MASTER

Demand forecasting and hierachical workforce staffing in Customer Service Chat (CSC) Centres

Leenen, W.E.M.

Award date:
2016

Eindhoven, June 2016

Demand forecasting and hierarchical workforce staffing in Customer Service Chat (CSC) Centres

by
W.E.M. (Wouter) Leenen

BSc Industrial Engineering & Management Science
Student identity number 0747825

In partial fulfillment of the requirements for the degree of

Master of Science
in Operations Management and Logistics

Supervisors:
dr. S.S. (Shaunak) Dabadghao, TUE, OPAC
dr.ir. N.P. (Nico) Dellaert, TUE, OPAC
Company Supervisors:
A. (Arno) de Wolf, ORTEC Consulting
J. (Joost) Rijlaarsdam, Web1on1

TUE. School of Industrial Engineering.
Series Master Theses Operations Management and Logistics

Subject headings: Regression Analysis, Time Series Forecasting, Capacity Planning, Queuing Theory, Workforce Staffing, Mathematical Programming

Abstract

A Customer Service Chat (CSC) centre is a new type of contact centre where Instant Messaging applications are used to serve customers over the internet. In this thesis, a workforce staffing model is developed to optimize the staffing decisions at a CSC centre located in The Netherlands: Web1on1. Web1on1 has multiple clients for which they operate the live chat. To deliver high quality conversations, chat agents need to have certain skills and knowledge. Those requirements differ between clients. Therefore, Web1on1 has clustered their clients such that agents can be trained to chat for a specific group of clients. Hence, there are multiple types of chats and consequently agents. Three models are developed. First, a forecasting model to predict the arrival rate of chats which is required to determine the expected workload. This model consists of a regression and time series model. Hereafter, a queueing model is used to translate the workload into capacity requirements. Due to the hierarchical structure of the workforce, there are multiple staffing solutions possible to achieve the required service level. The last model is a Mixed Integer Non-Linear Programming (MINLP) model that optimizes the staffing levels by making use of capacity transfers from one cluster to another. The forecasting and capacity model are implemented at the CSC centre Web1on1. The staffing model has been developed for scientific purpose.

Management summary

Contact centres and in particular call centres, have been around for decades. Although calling is still popular, new communication tools like chat and (instant-) messaging have emerged recently and will grow in market share in the future. Companies are adapting their contact centres to those new communication techniques.

Web1on1 is a provider of live chat. This is a functionality on websites which offers visitors the opportunity to contact the corresponding company via an Instant Messaging application. This type of contact centre is called a Customer Service Chat (CSC) centre.

Since the start in 2010, Web1on1 has enjoyed a sustainable growth. More and more clients are convinced that live chat adds value. Due the increased amount of work, it becomes more important to operate the centre efficiently. The live chat is operated by employees of Web1on1, who are called chat agents. The cost of this personnel is a large part of the operating budget (60-80% of the total operating budget), because of the labour-intensive character of the work (Aksin, Armony, & Mehrotra, 2007). Therefore contact centres benefit from an accurate prediction of the workload and an appropriate number of agents scheduled. The workload is the total amount of work that arrives per time unit. Having too few agents working in a certain period results in an under-performance, too many will unnecessarily increase costs.

The result of this thesis is three models - forecasting, capacity and staffing model - to support the decision making in staffing CSC centres like Web1on1. An overview of the modelling approach is graphically presented in Figure 1.

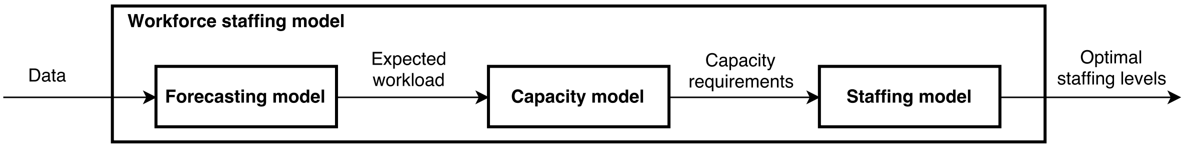


Figure 1: Modelling approach

As said, to make effective staffing decisions, it is important to have accurate forecasts of the workload arriving to the centre. Therefore, the first model developed, the forecasting model, is a model which predicts the workload. This model uses historic data to find aspects which have an effect on the number of chat arrivals. The most important findings of the data analysis are listed below.

- The amount of chats arriving to the centre is highly related to the website traffic;
- The arrival rate depends on the day of the week and hour of the day. Over the years, the distribution of chat arrivals over the day and week have changed;
- The number of chat arrivals is less on public holidays;
- There is no clear relation between the arrival rate and the weather. Although, there are some periods of the week which seem to be affected.

After the workload has been forecasted, the required number of agents to process this amount of work properly is determined by the capacity model. The CSC centre

Web1on1 can be modelled as a queueing system, see Figure 2. When all agents are busy, arriving customers are blocked. There is no waiting queue. Therefore, the most important indicator for the service level of Web1on1 is the probability of blocking customers. What an acceptable blocking probability is, must be decided by the management of the CSC centre. This is a trade-off between cost and quality. Combining this blocking probability with the expected number of arrivals from the forecasting model, the number of agents required to staff is determined.

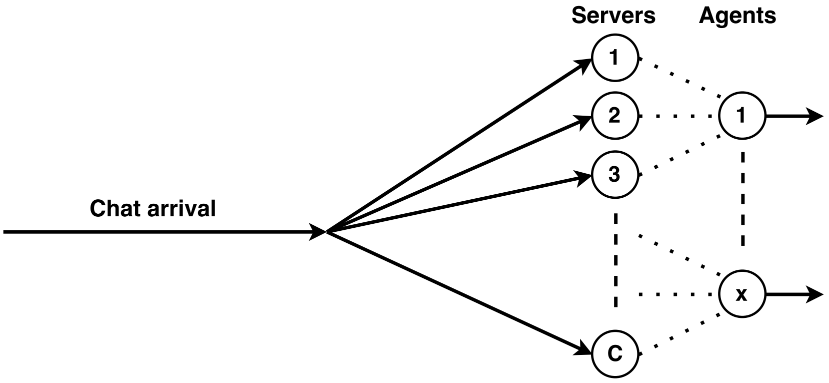


Figure 2: Queueing system at Web1on1

The third and final model optimizes the staffing levels per client cluster. Web1on1 has multiple clients and each of these require a chat agent to have a certain set of skills and knowledge. Web1on1 has clustered their clients such that agents can be trained to chat for a certain group of clients. The capacity model staffs each cluster independently, while the staffing model optimizes the server capacity over the different clusters by transferring excess capacity from one cluster to another. This staffing model provides a solution which takes the aggregated performance of the centre into account, while the capacity model staffs each cluster independently of each other.

The forecasting and capacity model are implemented at Web1on1 and the staffing model is only developed for scientific purposes. The performances of both models are compared to the historic performance of Web1on1. The up-time of the system (100% - blocking probability) and the number of agents staffed, is compared to the performance for the period January 2015 through September 2015. The required up-time is set to 90%. Furthermore, two staffing horizons are compared for the model implemented at Web1on1: 1 day and 7 days. The results are shown in Table 1. 'Model' refers the the model implemented at Web1on1. The results of the staffing model with a 1 day horizon, are shown in the last row.

Table 1: Performance of the models for the period January 2015 through September 2015

	Service level				Agents staffed	
	Mean	Variance	>90%	>95%	Amount	Saving (in%)
Historic	91.62	0.90	70.20	53.40	19695	
Model (7 days)	95.27	0.60	82.96	72.39	16988	13.745%
Model (1 day)	95.30	0.56	83.73	70.20	16993	13.719%
Staffing model	95.24	0.12	91.40	61.10	16934	14.019%

The model implemented at Web1on1 shows an improved performance when compared to the historical achievements. The average service level is increased while the number of agents staffed to achieve this, is reduced by more than 13%. The variance is also reduced which means that the service level is more stable. The amount of periods which achieve the required service level of 90% is increased as well.

Reducing the horizon from 7 days to 1 day does not improve the performance of the model much. The advice is to run the model every day and make staffing decisions one week in advance. The staffing levels can be adapted if the forecasts change substantially within this week.

When using the staffing model in addition to the forecasting and capacity model implemented at Web1on1, the average service level is not improved. The variance is further reduced which means that the service level is even more stable. Also the percentage of the periods in which the service level of 90% is achieved, is increased with almost 8 percentage points to more than 91%. The number of periods with an overcapacity (service level > 95%) is reduced, because of the transfer of excess capacity. The number of agents staffed is slightly decreased to a saving of more than 14% when compared to historic performance.

The models developed have shown to improve the performance of Web1on1 for the period considered. The service levels are improved while the labour capacity required is reduced. One of the important aspects of the solution developed in this thesis is the modular approach. It is possible to adapt each model independently when the conditions this require. This makes the model flexible and relatively simple to adapt to the circumstances of the CSC centre.

The models have underlying assumptions which are not entirely realistic. For example, the capacity model assumes that the arrival rate is constant within the time period. This is not the case. The arrival rate will vary within an hour. This is not taken into account by the capacity model. Due to a mismatch between assumptions and the real world, the model will perform not exactly the same as in theory. This should always be taken into account when making use of models. Another important drawback of the model, is the high sensitivity of changes in the data. The models base the decisions on data and are therefore dependent on the quality of the data delivered. Changes in the data set can cause errors which make the models useless. This indicates the importance of model maintenance.

Preface

First of all, I would like to thank Shaunak Dabadghao for being a constructive supervisor along the project. You always had time to help me structure my ideas, give me feedback on the progress I had made and to point me in the right direction. It has been a pleasure working under your supervision. Thank you for that! Nico Dellaert, I also want to thank you for the meetings we had. Those were very valuable and those helped me deliver this result. Thanks!

I want to thank Arno de Wolf for being my supervisor at ORTEC Consulting. You were an important discussion partner for the content of this project. In addition to that, you were also very helpful in the soft skills required as a consultant to deliver the service properly at the customer. Pascha Iljin, thank you for offering me the opportunity to work on this very interesting project. I want to thank the Analytics and Optimization team for making me feel comfortable and everybody at ORTEC who helped me achieving this result. A special thanks to Ronald Buitenhek who has always been very interested in the project. You have provided me with helpful insights.

Thank you to Joost Rijlaarsdam and the entire headquarter of Web1on1 for making these months great. You have welcomed me with an unforgettable business trip to Nice in France and since the very first moment I have felt part of the team. Thank you!

A special thanks to my parents, Paul and Carin, and sister Janou. Without your unconditional support, I would not have been able to achieve this. Also to my dear girlfriend Ellen for your endless love and giving me the confidence whenever I needed it. And last but not least, my friends and in particular Nik(e)s. Thank you for making this ride unforgettable.

With the delivery of this master thesis, my life as a student comes to an end. Sometimes it has been tough, but I enjoyed everything. It concludes a great period and marks the beginning of a new life. I am curious what the future will bring. Enjoy reading!

Wouter Leenen
May 31, 2016

Contents

List of Figures

List of Tables

List of Acronyms

ADF	Augmented Dickey–Fuller
ACF	Auto Correlation Function
ANOVA	Analysis of Variance
ARMA(p,q)	Autoregressive Moving Average model
BP	Blocking Probability
CP	Cutting Plane method
CSC	Customer Service Chat
ES	Exponential Smoothing
GPS	Generalized Pattern Search
HSD	Honest Significant Difference
IM	Instant Messaging
MIP	Mixed Integer Programming model
LP	Linear Programming model
MADS	Mesh Adaptive Direct Search
MP	Mathematical Programming
MINLP	Mixed Integer Non-Linear Programming model
OPTI	OPTimization Interface
PACF	Partial Auto Correlation Function
PI	Prediction Intervals
PS	Pattern Search
SL	Service Level
SRS	Square-Root Staffing rule
QoS	Quality of Service
WFM	Workforce Management

1 Introduction

The project for this thesis is conducted at Web1on1, a Customer Service Chat (CSC) centre. The first section of this chapter (1.1) is dedicated to the introduction of live chat and Web1on1 in particular. This new type of contact centre has emerged in the recent years and Web1on1 has grown rapidly since. The challenge Web1on1 faces, is introduced in section 1.2. Hereafter, the research questions and research design are addressed in section 1.3 and 1.4 respectively.

1.1 Web1on1, a live chat provider

Contact centres and in particular call centres, have been around for decades. Although calling is still popular, new communication tools like chat and (instant-) messaging have emerged recently and will grow in market share in the future. Companies are adapting their contact centres to those new communication techniques.

Nowadays, messaging is a communication tool widely used by individuals. Instant Messaging (IM) fits well in this trend. Web1on1 is a provider of live chat. This is a functionality on websites which offers visitors the opportunity to receive online service of the corresponding company via an IM application. Web1on1 delivers a complete service. This includes the implementation of the chat application on the client's website, the chat software itself and the employees who handle the chats. Most of the clients of Web1on1 are firms in the automotive industry. Next to this, Web1on1 is active in the real estate market. This thesis only focuses on the part of the centre related to the automotive industry, because this part dominates their operations. In the remainder of this thesis, the companies which make use of Web1on1's services, are called clients. Customers are the website visitors who make use of the live chat.

Since the start in 2010, Web1on1 has enjoyed a sustainable growth. More and more clients are convinced that live chat adds value. Due the increased amount of work, it becomes more important to operate the centre efficiently. The impact this has on the performance is significant and this step required to be able to continue the growth of the centre.

Before discussing the details of the design of the CSC centre, it is important to understand how the process of handling chats looks like in general. Customers who are using the chat functionality have a conversation with real people, so called chat agents. Customers can receive the service they require and get answers to their questions. After the website visitor starts the chat, it is assigned to an agent of Web1on1. After a while, when the conversation comes to an end, the chat is processed and stored in a database. Periodically, clients receive an update about the chats that have been processed at their website(s). This gives them insight in their customers and what is important to them.

As explained, Web1on1 has multiple clients. To serve customers of client X requires different knowledge and skills then for customers of client Y. Web1on1 has multiple types of chats and consequently different types of agents. Every agent has to have the fundamental skills to have a structured chat conversation and must be able to add value for the customer. In addition, client specific knowledge is required. To guarantee the quality of their service, Web1on1 educates their agents for a specific group of clients. These client groups are called clusters and Web1on1 has three of them: cluster 1, 2 and 3. Each cluster requires a certain set of knowledge and experience to serve the customers. There are three types of agents: type 1, 2 and 3. Table 1.1 shows which agent type can serve customers of which cluster.

Table 1.1: The relation between agent types and clusters

Cluster (i)	Agent type (j)		
1	1		
2		2	
3	1	2	3

1.2 Problem description

The total amount of work that arrives at the CSC centre per unit of time is called the workload. This load has to be processed such that the requirements as agreed upon with their clients are met. The level of service delivered is called the Service Level (SL) of the centre. Depending on the amount of workload arriving and the SL that has to be delivered, a certain agent capacity is required. As indicated by the paper of Aksin et al. (2007), labour costs determine a large part of the operating budget of contact centres. For a CSC centre like Web1on1, one of the important concerns is to staff chat agents with the right skills at the right time. Having too many agents staffed will increase labor cost unnecessarily while being understaffed will violate the SL requirements. Therefore, effective and efficient staffing of chat agents is one of the core objectives of Web1on1.

Given a certain capacity of agents staffed, there is a maximum amount of chats that can be served at the same time. Unlike call centre agents, one chat agent is able to handle multiple customers simultaneously. At Web1on1, each agent corresponds to three independent servers. Chapter 5 will address this aspect in more detail.

When all agents are serving the maximum number of customers simultaneously, the chat widget disappears on the website and customers cannot start a chat. The system is occupied and customers are blocked. The Blocking Probability (BP) is an important measure for the SL of the CSC centre. Web1on1 measures the percentage of the time the system is not occupied, the so called up-time. There is strived to an up-time of 90%. A more detailed description of this performance measure will be given in chapter 5. Figure 1.1 gives a graphical representation of the up-time historically. The four horizontal lines indicate the levels of 90%, 92.5%, 95% and 97.5%. Although the majority of the time, Web1on1 performs above or close to 90%, there is still a significant amount of hours which can be improved. Especially in the first four months, the SL fluctuates a lot and is often lower than the desired level of 90%. In the period May until June, technical issues have caused unreliable data which makes it hard to evaluate the performance in this period. In the last three months visualised, Web1on1 seems to perform better. There are less hours with a low SL and there is less fluctuation.

Before the CSC centre is able to staff agents, the workload has to be forecasted. If the actual amount of work deviates a lot from the prediction, the agent capacity staffed will be too little or too much and will cause fluctuations in the SL. To prevent Web1on1

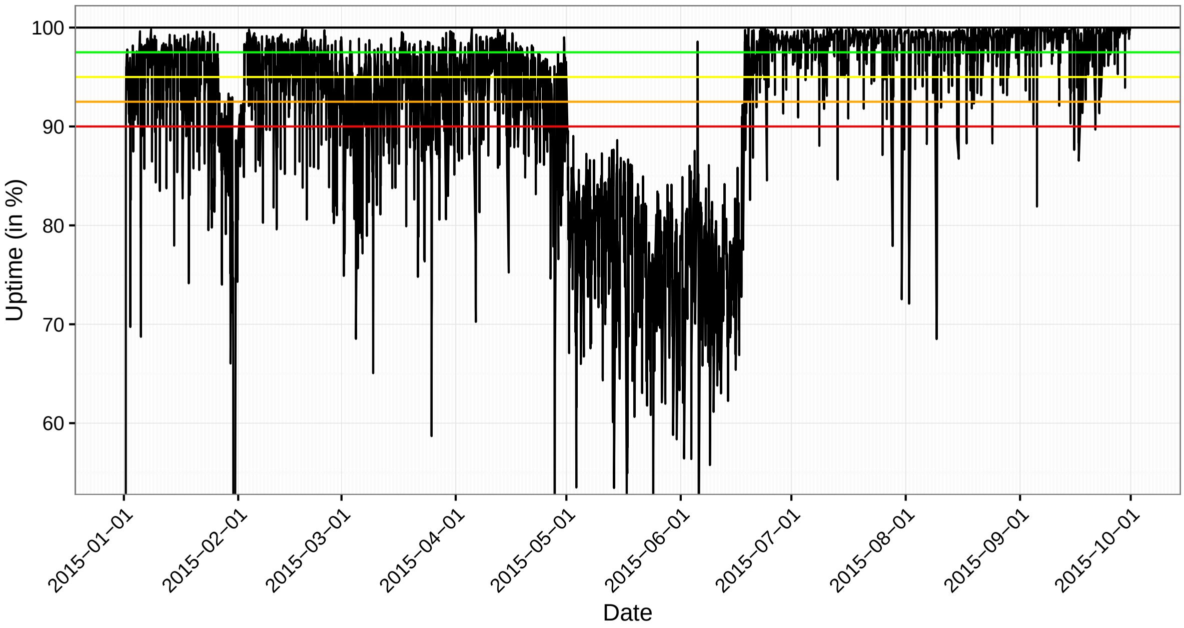


Figure 1.1: Realised SL Web1on1 Jan - Sep 2015

from this scenario, having accurate workload predictions is the first important step in the process of an effective staffing of agents. Figure 1.2 shows the prediction error of Web1on1 historically. This is the relative difference between the expected and actual workload (equation 1.1). The lack of accurate predictions is one of the reasons for the unstable performance of Web1on1.

$$\text{Prediction error (in \%)} = \frac{\text{Expected workload} - \text{Actual workload}}{\text{Actual Workload}}\text{x}100\% \tag{1.1}$$

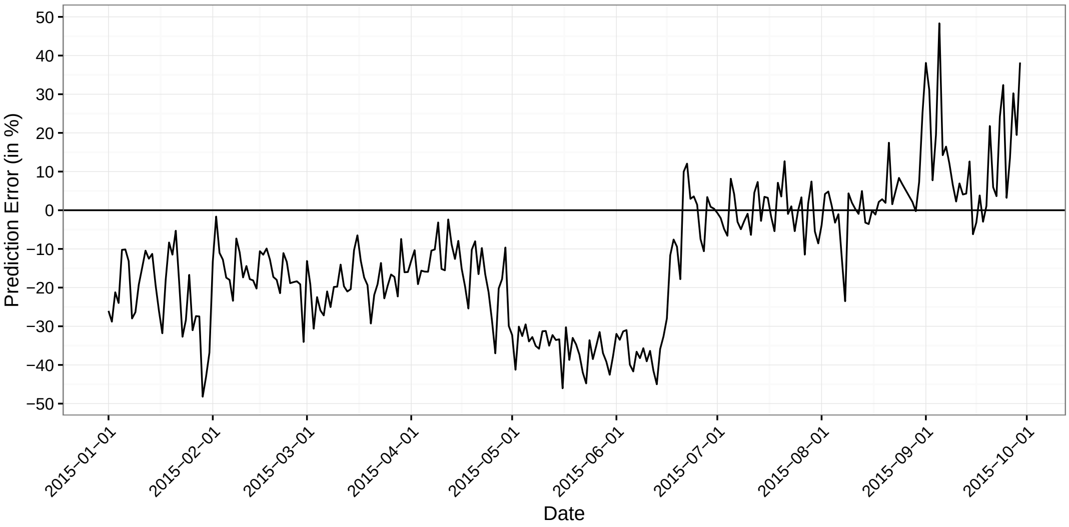


Figure 1.2: Prediction error Web1on1 Jan - Sep 2015

In order to be able to continue the growth of Web1on1, processes have to become more efficient and standardized. The processes have to be scalable. Currently, it takes the staffing employee at Web1on1 one day to come up with a schedule. A large part of this day is spent on analysing graphs and making the staffing decisions. If this process does not become more efficient and the centre continues growing, the centre is forced to hire more staffing employees. Making the process more efficient by supporting the employees with a decision support tool, contributes to the scalability of the process. Improving the performance is one aspect, but standardizing the staffing process itself is also important for Web1on1 to be able to expand their business.

Based on the findings elaborated, the following problem statement is defined:

The workload forecasting and the staffing of chat agents at the CSC centre are not effective and inefficient. This has a negative impact on the performance and growth potential of Web1on1. First, due to the inaccurate prediction of the workload, there is a mismatch between the number of agents staffed and the number of agents required. This results in under- or overstaffed periods which leads to insufficient performance or an increase of costs. The other problem is the inefficient and non-standardized staffing procedure. This makes it hard for Web1on1 to scale their operations and this threatens their growth potential.

1.3 Research questions

The previous section has outlined the challenge at Web1on1. This section will be used to explain what the objective of the thesis is by outlining the research questions. Chapter 7 summarizes the findings of the project and answers those questions.

The first research question and its sub question, is aimed at extracting information from the data available. Based on these insights, decisions about the design of the model can be made. The second research question and its sub question, focus on the first part of the staffing process, namely the workload prediction. The last research question and its sub questions, cover the capacity determination and the production of the staffing levels. Those questions define the models required to staff the chat agents efficiently.

