Capacity planning of emergency call centers

Author
R. P. Dwars

Supervisors
prof. dr. R. D. van der Mei
dr. S. Bhulai

Centrum Wiskunde en Informatica
Vrije Universiteit Amsterdam

August 27, 2013

Centrum wiskunde en informatica (CWI)
Science Park 123
1098 XG Amsterdam

Vrije Universiteit Amsterdam (VU)
De Boelelaan 1105
1081 HV Amsterdam

Preface

This paper is the final report of my internship at Centrum Wiskunde & Informatica CWI), as part of my master Business Analytics at Vrije Universiteit in Amsterdam. CWI is a renowned institute for mathematics and computer science in Amsterdam and it was founded in 1946. Business Analytics is a multi-disciplinary study program which focuses on improving business processes using economics, mathematics and computer science. During my internship I was able to practice these disciplines even further and I am grateful to CWI for providing me this opportunity.

Over the past six months I have had a wonderful time at CWI and I would like to thank everyone in Stochastics department for letting me feel at home. I would like to thank my supervisors prof. dr. Rob van der Mei and dr. Sandjai Bhulai for their support and coaching during these months. I would also like to thank Geert-Jan Kommer of RIVM (Rijksinstituut voor Volksgezondheid en Milieu) for revising my report and sharing his knowledge about emergency call centers. My gratitude also goes to Regionale Ambulance Voorziening Hollands-Midden (RAVHM) for giving me a tour at their emergency call center, which really improved my understanding of how emergency call centers work. Last but not least, I would like to thank Martin van Buuren and Jan-Pieter Dorsman for sharing their knowledge of their respective fields with me.

Contents

Management summary

This report provides an analysis of the impact of *staffing* on the *efficiency* and *performance* of emergency call centers. The performance is measured using the percentage calls that have a waiting time less that a certain time period. The efficiency is determined using the *minimum required centralists* to meet certain service level requirements. The methods used in this research are discrete event simulation and queueing theory. On top of the simulation, a user-friendly tool has been developed that helps management to analyze various 'what if' scenarios concerning staffing. This tool can answer questions such as:

- 'How much centralists would we need to have when there would be one nation-wide call center?'
- 'What is the impact of function differentiation on the performance of an emergency call center?'
- 'What would be the impact on the performance if the duration of the triage or dispatching could be improved by 10%?

The performance for different staffing combinations (i.e., combinations of call takers, dispatchers and generalists) has been analyzed using simulation. Results of this analysis show that combinations with only call takers and dispatchers are vulnerable to *bottlenecks*. Adding a generalist to such combinations solves this problem. Also, various scenarios for merging emergency call centers have been evaluated and the results show a *significant increase* in efficiency. However, a drawback of merging emergency call centers is that regional knowledge is reduced. Insufficient regional knowledge of centralists slows down the triage and the dispatching of ambulances. Therefore, a model has been created that takes this factor into account. The results show that efficiency can still be increased while regional knowledge is preserved (to some extent).

Main findings

- Merging call centers leads to staff reductions and and therefore an increase in efficiency. The plan of the government of going from 21 to 10 call centers will lead to a significant improvement in efficiency, but even more can be achieved by merging into one or two call center(s).
- Efficiency can still be gained, when regional knowledge is included in the model.
- The call volumes of the current 21 call centers *too low* to operate with call takers and dispatchers and therefore it is better to use generalists in these call centers. However, as call centers get bigger it becomes more beneficial to use function differentiation.
- Even when parameters have been under- or over estimated, the performance of A1- and A2-calls proves to be quite robust. However, for B-calls and follow-up calls, the performance indicators are less robust.

The results in this report are meant for *tactical* and *strategic* decision making only, because the models in this report don't take shift restrictions of centralists into account and also not the daily/weekly/seasonal fluctuations in the call frequency. The model used in this report could also be extended for usage on operational level, but more data would be needed in order to achieve this.

Abstract

Emergency call centers serve as the first point of contact between the caller and the emergency services. Therefore, it is vital that these call centers operate quickly while providing good quality of service. The main process within an emergency call center consists of two consecutive stages: first triage, followed by dispatching. Triage is about determining the location of the caller and assessing the severity of the accident. Dispatching is the process of coordinating ambulances in the region and assigning them to an accident. Sometimes a call may also lead to one or more follow-up calls, in which call center employees have to coordinate with ambulances and hospitals.

Problem statement At the moment, each of the 21 emergency call centers in the Netherlands use different types of staff combinations for handling calls. Some of them use only generalists, others use call takers and dispatchers instead. Moreover, the cooperation among emergency call centers is limited. Recently, the government introduced plans for merging emergency call centers and standardizing the processes within emergency call centers, but the exact impact of these plans on the *performance* and the *efficiency* of emergency call centers of this plan is yet unknown. Therefore, the impact of staffing on the performance and efficiency of emergency call centers for *ambulance services* has been analyzed using mathematical models and techniques. The performance is measured using the fraction of calls that is answered within a certain time frame and the efficiency is measured using the minimum required number of centralists. The main goal of this report is to provide a better understanding of the impact of *staffing* on the efficiency and performance at emergency call centers in the Netherlands. The results presented in this report are meant to support decisions at both tactical and strategic levels. Moreover, a user-friendly simulation tool has been developed for evaluating the impact of different staffing configurations on the performance and efficiency. This tool is able to analyze different scenarios for merging call centers and can also analyze the sensitivity of the performance indicators with regard to the input parameters.

Approach In order to provide a better understanding of the relation between staff configurations and the performance, both discrete event simulation and analytical models have been used. It turns out that analytical methods provide quite good approximations for measuring the performance of A1 and A2 calls. However, it is much more difficult to approximate the performance of B-calls and follow-up calls. The applicability of analytical methods is therefore limited and they turn out to be more suitable for doing quick capacity calculations. Therefore, simulation has been used for analyzing the impact of staffing on the efficiency and performance of emergency call centers.

Robustness of model A sensitivity analysis has been conducted in which the robustness of the simulation model with regard to its input parameters is tested. The goal of this analysis is to test the applicability of the model when the input parameters are under- or over estimated. It turns out that the performance of the call center for A1 and A2 calls is not very sensitive to deviations in the input parameters, because these calls yield the highest priority. However, B-calls and follow-up calls are much more sensitive, because their waiting time is dependent on call types with higher priority. Also, the accuracy of the simulation on the performance has also been tested. It turns out that the simulation converges quickly after simulating only a few days. The reason for this quick convergence is that emergency call centers are in general lightly-loaded systems.

Impact of staffing configurations on the performance With the use of simulation, three different staff configurations have been analyzed: one with function differentiations (i.e., call takers and dispatchers), one with only generalists (i.e., a multi-skilled employee) and a configuration with call takers, dispatchers *and* generalists. The impact of each of these configurations on the performance and utilization rate has been analyzed and the results show that configurations with only generalists

perform better for low call volumes. Using call takers, dispatchers *and* generalists avoids having bottlenecks at either the triage or the logistic part, because the generalists handle the overflow coming from call takers and dispatchers. Also, this staff configuration is more beneficial than using only generalists, because generalists have higher wages.

