest mean number of daily

B rides.

To determine the parameter μ_t for the non-homogeneous Poisson process we could test which day of the week, which month of the year, and which holidays show a significant high or low number of rides that day, and categorize these days of the week and months of the year for the different priorities. Then we could take the mean of the number of rides per day for each of the categories. The problem is that the categories influence each other, the busy weekends for the A1 rides could be less busy during August. By taking $12 \times 7 \times K$ categories, where K is the number of significant holidays, we could end up with taking the mean of 2 days. Hence we need to make use of a model which takes all categories into account altogether. Multiple linear regression does that. In Section 4.2.3 we will discuss a multiple linear regression model to forecast the number of daily rides:

$$Y_t = a + \sum_{i=1}^{12} b_i B_{i,t} + \sum_{j=1}^{7} c_j C_{j,t} + \sum_{k=1}^{K} d_k D_{k,t} + \epsilon_t, \tag{3.5}$$

where Y_t is the t-th observation of the number of daily rides, indicator $B_{i,t}$ has value 1 if the month of day t is the ith month of the year and value 0 otherwise, the indicator $C_{j,t}$ has value 1 if day t is the jth day of the week and 0 otherwise, the indicator $D_{k,t}$ has value 1 if day t is the kth significant holiday and 0 otherwise.

In this regression model *the significant factors for the parameters month of the year* $b_1, \ldots, b_{12}$, *and day of the week* $c_1, \ldots, c_7$, (with a significance level of 0.05) *will be for the A1 rides: January, April, August, November, Tuesday, Wednesday, Friday and Saturday.* The parameters d_1, d_3 and d_6 (*January-the-first, Queen's Day and Christmas Day*) denoting holiday effects obtained a p-value lower than 0.05. *The factors that are significantly for the regression model describing the number A2 requests are: the month August, days of the week Sunday, Monday, Friday, Saturday and the holidays Easter Monday and Whit Monday* (d_2 and d_5) are also of significantly influence on the number of A2 rides per day. *The months that are of influence on the number of B requests per day are: January, April, July and August. The numbers of daily B rides differ significantly over the week since all factors of the day-of-the-week effect are significantly. All the holidays we took into account in our regression model appeared to be significant for the B rides*, these are: January the first (denoted by parameter d_1), Easter Monday (d_2), Queen's Day (d_3), Ascension Day (d_4), Whit Monday (d_5), Christmas Day (d_6) and Boxing Day (d_7).

Let t denote a specific day, the value of the parameter μ_t for the non-homogeneous Poisson process is then given as a regular mean (the intercept in the regression model) plus or minus the significantly day-of-the-week, month and possible holiday effects. The estimates of the effects determining the parameter μ_t are given in Tables 3.1, 3.2 and 3.3. The results

show that for instance the factor Saturday causes an increase of 4.41 rides per day on the intercept, whereas being it August results in a decrease of 4.37 rides a day. For the B rides the general mean, a has value 34.90, the Wednesday effect causes an increase of 7.87, and days in the month April get an increase to the number of B rides per day with 2.01. When t denotes a Queen's Day which is on a Wednesday we would obtain as an estimate of μ_t: $34.90 + 7.87 + 2.08 - 20.49 = 24.36$.

Parameter:	a	b_1	b_4	b_8	b_{11}		
	General mean	Jan.	Apr.	Aug.	Nov.		
Estimate:	45.78	-4.05	0.96	-4.37	1.90		
Parameter:	c_3	c_4	c_6	c_7	d_1	d_3	d_6
	Tue.	Wed.	Fri.	Sat	Jan.1	Q.day	Chr. Day
Estimate:	-1.74	-1.44	1.81	4.41	44.77	66.01	-8.53

Table 3.1: *Significant effects on the mean of the non-homogeneous Poisson proces describing the number of daily A1 rides.*

Parameter:	a	b_8	c_1	c_2	c_6	c_7	d_2	d_5
	General mean	Aug.	Sun.	Mon.	Fri.	Sat.	Easter Mon.	Whit Mon.
Estimate:	9.36	-1.02	-1.38	0.89	0.47	-0.91	-4.25	-7.25

Table 3.2: *Significant effects on the mean of the non-homogeneous Poisson proces describing the number of daily A2 rides.*

Parameter:	a	b_1	b_4	b_7	b_8		
	Gen. mean	Jan.	Apr.	Jul.	Aug.		
Estimate:	34.90	1.82	2.08	-2.29	-3.58		
Parameter:	c_1	c_2	c_3	c_4	c_5	c_6	c_7
	Sun.	Mon.	Tue.	Wed.	Thu.	Fri.	Sat.
Estimate:	-22.16	6.24	9.44	7.87	7.96	9.87	-19.22
Parameter:	d_1	d_2	d_3	d_4	d_5	d_6	d_7
	Jan.1	Easter Mon.	Q.day	Asc.Day	Whit Mon.	Chr.Day	Box.Day
Estimate:	-16.75	-30.71	-20.49	-29.36	-27.64	-16.41	-30.42

Table 3.3: *Significant effects on the mean of the non-homogeneous Poisson proces describing the number of daily B rides.*

To check whether our data set is Poisson(μ_t) distributed we compare an overall density P_{overall} to our data set. For days t in November 1, 2005 to August 31, 2008 we calculate

μ_t by use of the effects given in Tables 3.1, 3.2 and 3.3, and determine the corresponding Poisson(μ_t) distribution. As the overall density we now take the mean of all these densities. Since the Poisson distribution takes only integer values we make use of contingency tables, comparing the expected number of days with n rides to the number of observed days with n rides.

$$H_0: \quad p_{1j} = p_{2j} = \cdots = p_{kj} \equiv p_j, \quad j = 1, \ldots, r = 24. \tag{3.6}$$

For the A1, A2 and B rides we found p-values of 0.7612, 0.2848 and 0.6948, respectively. Hence *for the three types of priorities we cannot reject the nullhypothesis of having the Poisson density with time-dependent arrival rate.*

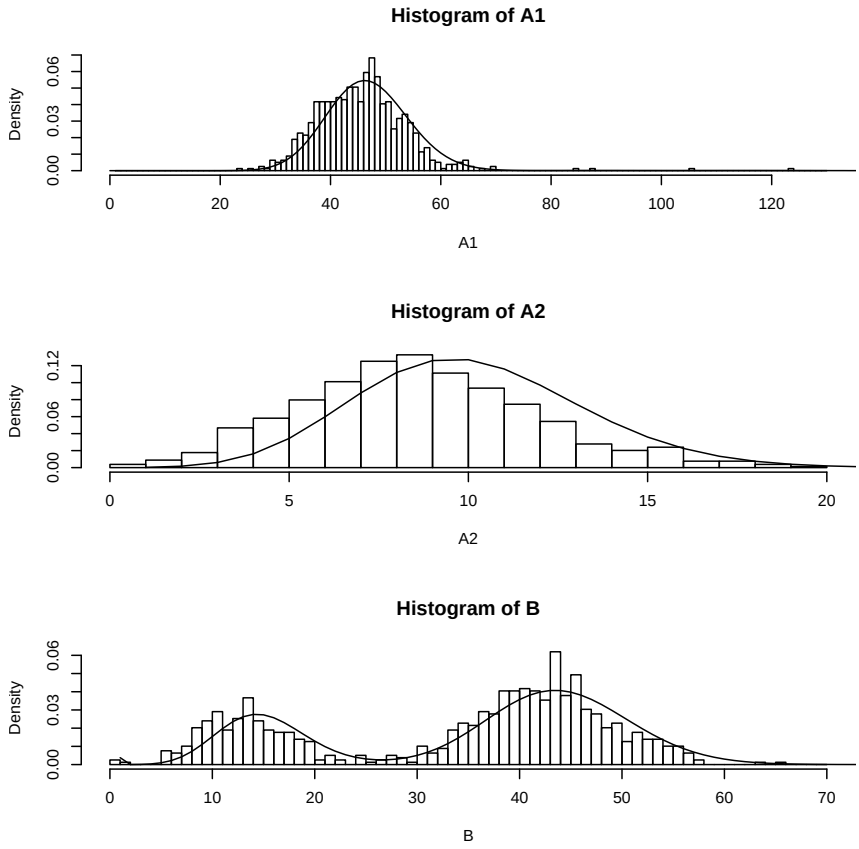

Figure 3.14: *Histogram of the number of rides per day, and the overall density of the Poisson model, for each of the ambulance priorities seperately.*

We obtained a non-homogeneous Poisson model with piecewise constant mean to describe the incoming calls per day for the three types of priorities separately, by summing a general mean with effects depending on the day. In Chapter 4 we will forecast the daily number of rides according to this model and compare with models found in the literature.

3.4.2 Intra-day Pattern

Since we would like to obtain staffing levels of the number of ambulances required per hour, we also need to model the incoming rides on the day itself, the daily pattern. In our research we assume that the rides which started during a certain hour follow a Poisson process and the rate of this Poisson process remains constant in this hour. We assume that for day t the distribution of the number of rides per hour $X_{t,i}$ conditional on $X_{t\cdot}$, the total rides of day t, is independent of the number of rides on other days. A distribution that satisfies to this assumption is the multinomial distribution with parameters $(N, p_1, p_2, \ldots, p_{24})$ where $N = X_{t\cdot}$ and p_i the probability that a ride occurs in hour i.

In the previous section we categorized the intra-day pattern of the rides per day for the three types of priorities separately, according to the day of the week. For the A1 rides we defined categories: $g_1^{(\mathrm{A1})}$ Sunday, $g_2^{(\mathrm{A1})}$ Monday-Thursday, $g_3^{(\mathrm{A1})}$ Friday, and $g_4^{(\mathrm{A1})}$ Saturday. The A2 rides can be grouped in two categories: $g_1^{(\mathrm{A2})}$ Sunday and Saturday, and $g_2^{(\mathrm{A2})}$ Monday-Friday. The categories we found for the B rides are: $g_1^{(\mathrm{B})}$ Sunday, $g_2^{(\mathrm{B})}$ Monday, $g_3^{(\mathrm{B})}$ Tuesday-Friday, and $g_4^{(\mathrm{B})}$ Saturday. Since we found different categories we define $(p_{g,1}, p_{g,2}, \ldots, p_{g,24})$ as the daily pattern vector with $g \in \{(\mathrm{A1}_1), (\mathrm{A1}_2), (\mathrm{A1}_3), (\mathrm{A1}_4), (\mathrm{A2}_1), (\mathrm{A2}_2), (\mathrm{B}_1), (\mathrm{B}_2), (\mathrm{B}_3), (\mathrm{B}_4)\}$. For the three priorities the probability $p_{g,i}$ is estimated as the fraction of rides in category g during hour i, hence

$$\hat{p}_{g,i} = \frac{\sum_t X_{t,i} 1_{t\in g}}{\sum_t \sum_{j=1}^{24} X_{t,j} 1_{t\in g}}, \tag{3.7}$$

where t ranges from the first of November 2005 to the 31$^{\mathrm{st}}$ of December 2007. Again we omitted the days defined as outliers on Page 21. For every day of the week the estimated $\hat{p}_{g,i}$ are plotted in Figure 3.15. Thus when the total number of rides per day is provided, we can use the daily pattern vector $\hat{p}_{g,i}$ to spread these rides over the hours of that day.

In the next chapter we will use a non-homogeneous Poisson(μ_t) model to forecast the rides per day and we will compare this model with forecasting models found in the literature.

3.5 Data Analysis of the Travel Times

Eventually we would like to obtain a decent staffing method of the ambulances. Therefore we need to investigate the travel times of the ambulances; we need an estimate how long an ambulance will be occupied after its departure of the ambulance station. Checking whether there exist significant differences between the three priorities, days of the week and the hour of the departure will provide insight in the data. According to the distinct behaviors we will categorize the empirical densities to different hours of different days.

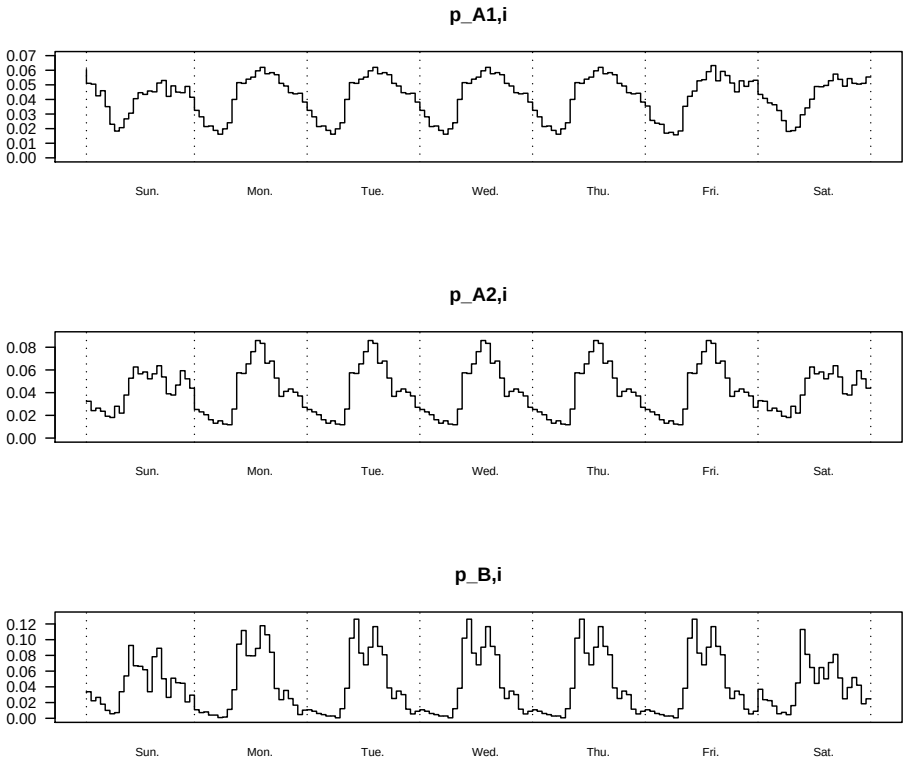

Figure 3.15: *The estimated daily pattern vector $\hat{p}_{g,i}$ given for each hour of the week, for each of the ambulance priorities separately.*

The data of the travel times contains some extreme values. *About 1 % of the A1 rides we came across has a ride duration of more than 100 minutes, and some of the ordered rides took more than six hours.* Examples of short rides are ride durations of less than a minute for a B ride; priority B rides can be aborted when a request which requires more urgent care is made. Also ride durations of more than six hours occurred; like the transport of a patient to a hospital in Groningen, or an ordered ambulance for a concert or a football match. Again we did not take such outliers into account in our data analysis.

The data of the travel times is continuous, hence we can use the Kolmogorov-Smirnov test [25] to determine whether the travel time distribution differs significantly for the three types of priorities, and if for each day of the week a different distribution exists. For notation, let $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$ have distributions F and G respectively, and consider the nullhypothesis :

$$H_0 : \quad F = G, \tag{3.8}$$

$$H_1 : \quad F \neq G.$$

Let $\hat{F}_m$ and $\hat{G}_n$, be the empirical distribution functions of $X_1, X_2, \ldots, X_m$, and $Y_1, Y_2, \ldots, Y_n$, respectively. The two-sample test statistic is given by:

$$D_{m,n} = \sup_{-\infty<x<\infty} |\hat{F}_m(x) - \hat{G}_n(x)|.$$

Rejection of the nullhypothesis (3.8) occurs when $D_{m,n}$ exceeds a critical value, which depends on the sample size and a chosen significance level.

As can be expected we found significant differences between the distributions of the travel times of the three priorities. *The A1 rides take a significantly shorter time to process than the A2 and B rides* (p-values smaller than 0.0001), *and the A2 rides take a significantly shorter time than the B rides*, we found a p-value smaller than 0.001.

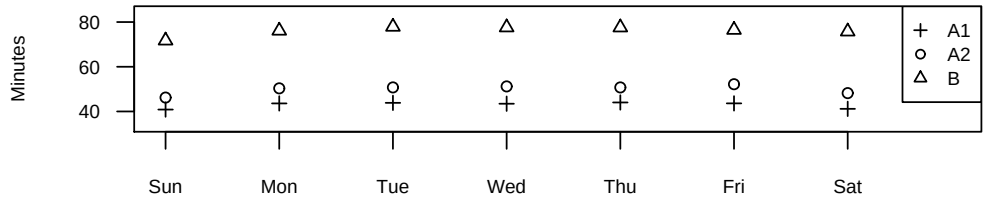

Figure 3.16: *The mean ride durations in minutes, of every day of the week for the three types of priorities separately.*

In Figure 3.16 we plotted the mean of the ride duration for each day of the week for the three types of priorities. Applying the Kolmogorov-Smirnov test to the travel times of each pair of days of the week we can determine for which days of the week the travel times have a unique distribution.

For the A1 and A2, rides comparing the distribution of the travel times of Sunday and the travel times of Saturday, leads to no rejection of the null hypothesis (p-value of 0.5064 and 0.4123), *and comparing the distributions of the travel times of Monday to Friday in pairs leads also to no rejection* of (3.8). We define $F_{\text{A1,SunSat}}$ and $F_{\text{A2,SunSat}}$ as the distribution of the travel times on Sunday and Saturday of the A1 respectively A2 rides, and let $F_{\text{A1,Mon-Fri}}$, $F_{\text{A2,Mon-Fri}}$ as the distribution of the travel times on Monday to Friday of the A1, respectively A2 rides. We can also apply the Kolmogorov Smirnov test with an alternative hypothesis to check whether distribution F is significantly lower than G;

$$H_0 : \quad F > G, \tag{3.9}$$

$$H_1 : \quad F \leq G.$$

When we test this hypothesis we get, with both p-values smaller than 0.001, that distribution $F_{\text{A1,SunSat}}$ is significantly smaller than distribution $F_{\text{A1,Mon-Fri}}$, and $F_{\text{A2,SunSat}}$ is significantly smaller than distribution $F_{\text{A2,Mon-Fri}}$.

The travel times of the B rides have a significant different distribution for the Sundays, Mondays and Saturdays, compared to the travel time distributions of the other days of the week, whereas the distribution of the travel times of Tuesday to Friday appear to be alike. Hence we obtain for the travel times of the B rides the following empirical distributions: $F_{\text{B,Sun}}$, $F_{\text{B,Mon}}$, $F_{\text{B,Tue-Fri}}$ and $F_{\text{B,Sat}}$. Applying the Kolmogorov-Smirnov test with the hypothesis in (3.9) we obtain which of these distributions is significantly the highest and which the lowest. We get that $F_{\text{B,Sun}}$ is significantly smaller than $F_{\text{B,Sat}}$, since the test gives a p-value of 0.0014. The distribution of the travel times of Saturday is significantly smaller than those of Monday (p-value is smaller than 0.0001). Because we obtained a p-value smaller than 0.0001, the distribution of the travel times on Monday is smaller than the distribution $F_{\text{B,Tue-Fri}}$.

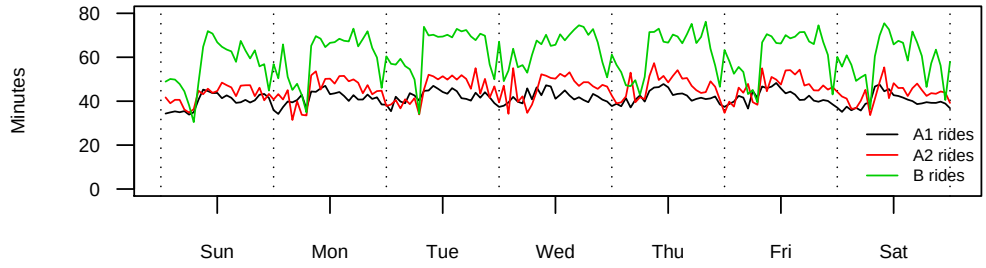

Figure 3.17: *The mean ride duration of the ambulance rides, separated by priority, starting in the same hour, for each hour of the week.*

In Figure 3.17 the mean of the ride durations are shown for each hour during the week. An investigation of the travel times of different hours of departure is necessary. By categorizing the empirical travel time distributions to in which hour the ambulance departed, we can be more specific.

