Department of Applied Mathematics
Faculty of Electrical Engineering,
Mathematics and Computer Science
University of Twente


Centrum Volksgezondheid Toekomst Verkenningen (VTV)
Rijksinstituut voor Volksgezondheid en Milieu(RIVM)


*Final Project*

# Optimal Capacity of Ambulance Service System in The Netherlands

Y.Chen
S0205435
September 2009 - June 2010

Department of Applied Mathematics
Faculty of Electrical Engineering,
Mathematics and Computer Science
University of Twente


Centrum Volksgezondheid Toekomst Verkenningen (VTV)
Rijksinstituut voor Volksgezondheid en Milieu(RIVM)


*Final Project*

# Optimal Capacity of Ambulance Service System in The Netherlands

Y.Chen
S0205435
September 2009 - June 2010

Supervisors:
Prof. dr. R.J. Boucherie(University of Twente)
Ir. G.J. Kommer(Rijksinstituut voor Volksgezondheid en Milieu)
Dr. ir. W.R.W. Scheinhardt(University of Twente)
Dr. ir. J.C.W. van Ommeren(University of Twente)

## Summary

Like many other developed countries, there is a sophisticated ambulance service management system in the Netherlands currently. People who dial 112 can get serviced within a extremely short time, in most of the cases the requests can be answered immediately. The well trained centralists will organize an effective triage when the phone gets through. The work procedure for a centralist in general consists of making a decision whether an ambulance is needed after the triage and giving an order to the ambulance station if an ambulance is needed. The main route of the ambulance is leaving the ambulance station, picking up the patients, transferring the patient to the hospital and coming back to the ambulance station. The limited budget for the ambulance service resources and the increasing request for the more efficient ambulance service calls for an improvement in the design of the schedules for the ambulance resources.

By using several queuing models, critical factors influencing the waiting time and staff workload in *call centers* are identified. Among these critical factors are the number of available centralists, the triage skills of centralists and occasionally unexpected events. The probability of delay in a call center can be deduced from these models if the incoming rate of the requests and the service rate of the centralists are known, based on the empirical data. The applications of the models have already been done for several regions in 2008. More investigations can be explored with these queuing models, an optimal centralist staffing level is decided when combination of several call centers is considered. The outcomes of these queueing models indicate there is a considerable decrease for the demanding of centralist if the combination was applied.

Several different research methods are introduced to determine an optimal scheduling for the *ambulances*. Continuous time Markov chain, M/M/c model, M/G/c/c models are applied to develop the suitable models. The result shows that the shortage of ambulances happens rarely, both theoretically and realistically, with the current scheduling level. With the mathematical model, the optimal scheduling levels of ambulances can be known with a prefixed satisfactory criterion that the probability a shortage occurs is less than a constant $\alpha$. In most of the cases, new optimal schedules are just close to the current schedules.

## Abbreviations and Terminology

### Urgency A1

- threat to the life or permanent disability of the patient (e.g. chest pains, breathing difficulties or cardiac arrest)
- with flashing light and siren (the ambulance has priority, road users move to the right or the left to let the ambulance through)
- within 15 minutes.

### Urgency A2

- no direct threat to life, but help is needed quickly (e.g. serious inflammations, such as appendicitis, or accidents with minor injuries)
- possibly with flashing light and siren
- within 30 minutes.

### Urgency B

- miscellaneous (e.g. transport of people to hospital for examinations or treatment).

**Centralist** The well trained worker who works in call center. Their main work includes two parts: to get information from the call maker and to give the order to the ambulance team.

**Delay** In the call centers, the delay means the arriving requests can not be served immediately. In the ambulance stations, the delay means the arriving requests can not get free ambulances immediately.

**Light Traffic** In queueing system, light traffic means that the occupation rate is small.

**LST** Laplace-Stieltjes transform, the Laplace-Stieltjes transform of a real-valued function $g$ is given by a Lebesgue-Stieltjes integral of the form: $\int e^{-sx}\, dg(x)$.

**Occupation Rate** The probability that a server(centralist/ambulance) is busy.

**PASTA** Poisson arrivals see time averages. PASTA states that the fraction of customers finding on arrival $n$ customers in the system is equal to the fraction of time there are $n$ customers in the system if the Poisson arrivals are satisfied.

**Response Time** The time needed from the call gets through until the ambulance arrives at the scene. This time period includes service time in the cal center, possible waiting time for an ambulance becoming available, preparation time in the ambulance station and driving time to the scene.

**Triage** In the call center, the word triage is used to describe a quick interview between the centralist and the call maker. The centralist should get the following information after the triage: what type of urgency it is and whether an ambulance is needed.

## List of Symbols

$c$: Number of servers(centralists/ambulances) in a queueing system

$\lambda_1$: Arriving rate of A1 customer

$\lambda_2$: Arriving rate of A2 customer

$\lambda_B$: Arriving rate of B customer

$\mu_1$: Service rate of A1 customer [1]

$\mu_2$: Service rate of A2 customer

$\mu_B$: Service rate of B customer

$P^c_{delay}$: Delay probability in call center

$P^a_{delay}$: Delay probability in ambulance station

$SA_{CC}$: Sample space of service time in the call center

$SA_{PR}$: Sample space of preparation time in the ambulance station

$SA1_{ij,Pk}$: Sample space of phase $k$ for A1 during time period $j$ on day type $i$.

$SA2_{ij,Pk}$: Sample space of phase $k$ for A2 during time period $j$ on day type $i$.

$SB_{ij,Pk}$: Sample space of phase $k$ for B during time period $j$ on day type $i$.

$i$: day types: 1, weekday; 2, Saturday; 3, Sunday.

$j$: time periods: 1, 0:00-8:00; 2, 8:00-16:00; 3, 16:00-24:00.

$k$: phases of the service time: 1, phase 1; 2, phase 2.

[1] $\mu_1, \mu_2, \mu_B$ denote the service rate in call center and ambulance trip in different models.

# Preface

The main goal of the project described in the proceeding chapters of this report is to determine the optimal capacity of the ambulance service system in the Netherlands. For me, finishing this project is the last step involved in obtaining a Master's Degree in Applied Mathematics. The last ten months provided me with a positive experiences. Therefore I would like to thank Geertjan Kommer and Richard Boucherie for providing me the opportunity to carry out this project.

I would like to thank all of the supervisors for their useful comments, honesty and interest in my project. I would like to thank Werner Scheinhardt, Geertjan Kommer and Jan-Kees van Ommeren for reviewing my report. Also I would like to thank all the colleges in RIVM and UT.

Of course all of my friends in and out of the Netherlands deserve my thanks. And last but not least, I would like to thank my parents for supporting me in almost everything I would like to do.

Yanting Chen
Enschede, July 2010

# Contents

# Chapter 1

# Introduction

This chapter provides the background information about this project and an overview of the report. Section 1.1 introduces the research institute where this project is carried out. Then, an overview of the ambulance service system in the Netherlands is introduced in section 1.2. Finally, the problem definition and outline of the report are presented in section 1.3 and 1.4 respectively.

## 1.1 Background

The purpose of this study is to develop a capacity model for the centralists and the ambulances, in other words, an optimal schedule of centralists and ambulances should be generated. The project is carried out in the VTV department of RIVM.

### 1.1.1 Centre for Public Health Forecasting

This department(Dutch: Centrum Volksgezondheid Toekomst Verkenningen) collects, evaluates, integrates and disseminates knowledge about health care and also explores the consequences for the health care system to give more support to the policy maker. The research result will support the RIVM, Ministry of Health, Welfare and Sport in policy making. The main tasks of this department are:

- Publish integrative research to assist policy modification;
- Present information on the official internet websites;
- Provide international research result in collaboration with EU, WHO and OECD.

### 1.1.2 RIVM

The Netherlands National Institute for Public Health and the Environment (Dutch: Rijksinstituut voor Volksgezondheid en Milieu or simply RIVM), is a Dutch research institute which is an independent agency of the Dutch Ministry of Health, Welfare and Sport.

RIVM performs tasks to promote public health and a safe living environment by conducting research and collecting knowledge worldwide. The results are used to support the Dutch government in formulating its policy. RIVM is located in Bilthoven and employs over 1400 people, many of whom work in multidisciplinary fields.

## 1.2 Introduction of The Ambulance Service System

An overview of the ambulance industry in the Netherlands is shown first. Then a short description of the working procedure for an ambulance trip is presented.

### 1.2.1 An Overview of The Ambulance Service in the Netherlands(Facts & Figures )

An overview of the ambulance service system in the Netherlands is introduced here. All the data used here is from [42].

There is 1003,050 ambulance trips in 2008 including A1, A2 and B trips. There is 439,725 A1 trips in 2008 and the average response time is 9:47 minutes, besides, the percentage of the A1 trips of which the response time is less than 15 minutes is 92.1% [1]. Similarly, there is 223,813 A2 trips in 2008 and the average response time is 15:53 minutes, the percentage of the A2 trips of which the response time is less than 30 minutes is 96.2%. Apart from A1 and A2 trips, the total number of B trips in 2008 is 339,512. Among all the ambulance trips in 2008, there are 786,667 billable trips, 169,977 EHGV trips and 46053 free trips.

The total number of workers in the ambulance service industry in the Netherlands in 2008 is 4865 and 4267 of them are working in the core functional departments. In 2008, there are 24 RAV regions in the Netherlands, 676 ambulances. The budget in 2008 for the ambulance service in the Netherlands is 363 million Euros.

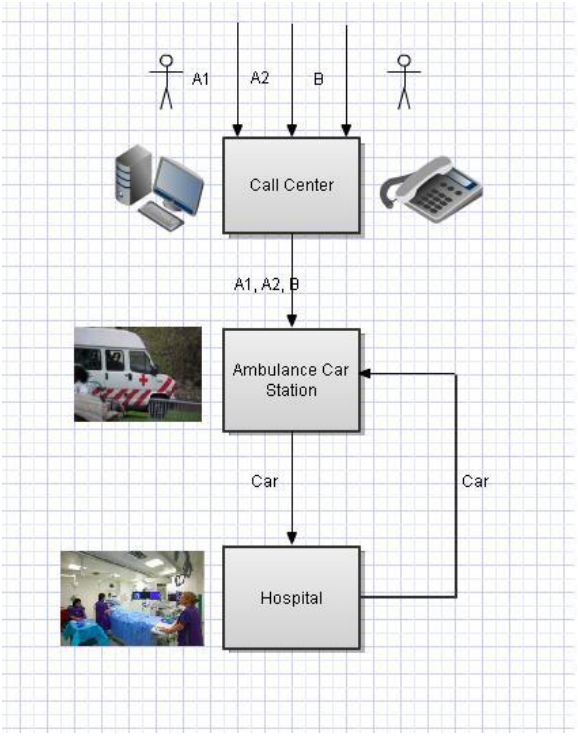
[1]The calculation is based on a specific filter, see the appendix of [42].

### 1.2.2 Ambulance Service System in the Netherlands

In this section, more information about the ambulance system in the Netherlands will be shown. There is a standard working procedure for offering the ambulance service in the Netherlands. A telephone call should be made first and after the triage the centralist will decide whether an ambulance is needed. Large waiting time is not allowed because it may leads people's death, so the most important feature of the ambulance call center is that abundant centralists are needed compared with other commercial call centers. Another point that deserve to mention is that the call centers in the Netherlands are working with other departments such as police, firefighters. This management collaborations can decrease the triage time because more resident information is available simultaneously.

If an ambulance is needed, the standard route for an ambulance trip is leaving the station, picking up the patients, transferring the patients to the hospital and coming back to the ambulance station, the flowchart in Figure 1.2 shows the procedure graphically.

Figure 1.1: Flowchart of Ambulance Service System




As mentioned previously, the optimal schedules of the centralists and the ambulances should be developed. The research conducted to propose a new schedule is guided by the research question stated next.

## 1.3 Problem Definition

The purpose of this study is to develop a capacity model for centralists and ambulances. The possibility of different types of jobs should be taken into consideration, such as A1, A2 and B. Besides, the efficiency of the model results should also be tested.

**Research Question**: *How to determine the optimal capacities and schedules for the centralists in the call centers and the ambulances in the stations?*

## 1.4 Outline

The research executed to improve the capacity and scheduling of the ambulance resources in the Netherlands is presented in the proceeding chapters of this report and organized as follows. An overview of the dataset for this project is presented in Chapter 2. In Chapter 3 an overview of the literature studies is given. The construction of the mathematical model, corresponding process of data calibration and calculation results for staffing the centralists are described in Chapter 4. A similar procedure for scheduling ambulances has been done in Chapter 5. In Chapter 6, two simulations are set up to evaluate the performance of the model results from Chapter 4 and Chapter 5. The report finishes with conclusions and recommendations in Chapter 7, followed by reference list and appendices.

# Chapter 2

# Data Overview

This chapter presents a data overview of this project. There is a short description of the dataset in section 2.1. Then, section 2.1 explains the structure of the dataset for this project. Finally, section 2.3 introduces the service criterions of this project and some general assumptions we will use in the following chapters.

## 2.1 Overview of the Dataset

The informations of the ambulance trips for region Groningen, Friesland, Drenthe are known in the dataset from year 2006 until 2008. We will give focus to the data in 2008. In 2008, there is 43478 ambulance trips in Groningen, 36823 ambulance trips in Friesland and 24601 ambulance trips in Drenthe.

## 2.2 Overview and Structure of the Dataset

The dataset we have in the project is for region Groningen, Friesland and Drenthe. For the applications of the theoretical models, only the dataset of Groningen and Drenthe are used because Friesland does not have a centralized system, which means, they have sub call centers and sub ambulance stations. So for Friesland, if we want to do the data experiment on it, we need to know more information about this trip such as which sub call center and which sub ambulance station are dealing with it instead of only knowing it occurs within Friesland.

Due to the reason that the dataset available in this project gives restrictions to the model construction, an introduction of the dataset available here is needed. There are 8 time moments recorded in the dataset, namely $t_1, ..., t_8$,

- $t_1$: time the phone call gets through;

- $t_2$: time the centralist in call center orders an ambulance [1];
- $t_3$: time the ambulance departs;
- $t_4$: time the ambulance arrives at the scene;
- $t_5$: time the ambulance departs from the scene;
- $t_6$: time the ambulance arrives at the hospital;
- $t_7$: time the ambulance leaves the hospital;
- $t_8$: time the ambulance comes back to the station.

So it is important to bear in mind that all the parameters needed for the proceeding models can only be estimated from the above time moments.

## 2.3 Service Criterions and General Assumptions

The criterions for the service quality in the call center and the ambulance stations are introduced first, then, two general assumptions used in the following chapters are also presented.

### 2.3.1 Service Criterions

The criterions for the service quality in the call center is that the delay probability in call center should be less then $\alpha$, $\alpha$ can be chosen at random(For example, 1%, 3%, 5%).

The criterions for the service quality in the ambulance station is that the response time should be short. The response time for at least 95% of the A1 customers should be less than 15 minutes. Similarly, the response time for at least 95% of the A2 customers should be less than 30 minutes.

- A1: $P(T < 15min) > 95\%$
- A2: $P(T < 30min) > 95\%$

Here is an explanation about the response time: response time($T$): $T = T1 + T2 + T3 + T4$

- $T1$: service time in call center;
- $T2$: possible waiting time for an ambulance becoming available;
- $T3$: preparation time for an ambulance to leave the station;
- $T4$: driving time to the scene.

[1] In calculation, this time moment is used as an approximation of the time moment that an call finishes for A1, A2 urgencies

### 2.3.2 General Assumptions

In the proceeding chapters, there are 2 general assumptions which will be used all the time,

**Assumption 1** *In the centralists staffing,* Only *A1 customers and A2 customers will be considered, the number of the centralists needed for the B services can be determined by the elementary calculation.*

Because the B services can be interrupted at any time when a call comes in, which means, the A1 and A2 services are preemptive to the B service. So it is reasonable to assume that only A1 and A2 will be considered in the model construction. If the rest of the capacity besides A1 and A2 is enough for the B services, then there is no need to add more centralists. If not, more centralists will be scheduled.

**Assumption 2** *In the ambulance scheduling,* Only *A1 customers and A2 customers will be considered, the number of the ambulances needed for the B services can be determined by the elementary calculation.*

There are three reasons why this assumption is set up. The first is that the requests for A1 and A2 service are stochastic but the B services are scheduled at least one day in advance, so the calculation for the ambulances needed by the B type of customers is trivial. The second reason explained below gives us more support to treat them separately. There are 2 types of ambulances in the station, the fully equipped ambulances are providing services for the A1 customers and the A2 customers, the other group of ambulances with lower standard of the equipments will serve B type of customers. Normally, this two groups will deal with their own requests independently. If a shortage happens, the B customers can borrow the ambulances from another group sometimes. The last reason is that the service criterions only give restrictions to the A1 and A2 services instead of all the services. So it is reasonable to assume that only A1 and A2 need to be taken in to account in the model constructions.