Merging emergency call centers Merging call centers is another topic that is discussed in this report. Recently, the government created a plan for merging the current 21 call centers into 10 call centers. However, the impact of this plan on the performance and efficiency of emergency call centers remains unclear. Most of the current 21 emergency call centers show very low utilization rates of employees, because the regions are quite small. Three scenarios have been analyzed, including the plan of the government. The other two scenarios concern 2 and 1 call center(s), respectively. If the plan of the government would be implemented, it would lead to a 28% percent staff reduction (for a staff configuration *without* function differentiation). The efficiency can be increased even more by merging to a nation-wide call center and the reduction in staff would then be 62%. One of the remarkable results is that the difference in staffing levels between a call center with and without function differentiation decreases when more call centers merge. The main reason for this result is that the call volumes are too low for using function differentiation in the current situation. A generalist is much more efficient for handling low call volumes, because he/she is more flexible. However, as more and more call centers merge, the utilization rate of call takers and dispatchers increases and function differentiation becomes more efficient. One of the disadvantages of merging call centers is that the aspect of regional knowledge is lost. In this report, a model has been provided that is able to preserve this knowledge to some extent. In this model, the merged call center has employees of the various subregions. A call is preferably handled by an employee of the same region, but when these employees are busy, the call is transferred to an employee of another region. The results show a significant increase in efficiency, but it is less than in the model without taking regional knowledge into account.

1 Introduction

Emergency call centers serve as the first point of contact between the caller and the emergency services and therefore it is vital that these call centers operate quickly while providing good quality of service. The main process within an emergency call center consists of two subsequent stages: first triage, followed by dispatching. Triage is about determining the location of the caller and assessing the severity of an accident. Dispatching is the process of coordinating ambulances in the region and assigning them to an accident. It may happen that a follow-up call occurs, in which coordination is provided between ambulances and emergency call centers.

1.1 Background information

At the moment, there are no clear regulations concerning service level requirements for handling emergency calls in the Netherlands. However, there are some rules about the minimum response times. Incoming calls can be classified into three urgency levels: A1, A2 and B. The A1 urgency is highly urgent and it concerns a life threatening situation. A2 is also urgent, but the situation is not life threatening. B is a code that indicates an order for the transportation from and to healthcare institutions. For A1 calls, the ambulance has to arrive within 15 minutes at the location of an accident and for an A2 call this timespan is 30 minutes [26]. These rules are quite general and not suitable for measuring the performance of an emergency call center. Performance is quite a general definition and can be looked at from a quantitative or qualitative point of view. The former mainly concerns about training, experience and knowledge of the centralists while the latter mainly concerns the waiting time, blocking probability and the utilization rate of employees. In this report, the focus will be on the *quantitative performance indicators* of emergency call centers.

In the call center literature, quite many articles have been published on the qualitative analysis of the performance in emergency call centers. These articles are mainly about methods for improving the triage and dispatching processes. However, with regard to the quantitative side of emergency call centers, only few articles have been written. G. Chin and J. Cok [1] provide a simulation model for evaluating the effectiveness of the current staffing levels at the Toronto Emergency Medical services call center. Barros et al. [23] provide a model using a Generalized Stochastic Petri Net (GSPN) for determining the capacity at emergency call centers in Brazil and include the abandonments and retrials for waiting calls. In this article recommendations are given for the maximum queue length and the number of employees but no attention is paid to the distribution of the waiting time. Finally, J. Puts provides a simulation model for measuring the performance of emergency call centers in the Netherlands [21]. Puts provides recommendations with regard to impact of different staff combinations on the waiting time time. Later on, this model has been extended by G.J. Kommer and M. van Buuren [17] and the model used this report is based on that model.

At the moment, each of the current 21 emergency call centers operate differently and their cooperation is limited. Recently, the government introduced plans for going to 10 emergency call centers and standardizing the processes within emergency call centers. The exact impact of these plans on the *performance* and the *efficiency* of emergency call centers of this plan is yet unknown. General call center theory could help answering these questions, but emergency call centers have some distinct features which make it hard to apply. These features include the presence of follow-up traffic, different types of employees, different priorities and inhomogeneous call durations. Therefore, discrete event simulation is used, in which the following three scenarios have been analyzed:

1. 10 call centers (plan of the government)
2. 2 call centers
3. 1 call center

An issue when emergency call centers are being merged, is the loss of regional knowledge. Regional knowledge can be crucial when answering 112 calls that come in by mobile phone. The location of these calls cannot be determined automatically, so the centralist has to ask for it. When call centers are merged, it gets harder for centralists to learn all the regional aspects and therefore the quality and efficiency of handling calls decreases. Modern techniques such as mapping software can speed up this process, but still much 'soft knowledge' is not captured entirely. Regional knowledge also plays a role for the logistic processes (i.e., dispatching and handling follow-up calls), because the centralist is less aware of the location of ambulances in the region and therefore it takes longer to coordinate ambulances.

1.2 Research questions

The goal of this report is to provide a better understanding of the impact of staffing on the efficiency and (quantitative) performance at emergency call centers in the Netherlands. The main research question of this report is therefore:

'What is the impact of staffing on the performance and efficiency of emergency call centers for ambulance services?'

Sub questions:

- What is the impact of different staff combinations on the performance indicators?
- What is the impact of merging call centers on the efficiency?

A part of this report concerns the minimum required staffing levels to meet certain service level requirements. These staffing levels are *average* numbers, in which practical issues such as shift lengths of employees and fluctuations (i.e., daily, seasonal etc.) in call volumes are not taken into account. However, these numbers still provide a good indication what would happen to the efficiency when call centers merge. The results in this report are therefore meant for supporting decisions on tactical level and strategic level. Examples of such decisions would be:

- 'How much centralists do I need when I merge emergency call centers?' (strategic)
- 'What is the impact of function differentiation on the performance of an emergency call center?' (tactical)
- 'What would be the impact on the performance if the duration of the triage or dispatching could be improved by 10%? (tactical)

1.3 Structure and approach

The structure of this report reflects the approach that has been taken for answering the research questions. First, a model is presented for modeling an emergency call center, followed by a chapter about the data analysis. In Chapter 4, an introduction to discrete event simulation is given and an overview of the simulation tool is provided. It also discusses the accuracy of the simulation as a function of the simulation duration. In Chapter 5, the robustness of the simulation model to its input parameters is evaluated in order to check whether the model is adequate for measuring the performance and efficiency of emergency call centers. In Chapter 6, analytical methods are introduced for approximating the performance indicators and these methods are compared with the simulation results. Chapter 7 provides an overview of the impact of different staffing combinations on the performance indicators. This chapter is followed by an analysis concerning the impact of merging emergency call centers on the minimum required staffing levels. Also, a model that includes regional knowledge is introduced in this chapter. The report ends with a conclusion in which the main results are discussed and recommendations are given.