We applied the Kolmogorov Smirnov test to look for differences in the empirical travel time distributions between the hours of departure of the days categorized in the previous alineas. And again we categorized the distributions which behave the same. The empirical distributions of the travel times of the A1 rides are plotted in Figure 3.18, the distributions of the A2 rides can be found in Figure 3.19, and the density of the travel times for B rides can be found in 3.20.

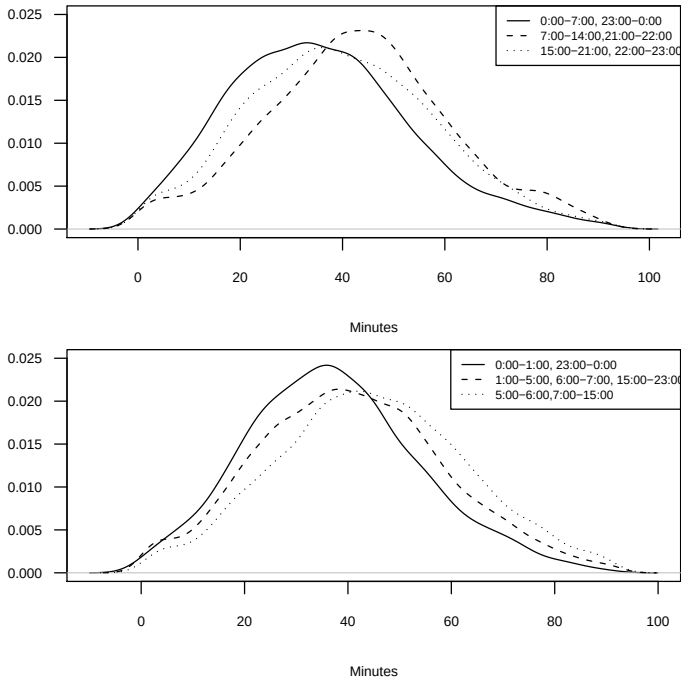

Figure 3.18: *Plots of the empirical densities of the travel times for the A1 rides on Sunday and Saturday, and Monday to Friday, separately, categorized for different hours of the day.*

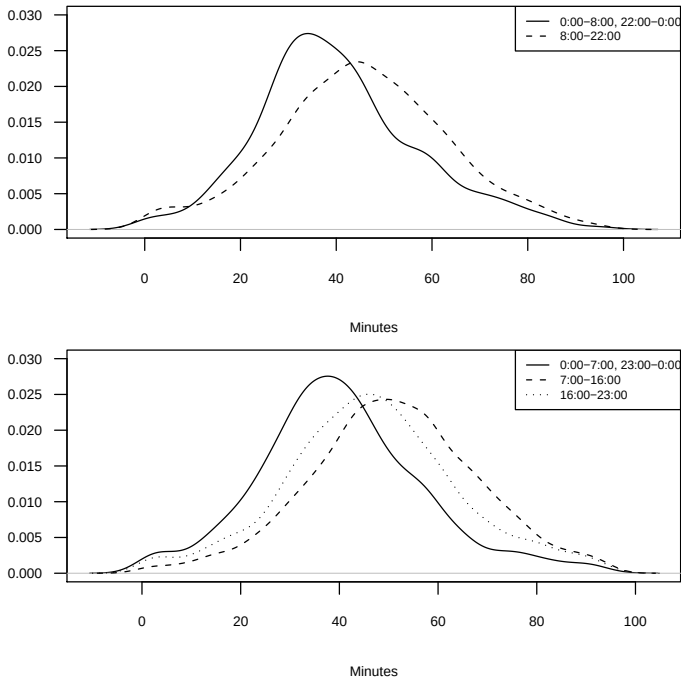

Figure 3.19: *Plots of the empirical densities of the travel times for the A2 rides on Sunday and Saturday, and Monday to Friday, separately, categorized for different hours of the day.*

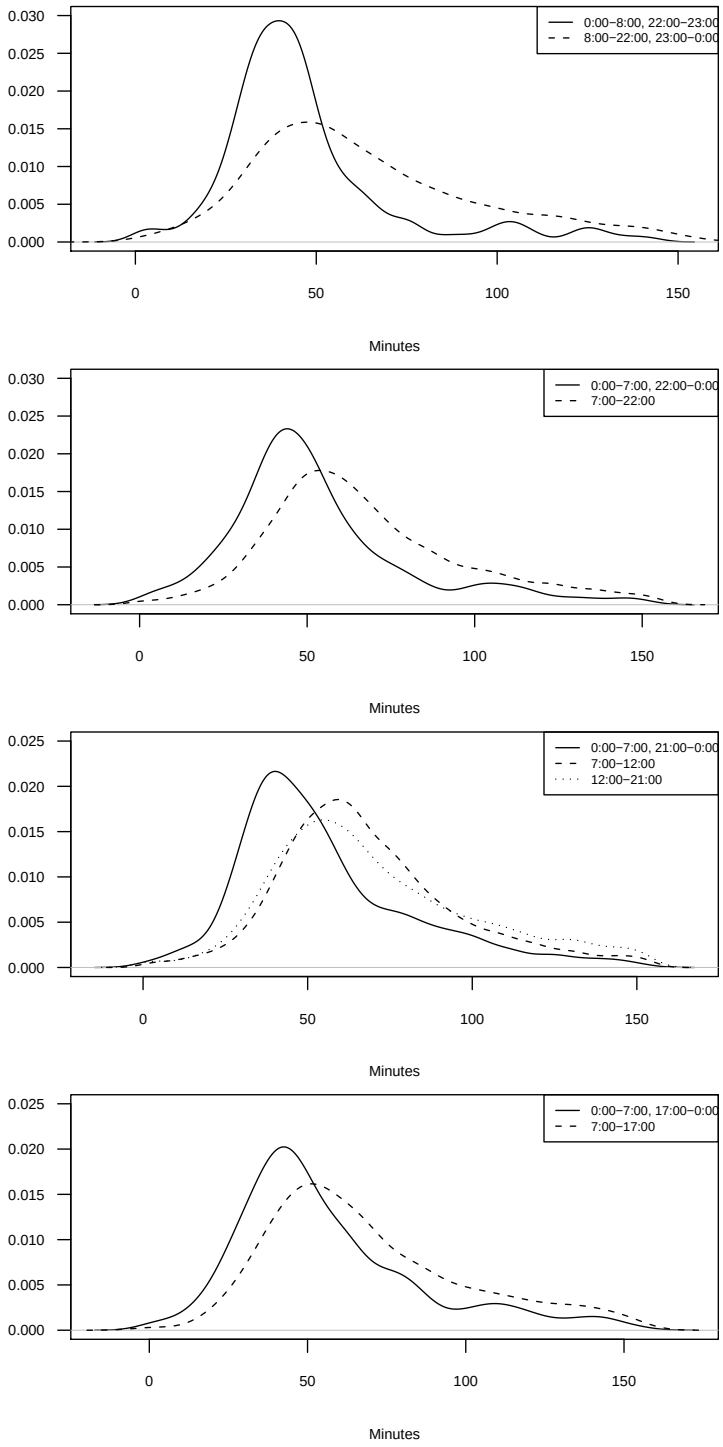

Figure 3.20: *Plots of the empirical densities of the travel times for the B rides on Sunday, Monday, Tuesday to Friday, and Saturday, separately, categorized for different hours of the day.*

3.6 Conclusions

In the beginning of this chapter we formulated several research questions. In this section we will try to provide some answers.

Does the data of the number of ambulance requests show correlations, typical yearly, weekly and daily arrival patterns?
Observing several plots, and applying a number of statistical tests provided insight in the characteristics of the number of ambulance requests. We summarize the prominent conclusions:

- The numbers of A1 rides per day have some explainable outliers: Queen's Day and the first of January.
- The number of A1 calls per day and the number of A2 calls per day are uncorrelated but the type B calls are negatively correlated to the number of A1 rides per day but positively to the A2 rides.
- The number of A1, A2 and B rides per day are positively correlated with the number of rides during the previous day.
- To the number rides two days ago the A1 rides are positively correlated, the A2 are not, whereas the B rides are significantly negatively correlated.
- A positive correlation between the number of rides per day with the number of rides seven days ago, for each of the priorities.
- During the weekend significantly higher number of A1 rides per day occur, whereas priority A2 and B rides intend to appear less during weekends.
- The numbers of A1 rides per hour over the day are significantly alike on Monday to Thursday. On Sunday, Friday and Saturday distinct daily patterns are visible.
- The daily patterns of the A2 rides can be grouped in two categories Monday to Friday, and Saturday and Sunday.
- The B rides show a significant difference in the daily patterns between Tuesday to Friday and the other days of the week separately.

Which factors are of influence when modeling a non-homogeneous Poisson process to the number of requested ambulances?
We made use of a multiple linear regression model to obtain estimates of the parameters of the non-homogeneous Poisson process. The factors month of the year, day of the week and a number of different Dutch holidays were taken as influential factors.

For the ambulance rides categorized as A1, we found the significant January, April, August, November, Tuesday, Wednesday, Friday and Saturday January-the-first, Queen's Day and Christmas Day. During those months and days a typical behavior of the number of A1 requests is indicated by the multiple linear regression model. The factors that are significant for the regression model describing the number A2 requests are: the month August, days of the week Sunday, Monday, Friday, Saturday and the holidays Easter Monday and Whit Monday. The months that are of influence on the number of B rides per day are: January, April, July and August. Each day of the week is of significant influence according to the multiple linear regression model, and the holidays which have a significant influence are: January the first, Easter Monday, Queen's Day, Ascension Day, Whit Monday, Christmas Day and Boxing Day. When we tested if the number of A1, A2 and B rides follow the non-homogeneous Poisson distribution we were not able to reject the nullhypothesis. With use of a multinomial distribution we obtained a model dividing the number of incoming rides per day over the day itself. The parameters of the multinomial distribution differ for different days of the week.

Does the data of the occupancy time of the ambulances show correlations, typical yearly, weekly and daily occupancy time patterns?
About 1 % of the A1 rides we came across has a ride duration of more than 100 minutes, and some of the ordered rides took more than six hours. The ambulance rides with priority A1 are significantly shorter than the A2 and B rides, and the A2 rides take a significantly shorter time than the B rides. Comparing the mean of the ride durations suggests that the day of the week is of influence to the occupancy time of an ambulance. Applying statistical tests showed us that the distribution of the ride durations is significantly distinct for different days of the week and different hours of the day. We obtained for the priority A1, A2 and B rides separately a number of empirical densities for the ride duration of the ambulances. First we categorized those densities according to distinct behavior on some weekdays. For the A1 and A2 rides the distributions of the travel times on Sunday and Saturday are significantly alike. For the other days of the week another density of the travel times is applicable. The travel times of the B rides have a significant different distribution for the Sundays, Mondays and Saturdays, compared to the travel time distributions of the other days of the week, whereas the distribution of the travel times of Tuesday to Friday appear to be alike. By investigating the travel times in one of the categories by hour of departure, we were able to obtain several significantly different empirical densities of the occupancy times for each of the priorities seperately.

Does the data provide reason to distinguish between ambulance rides of different priorities?.
The number of priority A1 ambulance rides per day and the number of A2 rides per day are uncorrelated, whereas the type B calls are negatively correlated to the number of A1 rides per day but positively to the A2 rides. This can be a result of the higher number of A1 rides, and the lower number of A2 and B rides, during the weekend. Although we established a

correlation, we investigated the three types of priorities separately since the travel times of rides of different priorities are significantly different. And about 51% of the dispatched rides is categorized as an A1 ride, about 10% are priority A2 and 38% obtained a B priority.

Chapter 4

Forecasting

In this chapter we will take a look at different forecasting methods predicting the number of ambulance rides per day. We maintain forecast horizons of two weeks, since this provides enough time to be able to decide whether extra ambulances need to be recruited, or on the contrary that ambulances can be scheduled for maintenance. Questions we will address are 'What forecasting methods employ the observations of the number of ambulance requests?', 'How can one measure the performance of a forecasting model?' and 'Which forecasting model is preferable?'

We start by introducing different measures to evaluate forecasting models by their goodness of fit to the data, and their forecasting performance in Section 4.1. Based on results from our data analysis to the number of incoming rides per day in Chapter 3, we introduce exponential smoothing methods, ARIMA models and multiple linear regression models in Section 4.2. These models and the non-homogeneous Poisson process introduced in Section 3.4.1, will be used to provide forecasts for the number of ambulance rides for each of the different priorities separately, based on a subset of our data the first 791 observations: November 1, 2005 through December 31, 2007. In Section 4.3 the models will be evaluated based on their goodness-of-fit and forecasting accuracy, by use of a test set, the remaining 244 observations (January 1, 2008 through August 31, 2008). In the next chapter we will develop staffing methods which can be applied to such forecasts.

Throughout this chapter Y_t denotes the number of rides on day t, for $t = 1, 2, \ldots, n$, where $n = 1035$. We define F_t as a forecast of the number of rides on day t.

4.1 Measuring Methods

To decide which of the considered models is the most accurate to forecast the ambulance calls per day we look at forecasts generated by our models and compare them to our test set of January 1, 2008 to August 31, 2008. We also need to check whether the models fit properly to the subset of the data we used to obtain our models. In this section we will introduce different ways to measure the goodness of fit of our models and the accuracy of the forecasts obtained by our models.

A measure of goodness-of-fit of the model is the Mean Squared Error (MSE) [25], which is the mean of the squared residuals $e_t = Y_t - \hat{Y}_t$. The MSE is defined as:

$$\mathsf{MSE}_{fit} = \frac{1}{n}\sum_{t=1}^{n} e_t^2. \tag{4.1}$$

The errors are squared, hence large errors are given more weight. We want the value of the MSE to be low. Some of the models discussed in this chapter use the MSE to determine their parameters.

Another value which will give an indication of how good the model is fitted to the data, is Akaike's information criterion [25]. It takes into account the maximum likelihood estimator L of the variance of residuals, and the number of estimated parameters N in the fitted model:

$$\mathsf{AIC} = -2\ln L + 2N, \tag{4.2}$$

The value of L cannot always be generated by software. In [25] an approximation is given for $-2\ln L$. By taking this approximation, the AIC becomes:

$$\mathsf{AIC} \approx n(1+\ln(2\pi)) + n\ln\sigma^2 + 2N, \tag{4.3}$$

where σ^2 denotes the sample variance of the residuals, which in its turn can be estimated by:

$$\hat{\sigma}^2 = \frac{1}{n-N}\,\mathbb{E}\sum_{t=1}^{n} e_t^2.$$

The AIC is a measure which can be used to compare fits of different models to the same data set; the AIC value itself does not provide much information on the fit of a model.

The MSE described as a measure of the fit of a model can also be used to measure the exactness of a forecast at day t of (in our case) the 14 upcoming days:

$$\mathsf{MSE}_{forecast}(t) = \frac{1}{14}\sum_{i=1}^{14}\left(Y_{t+i} - F_{t+i}(t)\right)^2. \tag{4.4}$$

Another useful way to compare our predictions is by use of the Mean Absolute Percentage Error (MAPE); it provides a more relative measure of correctness by standardizing the error in the forecast by the observed number of rides.

$$\mathsf{MAPE}(t) = \frac{1}{14}\sum_{i=1}^{14}\frac{|Y_{t+i} - F_{t+i}|}{Y_{t+i}} \times 100\%. \tag{4.5}$$

A negative quality of the MAPE is that a relatively large forecast error $|Y_{t+i} - F_{t+i}|$ can be canceled out by a large observation Y_{t+i}, making the forecast appear reasonable. Hence we introduce the Weighted Mean Absolute Percentage Error which multiplies the standardized errors with the influence of that observation compared to all observations of that forecast.

$$\begin{aligned}\mathsf{WMAPE}(t) &= \frac{1}{14}\sum_{t=1}^{14}\left(\frac{Y_{t+i}}{\frac{1}{14}\sum_{j=1}^{14} Y_{t+j}}\right)\frac{|Y_{t+i} - F_{t+i}|}{Y_{t+i}} \times 100\% \\ &= \frac{\sum_{i=1}^{14}|Y_{t+i} - F_{t+i}|}{\sum_{j=1}^{14} Y_{t+j}} \times 100\%.\end{aligned} \tag{4.6}$$

The MSE and the AIC of the residuals of the fitted models are discussed in the next sections, and the MSE, MAPE and WMAPE of the forecast provided by these models, will be given and compared in Section 4.3.

4.2 Daily Call Volumes

For each of the forecasting models we will provide a summary of its parameters for the three priorities separately, how we obtained these, and how to use the model to generate a forecast of the number of ambulance rides per day for the coming two weeks.

4.2.1 Exponential Smoothing; Holt-Winters

In the previous chapter we have seen that for the three types of priorities a correlation exists between the number of rides per day, and the previous day. Exponential smoothing methods have in common that recent values are given relatively more weight than older observations. Several exponential smoothing methods are described in [25]. One of these, the Holt-Winters method, can be used on data that exhibit a linear trend and a seasonal pattern. It is based on three smoothing equations; one for the level of the data set, one for the trend, and one for a (seasonal) pattern, and an additive error term with constant variance. The local deseasonalized level may be modified by the additive trend. There are two different Holt-Winters methods, depending on whether seasonality is modeled in an additive or a multiplicative way. The additive version assumes that the seasonal effects are of constant size, whereas the multiplicative version assumes that the seasonal effects are proportional to

the local deseasonalized mean level. We fit both models to the number of rides on day t, with t a date in the subset of our data

For the additive Holt-Winters model, the equations describing the level a_t of the data, the trend b_t and the seasonal component s_t at time t are defined follows:

$$\begin{aligned} a_t &= \alpha(Y_t - s_{t-s}) + (1-\alpha)(a_{t-1} + b_{t-1}), \\ b_t &= \beta(a_t - a_{t-1}) + (1-\beta)b_{t-1}, \\ s_t &= \gamma(Y_t - a_t) + (1-\gamma)s_{t-d}. \end{aligned}$$

where d is the length of the seasonality period. By obtaining new data points, the level, the trend and the seasonal component will be updated, and a_t, b_t and s_t will be smoothed. The model can generate forecasts for any number of days ahead; for m days ahead the forecast is given by:

$$F_{t+m} = a_t + b_t m + s_{t+1+(m-1) \mod d},$$

where a_t, b_t and $s_{t+1+(m-1) \mod d}$ are determined with the data available until day t. The smoothed average a_t does not include seasonality, it is deseasonalized by substracting s_{t-d}. The trend at time t, is based on weighting the most recent trend $a_t - a_{t-1}$ with β, and the previous one with $(1-\beta)$. To smooth the randomness of Y_t, the parameter s_t weights the newly computed seasonal component with γ and the one of one period ago with $(1-\gamma)$.

The seasonal component in the Holt-Winters method can also be modeled in a multiplicative way. One choses to model the seasonality in an additive manner when the data shows steady seasonal fluctuations, regardless of the overall level, whereas a multiplicative modeled seasonal component is chosen when the seasonal component is proportional to the average level of the series. The basic equations of the level, trend and seasonal aspect, and the formula for a forecast of m days are then given as follows:

$$\begin{aligned} a_t &= \alpha\frac{Y_t}{s_{t-s}} + (1-\alpha)(a_{t-1} + b_{t-1}), \\ b_t &= \beta(a_t - a_{t-1}) + (1-\beta)b_{t-1}, \\ s_t &= \gamma\frac{Y_t}{a_t} + (1-\gamma)s_{t-d}, \\ F_{t+m} &= (a_t + b_t m)s_{t+1+(m-1) \mod d}. \end{aligned}$$

Just as in the additional model d denotes the length of the seasonality period, a_t the level of the time series, b_t the trend, and s_t denotes the seasonal component at time t.