1. What insights does the chat data provide and which of these can improve the staffing performance of Web1on1?
   (a) What are the most important factors that have an influence on the arrival of chats and how are these related?
2. What forecasting model predicts the arrival of workload to the system accurate enough to generate valuable input for the capacity model?
   (a) What are the frequencies on which the forecasts and staffing decisions can be made?
3. How should Web1on1 staff their chat agents, such that the service level delivered is at least as agreed upon with the clients, while minimizing the total labour costs?
   (a) What queueing model fits the system the best?
   (b) What are the constraints that limit the solution space and define the staffing model?

1.4 Research design

The objective of this thesis is to develop a model which serves as a decision support tool for staffing employees of Web1on1. This model is called the workforce staffing model. It will consist of three mathematical models which each cover a part of the staffing process. This section is used to explain what the modelling approach is and how the performance of the models will be evaluated.

1.4.1 Modelling approach

Before describing each model in detail, the general approach is explained. The first step in the Workforce Management (WFM) process is the prediction of the workload to the centre for each cluster. The workload is the amount of work that arrives per time unit. When there are five chat arrivals which take 12 minutes each to be served, then the workload is equal to 1 hour. In queueing theory, this amount is also referred to as 1 Erlang, named after the founding father of queueing theory A.K. Erlang. When eight chats arrive and each of these chat requires 15 minutes of service, the workload is 2 hours or 2 Erlang. Hence, there are two values which determine the workload to the CSC centre: the number of chat arrivals and the expected service time for each chat. The expected service time is determined from historic data, but the expected number of arrivals is more difficult to predict. In order to make reliable predictions, the first part of the workforce staffing model is used to predict the number of chat arrivals and calculate the expected workload. The staffing is done per cluster per hour and therefore the workload is forecasted for each cluster and hour. The forecasting model predicts the total arrivals to the CSC centre and hereafter those aggregated arrivals are allocated to the clusters. The forecasting part of the model is described in detail in Chapter 4.

After the workload to the system is predicted by the forecasting model, the agent capacity required to process this work and achieve a SL of 90% up-time, has to be determined. Contact centres like call centres, are often analysed using queueing theory. Given the system characteristics of the centre, the expected workload and the required SL to be met, the agent capacity can be obtained. Therefore, the second model translates the workload into agent capacity by using a queueing model. A detailed description of this model is given in chapter 5. The queueing lay-out at Web1on1 is shown in Figure 1.3.

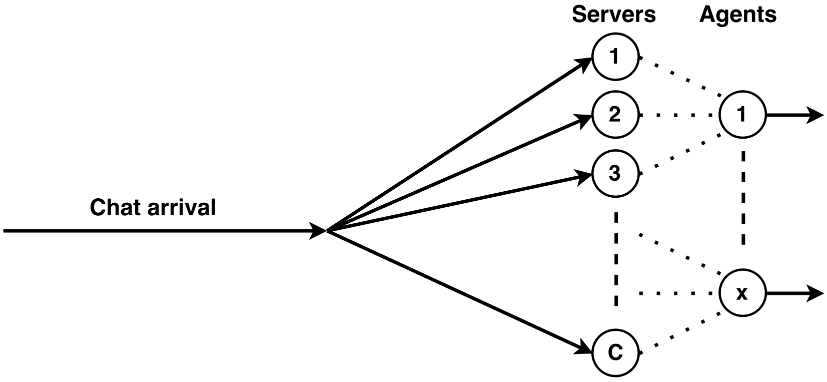


Figure 1.3: Queueing system at Web1on1

The final model is used to optimize the agent capacity over the different clusters. Before addressing the objective of this third model, it is important to understand the difference between servers and agents. In a call centre, one call agent can serve one customer at the same time and therefore those agents correspond to one server. In a CSC centre, a chat agent is able to serve multiple customers simultaneously. Hence, one agent represents multiple servers. It is essential to determine how many servers correspond to one chat agent of Web1on1. To do this, the departure rate of a chat agent is used. In the case of Web1on1, one agent is able to serve a maximum of 15 customers per hour and the expected service time is approximately 12 minutes. With a service time of 12 minutes, one independent server can serve up to 5 customers per hour. Therefore, one agent corresponds to three independent servers.

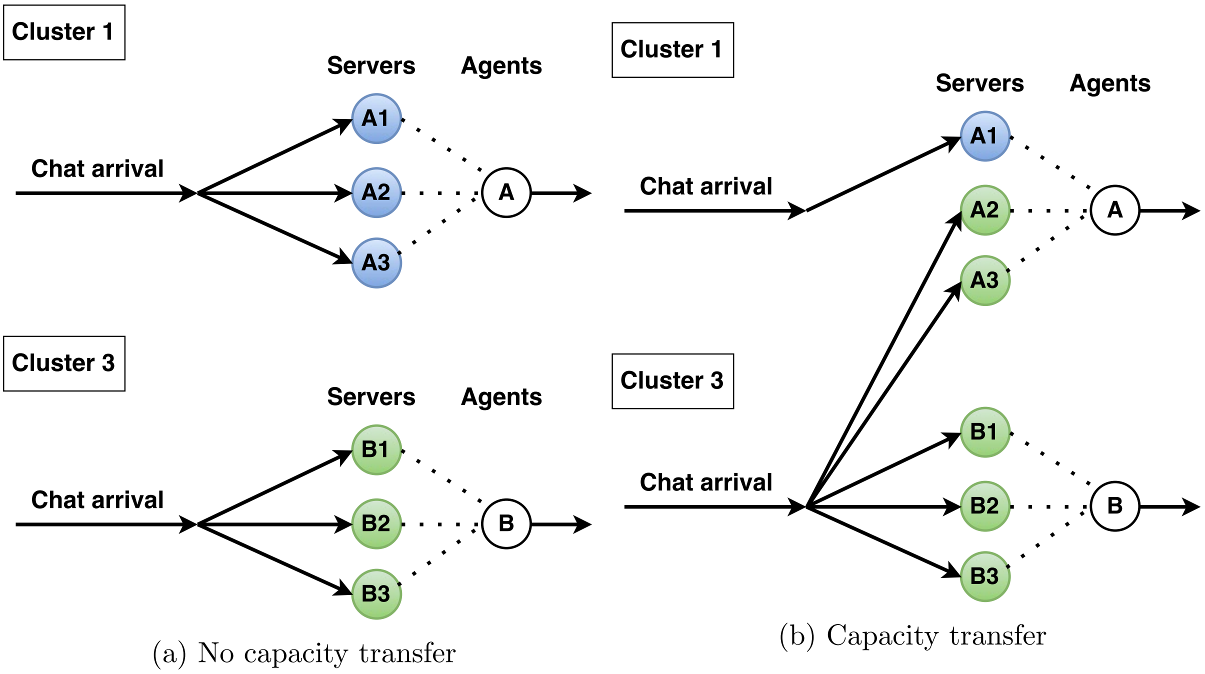
The output of the capacity model is the required number of servers to be staffed, given the workload expected and the SL demanded. It is only possible to staff one entire agent and because this agent corresponds to three independent servers, the amount of servers to staff per cluster is limited to a multiple of three. Although, the optimal amount of servers may be equal to 7 servers which is 2.33 agents. To prevent the sub-optimal staffing levels of either 2 or 3 agents, it is possible to transfer server capacity from one cluster to another. The 'transferred' servers move excess capacity from one cluster to a cluster with an under capacity. The third model determines the optimal number of agents to staff per cluster and the amount of servers that have to be transferred.

To illustrate how transferring server capacity can contribute to optimize the staffing at CSC centres, an example is given. In the first scenario, the optimal number of servers for cluster 1, 2 and 3 are 3, 6 and 3 respectively. This corresponds to 1, 2 and 1 agent of type 1, 2 and 3 respectively. It is possible to achieve an optimal solution without transferring any server capacity. For the second scenario, the optimal number of servers for cluster 1, 2 and 3 are 1, 3 and 5 respectively. In this case, the amount of agents is 0.33, 1 and 1.67 respectively. Of course, only entire agents can be staffed and therefore it is not possible to reach this solution without transfers. Staffing 1, 1 and 1 agent of type 1, 2 and 3 respectively, leads to 3 servers for each cluster. At cluster 1, there is an overcapacity of two servers, while cluster 3 requires two servers for the optimal solution. If two of the servers of agent type 1 are used to serve customers in cluster 3, the optimal staffing levels can be achieved. Table 1.2 shows both scenarios and Figure 1.4 illustrates the transfer of servers from cluster 1 to cluster 3 in scenario II.

Table 1.2: Example agent staffing optimization

Scenario	**I**			**II**		
Clusters	1	2	3	1	2	3
Servers required	3	6	3	1	3	5
Agents required	1	2	1	0.33	1	1.67
Agents staffed	1	2	1	1	1	1
Servers transferred	0	0	0	-2	0	+2
Servers staffed	3	6	3	1	3	5

The third model determines how many agents there have to be staffed per cluster and how many servers there have to be transferred. This will result in an optimal distribution of the agent capacity over the clusters. A mathematical program model is developed to



(a) No capacity transfer

(b) Capacity transfer

Figure 1.4: Illustration of capacity transfer

make these decisions. This model is discussed in detail in chapter 6. The three models - forecasting, capacity and staffing - form the workforce staffing model. A graphical representation of the modelling approach is given in Figure 1.5.

1.4.2 Performance determination

The objective of this thesis is to develop a model which staffs the agents such that the performance of the CSC centre is improved, while minimizing the labour costs. The forecasting model is developed to improve the prediction accuracy while the queueing and staffing model are used to determine the staffing capacity and the optimal capacity distribution over the clusters. To determine the effect of the model on the performance of the CSC centre, the use of the models is simulated and will be compared to the performance realised by Web1on1 for the period January 2015 until September 2015. There are two measures evaluated: the SL and the total amount of agent hours staffed to reach this level. The costs are directly related to the number of agents staffed. Benchmarking the model's performance with the historic performance of the CSC centre, indicates whether the model improves the staffing decisions.

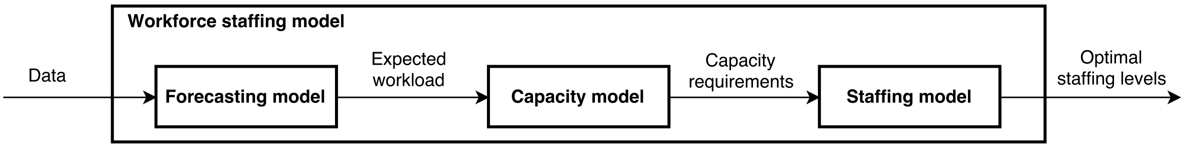


Figure 1.5: Modelling approach

Next to a possible better and more effective staffing of agents, the model will also add value to the standardization and scalability of the staffing process. The model will add value by the objectification of staffing decisions. In addition, the frequency at which decisions can be made, can be increased because a larger amount of data can be analysed in a short time period. How much value this adds for the centre is hard to determine and will not be part of the evaluation.

2 Literature review

The purpose of this chapter is to introduce the existing literature on Workforce Management (WFM) in CSC centres and related contact centres. First, a general introduction in WFM is given and hereafter the application to CSC centres is introduced. Section 2.3 is dedicated to the demand modelling phase of the WFM process. The final section of this chapter addresses to the optimization of hierarchical workforce planning in CSC centres.

The goal of WFM is to allocate personnel over time to satisfy arriving demand, while taking constraints into account. The output of the process is a timetable in which individual employees are assigned to working shifts. The first step of the WFM process (Figure 2.1), is modelling the demand that arrives at the centre. Predicting the arrival rate of customers and calculating what the required number of agents is, is widely studied for call centres. To some extend, those models can be applied to CSC centres. For the determination of the number of agents required, queueing theory is often used. The standard and widely used queueing model in call centres, is the Erlang-C model ($M/M/c$). The Erlang-A model ($M/M/c+M$) is an extended version of the Erlang-C model and it includes customer abandonment, as discussed by Mandelbaum and Zeltyn (2007). The third step of the WFM process, is scheduling shifts based on the expected demand. Many articles have been written about scheduling call centres. Most of them use Mathematical Programming (MP) models and some extend this by simulation.

There are only a few articles addressing CSC centres. There is expected that the speed at which these centres emerge, will accelerate in the near future, mainly because of the change in communication between individuals. Only Tezcan and Zhang (2014) addressed the staffing problem in CSC centres. Models developed for the application to call centres, can be useful as a starting point, but the system is slightly different. This has an impact on how to organize these centres efficiently. The main distinction between chat and call centres, besides the way of communicating, is the that chat agents can serve multiple customers simultaneously. Consequently, the productivity of an agent will be different. What effect this has on the performance of the contact centre and how these centres can be optimized, are questions that have not been answered and require additional research.

In short, many models have been developed to solve the staffing problem in call centres, but the emergence of CSC centres requires research to adapt these models and develop new ones for this specific type of contact centre. Almost no papers have addressed this application specifically, while there is expected that the demand to operate these centres efficiently, will increase in the upcoming years.

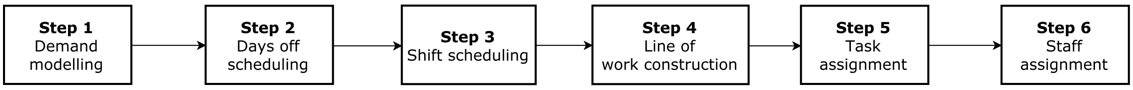


Figure 2.1: WFM process (Ernst et al., 2004)

2.1 Origin of workforce scheduling

In the early 1950s the first occurrence of workforce scheduling appeared. Edie (1954) published a paper about the scheduling of police officers at toll booths. The main objective was to investigate whether the staffing of police personnel at toll booths was sound and economical. Prior to the article, the schedules were created by police officers using their experience and gut feeling. Edie (1954) described in detail how data was collected and insight in the dynamics of the process at toll booths was gained. The behaviour of traffic arrivals, delays, service times, and traffic backups are the core findings of the analysis. After the analysis, the so called waiting-line problem was optimized and finally a method was proposed to deal with the scheduling problem. Dantzig (1954) showed how a Linear Programming model (LP) can help solve the problem introduced by Edie (1954). It focused solely on the scheduling problem and denies the first step; the arrival of traffic at the booths. Those two papers, Edie (1954) and Dantzig (1954), are seen as the origin of personnel planning and scheduling science.

Nowadays, the problems, methods and resources are very different from the one addressed by Edie (1954). Ernst, Jiang, Krishnamoorthy, and Sier (2004) stated that the process of WFM consists of six stages: demand modelling, days-off scheduling, shift scheduling, line of work construction, task assignment, and staff assignment. The entire process is called the WFM process and is shown in Figure 2.1.

The (expected) distribution of workload over time is essential information for the rest of the process. There are problems in which the workload is known or relatively easy to predict, for example when scheduling airline crews, factory workers, or nurses in a hospital. In other environments, like call centres, the workload is stochastic. This means that it is unknown when a call will arrive and how long it takes to process it. This information is essential in order to staff and schedule the agents effective and efficiently. The cost of this personnel is a large part of the operating budget (60-80% of the total operating budget), because of the labour-intensive character of the work (Aksin et al., 2007). Therefore call centres benefit from an accurate prediction of workload and an appropriate number of agents scheduled. Having too few agents working in a certain period (understaffed) results in an under-performance, too many (overstaffed) will unnecessarily increase costs.

2.2 Emergence of a new type of contact centre

In the recent decade, the communication tools used in society have developed rapidly. The emergence of mobile, online communication at the end of the last decade, has changed the communication between individuals tremendously. The speed of communication has increased and the classical channels, phone and email, have been accompanied by a third one: (online) messaging. Although communication between companies and customers are still mainly via the classical channels, CSC centres anticipating on this trend have emerged in the recent years. This development has an impact on how contact centres of companies are organized. There is expected that this trend will accelerate, or at least continue, in the near future.

Only a few papers have been written about this new type of contact centre. The more centres arise, the more important it becomes to operate them efficiently. The largest efficiency gains can be achieved at the largest part of the operating costs. Therefore, it is important to schedule chat agents as good as possible. In the literature, Tezcan and Zhang (2014) is the only paper addressed to the staffing problem in a CSC centre. They described how to route arriving chats to an agent such that those employees operate on their optimal level.

Tezcan and Zhang (2014) defined a CSC centre as a service where customers contact agents via an IM application over the internet. The speed of IM lies somewhere between email and a phone call. It demands a faster response than an email, but slower than a phone call. The model developed by Tezcan and Zhang (2014) is called the process sharing model. In this model there is assumed that agents handle multiple customers simultaneously. Hence, customers are served by a fraction of an agent's service capacity. The amount of service a customer needs, is stochastic. Customers' patience is limited. If the waiting time exceeds a certain amount, the customer abandons the queue and leaves the system. The service time is decreasing when the number of customers who share the agent's capacity is increasing, due to the increase of efficiency of the chat agent. Setting the service rate at a certain level and knowing the patience of customers, the rate at which customers leave the system is identified.

2.3 Demand modelling

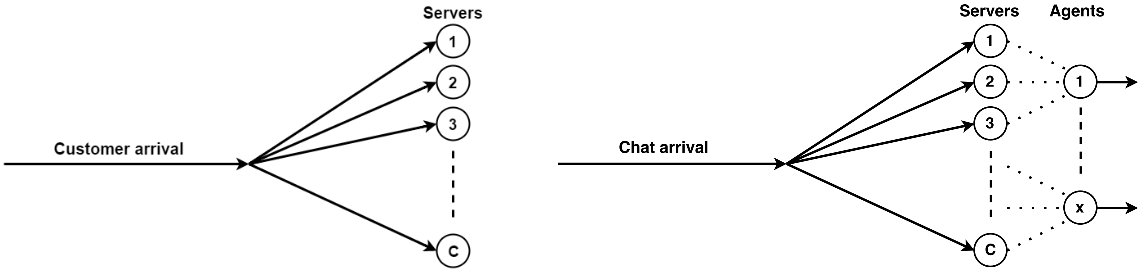
As argued in the beginning of this chapter, this section elaborates models that might be applicable to the first step of the WFM process: demand modelling. This step consists of two parts or sub-models which results in the number of servers required. First, a model to predict the number of chats arriving at the system. A second model to determine the number of servers to meet the SL requirement. Hence, in the WFM process, both models are captured in the first step, but in the workforce staffing model those are the first two models.

As stated before, forecasting how many customers will arrive at the CSC centre, is the starting point for the staffing problem. In the literature, no papers have been written about forecasting arrival rates at CSC centres. The models developed to predict the workload to other types of contact centres, are useful. In general, a call centre is similar to a CSC centre. Customers arrive to the centre and a certain service time is required to process the request. Therefore, it is reasonable to forecast the arrivals at a CSC centre, in the same way as arrivals at a call centre. Many models have been developed to model the arrival rate of customers at a call centre. Time series analysis, multivariate distributions or regression models are often used to predict the arrival rate. The modelling of arrivals is important to operate the centre effective and efficiently. The assumptions of the scheduling models regarding the arrival rate do not always hold, as shown by Jongbloed and Koole (2001). Therefore, it is important to keep an eye on the main objective when predicting arrival rates. Brown et al. (2005) and Aldor-Noiman, Feigin, and Mandelbaum (2009) are two papers that developed a model for the first step of the WFM process, but also payed attention to the scheduling part of the process. Avramidis, Deslauriers, and L'Ecuyer (2004); Taylor (2008) and Shen and Huang (2008) solely focused on predicting the workload in a call centre, without addressing the scheduling problem.

The next step in the demand modelling phase, is to translate the number of arrivals into the capacity requirements; the number of servers required. The number of agents is equal to the amount of servers in the case of a call centre. In the literature, call centre systems are often modelled as queueing systems in order to define the number of agents corresponding to a certain SL. Due to the similarities of the CSC centre and call centres, the CSC centre can be modelled as a queueing model. The purpose of this model is to define the relationships between the number of servers, SL and workload. This can be used to calculate the required staffing levels of the centre.

In the call centre industry, the most widely used queueing model is the $M/M/c/\infty$ (shortly $M/M/c$ or Erlang-C) model. The arrival rate follows a Poisson distribution and the service times are exponentially distributed. The SL of the system is measured by the long-term fraction of calls answered within a certain time, for example 90% of the calls answered within 45 seconds. The call centre has to staff sufficient agents such that the SL is met.

The Erlang-C model assumes that customers are very patient and willing to wait an unlimited time. Many papers addressing call centres extend the problem with the possibility of customer abandonment (Erlang A), see e.g., Mandelbaum and Zeltyn (2007); Feldman, Mandelbaum, Massey, and Whitt (2008); Kim and Ha (2012); Defraeye and Van Nieuwenhuyse (2013). The Erlang-B model ($M/M/c/c$), has no waiting queue. Customers which arrive when all servers are busy, are blocked. At the CSC centre Web1on1, there is also no waiting queue. When there are no servers available, the chat widget at websites disappears. Hence, arriving customers get blocked when all servers are occupied. Whether Web1on1 has all the requirements to be modelled as an Erlang-B model, is determined in chapter 3. Based on the outline of the system, the Erlang-B model seems to be the most appropriate queueing model for the CSC centre.



(a) Erlang-B model

(b) CSC centre system outline

Figure 2.2: Queueing models

Another approach often used by call centres instead of a queueing model, is the so called Square-Root Staffing rule (SRS). The objective for most contact centres, is to reduce the mean waiting time of customers while having a high utilization of agents. This is referred to as the 'Quality and Efficiency Driven' regime by Gans, Koole, and Mandelbaum (2003). Related to this, the SRS rule (equation 2.1) can be used to determine the required capacity. Halfin and Whitt (1981) formalized this rule for the Erlang-C model, after first being observed by Erlang in 1948 (Aksin et al., 2007). Jongbloed and Koole (2001), Aldor-Noiman et al. (2009) and Feldman et al. (2008) have demonstrated how this rule can be applied to real life data. It is a relatively simple method. This makes it understandable for the managers at contact centres and therefore very suitable to be implemented. Another benefit of this method is that it does not assume any distribution for the arrival rate or service times. The formula uses the workload (R) that arrives to the system and a decision parameter β. β is the parameter for the Quality of Service (QoS) of the system. The higher the value for β, the more weight there is put on the quality aspect of the system. Hence, the number of agents increases. The value of β is chosen by the CSC centre manager.

$$N \approx R + \beta\sqrt{R}, \qquad -\infty < \beta < \infty; \tag{2.1}$$

where N is the number of servers, R is the workload to the system (in Erlangs) and β is the service-grade parameter.

2.4 Workforce optimization in CSC centres

In call centres, each server is operated by one agent. In the case of the CSC centre Web1on1, one agent corresponds to three independent servers. After the required number of servers has been determined by the queueing model or SRS rule, the amount of agents to be staffed for the CSC centre is not obvious. Before moving on to the next step in the WFM process, the day-off staffing or scheduling step, an additional optimization step can be added. Only one agent can be staffed and therefore, the number of servers staffed per cluster is a multiple of three. Due to the hierarchical structure of the workforce of Web1on1, the server capacity can be re-distributed and this may result in a more optimal capacity planning for the CSC centre. This additional optimization step is called the capacity optimization step and is covered by the third model: the staffing model.