2 Modeling an emergency call center

This chapter describes the processes in an emergency call center and how they have been modeled. The first section describes the arrival process of incoming calls. The second section provides a description of the triage process. This chapter ends with an description of logistic processes that occur after the triage. Most parts of the model have been developed by Geert Jan Kommer, a health care researcher at RIVM and Martin van Buuren, a researcher at Center of Mathematics and Informatics (CWI). Their model and results are described in 'Modellen referentiekader Ambulancezorg' [17]. The differences between the model of Kommer-Van Buuren and the model presented in this report are the following:

- **Lognormal service durations:** Data analysis has shown that the lognormal distribution provides a better fit relative to the exponential distribution.
- **Interruption of service:** The extension of interupting service when all employees are busy has been added.
- **Priority assignment:** Due to practical constraints, only three priority levels were possible in the model of Kommer-Van Buuren. The model presented in this chapter has no limitations with regard to the priority of calls.
- **Performance measurement:** In the model of Kommer-Van Buuren, only the waiting time in the triage queue was measured but the time spent in the queue for dispatching was not included in the total waiting time.
- **112 as one group**: No distinction is made between 112 calls with medical assistance and 112 calls without medical assistance. Data-analysis has shown that the log-normal distribution is a good fit for all 112-calls.
- **Extended staffing possibilities:** Compared to the model of Kommer-Van Buuren, combinations of call takers, dispatchers *and* generalists are possible in this model.
- **Priority Queue:** In the model of Kommer-Van Buuren, A1-A2 calls and B calls were put in separate queues for triage and logistic services. The model presented here has two priority queues: one for triage and one for logistic calls and the priority of a call determines the position in the queue.

In this report, two different staff configurations have been analyzed: a configuration *with* function differentiation and one *without* function differentiation. In function differentiation, the triage and dispatching are performed by two different types of employees. The employee that performs the triage is called a *call taker* and the employee that does the dispatching and handling follow-up calls is called a *dispatcher*. In a configuration without function differentiation, only one multi-skilled employee is used which can perform both triage and logistic tasks (i.e., dispatching and follow-up calls). From now on, this employee is referred to as a *generalist*.

Figure 1 shows an overview of the major processes that take in a call center with function differentiation. When a call comes in, the so-called triage is started. Triage is the process in which the urgency of the call is determined and in which also medical assistance might be provided. When the triage is completed, an ambulance is dispatched if necessary. After an ambulance has been dispatched, the call is ended but for some calls, follow-up calls might occur.

Figure 2 shows an overview of the processes that take place at a call center with no function differentiation. The processes are still the same, but they are performed by different employees sequentially. The triage is performed by a call taker and the dispatching is done by a dispatcher. Separating these tasks has the advantage that incoming calls can be answered quicker, because the call takers are not occupied with the dispatching anymore. Function differentiation also has a financial benefit because the salary

of a dispatcher is in general lower than the one of a call taker because this task does not require any medical knowledge. A non-measurable effect is that triage employees and dispatchers specialize in their task, such that the quality and efficiency improves. An interesting paper related to this topic is 'A Review of Workforce Cross-Training in Call Centers from an Operations Management Perspective' by Aksin et al. [2]. This paper describes the advantages and disadvantages of cross training and provides several routing policies for call centers with cross trained agents.

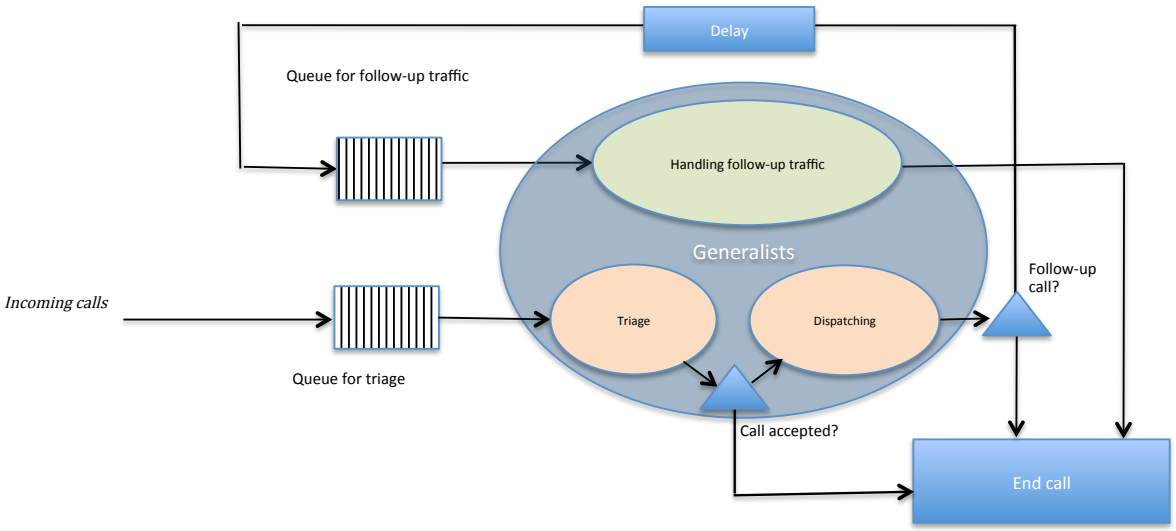


Figure 1: The sequence of events in an emergency call center without function differentiation.

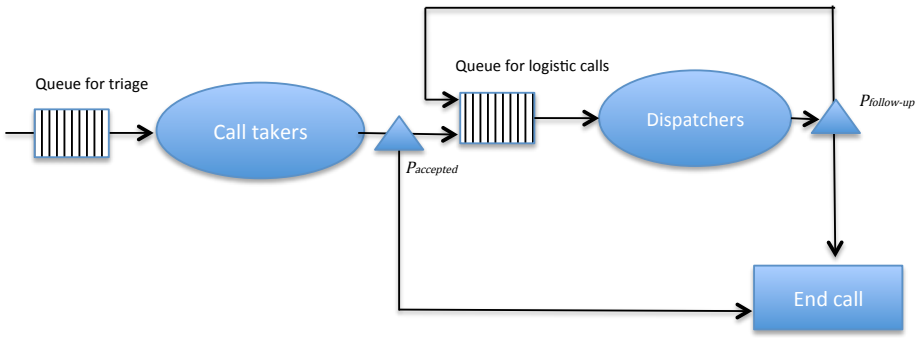


Figure 2: The sequence of events in an emergency call center with function differentiation.