The smoothing parameters α, β, and γ are determined by minimizing the mean square

error (4.1) of the model of the forecast one day ahead compared to the actual number of rides one day ahead. The estimated number of rides on day t is hence, according to the additive Holt-Winters model, given by

$$F_{(t-1)+1} = a_{(t-1)} + b_{(t-1)} \cdot 1 + s_{(t-1)+1+(1-1) \mod d} = a_{t-1} + b_{t-1} + s_{t \mod d}.$$

For the multiplicative model we get: $F_{(t-1)+1} = (a_{t-1} + b_{t-1}1)s_{(t-1)+1+(1-1) \mod d} = (a_{t-1} + b_{t-1})s_{t \mod d}$.

The parameters used in the additive model for the number of A1 rides per day are estimated by: $\alpha = 0.038$, $\beta = 0$ and $\gamma = 0.084$. The additive Holt-Winters model for the A1 rides becomes:

$$\begin{aligned} a_t &= 0.038\,(Y_t - s_{t-d}) + 0.962\,a_{t-1}, \\ b_t &= 0, \\ s_t &= 0.084\,(Y_t - a_t) + 0.916\,s_{t-d}. \end{aligned} \tag{4.7}$$

The forecast, according to this model, of the number of A1 rides per day is given by

$$F_{t+m} = a_t + s_{t+1+(m-1) \mod d}.$$

The multiplicative Holt-Winters model for the number of A1 rides per day is given by

$$\begin{aligned} a_t &= 0.041\frac{Y_t}{s_{t-s}} + 0.959 a_{t-1}, \\ b_t &= 0, \\ s_t &= 0.082\frac{Y_t}{a_t} + 0.918 s_{t-d}, \\ F_{t+m} &= a_t s_{t+1+(m-1) \mod d}. \end{aligned}$$

For the number of daily A2 rides per day we estimated the parameter values $\alpha = 0.012$, $\beta = 0$, and $\gamma = 0.116$. The corresponding additive Holt-Winters model is given by:

$$\begin{aligned} a_t &= 0.012\,(Y_t - s_{t-d}) + 0.988\,a_{t-1}, \\ b_t &= 0, \\ s_t &= 0.116\,(Y_t - a_t) + 0.884\,s_{t-d}, \end{aligned} \tag{4.8}$$

and the formula of the forecasts for m days ahead becomes:

$$F_{t+m} = a_t + s_{t+1+(m-1) \mod d}.$$

For the number of A2 rides we found the multiplicative model

$$\begin{aligned} a_t &= 0.00008\frac{Y_t}{s_{t-s}} + 0.99992a_{t-1}, \\ b_t &= 0, \\ s_t &= 0.143\frac{Y_t}{a_t} + 0.857s_{t-d}, \\ F_{t+m} &= a_t s_{t+1+(m-1) \mod d}. \end{aligned}$$

For the number of dispatched B rides per day we found parameter estimates 0.023, 0, and 0.074 for α, β, and γ. The additive Holt-Winters model describing the number of B rides on day t is given by:

$$\begin{aligned} a_t &= 0.023(Y_t - s_{t-d}) + 0.977a_{t-1}, \\ b_t &= 0, \\ s_t &= 0.074(Y_t - a_t) + 0.926s_{t-d}, \end{aligned} \tag{4.9}$$

and the formula of the forecasts for m days in the future then becomes:

$$F_{t+m} = a_t + s_{t+1+(m-1) \mod d}.$$

The multiplicative Holt-Winters model fitted to the number of dispatched B rides per day is given by:

$$\begin{aligned} a_t &= 0.011\frac{Y_t}{s_{t-s}} + 0.989a_{t-1}, \\ b_t &= 0, \\ s_t &= 0.072\frac{Y_t}{a_t} + 0.928s_{t-d}, \\ F_{t+m} &= a_t s_{t+1+(m-1) \mod d}. \end{aligned}$$

The analysis performed in Chapter 3 resulted in no suggestion of a trend in number of A1, A2 or B rides per day. Hence we should expect b_t to be zero for our 3 additive and the 3 multiplicative models we are trying to establish. Clearly this is the case since *for all models the estimate of β is zero.*

For the fit to the number of A1, A2 and B rides per day seperately, the level component a_t is alike for the additive and multiplicative model. The value of α indicates how fast the model reacts to a variation of the observed data. The obtained estimates of α are small, hence *the recently observed data does not play a prominent role to the fit* of the model to the subset of our data set.

The seasonal components differ for the additive and multiplicative model, which makes sense since in the additive models the seasonal component has an additive influence whereas in the multiplicative models s_t influences the model in a multiplicative manner.

The fitted additive and multiplicative Holt-Winters models appear to be alike for each of the three priorities. The model is not able to take into account (explainable) outliers, since omitting one data point will disturb the seasonal component. The high number of rides occurring on for instance Queen's Day or the first of January influence the model, when such days are implemented in the model.

The fits to the data of the additive and multiplicative models for each of the type of ambulance rides, appear alike; in a plot they are difficult to distinguish. The mean of their absolute differences is 0.14, 0.17, and 0.42 respectively, for the A1, A2 and B rides respectively. And the corresponding standard deviations are 0.17, 0.14, and 0.38.

These observations result in no preference for the multiplicative model compared to the additive Holt-Winters model and vice versa.

4.2.2 ARIMA Models

In Chapter 3 we investigated the correlation between the number of daily rides of successive days. The models we will discuss in this subsection, will take these correlations into account. We will try to fit an Autoregressive Integrated Moving Average model (ARIMA) [14] to the data of the number of rides per day.

ARIMA models are a mixture of an autoregressive part AR and an moving average MA part. The AR describes the relation between the value of the time-series at time t and its values on previous times. The basic equation of a p-th order autoregressive model is

$$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + e_t,$$

where c is a constant term, ϕ_i is the i-th autoregressive parameter and e_t denotes the error term at time t. The moving average part assumes that Y_t depends on the error term at time t and the previous error terms. A q-th order MA part of an ARIMA model is given by

$$Y_t \;=\; c + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q},$$

where c is again a constant term, θ_j denotes the j-th moving average parameter and with e_{t-k} we mean the error term at time $t-k$.

To fit an autoregressive moving average model we need the data to be stationary, that is, there is no growth or decline in the data [25]. One way to obtain a stationary time series is to difference the data: $Y'_t = Y_t - Y_{t-1}$. When we difference the data we can speak of an autoregressive integrated moving average model instead of an autoregressive moving average model. To check whether we have stationary data, we use the Dickey-Fuller test [25]. This

test estimates the regression model

$$Y_t' \quad = \quad \phi Y_{t-1} + \varphi_1 Y_{t-1}' + \varphi_2 Y_{t-2}' + \cdots + \varphi_6 Y_{t-6}',$$

by use of least squares. The estimated value of ϕ will be close to zero if Y_t needs differencing. For each of the priorities of the number of rides per day we obtain a p-value of 0.01, indicating the time-series are not stationary processes. Since the data shows a weekly pattern we will difference the data with a lag of 7 to obtain stationarity. A lag of 7 is chosen since our data shows a clear seasonality of 7 days.

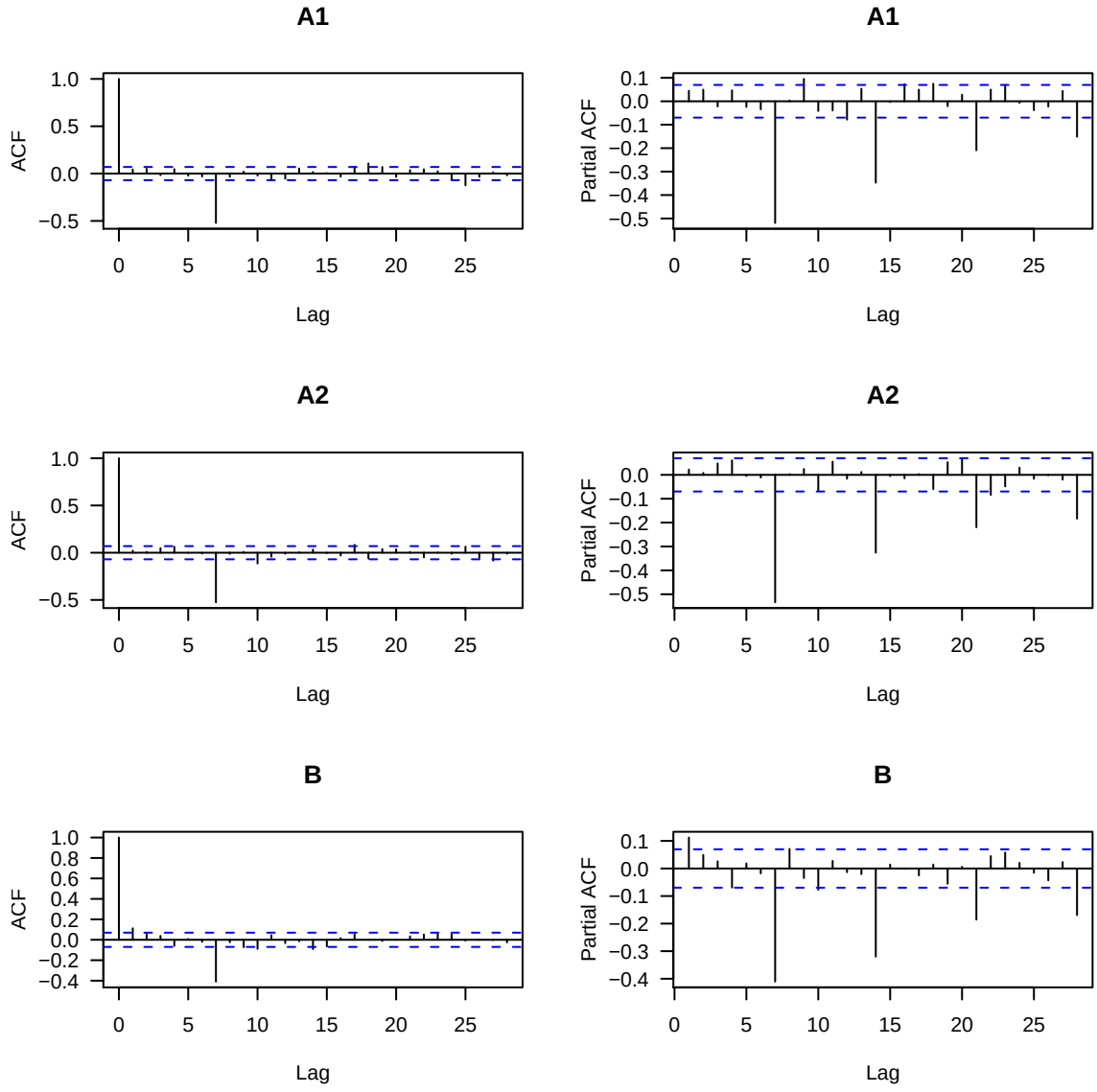

Figure 4.1: *Autocorrelation and partial autocorrelation function of the differenced data by lag 7.*

To determine the orders of an autoregressive or moving average model we can make use of the autocorrelation function (acf) and the partial autocorrelation function (pacf) of the differenced (by lag 7) data. This is explained in [25]. The acf and pacf of our differenced data do not provide a clear indication whether to use specifically an AR or a MA model. Hence we need to look for a mixture of these. The acf and pacf show pikes at lags 7, 14, 21, and 28, see Figure 4.1, which indicates that data separated by a week may exhibit the same properties. To take this into account in our model we need include seasonal autoregression of order P and or seasonal moving average parts of order Q in the model. The notation used to describe such a model is $\mathsf{ARIMA}(p,d,q)(P,D,Q)$, where d indicates the number of times the data is differenced by lag 1, and D the number of times the data is differenced by a lag of the period of the data.

The acf and pacf can help by determining which orders of AR and MA to choose, but this can be hard, since a mixture of the models is difficult to distinguish by just looking at the acf and pacf. Which of the models is preferred can be determined by minimizing Akaike's Information Criterion (see Section 4.1)[25][14]. We choose the values of p, q, P, and Q by minimizing

$$\mathsf{AIC} \quad = \quad -2\ln L + 2(p+q+P+Q),$$

where L denotes the maximum likelihood estimator. For all three priorities the acf and pacf do not give a clear indication whether we need to fit an autoregressive model or a moving average model to the data denoting the number of executed ambulance rides per day. Hence we determined for different combinations of p, q, P, and Q the value of AIC, and picked the model which resulted in the minimum AIC.

For the time-series of the number of A1 rides per day, the minimum AIC is attained by an $\mathsf{ARIMA}(2,0,1)(0,1,2)$ model:

$$Y_t \quad = \quad \phi_1(Y_{t-1} - Y_{t-8}) + \phi_2(Y_{t-2} - Y_{t-9}) + Y_{t-7} + \ldots \tag{4.10}$$

$$\ldots + e_t - \theta_1 e_{t-1} + \Theta_1(\theta_1 e_{t-8} - e_{t-7}) + \Theta_2(\theta_1 e_{t-15} - e_{t-14}). \tag{4.11}$$

The coefficients ϕ_i and Φ_i describe the non-seasonal and seasonal autoregressive part of the model respectively, where θ_i and Θ_i are the coefficients of the non-seasonal and seasonal contribution of the moving average part of the model respectively. *The order of the model takes the correlation between the number of A1 rides per day with the number of A1 rides yesterday, and with two days ago into account* (see Page 23). Also a seasonal influence is

modeled. We estimated the coefficients by use of the maximum likelihood.

$$\begin{aligned} Y_t &= 1.0314(Y_{t-1} - Y_{t-8}) - 0.0624(Y_{t-2} - Y_{t-9}) + Y_{t-7} + \dots \\ &\quad \dots + e_t + 0.9208e_{t-1} - 1.0360(-0.9208e_{t-8} - e_{t-7}) + 0.0360(-0.9208e_{t-15} - e_{t-14}). \end{aligned}$$

The time-series of the number of A2 rides per day is best fitted by the ARIMA(0,0,0)(0,1,1) model. *For the A2 rides only the number of rides one week ago appears to be of influence for the ARIMA model*:

$$Y_t = Y_{t-7} + e_t - \Theta_1 e_{t-7}. \tag{4.12}$$

Filling in the estimated parameters leads to the following equation describing the number of executed ambulance rides with priority A2:

$$Y_t = Y_{t-7} + e_t + 0.9638e_{t-7}.$$

The ARIMA model which scores the lowest AIC value when fitted to the number of B rides per day is the ARIMA$(2,0,1)(0,1,2)$:

$$Y_t = \phi_1(Y_{t-1} - Y_{t-8}) + \phi_2(Y_{t-2} - Y_{t-9}) + Y_{t-7} + \dots \tag{4.13}$$

$$\dots + e_t - \theta_1 e_{t-1} + \Theta_1(\theta_1 e_{t-8} - e_{t-7}) + \Theta_2(\theta_1 e_{t-15} - e_{t-14}). \tag{4.14}$$

With the parameter estimates we obtain:

$$\begin{aligned} Y_t &= 1.1019(Y_{t-1} - Y_{t-8}) - 0.1136(Y_{t-2} - Y_{t-9}) + Y_{t-7} + \dots \\ &\quad \dots + e_t + 0.9679e_{t-1} - 0.8869(-0.9679e_{t-8} - e_{t-7}) - 0.1130(-0.9679e_{t-15} - e_{t-14}). \end{aligned}$$

The ARIMA model fitted to the number of daily B rides is based on the relations with the number of rides, one, two and seven days before. See Page 23 for correlations.

Minimizing the AIC values is not the only measurement to determine whether the models we found are good fits to the data. When the residuals of the fit of a model are a white noise series, the model needs no further refinement [25]. Portmanteau tests investigate whether a set of autocorrelations r_k values is significantly different from a zero set. The Ljung-Box portmanteau [25] investigates the autocorrelations r_k at lag k in the following manner

$$Q = n(n+2)\sum_{k=1}^{h} \frac{r_k^2}{n-k},$$

because if the data is white noise, the statistic Q attains a distribution close to the chi-square distribution with $(h-m)$ degrees of freedom, where the maximum lag considered is denoted by h, m is the number of the used parameters to fit the model, and n denotes the number of

observations in the dataset. Usually $h \approx 20$ is chosen [25].

The p-values of the Portmanteau test for lag up to 30 are plotted in Figure 4.2.

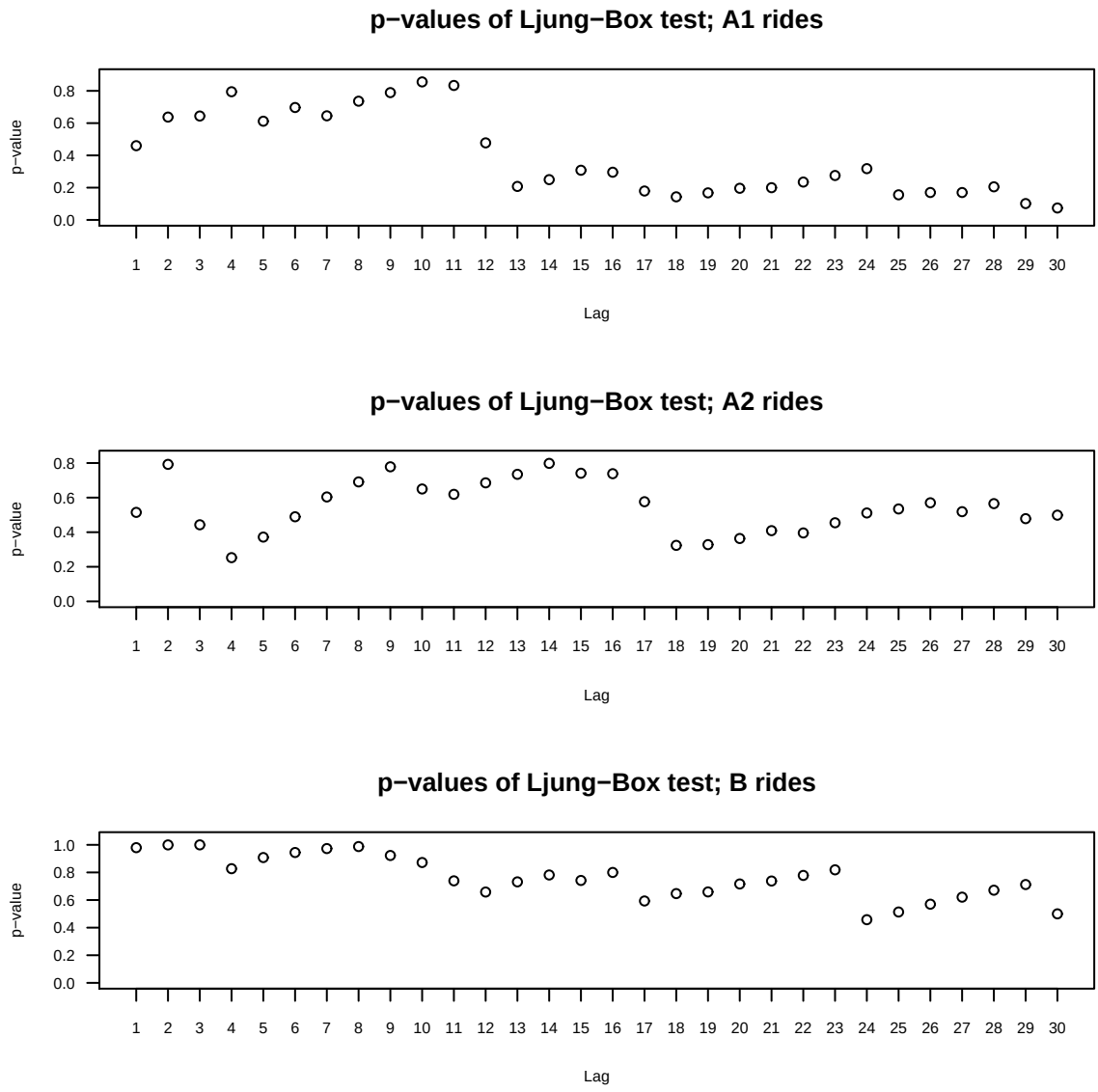

Figure 4.2: *p-values of the Ljung-Box test applied to the residuals obtained from fitting seasonal ARIMA models to our data.*

The p-values indicate that the obtained models (4.10), (4.12) and (4.15) *are indeed good fits to the data.*

The ARIMA models describing the number of ambulance rides per day for each of the three priorities separately, can be applied to forecasting. For the number of A1 rides per day

a forecast of one day ahead becomes

$$F_{t+1} = \phi_1(Y_t - Y_{t-7}) + \phi_2(Y_{t-1} - Y_{t-8}) + Y_{t-6} + \ldots \quad (4.15)$$

$$\ldots + e_{t+1} - \theta_1 e_t + \Theta_1(\theta_1 e_{t-7} - e_{t-6}) + \Theta_2(\theta_1 e_{t-14} - e_{t-13}). \quad (4.16)$$

A forecast for the number of A2 rides becomes:

$$F_{t+1} = Y_{t-6} + e_{t+1} - \Theta_1 e_{t-6}. \quad (4.17)$$

And for the number of ambulance B rides of priority B the forecast one day into the future becomes:

$$F_{t+1} = \phi_1(Y_t - Y_{t-7}) + \phi_2(Y_{t-1} - Y_{t-8}) + Y_{t-6} + \ldots \quad (4.18)$$

$$\ldots + e_{t+1} - \theta_1 e_t + \Theta_1(\theta_1 e_{t-7} - e_{t-6}) + \Theta_2(\theta_1 e_{t-14} - e_{t-13}). \quad (4.19)$$

In these equations the term e_{t+1} is unknown, its estimate is given by $\hat{e}_t = 0$. The values of e_{t-6}, e_{t-7}, etc. can be empirically determined from the fitted model; we take the residuals at times $t-6$ and $t-7$. For a forecast two days ahead, we need to know the number of rides on day $t+1$. Since we know how to forecast that number we implement this in the formula of F_{t+2}. We can continue this process to obtain a forecast for two weeks ahead. Since we difference the data, the forecasts will converge to the value of the last known data point.