# Chapter 3

# Literature Study

This chapter discusses the literature studied for this project. The first part covers the topic of models about the optimal capacity in call centers. The second part reviews articles involved with optimal scheduling of the ambulances. Section 3.1 gives introduction about the papers in staffing centralists, this introduction is composed by theoretical research, practical research in commercial and ambulance call centers and a conclusion is drawn from them. Similarly, section 3.2 presents the papers in scheduling ambulances, including theoretical and practical research, again, a conclusion is displayed at the end.

## 3.1 Staffing Centralist in Call Centers

Starting from 1950s, there is a long history in the research about centralist staffing in call centers, a large number of papers can be found about this topic. The following papers are discussing topics including heavy traffic limit, time varying control, resource sharing system, abandonments/retrials, statistical forecasting of calls, automated call distribution system and congestions.

### 3.1.1 Theoretical Investigations

The first paper investigating optimal distribution of resources in call centers is Erlang [19] in 1948. In the following two decades after this paper, more mathematician and engineer start to do research in this area. In 1974, Larson [34] use hypercube queuing model the first time to solve this problem. In general, the hypercube queueing model is a computer program, more details about this can be found in Larson's report. After this, many papers based on hypercube queuing models are published. This paper set up a hypercube queuing model to tackle the problems of facility locations and redistricting in urban emergency services, computationally efficient algorithms

are constructed for studying the analytical behavior of a multi-server queuing system with distinguishable servers. This model is aimed for analyzing the problems of vehicle location and response district design in urban emergency services. Apart from this, computation of several point-specific as well as area-specific performance measures for the call center are also allowed. In 1977, Segal and Weinberger [45] discuss both the analytical methods and some implementation considerations. The settings in the model assume that the operators should take full responsibility for all jobs within their own regions. They use a highly interactive software system which is a heuristic algorithm combining shortest 3 path, minimum cost flow, and enumerative techniques to tackle the problem. They also discuss a stochastic model of the work backlog in a region, based on the variability of the demand for service.

With the development of telecommunications, more and more commercial call centers show up in the 1980s. Compared with normal strategy of sales in shops, there is no doubt that selling online saves a lot of resources and earns more money. Although it is better to set up requirement for the service level in this situation, the fundamental goal for the businessman is still to use the least resources to satisfy the requirement. Due to this new situation, a new concept of *heavy traffic* shows up. In 1981, Shlomo and Whitt. [22] for the first time investigate this theoretically in the staffing research in call centers. Two different types of heavy-traffic limit theorems have been proved for $s$-server queues. Although heavy traffic is the last property we want to have in ambulance service centers, it is still interesting to take a look at this. Ward Whitt is a remarkable researcher in staffing problem in call center. After 1980, he wrote a lot of interesting papers about many aspects of call centers. In 1984, Whitt [49] investigates heavy traffic again. In this paper, approximations are discussed for the blocking probability and related congestion measures in service systems with $s$ servers, $r$ extra waiting spaces, blocked customers lost, and independent and identically distributed service times that are independent of a general stationary arrival process (the G/GI/s/r model).

In 1986, Sze [48] talks about a queuing model for telephone operator staffing. The goal of this research is to ensure that the customers receive good levels of service during normal load times and to protect them against very poor service during peak load periods. Several other features based on the Erlang-C Model are considered here: 1. very large number of servers involved at the same time; 2. bimodal service time distributions; 3. non-stationarity of customer arrivals, 4. customer abandonment and reattempts for service, and 5. nonpreemptive priority rules for service. Therefore, a new queuing model was developed to generate staffing tables for each operating system. This model is quite interesting because it contains several features of our

research questions. In the 1980s, the use of computers becomes more and more popular, so simulation methods start to play an important role then. In 1988, Kwan, Davis and Greenwood [32] use simulation to tackle the staffing level with time varying demand. Since the explicit formulae of time varying arrivals is not available, simulation is a good way to solve this problem. Especially in the service operation where the number of centralists has to be determined for short scheduling time periods with non stationary customer demand, the assumptions necessary for using steady-state solutions to elementary queueing models are usually violated. So this paper describes a simulation study of the behavior of such a service type. Finally, the results are compared with the steady-state solutions to queueing models. It is found that if the system utilization is below a derived maximum value (based on a service level criterion), then the steady-state solutions are robust enough to explain the behavior of the system and can be used to schedule worker requirements. This is a useful conclusion for our research because in ambulance car centers *light traffic* is a very important feature. More and more research about commercial call centers with large number of centralists and heavy load are investigated at this time period.

There are more and more papers discussing staffing in call center during 1990, the innovation of these papers lies not in the great changes in the mathematical models, but in slight modifications which can solve different aspects of real problems. In 1991, Levy and Arian[35] develop a dynamic algorithm for distributed queues with abandonments. They separates the distributing traffic to several parallell queues. The most important contributions in this paper are: First, a revenue-driven, Markovian decision model is set up. The model captures the essential elements of the problem; Second, the authors demonstrate a superior performance when implementing dynamic policies. In 1992, Falin and Yang[29] investigate the congestions in the research of call centers. The authors study the congestions and the recovering process of behavior in call center. The interesting system here under inspecting is still a heavy traffic system. The authors use queueing model to deal with subscriber retrials and investigate some of its properties. In addition, the authors derive the explicit formulas for the performance measure of the system and the limit theorems for systems under heavy traffic. In 1993, Bruce and Parsons [5] aim to use least money to find a staffing level to meet the requirement for the grade of service. The authors propose and implement an economic-optimization model for telephone-agent staffing at L.L. Bean. A cost objective function is set up based on queuing theory to determine the optimal staffing level. Apart from this, a regression model is used to count for retrials and potential caller abandonments. In 1994, Gordon and Fowler [21] try to find more accurate force to set up with more help from computer. In the real practice, the service providers constantly strive to cut costs for the resources while maintaining customer satisfaction.

Queueing theory gives them more possibility to make it come true. Consistency algorithms based on three models, Erlang C, M/G/c, and M/G/c with abandonments, are set up. More data tests have also been done after model exploration.

Jennings, Mandelbaum, Massey and Whitt [27] discuss mainly about the server staffing while the demand is time varying in 1996. A muti-server service system with general non-stationary arrival and service-time processes is considered. $S(t)$, the number of servers as a function of time, needs to be determined to meet required efficiency. The criterion of choosing $S(t)$ is that the probability of delay should below a target probability at all times. Finally the author demonstrate that this approximation is effective by making comparisons with the numerical solution of the Markovian $M_t/\mathrm{M}/S_t$ model. In 1997, Duffield and Whitt [16] investigate the control and recovery from rare congestion events in a large multi-server system. Deterministic fluid models are set up to describe the recovery from rare congestion events in a large multi-server system in which customer holding times have a general distribution. To do the approximation, large multi-server system can be presented as an $\mathrm{M/G/}\infty$ model. It has been proved in the paper that, under regularity conditions, the fluid approximations are asymptotically correct when the arrival rate increases. Numerical examples are shown to test the efficiency of these approximation models. In 1998, Mandelbaum and Pats [39] try to use a state-dependent stochastic networks to model the problem. In addition, Mandelbaum, Massey and Reiman [37]shows a strong approximation property for Markovian service network in 1998, they have proved the centre-limit theory in a weaker assumption compared with the previous papers. kolesar and Green [30] deduce the approximated Erlang's Delay formulae for normal distribution, staffing level based on this formulae has been identified. Mandelbaum and Pats [39] use a state-dependent queueing network to model the arrival and service rates, as well as routing probabilities. The state-dependent model captures the real situation better. For example, when the queue length is long, the service rate will increase. In 1999, Whitt [50] shows how to use dynamic model to staff in a call center which is aimed for immediately answering all calls . This paper develops practical models and analytical methods to dynamically determine the number of centralists with the objective of immediately answering all calls. Infinite server system is considered as an approximation model. However, another essential element influencing the efficiency of the model is the forecasting of what will happen in the near future.

In 2000, Aksin and Harker [3] investigate the performance measures in a multi-class multi-resource processor-shared loss system. This paper details a method to calculate performance measures in a specific type of loss system with multiple classes of customers with processor sharing property. Several

performance measures can be tested in this system such as delay probability. The author make some modifications to simplify the computation. In 2001, Jongbloed and Ger [28] model a call center as a queueing model, however, the Poisson arrivals have an unknown and varying arrival rates so prediction of arriving has to be done first, then Erlang formula is used to calculate waiting time and delay probability. The statistical estimation plays an important role in the efficiency. In 2001, Pinker and Shumsky [43] explore the efficiency-quality tradeoff of cross-trained workers. The authors use a stochastic service system to address this research. The outcome of the questions shows that which strategy will be used is depend on the situation met in different call center. From the point of queueing theory: Flexible or cross-trained servers provide more throughput with fewer workers than specialized servers. However, the quality of the service may decrease if only the economy of scales is considered. The general conclusion is as follow: for the small systems, the mixed schedule can be considered more optimal; for the large systems, the model leans to specialized models. Finally a case study is done to demonstrate the conclusion. In 2001, Cezik, Oktay and Hanan [15]. use integer programming model to determine a weekly schedule for the call center. Five working days and two days off are considered in this model. The objective is to determine a weekly staffing schedule to satisfy the demand for service grade, while minimizing the total labor cost of the centralists. An integer programming model is used to determine the weekly tour. The model is quite flexible then it can accommodate different daily models with varying requirements. This model can handle different days-off rules, the computational results are also demonstrated at the end.

In 2004, Atlason, Marina and Shane [7] use a method that combines simulation and cutting plane methods in service systems. At first the authors solve a relaxed linear (integer) program iteratively and transfer the original problem to a simulation, then, use the results of the simulation to generate constraints in the linear system. The conditions under which the solutions of the linear (integer) program converges to an optimal solution of the unrelaxed problem is derived. The concavity of the underlying service level function is critical for the method and a numerical test is presented. In 2005, Mandelbaum, Massey, Reiman and Rider. [38] investigate queue lengths and waiting times for multi-server queues with abandonments and retrials. Markovian multi-server queues with time dependent parameters are considered. Simple fluid and diffusion approximations are used to estimate the mean, variance, and density for both the queue length and possible waiting time. These approximations are solved by ordinary differential equations. The comparison with the simulation results shows the performance of the models are good.

### 3.1.2 Practical Research in Commercial Call Centers

In 1954 Edie [17] investigates the optimal number of vehicular tolls needed at the Port Authority tunnels and bridges, which is in principle a staffing problem. Similar to our research question, the Port Authority tries to handle traffic with a minimum number of toll collectors but at the same time keep good service to the public need. This objects requires that the staffing level gives a very good compromise between economical resource and service. Quantitative probability theory is used here to determine the relations between several critical elements. Again in 1959, Edie [18] shows more modifications to the previous model.

In 1974, Segal [44] gives more insight to the real problem. The number of telephone operators required on duty at switchboards fluctuates widely during the day, so this paper constructs a method for determining the number of operators needed in each hour based on network flow formulations. A real case application is done in 1976 by Buffa, Cosgrove and Luce [13]. An integrated working shift scheduling system is developed and applied in a very large call center in the General Telephone Company of California. The system can use the forecasted calls to determine the number of operators needed on a prefixed short time basis. In 1976, Henderson and Berry[24] propose a heuristic methods for staffing operators. Two types of heuristic methods are proposed for staffing operators to meet the varying demand over the whole day: The first method is to determine the work shift types when the operator shift schedule is known. The second method is to construct an operator shift schedule from a given set of work shift types. These heuristics are evaluated by solution quality and computational efficiency, using actual data.

More and more investigation is put in this area because telecommunication become much more important and developed after 1990. A new system shows up at this time: automated call distribution(ACD) system. In 1991, Agnihothri and Patricia [2] study the problem in staffing a centralized appointment scheduling department in Lourdes Hospital. Lourdes Hospital in Binghampton, New York, uses a telephone system to schedule appointments for outpatients, inpatients, and other ambulatory services. Queueing theory is used to plan optimal staffing levels to satisfy the estimated demand. Based on the results of this queueing model, staffing schedules were modified to meet the different demand in the whole day(including peak time and normal time). It was revealed that the current staffing level is enough to deal with all the requests. The author also shows that low server utilization(occupation rate) is quite important to provide a high level of service. Mason and Panton [41] use more simulation skills to do the staff scheduling in 1998. This authors describe a new simulation and optimal

based system for human resource scheduling in Auckland International Airport, New Zealand. A combination model refers to simulation and integer programming has been developed to determine near-optimal staffing levels. The application of these modeling result shows a significantly lower staffing levels, however, at the same time has a good performance.

### 3.1.3 Practical Research in Ambulance Call Centers

When it comes to 1970s, the rapid development in telecommunications calls for more sophisticated and customized models to deal with the real problems. In 1972, Larson [33] contributes a chapter about Improving the effectiveness of New York Citys 911. This chapter can be considered as the first case study about call center research. The results of a one month operational study of police emergency telephone operations in the central communications room of the New York City Police Department are displayed. This study serves as an example of elementary quantitative modeling to improve an ongoing operation.

As it is stated previously, there is a large amount of papers about call centers after 1980s, so here, we will focus on the research specially considering the ambulance calls service. In 1985, Mabert [36] pays more attention to the statistical area of the ambulance calls In 1987, Kuhn and Hoey [31] study how to improve police 911 operations in Washington, D.C.. In order to solve the increasing number of complaints from officials in the city and residents at the same time in recent years, new methods are set up to enhanced 911 which includes: 1. matching staff deployment with call demand, 2. improving call-handling performance, and 3. improving civilian pay equity. All of these ideas can be used besides only considering the capacity of the centralists.

### 3.1.4 Conclusion

With the prevalence of telephones, the research about staffing centralists in call center has a remarkable development since 1950s. During 1950-1980, the pioneer scientists start their exploration in this field. With the help of developing knowledge of queueing theory at the same time, at the end of 1970s, some of the most important properties of this research topic have already been characterized. Such as the relationship between the capacity and the performance measure can be identified. During 1980 and 2000, some mathematicians try to find more analytical model to tackle the questions raised in call center and the others try to apply the theory to the real problem and check how the models actually work. After 2000, there is few breakthrough improvements on the theoretical research of this topic. The papers are mainly coupling with the problems coming from different real situations.

## 3.2 Scheduling Ambulances in Stations

Compared with so much paper dealing with staffing problem in call centers, there is much less paper discussing exactly scheduling problem in ambulance stations. In most of the papers, the optimal locations of ambulance stations are discussed simultaneously.

### 3.2.1 Theoretical Investigations(Optimal Location)

In 1985, Jarvis [26] point out Erlang loss system is a good approximation for urban service systems. Later in 1989, Batta and Nirup [10] investigate maximal expected coverage location problem(MEXCLP). The model is required to maximize the expected coverage of demand with optimally locating servers while at the same time taking into account the possibility of servers being unavailable sometimes. The highlights of this paper is to relax the poisson arrivals and exponential service time requirements based on hypercube queueing models. In 1993, Ball and Lin [9] the first time use linear programming(LP) method to tackle this problem. After their research, plenty of paper get published based on the research of them. In this paper, the authors present an optimization model to determine the location of stations and the number of vehicles to place at each stations. To solve this, they propose the use of valid inequalities as a preprocessing technique to set up the integer programming(IP) and solve IP using a branch and bound procedure. The computational result shows that this techniques are quite effective. The objection is to minimize the number of facility stations needed with the constraints that every demanding point should be covered by at least one chosen stations within a target response time.

In 1996, Marianov and Revelle [40] investigat the maximal availability location problem for the siting of emergency vehicles based on queueing models. This paper formulates the probabilistic version of the maximal covering location problem, the author use queueing theory to obtain a more realistic model. With the limited number of ambulances, the distribution of ambulances maximize the reliability $\alpha$ that an ambulance become available within a time or distance standard by using a queueing theory model. In 1998, Serra and Marianov [46] use a very theoretical view to deal with this problem. The method is to locate P points on a graph in order to keep the predefined performance measure, such as, the probability that an ambulance is nearby is less than a coefficient. Graph theory is used here.