2.1 Arrival process

Not only 112 calls arrive at an emergency call center, but also calls originating from institutions such as police, fire department, hospitals, general practitioners and other healthcare institutions. As mentioned earlier, calls can be classified on their urgency. However, for A-calls the exact urgency (i.e., A1 or A2) has yet to be determined during the triage and therefore have the same level of priority. However, B-calls come in on other trunk lines such that the urgency is known in advance.

The arrival process of an emergency call center can be modeled as an inhomogeneous Poisson process because there is a large population of potential callers, of which each citizen might call with a small but non-neglectable probability [19]. The data analysis performed by M. Calinescu [28] confirms that

the number of incoming calls in a certain time interval has roughly a Poisson distribution. The analysis also showed that the arrival rate varies per hour. In this report, time dependent arrival rates are *not* included and the arrival rates are considered to be constant all day. The arrival process can be denoted in mathematical notation as follows:

λ_{ij}= the arrival rate of group i with urgency j
$N_{ij}(t)$= number of calls of group i with urgency j arrived between time 0 and time t, $t \geq 0$
$N_{ij}(t) \sim Poisson(\lambda_{ij}t)$
$i \in \{112$, politie, brandweer, thuiszorg, psychiatrie, Huisarts/HAP, Overig/verloskundig/GHOR$\}$
$j \in \{A1, A2, B\}$

There is also quite much background traffic among centralists and between emergency call centers. These calls are referred to as 'noise calls' and these calls do not lead to follow-up traffic. For a model with function differentiation, this noise has been equally spread among the call takers and dispatchers.

2.2 Triage and medical assistance

As mentioned earlier, the urgency of incoming calls is classified by three different categories: A1, A2 and B. The A1 code indicates a very urgent and life-threatening situation. A2 indicates an urgent but not life-threatening situation. Class B is the lowest urgency class and is an order for scheduled transportation of patients from and to health care institutions. The process of determining the urgency by a centralist is called 'triage'. Data analysis shows that the duration of the triage is dependent on both the origin (e.g., from the police or 112) and the urgency of a call. When a call comes in from 112, medical assistance ('meldkamer instructie') might be required. In general, centralists have medical knowledge and they have the ability of providing medical assistance to the caller. For example, if someone calls about a person suffering from a hearth attack, the centralist will give instructions for performing CPR.

The duration of the triage and medical assistance varies, because every accident is different. Therefore the duration of the triage (including medical assistance) has been modeled as a stochastic variable. The distribution of the triage duration are drawn from a lognormal distribution. The motivation for this decision is explained in the data analysis (chapter 3).

When all employees are busy and a higher priority call comes in, the call with the lowest priority will be interrupted and put in the queue. When all generalists or call takers are busy and no low priority calls are in service, the call is put in the queue. The concept of redials and abandonments of calls is not included in the model, because it is reasonable to assume that callers will keep waiting due to the high urgency of the calls. This might not apply for B-calls, but in order to keep the model simple, it is also assumed that these calls do not abandon or redial. The queue for triage has been modeled as an infinite priority queue, in which the position of the call in the queue is determined by its priority. This priority is indicated by a number between 0 and 10, in which 10 is the highest priority. Table 1 shows the priority of each caller group. The calls urgency A1 and A2 have been given equal priority levels, because the urgency of these calls has yet to be determined when a call comes in. B-calls come in on a separate line, which means that the urgency is already known. After performing triage, the priority is changed to either an A1 or A2 urgency. The 112 calls and unknown calls have been given the highest priority because these callers (usually) have no medical knowledge. The police, fire department have in general more medical knowledge compared to 112 callers. Calls originating from healthcare institutions have been given a lower priority because usually there is quite much medical knowledge present at these institutions.

Call group	A1	A2	B
112 Met Instructie	10	10	1
112 Zonder Instructie	10	10	1
Unknown calls	10	10	1
Brandweer	9	9	1
Politie	9	9	1
Ziekenhuis/MKA	8	8	1
Huisarts/HAP	8	8	1
Psychiatrie	8	8	1
Overig/Verloskundig/GHOR	8	8	1
Thuiszorg	8	8	1

Call type	**Priority**
Noise triage	0
Noise logistic	0

Table 1: Priority level for each caller group.

2.3 Dispatching and follow-up traffic

Dispatching is the process of sending an ambulance to the location of the caller. The caller never communicates with the dispatcher *directly*, because all communication goes through the call takers. The call takers sends requests (electronically) to the dispatcher. In the rest of this report, these requests will be referred to as: 'calls that require dispatching'. It might occur that one accident leads to multiple calls. In this case only one ambulance is dispatched to the accident. In other words, only one of these calls leads to the dispatching of an ambulance. To model this process, a Bernoulli process is used in which a call is accepted with probability $p_{accepted}$ and denied with probability $1-p_{accepted}$. If a call is accepted, an ambulance is dispatched, if not: the call is ended. The duration of the dispatching process is assumed to have an exponential distribution with an expected duration of 1 minute. The process of dispatching can also be interrupted when a call with a higher priority arrives. The interrupted call is put back in the queue for calls that require dispatching and follow-up calls. If a call is in the dispatching phase, it keeps the same priority as has been determined during triage. This way, calls with high priority (e.g., 112 calls) still yield priority in the logistic part.

After the dispatching of an ambulance, the call is ended. After a while, it might happen that a follow-up call takes place. These calls mainly concern the coordination among emergency call centers, ambulances and hospitals. The duration of these calls is also assumed to have an exponential distribution with an expected duration of 4 minutes. It must be noted that not every dispatched ambulance leads to a follow-up call. In order to model this, a Bernoulli process is used in which with probability $p_{follow-up}$ a follow-up call occurs. The priority of these follow-up calls is lower than incoming calls (a priority level of 7).

Conclusions

The main findings of this chapter are:

- Model is an expansion of work by G.J. Kommer and M. Van Buuren.
- Both the model with *call takers* and *dispatchers* and the model with only *generalists* have been presented.
- The arrival process is modeled as a homogenous Poisson process.
- The triage durations of each group have been modeled with a log-normal distribution, the duration of dispatching and follow-up calls with an exponential distribution.
- The service policy is *pre-emptive resume*: low-priority calls get interrupted when a high priority call arrives.

3 Data analysis

This chapter provides estimates for the input parameters, using 20.000 calls at 'Regionale Ambulance Voorziening Utrecht' (RAVU) in the time period of 1-3-2011 until 30-6-2011. One of the problems of determining these parameters, is the way of how call information is stored by emergency call centers. The ARBI dataset contains time stamps of certain moments in an emergency call center. However, this dataset does not contain any information about whether a call led to an actual dispatching of an ambulance. This information is stored in another database, which makes linking these two datasets difficult. In order to get estimates of the arrival frequencies and service durations, a trick has been used to link these datasets. This trick is based on linking both datasets with the use of the time stamps. For example, when a call came in at noon, the call must have led to the dispatching of an ambulance around the same time. However, it can't be said with certainty that this call really led to the dispatching of an ambulance. For 8.870 calls, two or more ambulance rides could be assigned to a call. For each call, only the first ambulance dispatching has been chosen. Using this approach, 91.8% of the ambulance rides could be linked to the ARBI dataset [17].