4.2.3 Regression

The data analysis carried out in Chapter 3 indicated that the number of ambulance rides on a certain day t is related to which day of the week, which month of the year, and which holiday it is (or is not) on day t. A regression model is an explanatory model relating the number of ambulance rides per day to these effects. The model we will try to fit to the data is called a multiple linear regression model; we have multiple explanatory variables which we model in a linear manner. If the values of the coefficients are known in the future, the regression model can be used to forecast the data. And since we know the day of the week, the month of the year and whether we look at a holiday, we can use this model to forecast the number of rides per day.

The regression model we take as a starting point, takes only into account the effects of the day of the week and the month of the year. In a general form the model looks like

$$Y_t = a + \sum_{i=1}^{12} b_i B_{i,t} + \sum_{j=1}^{7} c_j C_{j,t} + \epsilon_t, \quad (4.20)$$

where Y_t is the t-th observation of the number of daily rides, indicator $B_{i,t}$ has value 1 if the month of day t is the ith month of the year and value 0 otherwise, and the indicator $C_{j,t}$ has

value 1 if day t is the jth day of the week and 0 otherwise. The parameters $b_1, \ldots, b_{12}$ and $c_1, \ldots, c_7$ are real-valued unknown constants, and the ϵ_t are called the error terms. These error terms are independent and identically Gaussian distributed with mean 0, and with variance σ^2. Estimates of the parameters can be obtained by direct calculation, using least squares to minimize the sum of squares of the residuals. Without the following two constraints we would encounter the problem of multicollinearity,

$$\sum_{i=1}^{12} b_i = 0 \quad \text{and} \quad \sum_{j=1}^{7} c_j = 0.$$

The residuals obtained after fitting the regression model (4.20) to the data are plotted in Figure 4.3. The horizontal lines are at distance of $3\hat{\sigma}$ to zero, with $\hat{\sigma_2}$ the empirical variance of the residuals. A residual smaller than -3 or larger than 3 times $\hat{\sigma}$ can be classified as an outlier [25], and diserves extra investigation. *For the A1 rides we can see four residuals at a distance larger than $3\hat{\sigma}$ from zero: both Queen's Days April 29, 2006 and April 30, 2007, and the first of January of 2006 and 2007.* F*or the number of A2 rides per day the two residuals with a notably high value are:* February 13, 2006; and May 27, 2006. *Almost all of the fourteen residuals smaller than $-3\hat{\sigma}$ for the B rides appear on national holidays in the Netherlands*, and hence can be explained as outliers. These are: January 1, 2007; Easter Monday of 2006 and 2007; Queen's Day of 2007; Ascension Day of 2006 and 2007; Whit Monday of 2006 and 2007; Christmas Days of 2006 and 2007; Boxing Days of 2005, 2006 and 2007. The fourteenth residual smaller than $-3\hat{\sigma}$ corresponds to December 9, 2005.

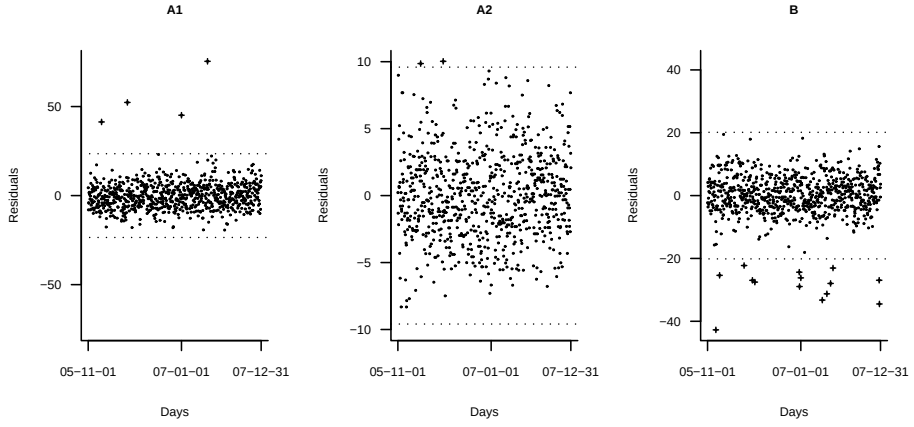

Figure 4.3: *Residuals for the regression model with factors day of the week and month of the year.*

To obtain a model that generates a better fit to the data we need to include new parameters $d_1, \ldots d_7$ to the model, denoting the holiday effects of January 1, Easter Monday, Queen's Queen's Day, Ascension Day, Whit Monday, Christmas Day, and Boxing Day. Even though we did not extreme values of the residuals at these holidays for all three priorities, we will include

them in the model for each priority to investigate their effects. For each of the ambulance priorities the next multiple linear regression model becomes:

$$Y_t = a + \sum_{i=1}^{12} b_i B_{i,t} + \sum_{j=1}^{7} c_j C_{j,t} + \sum_{k=1}^{7} d_k D_{k,t} + \epsilon_t. \tag{4.21}$$

The application of a regression model demands that the residuals are i.i.d. Gaussian distributed.

We applied the Ljung-Box test to the residuals after fitting the model in (4.21) to our data. The p-values for $h = 1, \ldots, 30$ are plotted in Figure 4.4 for the three types of priorities separately.

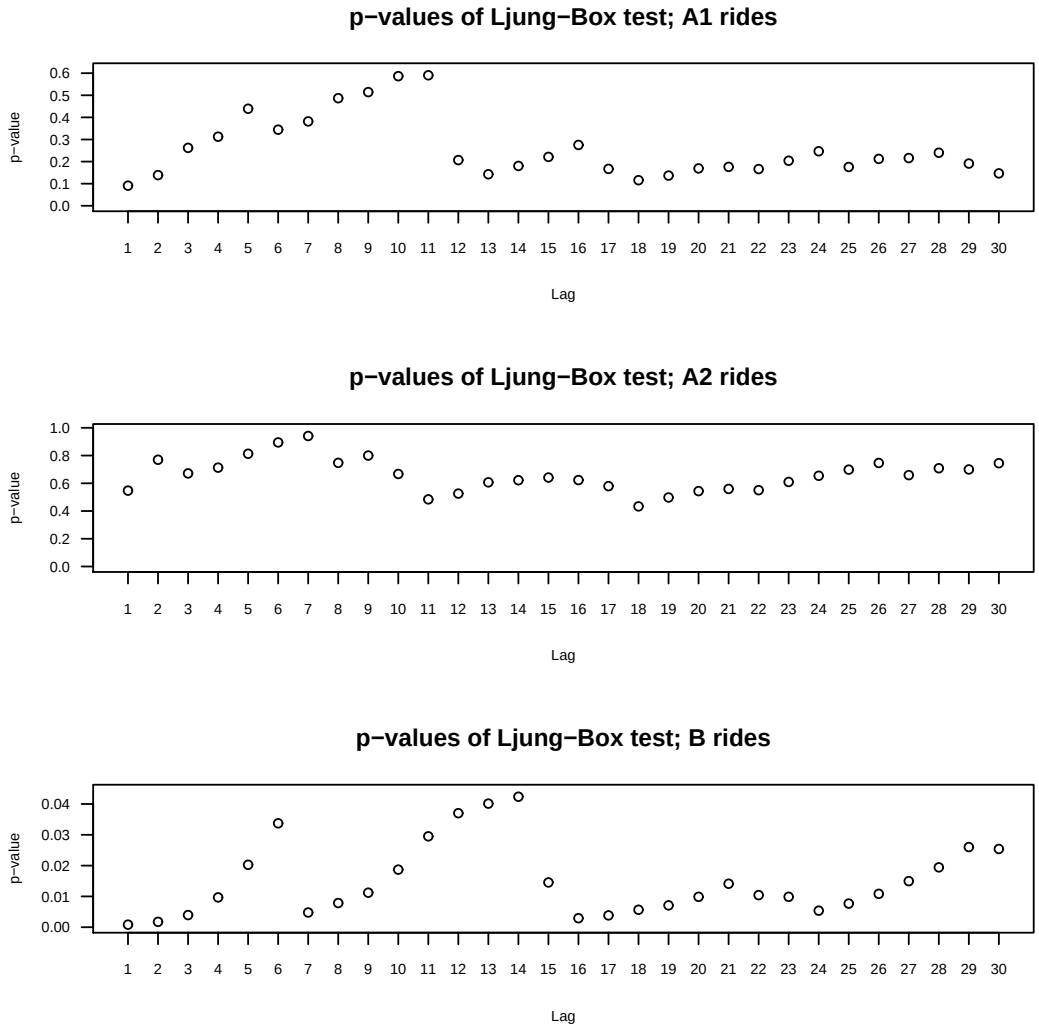

Figure 4.4: *p-values of the Ljung-Box test applied to the residuals for the regression models with factors day of the week, month of the year and some Dutch holidays, for each of the ambulance priorities separately.*

Apparently *the condition of no correlation between the residuals does not apply to the regression model fitted to the number of B rides per day*, hence we need to improve the regression model in (4.21). By fitting an AR model of order p to the residuals e_t, we take into account the correlation. (See Section 4.2.2 for an explanation of this AR(p) process). The improved regression model for the number of B rides per day becomes:

$$Y_t = a + \sum_{i=1}^{12} b_i B_{i,t} + \sum_{j=1}^{7} c_i C_{j,t} + \sum_{k=1}^{7} d_k D_{k,t} + \alpha_1 \epsilon_{t-1} + \ldots + \alpha_p \epsilon_{t-p} + e_t, \tag{4.22}$$

where e_t is i.i.d. normal distributed, with mean 0 and a variance of σ_e^2. In Chapter 3 we tested for correlation between the number of rides per day and the number of rides seven days earlier. The outcome was significant for the B rides, hence we will take p equal to 7, the parameters α_l which appear to be insignificant to the model will be omitted later on.

The estimates of the parameters and their corresponding p-values of a t test of the nullhypothesis that the given parameter is zero, for each parameter separately, are given in Tables 4.1, 4.2 and 4.3. For the number of *A1* rides per day, and the number of *A2* rides per day, we fitted the regression model given in (4.21). To the data of the number of B rides per day we fitted the model given in (4.22).

When we consider the models containing all introduced regression parameters, the formula describing the forecast at day t, m days ahead, for the number of A1 or A2 rides becomes:

$$F_{t+m} = a + \sum_{i=1}^{12} b_i B_{i,t+m} + \sum_{j=1}^{7} c_j C_{j,t+m} + \sum_{k=1}^{7} d_k D_{k,t+m}. \tag{4.23}$$

To forecast the number of rides per day by using only the significant parameters, we also make use of (4.23) but with some of the parameters $a, b_1, \ldots, b_{12}, c_1, \ldots, c_7$ omitted. For every day t in the test set, for which we provide a two weeks forecast, we redo the estimation of the parameters in (4.23). The formula determining the forecast of the number of B rides also contains the parameters of the AR(p) model fitted to the residuals.

In Section 4.3 we will discuss per priority two regression models, the first type is modeled as (4.21), and the second only takes the significant parameters of (4.21) into account, and some ARIMA parameters for the B rides. The parameter estimates of the models with only the significant ones will differ to those given in Table 4.1, 4.2, and 4.3.

Parameter:	a	b_1	b_2	b_3	b_4	b_5	b_6
	Intercept	Jan.	Feb.	Mar.	Apr.	May.	Jun.
Estimate:	45.76	-3.59	-0.35	-0.36	1.71	0.98	0.44
p-value:	< 0.0001	< 0.0001	0.6851	0.6656	0.0489	0.2486	0.6047
Parameter:	b_7	b_8	b_9	b_{10}	b_{11}	b_{12}	
	Jul.	Aug.	Sep.	Oct.	Nov.	Dec.	
Estimate:	0.14	-3.93	-0.08	1.54	2.34	1.17	
p-value:	0.8692	< 0.0001	0.9219	0.0639	0.0009	0.1023	
Parameter:	c_1	c_2	c_3	c_4	c_5	c_6	c_7
	Sun.	Mon.	Tue.	Wed.	Thu.	Fri.	Sat.
Estimate:	-0.07	-1.18	-2.22	-1.91	0.09	1.35	3.94
p-value:	0.9052	0.0517	0.0002	0.0014	0.8747	0.0235	< 0.0001
Parameter:	d_1	d_2	d_3	d_4	d_5	d_6	d_7
	Jan.1	Easter Mon.	Q.day	Asc.Day	Whit Mon.	Chr. Day	Box. Day
Estimate:	44.95	-9.30	66.14	-2.33	-3.79	-9.10	-1.49
p-value:	< 0.0001	0.0609	< 0.0001	0.6372	0.4405	0.0238	0.7109

Table 4.1: *Parameter estimates of the fitted multiple linear regression model of to the number of A1 rides per day, and their corresponding p-values of a t test of the nullhypothesis that the parameter is zero.*

Parameter:	a	b_1	b_2	b_3	b_4	b_5	b_6
	Intercept	Jan.	Feb.	Mar.	Apr.	May.	Jun.
Estimate:	9.14	0.72	0.14	0.42	0.78	-0.22	-0.39
p-value:	< 0.0001	0.0690	0.7257	0.2815	0.0577	0.5836	0.3240
Parameter:	b_7	b_8	b_9	b_{10}	b_{11}	b_{12}	
	Jul.	Aug.	Sep.	Oct.	Nov.	Dec.	
Estimate:	-0.53	-0.94	0.01	-0.35	0.40	-0.05	
p-value:	0.1761	0.0167	0.9757	0.3689	0.2271	0.8850	
Parameter:	c_1	c_2	c_3	c_4	c_5	c_6	c_7
	Sun.	Mon.	Tue.	Wed.	Thu.	Fri.	Sat.
Estimate:	-1.23	1.02	0.20	-0.27	0.49	0.59	-0.81
p-value:	< 0.0001	0.0003	0.4662	0.3441	0.0840	0.0345	0.0042
Parameter:	d_1	d_2	d_3	d_4	d_5	d_6	d_7
	Jan.1	Easter Mon.	Q.day	Asc.Day	Whit Mon.	Chr. Day	Box. Day
Estimate:	-1.26	-4.94	2.47	-2.91	-6.86	-1.09	0.25
p-value:	0.5878	0.0342	0.2870	0.2111	0.0031	0.5647	0.8939

Table 4.2: *Parameter estimates of the fitted multiple linear regression model of to the number of A2 rides per day, and their corresponding p-values of a t test of the null hypothesis that the parameter is zero.*

Parameter:	a	b_1	b_2	b_3	b_4	b_5	b_6
	Intercept	Jan.	Feb.	Mar.	Apr.	May.	Jun.
Estimate:	26.07	1.69	1.12	1.16	1.83	0.04	0.29
p-value:	< 0.0001	0.0204	0.1450	0.1043	0.0152	0.9604	0.6879
Parameter:	b_7	b_8	b_9	b_{10}	b_{11}	b_{12}	
	Jul.	Aug.	Sep.	Oct.	Nov.	Dec.	
Estimate:	-1.74	-2.81	-1.43	-1.21	-0.04	1.12	
p-value:	0.0148	0.0001	0.0556	0.0877	0.9430	0.066	
Parameter:	c_1	c_2	c_3	c_4	c_5	c_6	c_7
	Sun.	Mon.	Tue.	Wed.	Thu.	Fri.	Sat.
Estimate:	-18.56	8.13	9.00	6.30	5.55	7.77	-18.20
p-value:	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
Parameter:	d_1	d_2	d_3	d_4	d_5	d_6	d_7
	Jan.1	Easter Mon.	Q.day	Asc.Day	Whit Mon.	Chr. Day	Box. Day
Estimate:	-15.34	-31.19	-20.61	-29.19	-28.20	-18.91	-30.83
p-value:	0.0003	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
Parameter:	α_1	α_2	α_3	α_4	α_5	α_6	α_7
Estimate:	0.10	0.03	0.02	0.00	-0.01	0.02	0.10
p-value:	0.5878	0.0342	0.2870	0.2111	0.0031	0.5647	0.8939

Table 4.3: *Parameter estimates of the fitted multiple linear regression model of to the number of B rides per day, and their corresponding p-values of a t test of the null hypothesis that the parameter is zero.*

4.2.4 Non-homogeneous Poisson Process

In Chapter 3 we introduced non-homogeneous Poisson(μ_t) processes to describe the number of A1, A2, and B rides on day t separately. In this subsection we will show how to use these processes to forecast the number of rides per day and how to compare their fit and forecasts to the other models described in this chapter.

The demand of ambulances per day behaves like a non-homogeneous Poisson process (see Section 3.4.1); the number of ambulance rides during time interval t is Poisson(μ_t) distributed. A prediction of the number of rides on day t based on this distribution is the expectation of this distribution for day t. Hence the forecast m days ahead on day t becomes

$$F_{t+m} = \mathbb{E}(\text{Poisson}(\mu_{t+m})) = \mu_{t+m}.$$

The value of μ_t, for t a day in January 1, 2008 until August 31, 2008, is based on all observations made prior to January 1 2008. Hence our predictions based on the non-homogeneous Poisson process do not depend on the day t used as a starting point of our forecasts, and do not depend on the forecast horizon m.