In 2001, Gendreau and Semet [20] investigate the redeployment problem for the ambulance groups. The real time management of emergency medial services is encountered in this research. This paper use dynamic model to propose a parallel tabu search to determine the optimal location and capacity

for the resources. Simulation based on real data shows the high efficiency of this approach. The highlight of this research is that a lot of practical constraints are considered in the algorithm: such as, limited number of ambulances are located in a station; only a small portion of ambulances are flexible; repeated trips are not allowed, etc. In 2002, Aytug and Saydam [8] propose to use genetic algorithms(GA) to find the optimal locations for the limited resources. The authors focus their attention on a particular formulation model in which a nonlinear objective function is used to optimize the locations over a convex set. The advantage of GA method is that a near optimal solutions can always be found within a reasonable amount of time. However, there is some difficulties in how to get the proper parameters for the GA method, such as the percentile of the mutations, because the result of calculation is quite sensitive respect to it. In 2002, Borras and Pastor [11] start to consider the randomness in server availability when investigating the probabilistic siting models, for example, the nearest facility may not always be available at the time a call come in. Then the problem become not deterministic anymore and the author seek to locate the least number of facilities needed to cover all points of demand within a maximin time S. The objective function is to minimize the number of required facilities with the following constraints: the probability that at least one server is available to each demand mode when a new emergency situation arise should be greater than or equal to some reliability level $\alpha$, then a programming is constructed based on this. In 2003, Brotcorne and Semet [12] use a mixed model including deterministic model and probabilistic model to plan a schedule for ambulances.

In 2005, Snyder and Daskin [47] consider the dual problem of the original research questions, to minimize the expected failure cost is used here instead to maximize the successful achievement. A heuristic way to solve the problem is mentioned in 2006 because with non zero probability, some extreme case can happen. In 2006, Atkinson, Kovalenko, Kuznetsov and Mikhalevich [6] are dealing with this situation. In 2008, Ingolfsson, Budge and Erkut [25] describe an optimization model to solve the optimal ambulance location with random delays and travel times. The fraction of calls reached within a given response time is used as a performance measure. The response time is composed by a random delay plus a random travel time. The advantage of this model is that the randomness of the models has been increased.

### 3.2.2 Practical Research in Ambulance Stations

In 1999, Henderson and Manson [23] investigate the ambulance requirement problem in Auckland, New Zealand. As the city grows, roads become congests and the population demographics changes, some new problems show up: how many ambulances are needed and where should the ambulances

be placed to meet the targets efficiency if the economy of scale is considered. The questions are answered by 2 steps: first, a queueing model is set up to calculate how many ambulances are needed and the optimal places; Meanwhile, a modification has been done to refine the queueing model results. The result is calculated by a software BartSim. In 2004, Andersson, Petersson and Varbrand [4] investigate one real problem in the deployment of ambulance source. Similar to our project, the emergency scale is also divided by: life threatening, not life threatening and non-urgent.

### 3.2.3 Conclusions

Compared with the research of staffing centralists in call center, there is fewer research dealing with the deployment of ambulances. Apart from this, instead only consider the capacity problem, to consider the optimal location of these ambulances at the same time are discussed in nearly all the related papers. This phenomena indicates that the geographical distribution of the ambulances should also be an important aspect of this research.

# Chapter 4

# Modeling Capacity in Call Centers

The main goal in this chapter is to find and develop mathematical models that can be used to to determine the optimal number of centralists in an ambulance call center. A theoretical, insightful framework has to be found that models the working process as it currently occurs in the call center. Implicitly this means that a reliable calculation method must be constructed that does not solely give a model result, but also an explanation how it can be applied in real workforce for scheduling. Due to its corresponding nature of the process, it is plausible that the queueing models can be used here. In section 4.1-section 4.3, three mathematical models are set up. Tests of Hypothesis and parameter estimation are presented in section 4.4. Finally, model results are displayed in section 4.5.

## 4.1 Erlang C Model

The most commonly used model for commercial call centers is the Erlang C model, also called the M/M/c model [1]. M/M/c queue is a model with exponential interarrival times with mean $\frac{1}{\lambda}$, exponential service times with mean $\frac{1}{\mu}$ and $c$ parallel identical servers. Customers are served in order of arrival. The occupation rate per server can be defined as,

$$\rho = \frac{\lambda}{c\mu} < 1 \tag{4.1}$$

Under the assumptions of (3.1), the equilibrium distribution exists. The states of the system can be characterized by the number of customers in the system. Let $p_n$ denote the equilibrium probability that there are $n$ customers in the system. We can derive the equilibrium equations for the the probabilities $p_n$ by using a flow diagram. Instead of equating the flow

into and out of a single state $n$, we get simpler equations by equating the flow between the two neighboring states $n-1$ and $n$ yielding

$$\lambda p_{n-1} = \min(n, c)\mu p_n, n = 1, 2, ... \tag{4.2}$$

Iteration gives

$$p_n = \frac{(c\rho)^n}{n!} p_0, n = 0, ..., c \tag{4.3}$$

and

$$p_{c+n} = \rho^n \frac{(c\rho)^c}{c!} p_0, n = 0, 1, 2, ... \tag{4.4}$$

The probability $p_0$ can be derived from normalization, yielding

$$p_0 = (\sum_{n=0}^{c-1} \frac{(c\rho)^n}{n!} + \frac{(c\rho)^c}{c!} \times \frac{1}{1-\rho})^{-1} \tag{4.5}$$

An important quantity is the probability that a job has to wait. Denote this probability by $\Pi_W$. It is usually referred to as the *delay probability*. By PASTA it follows that

$$\begin{aligned} \Pi_W &= p_c + p_{c+1} + p_{c+2} + ... \\ &= \frac{p_c}{1-\rho} \\ &= \frac{(c\rho)^c}{c!}((1-\rho)\sum_{n=0}^{c-1} \frac{(c\rho)^n}{n!} + \frac{(c\rho)^c}{c!})^{-1} \end{aligned} \tag{4.6}$$

It will be clear that the computation of $\Pi_W$ by the above formulae leads to numerical problems when c is large. However, this problem can be neglected because of the small scale of the ambulance systems in the Netherlands. To apply this model, there are 3 key requirements that should be satisfied, namely,

- Poisson arrivals
- exponential service times
- Limited number of servers

The delay probability in a call center can be approximated by the delay probability from the M/M/c model, which is the summation of the equilibrium probabilities which indicate the number of customers in the system is equal to or greater than that of available servers in the system,

$$P^c_{delay} \approx \Pi_W = \sum_{i=c}^{\infty} P_i \tag{4.7}$$

where $P_i$ in the formulae is the equilibrium probability that there are $i$ customers in the system.

Whether this approximated $P_{delay}$ will be optimistic or pessimistic compared with real delay probability is the main concern of our interest. Due to the reason that the officers of other department such police and firefights will come to help sometimes when a shortage happens, we can predict that model result $P^c_{delay}$ will be larger than the probability a real delay happen.

Algorithm 4.1.1 is used to determine the optimal number of the centralists needed based on the M/M/c model. $\alpha$ is the upper bound of the delay probability.

**Algorithm 4.1.1** Algorithm for The Erlang C Model

c=1, Calculate $\Pi_W$
**while** $\Pi_W > \alpha$ **do**
  c=c+1;
**end while**
Print $\Pi_W$, c

The result of the algorithm indicates the minimum number of centralists needed to satisfy that at least $100 \times (1 - \alpha)\%$ percent of calls requesting an ambulance service can get through immediately. It will be shown later that unfortunately the exponential service time assumption is violated. A reasonable explanation about this is as follow: for the commercial call centers, most of the tasks have very short service time, so the exponential distribution may hold. However, in the ambulance car center circumstance, the service time can not be so short like this because a standard triage and dispatch procedure are needed for each customers, which makes the exponential service time impossible. Although the Erlang C model is not suited, there are still some directions we can try, such as

- insensitive systems which have no requirement to the distribution of service time at all, and the performance measure only depends on the first moment of the service time,
- simulation

Simulation is a general but slow method to solve this problem. Besides, not enough insights can be shown in simulation. So we still hope to find the analytical models to tackle this problem, several insensitive systems will be tried before simulations are done.

## 4.2 M/G/$\infty$ Model

The probabilities of different queue length in the *insensitive* systems only depend on the mean of the service time. M/G/$\infty$ queue is one of the examples. In the M/G/$\infty$ queueing model, the customers arrive according to a Poisson process with the rate $\lambda$. Their service times are independent and identically distributed with some general distribution function. The number of servers is infinite, in another words, there is always a server available for each arriving customer. Hence, the waiting time of each customer is zero and the sojourn time is equal to the service time. Thus by Little' s law we immediately obtain that

$$E(L) = \rho \tag{4.8}$$

where $\rho = \lambda E(B)$ denotes the mean amount of work that arrives per unit time, $L$ is the queue length and $B$ is the length of the service time. The probabilities $p_n$ that there are $n$ customers in the system can be obtained by the similar procedure as for the M/M/c queue. We obtain this formulae by equating the flow from state $n-1$ to $n$ and the flow from $n$ to $n-1$ that

$$p_{n-1}\lambda = p_n n\mu \tag{4.9}$$

Where $\lambda$ is the arriving rate and $\mu$ is the service rate. Thus

$$p_n = \frac{\rho^n}{n!}p_0 \tag{4.10}$$

Since the probabilities $p_n$ have to add up to one, it follows that

$$p_0^{-1} = \sum_{n=0}^{\infty} \frac{\rho^n}{n!} = e^{\rho} \tag{4.11}$$

Finally, we obtain:

$$p_n = \frac{\rho^n}{n!}e^{-\rho} \tag{4.12}$$

Since one of the features in call center that captures our interest, is that ambulance service groups are not the only group work in the call center, besides this, firefighters, polices are also working with them together. When a shortage of centralists occurs, people from other departments can come to help. So the M/G/$\infty$ model can be considered as a reasonable model.

It seems strange to use this model to determine the delay probability in the system because there is actually no delay in the system. So an equivalent "delay probability" is formulated here to estimate the probability that a delay may happen.

Use $P_i$ to denote the probability that there are $i$ customers in use,

$$P_i = \frac{\rho^i}{i!} \times e^{-\rho} \tag{4.13}$$

If we assume the number of the centralists available there is c, the equivalent delay probability can be calculated by the formulae:

$$P^c_{delay} \approx \sum_{i=c}^{\infty} P_i \tag{4.14}$$

Algorithm 4.2.1 is used to determine the staffing level by M/G/$\infty$ model.

**Algorithm 4.2.1** Algorithm of M/G/$\infty$ Model

c=1, Calculate $P^c_{delay}$
**while** $P^c_{delay} < \alpha$ **do**
  c=c+1;
**end while**
Print $P^c_{delay}$, c

## 4.3 M/G/c/c Loss Model

It is easy to see that M/G/c queue is a good model to describe the real situation. In this model, the arrivals follow the Poisson distribution, service time follow a general distribution which is unknown and there is limited number of servers. Although this model can describe the situation in call center well, no explicit formulae is available to calculate the performance measure of M/G/c queuing system. The intuitive explanation of this lies in the huge variation from service time distribution.

Another important feature of this queuing system is that the probability that a centralist is busy with a A1 or A2 case is really small, only this low utility property can guarantee the customers can get serviced in a really short time, which is important for the emergency service system. More standard mathematical theory can be formulated here based on the previous description. The arriving process is again a Poisson process with rate $\lambda$. The service time of the customer is independent and identically distributed with some general distribution function. The number of servers available there is $c$. Each newly arriving customer immediately goes into service if there is a server available, and the customer is lost if all servers are occupied. This system is therefore called a M/G/c *loss system*. Like the techniques we used in the previous model, we start to find the equilibrium probabilities $p_n$ of

$n$ customers in the system. Of special interest is the probability $p_c$, which, according to the PASTA property, describes the fraction of customers that are lost. It is obviously not the situation in the problem that we are investigating in, because the people who call 112 can never be lost. However, as we mentioned before, the low occupation rate makes this approximation plausible.

Before introducing the new model, it is important to notice again in a emergency call center, $\rho$ is small, which is the most important difference compared with other commercial call centers. This is called *light traffic* in queuing theory, light traffic theory is introduce by David and Donald in 1983 [14]. They describe a method for approximating the stationary distribution $\pi(k)$ for the number of customers in an M/G/c queuing system as the traffic goes to zero. Intuitively, when the traffic is light it is very unlikely that an arriving customer would see more than c customers in the system and hence the system might be well approximated by the M/G/c/c (Erlang Loss)model. Furthermore, most arriving customers who are delayed, arrive when the system contains exactly c customers. One might expect that the nonzero delay approaches the minimum of the c random variables. In fact, it has been proved in the paper

$$\frac{\pi(k)}{p(k)} \to 1 \text{ as } \lambda \to 0$$

where, for $k \leq c, p(k)$ is the stationary probability of having $k$ customers in an M/G/c/c system and, for $k > c, p(k)$ is $\rho^k/c!(k-c)!$ times the (k-c)th moment of the minimum of c independent equilibrium-excess service-time variable. Even further, it is reasonable to extend this proof when $\rho \to 0$ instead of $\lambda \to 0$.

**Theorem 1.** $\frac{\pi(k)}{p(k)} \to 1$ *as* $\rho \to 0$

*Proof.* $\rho = \frac{\lambda}{\mu}$
If $\mu < \infty$, then $\rho \to 0$ indicates $\lambda \to 0$
If $\mu \to \infty$, when $\rho \to 0$, $\exists$ n $\in$ N large enough, so that
$\frac{\mu}{2^n}$ = C(constant)
$\lambda$ = o($\mu$)
Mean while, $\frac{\lambda}{2^n} = \frac{o(\mu)}{2^n} = \frac{o(\mu)}{\mu} \times C \to 0$
So $\frac{\pi(k)}{p(k)} \to 1$ as $\rho \to 0$ □

So far, a conclusion can be draw is that: the waiting probability in the M/G/c queue can be well approximated by the blocking probability in M/G/c/c queue if the traffic is light. The *blocking probability* $B(c, \rho)$ is given by

$$B(c,\rho) = p_c = \frac{\rho^c/c!}{\sum_{n=0}^{c} \rho^n/n!}$$

The name of this formula is *Erlang's loss formula.* Here is a summary that why M/G/c/c model is a good choice here.

- Light traffic in M/G/c queuing system, which means the delay probability is quite small;
- M/G/c/c is an insensitive system, which means the delay probability is insensitive to the distribution of the service time, but only depends on its mean.

Apart from M/G/$\infty$ model, M/G/c/c loss system is still been taken into consideration because in real situation, the number of centralists can never be infinite There is also an other very important reason which inspires us to think about this model - light traffic. With the property of light traffic, it is intuitive that the probability of a congestion in the system happens really rare. Again it is also an insensitive system which asks no requirement for the distribution of service time. This two features give us the first impression that M/G/c/c loss model will be a good one to describe the situation.

$$P_{delay}^{c} \approx B(c,\rho) = p_c = \frac{\rho^c/c!}{\sum_{n=0}^{c} \rho^n/n!} \tag{4.15}$$

The algorithm 4.3.1 is used to determine the staffing level based on M/G/c/c model.

**Algorithm 4.3.1** Algorithm of M/G/c/c Model

c=1, Calculate $P_{delay}^{c}$
**while** $P_{delay}^{c} < \alpha$ **do**
  c=c+1;
**end while**
Print $P_{delay}^{c}$

## 4.4 Test of Hypothesis and Parameter Estimation

Tests of hypothesis and parameter estimation are needed before the final result can be displayed. The tests of hypothesis can indicate whether the models constructed are suitable ones and the parameter estimation will provide required parameters to get the results from the calculation. Section 3.4.1 covers the relevant tests of hypothesis, section 3.4.2 presents the useful parameters for the model calculation. A conclusion is drawn in section 3.4.3.

### 4.4.1 Test of Hypothesis

There are mainly two hypothesis tested here: Poisson arrivals and exponential service time. The data experiments of all these tests are catalogued into weekday, Saturday and Sunday by hours.

**Poisson Arrivals**
The first test is that whether the arrivals follow Poisson distribution in each hour. Of course other time segments can also be chosen.

The statistical descriptions for the tests of hypothesis are,

$H_0$: The arrivals follow poisson distribution in each hour;
$H_1$: The arrivals do not follow poisson process in each hour;

*Pearson's chi-square* statistical test is used here, the value of the test statistic is:

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i},$$

where

- $X^2$=the test statistic that asymptotically approaches a $\chi^2$ distribution;
- $O_i$=an observed frequency;
- $E_i$=an expected(theoretical) frequency, asserted by the null hypothesis;
- $n$=the number of possible outcomes of each event.