As introduced in chapter 1, there are multiple agent types at Web1on1. Chats arriving at client X require more specific knowledge than chats of client Y. Therefore, Web1on1 has clustered their clients based on the experience and knowledge an agent should have in order to be able to process the chats. Agents who have specific cluster knowledge and more experience, are also able to handle chats from clusters which do not require any specific knowledge or experience. Each type of agent has one or more cluster for which he or she is allowed to chat. Certain agent types can chat for multiple clusters, as shown in Table 1.1. Because some types of agent can take over the work of others, the workforce has a hierarchical structure.

One agent corresponds to three servers and each of these can be available for another cluster as long as it is part of the skill set of that agent. In other words, the capacity of one agent can be distributed over multiple clusters. Hence, the excess server capacity at one cluster can be transferred to another to balance the capacity over the clusters and accomplish the required SL without staffing additional agents. In chapter 1, an example is given to illustrate the transfer of servers.

In the current literature, there are only a few articles dedicated to live chat and CSC centres e.g., Shae et al. (2007); Tezcan and Zhang (2014). None of these addresses capacity optimization in a CSC centre with a hierarchical workforce. Articles addressing the capacity determination and planning in other contact centres are used as references. As explained in section 2.3, queueing theory can be used to translate the workload to the CSC system to the number of servers required to meet the SL, see e.g., Jongbloed and Koole (2001); Brown et al. (2005); Kim and Ha (2012); Tezcan and Zhang (2014). Although this thesis does not address the scheduling and rostering part of the WFM process, the methods used to determine optimal solutions, are suitable to solve the capacity optimization problem. Often MP models are used. Atlason, Epelman, and Henderson (2004); Bhulai, Koole, and Pot (2008); Cezik and L'Ecuyer (2008) and Avramidis, Chan, Gendreau, L'ecuyer, and Pisacane (2010) combine LP models with simulation. Queueing theory and algorithms are used by Feldman et al. (2008); Gurvich, Luedtke, and Tezcan (2010); Defraeye and Van Nieuwenhuyse (2013). An example of a more sophisticated model, a Cutting Plane method (CP), can be found in the paper of Avramidis et al. (2010). The CSC centre Web1on1 is an example of a multi-skilled, hierarchical contact centre. Multi-skilled call centres are addressed in the literature and often MP and simulation methods are combined to find a solution, see e.g., Atlason et al. (2004); Shumsky (2004); Bhulai et al. (2008); Cezik and L'Ecuyer (2008); Gurvich et al. (2010). The third models which optimizes the server capacity over the different clusters, is a MP model. Chapter 6 thesis gives a detailed description of the Mixed Integer Non-Linear Programming model (MINLP) model developed to determine the optimal distribution of the agents and servers at Web1on1.

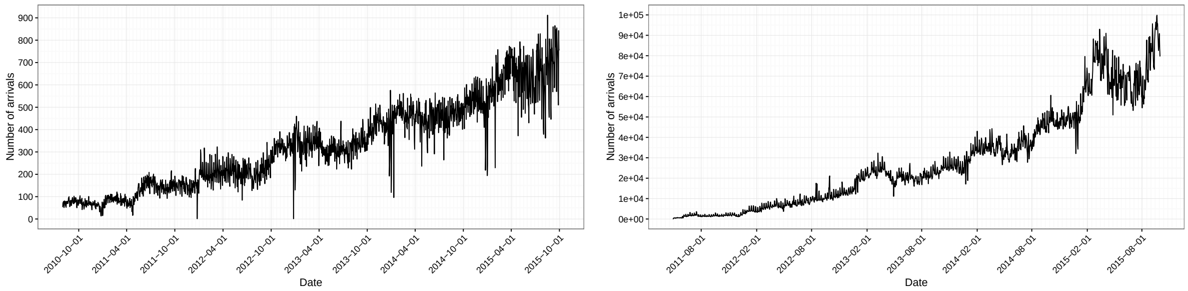
3 Centre analysis

This chapter discusses what the data has revealed about the operations at Web1on1. The data preparation and analysis conducted for this project are addressed. The raw data is not directly suitable to be used by the model. Before preparing the data and replacing the missing values in section 3.2, section 3.1 describes what raw data is available. Section 3.3 will then describe what information the data reveals. This information is used to make decisions about the design of the models.

3.1 Data description

There are multiple data sets which contain operational data from the CSC centre Web1on1. The data used in this thesis, is a merger of 4 data tables. The first set entails all the details about the chats that have been processed from August 2010 through September 2015. Next to that, there is a table which stores all the information about website traffic. Only the traffic on the web pages of clients at which the chat widget is active, is taken into account. The traffic at other parts of the website are not considered, because it is only possible to start a chat on a page where the chat widget is active. The remaining two data sets store the historic capacity planning and up-time of the system. The up-time is the percentage of time at least one server was available. This corresponds to 1 minus the blocking probability (BP) and is the realised SL of the centre.

Before the data can be analysed, the required information of each data set, is extracted and combined into a main data set. This contains all the information necessary to do the analyses and the modelling. The result is a data set containing 29,400 rows and 9 columns. The data ranges from 31 October 2013 until 30 September 2015. This is a different range than the available data from chats, because not all information has been gathered since the start of the CSC centre. Especially, the up-time of the system has only been measured since October 2013. For an illustration of the constructed data set, see Table A in Appendix A. Figure 3.1 shows the time series of the chat arrivals and website visitors.



(a) Chat arrivals

(b) Website visitors

Figure 3.1: Raw data Web1on1 from 2010 - 2015

3.2 Missing data

In order to explore and analyse the data on relationships over time, all time periods have to be included in the data set. Unfortunately, there is missing data for some periods. The data is completed by replacing the missing data with a best guess.

There are three types of the missing data:

1. No chat data available when CSC centre was closed;
2. Chat data missing of 9:00h - 10:00h for several dates;
3. Up-time of the system for a cluster.

If there is missing data on the chat data, the arrival rate and service time for that period is unknown. To come up with a reliable estimate, the data of website traffic is used. Fortunately, there is no missing data regarding the website traffic. Also the conversion rate from website traffic to chat arrivals, is known. This rate becomes relatively stable as the traffic increases. Combining the traffic and (average) conversion rate gives a fair guess for the missing arrival rates. Although the conversion rate converged to a constant level, there are some differences in the conversion rates for some time periods. To take this into account, the average conversion rate for each month/year combination has been used. The result of combining the website traffic and conversion rate is rounded to the nearest integer.

The other chat data missing, the average chat duration, is replaced by the last observation of the similar period and cluster. For example, if the chat duration for a Monday 17:00 in cluster 1 is missing, the average chat duration of the last Monday 17:00 in cluster 1 is used.

The third type of missing data, is the up-time of the system. Not for every hour and cluster the up-time has been stored. The missing values are replaced with the average up-time of the other cluster(s) in that same period.

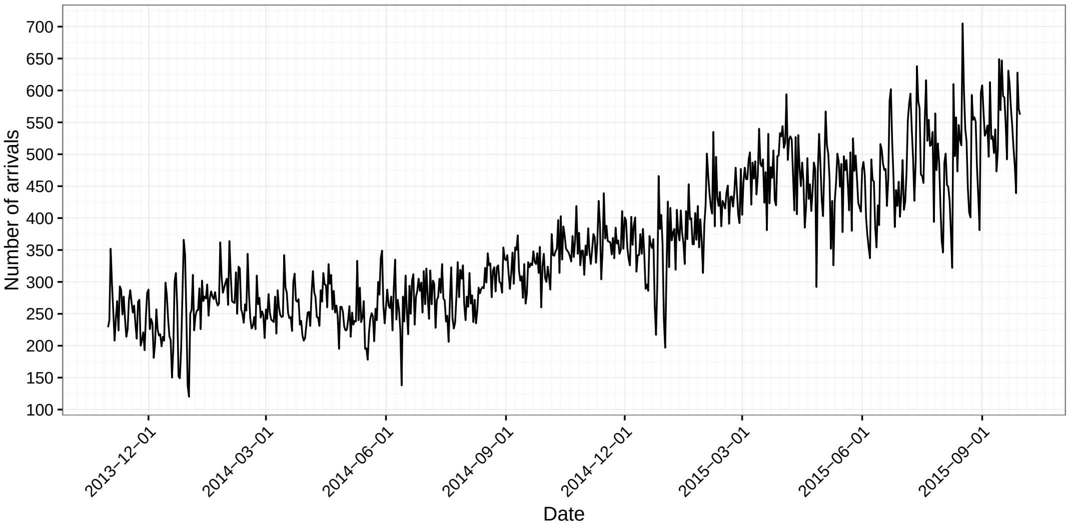


Figure 3.2: Time series of chat arrivals per day after data preparation

3.3 Data exploration

The first step of the data analysis, is a graphical exploration to indicate what information the data might entail. This can be useful to decide what models to use and what variables might have an effect on the arrival rate of chats. Two aspects that have become clear are the distribution of arrivals over the different clusters and the probability distributions of the arrival rate and service times. Next to this, the data reveals that those distributions vary over time. Those observations will be discussed in this section.

The prepared time series of the arrival rates and the website traffic are represented in Figure 3.2 and Figure 3.3, respectively. As expected, both series show a similar behaviour. The upward sloping trend of chats is mainly caused by the increase of website traffic. By increasing the number of clients, the total website traffic of the CSC centre increases. Another important aspects are shifts in the trend line. Not all clients have the same amount of website traffic. When a new client with a lot of website traffic is added to the centre, a shift appears in both the traffic and chat arrivals. Therefore, the size of all clients together, quantified as the total website traffic of the Web1on1, is an important indicator for the amount of chats. This so called ‘Centre Weight’ is the first variable to be captured in the forecasting model.

A closer look at Figure 3.2 reveals that there might be a long term seasonality (yearly) present in the data. The amount of chats arriving in the summer is less than during other periods of the year. It is hard to determine, by only looking at the time series in Figure 3.2, whether there is a significant monthly effect in the arrival rates of chats. Therefore, in Chapter 4, a more detailed analysis is performed to find out whether this effect is significant enough to be taken into account by the forecasting model.

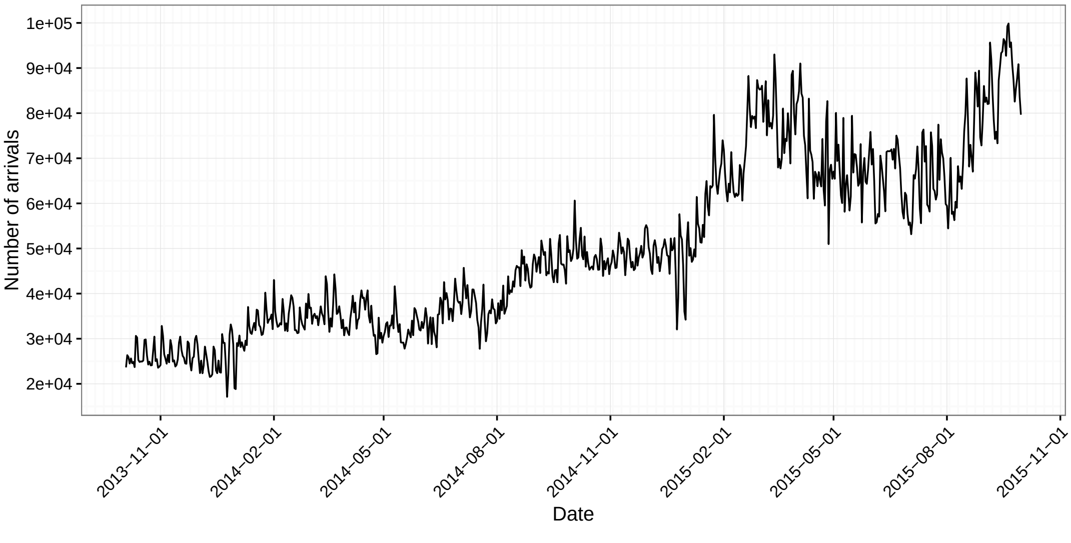


Figure 3.3: Time series of website visitors per day after data preparation

Figure 3.4 shows the hourly chat arrivals for a certain week in 2014. The short term seasonality (weekly and daily) is easier to detect graphically than its long term counterpart. The hour of the day and day of the week, clearly have an effect on the arrivals to the CSC centre. On weekdays, the hours with the highest load to the system are between 17:00h and 23:00h. In the weekend, most chats arrive in the morning and throughout the day. The total number of chats arriving on an entire day, is decreasing throughout the week.

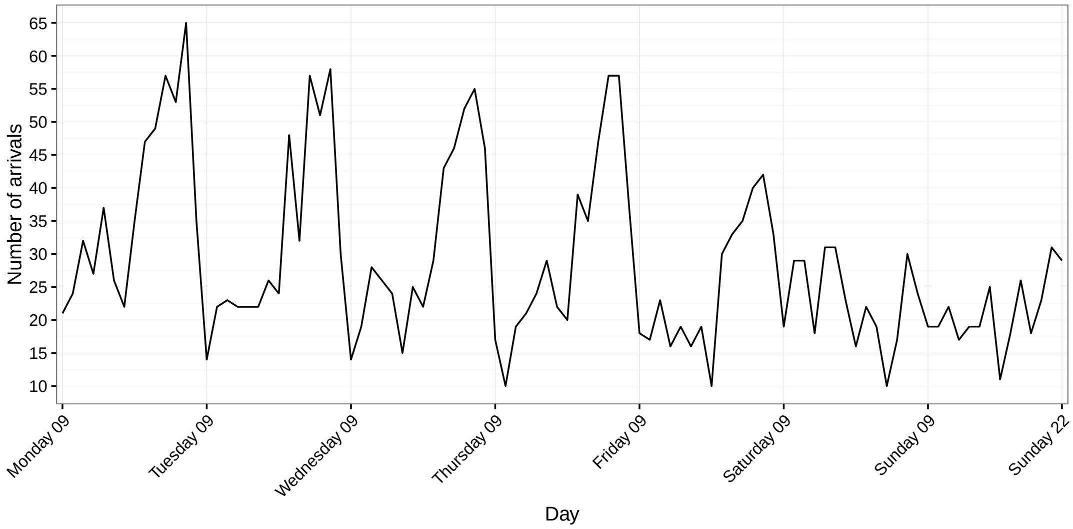


Figure 3.4: Time series of chat arrivals of one week

Another important characteristic is visualized in Figure 3.5. As mentioned, the arrivals over the day and week are not equally distributed. Most of the chats arrive during the evening of working days. But the distributions have not been constant over time. This is caused by the changing group of clients. Apparently, the more recent the data, the larger the CSC centre and the more evenly the arrivals are distributed over the day and week. As will become clear in the remainder of this section, this is not the only aspect which has evolved over time.

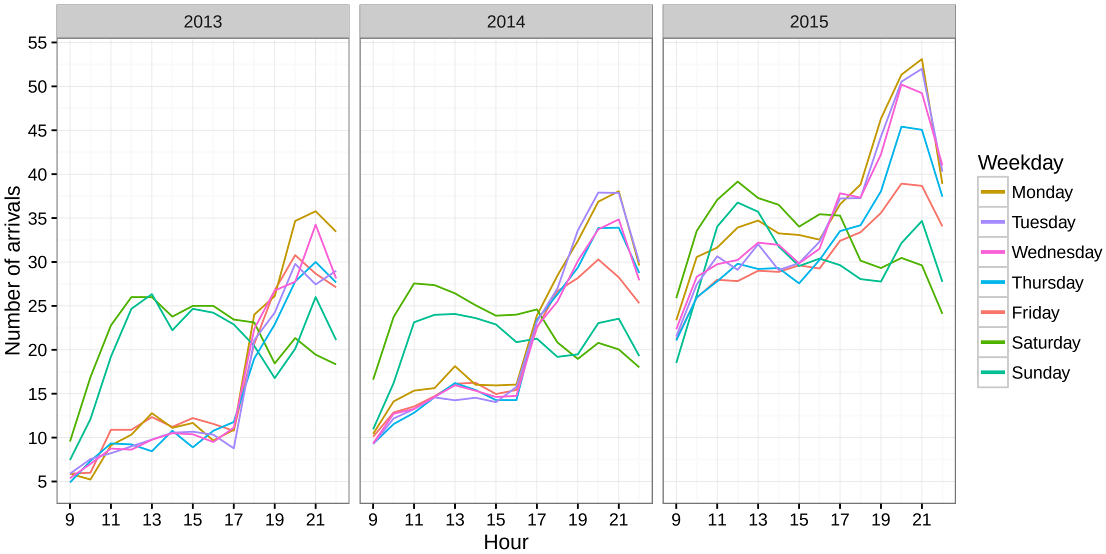


Figure 3.5: Average number of arrivals per weekday per year

Public holidays seem to have an effect on the arrival rates of chats. There are some days throughout the year that show extreme drops in the total number of chats handled. Most of those days turn out to be public holidays. The holidays considered for the analysis are presented in Table A.2 in Appendix A. Figure 3.6 shows the average arrivals per hour for regular days and public holidays. Throughout the entire day, on average the arrival rate is lower than a regular day. In order to improve the accuracy of the model, a variable for public holidays might be beneficial. Whether the difference between holidays and regular days is significant, will be determined during the construction of the forecasting model in Chapter 4.

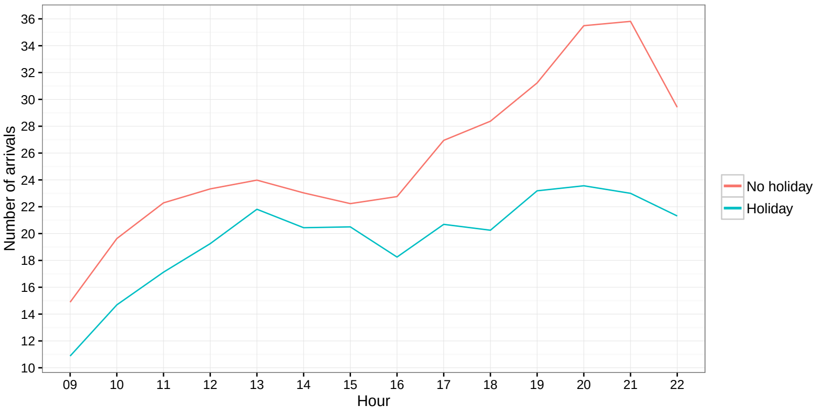


Figure 3.6: Average number of arrivals on public holidays

Based on the experience of the employees at the CSC centre, an external effect that might influence the amount of chats arriving, is weather. When the sun is shining, it is less likely that people are visiting websites. On a rainy day, the opposite is probably the case. For each day, the weather grade, based on four meteorological elements, is calculated. The higher the grade, the better the weather. The weather grade is described in more detail in chapter 4. Figure 3.7 shows the average arrival rate for each weather grade. Based on this graph, it cannot immediately be concluded whether the weather has a significant effect.

To have a better understanding of the impact of the weather on the arrivals, the weather grades are clustered. Bad weather grades (1 - 4) are aggregated in the group “Bad”, average weather grades (5 - 6) represent group “Average” and the group “Good” is a collection of the grades 7 - 10. This gives the opportunity to compare more extreme weather effects. The average arrivals per weather group are shown in Figure 3.8. On weekdays, most of the hours do not seem to be affected by the weather. Only on Mondays and Wednesday there may be some effect of bad weather. Weekends also may be affected. Whether these effects are significant, is determined in Chapter 4.

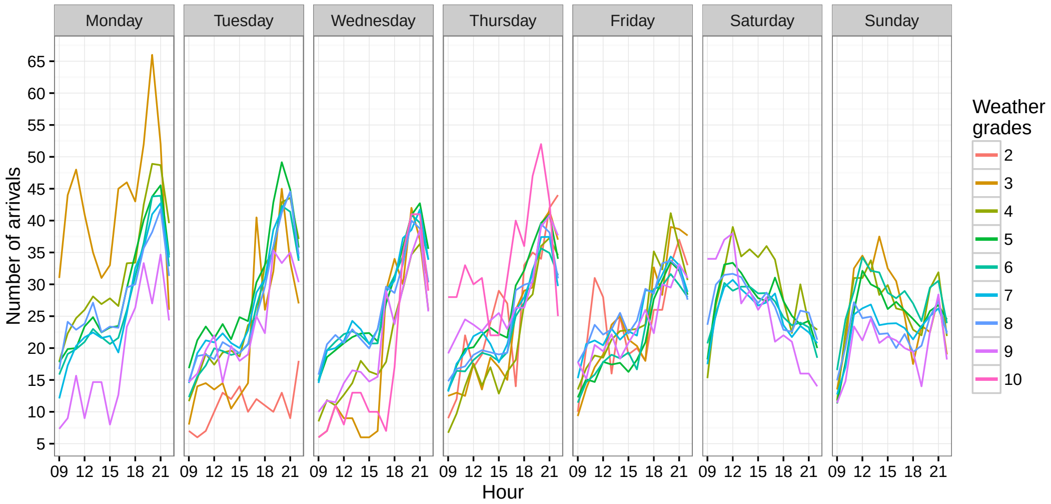


Figure 3.7: Average number of arrivals per weather grade per weekday

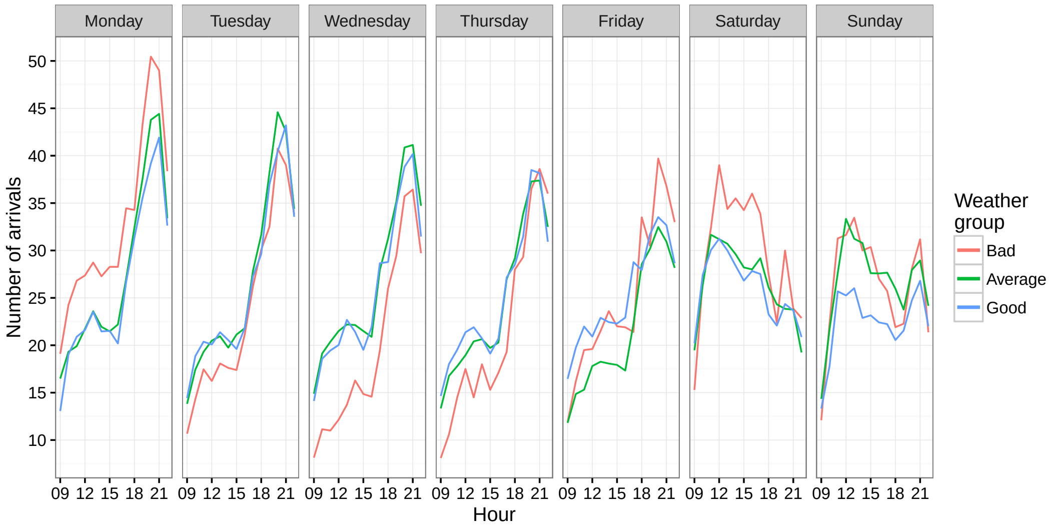


Figure 3.8: Average number of arrivals per weather group per weekday

The workforce schedule has to be determined per client cluster. As mentioned in Chapter 1, the clients of Web1on1 are grouped based on the skills and knowledge an agent requires to be able to serve customers of a client well. Therefore, the arrival rates have to be forecasted per cluster. As can be seen in Figure 3.9, the distribution is not the same for all clusters. The relative chat amount that belongs to each cluster is not constant over time, but fluctuates between 20% and 40% per cluster. When determining the cluster distribution, the allocation of arrivals to a cluster, it is important not to base this on old data. This data may contain outdated information. This would result in a misleading arrival rate per cluster. More recent observations are more relevant. Next to this, also the effect of the hour and day of the have to be taken into account when determining the cluster distribution.