3.1 Arrival frequencies

Analysis of the ARBI data shows that incoming calls come from nine different caller groups: 112 (the emergency telephone number in the Netherlands), the police, the fire department, thuiszorg, psychiatrie, Huisarts/HAP and Overig/verloskundig/GHOR. The last group is a mixed group which has similar arrival and service durations. The origin of some calls could not be determined and these calls have been added as a residual group called 'unknown'. As mentioned before, the noise calls concern traffic among centralists and between emergency call centers. Tables 2 and 3 show the frequency of calls for each group and it shows that the noise group is significantly higher compared to the other groups.

Origin	A1	A2	B
112	0.83	0.58	0.04
Brandweer	0	0.01	0
Huisarts/HAP	0.95	0.75	1.26
Unknown	0.12	0.37	0.11
Overig/Verloskundig/GHOR	0.05	0.07	0.09
Politie	0.09	0.11	0
Psychiatrie	0	0.04	0.09
Thuiszorg	0	0.01	0.26
Ziekenhuis/MKA	0.13	0.66	1.38
Total	2.18	2.60	3.24

Table 2: Arrival frequencies (calls/hr) of the different caller groups.

Origin	Calls/hr
Noise Triage	12.88
Noise Logistic	12.88
Total	25.76

Table 3: Arrival frequencies (calls/hr) for noise calls.

To give an impression about the time dependency of the call streams, a brief data analysis has been done with regard to the time dependency of the arrival process. However, in this report only average arrival rates are considered (i.e, time independent). For an extensive analysis about forecasting call

streams in emergency call centers, the papers of M. Mahfoud [15] and Lewis et al. [16] provide an overview of various techniques. Figures 3 and 4 display the arrival rate per weekday for A1, A2 and B calls respectively. It shows that during the night, the arrival rate is lowest although still some 112 calls come in. From 7:00 the arrival increases and the maximum arrival rate is reached between 11:00 and 13:00. The average arrival rate on Saturdays and Sundays appears to be slightly lower compared to the other days.

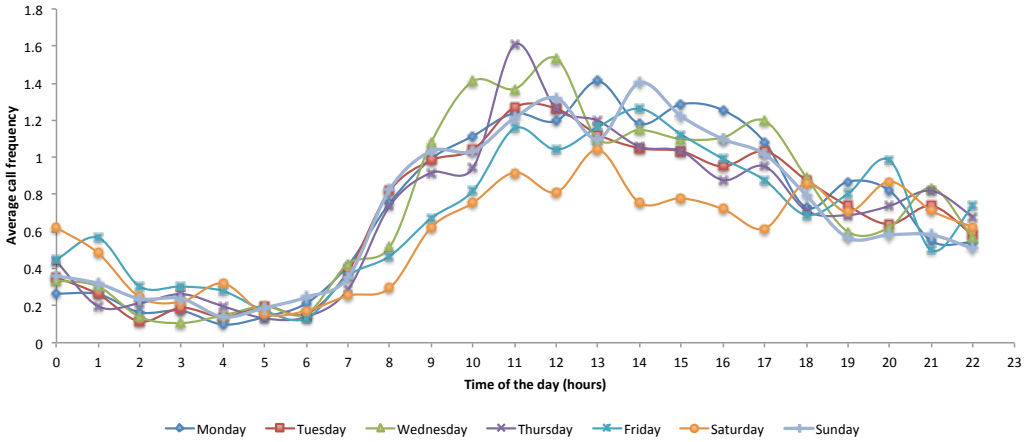


Figure 3: Arrival frequency per hour of A1,A2 calls for each weekday.

Figure 4 shows that the number of B-calls is significantly lower on Fridays and Saturdays. The reason for this phenomenon is that surgeries don't take place in the weekend. On Sunday however, patients have to be transported by ambulance such that they can undergo surgery on Monday.

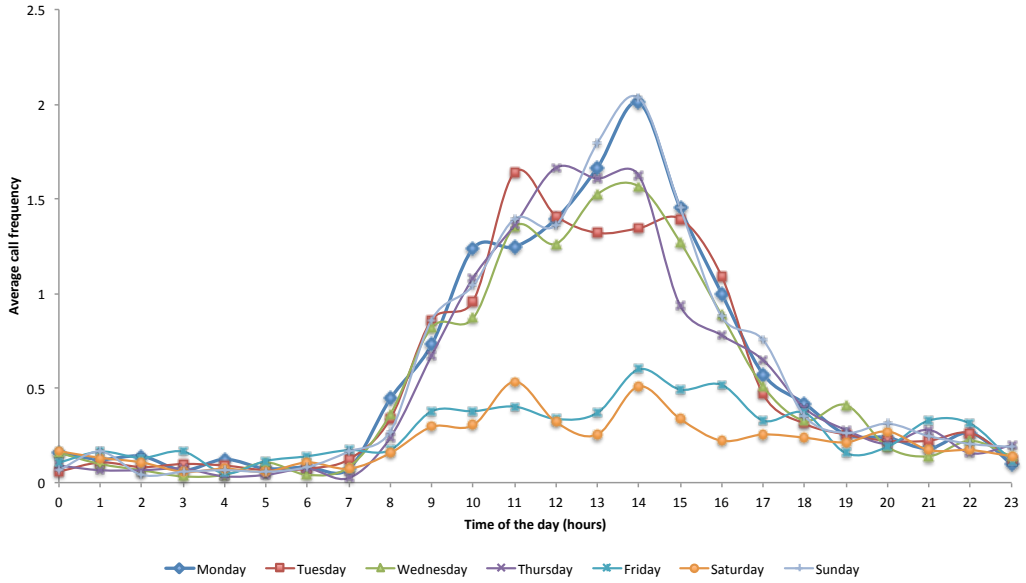


Figure 4: Arrival frequency per hour of B calls for each weekday.

3.2 Call durations

Table 4 shows the average call durations for the nine different caller groups. This service duration measures only the duration of the triage and does not include the duration of dispatching and handling follow-up traffic. The average durations for dispatching of an ambulance and handling a follow-up call could not be derived from the data and therefore these values have been estimated by experts in the field. The average duration for dispatching is estimated to be 1 minute and handling a follow-up call takes 4 minutes. Both the duration of dispatching and handling follow-up calls have been modeled with an exponential distribution.

Origin	A1	A2	B
112	1.7	1.7	1.8
Brandweer		1.8	
Huisarts/HAP	1.6	1.9	1.9
Unknown	1.5	1.6	1.5
Overig/Verloskundig/GHOR	1.5	1.7	1.8
Politie	1.4	1.3	2.1
Psychiatrie	1.4	2.5	2.4
Thuiszorg	2.9	1.7	1.6
Ziekenhuis/MKA	1.5	1.7	2.6

Origin	Average call duration (min)
Noise Triage	0.98
Noise Logistic	0.98

Table 4: Duration of the triage (in minutes) for the different call groups.