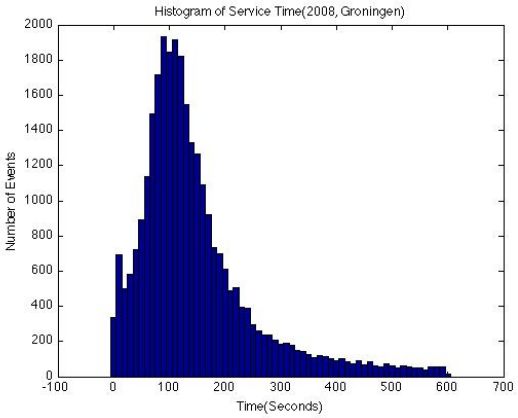
The intuition of Pearson's chi square test is that if data under test follow theoretical distribution, the empirical frequency should be close to the theoretical frequency on each intervals.

Data from 2008 are applied to this test. The test result indicates null hypothesis can not be rejected. So the conclusion is that Poisson arrivals are satisfied. Actually the result is reasonable because the requests for the ambulance service come at random.

**Exponential Service Time**
Similar statistical tests can be constructed to test the distributions of the service time in the call center and the ambulance stations. Unfortunately, the conclusions here are that service times do not follow the exponential distribution. Figure 4.1 display the histogram of the service time which is less than 600 seconds(10 min) for Groningen in 2008. From this histogram, an obvious contradiction is shown. Due to the standard triage procedures in

ambulance call center, most of the calls can only be finished after a certain amount of time. Therefore, it is not difficult to understand why exponential service time can not be true in ambulance call center.

Figure 4.1: Histogram of The Service Time(2008,Groningen)




### 4.4.2 Parameter Estimation

Before calculation, parameter estimation is needed for the required parameters in the models. Maximal likelihood estimation(MLE) method is used here. Here is a short introduction about maximum likelihood estimation. The idea behind maximum likelihood parameter estimation is to determine the parameters that maximize the probability (likelihood) of the sample data. From a statistical point of view, the method of maximum likelihood is considered to be more robust (with some exceptions) and yields estimators with good statistical properties. In other words, MLE methods are versatile and apply to most models and to different types of data. In addition, they provide efficient methods for quantifying uncertainty through confidence bounds. Although the methodology for maximum likelihood estimation is simple, the implementation is mathematically intense. Using today's computer power, however, mathematical complexity is not a big obstacle.

#### Arriving Rates

The maximal likelihood for the Poisson distribution is:

$$\hat{\lambda}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} t_i$$

$t_i$ is the observation.

The parameter estimation can be done base on this formulae, the applications for the arriving rate($\lambda$) during weekday in 2008 are done. Here are the examples in region Groningen.

Table 4.1: The Arriving Rate for Weeday, Region Groningen,2008

| Time Period | $\lambda$ | Time Period | $\lambda$ |
|---|---|---|---|
| 0-1 | 2.0916 | 12-13 | 5.2137 |
| 1-2 | 1.8321 | 13-14 | 4.7977 |
| 2-3 | 1.5344 | 14-15 | 4.6183 |
| 3-4 | 1.3397 | 15-16 | 4.5076 |
| 4-5 | 1.2519 | 16-17 | 4.6412 |
| 5-6 | 1.2634 | 17-18 | 4.2023 |
| 6-7 | 1.4733 | 18-19 | 3.2901 |
| 7-8 | 2.1908 | 19-20 | 3.1565 |
| 8-9 | 4.0573 | 20-21 | 3.3397 |
| 9-10 | 4.3168 | 21-22 | 3.2977 |
| 10-11 | 4.5840 | 22-23 | 2.8550 |
| 11-12 | 5.4695 | 23-24 | 2.4351 |

Similarly, the arriving rate for the ambulance services on weekday in Friesland and Drenthen, 2008 can also be estimated. The requests for the ambulance services have a peak at daytime but there is an obvious decrease during the early morning and midnight. The outcomes are reasonable. First, there is hardly dangerous things happening during early morning or midnight when people are sleeping or relaxing at home; Second, during the daytime, the probability that an accident occurs increase because of the potential damage to human activities.

**Service Time in Call Center**

The distribution of the service time is unknown, so the MLE can not be deduced here. Therefore, we just simply calculate the average service time($\frac{1}{\mu}$) in the call centers.

$$\frac{1}{\mu} = \frac{\sum_{i=1}^{n} t_i}{n} \tag{4.16}$$

the average service time for A1 and A2 services in call center(region Groningen, workday, 2008) is 3.40 minutes, the average service time in Friesland is

2.41 min. Similarly, the average service time in call center(region Drenthe, workday, 2008) is 3.29 minutes.

### 4.4.3 Conclusion

The tests of hypothesis indicates that the assumption of Poisson arrivals holds but that of exponential service time fails to be true. The possible reasons of these results are that the requests come at random and several standard questions should be asked before a decision of dispatching an ambulance can be made. Since the poisson arrival holds, the MLE can be applied to get the arriving rates. However, only the average service time is calculated because the distribution of this service time is unknown.

## 4.5 Applications of the Theoretical Models and Results

The model results of all the regions are displayed in this section. First, the model results for A1 and A2 customers are displayed. Second, the method of calculating the capacity for dealing with B customers are introduced. At the end of the section, an exploration about the combination of the call centers has been done.

### 4.5.1 Model Result of Region Groningen(Example)

#### Model Result(A1/A2)

The conclusion has been drawn previously is that Poisson arrival holds but not exponential service time. Then it is easy to find out that M/M/c model is not suitable anymore. However, because of the insensitivity of M/G/$\infty$ model and M/G/c/c model, these two models still can be considered as good models to determine optimal number of centralists needed in the ambulance call center. When $\alpha = 0.01$, Table 4.2 displays the optimal number of centralists needed in each hour on weekday in Groningen based on M/G/$\infty$ model. Similarly, Table 4.3 shows the optimal number of centralists needed in each hour on weekday in Groningen based on M/G/c/c model.

Table 4.2: The Staffing Level for Weekday, Region Groningen (M/G/$\infty$)

| Time Period | Staffing Level | Time Period | Staffing Level |
|---|---|---|---|
| 0-1 | 2 | 12-13 | 3 |
| 1-2 | 2 | 13-14 | 3 |
| 2-3 | 2 | 14-15 | 3 |
| 3-4 | 2 | 15-16 | 3 |
| 4-5 | 2 | 16-17 | 3 |
| 5-6 | 2 | 17-18 | 3 |
| 6-7 | 2 | 18-19 | 3 |
| 7-8 | 2 | 19-20 | 3 |
| 8-9 | 3 | 20-21 | 3 |
| 9-10 | 3 | 21-22 | 3 |
| 10-11 | 3 | 22-23 | 3 |
| 11-12 | 3 | 23-24 | 2 |

Table 4.3: The Staffing Level for Weekday, Region Groningen (M/G/c/c)

| Time Period | Staffing Level | Time Period | Staffing Level |
|---|---|---|---|
| 0-1 | 2 | 12-13 | 3 |
| 1-2 | 2 | 13-14 | 3 |
| 2-3 | 2 | 14-15 | 3 |
| 3-4 | 2 | 15-16 | 3 |
| 4-5 | 2 | 16-17 | 3 |
| 5-6 | 2 | 17-18 | 3 |
| 6-7 | 2 | 18-19 | 3 |
| 7-8 | 2 | 19-20 | 3 |
| 8-9 | 3 | 20-21 | 3 |
| 9-10 | 3 | 21-22 | 3 |
| 10-11 | 3 | 22-23 | 3 |
| 11-12 | 3 | 23-24 | 2 |

However, the real staffing in call center have 3 shifts in one day, 8 hours per each, then the final staffing table can be determined in table 4.4. The general probability of delay in this table is calculated by the weighted average of $P^c_{delay}$ in each hour,

$$P^c_{delay} = \sum_{i=1}^{24} w_i (P^c_{delay}) \tag{4.17}$$

Where $w_i$ is the weight.

Table 4.4: The Staffing Level of Weekday, Region Groningen

| Model Type | 0-8 | 8-16 | 16-24 | $P^c_{delay}$ |
|---|---|---|---|---|
| M/G/$\infty$ | 2 | 3 | 3 | 0.0028 |
| M/G/C/C | 2 | 3 | 3 | 0.0026 |
| Real Staffing Level | 2 | 3 | 3 | Unknown |

**Elementary Calculation for B Type of Customers**

Most of the B type of customers are coming between 8:00 and 16:00 during the weekday. Now we are going to calculate the total busy time for the A1 and A2 service which we shall call $BT_1$ during 8:00-16:00 and the total busy time for the B services $BT_2$ during this time period.

- If $BT_1 + BT_2 \leq 8 * c$, then there is no need to add more centralists.
- If $BT_1 + BT_2 > 8 * c$, then more centralists are needed.

It is a pity that in the dataset, we only know $t_1$(time a call comes through) and $t_2$(time an ambulance is ordered) instead of the time a call finishes. For most of the B services, the time that the ambulances are ordered happens several hours later after the call finishes. So we are lack of the accurate data to calculate this although the method is easy.

### 4.5.2 Model Results

More results for the staffing level on weekday, Saturday and Sunday for all the regions are displayed here:($\alpha = 0.01$)

**Groningen**

Table 4.5: The Staffing Level in Groningen($M/G/\infty, \alpha = 0.01$)

| Time Period | 0-8 | 8-16 | 16-24 |
|---|---|---|---|
| Weekday | 2 | 3 | 3 |
| Saturday | 2 | 3 | 3 |
| Sunday | 3 | 3 | 3 |

Table 4.6: The Staffing Level in Groningen ($M/G/c/c, \alpha = 0.01$)

| Time Period | 0-8 | 8-16 | 16-24 |
|---|---|---|---|
| Weekday | 2 | 3 | 3 |
| Saturday | 2 | 3 | 3 |
| Sunday | 3 | 3 | 3 |

**Friesland**

Table 4.7: The Staffing Level in Friesland ($M/G/\infty, \alpha = 0.01$)

| Time Period | 0-8 | 8-16 | 16-24 |
|---|---|---|---|
| Weekday | 2 | 3 | 2 |
| Saturday | 2 | 3 | 2 |
| Sunday | 2 | 3 | 2 |

Table 4.8: The Staffing Level in Friesland ($M/G/c/c, \alpha = 0.01$)

| Time Period | 0-8 | 8-16 | 16-24 |
|---|---|---|---|
| Weekday | 2 | 3 | 2 |
| Saturday | 2 | 3 | 2 |
| Sunday | 2 | 2 | 2 |

**Drenthe**

Table 4.9: The Staffing Level in Drenthe ($M/G/\infty, \alpha = 0.01$)

| Time Period | 0-8 | 8-16 | 16-24 |
|---|---|---|---|
| Weekday | 2 | 3 | 3 |
| Saturday | 2 | 3 | 2 |
| Sunday | 2 | 3 | 2 |

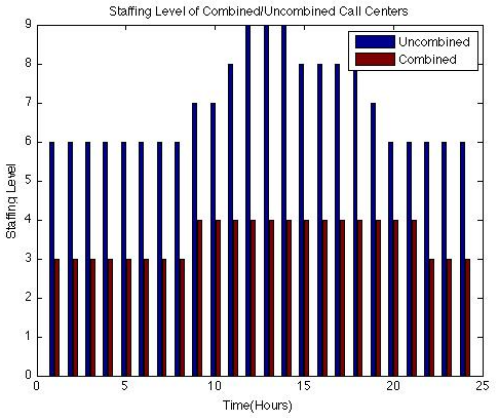
Table 4.10: The Staffing Level in Drenthe ($M/G/c/c, \alpha = 0.01$)

| Time Period | 0-8 | 8-16 | 16-24 |
|---|---|---|---|
| Weekday | 2 | 3 | 3 |
| Saturday | 2 | 3 | 2 |
| Sunday | 2 | 3 | 2 |

## 4.6 Economical Scales of the Call Centers

One practical question raised in call center is that whether it is more efficient to combine the call center. Intuitively, a centralized system means less blocking because when all the centralists are occupied in one call center, the free centralist in another call center can come to help if combination is applied. The following figure shows the staffing level needed for the uncombined and combined call centers. M/G/c/c model is applied on region Groningen, Friesland and Drenthe during 8:00 and 16:00 on weekday when $\alpha = 0.01$. The blue bar shows the number of centralists needed without combination and the red bar indicates the number of centralists needed with combination.

Figure 4.2: Comparison of Combined and Uncombined Call Centers




During the investigation, only A1 and A2 are considered. However, due to the lack of the data for the time that a call finishes, the accurate busy time for the B customers can not be calculated. Perhaps more centralists are needed after combination because of the large amount of B services.

# Chapter 5

# Modeling Capacity of Scheduling Ambulances in Stations

In this chapter, three models aimed for optimal scheduling of ambulances are discussed. The first model is set up to meet the criterion 1 (response time should be short), and the following 2 models are built up to meet criterion 2 (delay probability should be small). The calculations have been done apart from model construction. In section 5.1-5.3, three models are presented to determine the optimal schedules for the ambulances. The method to determine the capacity needed for B type of customers is displayed in section 5.4. Finally, the model results are shown in section 5.5.

## 5.1 M/M/c+CTMC/Multinomial Model

From chapter 2, we can know the original criterion for the scheduling of ambulances is,

- A1: $P(T < 15min) > 95\%$
- A2: $P(T < 30min) > 95\%$

Response Time($T$): $T = T1 + T2 + T3 + T4$

- $T1$: service time in call center;
- $T2$: possible waiting time for ambulance becoming available;
- $T3$: preparation time for ambulance to leave the station;
- $T4$: driving time to the scene.

We notice that for all these time intervals, only $T2$ can be influenced by the number of ambulances, which means, the number of ambulances can not influence the service time in call center, the preparation time for ambulance to leave the station and driving time at all. However, compared with other time period, $T2$ is extremely short because the probability that $T2$ is zero is high, which makes it risky to use the distribution of $T$ to determine the number of ambulances needed because the distribution of $T$ is not sensitive when the number of ambulances changes. Although we are puzzled slightly here, we still need to explore whether this criterion can be satisfied. So our main objective is to find the distribution of response time ($T$) if the schedules of ambulances are known.

The mathematical model set up here can only serve as an approximate model because exponential service times are not realistic. But this model can still give some insight what will happen in the real world. First, M/M/c model is used to calculate the equilibrium distribution of the numbers of customers in the system. When the number of the customers in the system exceeds the number of servers in the system, the newly arriving customers have to wait. Then how long should be wait ($T2$) is our main focus. The distribution of $T2$ depends on how many A1 and A2 customers are in the system and which phase they are in, it is obviously that the ambulance evolved in A1 on the way back to the station will become available soon but the ambulance evolved in A2 which is still on the way to the scene will take more time to come back. Then based on the distribution of $T2$, an convolution can be applied to $T1, T2, T3$ and $T4$ to get the distribution of response time ($T$). We want to solve the problem by 2 steps, first we use M/M/C model to calculate the delay probability and then we can make use of the tool (CTMC) or a direct method(Mutinomial distribution) to get the LST of the distribution of response time $T$.

### 5.1.1 Mathematical Models When All Ambulances Are Occupied

The main mathematical theory used to model the situation when all the ambulances are busy is continuous time Markov chain(CTMC) and multinomial distribution. The method based on continuous time Markov chain is introduced first.

**Continuous Time Markov Chain Model**
Now we start to construct a continuous time Markov Chain. The concepts of phase 1 and phase 2 are used in this Markov Chain, which are defined as follows:

- phase 1 ($t_4 - t_2$): the time period from ordering an ambulance until the ambulance arrives at the scene;

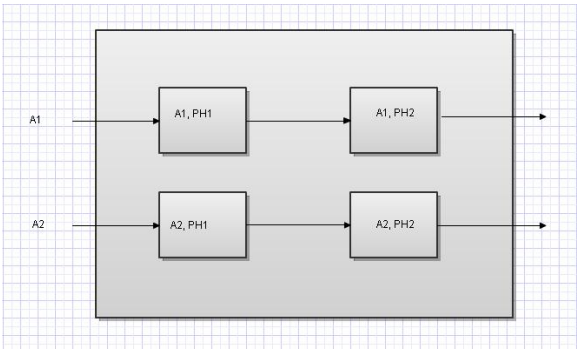
- phase 2 ($t_8 - t_4$): the time period from when the ambulance arrives at the scene until coming back to the station.