As discussed before, an important insight the data provides, is that relationships have changed over time. Due to changing circumstances of the CSC centre, such as the amount and types of clients, the patterns observed in 2015 are not the same as in 2014 (Figure 3.5). Another example of this behaviour is presented in Figure 3.10. It shows how the

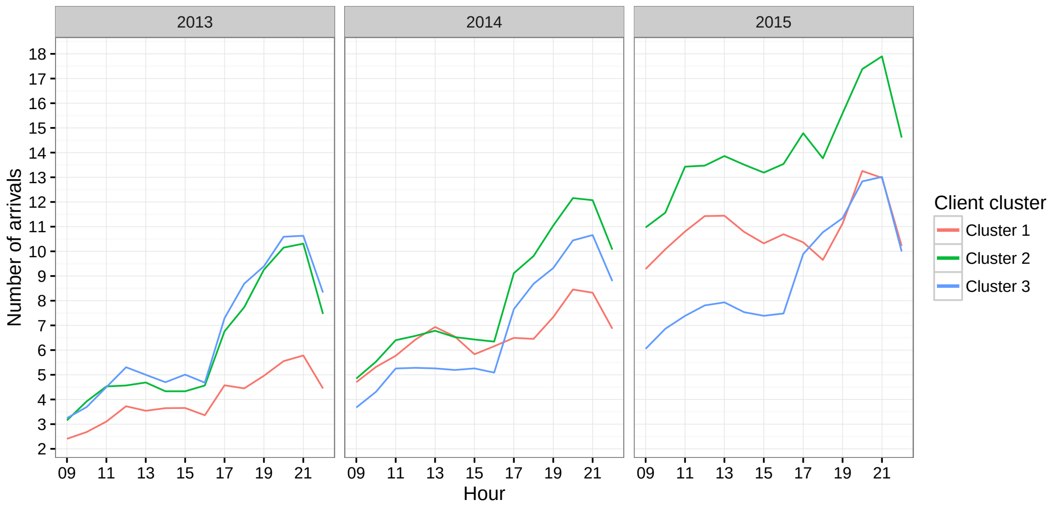
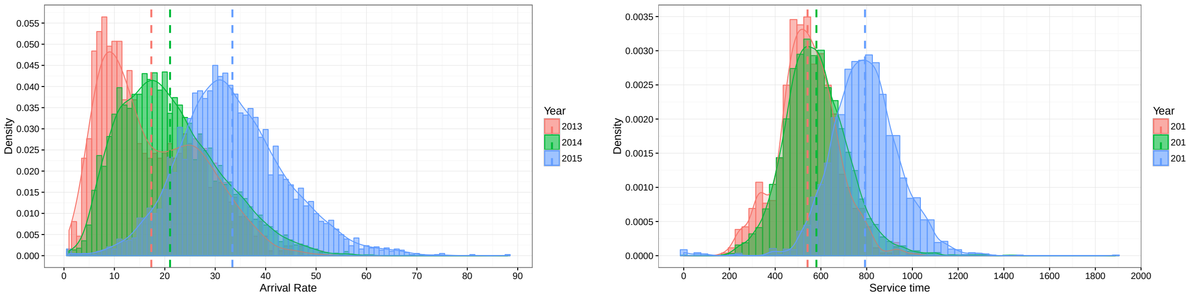


Figure 3.9: Average number of arrivals per cluster per year

distribution of number of chats arriving per hour has changed over the years. Both the mean and variance of the number of arrivals has increased. The same applies to the service times. It is important to take this behaviour into account when developing the model. Old data might be outdated and may have a negative influence on the performance of the model.



(a) Arrival rate

(b) Service time (sec)

Figure 3.10: Histogram of arrival rate and service time distribution per year

In order to determine the capacity requirements per cluster, the queueing system of the CSC centre has to be modelled. The distribution of the arrival rates and the service times are important in the definition of this queueing model. The characteristics and performance of the queue, depend on how the customers arrive and how they are served. The Erlang B ($M/M/c/c$), also known as the Erlang loss model, assumes that both the arrival rate and service time follow a Poisson distribution. One of the key characteristics of this distribution is that the mean is equal to the variance. Figure 3.10 shows the distributions of the arrival rates and service time of each year with the dashed line as the mean value. Given the means and variances in Table 3.1, it is clear that the arrival rate cannot follow this distribution, because the variance is much larger than the mean. Figure 3.10b shows the distribution of the service time. Again, these do not follow a Poisson distribution because the variance dominates the mean. Consequently, the queueing model used to determine the capacity requirements cannot be an Erlang B model. Which queueing model does fit the CSC centre the best, is described in Chapter

5.

Table 3.1: Mean and variances of arrival rate and service time per year

Year	Arrival rate		Service time (sec)	
	Mean	*Variance*	*Mean*	*Variance*
2013	17.3	96.5	541.6	16754
2014	21.0	96.5	580.3	19608
2015	33.4	112.9	792.8	23168

4 Forecasting model

The purpose of this chapter is to outline the forecasting model which is developed to predict the arrival of chats to the CSC centre. Section 4.1 starts with the description of the forecasting approach. Afterwards, the regression model and time series model are described in section 4.2 and 4.3, respectively. How these two models are used to make the aggregated forecasts, can be found in section 4.4. The last part of this chapter (4.5) is used to explain how the aggregated forecast is distributed to have the forecast per cluster.

4.1 Modelling approach

Before describing each part of the model in detail, the modelling approach and the reasons for making these decisions, are described. The forecasting model consists of two sequential models. First, a regression model to correct for long term trends, long term seasonality, and external effects. The residuals of this model are the input for a time series model to capture the short term seasonality and the relationships of arrivals over time. There are two reasons why there has been chosen for this approach.

1. The time series of the observed arrivals is not stationary and is therefore not suitable to be used by a time series model directly. There are mathematical solutions available to overcome this, like a logarithmic transformation or by differentiating the series. It is more convenient to correct the data by modelling the effects that cause the non-stationarity. This is the purpose of the regression model;
2. The effects modelled by the regression model, are the long term seasonality and external effects, while the time series model focuses on the short term relationships. Splitting the model in those two sub-models offers the opportunity to determine the amount of historic data to be taken into account, for each sub-model separately.

The second reason needs some additional explanation. The day of the week or hour of the day are examples of short term seasonality. The data shows that seasonality effects change over time. In 2015, the arrivals of chats were more equally distributed throughout the day than in 2014. The difference between the busiest and calmest hours has decreased. Also the distribution of arrivals throughout the week has not been the same. The reasons for these changes will not be elaborated on in this thesis.

It is likely that the distributions of arrivals over the week or day in the near future, are similar to the distributions observed in the recent past. Data from several months or years ago differ significantly and are not appropriate to base predictions upon. The effect of weekday and hour are not included in the regression model, because this model also includes long term effects and requires more historic data. This historic data is outdated for the weekly and daily seasonality and would bias the forecast. Therefore, those short term frequencies will be captured by the second part of the model, the time series model which only uses recent, relevant data.

4.2 Linear regression model

The first step in the forecasting model is to correct the data for the long term trend, long term seasonality, and external effects. Combining the information revealed by the data with the experience of the CSC centre employees, determines what causes the trend and trend shifts. There are yearly seasonality patterns and external variables which might have an effect on the arrivals of chats. In the remainder of this section, the independent variables for the regression model are determined and described.

4.2.1 Long term trend and trend shifts

The upward sloping trend is caused by the increase of clients for which the CSC centre chats. The more clients, the more website visitors and accordingly more potential chat arrivals. The impact of adding a new client to the CSC centre, depends on the size and type of client. A client with more website traffic will in general cause a large shift in the number of chat arrivals, while a small one will not be noticed at all. In the remainder of this thesis, the weight of a client or the centre is the quantification of the chat potential. The higher the potential, the higher the weighting. Hence, the more clients and the more website traffic those clients have, the higher the weight of the CSC centre, also called 'Centre Weight'. The aspects on which the weight of an individual client dependents, are listed below.

1. The profile of a client. There are two types of client: full coverage and evening-weekend coverage. Clients with a full coverage profile have the live chat available from 9:00 till 23:00, 7 days a week. For clients with the evening/weekend profile, the CSC centre only chats outside opening hours, which is from 17:00 till 23:00 on weekdays and 9:00 till 23:00 in the weekend.
2. The total amount of website traffic a client's website receives. Large companies have more website visitors than smaller ones and consequently more potential chat arrivals.
3. The type and number of website pages on which the live chat application is available. Those are the web pages from which a visitor can start a chat. If the chat widget is active on the homepage of a website, the amount of chats will be larger than if the widget is located on a more hidden part of the website. Also the number of pages at which the chat widget is available, has an impact on the chat potential of a client.
4. The total budget available to spend on live chat. Based on the traffic, many clients can do more chats than Web1on1 is allowed to do. The budget of clients on live chat is limited and therefore the amount of chats the CSC centre can execute, is limited as well.

All aspects mentioned above are covered in the quantification of the chat potential: the weight of a client. The accumulated weight of all clients is the weight for the CSC centre, the so called 'Centre Weight'. Based on the four aspects mentioned above, the sales department predicts what the weight will be for each client in the upcoming three months. This prediction is also made for new clients. Aggregating these predictions result in the Centre Weight for the upcoming months. It is assumed that the forecasts made by the sales department are accurate enough. This is a fair assumption because the centre 'targets' more or less on their own predictions. Mainly because of budget constraints, Web1on1 has to aim for a number of chats within a period for each client. When the number of arrivals deviates too much from the predetermined levels, Web1on1 can affect the conversion rate from website traffic to chats. By changing this rate the amount of chat arrivals can be influenced as well.

For historic data, it is possible to determine what the Centre Weight has been for each period. The Centre Weight is used to correct for the increasing trends and trend shifts. Therefore, the first variable of the regression model (x_1) is the Centre Weight.

The regression model including a constant and the Centre Weight as independent variable, is shown in Table B.1 in Appendix B. The coefficient for the variable Centre Weight is very small but positive, as expected. The variable is significant at the highest level. The adjusted R-squared of the model is 0.325, which means that 32.5% of the variance in the data is captured by the model with only the centre weight as an independent variable. The residuals of this model are depicted in Figure 4.1. The upward sloping trend and the shifts have disappeared. The growth of the CSC centre that was present in the data, is captured by the Centre Weight variable.

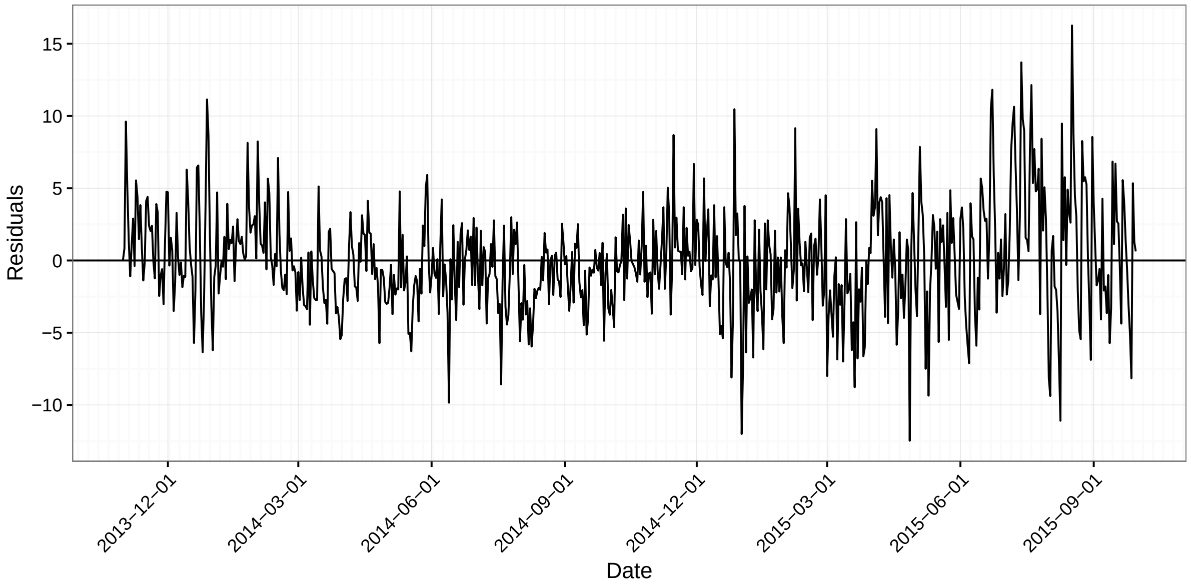


Figure 4.1: Residuals regression model with centre weight as dependent variable

4.2.2 Long term seasonality

Figure 4.1 reveals that the residuals of the linear regression model with only the Centre Weight as independent variable, contain a cyclical behaviour. Although the cycles do not fit the meteorological seasons perfectly, the amount of chats arriving to the centre do not seem to be equally distributed over the year (Figure 4.2). The box-plot shown in Figure B.1 in Appendix B does not indicate clear differences between the monthly average residuals. The Kruskal-Wallis test is performed to determine whether there is a significant difference. This test is a non-parametric test which does not have the assumption that the data in each group are from a Normal distribution and have a homogeneous variance. The p-value of the Kruskal-Wallis test is 0, which indicates that there is a significant difference between some of the average residuals per month. Because of this result, the second independent variable (x_2) of the regression model, is the Month. The regression model is given in Table B.2 in Appendix B. The adjusted R-squared is slightly increased to 0.334.

Figure 4.3 shows the residuals of the regression model with both the centre weight and month as independent variables. The deviation from zero is less when compared to the residuals in Figure 4.1. Some of the waves have been removed, but there still might be some seasonality in the data. Despite that, the error of the regression model is reduced.

Besides the statistical evidence, it is important to be able to declare the patterns observed in the data and relate these to human behaviour. If it cannot be declared, it might be better to exclude this from the model. The reason for a yearly seasonality in chat arrivals, can probably be declared by the seasonality in car sales. As mentioned in chapter 1, the clients of Web1on1 considered in this thesis, all operate in the automotive industry. If more people are willing to buy a car, more people will visit car brand and car dealer websites and this will lead to an increase in the number of chats. The long term seasonality, the effect of the month of the year, can be explained and it is therefore decided to include the month as an independent variable in the regression model.

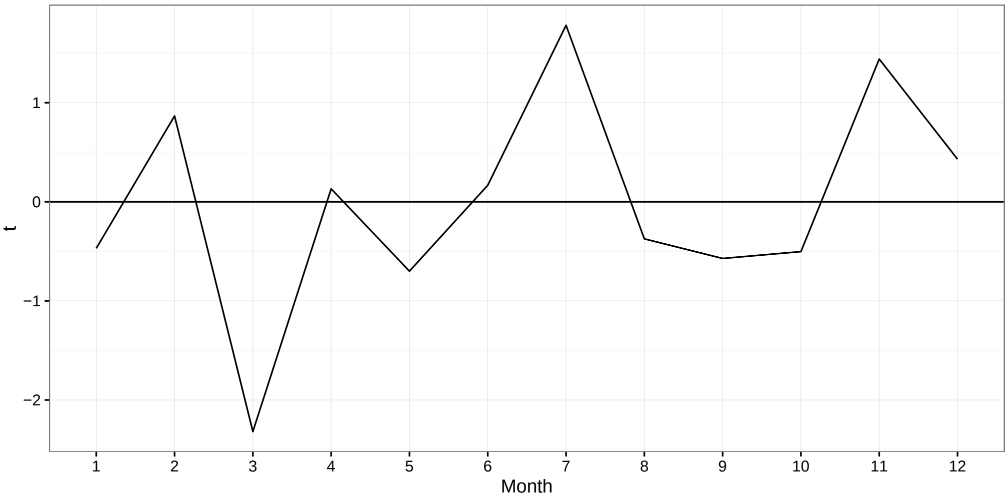


Figure 4.2: Residuals after regression model with centre weight per month

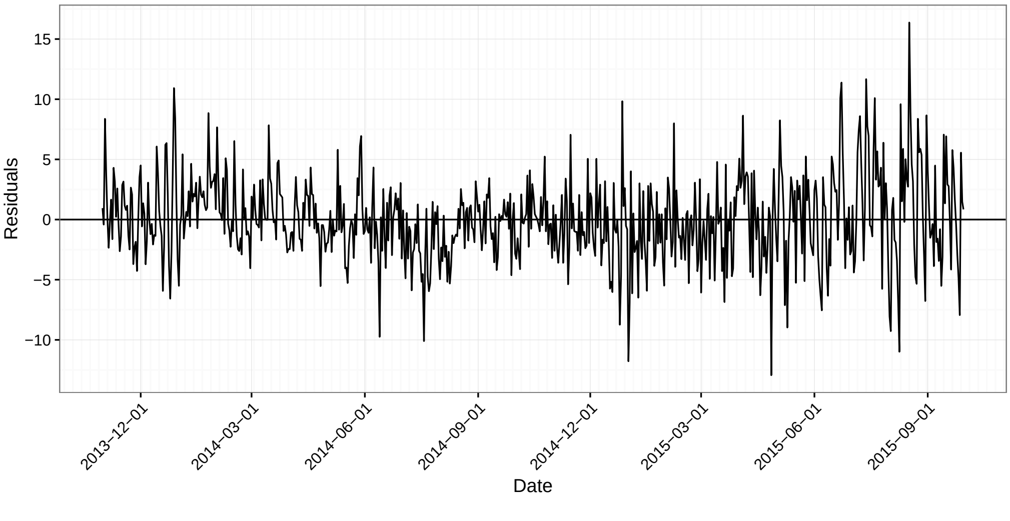


Figure 4.3: Residuals regression model with centre weight and month as independent variables

4.2.3 Public holidays

The data exploration in chapter 3 already indicated that there might be an effect of public holidays on the arrival rate of chats. A closer look at the residuals, reveals that a couple of days have a low amount of chat arrivals. It turns out that most of these are public holidays. It does make sense that, for example, during Christmas less people are visiting websites and consequently less people are starting a chat. Before including an independent variable for public holidays in the model, it is determined whether the effect is significant. The public holidays considered, are shown in Table A.2 in Appendix B.

Figure 3.6 in chapter 3 shows the average chat arrivals at public holidays and regular days. During the entire day, the amount of chats arriving per hour, is lower on a public holiday. A one-sided sample t-test confirms that the average arrival rate is significantly lower during public holidays. The p-value of the t-test is 0, which indicates that the average arrival rates on public holidays are not equal to the arrival rates on other days. For this reason, a dummy variable for public holiday (x_3) is added to the regression model.

As can be seen in Table B.3 in Appendix B, adding the dummy variable to the regression model, does not have a significant impact on the entire model performance. The adjusted R-squared is slightly increase to 0.336. Although the dummy variable is 0 on regular days and therefore not present in the model during those days, adding a dummy variable for holidays does not only effect the holidays itself. The dummy variable ensures that the coefficients have a better approximation and less biased by extreme values. Because there are only a couple of holidays per year, the residuals are the same in general. Therefore, the series is very similar to one displayed in Figure 4.3.

4.2.4 Weather effect

Based on the experience of the CSC centre employees, an external effect that might influence the arrival of chats is weather. To be able to correct the data for this effect, the weather is quantified. This is done by grading the weather with a number between 1 and 10, where 10 (1) is day with very good (bad) weather conditions. This weather grade used for this analysis has been developed by the website Weeronline.nl in 2008 and the method is explained at Onweer-online (2008). The weather grade is explained in more detail, before determining whether it affects arrivals enough to include it as an independent variable in the regression model.

The grade is based on four meteorological elements: wind, overcast, rainfall and fog. The grade starts with a 10 and there are points subtracted for each of the elements present. For example, if the total duration of rainfall throughout the day is between the 10 and 90 minutes, 1 point is subtracted. If the maximum wind speed is more than 6 Beaufort (Bft.), 3 points are subtracted. A complete overview of the quantification of the meteorological effects, is given in Table B.4 in Appendix B. For each day the weather grade is determined. There is no distinction made between the weather throughout the day. The weather station “De Bilt” is used for national meteorological measurements and is located near Utrecht. For this analysis, the measurements at this station are used to determine the weather grade for a particular day. The historic data is extracted from the data base of the Royal Netherlands Meteorological Institute (KNMI) which can be accessed via KNMI (2016).

For the range of the entire data set, there were no days with a grade of 1. Three days had scored a 2 and only two days scored a 10. A first look at Figure 3.7 in chapter 3 does not immediately show a clear difference between days with high grades and days with lower ones. On Mondays and in the weekend, it seems to have a larger impact than on other days of the week. The box plot of the average number of arrivals per weather grade, is shown in Figure B.2 in Appendix B. To determine whether the arrivals differ significantly during good and bad weather conditions, there are two tests applied. First, an Analysis of Variance (ANOVA) is performed to test whether the average arrival rates differ if the weather is different. The results of this test are given in Table B.5 in Appendix B. The p-value is 0, which shows that the mean arrival rate is not equal for all weather grades. To discover which grades differ significantly, a post-hoc test, the Tukey's HSD test, is used. The output of the test for each weather grade combination, is listed in Table B.6 in Appendix B. Only very good weather (grade 9), seem to have significantly less arrivals then most of the other grades.

Based on these results, including a variable for the weather grade would not improve the forecasting accuracy much. Adding a new parameter to the model requires an additional parameter estimate. This increases the uncertainty of the model and this is insufficiently compensated by an increased forecast accuracy.

Another option to consider is to use the clustered weather grades. These group bad, average and good grades. The results of the ANOVA (Table B.7 in Appendix B) and Tukey's HSD test (Table B.8 in Appendix B) show that there are no significant differences of the arrival rates between different weather clusters.

The uncertainty of weather forecasts has not been taken into account when considering the weather as an independent variable for the regression model. A few days ahead forecast, is probably accurate, but a week or two week ahead forecast probably not. Combining the results of the statistical tests and the fact that weather forecasts are uncertain, it is decided not to include the weather grade or the weather cluster as independent variables to the regression model. Nevertheless, the information revealed by the data analysis about the extreme effects in specific periods, are valuable for the CSC centre.

4.2.5 Stationary residuals regression model

Using a regression model with independent variables that explain the effects that cause the non-stationarity in the data, is preferred over a mathematical transformation. The objective of this model is to end up with a stationary series which is suitable to be used by the time series model. To see whether the series of the residuals of the regression model are stationary, the Augmented Dickey–Fuller (ADF) test is used.

The test reveals that the entire series is not stationary. There is no trend present in the data, but the variance is not constant. Fortunately, it is not necessary to have the entire series stationary because not the entire series will be used in the time series model. Only a time frame of the most recent weeks will be used as input for the model.