To determine whether the estimated $\hat{\mu_t}$ provide a good forecast we determine the forecast measures discussed in Section 4.1. The MSE of the forecast becomes

$$\begin{aligned}\mathsf{MSE}_{forecast}(t) &= \frac{1}{14}\sum_{i=1}^{14}(Y_{t+i}-F_{t+i})^2 = \frac{1}{14}\sum_{i=1}^{14}(Y_{t+i}-\mu_{t+i})^2 \\ &= \frac{1}{14}\sum_{i=1}^{14}Y_{t+i}^2 - 2Y_{t+i}\mu_{t+i}+\mu_{t+i}^2.\end{aligned}$$

The MAPE can be determined in the same manner:

$$\mathsf{MAPE}(t) = \left(\frac{1}{14}\sum_{i=1}^{14}\frac{|Y_{t+i}-F_{t+i}|}{Y_{t+i}}\times 100\%\right) = \left(\frac{1}{14}\sum_{i=1}^{14}\frac{|Y_{t+i}-\mu_{t+i}|}{Y_{t+i}}\times 100\%\right).$$

Determining the expectation of the weighted mean absolute percentage error of the non-homogeneous Poisson process goes alike:

$$\mathsf{WMAPE}(t) = \left(\frac{\sum_{i=1}^{14}|Y_{t+i}-F_{t+i}|}{\sum_{j=1}^{14}Y_{t+j}}\times 100\%\right) = \left(\frac{\sum_{i=1}^{14}|Y_{t+i}-\mu_{t+i}|}{\sum_{j=1}^{14}Y_{t+j}}\times 100\%\right).$$

After determining the value of μ_t with use of Tables 3.1, 3.2 and 3.3, we can calculate the MSE, MAPE and WMAPE of the forecasts provided by the non-homogeneous Poisson process.

Since the non-homogeneous Poisson process does not provide the number of rides per day but a distribution as such, we cannot calculate the residuals of any realization. Hence

the MSE and variance of the residuals cannot be determined as such. What we can do is determine the expected residuals, and hence the MSE and variance of residuals of the fit, to compare the non-homogeneous Poisson process with the other models. The expected residual at day t is now given by:

$$\mathbb{E}\, e_t = \mathbb{E}\,(Y_t - X_t) = Y_t - \mathbb{E}X_t = Y_t - \mu_t.$$

The MSE of the goodness-of-fit is calculated as:

$$\begin{aligned}
\mathbb{E}\,\mathsf{MSE}_{fit} &= \mathbb{E}\left(\frac{1}{n}\sum_{t=1}^{n} e_t^2\right) = \frac{1}{n}\sum_{t=1}^{n}\mathbb{E}\,(Y_t - X_t)^2 = \frac{1}{n}\sum_{t=1}^{n}\mathbb{E}\left(Y_t^2 - 2Y_tX_t + X_t^2\right) \\
&= \frac{1}{n}\sum_{t=1}^{n} Y_t^2 - 2Y_t\mathbb{E}X_t + \mathbb{E}X_t^2 = \frac{1}{n}\sum_{t=1}^{n} Y_t^2 - 2Y_t\mathbb{E}X_t + \mathbb{V}arX_t + (\mathbb{E}X_t)^2 \\
&= \frac{1}{n}\sum_{t=1}^{n} Y_t^2 - 2Y_t\mu_t + \mu_t + \mu_t^2,
\end{aligned}$$

since $\mathbb{E}\,X_t^2 = \mathbb{V}arX_t + (\mathbb{E}X_t)^2$. Akaike's information criterion is another measure of fit which we want to investigate. An approximation of the AIC is given by

$$\mathsf{AIC} \approx n(1 + \ln(2\pi)) + n\ln\sigma^2 + 2N.$$

The non-homogeneous Poisson(μ_t) process is characterized by one parameter μ_t, hence $N = 1$. The sample variance, σ^2, can again be estimated by:

$$\hat{\sigma}^2 = \frac{1}{n-N}\,\mathbb{E}\,\sum_{t=1}^{n} e_t^2.$$

For the non-homogeneous Poisson process the sample variance becomes:

$$\begin{aligned}
\hat{\sigma}^2 &= \mathbb{E}\,\frac{1}{n-1}\sum_{t=1}^{n}(Y_t - X_t)^2 \\
&= \frac{1}{n-1}\sum_{t=1}^{n}\left(Y_t^2 - 2Y_t\mathbb{E}\,X_t + \mathbb{E}X_t^2\right) \\
&= \frac{1}{n-1}\sum_{t=1}^{n}\left(Y_t^2 - 2Y_t\mathbb{E}\,X_t + \mathbb{V}arX_t + (\mathbb{E}X_t)^2\right) \\
&= \frac{1}{n-1}\sum_{t=1}^{n}\left(Y_t^2 - 2Y_t\mu_t + \mu_t + \mu_t^2\right).
\end{aligned}$$

Hence the AIC can be calculated by

$$\mathsf{AIC} \approx n(1 + \ln(2\pi)) + n\ln\left(\frac{1}{n-1}\sum_{t=1}^{n}\left(Y_t^2 + \mu_t(1 - 2Y_t + \mu_t)\right)\right) + 2.$$

In the next section we will provide the value of the MSE and AIC of the fit, and the forecast measures belonging to our non-homogeneous Poisson process.

4.3 Evaluation

In Section 4.1 we introduced several statistical measures to compare forecast methods. In this section we will give an overview of the results of each of the models discussed in Section 4.2. A good forecasting model fits the data and overall provides an accurate forecast. Both are important but a sufficient forecast is of more importance, since that is our goal.

Models:	MSE	AIC
Holt-Winters additive	69.19	10871.71
Holt-Winters multiplicative	69.19	10871.67
ARIMA	62.35	5516.31
Multiple Linear Regression	45.14	5310.28
Multiple Linear Regression (only significant parameters)	45.87	5296.99
Non-homogeneous Poisson process	91.96	5824.17

Table 4.4: *Table of the MSE and AIC of the fit for each of the fitted models to the number of A1 rides per day.*

Models:	MSE	AIC
Holt-Winters additive	11.29	9437.92
Holt-Winters multiplicative	11.44	9448.14
ARIMA	10.62	4106.77
Multiple Linear Regression	10.00	4117.71
Multiple Linear Regression (only significant parameters)	10.24	4103.21
Non-homogeneous Poisson process	19.36	4591.64

Table 4.5: *Table of the MSE and AIC of the fit for each of the fitted models to the number of A2 rides per day.*

To decide which model fits well to the data we compare the values of the MSE and AIC of their residuals; we want these values to be low. *For all three priorities the regression models provide the fit with the lowest MSE and AIC, whereas the MSE and AIC value for the non-homogeneous Poisson processes are the highest.* The number of A2 rides is significantly lower than the number of A1 or B rides per day. This explains a lower MSE of the residuals of the models describing the number of A2 rides per day. Also the existence of some outliers

Models:	MSE	AIC
Holt-Winters additive	48.93	10597.56
Holt-Winters multiplicative	48.98	10598.40
ARIMA	45.198	5262.00
Multiple Linear Regression	32.89	5059.89
Multiple Linear Regression (only significant parameters)	33.98	5071.61
Non-homogeneous Poisson process	68.22	5588.00

Table 4.6: *Table of the MSE and AIC of the fit for each of the fitted models to the number of B rides per day.*

for the number of A1 and B per day is of influence. Unexplainable outliers are difficult (if not impossible) to model. *And even though most of the found outliers are explainable, the Holt-Winters, and ARIMA models are not able to consider them as outliers, whereas it is easy to adjust the regression models in such a manner that it takes these outliers into account.* One could consider to eliminate the outliers and replace their extreme values with a value more expectable and/or explainable. The results as shown in Tables 4.4, 4.5, and 4.6 could become very different.

With each of the forecasting models discussed earlier in this chapter, we produced forecasts for the next two weeks, for January 1, 2008 until August 17, 2008. For each of these two weeks forecasts we derived the MSE (4.4), MAPE (4.5) and WMAPE (4.6). In Figures 4.5, 4.6, and 4.7 we plotted these values for each of the earlier mentioned dates. *The forecast measures for the A2 and B rides show some peaks for the Holt-Winters additive and multiplicative forecasting model.* Further investigation showed that the data until that day resulted in the parameter estimate $\alpha = 1$ (see Section 4.2.1). Hence the predicted number of ambulance rides per day is constant over the week, which results in high prediction errors. Just as in the evaluation about the fits of each of the models, some outliers are of big influence. The high number of A1 rides, and the small number of B rides on the first of January 2008, and Queen's Day (April 30, 2008), are difficult to forecast for the Holt-Winters and ARIMA models. Hence the forecasting measures of the two week forecast, prior to these two holidays, have increased. But since these models put more weight to recent observations, the high, or low, numbers of rides on the mentioned days affect also the forecasts for the next two weeks. Omitting the explainable outliers (for instance, the holidays), and replacing them with expectable numbers could lead to better forecasting results.

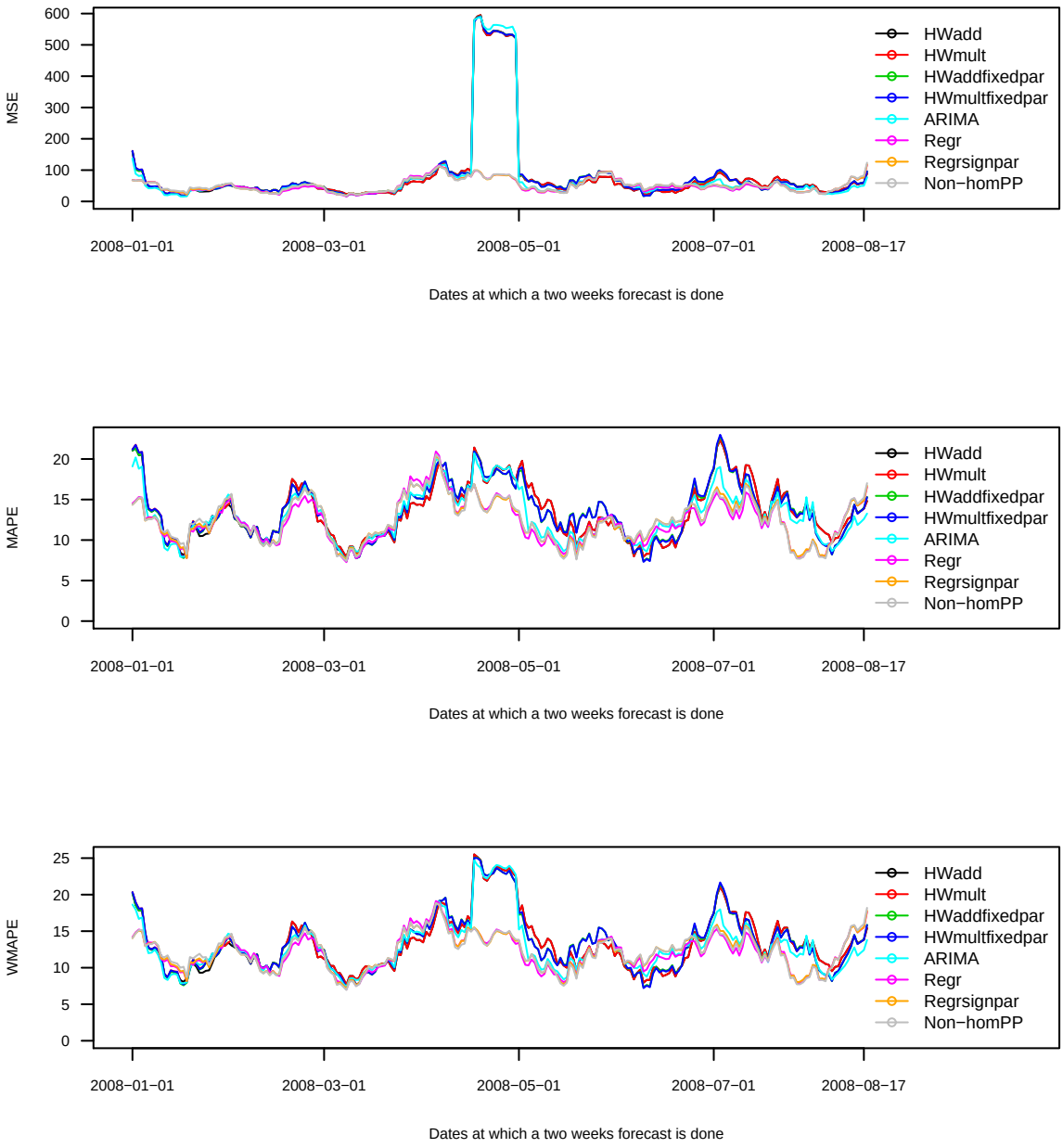

Figure 4.5: *For eight forecasting models the MSE, MAPE and WMAPE are determined for each day of the test set at which a 2 week forecast is provided of the number of A1 rides per day.*

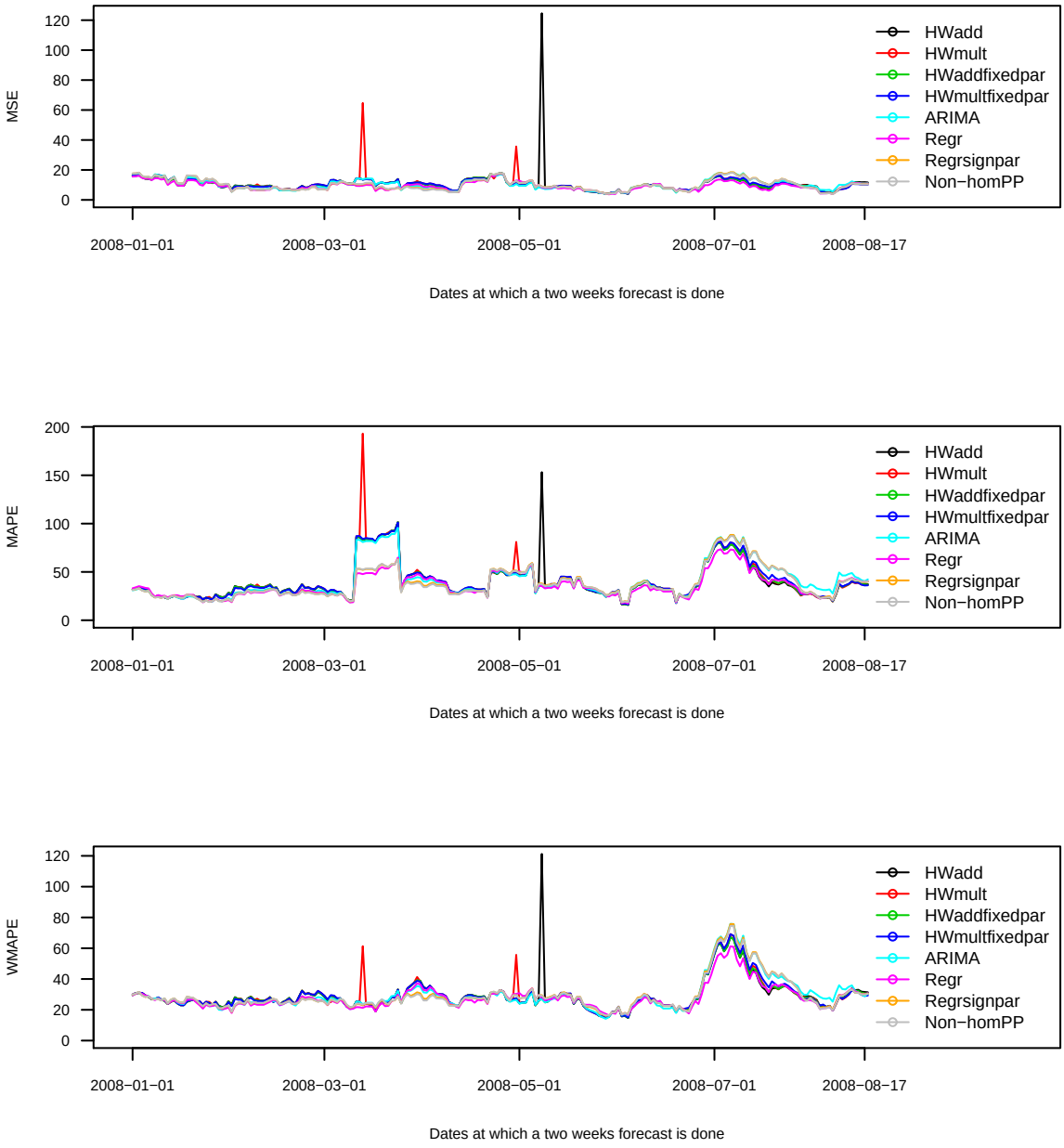

Figure 4.6: *For eight forecasting models the MSE, MAPE and WMAPE are determined for each day of the test set at which a 2 week forecast is provided of the number of A2 rides per day.*

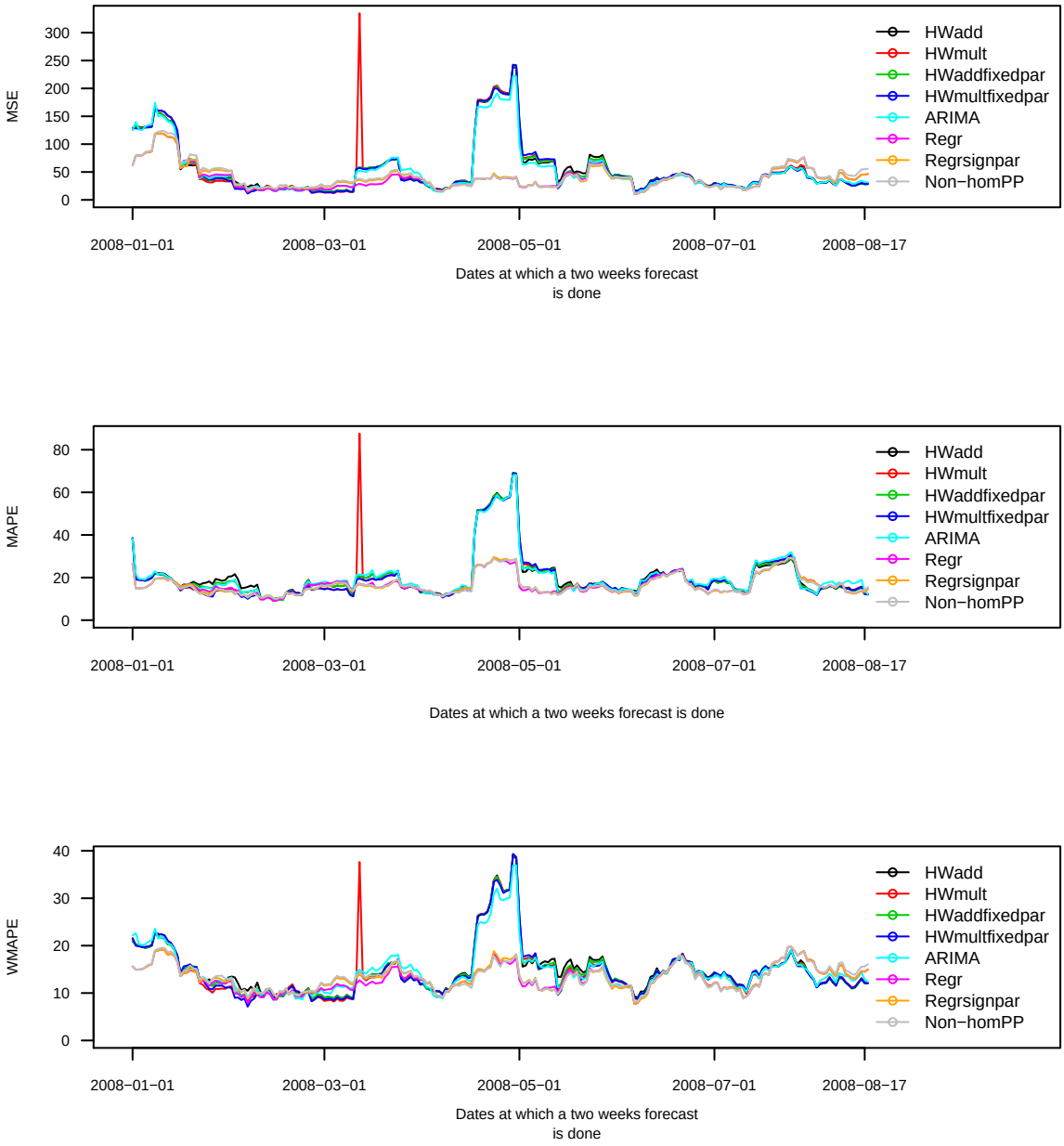

Figure 4.7: *For eight forecasting models the MSE, MAPE and WMAPE are determined for each day of the test set at which a 2 week forecast is provided of the number of B rides per day.*

The plots of the MSE, MAPE, and WMAPE, in Figures 4.5, 4.6, and 4.7, *are not sufficient enough to make it possible to decide which model provides the most accurate forecasts*, id. which model generates forecasts with the lowest MSE, MAPE, and WMAPE. That is why we also compared the mean and the median of the obtained MSEs, MAPEs, and WMAPEs. (We also consider the median since some of the extreme values of the forecast measures were caused by explainable outliers in the data.)