It is obvious that this continuous time Markov chain is *only* considered under the case that when all the ambulances are occupied. We now describe the 4-dimential CTMC as $(N_1(t), N_2(t), N_3(t), N_4(t))$,

- $N_1(t)$: number of ambulances used by A1 in phase 1 at time t;
- $N_2(t)$: number of ambulances used by A1 in phase 2 at time t;
- $N_3(t)$: number of ambulances used by A2 in phase 1 at time t;
- $N_4(t)$: number of ambulances used by A2 in phase 2 at time t;

The graphical description of these 4 states is in Figure 5.1:

Figure 5.1: States in Continuous Time Markov Chain




At the end of the state description, the whole state space can also been known. If the number of the available ambulances is $c$, the number of the different states in this Markov Chain is $\binom{c+3}{3}$, which can be considered as a partition of the ambulances groups into 4 parts.

We are interested in the invariant measures when the Markov chain achieves the equilibrium state, so we only focus on the state $\vec{n}$, where $\vec{n} = (n_1, n_2, n_3, n_4)$. Then we need to construct the transition rates for this 4-dimensional Markov chain, the following list is an explanation for the notations, all the variables are assumed to be exponentially distributed. The explanations are also available in the section of notations before the report starts.

- $\lambda_1$: The arriving rate for A1;
- $\lambda_2$: The arriving rate for A2;
- $\gamma_{11}$: The service rate for the phase 1 of A1;

- $\gamma_{12}$: The service rate for the phase 2 of A1;
- $\gamma_{21}$: The service rate for the phase 1 of A2;
- $\gamma_{22}$: The service rate for the phase 2 of A2;

The transition rates are as follow:

- $(n_1, n_2, n_3, n_4) \to (n_1 - 1, n_2 + 1, n_3, n_4) = \gamma_{11}$
- $(n_1, n_2, n_3, n_4) \to (n_1 + 1, n_2 - 1, n_3, n_4) = \frac{\lambda_1}{\lambda_1+\lambda_2}\gamma_{12}$
- $(n_1, n_2, n_3, n_4) \to (n_1, n_2 - 1, n_3 + 1, n_4) = \frac{\lambda_2}{\lambda_1+\lambda_2}\gamma_{12}$
- $(n_1, n_2, n_3, n_4) \to (n_1, n_2, n_3 - 1, n_4 + 1) = \gamma_{21}$
- $(n_1, n_2, n_3, n_4) \to (n_1 + 1, n_2, n_3, n_4 - 1) = \frac{\lambda_1}{\lambda_1+\lambda_2}\gamma_{22}$
- $(n_1, n_2, n_3, n_4) \to (n_1, n_2, n_3 + 1, n_4 - 1) = \frac{\lambda_2}{\lambda_1+\lambda_2}\gamma_{22}$

What we need from this Markov chain is to get the invariant measures of the different states in the state space. Then we can know that when a delay happens, how long you have to wait. We denote the probability that the equilibrium system is in state $\vec{n} = (n_1, n_2, n_3, n_4)$ by $p(n_1, n_2, n_3, n_4)$.

**Multinomial Distribution Model**

Although the CTMC model can give us the probabilities $p(n_1, n_2, n_3, n_4)$ as we want, there is one shortage of this method, that is, it always takes quite a long time to achieve an equilibrium state which means the delay should happen for quite a long time, this is different from the real situation. So another direct method, multinomial distribution method, is also considered here to determine probability $p(n_1, n_2, n_3, n_4)$.

$$\begin{aligned} p(n_1, n_2, n_3, n_4) &= Pr(N_1 = n_1, N_2 = n_2, N_3 = n_3, N_4 = n_4) \\ &= \frac{(n_1 + n_2 + n_3 + n_4)!}{n_1! n_2! n_3! n_4!} \times \\ &\quad (\frac{\lambda_1}{\lambda_1 + \lambda_2} \frac{(\gamma_{11})^{-1}}{(\gamma_{11})^{-1} + (\gamma_{12})^{-1}})^{n_1} \times \\ &\quad (\frac{\lambda_1}{\lambda_1 + \lambda_2} \frac{(\gamma_{12})^{-1}}{(\gamma_{11})^{-1} + (\gamma_{12}) - 1})^{n_2} \times \\ &\quad (\frac{\lambda_2}{\lambda_1 + \lambda_2} \frac{(\gamma_{21}) - 1}{(\gamma_{21})^{-1} + (\gamma_{22})^{-1}})^{n_3} \times \\ &\quad (\frac{\lambda_2}{\lambda_1 + \lambda_2} \frac{(\gamma_{22})^{-1}}{(\gamma_{21})^{-1} + (\gamma_{22})^{-1}})^{n_4} \end{aligned} \tag{5.1}$$

The parameters in this formulae are explained here: when a customer comes, with probability $\frac{\lambda_1}{\lambda_1+\lambda_2}$, it is an A1 customer, and for an A1 customer, with

probability $\frac{(\gamma_{11})^{-1}}{(\gamma_{11})^{-1}+(\gamma_{12})^{-1}}$, it is in phase 1. So the probability that this ambulance is used by an A1 customer in phase 1 is $\frac{\lambda_1}{\lambda_1+\lambda_2}\frac{(\gamma_{11})^{-1}}{(\gamma_{11})^{-1}+(\gamma_{12})^{-1}}$. Similarly, the probabilities that the ambulances are used by A1 customer in phase 2, A2 customer in phase 1 and A2 customers in phase 2 can be deduced respectively.

### 5.1.2 Mathematical Model without Considering Number of Ambulances in Use

Apart from the exploration of a delay happen, the probability that there is no delay in the system is also important because the previous two methods can only help us to derive the conditional distribution of $T2$ when $T2$ is not zero. The M/M/c model is used here to determine the probabilities that there are no delay. Before model descriptions, the notations used here will be introduced, which can also be found in the "Notation" section.

- $\lambda_1$: the arriving rate for A1;
- $\lambda_2$: the arriving rate for A2;
- $\mu_1$: the service rate for A1;
- $\mu_2$: the service rate for A2.

**Arriving Rate**
The arriving rate for the M/M/c system is $\lambda_1 + \lambda_2$ because this nothing but a combination of two Poisson process.

**Service Rate**
The service rate is determined as follow,

$$\frac{1}{\mu} = \frac{\lambda_1}{\lambda_1+\lambda_2} \times \frac{1}{\mu_1} + \frac{\lambda_2}{\lambda_1+\lambda_2} \times \frac{1}{\mu_2} \tag{5.2}$$

The probability that a newly arrival is A1 customer is $\frac{\lambda_1}{\lambda_1+\lambda_2}$, similarly, the probability that a newly arrival is A2 customer is $\frac{\lambda_2}{\lambda_1+\lambda_2}$. So in general, the mean service time is the weighted sum of the mean service times of A1 and A2 customers.

By using M/M/c model, we can get the distribution of the number of customers in the system:

$$p_i = \frac{(c\rho)^i}{i!}p_0, i = 0, ..., c \tag{5.3}$$

and

$$p_{c+i} = \rho^i\frac{(c\rho)^c}{c!}p_0, i = 0, 1, 2, ... \tag{5.4}$$

where occupation rate $\rho$ is,

$$\rho = \frac{\lambda}{c\mu} \tag{5.5}$$

$c$ is number of servers(ambulances), $p_0$ is the probability that the system is idle, which means all the ambulances are waiting in the station.

So the probabilities that $T2$ are zero and non-zero can be got as follow:

- $q_1$: $\sum_{i=0}^{c-1} p_i$, probability that $T2$ is zero;
- $q_2$: $1 - q_1$, probability that $T2$ is non-zero.

### 5.1.3 Combined Mathematical Model

Denote the LST of $T2$ by $S_{T2}(\tilde{s})$. If the number of available ambulances is $c$, then the number of combination of the states when all the ambulances are occupied is $\binom{c+3}{3}$. Let $P_{\vec{n}}$ denote the probability that the system is in state $\vec{n}$, and $\sum_{\vec{n}} P_{\vec{n}} = 1$. When the system is in state $\vec{n} = (n_1, n_2, n_3, n_4)$ , we use $r_{\vec{n}}$ to denote the rate that one ambulance becomes available: $r_{\vec{n}} = n_1 \times \mu_1 + n_2 \times \gamma_{12} + n_3 \times \mu_2 + n_4 \times \gamma_{22}$. Then the LST of $T2$ is:

$$S_{T2}(\tilde{s}) = q_1 + \sum_{i=c}^{\infty} p_i \times (\sum_{\vec{n}} P_{\vec{n}} \frac{r_{\vec{n}}}{r_{\vec{n}} + s})^{i-c+1} \tag{5.6}$$

where $c$ is the number of ambulances, $p_i$ is the probability that there are $i$ customers in the system, $P_{\vec{n}}$ is the probability that the system is in state $\vec{n}$. So far, the distribution of $T2$ can be derived theoretically. Then the distribution of response time $T$ can be got by the convolution of $T1, T2, T3$ and $T4$ when the number of the ambulances $c$ is known.

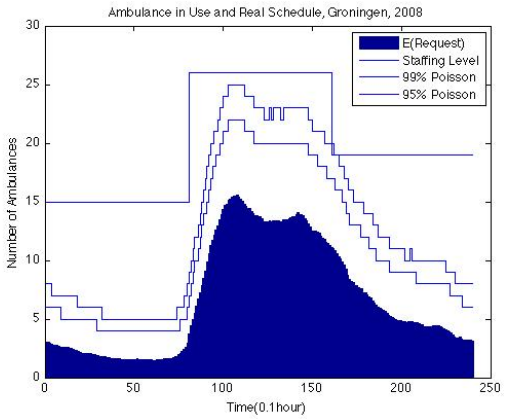
### 5.1.4 Problem Encountered

The first data exploration starts to show some problems: again, we use the data from 2008, weekday, region Goningen. The real data shows the percentage of A1 can be achieved within 15 min is 92.64% and the percentile of A2 can be achieved within 30 min is 88.93% without knowing any information about whether there is a delay and when a delay happens. For region Drenthe, 93.31% of A1 and 93.14% of A2 can be achieved in time. The reasons why these result are different from the introduction in Chapter 1 are: 1, these investigations are done in specific regions, maybe the other regions have a better performances then the overall performance is still good. 2, It can be found in the appendix of [42] that the data are filtered by some criterion before the calculation but my calculations are based on the raw datasets. However, the discussion with manager in ambulance station indicates a delay happens really rare. This strange phenomena drive us to

think about which is the essential reason for the prolonged response time, shortage of ambulance or long driving time? Then an rough calculation is done to investigate this problem.

First we are checking about the data on weekday, region Groningen, 2008. The Figure 5.2 shows the average number of ambulances in use, staffing level and 2 stairs functions depicted based on Poisson cumulative density function with parameter 0.95 and 0.99.

Figure 5.2: Ambulance in Use and Real Schedule, 2008, region Groningen




A more precise explanation of the elements in this figure is as follow:

- Average number of ambulances in use during weekday per day;
- Staffing Levels of Ambulances;
- Efficiencies Achieved by Poisson Assumptions.

If we consider the whole year as time series, year 2008 can be interpreted by the interval [0, 8784] (hour) because there are $366 \times 24 = 8784$ hours for year 2008. Denote the total number of ambulance trips by M (M=43478), then we can investigate how many ambulances are in use at different time epochs every 0.1 hour(0.1, 0.2, 0.3, ..., 8784). The total number of time epochs is $N$, $N = 8784 \times 10 = 87840$ and the ordring time and release time for all M trips are known from the dataset. The pseudocode to detect how many ambulances are in use at each time epoch is in algorithm 5.1.1:

**Algorithm 5.1.1** The Algorithm to Detect Number of Ambulances in Use

```
M; N; Z=zeros(1,N);
for i=1 to N do
  for j=1 to M do
    if t2(j) ≤ 0.1*N ≤ t8(j) then
      Z(i)=Z(i)+1;
    end if
  end for
end for
```

Then Z(i) is the number of ambulances in use at time epoch 0.1*i. Then we need to use more algorithms to determine whether a delay occurs for each time epoch. Assume there are $c$ ambulances in the station.($c$ is varying during different time periods). Algorithm 5.1.2 gives the first step:

**Algorithm 5.1.2** The Algorithm to Detect Delay(1)

```
for i=1 to N do
  if Z(i) > c then
    D(i)=2; {If the number of the ambulances in use is greater than
    the number of the ambulances they have, we declaim that there is a
    delay}
  end if
  if Z(i) == c then
    D(i)=1; {If the number of the ambulances in use is the same as the
    number of the ambulances they have, we need to check the prepara-
    tion time before getting the final conclusion}
  end if
  if Z(i) < c then
    D(i)=0; {If the number of the ambulances in use is less than the
    number of the ambulances they have, we declaim that there is no
    delay}
  end if
end for
```

Based on the previous 2 algorithms, it is possible to check whether a delay occurs in each ambulance trip. There is one more thing needed to be mentioned here: although the ambulance management is different from region to region, the average time needed for the preparation of an ambulance is less than 2 minutes in general, so we assume that when $t_3 - t_2 < 5$min, we claim that there is no delay.

**Algorithm 5.1.3** The Algorithm to Detect Delay(2)

```
R=zeros(1, M);
for i=1 to N do
  for j= 1 to M do
    if t2(j) < 0.1*i <t3(j) & D(i)==2 then
      R(j)=1;
    end if
    if t2(j) < 0.1*i <t3(j) & D(i)==1 & t3(j)-t2(j) > 5 min then
      R(j)=1;
    end if
  end for
end for
```

Based on this algorithm, we can find when a delay occurs and in which trip there is a delay. For all the time epochs, $Z(i)$ is the number of the ambulances used at time $0.1 * i$, $D(i)$ is an indicator to show whether the number of the ambulances in use exceeds the number of the available ambulances at time $0.1 * i$. Based on these two vectors, the new indicator $R(j)$ can be determined. When $R(j)$ is 1, there is a delay in the $j$th ambulance trip, otherwise there is no delay.

**Probability of Delay**

The specific day and time when a delay occurs is detected. The probability of delay can be calculated by calculating the fraction of the number of trips met a delay and the total number of the trips. In Table 5.1, the numbers of the trips met a delay are displayed. In stead of pointing out in which trip a delay occurs, this table gives more insight such as when a delay occurs. From the table, we can see that number of the trips met a delay is only 17, the total number of the trips is 43478, then the probability of delay is $3.91 \times 10^{-4}$.

Table 5.1: Number of Delay Trips

| Time | 16:06 | 16:12 | 16:18 | 16:24 | 16:30 | 16:36 | 16:42 | 16:48 | 16:54 |
|---|---|---|---|---|---|---|---|---|---|
| Number | 3 | 2 | 3 | 2 | 2 | 2 | 1 | 1 | 1 |

So far, the problem has become more clear: the shortage of ambulances happens really rare but the criterion from the government still can not be met. We suggest that the failure of satisfying criterion 1 is not due to the shortage of ambulances but the long driving time eventually. In order to make this conclusion more concrete, several investigations have been done:

- Leave these 17 trips out, there is still a failure to meet the criterion;
- If the preparation time($t_3 - t_2$) for a delay trip is larger than 5 min, substitute it by 5 min, then calculate how often A1 and A2 can meet the criterion 1, the result still shows a failure;
- regardless whether there is a delay or not, if $t_3 - t_2 > 5min$, then substitute it by 5 min, the result still shows a failure to meet criterion 1.

So far, there are following reasons to convince us that criterion 1 is not a good criterion to determine the optimal number of ambulances:

- $T2$ is very small compared to the response time $T$, then it is not clear whether the change for the number of ambulances has great effect on the distribution of $T$ (response time);
- Data explorations have been done to show that a delay happens really rare, but criterion 1 still can not be achieved;
- Discussion with managers in ambulance stations shows most of the time the ambulances are waiting in the station which means the shortage happens really rare, but the criterion still can not be satisfied.

However, the optimal capacities of ambulances still need to be dermined, then another criterion is needed to determine the optimal number of ambulances. So we set up a new criterion here which we shall call criterion 2: $P(T2 > 0)$ should be small, more specifically, we need to get the minimum number of ambulances which satisfies $P(T2 > 0) < \epsilon$ ($\epsilon$ is a predefined parameter). We denote $P(T2 > 0)$ by $P^a_{delay}$ which the probability that the patients can not get the ambulances immediately. So the following models are constructed to calculate the minimum number of ambulances needed such that $P^a_{delay}$ is small.