The ADF test is applied to determine what the maximum length of the series is, before it becomes non-stationary. This is the maximum amount of data that can be used as input for the time series model. It turned out that a series of 4 weeks or longer, is non-stationary. This limits the maximum length of data used for the time series to three weeks. This is a short period and might limit the performance of the time series model. Whether this is the case, will become clear in section 4.3. If more data is required than three weeks, the data still have to be transformed.

4.3 Time series model

The second forecasting model is applied to the residuals of the regression model. The yearly seasonality and the growth of the CSC centre are captured by the regression model. The weekly and daily patterns will be modelled by the second forecasting model, a time series model. Because both the weekly and daily seasonality have to be included in the model, the time series model must be able to model this double seasonality.

The model selected is a so called BATS model, which stands for Box-Cox transform, Autoregressive Moving Average model (ARMA(p,q)) errors, Trend and Seasonal components introduced by De Livera, Hyndman, and Snyder (2011). It is an extended version of the standard Exponential Smoothing (ES) model and consists of similar elements as those models. The three most important additional capabilities the BATS model has, are (i) to transform non-stationary data with a Box-Cox transformation, (ii) to include multiple seasonal periods and (iii) model the error terms as an ARMA(p,q) process. The BATS model is given in equation 4.1.

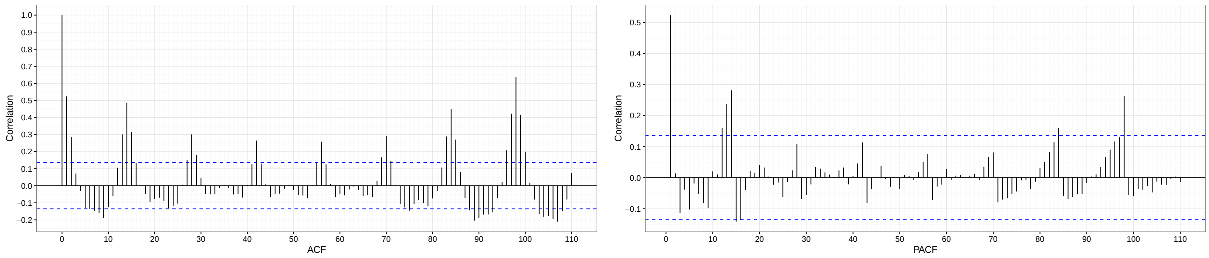
$$
\begin{aligned}
y_t^{(\omega)} &= l_{t-1} + \phi b_t + \sum_{i=1}^{T} s_t^{(i)} + d_t, \\
y_t^{(\omega)} &= \begin{cases} \frac{y_t^{\omega}-1}{\omega}, & \text{if } \omega \neq 0 \\ \log(y_t), & \text{if } \omega = 0 \end{cases} \\
l_t &= l_{t-1} + \phi b_{t-1} + \alpha d_t, \\
b_t &= (1-\phi)b + \phi b_{t-1} + \beta d_t, \\
s_t^i &= s_{t-m_i}^i + \gamma_{m_i} d_t, \\
d_t &= \sum_{i=1}^{p} \psi_i d_{t-i} + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} + \epsilon_t
\end{aligned}
\tag{4.1}
$$

where y_t are the residuals of the regression model in period t, m_i denotes the seasonal period i, l_t is the local level, b the long-run trend, b_t is the short run trend in period t, $s_t^{(i)}$ is the i^{th} seasonal component at time t, d_t denotes the ARMA(p,q) process for the error terms and ϵ_t is a Gaussian white noise process. ψ_i and θ_i are the parameters of the ARMA(p,q) process. ϕ is the damping parameter of the short run trend. ES models with a trend but without a damping parameter, assume that the trend continues indefinitely into the future. Gardner Jr and McKenzie (1985) introduced this parameter to dampen the trend. α, β and γ are the smoothing parameters.

As discussed in chapter 3, if the series has a length of 3 weeks or less, the stationary requirement is satisfied and the Box-Cox transformation does not have to be used. When more historic data has to be used to come up with an accurate forecast, the transformation can be used to achieve a stationary series. The long-run trend (b) and short-run trend (b_t) are not included in the model. The b is captured by the regression model and it is assumed that within a time period (t), the arrival rate is time homogeneous. This means that the arrival rate does not change within period t and the short-run trend (b_t) is zero.

The first seasonality component included in the time series model is the daily seasonality. For the CSC centre, a day consists of 14 hours (09:00-23:00) and hourly data is available. Consequently, the corresponding frequency of the daily seasonality is 14. The second one is the weekly seasonality and here the frequency is equal to 98. As can be seen in the Auto Correlation Function (ACF) (Figure 4.4a) and Partial Auto Correlation

Function (PACF) (Figure 4.4b) plots of the regression residuals, the lags at 14 and 98 are highly correlated. This means that an hour is significantly affected by what has happened 14 and 98 time periods ago. This confirms the presence of daily and weekly seasonality. Not only the seasonal lags are significant, but also the first lags of the spectrum and the lags surrounding the seasonal periods. It cannot be assumed that the error terms of a model with only the seasonal components, are white noise. Therefore, the error terms are modelled as an ARMA(p,q) process.



(a) ACF (b) PACF

Figure 4.4: ACF and PACF plot of residuals regression model

The last step in fitting the time series model, is estimating the parameter values α, θ, γ, p and q. Those depend on what data is used as input to fit the model. In general, it turned out that the values for α and γ are very small, which indicates that the estimates for the level and seasonality are very similar to the last observation. Hence, there is no upward or downward sloping trend and there is expected that the next observation is very similar to the last corresponding seasonal observation. The model is trained every time it has to forecast. Training the model with different data result in slightly different parameter values. The design of the time series model used, is given in formula 4.2.

$$\begin{aligned} y_t &= l_{t-1} + s_t^1 + s_t^2 + d_t, \\ l_t &= l_{t-1} + \alpha d_t, \\ s_t^1 &= s_{t-14}^1 + \gamma_{14} d_t, \\ s_t^2 &= s_{t-98}^2 + \gamma_{98} d_t, \\ d_t &= \sum_{i=1}^{p} \psi_i d_{t-i} + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} + \epsilon_t \end{aligned} \tag{4.2}$$

The last step in developing this model, is to determine whether using three weeks of historic data, is sufficient to enough to train a time series model which generates accurate forecast. Otherwise, non-stationary data has to be used and the Box-Cox transformation has to be applied. Models trained by two to twelve weeks of historic data are compared. It turned out that the length of the training set, did not have a significant impact on the forecasting accuracy. There is decided to train the time series model on the last three weeks of historic data, which is as long as possible while meeting the stationary requirement. Hence, no transformations have to be used.

4.4 Combined forecast

Both sub-models, the regression and time series model, give an hourly forecast. The value predicted by the regression model is the same for each hour of a particular day, because the independent variables do not depend on the hour of the day. The time series model actually predicts the error of the point forecast of the regression model. To construct the combined hourly forecast the two sub-forecasts are added. This results in a point forecast for the arrival rate of each hour.

In addition to the point forecasts, the Prediction Intervals (PI) are determined. The formula for 100(1-α)% PI for h time periods ahead from period N, $\hat{X}_N(h)$, is given in formula (4.3). This formula shows that the width of the PI depends on the error term of the forecasting model ($e_N(h)$). If there is any correlation between error terms of both sub-models, the width of the PI are also correlated. Therefore, it is important to determine whether the error terms of the sub-models are dependent or not.

The sub-models are executed sequentially and the input for the time series model depends on the output of the regression model. The forecast of the regression model does not affect the uncertainty or variance present in the data. It only determines the level of the time series model. This means that the level parameter of the time series model (long term (l)) are related to and dependent on the forecast of the regression model. The error terms of both models are not affected and therefore independent. As can be seen in equation 4.2, the uncertainty is captured in the error term d_t, while the effect of the regression forecast is absorbed by l_t. The term d_t determines the forecasting accuracy and consequently the width of PI interval for the time series model. Because this is not affected by the regression output, the error terms are independent and so are the width of the PI. Hence, the PI of the combined forecast is determined by simply adding the PI of the two sub-models together.

$$\hat{X}_N(h) \pm z_{\alpha/2}\sqrt{Var\big[e_N(h)\big]} \tag{4.3}$$

where $\hat{X}_N(h)$ is the point forecast of an h-step ahead forecast given data up to point N, $z_{\alpha/2}$ is the $1-\alpha/2$ quantile of the standard Normal distribution, and $e_N(h)$ is the forecasting error of the h-step ahead forecast.

The width of the prediction interval is determined by the prediction error of the forecasting model(s). The prediction error depends on two types of uncertainty:

1. uncertainty of the parameter estimates;
2. uncertainty of the input values.

In the regression model, only the parameters are uncertain. Those have to be estimated, but the input value is certain. For instance, if today is a Thursday, it is 100% certain that tomorrow will be Friday. The same applies to the other independent variables.

The input values for the time series model rely on (recent) events. What will happen in two hours, partially depends on what will happen in the next hour. Therefore, an h-step ahead forecast contains more uncertainty than a 1-step ahead forecast. With each step, the uncertainty increases.

The formulas in (4.5) and (4.6) give the (variance of the) prediction error of an h-step ahead forecast for the ARMA(p,q) model in formula (4.4). X_t is the observation at time t, L is the lag operator ($L^iX_t = X_{t-i}$), ϕ_i and θ_i are the parameters of the ARMA(p,q) process, and ϵ_t is a white noise process with variance σ^2_ϵ. It shows how $e_N(h)$, and consequently the width of the PI, has a positive relation with the forecasting horizon (h).

The PI of the time series model widens if h increases. Granger, White, and Kamstra (1989) give a more detailed description of the construction of forecasting intervals.

$$(1 + \sum_{i=1}^{p} \phi_i L^i) X_t = \mu + (1 + \sum_{i=1}^{q} \theta_i L^i) \epsilon_t \tag{4.4}$$

$$e_N(h) = \left[X_{N+h} - \hat{X}_{N+h}\right] = \epsilon_{N+h} + \sum_{j=1}^{h-1} \theta_j \epsilon_{N+h-j} \tag{4.5}$$

$$Var(e_N(h)) = (1 + \sum_{i=1}^{h-1} \theta_i) \sigma_\epsilon^2 \tag{4.6}$$

The width of the PI of the regression model is constant and the width of the PI of the time series model increases when h increases. The prediction errors of the sub-models are independent which makes it justifiable to add the two PI together to get the combined PI. Consequently, the width of the PI of the combined forecast increases when h increases.

4.5 Forecast per cluster

As described in chapter 1, the CSC centre has clustered their clients and for each cluster, different agent skills are required. Therefore, the staffing decisions and thus the forecasts, have to be on cluster level. In order to determine the capacity requirements for each cluster, the distribution of chat arrivals over the clusters is determined. During the data exploration in chapter 3, three aspects were discovered:

1. the distribution of chats is not equally divided over the three clusters;
2. the distribution of chats has changed over time;
3. the distribution of chats depends on the period (day and hour);

Because of these observations, the distribution of arrivals over the clusters is determined for each period (day and hour).

In general, the weight of the centre changes every month. The amount of work per client, depends on the elements explained in section 4.2.1 and these can vary a lot in a short time period. Therefore, the cluster distribution will fluctuate more than the centre weight. Due to these short term movements, the data on which the expected distribution is based, should not be too old. The expected cluster distribution is based on the average arrival rate distribution in the last four weeks. For instance, the distribution between the three clusters for next Monday 17:00h is the average of the cluster distribution of the last four weeks on Monday 17:00h. If this was 35%, 40% and 25% for the clusters 1, 2 and 3 respectively, the aggregated arrival rate is distributed over the clusters with the same ratio. The length of the period on which the average ratio is determined, is relatively short, but very similar to the length of the data used to fit the time series model (three weeks). The combination of the cluster distribution and the aggregated forecast, results in the forecast per cluster.

5 Capacity model

This chapter is dedicated to the second model which determines the capacity requirements based on the predictions made by the forecasting model. In chapter 1, it is mentioned that the system at the CSC centre will be modelled as a waiting queue. The system characteristics determine which queueing model is the most appropriate. This will be discussed in section 5.1. Given a certain queueing model, the corresponding performance indicators are discussed in section 5.2. In section 5.3, it is explained how the model is used to find the required number of agents to be scheduled. The forecasting model and this capacity model are implemented at the CSC centre Web1on1. In the last section of this chapter, the performance of those models is discussed.

5.1 Queueing model

The design of the waiting queue at Web1on1 is most similar to the M/M/c/c queue (Erlang B) as can be seen in Figure 2.2 in chapter 2. Arriving customers do not have to queue and are immediately assigned to a free server. If there are no servers available, the chat widget disappears and arriving customers are blocked. The largest difference in structure between the Erlang B model and the CSC centre is that one agent handles multiple customers simultaneously. One chat agent corresponds to three independent servers, independent of the agent type.

In the Erlang B system, customers arrive to the system according to a Poisson distribution. In chapter 3, it is shown that the variance of the arrival rate is larger than the mean. One of the characteristics of a Poisson process is that the mean is equal to the variance. Hence, the arrivals cannot follow a Poisson distribution. Based on historic data, there is no probability distribution which fits the data properly. Therefore, it is assumed that chats arrive independently and follow a general distribution (G or GI to be explicit).

The service time of a single chat is more difficult to determine, because there are multiple aspects that have an effect. Especially in the case of this CSC centre, where one agent can serve multiple customers simultaneously, the number of chats an agent is handling at the same time, will affect the service time. The response time of an agent will increase and this results in a longer chat duration. In addition to this, also the response speed of the customer and in which cluster the chat arrives, have an effect on the total service time of a chat. The data does not include any detailed information about the duration of a chat.

Fortunately, the time of interest is the total time a customer is occupying a server, independent whether the agent is actually busy with that specific customer or not. The time a customer has to wait for an answer from an agent or the time it takes for a customer to respond, is part of the total service time. There is data available on the total chat duration and this shows that the variance of the service time is larger than the mean (Table 3.1 in chapter 5). The distribution of the historic service times did not follow an exponential distribution or another standard distribution used in queueing theory. Therefore it is assumed that the service time follows a general distribution (G).

Based on the structure of the system, the distribution of the arrivals and the distribution of the service times, the queue at the CSC centre is modelled as a G/G/c/c system, where c corresponds to the number of servers.

In addition to the design of the queueing system, the number of servers one agent represents has to be determined. The expected service time of a chat is approximately 10 to 12 minutes. This means that one independent server is able to serve five to six chats per hour with an occupancy rate of 100%. Based on the experience of the CSC centre employees and the available data, the maximum number of chats one agent is able to handle per hour, is approximately 15 chats. Hence, the departure rate of one agent when having an occupancy rate of 100%, is approximately 15 chats per hour. Based on this, there is decided that one agent corresponds to three independent servers.

5.2 Performance indicators

Based on the queueing model of the CSC centre, the performance indicators for the system can be defined. Because there is no queue, there is no performance measure of the queueing or waiting time of customers. The most important indicator for the performance of the system, is the probability that an arriving customer is blocked (BP). The lower this probability, the better the service delivered to the customer and the higher the SL.

The system of the CSC is also referred to as the G/G/c loss system, because there are customers lost when all servers are occupied. Not surprisingly, the aspect that is related to the BP of the system, is the occupancy rate of a server. This is the utilization of the servers, also defined as the percentage of time a server is busy. The higher this percentage, the higher the probability that a customer will be blocked. The occupancy rate (ρ) can be calculated by dividing the workload (W) by the total number of servers scheduled (5.1 and 5.2). If the occupancy rate increases and approaches 1, the probability of blocking will also increase and approach 1.

$$W = \frac{\lambda}{\mu} \tag{5.1}$$

$$\rho = \frac{W}{c} \tag{5.2}$$

where λ is the arrival rate, $1/\mu$ is the expected service time and c is the number of servers.

For the M/M/c/c system the blocking probability can be determined. Unfortunately, there is no closed formula to calculate the blocking probability of the G/G/c/c system. Effective approximations for the steady-state blocking probability in the G/G/c/c model have been discussed in Li and Whitt (2014). For the capacity model at the CSC centre, the Hayward approximation is used to approximate the BP (Formula 5.3). This approximation corrects the BP of the M/M/c/c system by using the so called peakedness (z) of the system. This is defined as the ratio of the variance to the mean of the steady-state number of busy servers. Hence, if the peakedness is equal to 1, the blocking probability resulting from the Hayward approximation, is the same as the blocking probability in the

M/M/c/c model.

$$B(W/z, c/z) = \frac{(W/z)^{(c/z)}/(c/z)!}{\sum_{i=1}^{(c/z)} (W/z)^i/i!} \tag{5.3}$$

where W is the workload arriving to the system (5.1), z is the peakedness ($Var(N)/E(N)$) and N is the expected number of busy servers (λ/μ).

The realised blocking probability can deviate from the expected blocking probability, because of an error in one of the three input parameters (λ, μ, z). The better the forecast, the less the realised up-time will deviate from the expected up-time, the more stable the SL and the more constant the workload will be for the agents. As it has been shown in chapter 3, the expected service time of a chat ($1/\mu$) and the peakedness (z) depend on the position in the week (weekday and hour) and the client cluster. To have a reliable expected service time and peakedness, the average values of the last 4 weeks (per weekday, hour and cluster) are used as an approximation for the service time and peakedness.

To explain the behaviour of the capacity model and the relationships between the parameters, Figure 5.1 shows the relationships between the SL, the number of servers c and the workload W, given that the peakedness (z) is equal to 1. Hence, this is 1 - BP for a M/M/c/c queue. The more servers staffed, the better the system is able to achieve a high SL when the workload increases.

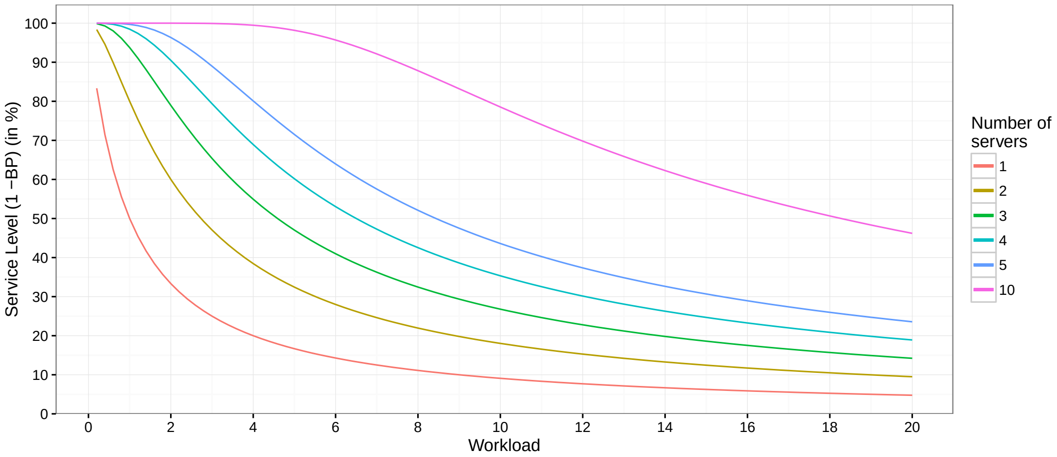


Figure 5.1: Effect of number of servers on Blocking Probability, given z = 1

Figure 5.2 shows the relationships between the SL, the workload W and the peakedness z, given that the number of servers (c) is equal to 6. For the value of z equal to 1, the behaviour is the same as in Figure 5.1. When z is more than 1, the variance of the workload is larger than the mean and the SL is lower when the workload increases. For z smaller than 1, the opposite is the case.

5.3 Capacity determination

Using the Hayward approximation for the BP (equation 5.3), the number of servers required to achieve a certain SL, can be determined. The four input variables required

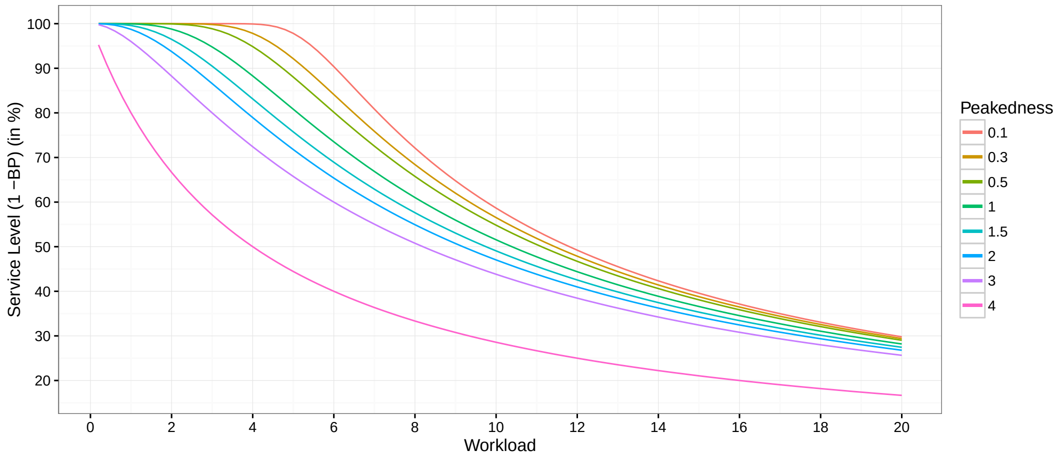


Figure 5.2: Effect of peakedness on Blocking Probability, given c = 6

are the forecasted arrival rates, the required service level, the expected service times and the peakedness (λ, SL, $1/\mu$ and z).

The capacity is determined by recursion. It starts with one server and as long as the SL is lower than the required level, one server is added. This results in the required number of servers per cluster for the centre. Each agent corresponds to three independent servers. Dividing the number of servers required by this amount, gives the agent capacity. This might not be an integer. The third model, addressed in Chapter 6, will determine how to achieve an optimal number of servers while only staffing an integer number of agents and how to distribute the capacity over the clusters.

As mentioned in section 5.2, the more accurate the forecast for the arrival rate (λ), the better the staffing levels can be determined and the closer the SL will be to the required level. Unfortunately, the forecasts of the arrival rate do include uncertainty. The capacity model takes into account the uncertainty in the distribution of the arrival rates and the expected service times, but does not correct for any uncertainties in the forecasts for the arrival rates. Therefore, in addition to the point forecasts, the upper and lower bounds of the 80% and 90% PI are part of the output of the capacity model. This gives Web1on1 insight in the amount of uncertainty of each forecast and offers them the opportunity to decide what the level of uncertainty is, they want to accept. When more agents are staffed, the probability of achieving the desired SL is higher, but this involves higher costs. Now, the CSC centre is able to make a trade off between quality and cost. In discussing the results, the point forecasts of the arrival rates are used. The performance of the model when using the upper and lower bounds of the PI, will not be addressed.

5.4 Results at Web1on1

The model implemented at Web1on1 consists of the forecasting and capacity model. The third model, the staffing model, is only developed for scientific purpose and not implemented at the CSC centre. The output of the first two models, the forecasting and capacity model, is the number of servers required to meet the SL. To determine the number of agents required, the the amount of servers is simply divided by three. In this case, the amount of agents might not be an integer. This optimization model redistributes the server capacity over the clusters to optimize the solution. Before addressing this model, the performance of the models implemented at Web1on1 is discussed.