A1 rides		
MSE:	Lowest mean:	Regression
	Lowest median:	Regression
MAPE:	Lowest mean:	Regression
	Lowest median:	Regression
WMAPE:	Lowest mean:	Regression
	Lowest median:	Regression

A2 rides		
MSE:	Lowest mean:	Regression
	Lowest median:	Regression
MAPE:	Lowest mean:	Regression
	Lowest median:	Regression
WMAPE:	Lowest mean:	Regression
	Lowest median:	Regression

B rides		
MSE:	Lowest mean:	Regression
	Lowest median:	Multiplicative Holt-Winters
MAPE:	Lowest mean:	Regression (only sign. parameters) & Non-homogeneous Poisson process
	Lowest median:	Regression (only sign. parameters)
WMAPE:	Lowest mean:	Regression
	Lowest median:	Regression

Table 4.7: *For each of the ambulance ride priorities, the forecasting models which provided the lowest median and mean of the MSE, MAPE and WMAPE, based on a prediction of the number of rides per day for a two week forecast horizon, are shown.*

4.4 Conclusions

In the beginning of this chapter we formulated several research questions. In this section we will try to provide some answers.

What forecasting methods employ the observations of the number of ambulance requests?
We used four different approaches to model and forecast the number of rides per day, demanded for ambulances stationed in Amsterdam. These are: the exponential smoothing methods of Holt and Winters with a seasonality of seven days, multiple linear regression models that consider the day of the week, the month of the year and holiday effects, ARIMA models which combine autoregressive and moving average models on possibly integrated data, and the non-homogeneous Poisson process, based on queueing theory.

The choice of these models is based on literature and data analysis applied to the number of ambulance rides. The exponential smoothing methods, and the ARIMA models are based on correlations between successive days, the difference between the number of rides on a day and the number of rides two days ago, and between the number of rides seven days ago, and a seasonality factor. The modeled multiple linear regression models, and the non-homogeneous Poisson process take into account the influence of the month of the year, the day of the week and a number of holidays.

How can one measure the performance of a forecasting model?
A forecasting model preferably provides accurate forecasts, but we also want the model to provide a decent fit to the data on which the model is based. The goodness of fit to the subset of our data, is determined by measures based on the difference between the value of the fit and the corresponding value in our data set. The measures we used to determine the goodness-of-fit are the mean squared error and Akaike's information criterion. The first puts more weight to large errors of the fit since these are squared. Akaike's information criterion takes into account the number of estimated parameters, and the maximum likelihood of the variance of the residuals. For each parameter that needs to be estimated, an error can occur which can have a negative effect on the fit. The variance of the residuals indicates whether the size of the errors of the fit are equally spread. To determine the accuracy of different forecasts we make use of the difference between the predicted number of rides per day, and the actual number of dispatched rides on the corresponding day in our test set. The measures we used are the mean squared error, the mean absolute percentage error, and the weighted mean absolute percentage error. The mean absolute percentage error has the negative quality that a relatively large forecast error can be overlooked when the corresponding actual number of rides is large, and a small forecast error can be given much weight when the corresponding actual number of rides is also small. Hence we introduced the weighted mean absolute percentage error, which multiplies the standardized errors with the influence of that observation compared to all observations of that forecast.

Which forecasting model is preferable?
The performance of a forecasting model depends on the model, the choices made according to the model (for instance, which parameters need to be included) and above all, the data.

When determining the two-weeks forecasts according to the Holt-Winters additive and multiplicative models, we recalculate its parameters for every day t. For some days in the test set it occurred that the obtained parameters, in our context, had meaningless values, so to speak. Hence we also produced the forecasts with constant parameter values, based on the number of ambulance rides per day of the days November 1, 2005 until December 31, 2007 in our data set.

We calculated the AIC values for different orders of the autoregressive, integrated, and moving average parts of the ARIMA model, respectively, to decide which ARIMA model fitted best to our data. The statistical software we used is R, which for some combinations of parameters took a long time calculating the AIC value, or produced an error. Hence determining which order to choose can be complicated to program, since the best working order can differ in the future.

To determine which parameters to in- or exclude in the regression models we made use of the data-analysis carried out in Chapter 3. When more data is available (which will happen in the future), those decisions can be made with more certainty, and probably even more categories/effects can be included, for instance, the summer holidays and big events organized in Amsterdam. The residuals obtained after modeling a regression model to the number of B rides per day showed correlation. Hence we had to include an ARIMA model. When one wants to automate the forecasting process a check for correlation needs to be added. It is most probable that the performance of the non-homogeneous Poisson process can be increased. The parameter μ_t we used is only based on the first part of our data set, we did not update these. The data fits well to the non-homogeneous Poisson distributions, hence we can make use of this model in the next chapter where we are going to decide how many ambulances should be staffed per hour.

Our data contains some (explainable) outliers, and some models cannot cope with them in a profitable manner as explained in the previous section. One can choose to ignore these outliers, and the incorrect forecast caused by them. This because the ambulance planner already decides to schedule extra ambulances on, for instance, New Year's eve and Queen's Day.

Our research showed that the best models fitting to our data, and perform best on predicting the number of ambulance rides on a certain day are overall the regression models. This can be explained by that our data that showed month of the year, day of the week, and holiday effects, which are all taken into account by the regression models.

In the next chapter we will use the obtained non-homogeneous Poisson processes, describing the number of A1, A2, and B rides seperately, to staff the ambulances.

Chapter 5

Staffing

Based on the Erlang loss model, different staffing models are introduced in Section 5.1. By use of the modified offered load approximation, we apply traffic characterizations of the ambulance requests into these staffing models. We consider the located patterns in incoming requests, and the distinct travel time distributions. In Section 5.2 we will determine for the current schedule the costs and the performance for an average week, and compare those to the obtained staffing levels. In Section 5.2.1 we investigate the performance of staffing levels based on each of the priority requests exclusively and a combination of these. We will finish in Section 5.3 with conclusions.

The research questions are: 'How to determine the quality of a certain staffing method?', 'What is the performance of the current staffing method?', 'Which staffing models are applicable to our data?', 'Should the ambulance rides be staffed for each of the priorities separately?' and 'Which of the considered staffing models provides the best staffing levels?'

The ambulance service of the GGD schedules the same number of ambulances each week of the year. On special days extra ambulances are scheduled. Examples are Queen's Day and New Year's Eve. (Our data analysis in Chapter 3 indicated that on those days more ambulance rides were executed.) For each of the weekdays the same schedule is maintained, but on the nights of Friday to Saturday and Saturday to Sunday a small number of extra ambulances are scheduled. In Figure 5.1 the current staffing levels are given for the ambulance rides of a regular week by a thick line, and the thin line denotes the average number of dispatched rides for each hour of the week. Compared to the other days of the week, during the weekend less ambulance requests can be seen, and in Figure 5.1 it is shown that the average travel time is shorter. Based on these observations one would expect that during the weekend less ambulances should be scheduled, instead of more which is the case. An explanation for this staffing decision could be that during the weekend more high priority

requests enter the dispatching system. To secure the criterion of reaching the emergency location within fifteen minutes, sufficient ambulances should be scheduled.

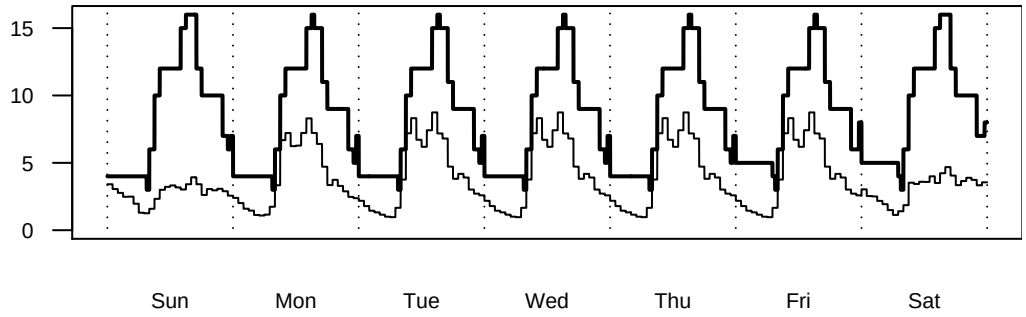

Figure 5.1: *Current staffing levels for each week, denoted by the thick line. For each hour of the week the average of the number of dispatched ambulances per hour is denoted by the thin line.*

5.1 Staffing Models Based on the SRSS Rule

The staffing models we will address are based on queueing theory. Literature shows compelling reasons to assume that the incoming emergency requests for ambulances follow a non-homogeneuous Poisson process. The data we investigated contains just the ambulance rides dispatched to one ambulance station. Incoming ambulance requests immidiately go into service when an ambulance is free. When all ambulances are occupied, another ambulance provider is contacted by the emergency call center. *The queueing model describing the process where incoming calls get lost in the system, is the Erlang loss model, $M/G/c/c$.* In this model the ambulance requests arrive according to a Poisson process with a constant rate, the occupancy times (travel times) are independent and identically distributed with some general distribution G, and c ambulances are available [28].

To optimize the staffing of personnel a lot of research has been done for, for instance, call centers, see [5], [21], and [22]. A simple staffing method is the square root safety staffing (SRSS) rule, which we will discus briefly, for more information see [4], [37].

An exceptionally accurate and robust method to determine the number of servers for the $M/M/c$ queue, such that the corresponding delay probability stays below a certain α, is the square-root safety staffing rule (SRSS) [10]. The queueing model on which this rule is applicable, allows incoming calls to be put in the queue when all c servers are busy. Although we do not consider such a model since requests for ambulance rides cannot be put on hold, the SRSS rule is still of great interest [4].

The SRSS rule assumes a fixed offered load $\rho = \lambda/\mu$, with arrival rate λ, and with μ the expectation of the travel times. In Chapter 3 we determined that the demand for ambulances, and the duration of their occupancy varies over time. In [4] the use of a modified offered load (MOL), according to the arrival rate, is discussed, which we will implement to obtain staffing levels that take the patterns of ambulance requests into account. Since the travel times of the ambulance rides differ over time, we adjust the MOL by applying the different travel time distributions categorized in Section 3.5 to obtain a more reliable staffing method.

5.1.1 Square Root Safety Staffing

The SRSS rule determines for a fixed service grade $\beta \in \mathbb{R}_+$ and an offered load $\rho = \lambda/\mu$, (with λ the arrival, and μ the service rate), a recommendation of the number of servers to schedule. When the requests for ambulance rides would enter the system regularly over time, and the travel time of the ambulances would have a constant value μ, the number of necessary ambulances would be ρ, the mean number of occupied ambulances. But an extra number of ambulances should be staffed, since in reality this is not the case. The SRSS rule regulates this; when the offered load and the number of ambulances are large enough, the staffing level s can be determined according to:

$$s = \rho + \beta\sqrt{\rho}. \tag{5.1}$$

A measure of the performance of a staffing level s for the $M/G/c/c$ model is the corresponding Erlang's blocking probability [28]:

$$B(s, \rho) = \frac{\rho^s/s!}{\sum_{k=0}^{s} \rho^k/k!}. \tag{5.2}$$

Let α be a pre-specified maximum level for the blocking probability, the optimal staffing level can be determined by

$$s^* = \arg\min\{s \in \mathbb{N}_+ : B(s, \rho) \leq \alpha\}. \tag{5.3}$$

Hence for a known ρ and a fixed α, by varying the value of s, the minimal staffing level such that the corresponding blocking probability is less than α can be determined. By filling in ρ and the determined staffing level s in Equation (5.1), the service grade β can be calculated.

5.1.2 Modified Offered Load Approximation

Our data of the executed ambulance rides does not behave like a $M/G/c/c$ model, since the rate of the number of ambulance rides is time-dependent. Hence we schould aply a $M_t/G/c/c$ model, the number of ambulance requests in this model follows a Poisson distribution with a time-dependent mean $m(t)$. The Modified Offered Load (MOL) approximation for the $M_t/G/c/c$ queue, based on the infinite-server queue $M_t/G/\infty$, performs well especially when the blocking probability is not too large [4]. With MOL we use the stationary loss model but replace the instantaneous offered load by the "modified" offered load $m(t)$ of the infinite-server model with time-dependent arrivals. The mean of the number of occupied ambulances in the system at time t can be expressed as

$$m(t) = \int_{v=0}^{\infty} \lambda(t-v)\mathbb{P}(T > v)\mathrm{d}v, \tag{5.4}$$

where T denotes the travel time distribution. In Section 3.14 we determined the parameters for the non-homogeneuous Poisson process describing the number of incoming ambulance requests per hour, hence we choose $\lambda(t)$ piecewise constant for each hour. For T we select the empirical travel time distribution. We determined the MOL approximation of the number of occupied ambulances during an average week for each of the ambulance priorities separately (see Page 73 for a clarification of this average week.) In Figure 5.2 we plotted the total number of occupied ambulances throughout this average week.

To staff the number of required ambulances we apply the SSRS rule (5.1), but we replace the instantaneous offered load by the"modified" offered load $m(t)$ of the infinite server model with time-dependent arrivals. The approximate blocking probability at time t becomes:

$$B_t \approx B(s, m(t)) = \frac{m(t)^s/s!}{\sum_{k=0}^{s} m(t)^k/k!} \tag{5.5}$$

Given the modified offered load, $m(t)$, and the pre-specified level α, we take the approximate blocking probability, defined in Equation (5.5), to determine the service grade β as in Equation (5.1), by considering the current staffing level and by taking $\bar{m}(t)$, the average value of $m(t)$, as ρ.

5.1.3 Adjusted Modified Offered Load Approximation

Not only does the number of ambulance requests differ significantly over time, also the ride durations show day of the week and intra-day patterns. Hence we adjust formula (5.4) to obtain a modified offered load $M(t)$ which also takes the different travel time distributions

into account:

$$M(t) = \int_{v=0}^{\infty} \lambda(t-v)\mathbb{P}(T_{t-v} > v)\mathrm{d}v, \tag{5.6}$$

where T_{t-v} denotes the empirical travel time distribution applied for time $t-v$. To determine $\mathbb{P}(T_{t-v} > v)$ we make use of the data analysis executed in Section 3.5. Here we divided the days of the week and hours of the day into different categories, each having the same travel time distribution, for the A1, A2, and B rides seperately.

In Figure 5.2 we plotted the total number of occupied ambulances throughout an average week, based on the MOL approximation, and based on the adjusted MOL approximation. *Taking into account the difference in travel time distributions does not result in a remarkable different approximation of the number of occupied ambulances.*

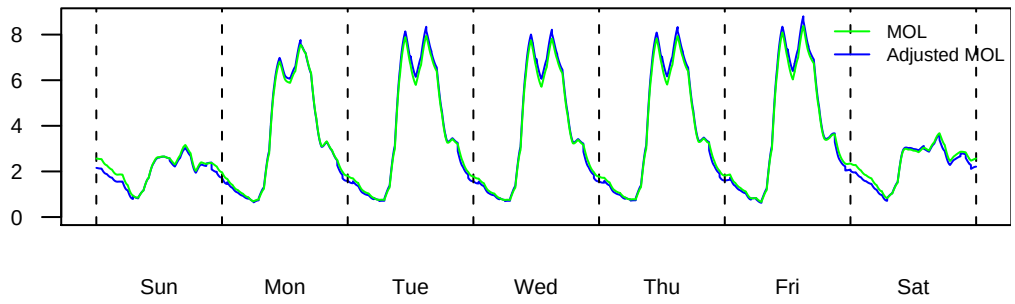

Figure 5.2: *For each of the priorities the modified offered load approximation and the adjusted adjusted modified offered load approximation are determined separately for an average week. The combined MOL and adjusted MOL approximations are denoted in respectively green and blue.*

5.2 Numeric Results

The discussed staffing methods generate a staffing level which considers the pre-specified permitted blocking probability of the incoming request of the call center α. To determine the value of α we analyze the current staffing of the ambulance service for an average week. As the "average week" we take the mean number of A1, A2 and B rides factorized by the hourly patterns obtained in Section 3.3, and hence obtain for each hour of the week the average of the number of requested ambulance rides. The analyzed travel times are given in minutes, hence we will use in this chapter time steps of one minute. To determine the blocking probability (5.2) per minute for an average week of the current staffing level, we use $M(t)$ (defined in Equation (5.6)) as the load ρ_t at time t.

The performance of a staffing method can be discussed by the corresponding blocking probability, defined in Equation (5.2), we want the blocking probability to be low. We determine the performance of a staffing method not just by the blocking probability on time t, but with the blocking probability at time t multiplied by the number of expected incoming rides on time t, $\lambda(t)$. The occurence of all ambulances being occupied is more unfavorable on busy hours.

$$\text{Performance}(s) = \sum_t \lambda(t) \cdot B(s, m(t)). \tag{5.7}$$

We will also compare the costs of the staffing levels to costs of the current schedule. We define these costs as the sum of the number of ambulances scheduled per hour over the regular week:

$$\text{Costs}(s) = \sum_t s_t. \tag{5.8}$$

A plot of the current staffing level, the value of $M(t)$, and the corresponding blocking probability for the considered average week, are given in Figure 5.3. *Based on the current ambulance schedule, the maximum attained blocking probability is* 0.1174, hence we take $\alpha \in \langle 0, 0.12]$, and investigate for different α the performance and costs of the maintained staffing levels. *The performance*, see Equation (5.7), *of the current staffing level is* 10.28, a value which in itself does not give much information, but can be compared to the performances of other staffing levels. *The costs*, see Equation (5.8), *of the current schedule are 1432 ambulances.*

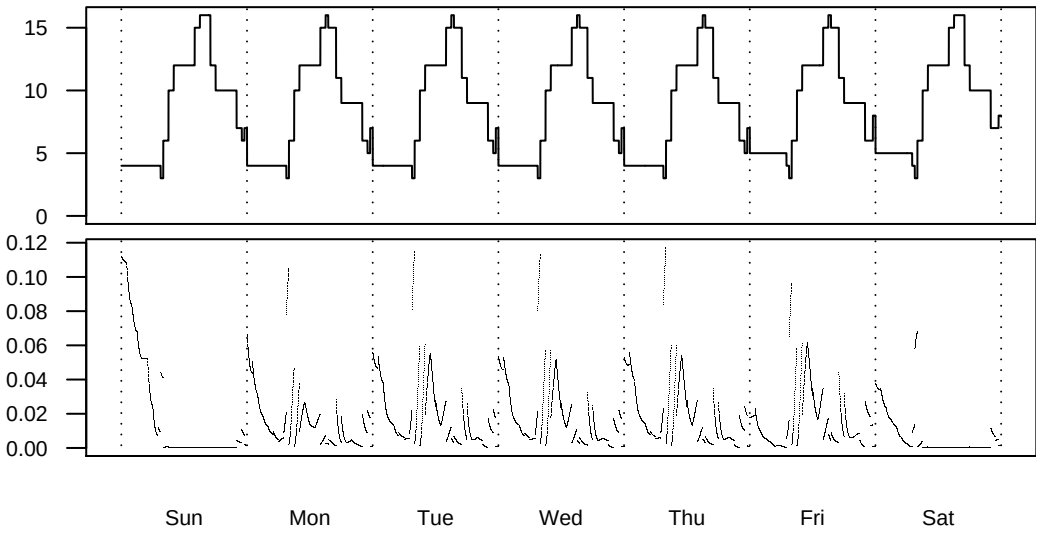

Figure 5.3: *Current staffing level, denoted by the thick line, and an average of the number of ambulance rides per half hour, denoted by the thin line.*

One could consider to distinguish between the priority of the ambulance rides, and establish a staffing method for each of them separately. During both A2 and B rides, the ambulance driver is not allowed to make use of light signals and sirens, and should abide the standard speed limits. An ambulance driver is only allowed to ignore the regular speed limits when he or she is trained to drive at high speed during busy traffic. An ambulance dispatched to a B ride, to transport patients between hospitals or a hospital and the patient's home, does not necesarily need to be equipped with all medical tools necessary during an A1 ride. For each of the discussed staffing methods, we provide staffing levels of the total number of required ambulances. We make a distinction between the staffing levels based on the number of occupied A1, A2, and B rides seperately, and staffing levels based on the total number of occupied ambulances.