## 5.2 4-d CTMC Model

Now we move to Criterion 2: $P^a_{delay} < \epsilon$. When we start to use Markov chain model, we have already taken it for granted that all the service times follow the exponential distribution, but this requirement is almost impossible to satisfy, so this model again can only be considered as an approximation model. Since we know from real observations that a delay occurs really rare, it is reasonable to assume at most only 1 customer per type (A1 or A2) will wait for an ambulance. This seems also true from the practical experience, that is, if there was more than 1 or 2 customers waiting for an ambulance, the manager in the ambulance station will start to arrange a request by asking help from nearby regions. Then we have constructed the following 4

dimensional continuous time Markov chain to tackle this problem. Denote the number of available ambulances by $c$. The states description are as follow:

- $N_1(t)$: number of A1 in service at time t; $(n_1 = 0, 1, 2, ..., c)$
- $N_2(t)$: number of A1 in queue at time t; $(n_2 = 0, 1)$
- $N_3(t)$: number of A2 in service at time t; $(n_1 = 0, 1, 2, ..., c)$
- $N_4(t)$: number of A2 in queue at time t; $(n_2 = 0, 1)$

Because we are interested in the equilibrium state of this Markov chain instead of time independent behavior, vector $\vec{n} = (n_1, n_2, n_3, n_4)$ is used to denote the state that the number of A1 customers in service is $n_1$, the number of A1 customers in queue is $n_2$, the number of A2 customers in service is $n_3$ and the number of A2 customers in queue is $n_4$. Denote the state space of this Markov chain by $S$, We will explain the state space by introducing four subsets of $S$, namely $S_1$, $S_2$,$S_3$ and $S_4$.

- $S_1$: $\vec{n}$ satisfy $||\vec{n}||_1 < c$ [1] and $n_2 = 0, n_4 = 0$. The elements can be denoted by $(n_1, 0, n_3, 0)$; The explanation for these states are not all the ambulances are in use, the number of the states in space $S_1$ is $\binom{0+1}{1} + \binom{1+1}{1} + \binom{2+1}{1} + ... + \binom{c}{1}$.
- $S_2$: $\vec{n}$ satisfy $||\vec{n}||_1 = c$ and $n_2 = 0, n_4 = 0$. The elements can be denoted by $(n_1, 0, n_3, 0)$; The explanation for these states are that all the ambulances are in use but there is no waiting customers in the queue. The number of the states in space $S_2$ is $\binom{c+1}{1}$.
- $S_3$: $\vec{n}$ satisfy $||\vec{n}||_1$=c+1, the elements can be denoted by $(n_1, 1, n_3, 0)$ and $(n_1, 0, n_3, 1)$; The number of the states in $S_3$ is $2 \times (c + 1)$.
- $S_4$: the states with delay, there is one A1 customer and one A2 customer in the queue, which means both $n_2$ and $n_4$ are 1, so the number of the state in space $S_4$ is $c + 1$.

So the total number of states of $S$ is: $\frac{(c+2)(c+1)}{2} + 3 \times (c+1) = (4 + \frac{c}{2})(c+1)$.

In the following list, the notations used for this Markov chain are explained,

- $\lambda_1$: arriving rate for A1 customers;
- $\mu_1$: service rate for A1 customers;
- $\lambda_2$: arriving rate for A2 customers;

[1] $||\vec{n}||_1$ is the $L^1$ norm of the vector $\vec{n}$

- $\mu_2$: service rate for A2 customers.

If the total number of ambulances is $c$. The transition rates of this continuous time Markov chain are displayed by 4 group respectively,

For the states of $S_1$, the number of the customers in service can either be increased or decreased.
$S_1$

- $(n_1, 0, n_3, 0) \rightarrow (n_1 + 1, 0, n_3, 0) : \lambda_1$
- $(n_1, 0, n_3, 0) \rightarrow (n_1 - 1, 0, n_3, 0) : n_1 * \mu_1$
- $(n_1, 0, n_3, 0) \rightarrow (n_1, 0, n_3 + 1, 0) : \lambda_2$
- $(n_1, 0, n_3, 0) \rightarrow (n_1, 0, n_3 - 1, 0) : n_3 * \mu_2$

For the states of $S_2$, the newly arriving customers can only wait in the queue.
$S_2$

- $(n_1, 0, n_3, 0) \rightarrow (n_1, 1, n_3, 0) : \lambda_1$
- $(n_1, 0, n_3, 0) \rightarrow (n_1 - 1, 0, n_3, 0) : n_1 * \mu_1$
- $(n_1, 0, n_3, 0) \rightarrow (n_1, 0, n_3, 1) : \lambda_2$
- $(n_1, 0, n_3, 0) \rightarrow (n_1, 0, n_3 - 1, 0) : n_3 * \mu_2$

For the states of $S_3$, the customers in queue can come into the service and a newly arriving customer of another type is allowed to join in the queue.
$S_3$

- $(n_1, 1, n_3, 0) \rightarrow (n_1, 0, n_3, 0) : n_1 * \mu_1$
- $(n_1, 1, n_3, 0) \rightarrow (n_1 + 1, 0, n_3 - 1, 0) : n_3 * \mu_2$
- $(n_1, 1, n_3, 0) \rightarrow (n_1, 1, n_3, 1) : \lambda_2$
- $(n_1, 0, n_3, 1) \rightarrow (n_1, 0, n_3, 0) : n_3 * \mu_2$
- $(n_1, 0, n_3, 1) \rightarrow (n_1 - 1, 0, n_3 + 1, 0) : n_1 * \mu_1$
- $(n_1, 0, n_3, 1) \rightarrow (n_1, 1, n_3, 1) : \lambda_1$

For the states of $S_4$, the customers in queue can get serviced when an ambulance becomes available.
$S_4$

- $(n_1, 1, n_3, 1) \rightarrow (n_1, 0, n_3, 1) : n_1 * \mu_1$
- $(n_1, 1, n_3, 1) \rightarrow (n_1 + 1, 0, n_3 - 1, 1) : n_3 * \mu_2$

- $(n_1, 1, n_3, 1) \rightarrow (n_1, 1, n_3, 0) : n_3 * \mu_2$
- $(n_1, 1, n_3, 1) \rightarrow (n_1 - 1, 1, n_3 + 1, 0) : n_1 * \mu_1$

Based on this 4 dimensional continuous time Markov chain, $P^a_{delay}$ can be determined by the following formula, which means at least one of the queues is not empty, then we claim that a delay occurs:

$$P^a_{delay} = 1 - Pr(N_2 = 0, N_4 = 0) \tag{5.7}$$

We expect that the model result will be more optimistic compared with the real situation because the ambulances circumstance is quite different from the call center case. There is much more possibility that unexpected case happen which may lead to a failure to meet the 15 min or 30 min criterion, such as bad whether, bad road condition, unexpected failure of an ambulance or an ambulance is under maintenance during the daytime.

This model can be used to evaluate a given schedule or determine a schedule. Algorithm 5.2.1 is set up to determine a new schedule: Although this model

**Algorithm 5.2.1** The Algorithm to Determine New Schedule(4-d CTMC)

```
N=C;(current number of ambulances)
Calculate P^a_delay
if P^a_delay > ε then
  while P^a_delay > ε do
    N=N+1;
  end while
end if
if P^a_delay < ε then
  while P^a_delay < ε do
    N=N-1;
  end while
  N=N+1;
end if
```

seems to be a plausible model to determine the number of ambulances in the system, the exponential service time distribution is still away from the real world, so, seeking for other models which can escape from exponential service requirement is still our main focus. Again, insensitive systems seems to be a good direction to explore.

## 5.3 M/G/c/c Loss Model

A standard M/G/c queue is set up here again because the arrivals follow a Poisson distribution and the exponential service time is not available.

Although the traffic here is not as light as that in the case of call centers, the traffic is still moderate. The most important thing we notice is that: $\rho^c$ is very small, that means the probability that all the servers are occupied is really small, where $\rho$ is the probability that an ambulance is busy and $c$ is the number of the ambulances available in the station. Then the probability that the a delay happens is almost 0. Due to this intuition, an M/G/c/c model may be again a good approximation of this system. So the following are the reasons why we move from CTMC model to M/G/c/c model,

- Several data explorations show the exponential service time assumption is not satisfied, but the Poisson arrivals still hold, then an insensitive system is a good direction;
- M/G/c/c queue have the formulae to evaluate the delay probability and it is also a good approximation for M/G/c queue when the traffic is light.

Here again we only consider about A1 and A2 case, the explanation of the notation is explained here again:

- $\lambda_1$: the arriving rate for A1;
- $\lambda_2$: the arriving rate for A2;
- $\mu_1$: the service rate for A1;
- $\mu_2$: the service rate for A2.

Then for the M/G/c/c model, the parameters can be determined in the following way:

- arriving rate: $\lambda = \lambda_1 + \lambda_2$;
- service rate: $\frac{1}{\mu} = \frac{\lambda_1}{\lambda_1+\lambda_2}\frac{1}{\mu_1} + \frac{\lambda_2}{\lambda_1+\lambda_2}\frac{1}{\mu_2}$.

The probability that a delay happens in an ambulance station can now be found by the following formula:

$$P^a_{delay} \approx B(c,\rho) = p_c = \frac{\rho^c/c!}{\sum_{n=0}^{c}\rho^n/n!} \tag{5.8}$$

where $B(c,\rho)$ denotes the *Blocking Probability* of the system, in another words, $p_c$, the probability that all the ambulances in the system are busy. Again, occupation rate $\rho$ is the probability that an ambulance is busy and c is the number of available ambulances.

The algorithm to determine the staffing level is as follow:

**Algorithm 5.3.1** Algorithm to determine staffing level by M/G/c/c Model

c=1;
Calculate $P^c_{delay}$
**while** $P^a_{delay} > \epsilon$ **do**
  c=c+1;
**end while**
Print $P^a_{delay}$

### 5.3.1 M/G/c/c Model with Preference

Because in the previous models, the ambulances for A1/A2 and B are separated absolutely which is not exactly the same as the real case, so in this model, a preference is considered instead of totally separating these 2 groups. Jarvis [26], develop a fast convergent algorithm to calculate performance measure of this kind of ambulance systems.

Advantage of the Model:

- The idea of separating ambulances for A1/A2 and B is considered here, but absolute separation is avoided.
- The different service times for A1, A2 and B can be considered simultaneously in one model;
- A loss System is a good approximation for the ambulance service system because the waiting occurs really rare, which means the probability that the number of the customers in the system is more than the number of the servers is extremely small.

Disadvantage of the Model:

- In the model, there is an extra preference within the group of the ambulances for A1/A2 and B. But in the real case, the ambulances used in each group are chosen at random(there is no preferences within the group). More modifications should be considered here: such as, every time assign the preferred ambulances within each group randomly, then repeat this procedure several times to get the average occupation rate for each of the ambulancess;
- An ambulance car is assigned to the B customers immediately after a request coming up. However, in real case, B customer can wait or be scheduled to do the tasks later.
- In [26], no analytical bounds on the accuracy or convergence properties of the approximation procedure have been developed, only simulation results are used as a comparison;

## 5.4 The Model to Calculate Capacity for B customers

It can be noticed from real dataset that nearly all the B services are done between 8:00 and 16:00. A discussion with the manager in the ambulance station indicates the ambulances used for B type customers should be used as much as possible. In another words, for efficiency reasons, the ambulances must be scheduled such that they have a high occupation rate. Normally the service for B customer should be scheduled at least one day in advance. It is also mentioned by the manager in ambulance stations that in order to use least resource to solve the problem, it is preferred to use the number of ambulances which has a little bit shortage comparing with the real request, because sometimes they can borrow ambulances from group A1 and A2. So it is reasonable to assume there is no stochastic process for scheduling B customers. Then there is an easy formulae to determine the number of ambulances needed for B type customers. Denote the average number of requests per hour from B customers by $\lambda_B$ and let $\lambda_B$ round up to the next integer. Let the average service time for 1 trip be $\frac{1}{\mu_B}$ (hours). Then the number of ambulances needed for B type customers is:

$$C_B = \frac{\lambda_B}{\mu_B}$$

## 5.5 Applications of the Theoretical Models and Results

The model results of the 4-d Markov Chain and the M/G/c/c model are displayed here. Again, the examples of region Groningen and region Drenthe in 2008 are displayed here. The reason why the results of Friesland are not implemented is that the ambulances system in Friesland is not centralized. In addition, we do not know the data for each of the sub regions in Friesland.

### 5.5.1 Model Result of Region Groningen

- M/M/C+CTMC/Multinomial Model: As we stated earlier, this model is considered to be not a proper model to solve the problem because the failure to meet the government criterion is not because of the shortage of ambulances. Therefore we do not implement this model. Therefore, we did not apply this.

- 4 dimensional CTMC Model: There are 3 shifts in one day and a different numbers of ambulances are needed in each shift. As an example we chose to apply this model to the data between 8:00 and 16:00, region Groningen, 2008. The delay probability based on 4 dimensional

CTMC is:

$$P^a_{delay} = 1.6768 \times 10^{-5}$$

and the "real delay" can be determined by the algorithm stated in section 5.1.4. The number of delay trips in 17, and the total number of trips between 8:00 and 16:00 is 24202, so

$$P_{real} = 7.0242 \times 10^{-4}$$

Just as expected in the earlier sections, this probability is higher than the theoretical result.

- M/G/c/c Model Because $P_{real}$ is around $10^{-4}$, it is reasonable to choose $\epsilon$ to be $10^{-4}$. With $P^a_{delay} < 1 \times 10^{-4}$, the schedule of ambulances based on M/G/c/c model are displayed in Table 5.2.

Table 5.2: The Schedule of A1/A2 Ambulances

| Time Interval | 0-8 | 8-16 | 16-24 |
|---|---|---|---|
| # Ambulances Needed For $P^a_{delay} < 10^{-4}$ | 12 | 20 | 18 |

- With Mathematical model described in section 5.4, we can calculate the schedules of the ambulances for B type customers:

Table 5.3: The Schedule of B Ambulances

| Time Interval | 0-8 | 8-16 | 16-24 |
|---|---|---|---|
| # Ambulances Needed For B Customers | 1 | 8 | 1 |

Then now we can calculate the new schedule for ambulances needed by summing up the result of these two tables in Table 5.4:

Table 5.4: The Schedule of Ambulances(Total of A1/A2 and B)

| Time Interval | 0-8 | 8-16 | 16-24 |
|---|---|---|---|
| # Ambulances Needed For $P^a_{delay} < 10^{-4}$ | 13 | 28 | 19 |

### 5.5.2 Model Results of Region Drenthe

The model results of region Drenthe are listed here,

- M/M/C+CTMC/Multinomial Model: With the same reason stated previously, we did not implement this model.

- 4 dimensional CTMC Model: Similar to the procedure in calculating model results in region Groningen, the $P^a_{delay}$ during 8:00 and 16:00 during weekday can be got as follow,

$$P^a_{delay} = 1.2832 \times 10^{-7}$$

and the "real delay" determined by the algorithm stated earlier is:

$$P_{real} = 4.9568 \times 10^{-5}$$

Just as expected in the earlier sections, this probability is higher than the theoretical result.

- M/G/c/c Model: Because the $P_{real}$ is around $10^{-5}$, it is reasonable to choose $\epsilon$ to be $10^{-5}$. With $P^a_{delay} < 1 \times 10^{-5}$, the schedule of ambulances based on M/G/c/c model are displayed in Table 5.5.

Table 5.5: The Schedule of A1/A2 Ambulances

| Time Interval | 0-8 | 8-16 | 16-24 |
|---|---|---|---|
| # Ambulances Needed For $P^a_{delay} < 10^{-5}$ | 10 | 16 | 13 |

- Similarly, the model in section 5.4 is used to determine the number of ambulances needed for the B type customers.