The forecasting and capacity model are applied to the data of January through September 2015. The model is run for each day in this period. Every run, the forecasting models are trained with the most recent data: one year of historic data for the regression model and three weeks for the time series model. In this simulation, the amount of agents is rounded to the next integer to assure that the expected SL is at least the required level. The realised SL is determined by the actual arrivals. The SL (1-BP, the up-time of the system) and the amount of agents staffed are compared to the historic performances of Web1on1. The target SL is set to 90% up-time.

As mentioned in one of the research questions, it is important to investigate what impact the forecasting horizon has on the performance of the model. This gives insight in how the model should be used. There are two staffing horizons examined: 7 and 1 day (5.4.1 and 5.4.2 respectively).

5.4.1 7 day staffing horizon

The current staffing frequency of Web1on1 is once per week. In the beginning of each week the agent capacity and corresponding schedules for the next week are produced. Hence, this procedure causes a varying forecasting horizon. For Mondays it is 8 days, but for Sunday 14 days. This is due to the availability of data. For each day, the data until the previous day is available. Hence, running the model on Monday to staff next weeks' hours, results in a forecast between 8 and 14 days. This is graphically presented in Figure 5.3.

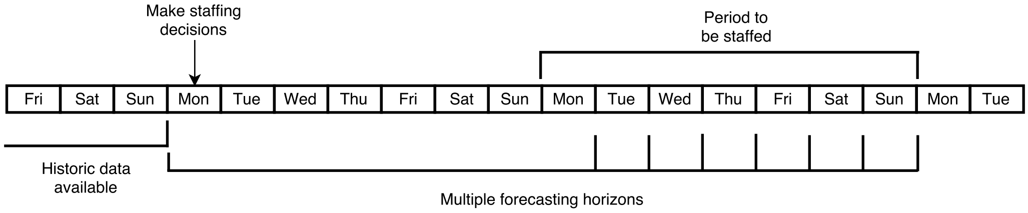


Figure 5.3: Staffing approach Web1on1

At the moment, the process is labour intensive and it is not possible to increase the staffing frequency. The models can be run on a daily basis and give Web1on1 the opportunity to make staffing decisions every day. This ensures that the forecasting and staffing horizon are the same for each weekday.

The first simulated scenario is the 7 day staffing horizon. This means that for each day, the model is trained and run and staffing decisions will be made for the same weekday next week. For example, on Monday the decisions for the staffing levels for next Monday are made. On Tuesday, the staffing levels for next week's Tuesday are determined, and so on. Hence, the forecasting horizon is 8 days, because there is only data available until Sunday. For Tuesdays there is data available until Monday, etc. This process is shown in Figure C.1 in Appendix C. This process has been simulated for the period January 2015 until September 2015. The model advices how many agents of each type should be staffed. The SL realised historically and by the model are graphically represented in Figure 5.4.

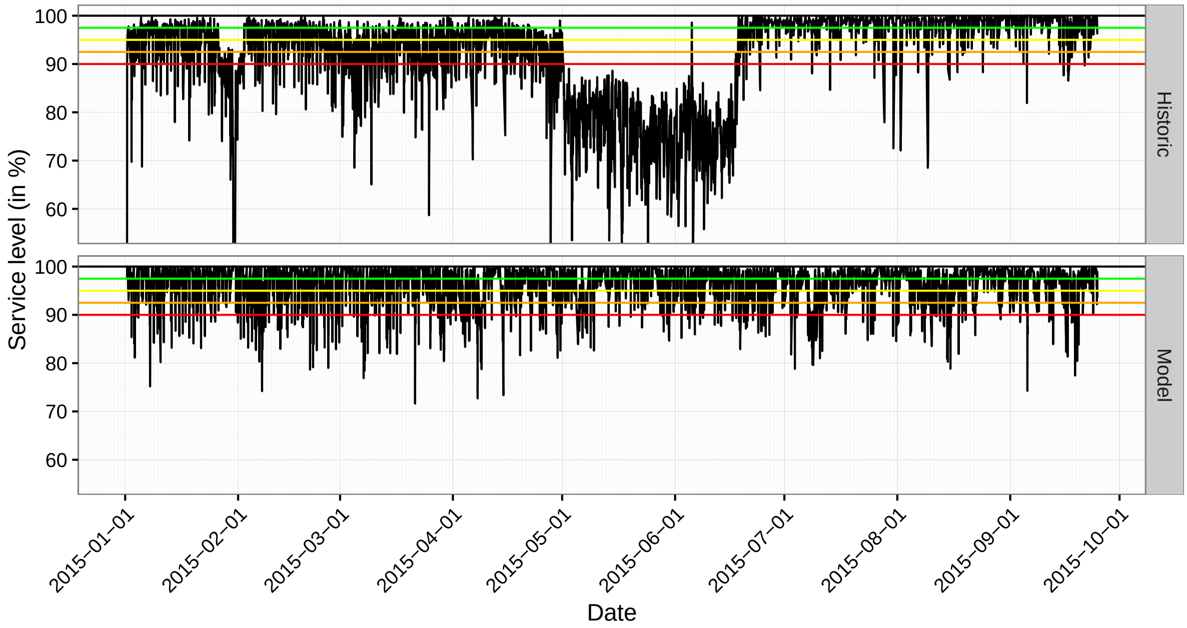


Figure 5.4: SL comparison 7 day staffing horizon

Based on the graph, it is hard to state whether and how much both series differ. The SL realised by the model seems to be more stable than it has been historically. There are no hours with a very low SL and the spread also seems to be reduced. To have a better understanding of the differences between the historical and model's performance, Table 5.1 gives some key measurements. The mean and variance of the SL, the amount of agents staffed and the percentage of periods that achieved a certain service level (90% and 95% respectively) are given. Historically, the period May - June has suffered from technical issues. Table C.1 in Appendix C shows the results when those months are excluded from the analysis.

Table 5.1: Performance of the model implemented at Web1on1 with 7 day staffing horizon

	Service level (%)				Agents staffed	
	Mean	Variance	>90%	>95%	Amount	Saving
Historic	91.62	0.9	70.2	53.4	19695	
Model	95.27	0.593	82.96	72.39	16988	13.745%

The average SL realised is increased and the variance is also reduced compared to the historic performance. Hence, the spread of the SL is decreased. Although there is targeted at a SL of 90%, the average is more than 95%. Next to those two measures, there is determined what percentage of the hours achieved a SL of at least 90% and 95%. Again the model has increased both percentages with respect to the historic performance. The final two columns of Table 5.1 show the amount of agents used to realise the performance. There is a 13.75% reduction in the amount of labour used. The model with a 7 day staffing horizon shows a higher and more stable SL, while the amount of agents required is reduced. When excluding the period May - June 2015 from the analysis, the historic SL shows a slightly better performance, but the model is very close while it staffed 11.57% agents less.

5.4.2 1 day staffing horizon

In the previous section, the 7 day staffing horizon is addressed. In this section the performance of the model when having a 1 day horizon is examined. This approach is visualized in Figure C.2 in Appendix C. Due to the reduced forecasting horizon, there are more accurate predictions expected which should result in a better staffing performance. The SL realised historically and by the model with a 1 day horizon are graphically represented in Figure 5.5.

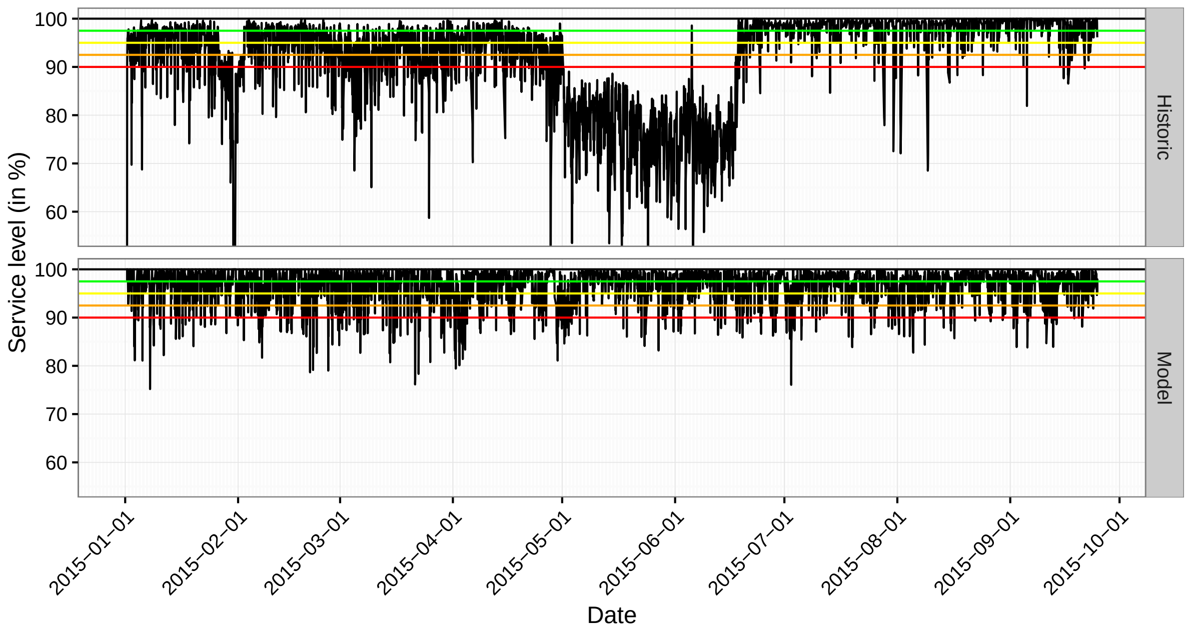


Figure 5.5: SL comparison 1 day staffing horizon

The series of the model seems to have an increased SL and a reduced spread. Again, it is hard to identify the differences between both series on the graphs itself. Table 5.2 shows the same measures as Table 5.1, but the results of the 1 day horizon are added to the table. Table C.1 in Appendix C shows the analysis without May - June 2015.

Table 5.2: Performance of the model implemented at Web1on1 with 1 and 7 day staffing horizon

	Service level (%)				Agents staffed	
	Mean	Variance	>90%	>95%	Amount	Saving
Historic	91.62	0.90	70.20	53.40	19695	
Model (7 days)	95.27	0.60	82.96	72.39	16988	13.745%
Model (1 day)	95.30	0.56	83.73	70.20	16993	13.719%

The decrease of the staffing horizon and consequently the forecasting horizon, shows a further improvement the performance of the SL on all measures, except the number of agents staffed. The amount of labour used to achieve the better SL, is slightly increased. Therefore, the performance of the model does not seem to improve a lot when the horizon is reduced from 7 to 1 day. When excluding the period May - June 2015 from the analysis

(Table C.1 in Appendix C), the BP is more or less the same and there is still a reduction of 11.43% in the amount of agents staffed.

It is important for chat agents to know a sufficient time in advance when they are expected to work. It is not possible to announce the schedules 1 day in advance. Given the results of this and the previous section, there is advised to make daily staffing decisions with a horizon of 7 days. This schedule can be communicate with the related agents. In the remaining time between the first decision and the actual moment of the schedule, up until one day in advance, the forecasts have to be monitored. When there are changes in the amount of workload expected and this requires different staffing levels, Web1on1 is able to adapt those levels. This approach gives the CSC centre the flexibility of making adaptions relatively short before the execution of the schedule, while agents are informed sufficiently in advance when they are expected to work. Given the results of this and the previous section, there is expected that no large changes have to be made shortly before the beginning of the schedule.

Although the model showed to be capable of at least maintaining the SL with reducing the amount of agents, it is interesting to determine whether the performance can be optimized even further. Therefore, in the next chapter, a capacity optimization model is developed to distribute the agent capacity over the different clusters. In Chapter 6, the performance of the model implemented at the CSC centre Web1on1 is compared to the performance of the model with the addition of an staffing optimization model.

6 Staffing model

This chapter is dedicated to the third and final model developed for the staffing of chat agents at CSC centres like Web1on1. It determines the optimal agent and server distribution for the different clusters where it is possible to transfer server capacity from one cluster to another. The objective is to maximize the profit of the CSC centre while achieving a predetermined SL. Next to this SL constraint, the model has to find a solution which also satisfies other centre specific constraints. The solution depends on two decision variables: (i) the number of agents of each type to be staffed and (ii) the number of servers to be transferred. These decisions are not trivial in any situation, as will be described in section 6.1. In the remainder of this chapter, the MINLP developed to optimize the staffing levels for each cluster per period is addressed. As said, section 6.1 outlines the system of the CSC centre to show the environment for which the model is developed. Thereafter, in section 6.2, the model is described in detail. Section 6.3 gives a short description of the algorithm used to solve the program. Section 6.4 discusses the performance of the model and compares it to the model implemented at Web1on1.

6.1 Essential centre characteristics

After the forecasting and capacity model, a third model is developed to determine the optimal staffing levels for each cluster, such that the total profit of the CSC centre is maximized within the constraints of the centre. As outlined in chapter 1 and 5, the chat agents correspond to three independent servers. Given the arrival rate predictions, expected service times, peakedness and the Hayward approximation, the BP related to the number of servers, is known.

Before describing the model in detail, the most important aspects of the CSC centre related to model are described. The model is developed for this specific situation, but it is possible to apply it in another setting by adapting the environment specific constraints of the model.

- The CSC centre has multiple clients. For each client, there is certain knowledge and experience required to serve customers in a proper way. The CSC centre has clustered the clients in three groups: cluster 1, 2 and 3. Cluster 1 and 2 contain clients which require more specific client knowledge and experience to be handled. Cluster 3 clients do not require specific knowledge and can be handled by less experienced agents.

- Web1on1 has three types of agents: type 1, 2 and 3. Each type corresponds to the cluster they can serve. Every new agent start as a type 3 agent. Those only chat for cluster 3 clients. When agents are more experienced, they are educated to chat for either cluster 1 or cluster 2. But next to this, they remain appropriate to chat for cluster 3 clients. This makes it a hierarchical workforce.

- Each Web1on1 agent is able to process multiple chats simultaneously and each agent represents three independent servers. Agents of type 1 or 2 can handle chats from multiple clusters. In the case of a type 1 agent, it is possible to have a part of the servers available for cluster 1 clients, and the remaining for cluster 3.

Table 1.1 in chapter 1 shows which agent type is able to serve customers of which cluster. The characteristic described in the last note, corresponds to the transfer of server capacity from one cluster to another. This is an important difference between the CSC centre and call centres. It makes the system more flexible and able to reach optimal solutions, while the number of servers is limited to a multiple of the number of servers one agent represents. In the staffing model, this is one of the key characteristics and therefore important to understand the concept. In chapter 1, an example is given to explain the principle of server transfers.

Depending on the cost structure of the agent types, it may be profitable to schedule extra agents of type 1 or 2 and transfer all those servers to cluster 3. For agent type 3 agents, it is important to gain experience. In the described scenario, a large part of the work these agents can do, is taken over by agents which are higher in hierarchy. This is undesirable and therefore the amount of transfers is limited. The strictness of the transferability of servers, is controlled by a constraint in the model.

6.2 Model description

This section is dedicated to the formulation of the model elements. The sets, parameters and variables used to design the MINLP model are described in section 6.2.1, 6.2.2 and 6.2.3 respectively. Hereafter, the objective function is given in section 6.2.4 and in section 6.2.5 the constraints of the model are explained in detail.

6.2.1 Sets

- $\mathcal{D} = \{1, \dots, 7\}$ denotes the first time index, the day of the week. The CSC centre operates 7 days a week. The day of the week is indexed by d.
  $d = 1$ corresponds to Monday
  $d = 2$ corresponds to Tuesday
  $\dots$
- $\mathcal{T} = \{1, \dots, 14\}$ denotes the second time index, the hour of the day. The CSC centre operates 14 hours a day (09:00 - 23:00). The hour of the day is indexed by t.
  $t = 1$ corresponds to the period [09:00 - 10:00)
  $t = 2$ corresponds to the period [10:00 - 11:00)
  $\dots$
- $\mathcal{I} = \{1, 2, 3\}$ denotes the set of cluster types, indexed by i.
- $\mathcal{J} = \{1, 2, 3\}$ denotes the set of employee types, indexed by j. Table 1.1 in chapter 1 shows which employee type j is able to serve customers at cluster i in case of Web1on1.

6.2.2 Parameters

- α_j is the number of servers that correspond to one agent of type j.

- p_i is the price per chat served in cluster i in Euros.
- c_j is the hourly wage of agent type j in Euros.
- $\lambda_{i,t,d}$ is the arrival rate of chats in cluster i at time t on day d.
- $1/\mu_{i,t,d}$ is the expected service rate in cluster i at time t on day d.
- $z_{i,t,d}$ is the peakedness of the queueing system as described in section 5.2 in cluster i at time t on day d.
- SL_i is the maximum BP allowed in cluster i.
- β_i is the maximum number of servers in cluster i that is allowed to be handled by an agent which is higher in hierarchy. In case of Web1on1, it is not allowed to have one or more agents replaced by transferred servers.

6.2.3 Variables

- $W_{j,t,d}$ is the number of agents of type j staffed at time t on day d.
- $X_{i,j,t,d}$ is the number of transferred servers of agent type j to cluster i at time t on day d.
- $C_{i,t,d}$ is the number of servers serving in cluster i at time t on day d. This depends on the variables $W_{j,t,d}$ and $X_{i,j,t,d}$.
- $B_{i,t,d}$ is the blocking probability in cluster i at time t on day d. This depends on the values of $C_{i,t,d}$, $\lambda_{i,t,d}$, $\mu_{i,t,d}$ and $z_{i,t,d}$.
- $N_{i,t,d}$ is the number of chats processed in cluster i at time t on day d. This depends on the values of $\lambda_{i,t,d}$ and $B_{i,t,d}$.

The total number of chats that are processed in cluster i at time t on day d, depends on the realised arrival rate ($\lambda_{i,t,d}$) and the blocking probability ($B_{i,t,d}$). The Hayward approximation is used to approximate $B_{i,t,d}$ and is determined by $C_{i,t,d}$, $\rho_{i,t,d}$ and $z_{i,t,d}$. The value for $z_{i,t,d}$ is known for all i, t and d, $\rho_{i,t,d}$ is the workload (equation 6.1) and relies on the forecasted value of $\lambda_{i,t,d}$, and $C_{i,t,d}$ is one of the decision variables.

$$\rho_{i,t,d} = \frac{\lambda_{i,t,d}}{\mu_{i,t,d}}, \qquad \forall i \in \mathcal{I},\ t \in \mathcal{T},\ d \in \mathcal{D} \tag{6.1}$$

The variables $W_{j,t,d}$, $X_{i,j,t,d}$, $C_{i,t,d}$ and $N_{i,t,d}$ are all integers and the function for $B_{i,t,d}$ (5.3) is non-linear. It is important to include this non-linear, queueing behaviour in the staffing model. Combining this with the constraints, the staffing model is a constrained MINLP. These type of programs are hard to solve and require algorithms to find (optimal) solutions. Even if appropriate algorithms are used, it is not certain that always an optimal solution is found and that the solution is always the same. Which algorithm is used to solve this MINLP and how there can be dealt with finding a solution, will be discussed in section 6.3.

6.2.4 Objective function

The objective of this optimization model is to staff the CSC centre such that the profit is maximized and the constraints such as the SL, are met. Maximizing profit is a trade-off between revenue and costs. Staffing additional agents will increase labour cost, but may also increase the total number of chats processed. The CSC centre receives a revenue per chat handled. The more chats, the higher the revenue. The overhead and fixed costs of the centre do not depend on the number of chats or the amount of agents staffed and are not considered in this model. The total profit of the centre is defined by Π (6.2).

$$\Pi = \sum_{\mathcal{D}} \sum_{\mathcal{T}} \sum_{\mathcal{I}} \sum_{\mathcal{J}} N_{i,t,d} \cdot p_i - W_{j,t,d} \cdot c_j \tag{6.2}$$

6.2.5 Constraints

In addition to the objective function, several constraints are defined to complement the MINLP model. Those serve three different purposes. The first type of constraints define the relationships between the variables and parameter of the CSC centre. Next to these, there are constraints to make sure that the staffing solution meets all user desired requirements, for example the minimum SL or the maximum number of servers allowed to be transferred to cluster i (β_i). The last type defines the feasible upper/lower bounds and integer restrictions for the model. All constraints will be discussed in detail.