In the field of call center modeling, it is common to model call durations with either an exponential distribution or with a lognormal distribution. Figure 5 shows the general shape of both distributions. Due to its bell-shaped curve, the log-normal distribution looks similar to the normal distribution, but it is skewed to the left. The exponential distribution is a relative simple distribution, with only one parameter and it is not bell-shaped. For theoretical purposes, it is convenient to use this distribution because it has only one parameter and it has the memoryless property [19]. However empirical studies have shown that call durations are better approximated by using the log-normal distribution. Bolotin describes in the article 'Telephone Circuit Holding Time Distributions' [25] that the logarithm is a good way to represent the human perception of time, because time and time-value are not linearly related. For example, humans perceive the difference between 40 seconds lasting conversations and 30 seconds conversations just the same as the difference between 4 minutes and 3 minutes. The logarithm function values small values higher than big values and therefore it matches with the human perception of time. Due to this psychological phenomenon, the log-normal distribution provides a good fit for call durations in general.

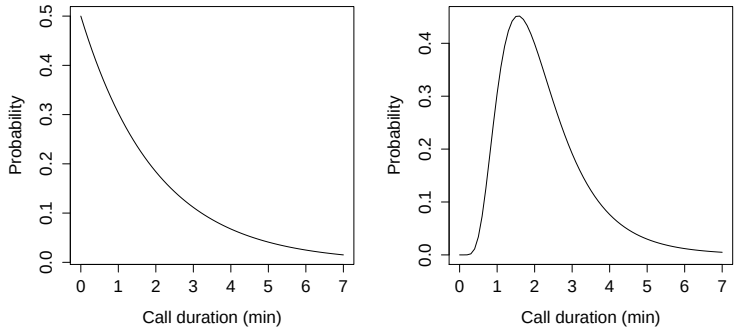


Figure 5: Left: exponential distribution, Right: log normal distribution.

Data analysis has been performed in order to determine the best fitting distribution to the call durations. The analysis is performed on the groups that have the highest arrival frequencies: huisarts/HAP, calls from hospital, 112 calls and noise calls. Each of these groups have been fitted with a log normal distribution and an exponential distribution.

Huisarts/HAP Figure 6 displays the distribution of the call durations for each urgency level (A1, A2, B) respectively. The red line represents the fit of the log-normal distribution and the blue line represents the fit of the exponential distribution. The figure shows that the log-normal distribution provides a better fit than the exponential distribution.

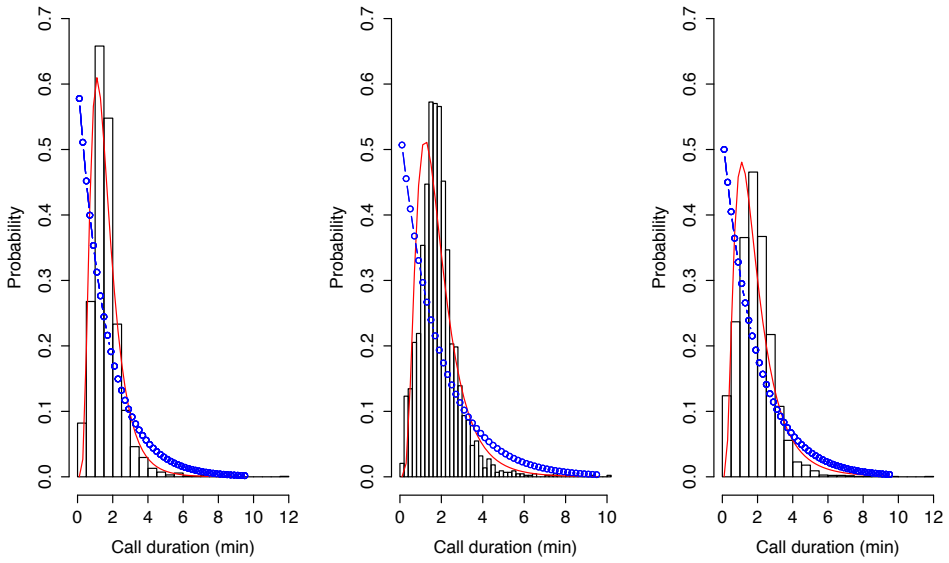


Figure 6: Left: A1 calls - Center: A2 calls - Right: B calls.

112 calls The second group that has been investigated are the 112 calls. Figure 7 shows the distribution for 112 calls for the A1 and A2 urgencies respectively. The red line represents the fit of the lognormal distribution and the blue line represents the fit of the exponential distribution. Also for these calls, the log-normal distribution provides a better fit.

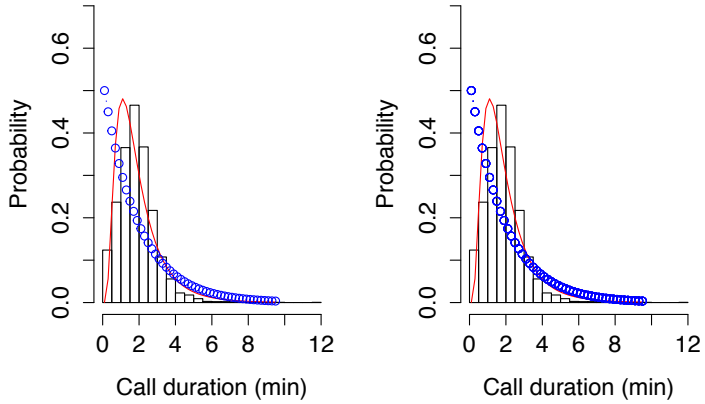


Figure 7: 112 calls , Left graph: A1-calls, Right graph: A2 calls.

Hospital calls Figure 8 shows the distribution of the call duration for calls originating from hospitals. The red line represents the fit of the lognormal distribution and the blue line represents the fit of the exponential distribution. Also for calls from the hospitals, the log-normal distribution provides a better fit.