In Figures 5.4, 5.5, 5.6, 5.7, 5.8, and 5.9 the relationship between the pre-specified α and the corresponding costs, and α and the performance, based on the different staffing levels, is given. The costs and performance of the current staffing are denoted by a dotted line.

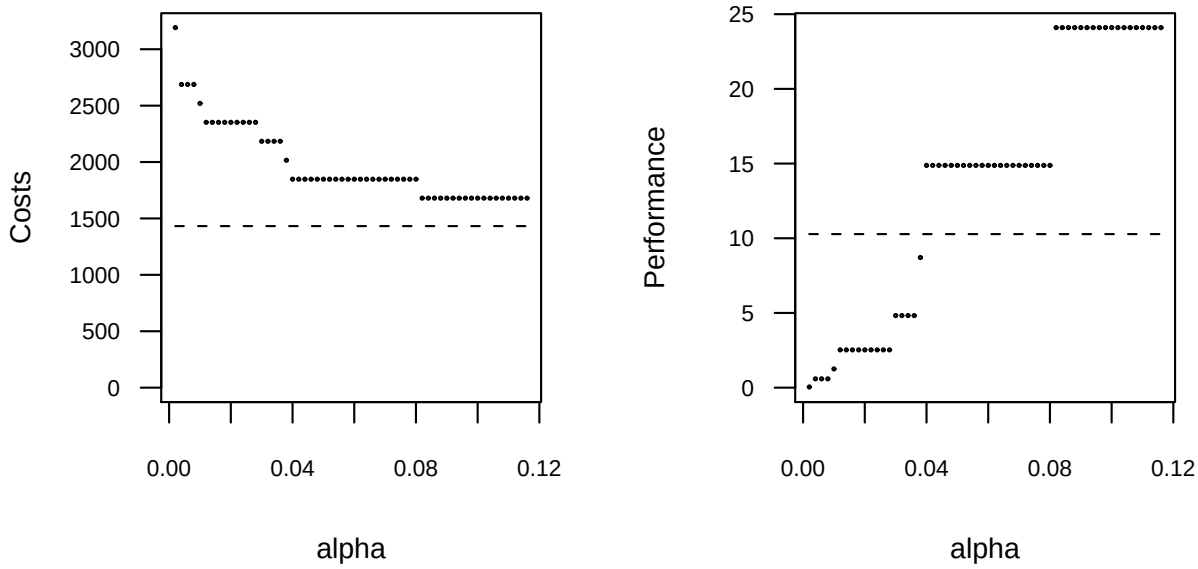

Figure 5.4: *Relation between α and costs, and between α and the performance for a staffing level based on the SRSS rule; the three priorities are staffed separately. The dotted lines denote the costs and performance of the current staffing levels.*

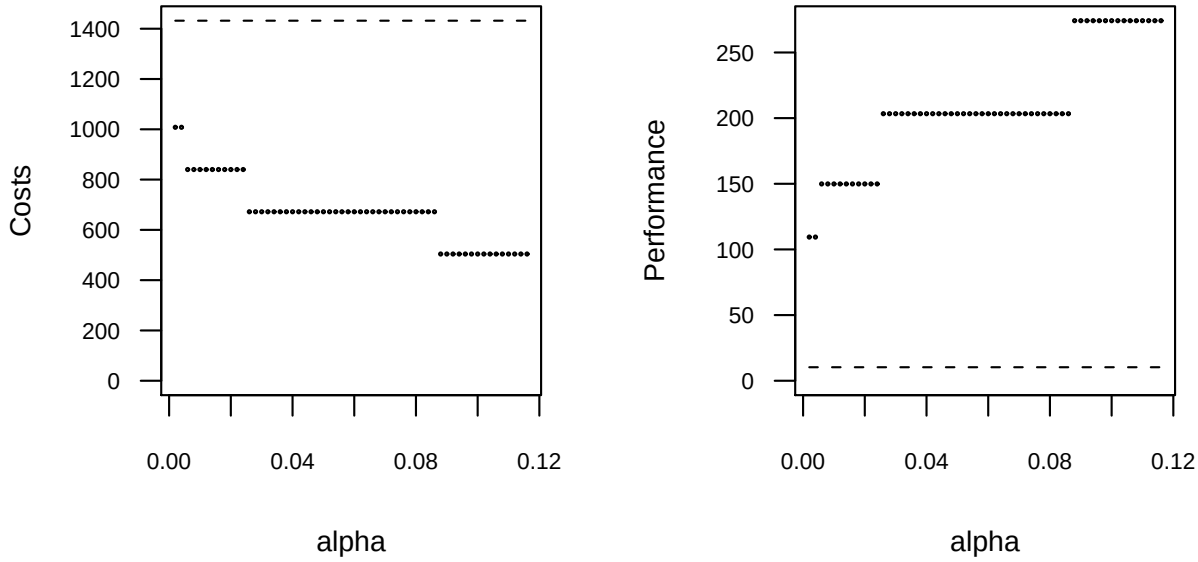

Figure 5.5: *Relation between α and costs and between α and the performance for a staffing level based on the SRSS rule; the three priorities are combined. The dotted lines denote the costs and performance of the current staffing levels.*

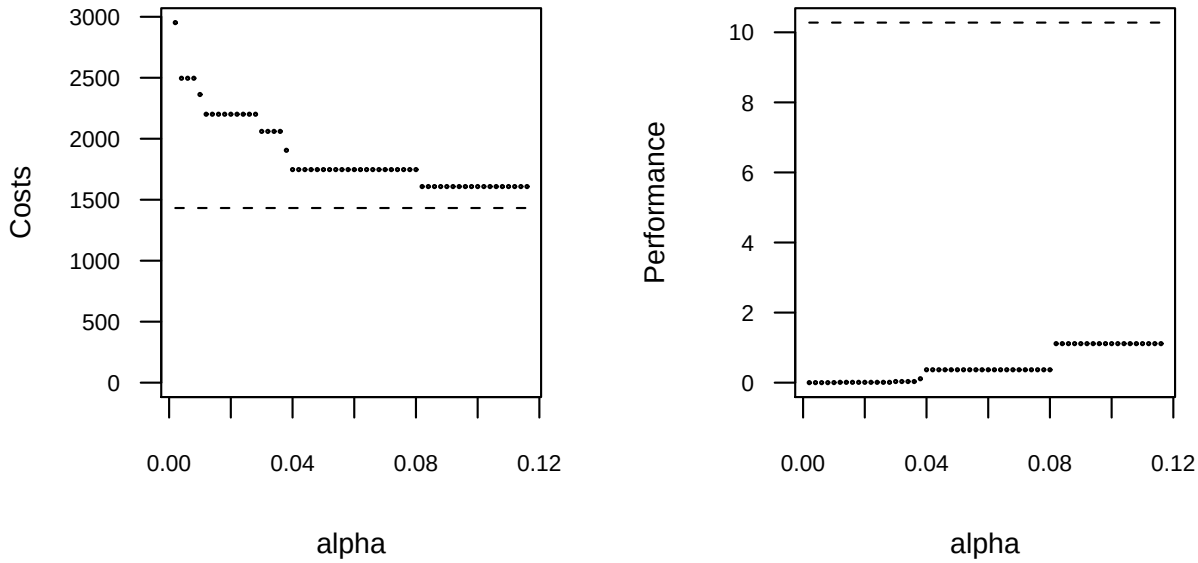

Figure 5.6: *Relation between α and costs and between α and the performance for a staffing level which made use of the MOL approximation; the three priorities are staffed separately. The dotted lines denote the costs and performance of the current staffing levels.*

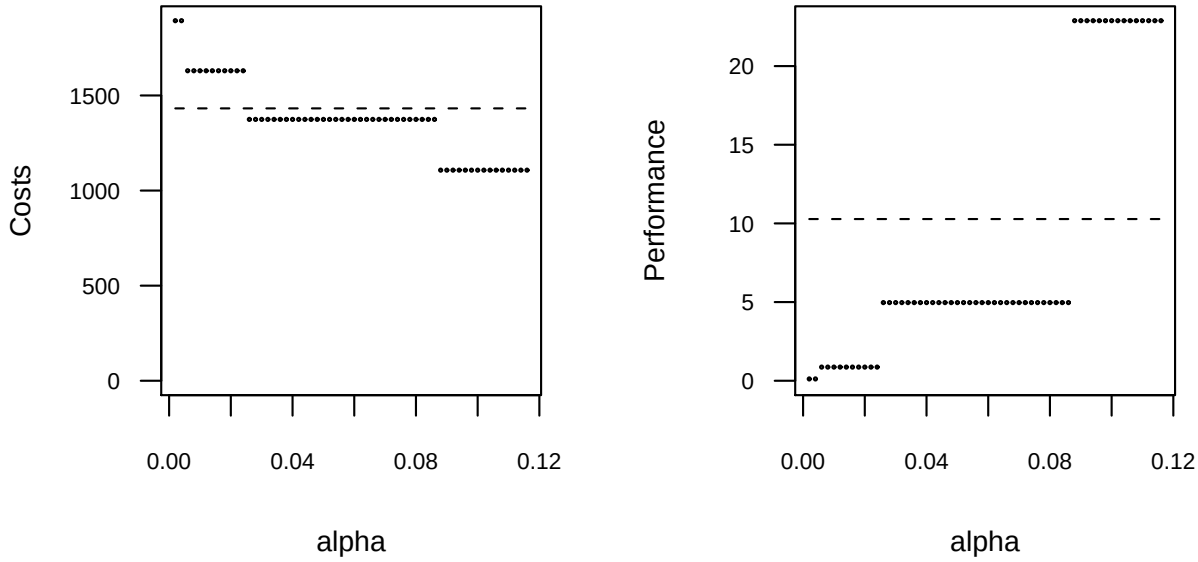

Figure 5.7: *Relation between α and costs and between α and the performance for a staffing level which made use of the MOL approximation; the three priorities are combined. The dotted lines denote the costs and performance of the current staffing levels.*

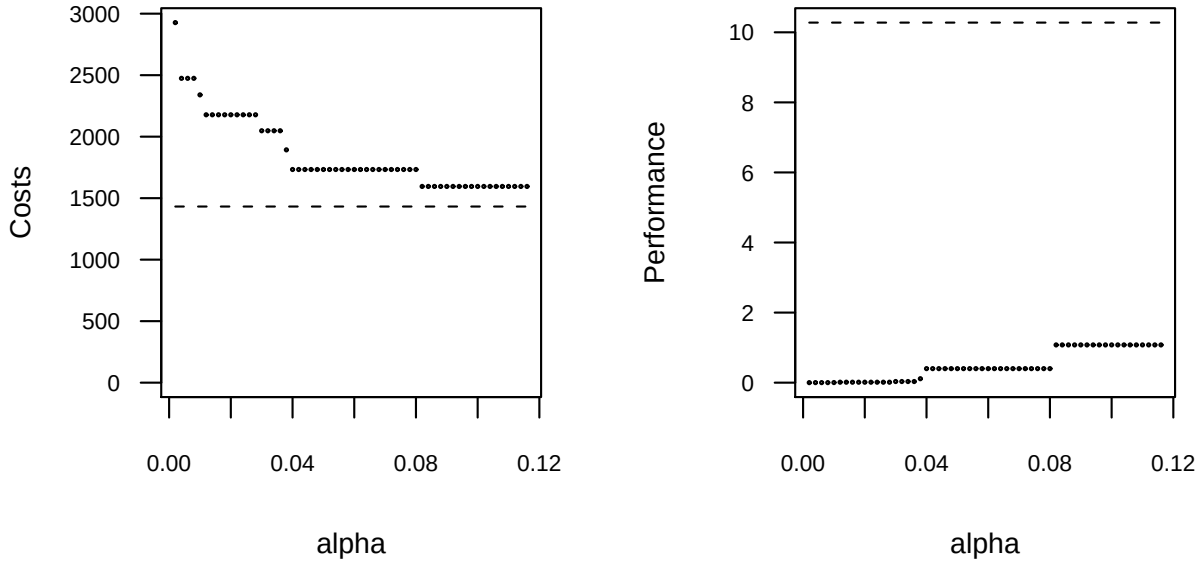

Figure 5.8: *Relation between α and costs and between α and the performance for a staffing level which made use of the adjusted MOL approximation; the three priorities are staffed separately. The dotted lines denote the costs and performance of the current staffing levels.*

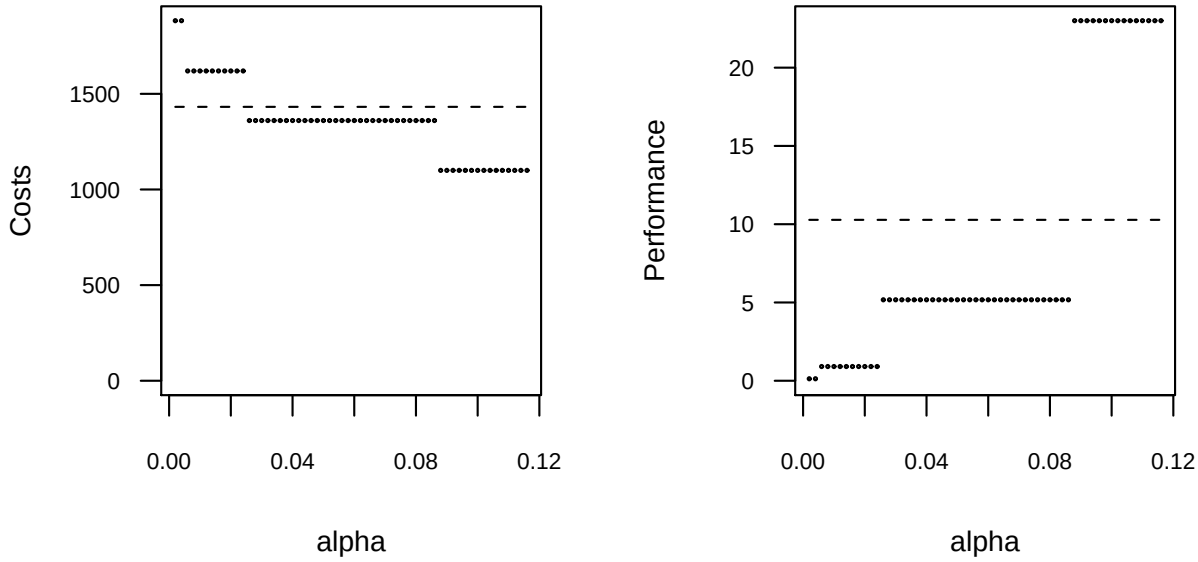

Figure 5.9: *Relation between α and costs and between α and the performance for a staffing level which made use of the adjusted MOL approximation; the three priorities are combined. The dotted lines denote the costs and performance of the current staffing levels.*

A staffing method which we would suggest to the ambulance service is derived from a small α, generates equal or lower costs (5.8) than the current maintained schedule, and scores low on performance defined in Equation (5.7).

For $\alpha \in \langle 0, 0.12]$, the staffing levels based on the SRSS rule score bad in performance or in costs. The staffing levels based on the MOL approximation, where only the difference in arrival patterns is taken into account, and the staffing level is based on the values of $m(t)$ of A1, A2 and B combined, score well for α between 0.025 and 0.085. When for each of the priorities a staffing based on the MOL approximation is determined, these staffing levels altogether score well on performance, but bad on costs. Taking into account the differences in travel times of ambulance rides on different hours, does not contribute that much apparantly. *The performances and costs of the staffing levels where the MOL approximation was used are alike compared to those staffing levels based on the adjusted MOL approximation.*

The maximum value of α we considered, 0.12, is the current maximum attained blocking probability. By obtaining the different staffing levels we use α to determine β (see Equations (5.1, (5.2), and (5.3)). Then β and ρ, $m(t)$ or $M(t)$ provide us our staffing levels. Since β is determined by the average number of busy ambulances, it is possible that an obtained staffing level causes higher blocking probabilities than the pre-specified α. To be sure we take a reasonable value of α when evaluating the different staffing methods.

5.2.1 Staffing of Ambulances with Different Priorities

One could consider to staff the ambulance rides of the different priorities separately. When an ambulance executes an A2 or B ride, the driver is not allowed to make use of the sirens or light signals, see Page 16. The B rides are rides where a patient is transferred to and/or from a hospital. Hence the costs for the B rides could be reduced by dispatching ambulances that do not contain all the medical equipment which is necessary for a A1 or A2 ride, and the possibility to use sirens and light signals.

As a starting point we take $\alpha = 0.02$, the mean blocking probability for the average week when the current staffing level is employed (see Figure 5.1). To staff the number of ambulances on time t we used the adjusted MOL approximation. Hence we determine the service grade β such that a staffing level determined for the average $M(t)$ generates a blocking probability smaller than the pre-specified α.

In Figure 5.10 the staffing level for the average week is given based on determining $M(t)$ for each of the three priorities separately, and then using their sum to determine the staffing level. The lighter green line denotes the obtained staffing , the darker green line denotes the ceiling of the obtained staffing level. In the second plot the corresponding blocking probabilities are shown, the lighter blue corresponding to the lighter green staffing level, and the darker blue corresponding to the rounded off staffing level.

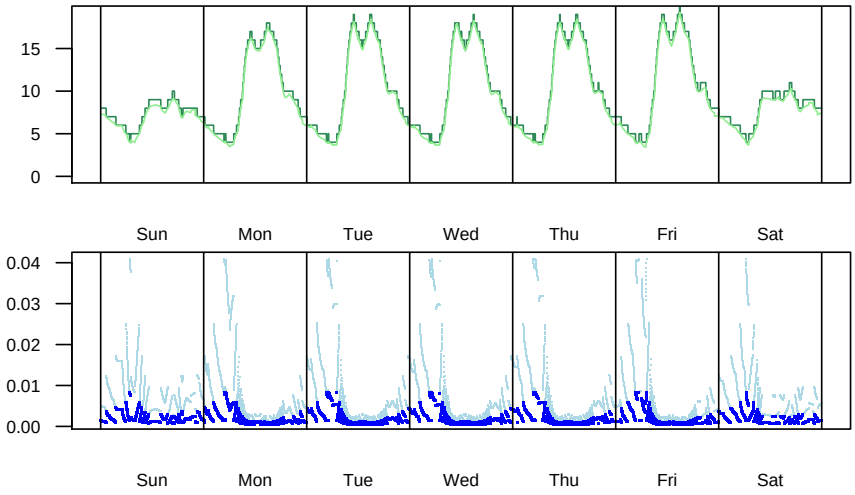

Figure 5.10: *The staffing level, its ceiling, and their corresponding blocking probabilities, for a regular week determined for $\alpha = 0.02$, based on the adjusted MOL approximation determined as the sum of the adjusted MOL for the A1, A2, and B rides separately.*

The blocking probability corresponding to the rounded off staffing level, based on the combination of adjusted MOL approximations obtained for each of the priority rides separately, appears to stay below 0.01, which is a result we like.