Table 5.6: The Schedule of B Ambulances

| Time Interval | 0-8 | 8-16 | 16-24 |
|---|---|---|---|
| # Ambulances Needed For B Customers | 1 | 3 | 1 |

Then now we can calculate the new schedule for ambulances needed by summing up the result of these two tables in Table 5.7:

Table 5.7: The Schedule of Ambulances(Total of A1/A2 and B)

| Time Interval | 0-8 | 8-16 | 16-24 |
|---|---|---|---|
| # Ambulances Needed For $P^a_{delay} < 10^{-5}$ | 11 | 19 | 14 |

### 5.5.3 More Results

More results based on M/G/c/c model for region Groningen and Drenthe on weekday, Saturday and Sunday are displayed here,

**Groningen**

Table 5.8: The Schedule of Ambulances in Groningen ($M/G/c/c, \epsilon = 10^{-4}$)

| Time Period | 0-8 | 8-16 | 16-24 |
|---|---|---|---|
| Weekday | 13 | 28 | 19 |
| Saturday | 14 | 19 | 18 |
| Sunday | 15 | 19 | 17 |

**Drenthe**

Table 5.9: The Schedule of Ambulances in Drenthe ($M/G/c/c, \epsilon = 10^{-5}$)

| Time Period | 0-8 | 8-16 | 16-24 |
|---|---|---|---|
| Weekday | 11 | 19 | 14 |
| Saturday | 12 | 17 | 15 |
| Sunday | 13 | 16 | 14 |

**Conclusion**

The models set up in the previous section have been used to determine the optimal schedules of ambulances for region Groningen and Drenthe. The 4-d CTMC model can give an approximation of the $P^a_{delay}$ in the ambulance station. In addition, the optimal schedules of ambulances can be generated by the M/G/c/c model. The $P^a_{delay}$ of these new schedules can be under control by some given parameters, such as, $10^{-4}, 10^{-5}$.... From a theoretical point of view, these schedules are quite useful. In order to make sure the performance of these schedules is still good in practice, the relevant simulations will be introduced in the next chapter. There is also another conclusion can be drawn from the model results: the current schedules of ambulances are near optimal, only slightly adjustment is needed if more efficiency is required.

# Chapter 6

# Simulation

Due to the lack of data, the probabilities of delay in call center and ambulance stations are still unknown. So apart from all the deterministic models defined in chapter 4 and 5, we also performed a stochastic simulation to explore $P^a_{delay}$ and $P^c_{delay}$. This chapter contains the goals of the simulations, the parameter estimation before the simulations and the pseudocode of the simulations followed by the simulation results.

## 6.1 Goals

The general goals of the simulations are to get the performance measures for call centers and ambulance stations when the staffing levels and schedules of ambulances are given. The details are:

- $P^c_{delay}$ when staffing levels are known;
- $P^a_{delay}$ when ambulance schedules are known.

## 6.2 Descriptions of Simulations

The general descriptions of simulations are introduced here:

### 6.2.1 Desctiption of Simulation in Call Center

The procedure of simulation in call center is to generate a matrix which has 3 columns, the details of them are:

- Column 1: the arriving time for a call;
- Column 2: the time that the call comes through;
- Column 3: the time that the call finishes.

### 6.2.2 Description of Simulation in Ambulance Station

The method of simulation in ambulance station is to continue to generate the matrix which has 5 columns, the details of them are:

- Column 1: the time that an ambulance is ordered;
- Column 2: the time that an ambulance becomes available;
- Column 3: the time that the ambulance leaves the station;
- Column 4: the time that the ambulance arrives at the scene;
- Column 5: the time that the ambulance comes back to the station.

## 6.3 Parameters

The parameters used in the simulations are explained here:

1. The staffing level in call center: there are 3 shifts a day and the schedules are different for weekday and weekend. See Table 6.1. For the staffing levels, we can choose them at random. In the following simulation, the model results from chapter 4 will be used here.

Table 6.1: The Staffing Level

| Time | 0-8 | 8-16 | 16-24 |
|---|---|---|---|
| Weekday | $C_{11}$ | $C_{12}$ | $C_{13}$ |
| Saturday | $C_{21}$ | $C_{22}$ | $C_{23}$ |
| Sunday | $C_{31}$ | $C_{32}$ | $C_{33}$ |

2. The schedule of ambulances: see Table 6.2. For the schedules, we can choose them at random too. In the following simulation, the model results from chapter 5 will be used here.

Table 6.2: The Ambulance Schedule

| Time | 0-8 | 8-16 | 16-24 |
|---|---|---|---|
| Weekday | $A_{11}$ | $A_{12}$ | $A_{13}$ |
| Saturday | $A_{21}$ | $A_{22}$ | $A_{23}$ |
| Sunday | $A_{31}$ | $A_{32}$ | $A_{33}$ |

3. The arriving rate: we divide the whole day into 24 different time slots by hour, so we need to estimate them by hours. See Table 6.3, Again,

in fact, these parameters can also be chosen at random. A stochastic Poisson arrivals can be generated based on these parameters.

Table 6.3: Parameter Estimation of Arriving Rates

| Type | Parameters |
|---|---|
| A1 | $\lambda_{1,1}$, ... , $\lambda_{1,24}$ |
| A2 | $\lambda_{2,1}$, ... , $\lambda_{2,24}$ |

4. The service times: sample them from empirical datasets, the reason why we do not estimate the probability density function of the service time distribution is that the distribution is really wired and it is hardly to find any theoretical distribution to fit it.

   - Call centers: the service times for A1and A2 are not distinguishable, so there is only one sample set $SA_{CC}$, which denotes the sample space.
   - Ambulance stations:
     There is no obvious difference in the preparation time in ambulance stations, this sample space is called $SA_{PR}$. Because the differences of emergency scales and road conditions, the whole driving time sample sets can be divided into following sub-samples in Table 6.4, 6.5, 6.6:

Table 6.4: The Samples of Driving Time for A1

| Time Period | 0-8 | 8-16 | 16-24 |
|---|---|---|---|
| Weekday | $SA1_{11,P1}, SA1_{11,P2}$ | $SA1_{12,P1}, SA1_{12,P2}$ | $SA1_{13,P1}, SA1_{13,P2}$ |
| Saturday | $SA1_{21,P1}, SA1_{21,P2}$ | $SA1_{22,P1}, SA1_{22,P2}$ | $SA1_{23,P1}, SA1_{23,P2}$ |
| Sunday | $SA1_{31,P1}, SA1_{31,P2}$ | $SA1_{32,P1}, SA1_{32,P2}$ | $SA1_{33,P1}, SA1_{33,P2}$ |

Table 6.5: The Samples of Driving Time for A2

| Time Period | 0-8 | 8-16 | 16-24 |
|---|---|---|---|
| Weekday | $SA2_{11,P1}, SA2_{11,P2}$ | $SA2_{12,P1}, SA2_{12,P2}$ | $SA2_{13,P1}, SA2_{13,P2}$ |
| Saturday | $SA2_{21,P1}, SA2_{21,P2}$ | $SA2_{22,P1}, SA2_{22,P2}$ | $SA2_{23,P1}, SA2_{23,P2}$ |
| Sunday | $SA2_{31,P1}, SA2_{31,P2}$ | $SA2_{32,P1}, SA2_{32,P2}$ | $SA2_{33,P1}, SA2_{33,P2}$ |

Table 6.6: The Samples of Driving Time for B

| Time Period | 0-8 | 8-16 | 16-24 |
|---|---|---|---|
| Weekday | $SB_{11,P1}, SB_{11,P2}$ | $SB_{12,P1}, SB_{12,P2}$ | $SB_{13,P1}, SB_{13,P2}$ |
| Saturday | $SB_{21,P1}, SB_{21,P2}$ | $SB_{22,P1}, SB_{22,P2}$ | $SB_{23,P1}, SB_{23,P2}$ |
| Sunday | $SB_{31,P1}, SB_{31,P2}$ | $SB_{32,P1}, SB_{32,P2}$ | $SB_{33,P1}, SB_{33,P2}$ |

## 6.4 Simulation in Call Center

This section contains the pseudocode of the simulation of the call center and simulation result. Only A1 and A2 are considered in this simulation. The comparison of model results and simulation result is also displayed.

### 6.4.1 Pseudocode of Simulation in Call Center

Here is one small part of pseudocode of simulation in call center during weekday, 0:00-8:00. The reason why not all the pseudocode is presented here is that for the other time period and day type, the procedure is similar.

**Algorithm 6.4.1** Pseudocode of Simulation in Call Center

Matrix S is used to record all the time point in this simulation;
Generate arriving times for A1 and A2 (1 year), these times are put in the first column of S;
{Denote the total number of ambulance trips by R}
**for** j=1:365 **do**
  **for** i=1:R **do**
    **if** j is Weekday **then**
      **if** The arriving time lies in (0, 8] **then**
        $T_{11}$=zeros($C_{11}$, 2); {$T_{11}$ is used to record start time and finish time of the current jobs for centralists}
        ($min_t, ind_t$)=min($T_{11}$(:,2)); {$min_t$ is the minimal finish time and $ind_t$ is the corresponding centralist}
        **if** S(i,1)> $min_t$ **then**
          S(i,2)=S(i,1); {If there is at least one centralist free before a request comes in, then he/she can get serviced immediately}
          $T_{11}$($ind_t$,1)=S(i,1); {the record for the centralist changes to the current task}
        **end if**
        **if** S(i,1)<= $\min_t$ **then**
          S(i,2)=$min_t$; {If all the centralists are busy when a call comes in, then the call maker has to wait until one centralist finishes his/her work}
          $T_{11}$($ind_t$,1)=T($ind_t$,2) {The record for the centralist changes to the current one}
        **end if**
        S(i,3)=S(i,2)+rand($SA_{CC}$);
        $T_{11}$($ind_t$,2)=T($ind_t$,1)+(S(i,3)-S(i,2)); {The records for current centralist are complete}
      **end if**
    **end if**
  **end for**
**end for**

### 6.4.2 Result of Simulation in Call Center

The dataset of 2008 region Groningen is again used here. This simulation has been run for 100 times, the probability of delay is calculated as follows:

$$P^c_{delay} = \frac{\sum_{i=1}^{R} 1(S(i,1)! = S(i,2))}{R} \tag{6.1}$$

$R$ is the total number of the ambulance trips in one year.

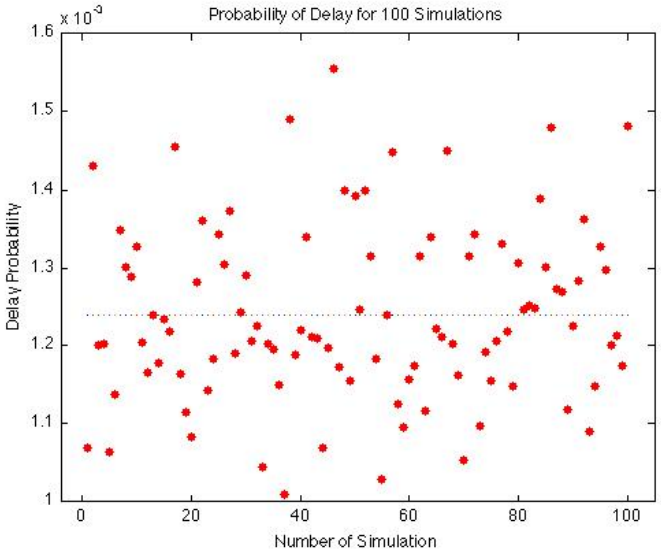
If the arriving time for a call is different from the time a call comes through, we claim there is delay, this probability is calculated by the total number of delay trips divided by the total number of trips. For region Groningen, the model result in Chapter 4 is used here,

Table 6.7: The Staffing Level in Groningen ($M/G/c/c, \alpha = 0.01$)

| Time Period | 0-8 | 8-16 | 16-24 |
|---|---|---|---|
| Weekday | 2 | 3 | 3 |
| Saturday | 2 | 3 | 3 |
| Sunday | 3 | 3 | 3 |

The result is shown in Figure 6.1.

Figure 6.1: $P^c_{delay}$ in Simulation(Groningen)



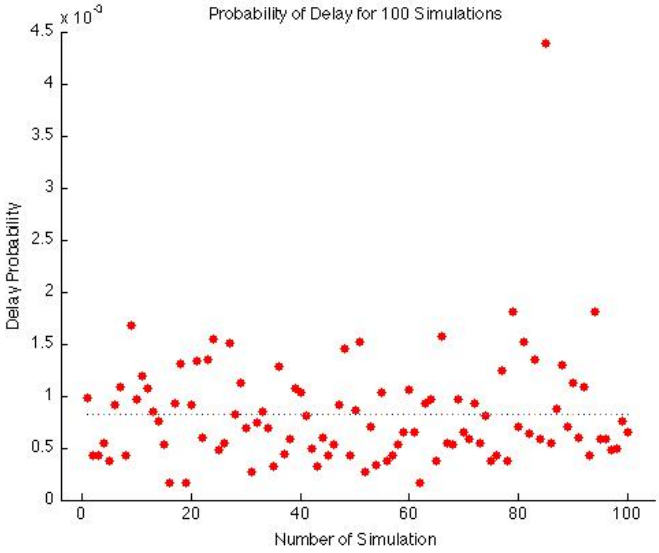

The results of $P^c_{delay}$ can be seem in the graph, all of them lie between 0.0010 and 0.0015, the variation is quite small and the average of them is 0.00124. The simulation indicates the performances of the model results got in Chapter 4 are quite good.

The simulation result for region Drenthe is in figure 6.2 if the staffing level in Chapter 4 is used here,

Table 6.8: The Staffing Level in Drenthe ($M/G/c/c, \alpha = 0.01$)

| Time Period | 0-8 | 8-16 | 16-24 |
|---|---|---|---|
| Weekday | 2 | 3 | 3 |
| Saturday | 2 | 3 | 2 |
| Sunday | 2 | 3 | 2 |

Figure 6.2: $P^c_{delay}$ in Simulation(Drenthe)




The results of $P^c_{delay}$ can be seem in the graph, all of them lie between 0.0001 and 0.0045, the variation is quite small and the average of them is 0.0008. The simulation indicates the performances of the model results got in Chapter 4 are quite good.

## 6.5 Simulation in Ambulance Station

This section contains the pseudocode of simulation in ambulance station and the result of simulations. Again, a comparison between model result and simulation result is also displayed here. The simulation is dealing with A1, A2 and B all together.

### 6.5.1 Pseudocode of Simulation in Ambulance Station

Here is one small part of the pseudocode for the simulation in the ambulance station during weekday, 0:00-8:00. The pseudocode of other time intervals and other day types is similar.

**Algorithm 6.5.1** Pseudocode of Simulation in Ambulance Station

Matrix W is used to record all the time point in this simulation;
Use the times that the ambulances are ordered from the dataset of region Groningen, 2008;
{Denote the total number of task by R}
**for** j=1:365 **do**
  **for** i=1:R **do**
    **if** j is Weekday **then**
      **if** The arriving time lies in (0, 8] **then**
        $T_{11}$=zeros($A_{11}$, 2); {$T_{11}$ is used to record start time and finish time of the current jobs for ambulances}
        ($min_t, ind_t$)=min($A_{11}$(:,2)); {$min_t$ is the minimal finishes time and $ind_t$ is the corresponding ambulance}
        **if** W(i,1)> $min_t$ **then**
          W(i,2)=W(i,1); {If there is at least one ambulance is free before a request come in, then he/she can get serviced immediately}
          $T_{11}$($ind_t$,1)=W(i,1); {the record for the ambulance changes to the current task}
        **end if**
        **if** W(i,1)<= $\min_t$ **then**
          W(i,2)=$min_t$; {If all the ambulances are busy when a call come in, then the request have to wait until one ambulance finish its work}
          $T_{11}$($ind_t$,1)=T($ind_t$,2) {The record for the ambulance change to the current one}
        **end if**
        W(i,3)=W(i,2)+rand($SA_{PR}$); {The preparation time can be got from the sample}
        **if** Customer is type A1 **then**
          W(i,4)=W(i,3)+rand($SA1_{11}^{P1}$);
          W(i,5)=W(i,4)+rand($SA1_{11}^{P2}$);
          $T_{11}$($ind_t$,2)=T($ind_t$,1)+(W(i,5)-W(i,2));
        **end if**
        **if** Customer is type A2 **then**
          W(i,4)=W(i,3)+rand($SA2_{11}^{P1}$);
          W(i,5)=W(i,4)+rand($SA2_{11}^{P2}$);
          $T_{11}$($ind_t$,2)=T($ind_t$,1)+(W(i,5)-W(i,2));
        **end if**
        **if** Customer is type B **then**
          W(i,4)=W(i,3)+rand($SB_{11}^{P1}$);
          W(i,5)=W(i,4)+rand($SB_{11}^{P2}$);
          $T_{11}$($ind_t$,2)=T($ind_t$,1)+(W(i,5)-W(i,2));
        **end if**
        {The records for the current ambulance and task are complete}
      **end if**
    **end if**
  **end for**
**end for**

### 6.5.2 Results of Simulation in Ambulance Station

The probability of delay is calculated as follow:

$$P^a_{delay} = \frac{\sum_{i=1}^{R} 1(W(i,1)! = W(i,2))}{R} \tag{6.2}$$

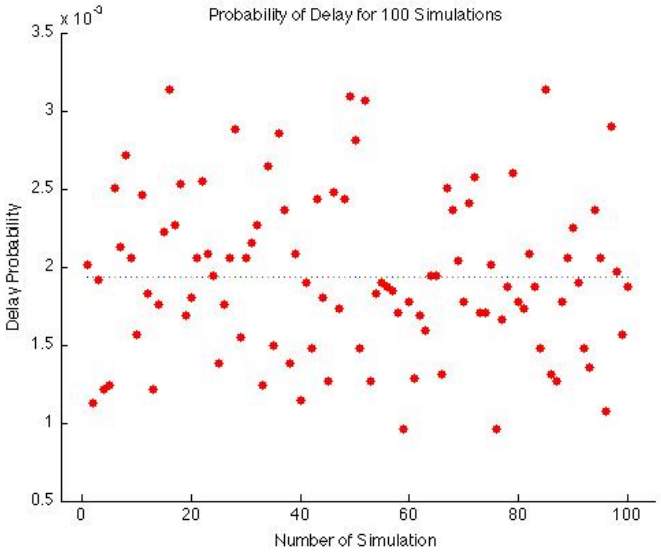
where $R$ is the number of the trips. If the time of ordering an ambulance is different from the the time an ambulance become available, we claim there is delay, this probability is calculated by the total number of delay divided by the total number of cases. The model result for region Groningen in Chapter 5 is use here,

**Groningen**

Table 6.9: The Schedule of Ambulances in Groningen ($M/G/c/c, \epsilon = 10^{-4}$)

| Time Period | 0-8 | 8-16 | 16-24 |
|---|---|---|---|
| Weekday | 13 | 28 | 19 |
| Saturday | 14 | 19 | 18 |
| Sunday | 15 | 19 | 17 |

Figure 6.3: $P^a_{delay}$ in Ambulance Station(Groningen)




It can be shown that $P^a_{delay}$ for 100 runs, the variation is quite small and all the results lies in $4.01 * 10^{-5}$ and $3.66 * 10^{-4}$, and the average is $9.27 * 10^{-5}$.