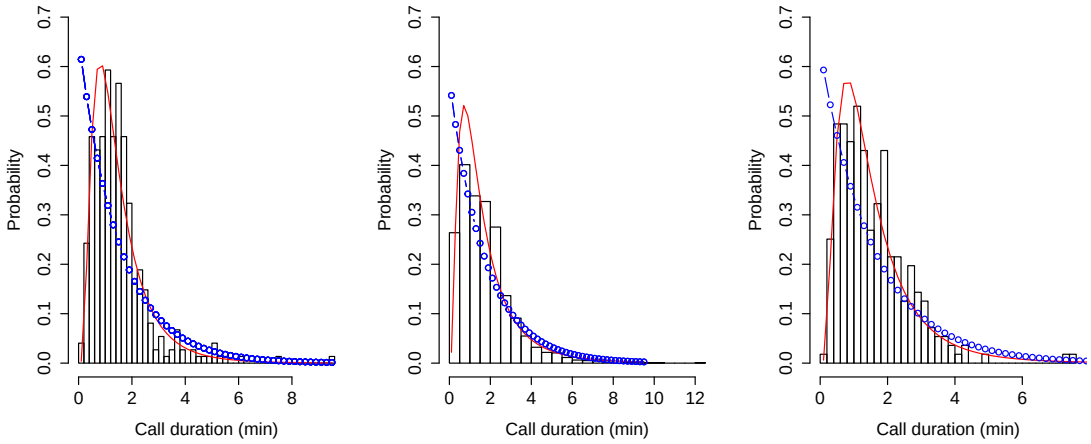


Figure 8: Calls from hospitals, Left: A1 calls - Center: A2 calls - Right: B calls.

Noise calls Figure 9 shows the distribution of the call duration for noise calls. Also here, the red line represents the fit of the log-normal distribution and the blue line represents the fit of the exponential distribution. Also for noise calls, the log-normal distribution provides a better fit.

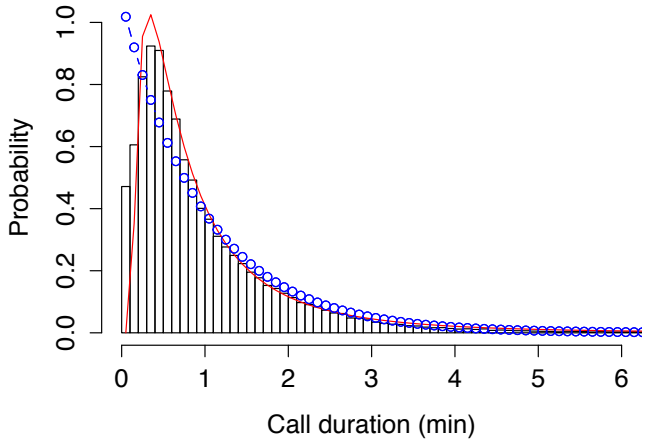


Figure 9: Distribution of call durations for noise calls.

Figures 6 to 9 show that the log-normal distribution is a reasonable fit. In order to statistically validate the assumption of log-normally distributed call durations, a Kolmogorov-Smirnov test has been performed to test the equality of distributions. The Kolmogorov-Smirnov is a non-parametric test which compares the empirical cumulative distribution of the data with the one of the theoretical distribution [13]. In this case, the theoretical distribution is the log-normal distribution with parameters:

μ (location), σ (scale) and their maximum likelihood estimators are given in (1) and (2).

$x_{hij} = h^{th}$ call duration of group i with urgency j

$n_{ij} =$ number of observations in group i with urgency j.

$$\hat{\mu}_{ij} = \frac{1}{n_{ij}} \sum_{h=1}^{n_{ij}} \ln(x_{hij}) \tag{1}$$

$$\hat{\sigma}_{ij} = \sqrt{\frac{1}{n_{ij}-1} \sum_{h=1}^{n_{ij}} (\ln(x_{hij}) - \frac{1}{n} \sum_{h=1}^{n_{ij}} \ln(x_{hij}))^2} \tag{2}$$

$$\hat{\lambda} = \frac{1}{n_{ij}} \sum_{h=1}^{n_{ij}} x_{hij}$$

$i \in \{112,$ politie, brandweer, thuiszorg, psychiatrie, Huisarts/HAP, Overig/verloskundig/GHOR$\}$

$j \in \{$A1, A2, B$\}$

The following expressions describe the hypotheses and the test statistic for the Kolmogorov-Smirnov test. The null hypothesis is displayed in (3) and when the null-hypothesis is rejected, it can be said that the call duration of that group is *not* log-normally distributed. The test statistic is denoted by $D_{n_{ij}}$ and it measures the biggest distance between F_{ij} and F_{ij0}.

$F_{ij0} =$ log-normal distribution with parameters $\hat{\mu}_{ij},\ \hat{\sigma}_{ij}$.

$F_{ij} =$ distribution of the data of group i with urgency j.

$$\begin{aligned} H_0 &:= \quad F_{ij} = F_{ij0} \\ H_1 &:= \quad F_{ij} \neq F_{ij0} \end{aligned} \tag{3}$$

$$D_{n_{ij}} = \sup_{x_{hij}} \|F_{n_{ij}}(x_{hij}) - F_{ij0}(x_{hij})\| \tag{4}$$

$h \in \{1, \ldots, n_{ij}\}$

$i \in \{112,$ politie, brandweer, thuiszorg, psychiatrie, Huisarts/HAP, Overig/verloskundig/GHOR$\}$

$j \in \{$A1, A2, B$\}$

Tables 5 and 6 show the p-values of the Kolmogorov-Smirnov test for each caller group. Some groups have less than 30 observations and the p-values for these groups are omitted and they are indicated by 'N/A'. When using a significance level of 0.05, the null hypothesis is rejected for a few groups. For example, huisarts/HAP group with an A1/A2 urgency have a p-value equal to zero, which means that these call durations are *not* lognormal distributed. However, Graphs 6 to 9 show a good fit of the log-normal distribution. These are contradicting results, but they can be explained by looking at number of observations for each caller group. (see Tables 5 and 6). For groups with many observations, the statistical power of the test becomes large and there is more evidence to either reject or accept the null-hypothesis. For groups with less observations, the statistical power is lower and the null hypothesis is rejected less quickly. Although some null hypotheses have been rejected, assuming log-normal distributed call durations for these groups is still a good approximation (as can be seen in Figures 6 to 9).

	Frequencies			P-values		
	A1	A2	B	A1	A2	B
112	2417	1712	116	0	0	0.063
Brandweer	13	18	0	N/A	N/A	N/A
Huisarts/HAP	2778	2192	3699	0	0	0
Onbekend	348	1096	328	0.064	0	0.35
Overig	160	193	259	0.493	0.065	0.011
Politie	268	322	4	0.078	0.326	N/A
Psychiatrie	14	114	267	N/A	0.003	0
Thuiszorg	9	31	770	N/A	0.7	0
Ziekenhuis/MKA	371	1933	4039	0.082	0	0

Table 5: Frequency of different caller groups and the p-values of the Kolmogorov-Smirnov test.

	Frequency	P-value
Noise calls	89122	0.0

Table 6: Frequency and p-value of noise calls.

3.3 Other parameters

There are a few other model parameters that play a role in an emergency call center. These parameters are:

- **Fraction of rejected calls**: If an accident happens somewhere, multiple calls may arrive from citizens. A part of these calls is already filtered by the 112 call center of the KLPD in Driebergen, but it is still possible that these calls arrive at the ambulance call center. Therefore, a part of the incoming calls is rejected because an ambulance has already been dispatched. Data analysis showed that 10% of the calls do not lead to the dispatching of an ambulance.
- **Fraction of incoming calls that lead to follow-up traffic**: Not every incoming call leads to follow-up traffic. It has been estimated by experts that about 11.1% of the incoming calls leads to follow-up traffic.
- **The time gap between dispatching and follow-up traffic**: A follow-up call occurs some time after an ambulance has been dispatched. This delay could not be derived from the data, but the average delay has been estimated to be 25 minutes by experts in the field.