When considering the three types of ambulance rides separately, suggesting the staffed ambulances can only execute one of the three types of priorities, we obtain the staffing levels shown in Figures 5.11, 5.12 and 5.13. Again we derived β such that the average number of staffed ambulances results in a blocking probability less than $\alpha = 0.02$.
The blocking probabilities of the rounded off staffing for the A1 and A2 rides behave as desired, they appear to stay under α, whereas the staffing for exclusively B rides generates blocking probabilities of value 0.7. This is caused by the high difference of $M(t)$ during the weekend and weekdays. The service grade β of the staffing level is based on the average value of $M(t)$, and hence the staffed number of ambulances during weekdays, based on this β, thus is too small and understaffing occurs. During these moments of understaffed B rides, the call center operator can decide to deploy ambulances kept for A1 or A2 ambulances to B rides.

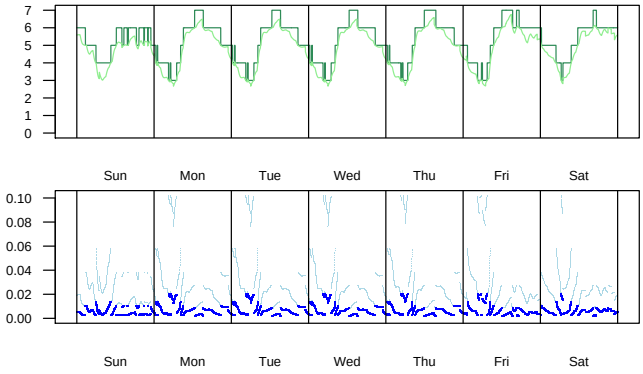

Figure 5.11: *The staffing levels exclusively for the A1 rides, its ceiling, and their corresponding blocking probabilities, for a regular week determined for $\alpha = 0.02$, based on the adjusted MOL approximation.*

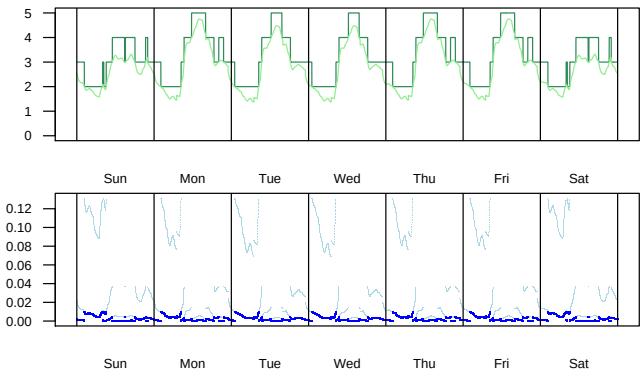

Figure 5.12: *The staffing levels exclusively for the A2 rides, its ceiling, and their corresponding blocking probabilities, for a regular week determined for $\alpha = 0.02$, based on the adjusted MOL approximation.*

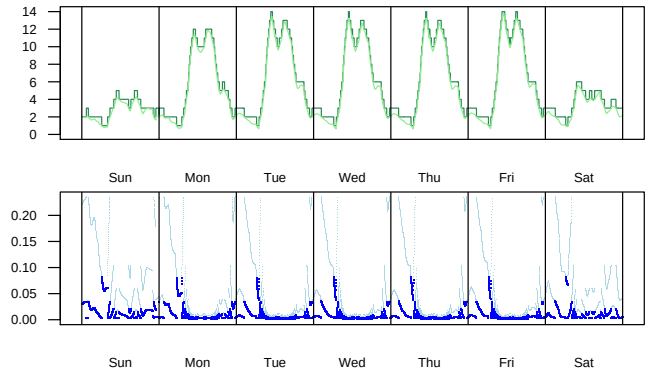

Figure 5.13: *The staffing levels exclusively for the B rides, its ceiling, and their corresponding blocking probabilities, for a regular week determined for* $\alpha = 0.02$*, based on the adjusted MOL approximation.*

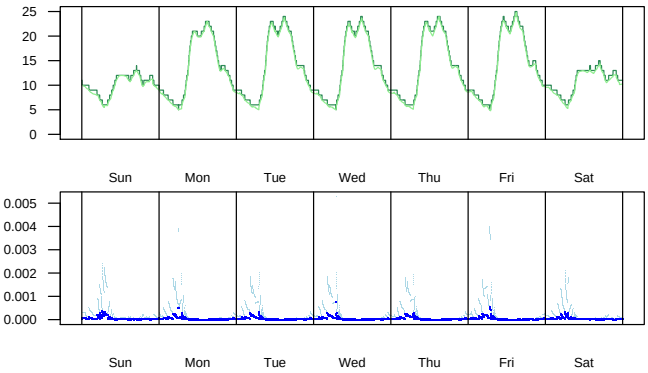

Figure 5.14: *The staffing level, taken as the sum of the staffing levels based on the adjusted MOL approximation determined for* $\alpha = 0.02$ *for the A1, A2, and B rides separately, its ceiling, and their corresponding blocking probabilities.*

When taking the staffing levels for each of the different urgent ambulances combined, and letting the ambulances be able to be dispatched to every requested ride, we obtain overstaffing, see Figure 5.14. The blocking probabilities for each of the instants over the week, have a value less than 0.001, which is desirable. But compared to the current staffing level (see Figure 5.1), the total number of planned ambulances is clearly higher, and hence the costs. When the staffing levels of each of the separate priorities are combined, overstaffing is probable. The buffer generated by β for each staffing level results in an excessive buffer.

One could decide to determine more than one β for the B rides. Based on just the first a staffing level for the weekdays can be obtained, and the second β can be used to obtain a staffing level for the B rides during the weekend. Reserving ambulances on weekdays just for B rides can result in a redundancy of costs. Since on those days enough requests for B rides enter the dispatching center.

5.3 Conclusions

In the beginning of this chapter we formulated several research questions. In this section we will try to provide some answers.

How to determine the quality of a certain staffing method?
To compare different staffing models we provided staffing levels for one week, based on the mean number of ambulance rides per priority for every hour of the week. The dispatching process of ambulance rides can be considered as an $M/G/c/c$ queue, where c denotes the number of scheduled ambulances [7]. Based on queueing theory the probability of all ambulances being occupied, the blocking probility can easily be derived. We determined the performance of a staffing level not just by the blocking probability, but with the blocking probability multiplied by the number of expected incoming rides. We chose for this performance measure to put extra weight on the blocking probability on busy moments. An economic measure of a certain staffing is provided by taking the total number of ambulances staffed in one week, by summing the number of scheduled ambulances per hour. We defined these as the costs of a staffing method. A staffing method which we would present to the ambulance service would generate lower costs than the current maintained schedule, and score low in performance compared to the performance of the current staffing levels.

What is the performance of the current staffing method?
To obtain the blocking probability generated by the current staffing method we determined an adjusted modified offered load approximation (adj. MOL). Based on the time-dependent arrival rate of ambulance requests, and the empirical densities (categorized by priority, day of the week, and hour of the day), this approximation provides the number of occupied ambulances on any time t during the week. Based on the current ambulance schedule, the maximum attained blocking probability is 0.12, the performance has a value of 10.28, and the costs of the staffing levels are 1432 ambulances.

Which staffing models are applicable to our data?
For the $M/M/c$ queue an exceptionally accurate and robust method to determine the number of servers, such that the corresponding delay probability stays below a pre-specified value, is the square-root safety staffing rule [37]. The dispatching process of ambulance rides cannot be seen as such a queueing model, since ambulance requests cannot be put on hold. Even though this rule is not applicable to our dispatching process, the square root safety staffing rule is still of great interest [4]. By using the modified offered load (MOL) approximations, the square root safety staffing rule is applicable to time-dependent travel rates. By adjusting the MOL approximation, even different empirical densities of the travel times can be taken into account.

Should the ambulance rides be staffed for each of the priorities separately?
Based on practical and economical reasons one could decide to staff the ambulance rides for each of the ride priorities separately. An example is that driving at high speed with ambulances through urban areas requires special training. When not enough trained personnel is scheduled, the ambulance service could decide to point out several ambulances that are only allowed to carry out A2 and B requests. An economic reason could be that for an A1 request a lot of medical equipment should be applicable compared to B rides. An ambulance reserved for B rides could be equiped with less medical instruments.

When the (adjusted) MOL approximation is determined for each of the priorities seperately, the total number of occupied ambulances during the week can be determined. Staffing levels based on the total number of occupied ambulances appear to stay under the pre-specified level of the blocking probability. When a staffing level is determined for each of the priorities separataly, and the total number of required ambulances is determined, we obtain overstaffing, which leads to a low performance, but higher costs. In case the costs of ambulances which will be specifically utilized for B rides are much lower than normal ambulances, one could consider to staff the B rides seperately.

A problem which occurs when the ambulance rides are staffed per priority is that the staffing level for exclusively B rides generates unacceptable high blocking probabilities of 0.7. This can be explained by the big difference of the number of exclusively occupied ambulances for the B rides during the weekend and weekdays. The service grade β of the staffing level is based on the average value of the (adjusted) MOL approximation. Hence the staffed number of ambulances during weekdays, based on this β is too small and the ambulance provider is understaffed. During these moments of understaffed B rides, the call center operator can decide to deploy ambulances reserved for A1 or A2 rides, to B rides. One could also decide to determine a staffing level just for the weekdays, and one just for the weekend, based on two different β. Reserving ambulances on weekdays just for B rides can result in a redundancy of costs, since on those days enough requests for B rides enter the dispatching center.

Which of the considered staffing models provides the best staffing levels?
For different values of the pre-specified blocking probability α, we determined staffing levels based on the square root safety staffing rule. We applied a modified offered load approximation, and an adjusted modified offerd load approximation. The first considers a constant arrival rate of the dispatched requests, and a constant rate of the travel time distribution. The second considers a time-dependent arrival rate of the dispatched ambulance requests, and a constant rate of the travel time distribution. The latter considers a time-dependent arrival rate of the dispatched ambulance requests, and different categories for the travel times, based on the time of departure.

For different values of α, the staffing levels based on the SRSS rule score bad in performance or in costs. The staffing levels based on the MOL approximation, and the staffing levels based on the values of $m(t)$ of A1, A2 and B combined, score well for α between 0.025 and 0.085. The sum of the staffing levels for the A1, A2, and B rides, each based on the MOL approximation, scores well on performance, but bad on costs. The performances and costs of the staffing levels where the MOL approximation was used are alike compared to those staffing levels based on the adjusted MOL approximation.

Distinguishing in the travel time distributions for different hours of the week, does not result in a remarkable different approximation of the number of occupied ambulances i.e., the adjusted MOL compared to the MOL approximations. For the average week the maximum difference between these two approximations is 0.42, and the standard deviation of the differences is 0.19. Obtaining the different departure time categories of travel times, and the travel time distribution per category, takes time, and appeared to be of no outstanding improvement.

Chapter 6

Conclusions

The ambulance requests show for each of the priorities, A1, A2, and B, distinct correlations and arrival patterns. A model to describe the incoming requests should take these correlations, the month of the year, the day of the week, certain holidays, and hour of the day into account.
The non-homogeneous Poisson processes describing the number of A1, A2, and B rides per day in Section 3.4.1 consider arrival pattern effects, but neglect correlations. Even though the number of ambulance rides per day is significantly alike to the Poisson distributions.
To model the number of requests for a certain hour a multinomial distribution is used, conditional on total number of rides that day. In [7] data analysis is performed on ambulance requests. Test results indicated correlations between different hours of the same day. By taking such correlations into account the estimates of the number of requests per hour could become more precise.

The occupancy time of ambulance rides differs significantly according to in which hour the ambulance took of. We first categorized the empirical travel time densities by day of the week, to continue by investigating which hours in a category show significantly distinct distributions. We did not consider month of the year as an effect or certain holidays. While these factors could lead to different categories. Since an empirical density was sufficient to obtain a MOL and adjusted MOL approximation, we neglected to try to fit known densities to the travel time data.

For each of the priorities of the ambulance rides, the multiple linear regression method provided the most accurate fit to the data, and generated the most accurate predictions of a two week forecast horizon. The multiple linear regression model was based on the factors month of the year, day of the week and being it a certain holiday. The model is easy to understand, easy to implement in statistical software, and explainable outliers can be taken into account.

A Poisson distribution with time-dependent mean is a suitable candidate for describing the number of ambulances per day. But the forecasts we provided based on this distribution resulted in large forecasting errors. This is among other things explainable by the fact that we did not update the parameters while generating forecasts based on the data in the test set.

The square root safety staffing rule with a modified offered load approximation, provided staffing levels which score better compared to performance and costs than the current staffing level. We adjusted the modified offered load approximations in such a manner that besides arrival patterns, also differences in empirical travel time distributions were considered. This adjustment did not lead to remarkable differences of the generated staffing levels. Categorizing the travel time distributions is time-consuming, hence one should first investigate whether staffing levels based on just the modified offered load approximation are sufficient enough.
The staffing models discussed in this thesis were based on queueing theory. One could also decide to use discrete optimization to obtain staffing levels. When discretizing is used, not only the number of necessary ambulance can be determined, also an extension to scheduling the shifts of ambulance personnel is posible.

Bibliography

[1] Alsalloum, O. I., & Rand, G. K. (2003). A goal-programming model applied to the EMS system at Riyadh City, Saudi Arabia. Lancaster University Management School. Working paper.

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

[3] Baker, J. R., & Fitzpatrick, K. E. (1986). Determination of an optimal forecast model for ambulance demand using goal programming. *Journal of Operational Research Society 37*(11), 1047-1059.

[4] Bekker, R., & de Bruin, A. M. (2010). Time-dependent analysis for refused admissions in clinical wards. *Annals of Operations Research 178*(1), 45-65.

[5] Bhulai, S., Kan, W. H. & Marchiori, E. (2005). Nearest neighbour algorithms for forecasting call arrivals in call centers. *Technical Report* WS2005-12, VU University Amsterdam.

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

[7] Channouf, N., L'Ecuyer, P., Ingolfsson, A., & Avramidis, A. N. (2007). The application of forecasting techniques to modeling emergency medical system calls in Calgary, Alberta. *Health Care Management Science 10*(1), 25-45.

[8] Dobson, J. (2007). Understanding Failure: The London Ambulance Service Disaster. In G. Dewsbury, & J. Dobson (Eds.), *Responsibility and dependable systems* (pp. 130161). Secaucus, NJ: Springer.

[9] Erdoğan, G., Erkut, E., & Ingolfsson, A. (2010). Scheduling ambulance crews for maximum coverage. *Journal of the Operational Research Society 61*, 543-550

[10] Gans, N., Koole, G., & Mandelbaum, A. (2003). Telephone Call Centers: Tutorial, Review, and Research Prospects. *Manufacturing & Service Operations Management 5*, 79-141.

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

[12] Goldberg, J. (2004). Operations research models for the deployment of emergency services vehicles. *EMS Management Journal 1*(1), 20-39.

[13] Green, L. V., Kolesar, P. J. (2004) Improving emergency responsiveness with management science. *Manage Sci 50* 1001-1014

[14] de Gunst, M.C.M. (2007). *Statistical Models*, Faculty Exact Sciences, vrije Universiteit amsterdam.

[15] de Gunst, M. C. M. & van der Vaart, A. W. (2007). *Statistische Data Analyse*, Faculty Exact Sciences, vrije Universiteit amsterdam.

[16] Harewood, S. I. (2002). Emergency ambulance deployment in Barbados: a multi-objective approach. *Journal of the Operational Research Society 53,* 185-192.

[17] Hayes, J., Moore, A., Benwell, G., & Wong, B. L. W. (2004). Ambulance dispatch complexity and dispatcher decision strategies: implications for interface design. In Masoodian, M., Jones, S., & Rogers, B. (Eds.), *Computer Human Interaction* [Lecture notes] (pp 589-593) Springer.

[18] Henderson, S. G., & Mason, A. J. (1999). Estimating ambulance requirements in Auckland, New Zealand. *In Proceedings of the 1999 Winter Simulation Conference: Simulation: a bridge to the future.* 1670-1674.

[19] Henderson, S. G. & Mason, A. J. (2004). Ambulance service planning: simulation and data visualisation. In: *Handbook of Operations Research and Healthcare Methods and Applications* (eds. F. Saintfort, M.L. Brandeau and W.P. Pierskalla). International series in *Operations Research and Management Science 70,* 77-102.

[20] Ingolfsson, A., Budge, S., & Erkut, E. (2003). Optimal ambulance location with random delays and travel times. *Health Care Management Science 11,* 262-274.

[21] Ingolfsson, A., Cabral, E. & Wu, X. (2002). Combining integer programming and the randomization method to schedule employees. *University of Alberta Research Report 02-1.*

[22] Koole, G. (2005). Call Center Mathematics a scientific method for understanding and improving contact centers. Retrieved from http://www.math.vu.nl/koole/ccmath/book.pdf

[23] Larson, R. C., & Odoni, A. R. (1981). Urban operations research - logistical and transportation planning methods. First published by Prentice-Hall. Retrieved from http://web.mit.edu/urban_or_book/www/book/

[24] Lokhorst, F. (2003). *Ambulancevervoer in de regio Amsterdam.*

[25] Makridakis, S. G., Wheewright, S. C., & Hyndman, R. J. (1998). *Forecasting: Methods and applications* (3rd edition). New York: John Wiley & Sons.

[26] Mendonça, F. C., & Morabito, R. (2001). Analysing emergency service ambulance deployment on a Brazilian highway using the hypercube model. *Journal of the Operational Research Society 52*(3), 261-270.

[27] Miljoenennota 2009, http://www.prinsjesdag2008.nl (outlined in *De Volkskrant*, 17 september 2008, pp.2).

[28] Adan I., & Resing, J. (2001) *Queueing Theory* [Lecture notes].

[29] Restrepo, M. (2008). *Computational methods for static allocation and real-time redeployment of ambulances* (Ph.D. thesis) Cornell University, Ithaca, New York.

[30] Restrepo, M., Henderson, S. G. & Topaloglu, H. (2009). Erlang loss models for the static deployment of ambulances. *Health Care Management Sci. v12*, 67-79.

[31] Setzler, H., Saydam, C. & Park, S. (2009). EMS call volume predictions: a comparative study. *Computers & Operations Research 36*(6), 1843-1851.

[32] Shen, H. & Huang, J. Z. (2005). Analysis of call centre arrival data using singular value decomposition. *Applied Stochastic Models in Business and Industry 21,* 251-263.

[33] Shen, H., & Huang, J. Z. (2008). Interday forecasting and intraday updating of call center arrivals. *Manufacturing & Service Operations Management 10*(3), 391-410.

[34] Singer, M., & Donoso, P. (2008). Assessing an ambulance service with queuing theory. *Computers & Operations Research 35,* 2549-2560

[35] Swersey, A.J. (1994). The deployment of police, fire and emergency medical units. In Pollock, S. M., Rothkopf, M. H. & Barnett, A. (Eds.) *Operations Research and the Public Sector* (pp. 151-200).

[36] Taylor, J. W. (2008). A comparison of univariate time series methods for forecasting intraday arrivals at a call center. *Management Science 54*(2), 253-265.

[37] Tijms, H. C. (2003). A first course in stochastic models. Wiley, New York.