The first four constraints define the relationships between variables and parameters. Those relationships translate the processes at the CSC centre to mathematical equations. Constraint (P.1) has two objectives. First it defines that the number of chats that will be served in cluster i, is equal to the product of the number of arrivals at cluster i and the blocking probability at cluster i. Next to this and in combination with constraint (P.13), it ensures that $N_{i,t,d}$ is rounded to the nearest integer value. $B_{i,t,d}$ is the approximation of the BP as given in equation 6.3. This is determined by three input parameters: $C_{i,t,d}$, $\rho_{i,t,d}$ and $z_{i,t,d}$. For each period t, d and cluster i, $\rho_{i,t,d}$ and $z_{i,t,d}$ are known. $C_{i,t,d}$ is a decision variable. The more servers staffed, the lower the BP. Constraint (P.2) - (P.4) represent how the number of agents ($W_{j,t,d}$) and the number of transfers ($X_{i,j,t,d}$) establish the number of servers in cluster i ($C_{i,t,d}$). It is very specific for the situation at Web1on1. In Appendix D a more general description of the MINLP model is given.

$$B_{i,t,d} = \frac{(\rho_{i,t,d}/z_{i,t,d})^{(C_{i,t,d}/z_{i,t,d})} / (C_{i,t,d}/z_{i,t,d})!}{\sum\limits_{i=1}^{(C_{i,t,d}/z_{i,t,d})} (\rho_{i,t,d}/z_{i,t,d})^i / i!} \tag{6.3}$$

$$\begin{aligned}
&N_{i,t,d} \leq \lambda_{i,t,d}(1 - B_{i,t,d}) + 0.5 && \forall \quad i \in \mathcal{I},\ t \in \mathcal{T},\ d \in \mathcal{D} && \text{(P.1)}\\
&C_{1,t,d} = W_{1,t,d} \cdot \alpha_1 - X_{3,1,t,d} && \forall \quad t \in \mathcal{T},\ d \in \mathcal{D} && \text{(P.2)}\\
&C_{2,t,d} = W_{2,t,d} \cdot \alpha_2 - X_{3,2,t,d} && \forall \quad t \in \mathcal{T},\ d \in \mathcal{D} && \text{(P.3)}\\
&C_{3,t,d} = W_{3,t,d} \cdot \alpha_3 + X_{3,1,t,d} + X_{3,2,t,d} && \forall \quad t \in \mathcal{T},\ d \in \mathcal{D} && \text{(P.4)}
\end{aligned}$$

Constraint (P.5) restricts that the SL is at least equal to SL_i for cluster i. This ensures that the solution meet the performance requirement as demanded by the CSC centre. Constraints (P.6) and (P.7) restrict that a maximum of $\alpha_3 - 1$ servers are allowed to be transferred. This prevents the model from replacing one or more entire agents of type 3.

$$\begin{aligned} B_{i,t,d} &\leq SL_i && \forall \quad i \in \mathcal{I},\ t \in \mathcal{T},\ d \in \mathcal{D} && \text{(P.5)} \\ X_{3,1,t,d} + X_{3,2,t,d} &\leq \beta_3 && \forall \quad t \in \mathcal{T},\ d \in \mathcal{D} && \text{(P.6)} \\ \beta_3 &< \alpha_3 && \forall \quad t \in \mathcal{T},\ d \in \mathcal{D} && \text{(P.7)} \end{aligned}$$

Constraints (P.8) - (P.10) are the boundary conditions for the variables. Those make sure that at least one agent and server is staffed per cluster per period and that transfers are non-negative. The constraints (P.11) - (P.13) restrict the variables $W_{j,t,d}$, $X_{i,j,t,d}$ and $N_{i,t,d}$ to integer values. $C_{i,t,d}$ has to be an integer as well, but this is forced by the integer constraints of the other variables and the fact that α_j is an integer as well.

$$\begin{aligned} W_{j,t,d} &\geq 1 && \forall \quad i \in \mathcal{J},\ t \in \mathcal{T},\ d \in \mathcal{D} && \text{(P.8)} \\ X_{i,j,t,d} &\geq 0 && \forall \quad i \in \mathcal{I},\ j \in \mathcal{J},\ t \in \mathcal{T},\ d \in \mathcal{D} && \text{(P.9)} \\ C_{i,t,d} &\geq 1 && \forall \quad i \in \mathcal{I},\ j \in \mathcal{J},\ t \in \mathcal{T},\ d \in \mathcal{D} && \text{(P.10)} \\ W_{j,t,d} &\in \mathbb{N} && \forall \quad j \in \mathcal{J},\ t \in \mathcal{T},\ d \in \mathcal{D} && \text{(P.11)} \\ X_{i,j,t,d} &\in \mathbb{N} && \forall \quad i \in \mathcal{I},\ j \in \mathcal{J},\ t \in \mathcal{T},\ d \in \mathcal{D} && \text{(P.12)} \\ N_{i,t,d} &\in \mathbb{N} && \forall \quad i \in \mathcal{I},\ t \in \mathcal{T},\ d \in \mathcal{D} && \text{(P.13)} \end{aligned}$$

6.3 Direct search algorithm

The non-linear constraints and non-differentiable approximation for the BP, significantly increase the difficulty of solving the problem. In the relatively small setting of Web1on1, it has been tried to translate the problem into an easier to solve MP model, but unsuccessful. MINLP problems have the complexity of both Mixed Integer Programming model (MIP) problems and non-linear constraints. The most appropriate approach for solving these type of NP-hard problems, is by the use of heuristics.

In order to solve the MINLP model, the software MATLAB (2016) and the solver NOMAD are used. This solver implements the Mesh Adaptive Direct Search (MADS) algorithm to solve black-box optimization problems. It is programmed in the programming language C++, but is available in a free MATLAB (2016) toolbox, called OPTimization Interface (OPTI) toolbox. A detailed description of the NOMAD solver is given in the article of Le Digabel (2011) and more information about the OPTI toolbox can be found at Currie and Wilson (2012).

The MADS algorithm used to solve the MINLP belongs to the family of Pattern Search (PS) algorithms. These do not require the problem to have a gradient in order to be find a solution. To have a better understanding how the MADS algorithm operates towards a solution, the algorithm is briefly described.

An iteration can consist of two steps: a search and polling step. Given an initial point x_0, the algorithm starts each iteration by evaluating some trial points. These trial points lie on the current mesh. This is a grid constructed by N directions from the initial point or the current best feasible solution. The value of the objective function in each of these points, are compared with the current best feasible solution. If there is a feasible point on this mesh which has a better solution than the best objective value found so far, this will be the new best feasible solution. An iteration that found such an improved value, is called successful and the next iteration will start from this new best solution. If the iteration fails to find an improved point on the mesh, the algorithm moves to the polling step before terminating the iteration. Again, trial points are generated, but the magnitude of the distance from these points to the current best feasible point, decreases each iteration. This forces the algorithm to search more locally each iteration. The maximum distance the algorithm is allowed to search for better solutions, is defined by the poll size parameter. Next to this parameter, there is the mesh size parameter which defines the step size on the mesh. The difference between Generalized Pattern Search (GPS) algorithms and MADS is in the polling step. In the GPS algorithm, it is required to set the mesh parameter equal to the polling parameter. In other words, the step size of the mesh is equal to the maximum distance of trial points from the current best feasible solution. For the MADS algorithm, this is not required. This gives the algorithm more flexibility to find optimal solutions. A more detailed description and graphical representation of the MADS algorithm and the difference between GPS algorithms, can be found in the article of Audet and Dennis Jr (2006).

6.4 Performance evaluation

In the previous chapter, the performance of the model implemented at Web1on1 is evaluated. It has been shown how the forecasting and capacity model improves the staffing performance for Web1on1. The model described in this chapter optimizes the staffing solution by redistributing the server capacity. This section describes the performance of the staffing model and compares it to the performance realised by the model implemented at Web1on1. The same simulation approach is used. For each day in the period January 2015 through September 2015, the model is run and the staffing levels are determined. For this analysis only the 1 day staffing horizon is examined. Due to the small difference between the 1 and 7 day horizon observed in Chapter 5, it is expected that the performance with a 7 day horizon is not much different from the performance with a 1 day horizon.

The MINLP model requires the revenue per chat handled (p_i) and the hourly wages for the chat agents (c_i) as input parameters. For the analysis performed in this section, the revenue per chat is set to € 4,- and the wage is € 10,- per agent per hour. The revenue and wages are independent of the cluster and agent type. Changing these values has an impact on the model and may impact the staffing decisions. The higher the price of a chat, the more profitable it is to staff an additional agent. Having wages dependent on the type of agent, affects the decisions on staffing levels and the number of transfers. The target SL is set to 90% for all clusters ($SL_i = 0.9$).

Figure 6.1 shows the SL historically and achieved by the staffing model with a 1 day staffing horizon. It is clear that the spread of the SL has decreased and there are fewer hours with a low SL. Table 6.1 shows the key measures of the performance and the total amount of agents staffed. The performance of the model implemented at Web1on1 is also shown for comparison. As expected, the variance of the SL is lower than historically and achieved by the model implemented at Web1on1. The amount of hours achieving the minimum SL of 90% is increased, but the amount of hours exceeding a SL of 95% is decreased. Hours which have an overcapacity result in hours with a high SL (> 95%). Due to the redistribution of the overcapacity, the amount of hours with a SL over 95% is reduced. The total amount of agents staffed is even slightly further decreased when compared to the model implemented at Web1on1. Compared to the historic performance, the amount of agents is reduced by more than 14%. Table D.3 in Appendix D shows the same measurements as Table 6.1, but the period May - June is excluded from the analysis. The results of the staffing model are similar and the savings in labour is reduced 11.8% when compared to the historic performance.

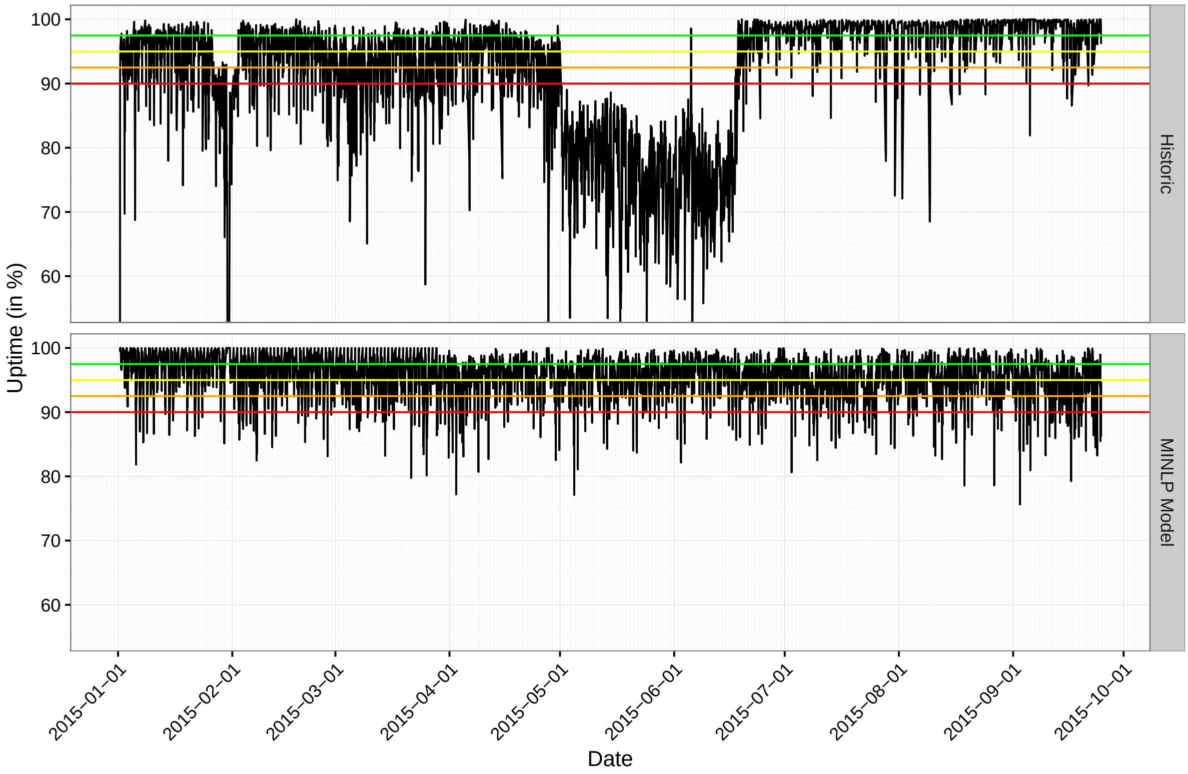


Figure 6.1: SL realised historically and by staffing model

Table 6.1: Performance of the staffing model with 1 day staffing horizon

	Service level				Agents staffed	
	Mean	Variance	>90%	>95%	Amount	Saving (in%)
Historic	91.62	0.90	70.20	53.40	19695	
Model (7 days)	95.27	0.60	82.96	72.39	16988	13.745%
Model (1 day)	95.30	0.56	83.73	70.20	16993	13.719%
MINLP	95.24	0.12	91.40	61.10	16934	14.019%

7 Conclusion

This chapter concludes this thesis by discussing the research questions introduced in chapter 1 (section 7.1). This summarizes of the findings of this study. Section 7.2 discusses the impact of the models on Web1on1 and other CSC centres. Hereafter, the limitations of the models and the research are addressed in section 7.3. In the last section of this chapter (7.4), both scientific and company specific future research directions are given.

7.1 Research questions

In chapter 1, the research questions are formulated. The answers on each of these questions are given throughout the thesis. In this section, they are brought together to summarize the findings of this research. The results are given per research question.

1. What insights does the chat data provide and which of these can improve the staffing performance of Web1on1?
   (a) What are the most important factors that have an influence on the arrival of chats and how are these related?

The data analysis revealed that there is a relationship between the website traffic and the arrival rate of chats. Having more clients will result in more website traffic and this will generate more chats. If the number of website traffic of the centre increases with 100,000 monthly visitors, the number of arrivals per hour will increase by 1,5 chat. This is equal to a conversion rate from traffic to chats of 0.6%.

The month of the year also has an effect on the number of arrivals. Not all months do differ significantly, but there is a yearly seasonality present in the data. There is only one entire year of data available and this may influence the outcome of the analysis. If there is more data available in the future, it would be interesting to repeat the analysis and see whether the effects still occur with the same magnitude.

Holidays have a negative effect on the number of chat arrivals. Throughout the entire day, there arrive less chats than on a regular day. The forecasting accuracy is affected on both general and public holidays. By including a dummy variable for holidays, the coefficients of regular days are not biased by the arrivals on holidays.

The short term seasonality of the week and day was clearly present in the data. Working days show a very similar pattern, but the weekend is very different. On working days, the evening shows the highest peaks while in the weekend the mornings show higher arrival rates. Over time, this short term seasonality has changed.

The weather did not have a significant and clear effect on the arrivals of chats, but some periods of the week seem to be affected. On Mondays, there are more chat arrivals when the weather is bad (weather grade 1 - 4), while Wednesday and Thursday show an opposite behaviour under similar circumstances. The effect of the weather on arrivals on Saturday and Sunday are similar, but initiated in a different way. On Saturdays, there are more arrivals when the weather is bad, but the arrival rate is not reduced when the weather is good. On Sundays, there are less arrivals when the weather is good, but not more than average when the weather is bad. Due to these ambiguous observations and the uncertainty of weather forecasts, the weather is not included as an independent variable in the regression model. Still, for Web1on1 it is valuable and important to keep in mind the effects discovered when making staffing decisions.

The last important aspect to be mentioned, is that patterns and relationships have changed over time. They are strongly related to the number and type of clients. The advice is to monitor those relationships closely and make sure that the model is reviewed periodically to maintain and improve the performance.

2. What forecasting model predicts the arrival of workload to the system accurate enough to generate valuable input for the capacity model?

   (a) What are the frequencies on which the forecasts and staffing decisions can be made?

The forecasting model consist of two parts, first a regression model which models the long term trend, long term seasonality and the effect of holidays. Hereafter, a time series model is used to capture the short term relationships and dependencies in time.

Chapter 5 revealed that the SL of the CSC centre is not much improved when the staffing horizon is reduced from 7 days to 1 day. The reason for this is that the most valuable hours to improve the prediction accuracy, are at the time lags less than 15. In other words, 1 day or less in advance of the beginning of the period. By reducing the staffing horizon from 7 days to 1 day, the forecasting horizon is only reduced to 2 days. If Web1on1 is able to be more flexible in adapting the staffing levels just before the beginning of the period, the forecasts will become more accurate and more advantage can be gained.

Although, it is recommended to run the model daily. Make the first staffing decisions one or two weeks in advance and adapt those during the passage of time. This approach offers to opportunity to make daily staffing decisions, but still have staffing levels well in advance to inform agents. Due to the relatively small improvement from a 7 day horizon to 1 day, there is expected that it is not required to make large adaptions in the staffing levels shortly before the beginning of the schedule.

3. How should Web1on1 staff their chat agents, such that the service level delivered is at least as agreed upon with the clients, while minimizing the total labour costs?

   (a) What queueing model fits the system the best?
   (b) What are the constraints that limit the solution space and define the staffing model?

The design of the system at Web1on1 is most similar to an Erlang B model, because there is no queue and customers can only arrive when there is at least one server available. When all servers are busy, customers are blocked. Different from the models in call centres, is that one chat agent serves multiple customers simultaneously. Therefore one agent represents multiple servers. In the case of Web1on1 this is equal to three independent servers.

The arrival rates do not follow a Poisson distribution and the service times are not exponentially distributed. Therefore, it is decided to use a G/G/c model to determine the server capacity. For this type of centre, a closed formula for the blocking probability does not exist. The Hayward approximation is used to estimate this probability.

The solution space of the optimization model is limited by the constraints of the MINLP model as described in section 6.2.5. The most important ones are the constraints (P.2) - (P.4). Those model the relationships between the type of agent and the clusters. In addition to that, the transferring of servers is also captured in these constraints. This is one of the important aspects of the staffing model.

7.2 Research implications

In this section, the results of the models developed will be summarized and the implications for Web1on1 are elaborated. In addition, the application of those techniques to other contact centres is discussed as well.

The forecasting and capacity model have both been implemented at Web1on1 to support the decision making of staffing chat agents. The first important contribution this model delivers, is the objectification and standardization of the decision making. Due to the rapid growth of the CSC centre in the recent years, it is important to make processes more efficient. The model can be run on a daily basis and this offers Web1on1 the opportunity to intensify the frequency of decision making. This will result in more accurate forecasts, better staffing levels, an improved performance and a reduction in the labour costs. As shown in Tables 5.1 and 5.2 in Chapter 5, the SL is slightly improved and more stable, while the amount of agents staffed has been reduced.

Next to the forecasting and capacity model, a staffing model has been developed to optimize the server capacity over the different clusters. The hierarchical structure of the workforce at Web1on1 makes it possible to transfer overcapacity from one cluster to another. As shown in Table 6.1 in Chapter 6, the percentage of hours achieving the required SL (a blocking probability of 10% or less) has further increased to over 90%. Compared to the performance of the model implemented at Web1on1, this almost an improvement of 8 percentage points. The amount of hours with an overcapacity (SL more than 95%) has been reduced and the amount of agents staffed has slightly further decreased. Those results show the additional value of having the possibility to transfer server capacity between client clusters.

An important feature of the model which has not been mentioned yet, is the independence between the three different models: forecasting, capacity and staffing model. Each of these can be adapted separately, as long as the input and/or output remains the same. For example, in order to implement the model at Web1on1 the queueing part had to be adapted. Web1on1 has developed their own chat software and this is operational since November 2015. The consequence of switching the software, is that the design of the queueing system also changed. In the new system, the queueing behaviour is very similar to an Erlang A model (Mandelbaum & Zeltyn, 2007), a model which includes a waiting queue and customer abandonment. The capacity model has been adapted to be able to implement the model at Web1on1. The forecasting model has not been changed. This example demonstrates the flexibility and adaptability of the model.

The modular design of the total workforce staffing model - which consists of the forecasting, capacity and staffing model - makes it relatively easy to adapt the model when specific centre characteristics change. In addition, applying the model to another CSC centre with a different design or workforce hierarchy belongs to the possibilities. The staffing model developed in Chapter 6 is developed for the situation at Web1on1. When having a different structure of the client clusters or workforce, the staffing model can be adapted. In Appendix D, a general description of the staffing model is given. This can be used as the starting point when applying or implementing the staffing model at other contact centres.

The design of the workforce staffing model makes it possible to adapt it towards the circumstances. This makes it applicable to other CSC and contact centres. The results achieved at Web1on1 are promising. The emergence of new types of contact centre will increase the demand for models to operate those centres efficiently. The dynamic environment and rapid changes in communication tools in the recent years, require to have flexible models which are able to adapt to changes.

7.3 Modelling limitations

The methods used in this thesis and the models developed involve limitations. This section will address the limitations of each model separately. Before going into the model specific limitations, it is important to mention a general limitation. Some minor changes in the environment or data structure can make a model useless. Therefore, it is important to realise that those models have to be maintained and updated frequently.

In the forecasting model, the weight of the CSC centre is an important indicator for the arrival rates. Historically, the centre weight can be determined, but there is an uncertainty about what this will be in the future. Having the forecasting model relied on a prediction, is a downside of this approach.

The time series of the residuals of the regression model is not stationary when a period more than three weeks is considered. This limits the possibilities of using data to train model. The objective of the regression model has only partly been achieved.

The capacity model assumes that within each time period of an hour, the system is in steady state and the chance of a chat arrival is equally likely. In practice, this is not the case. Within an hour, there can be fluctuations in the number of arrivals. For example, if a popular television show on Saturday evening or the daily news broadcast has been finished, there will be more website traffic and it is more likely that there will be a chat arrival. Due to data limitations, it was not possible to tighten the time window of each period. The capacity model does not take into account fluctuations of the arrival rate within an hour.

It is assumed that the servers of an agent operate independently. In addition, the service times used by the queueing model do not depend on the number of customers that share the agent's capacity, while this will have an effect. For each individual chat agent, it is an optimal productivity level, as discussed by Tezcan and Zhang (2014). If this level is exceeded, the service time will increase. To determine what the effect is of capacity sharing on the service time, requires more detailed information on the chat duration. This was not available in the case of Web1on1, but worth investigating in the future.

The capacity model uses the forecasts of the arrival rate made by the forecasting model. The uncertainty in the predictions is not taken into account by the capacity and staffing model. In the analysis performed in this thesis, the point forecasts are used. The output of the model implemented at Web1on1, includes the 80% and 90% prediction intervals of the forecasts and in addition also the required agent capacity for each level. This gives them insight in the consequences of the forecasting uncertainty for the staffing levels and offers them the opportunity to make a trade-off between cost and quality.

Another limitation of the capacity model is the assumption that all agents are the same. A distinction is made between the average service time per cluster per period, but not between individual agents. Not every agent has the same productivity and this uncertainty is not included in the model. While making the scheduling decisions and assigning individuals, it is important to keep this aspect in mind.

The staffing model uses the MADS algorithm to find optimal solutions. Unfortunately, due to the complex formula for the blocking probability and the integer constraints, it is difficult to solve the problem and it takes some computation time to find the optimal solution. For each period, it requires between 60 and 90 seconds to find a solution. To have the model find a solution for the next 7 days (98 periods), it takes between 1.5 and 2.5 hours. With a time window of daily decision making, this is acceptable, but when the frequency is further increased the algorithm is not fast enough and has to be improved to reduce computation time.

7.4 Future research directions

The future research directions are split into scientific and company specific recommendations.

As mentioned in chapter 2, only a few papers have been written about CSC centres. It is expected that the interest in how to operate these centres efficiently, will increase in the future. An important aspect which requires further research is the effect of capacity sharing on the productivity of agents and consequently on the service times. Tezcan and Zhang (2014) showed how to route arriving customers towards agents to optimize their productivity. Further research on agent productivity is required.

There are several expectations about how contact centres will develop in the future. One possibility is that those centres will become messaging centres. The main difference between chatting and a messaging, is the structure of the conversation. A chat has a start and end point and is conducted within a relatively short time period. A message is just a part of a conversation send from one individual or company to another. The conversation does not have to have a clear start or end point. A response does not have to be given immediately, but can be sent later. For a contact centre, this would be a new dimension. A conversation does not have to be processed by one agent, but can be postponed to someone else who is more specialized in the topic of the message, for example. Messaging is similar to email to some extent, but shorter and send via Instant Messaging applications on smartphones, like Facebook Messenger and WhatsApp Messenger. How to design messaging centres efficiently and what impact this has on the staffing and scheduling decisions, are future research directions.

For Web1on1 it is important to continue the quantification of relationships between the most important aspects in the CSC centre: the arrival rate, service time, agent productivity and the SL. Incorporating data analysis in the decision making will be the key to improve the performances and continue growing.

The workforce staffing model quantifies some of the relationships. It is possible to determine the most profitable SL, given the wages of the agents and the revenue of a chat. It is also interesting to do the analysis the other way around; what is the appropriate price of chat when a client requires a certain SL, such that the profit margin is still met? Dynamic pricing may be worth considering in the future.

An aspect which has not been mentioned in this project, but what is an interesting research direction, is Conversion Ratio Optimization (CRO). Which decisions can be made to affect the conversion ratio from website visitors to chats and what is the impact of such a decision? This is an important mechanism to have an influence on the workload arriving to the system and finding a balance between demand and supply.

A final, more innovative direction is the use of text mining. The amount of text produced by the chats, contain a lot of information. The larger the centre becomes, the more difficult it is to have insight in the content of conversations. Automated procedures which monitor chats and the performance of agents, is an interesting direction to investigate.