Compared to the model result of this schedule in Chapter 5, which is less than $1 \times 10^{-4}$, this probability is larger. Although the $P^a_{delay}$ in simulation is larger, the simulation still indicates a good performance of the model results in Chapter 5.

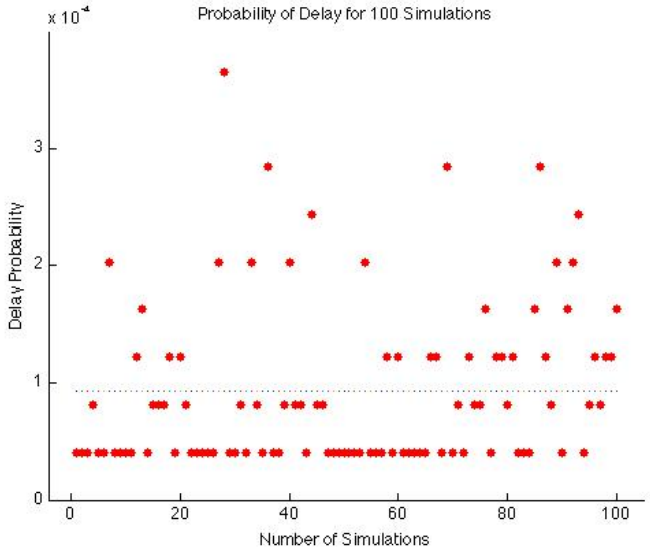
The similar result for Drenthe is in Figure 6.4, again, the model result from Chapter 5 is used here,

**Drenthe**

Table 6.10: The Schedule of Ambulances in Drenthe ($M/G/c/c, \epsilon = 10^{-5}$)

| Time Period | 0-8 | 8-16 | 16-24 |
|---|---|---|---|
| Weekday | 11 | 19 | 14 |
| Saturday | 12 | 17 | 15 |
| Sunday | 13 | 16 | 14 |

Figure 6.4: $P^a_{delay}$ in Ambulance Station(Drenthe)




It can be shown that $P^a_{delay}$ for 100 runs, the variation is quite small and all the results lies in $4 * 10^{-5}$ and $4 * 10^{-4}$, and the average is $9.03 * 10^{-5}$. Compared to the model result of this schedule in Chapter 5, which is less than $1 \times 10^{-5}$, this probability is not far away. The simulation indicates a good performance of the model results in Chapter 5.

## 6.6 Conclusion

The simulations are set up to evaluate the performance of the systems in practice if the model results are used. Although there are differences between model results and simulation results, the efficiency of the new schedules is still kept. So the most important conclusion we can draw from this chapter is the simulations also guarantee the good performances of the schedules got from the mathematical models in Chapter 4 and Chapter 5.

# Chapter 7

# Conclusions and Recommendations

This chapter summarizes the conclusions that can be drawn from the study described in this report. As a model for future tools to determine optimal capacity of ambulance resources, the queueing models are shown to be quite useful. Therefore a lot of research has been done to find models that can give insight in the advantages and limitations of queueing models. In this thesis, the mathematical models used to determine the optimal staffing levels in call centers and the optimal schedules in ambulance stations are presented followed by the applications of these models. The conclusions and recommendations for centralists staffing and ambulances scheduling are displayed in section 7.1 and 7.2 respectively. Finally, the thoughts of future research are listed in section 7.3.

## 7.1 Conclusions and Recommendations in Centralists Staffing

Since there are many different types of queueing models, the most related ones are presented. The Erlang C model serves as a start because this is the most commonly used queuing model in call centers currently. In this model many characteristic properties of ambulance service system in the project have been included, like the Poisson arrivals and limited number of centralists. Nevertheless there also are properties that have not been included, like non-exponential service times. The later exploration of the M/G/$\infty$ and M/G/c/c models gives more freedom to investigate systems with non-exponential service times.

Considering the performance measure of these systems, it was shown that light traffic is important in satisfying rare delay when a centralist is needed.

The delay here means the requests can be answered immediately. When the number of centralists becomes larger, the probability of delay becomes smaller. A simulation has been set up later to evaluate the performance of the staffing level got in mathematical models. The result reveals the current staffing level is quite effective. Due to the small scale of call centers in the Netherlands, a decrease of the number of centralists will have considerable effect on the probability of delay. So a recommendation is to keep the current staffing level. Another practical question asked from the experts in ambulance system is whether it is a wise idea to combine several call center. The mathematical model results show the system will be much more efficiency if combinations are applied.

Finally, we can see that the mathematical models can also used to predict the optimal staffing level is the prediction of the incoming rates and service rates are done.

## 7.2 Conclusions and Recommendations in Ambulance Scheduling

Apart from research for the staffing centralists, it is also investigated that how to generate an optimal schedule for ambulances. Several mathematical models are constructed to solve this problem. The probability of delay for ambulances depends on many factors, but mostly on the driving time of an ambulance trip, which indicates that the geographical aspect should be included in future research after this project. The optimal schedules of ambulances should satisfy that the percentage of the cases that the response time is less than 15 min and 30 min for A1 and A2 respectively should be at least 95%. The model construction starts with a queueing model again. Unfortunately, more data investigation indicates this constraint is hard to meet if only the capacities are discussed, therefore, a more efficient performance measure, delay probability is used to determine whether this system can provide effective service to all the requests. Based on this new constraint, continuous time Markov chain model and queueing models are constructed to determine the optimal number of ambulances. The calculation based on these models reveals that the current schedules are almost optimal, only slightly adjustment is needed to make it more economical. The calculation also indicates other rules of thumb: it is not wise to schedule too many B trips during 16:00-17:00 because this is the most busy time for possible A trips.

Again, we can see that the mathematical models can also used to predict the optimal staffing level is the prediction of the incoming rates and service rates are done.

## 7.3 Future Work

It can never be concluded that the research on this topic is complete. However, we are always allowed to give more insight to what can be approved in the future both in practice and in theory. From a practical point of view, more time points which can give more support to the performance measure, such as, the time that a phone call gets through and an ambulance becomes available, can be recorded. There results can used to proceed the research much more efficiently. From a theoretical point of view, the theory of light traffic and the optimal locations of ambulances resources can be considered in the future research.

# Bibliography


[1] I. Adan and J. Resing. Queueing theory. *Eindhoven University of Technology, The Netherlands*, 2001.

[2] S.R. Agnihothri and P.F. Taylor. Staffing a centralized appointment scheduling department in Lourdes Hospital. *Interfaces*, pages 1–11, 1991.

[3] O.Z. Akin and P.T. Harker. Computing performance measures in a multi-class multi-resource processor-shared loss system. *European Journal of Operational Research*, 123(1):61–72, 2000.

[4] T. Andersson, P. Varbrand, and S. Petersson. Dynamic ambulance relocation for a higher preparedness. In *Proc. of the 35lh Annual Meeting of the Decision Sciences Institute*, 2004.

[5] B. Andrews and H. Parsons. Establishing telephone-agent staffing levels through economic optimization. *Interfaces*, 23(2):14–20, 1993.

[6] J.B. Atkinson, I.N. Kovalenko, N.Y. Kuznetsov, and K.V. Mikhalevich. Heuristic methods for the analysis of a queuing system describing emergency medical service deployed along a highway. *Cybernetics and Systems Analysis*, 42(3):379–391, 2006.

[7] J. Atlason, M.A. Epelman, and S.G. Henderson. Call center staffing with simulation and cutting plane methods. *Annals of Operations Research*, 127(1):333–358, 2004.

[8] H. Aytug and C. Saydam. Solving large-scale maximum expected covering location problems by genetic algorithms: A comparative study. *European Journal of Operational Research*, 141(3):480–494, 2002.

[9] M.O. Ball and F.L. Lin. A reliability model applied to emergency service vehicle location. *Operations Research*, 41(1):18–36, 1993.

[10] R. Batta, J.M. Dolan, and N.N. Krishnamurthy. The maximal expected covering location problem: Revisited. *Transportation Science*, 23(4):277, 1989.

[11] F. Borrás and J.T. Pastor. The ex-post evaluation of the minimum local reliability level: an enhanced probabilistic location set covering model. *Annals of Operations Research*, 111(1):51–74, 2002.

[12] L. Brotcorne, G. Laporte, and F. Semet. Ambulance location and relocation models. *European Journal of Operational Research*, 147(3):451–463, 2003.

[13] Cosgrove M.J. Buffa, E.S. and B.J. Luce. An integrated work shift scheduling system. *Decision Sciences*, 7(4):620–630, 2007.

[14] D.Y. Burman and D.R. Smith. A light-traffic theorem for multi-server queues. *Mathematics of Operations Research*, 8(1):15–25, 1983.

[15] T. Cezik, O. Gunluk, and H. Luss. An integer programming model for the weekly tour scheduling problem. *Naval Research Logistics*, 48(7):607–624, 2001.

[16] N.G. Duffield and W. Whitt. Control and recovery from rare congestion events in a large multi-server system. *Queueing Systems*, 26(1):69–104, 1997.

[17] L.C. Edie. Traffic delays at toll booths. *Journal of the Operations Research Society of America*, 2(2):107–138, 1954.

[18] L.C. Edie. Case Histories Five Years after-A Symposium: Review of Port of New York Authority Study. *Operations Research*, 8(2):263–277, 1960.

[19] A.K. Erlang. On the rational determination of the number of circuits. *The life and works of AK Erlang*, pages 216–221, 1948.

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

[21] J.J. Gordon and M.S. Fowler. Accurate force and answer consistency algorithms for operator services. In *The fundamental role of teletraffic in the evolution of telecommunications networks: proceedings of the 14th International Teletraffic Congress, ITC 14, Antibes Juan-les-Pins, France, 6-10 June 1994*, page 339. Elsevier, 1994.

[22] S. Halfin and W. Whitt. Heavy-traffic limits for queues with many exponential servers. *Operations research*, 29(3):567–588, 1981.

[23] S.G. Henderson and A.J. Mason. Estimating ambulance requirements in Auckland, New Zealand. In *Proceedings of the 31st conference on Winter simulation: Simulation—a bridge to the future-Volume 2*, page 1674. ACM, 1999.

[24] W.B. Henderson and W.L. Berry. Heuristic methods for telephone operator shift scheduling: an experimental analysis. *Management Science*, 22(12):1372–1380, 1976.

[25] A. Ingolfsson, S. Budge, and E. Erkut. Optimal ambulance location with random delays and travel times. *Health Care Management Science*, 11(3):262–274, 2008.

[26] J.P. Jarvis. Approximating the equilibrium behavior of multi-server loss systems. *Management Science*, 31(2):235–239, 1985.

[27] O.B. Jennings, A. Mandelbaum, W.A. Massey, and W. Whitt. Server staffing to meet time-varying demand. *Management Science*, pages 1383–1394, 1996.

[28] G. Jongbloed and G. Koole. Managing uncertainty in call centres using Poisson mixtures. *Applied Stochastic Models in Business and Industry*, 17(4):307–318, 2001.

[29] Z. Khalil, G. Falin, and T. Yang. Some analytical results for congestion in subscriber line modules. *Queueing Systems*, 10(4):381–402, 1992.

[30] P.J. Kolesar and L.V. Green. Insights on service system design from a normal approximation to Erlang's delay formula. *Production and Operations Management*, 7(3):282–293, 2009.

[31] P. Kuhn and T.P. Hoey. Improving police 911 operations in Washington, DC. *National Productivity Review*, 6(2):125–133, 2006.

[32] S.K. Kwan, M.M. Davis, and A.G. Greenwood. A simulation model for determining variable worker requirements in a service operation with time-dependent customer demand. *Queueing Systems*, 3(3):265–275, 1988.

[33] R.C. Larson. Improving the Effectiveness of New York City's 911. *Analysis of Public Systems, MIT Press, Cambridge, Mass*, 1972.

[34] R.C. Larson. A hypercube queuing model for facility location and redistricting in urban emergency services. *Computers & Operations Research*, 1(1):67–95, 1974.

[35] Y. Levy, S. Durinovic-Johri, and R.A. Milito. Dynamic network call distribution with periodic updates. In *The Fundamental Role of Teletraffic in the Evolution of Telecommunications Networks: Proceedings of the 14th International Teletraffic Congress (ITC 14), Antibes Juan-les-Pins, France, 6-10 June 1994*, page 85. Elsevier Science Ltd, 1994.

[36] V.A. Mabert. Short interval forecasting of emergency phone call (911) work loads. *Journal of Operations Management*, 5(3):259–271, 1985.

[37] A. Mandelbaum, W.A. Massey, and M.I. Reiman. Strong approximations for Markovian service networks. *Queueing Systems*, 30(1):149–201, 1998.

[38] A. Mandelbaum, W.A. Massey, M.I. Reiman, A. Stolyar, and B. Rider. Queue lengths and waiting times for multiserver queues with abandonment and retrials. *Telecommunication Systems*, 21(2):149–171, 2002.

[39] A. Mandelbaum and G. Pats. State-dependent stochastic networks. Part I: Approximations and applications with continuous diffusion limits. *The Annals of Applied Probability*, 8(2):569–646, 1998.

[40] V. Marianov and C. Revelle. The queueing maximal availability location problem: a model for the siting of emergency vehicles. *European Journal of Operational Research*, 93(1):110–120, 1996.

[41] A.J. Mason, D.M. Ryan, and D.M. Panton. Integrated simulation, heuristic and optimisation approaches to staff scheduling. *Operations Research*, 46(2):161–175, 1998.

[42] Peter Duijf, Erik Grummels, Gerard Leerkes. *Ambulances in-zicht 2008*. 2009.

[43] E.J. Pinker and R.A. Shumsky. The efficiency-quality trade-off of cross-trained workers. *Manufacturing & Service Operations Management*, 2(1):32–48, 2000.

[44] M. Segal. The operator-scheduling problem: A network-flow approach. *Operations Research*, 22(4):808–823, 1974.

[45] M. Segal and DB Weinberger. Turfing. *Operations Research*, 25(3):367–386, 1977.

[46] D. Serra and V. Marianov. The p-median problem in a changing network: the case of Barcelona* 1. *Location Science*, 6(1-4):383–394, 1998.

[47] L.V. Snyder and M.S. Daskin. Reliability models for facility location: The expected failure cost case. *Transportation Science*, 39(3):400–416, 2005.

[48] D.Y. Sze. A queueing model for telephone operator staffing. *Operations Research*, 32(2):229–249, 1984.

[49] W. Whitt. Heavy traffic approximations for service systems with blocking. *AT&T Bell Lab. Tech. J*, 63:689–708, 1984.

[50] W. Whitt. Dynamic staffing in a telephone call center aiming to immediately answer all calls. *Operations Research Letters*, 24(5):205–212, 1999.

# List of Figures

# List of Tables