References

Aksin, Z., Armony, M., & Mehrotra, V. (2007). The modern call center: A multi-disciplinary perspective on operations management research. *Production and Operations Management*, *16*(6), 665–688.

Aldor-Noiman, S., Feigin, P. D., & Mandelbaum, A. (2009). Workload forecasting for a call center: Methodology and a case study. *The Annals of Applied Statistics*, 1403–1447.

Atlason, J., Epelman, M. A., & Henderson, S. G. (2004). Call center staffing with simulation and cutting plane methods. *Annals of Operations Research*, *127*(1-4), 333–358.

Audet, C., & Dennis Jr, J. E. (2006). Mesh adaptive direct search algorithms for constrained optimization. *SIAM Journal on optimization*, *17*(1), 188–217.

Avramidis, A. N., Chan, W., Gendreau, M., L'ecuyer, P., & Pisacane, O. (2010). Optimizing daily agent scheduling in a multiskill call center. *European Journal of Operational Research*, *200*(3), 822–832.

Avramidis, A. N., Deslauriers, A., & L'Ecuyer, P. (2004). Modeling daily arrivals to a telephone call center. *Management Science*, *50*(7), 896–908.

Bhulai, S., Koole, G., & Pot, A. (2008). Simple methods for shift scheduling in multiskill call centers. *Manufacturing & Service Operations Management*, *10*(3), 411–420.

Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S., & Zhao, L. (2005). Statistical analysis of a telephone call center: A queueing-science perspective. *Journal of the American statistical association*, *100*(469), 36–50.

Cezik, M. T., & L'Ecuyer, P. (2008). Staffing multiskill call centers via linear programming and simulation. *Management Science*, *54*(2), 310–323.

Currie, J., & Wilson, D. I. (2012, 8–11 January). OPTI: Lowering the Barrier Between Open Source Optimizers and the Industrial MATLAB User. In N. Sahinidis & J. Pinto (Eds.), *Foundations of Computer-Aided Process Operations.* Savannah, Georgia, USA.

Dantzig, G. B. (1954). Letter to the editor-a comment on edie's "traffic delays at toll booths". *Journal of the Operations Research Society of America*, *2*(3), 339–341.

Defraeye, M., & Van Nieuwenhuyse, I. (2013). Controlling excessive waiting times in small service systems with time-varying demand: an extension of the isa algorithm. *Decision Support Systems*, *54*(4), 1558–1567.

De Livera, A. M., Hyndman, R. J., & Snyder, R. D. (2011). Forecasting time series with complex seasonal patterns using exponential smoothing. *Journal of the American Statistical Association*, *106*(496), 1513–1527.

Edie, L. C. (1954). Traffic delays at toll booths. *Journal of the operations research society of America*, *2*(2), 107–138.

Ernst, A. T., Jiang, H., Krishnamoorthy, M., & Sier, D. (2004). Staff scheduling and rostering: A review of applications, methods and models. *European journal of operational research*, *153*(1), 3–27.

Feldman, Z., Mandelbaum, A., Massey, W. A., & Whitt, W. (2008). Staffing of time-varying queues to achieve time-stable performance. *Management Science*, *54*(2), 324–338.

Gans, N., Koole, G., & Mandelbaum, A. (2003). Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management*, *5*(2), 79–141.

Gardner Jr, E. S., & McKenzie, E. (1985). Forecasting trends in time series. *Management Science*, *31*(10), 1237–1246.

Granger, C. W. J., White, H., & Kamstra, M. (1989). Interval forecasting: an analysis based upon arch-quantile estimators. *Journal of Econometrics*, *40*(1), 87–96.

Gurvich, I., Luedtke, J., & Tezcan, T. (2010). Staffing call centers with uncertain demand forecasts: A chance-constrained optimization approach. *Management Science*, *56*(7), 1093–1115.

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

Jongbloed, G., & Koole, G. (2001). Managing uncertainty in call centres using poisson mixtures. *Applied Stochastic Models in Business and Industry*, *17*(4), 307–318.

Kim, J. W., & Ha, S. H. (2012). Advanced workforce management for effective customer services. *Quality & Quantity*, *46*(6), 1715–1726.

KNMI. (2016). *Daggegevens van het weer in nederland.* Retrieved 2016-02-08, from `https://www.knmi.nl/nederland-nu/klimatologie/daggegevens`

Le Digabel, S. (2011). Algorithm 909: Nomad: Nonlinear optimization with the mads algorithm. *ACM Transactions on Mathematical Software (TOMS)*, *37*(4), 44.

Li, A. A., & Whitt, W. (2014). Approximate blocking probabilities in loss models with independence and distribution assumptions relaxed. *Performance Evaluation*, *80*, 82–101.

Mandelbaum, A., & Zeltyn, S. (2007). Service engineering in action: the palm/erlang-a queue, with applications to call centers. In *Advances in services innovations* (pp. 17–45). Springer.

MATLAB. (2016). *version 9.0 (r2016a).* Natick, Massachusetts: The MathWorks Inc.

Onweer-online. (2008). *Het weercijfer, zo werkt het.* Retrieved 2015-04-13, from `http://www.onweer-online.nl/forum/topic/9275/het-weercijfer-zo-werkt-het/`

Shae, Z.-Y., Garg, D., Bhose, R., Mukherjee, R., Guven, S., & Pingali, G. (2007). Efficient internet chat services for help desk agents. In *Services computing, 2007. scc 2007. ieee international conference on* (pp. 589–596).

Shen, H., & Huang, J. Z. (2008). Forecasting time series of inhomogeneous poisson processes with application to call center workforce management. *The Annals of Applied Statistics*, 601–623.

Shumsky, R. A. (2004). Approximation and analysis of a call center with flexible and specialized servers. *OR Spectrum*, *26*(3), 307–330.

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

Tezcan, T., & Zhang, J. (2014). Routing and staffing in customer service chat systems with impatient customers. *Operations research*, *62*(4), 943–956.

A Data

Table A.1: Input data

Date	Hour	Cluster	Chat arrivals	Avg. Duration (in sec)	Website visitors	Agents scheduled	Uptime system (%)
31-10-13	15:00 - 16:00	BC1	6	433.83	475	1	99.43
31-10-13	15:00 - 16:00	BC2	1	541	396	1	100
31-10-13	15:00 - 16:00	BCD	1	996	412	1	100
31-10-13	16:00 - 17:00	BC1	6	389.67	468	1	96.55
31-10-13	16:00 - 17:00	BC2	1	996	412	1	100
31-10-13	16:00 - 17:00	BCD	5	484.6	457	1	100
31-10-13	17:00 - 18:00	BC1	6	505.67	546	2	92.17
31-10-13	17:00 - 18:00	BC2	5	474.4	886	2	100
31-10-13	17:00 - 18:00	BCD	5	484.6	1069	2	89.47
31-10-13	18:00 - 19:00	BC1	1	478	440	1	99.15
31-10-13	18:00 - 19:00	BC2	7	531	773	1	100
31-10-13	18:00 - 19:00	BCD	9	499.44	958	1	100
31-10-13	19:00 - 20:00	BC1	1	478	417	1	100
31-10-13	19:00 - 20:00	BC2	15	552	644	3	100
31-10-13	19:00 - 20:00	BCD	14	409.21	915	2	100

Table A.2: Public holidays included

New years day	January 1
Kingsday (Dutch holiday)	April 27
Good Friday	*depends*
Easter	*depends*
Easter Monday	*depends*
Christmas eve	December 24
Christmas	December 25
2^{nd} Christmas day	December 26
New years eve	December 31

B Forecasting model

Table B.1: Regression model with Centre weight as independent variable

Coefficients				
	Estimate	Std. Error	t-value	p-value
(intercept)	6.14E+00	3.00E-01	20.48	0
Weight	1.26E-07	1.84E-09	68.64	0
Adj. R-squared	0.3246			
p-value	0			

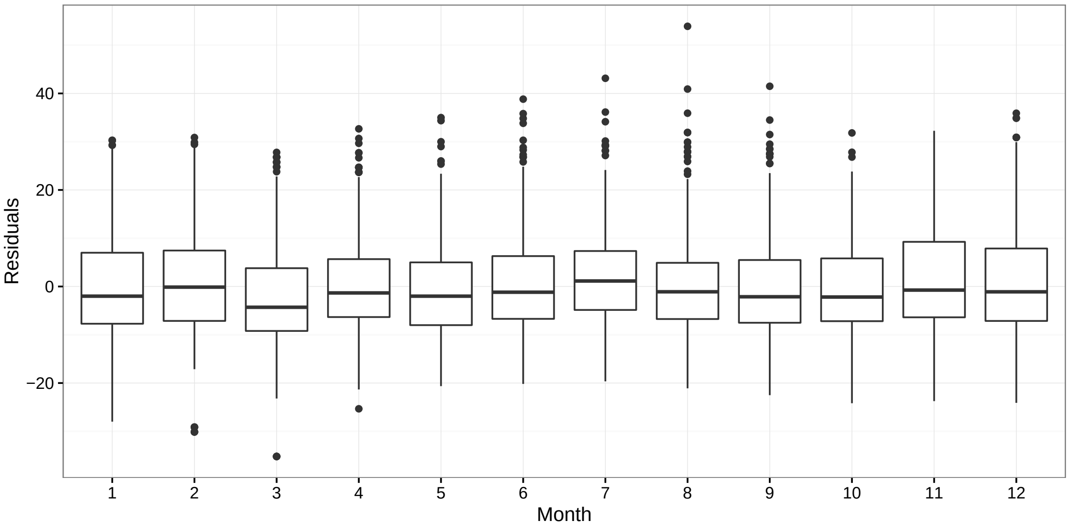


Figure B.1: Boxplot residuals per month

(Onweer-online, 2008)

Table B.2: Regression model with Centre weight and Month as independent variables. January as reference month

Coefficients				
	Estimate	Std. Error	t-value	p-value
(intercept)	4.59E+00	4.38E-01	10.483	0
Weight	1.32E-07	2.05E-09	64.164	0
Month02	1.26E+00	4.82E-01	2.613	0.009
Month03	-2.07E+00	4.76E-01	-4.341	0
Month04	9.70E-01	4.75E-01	2.042	0.0411
Month05	1.43E-01	4.71E-01	0.303	0.7621
Month06	1.02E+00	4.74E-01	2.147	0.0319
Month07	2.64E+00	4.71E-01	5.607	0
Month08	3.56E-01	4.75E-01	0.749	0.4539
Month09	1.92E-02	4.91E-01	0.039	0.9688
Month10	5.11E-01	5.69E-01	0.898	0.3691
Month11	2.47E+00	4.76E-01	5.188	0
Month12	1.18E+00	4.73E-01	2.499	0.0125
Adj. R-squared	0.3335			
p-value	0			

Table B.3: Regression model with Centre weight, Month and Holiday as independent variables.

Coefficients				
	Estimate	Std. Error	t-value	p-value
(intercept)	4.59E+00	4.38E-01	10.483	0
Weight	1.32E-07	2.05E-09	64.164	0
Month02	1.26E+00	4.82E-01	2.613	0.009
Month03	-2.07E+00	4.76E-01	-4.341	0
Month04	9.70E-01	4.75E-01	2.042	0.0411
Month05	1.43E-01	4.71E-01	0.303	0.7621
Month06	1.02E+00	4.74E-01	2.147	0.0319
Month07	2.64E+00	4.71E-01	5.607	0
Month08	3.56E-01	4.75E-01	0.749	0.4539
Month09	1.92E-02	4.91E-01	0.039	0.9688
Month10	5.11E-01	5.69E-01	0.898	0.3691
Month11	2.47E+00	4.76E-01	5.188	0
Month12	1.18E+00	4.73E-01	2.499	0.0125
Holiday1	-3.95E+00	6.90E-01	-5.726	0
Adj. R-squared	0.3356			
p-value	0			

Table B.4: Grades subtracted per weather element

Rainfall	Duration	Overcast	Exp. Coverage
Substract 0	< 10 minutes	Substract 0	<1/8
Substract 1	10-90 minutes	Substract 1	1/8 - 5/8
Substract 2	90-300 minutes	Substract 2	5/8 - 7/8
Substract 3	300-500 minutes	Substract 3	>7/8
Substract 4	>500 minutes		
Wind	**Speed**	**Visibility**	
Substract 0	0-2 Bft.	Substract 0	No fog
Substract 1	3 Bft.	Substract 1	< 6 hours fog
Substract 2	4-5 Bft.	Substract 2	> 6 hours of fog
Substract 3	>5 Bft.		

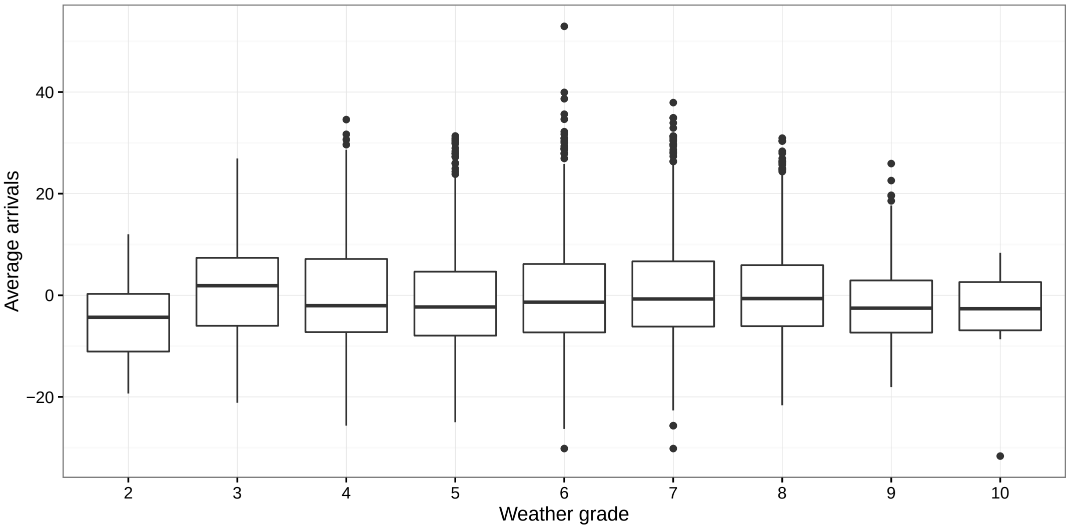


Figure B.2: Boxplot residuals per weather grade

Table B.5: Output ANOVA test weather grades

	df	Sum of squares	Mean square	F Value	P-value
Weather grade	8	5273	659.2	4.611	1.24E-05
Residuals	9771	1396738	142.9		

Table B.6: Output Tukey's HSD test weather grade

Grade	Difference	Lower	Upper	p-value
3-2	4.818	-1.639	11.275	0.333
4-2	5.828	-0.048	11.704	0.054
5-2	5.933	0.138	11.727	0.040
6-2	5.075	-0.700	10.849	0.139
7-2	5.038	-0.730	10.807	0.144
8-2	5.412	-0.382	11.207	0.089
9-2	2.566	-3.456	8.589	0.925
10-2	5.214	-3.835	14.264	0.691
4-3	1.010	-2.262	4.282	0.989
5-3	1.114	-2.008	4.237	0.973
6-3	0.257	-2.829	3.342	1.000
7-3	0.220	-2.854	3.294	1.000
8-3	0.594	-2.529	3.717	1.000
9-3	-2.252	-5.779	1.276	0.557
10-3	0.396	-7.224	8.017	1.000
5-4	0.104	-1.503	1.712	1.000
6-4	-0.753	-2.288	0.782	0.845
7-4	-0.790	-2.302	0.722	0.794
8-4	-0.416	-2.025	1.194	0.997
9-4	3.262	-5.560	-0.964	0.000
10-4	-0.614	-7.749	6.521	1.000
6-5	-0.858	-2.041	0.325	0.373
7-5	-0.894	-2.047	0.259	0.281
8-5	-0.520	-1.798	0.758	0.942
9-5	-3.366	-5.446	-1.287	0.000
10-5	-0.718	-7.786	6.349	1.000
7-6	-0.036	-1.086	1.013	1.000
8-6	0.338	-0.848	1.523	0.994
9-6	-2.508	-4.532	-0.485	0.004
10-6	0.140	-6.912	7.191	1.000
8-7	0.374	-0.782	1.530	0.986
9-7	-2.472	-4.479	-0.466	0.004
10-7	0.176	-6.871	7.222	1.000
8-9	-2.846	-4.927	-0.765	0.001
8-10	-0.198	-7.266	6.870	1.000
10-9	2.648	-4.608	9.904	0.969

Table B.7: Output ANOVA test weather group

	df	Sum of squares	Mean square	F Value	P-value
Weather group	8	661	330.6	2.303	0.1
Residuals	9797	1405946	143.5		

Table B.8: Output Tukey's HSD test weather group

Grade	Difference	Lower	Upper	p-value
average-bad	0.017	-0.986	1.020	0.999
good-bad	-0.506	-1.493	0.482	0.453
good-average	-0.523	-1.123	0.077	0.102

C Capacity model

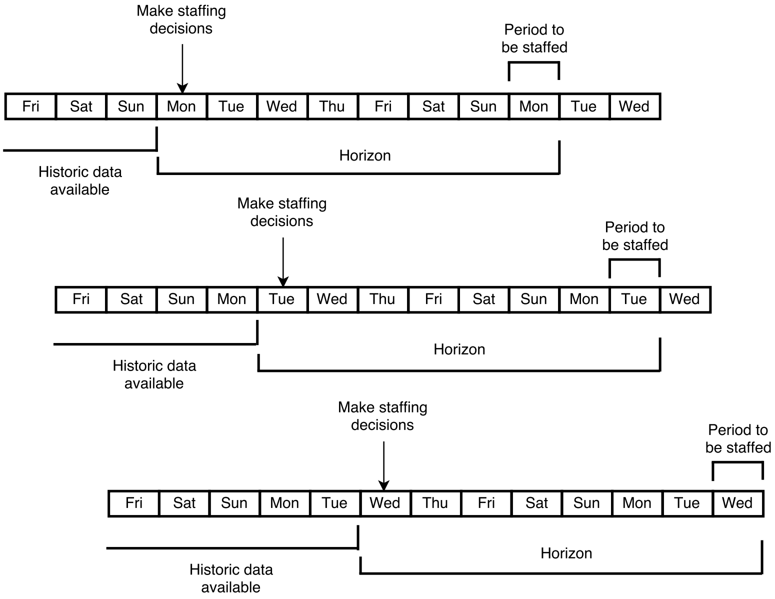


Figure C.1: Staffing approach 7 day staffing horizon

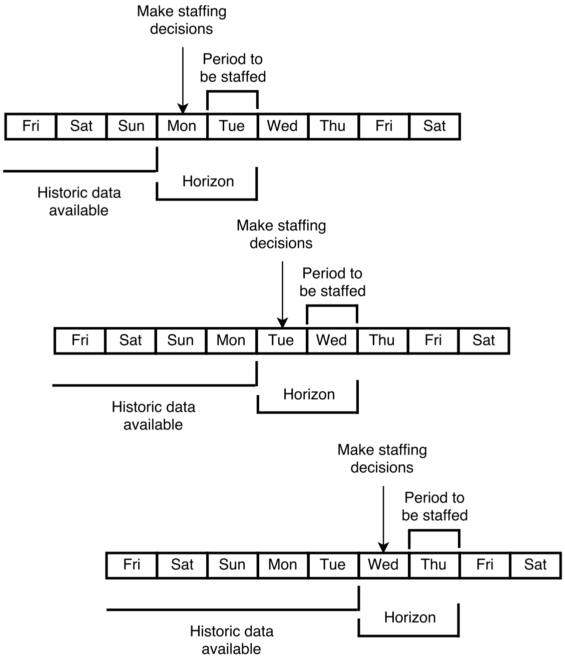


Figure C.2: Staffing approach 1 day staffing horizon

Table C.1: Performance of the model at Web1on1 with 1 and 7 day staffing horizon, except May - June 2015

	Service level (%)				Agents staffed	
	Mean	Variance	>90%	>95%	Amount	Saving (in%)
Historic	94.97	0.34	85.28	64.39	14711	
Model (7 Days)	95.10	0.61	82.16	71.32	13030	11.427%
Model (1 day)	95.17	0.56	82.87	71.57	13009	11.570%

D Staffing model

$$
\begin{aligned}
\max \sum_{\mathcal{D}} \sum_{\mathcal{T}} \sum_{\mathcal{I}} \sum_{\mathcal{J}} & N_{i,t,d} \cdot p_i - W_{j,t,d} \cdot c_j \\
& \text{s.t.} \\
N_{i,t,d} & \leq \lambda_{i,t,d}(1 - B_{i,t,d}) + 0.5 \\
C_{i,t,d} & = W_{j,t,d} \cdot \alpha_j \cdot \mathbb{1}_{i=j} + X_{i,j,t,d} \cdot \mathbb{1}_{\text{to cluster i}} - X_{i,j,t,d} \cdot \mathbb{1}_{\text{from cluster i}} \\
X_{i,j,t,d} & = X_{j,i,t,d} \\
B_{i,t,d} & \leq SL_i \\
\sum_{\mathcal{I}} X_{i,j,t,d} & \leq \beta_j \\
\beta_j & < \alpha_j \\
W_{j,t,d} & \geq 1 \\
X_{i,j,t,d} & \geq 0 \\
C_{i,t,d} & \geq 1 \\
W_{j,t,d} & \in \mathbb{N} \\
X_{i,j,t,d} & \in \mathbb{N} \\
N_{i,t,d} & \in \mathbb{N}
\end{aligned}
$$

where $\mathbb{1}_{i=j}$ is the indicator function whether i is equal to j, $\mathbb{1}_{\text{to cluster i}}$ is the indicator function whether it is allowed to transfer server capacity of agent type j *to* cluster i and $\mathbb{1}_{\text{from cluster i}}$ is the indicator function whether it is allowed to transfer server capacity *from* cluster i to support agents of type j. Hence, $\mathbb{1}_{\text{from cluster i}}$ is the transposed matrix of $\mathbb{1}_{\text{to cluster i}}$. The indicator functions for the case of Web1on1 are given in Table D.1 and D.2.

Table D.1: Indicator function $\mathbb{1}_{\text{to cluster i}}$ at Web1on1

	Agent type		
Cluster	1	2	3
1	0	0	0
2	0	0	0
3	1	1	0

Table D.2: Indicator function $\mathbb{1}_{\text{from cluster i}}$ at Web1on1

	Agent type		
Cluster	1	2	3
1	0	0	1
2	0	0	1
3	0	0	0

Table D.3: Performance of the staffing and model at Web1on1, except May - June 2015

	Service level (%)				Agents staffed	
	Mean	Variance	>90%	>95%	Amount	Saving (in%)
Historic	94.97	0.34	85.28	64.39	14711	
Model (7 Days)	95.10	0.61	82.16	71.32	13030	11.427%
Model (1 day)	95.17	0.56	82.87	71.57	13009	11.570%
MINLP	95.19	0.13	90.88	60.33	12981	11.760%