Conclusions

The main findings of this chapter are:

- Data analysis shows that call durations vary per origin and urgency.
- The distribution of the triage duration is best fitted by the log-normal distribution.
- The fraction follow-up traffic and the fraction of denied calls has been estimated by experts in the field.

4 Simulating an emergency call center

This chapter describes 'ECCSIM' (Emergency Call Center simulator), an emergency call center simulation tool that has been developed for supporting decisions related to staffing in emergency call centers. First, an introduction to discrete event simulation is given, followed by a description of the features of ECCSIM. This chapter ends with a section that demonstrates the accuracy of the simulation tool.

4.1 Introduction to simulation

Simulation has proven to be quite an effective instrument over the years and is used more and more because computational power is increasing. With the use of simulation, real-world processes are imitated and the performance metrics are measured subsequently. Simulation is used when the processes become too complex to be solved by analytical methods. However, simulation can be less traceable and it can be time consuming to build. Mehrotra et al. describe three different purposes of how simulation is used within the call center industry [22]:

1. **Traditional Simulation Analysis:** A simulation model is built to analyze a specific operation, with inputs obtained from a variety of data sources.
2. **Embedded Application - routing:** Many of the leading ACD (Automatic Call Distributing) and CTI (Computer Telephony Integration) applications include a routing simulation to provide insights to routing design engineers about the impact of different decisions.
3. **Embedded Application - Agent Scheduling:** Creating an optimal schedule for center agents. This becomes (even more) difficult when both agents and calls are non-homogeneous.

Emergency call centers have some unique properties which make it difficult to apply analytical methods. The presence of multi-priority call types, follow-up calls, multi-type employees and the different call durations are a few aspects that make applying analytical methods difficult. In most the articles that have been written on emergency call center capacity planning, simulation is used instead of analytical methods. One of the more recent and elaborate papers is 'Simulation Modeling for Staff Optimization of the Toronto Emergency Medical Services Call Centre' by G. Chin et al., who provides a simulation model for evaluating the effectiveness of the current staffing policy [1]. This paper also provides arguments for when simulation should be used instead of queueing theory.

The emergency call center has been simulated using discrete event simulation. Discrete event simulation is based on so-called 'events', such as an arriving call or a service completion. Typically, each event triggers other events. Typically, simulation takes some time to run. The duration of a simulation can be measured in multiple ways, such as the number of days and the number of calls that have been processed. In this report, the first option has been chosen because this is more intuitive than the number of calls. A problem occurs when starting the simulation, because then the call center is empty. Two solutions are possible: either start with a representative number of calls in the call center or start with an empty call center and start recording after some time. This start-up period of the latter solution is also called 'warm-up' period in literature. Due to its simplicity, the warm-up period has been chosen for simulating an emergency call center.

4.2 Features of ECCSIM

ECCSIM has been designed for emergency call center management in order to provide support for making decisions on both tactical and strategic level with regard to the following aspects:

- The impact of staffing levels and configurations on the performance indicators.
- The impact on performance and efficiency for when two or more call centers merge.

ECCSIM has been written in Java, which ensures great speeds compared to simulation software such as Rockwell Arena. Moreover, a friendly user interface has been developed that can be used by virtually everyone.

ECCSIM consists of the following modules:

- Emergency call center simulation
- Scenario analyzer
- Sensitivity analysis

Each of these modules have specific functions and these are explained in the next subsections.

4.2.1 Emergency call center simulation

Figure 11 shows the user interface of ECCSIM. The main screen enables the user to manually configure the parameters and the staffing configuration. The input is provided using a wizard of three steps. The first step requires users to select the regions in which they are interested. Simultaneously, the program provides an estimate of the number of calls per day, the area size and the total number of citizens living in that region. The second step is for changing the average arrival rates and service durations for each caller group.

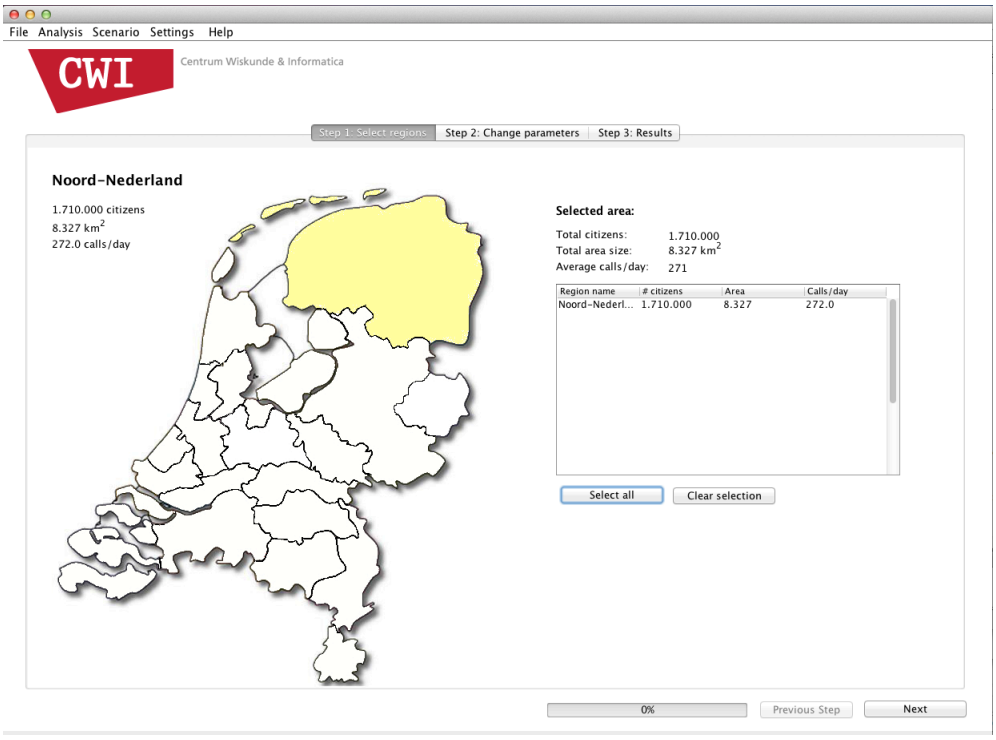


Figure 10: First page of wizard: selecting the regions.

Figure 11 shows the last page of the wizard and it shows the most important processes within an emergency call center. This page enables the user to change the staffing configuration and other parameters. When the simulation finishes, the program displays the results of the simulation in the picture. An explanation of each input/output field is given under the picture.

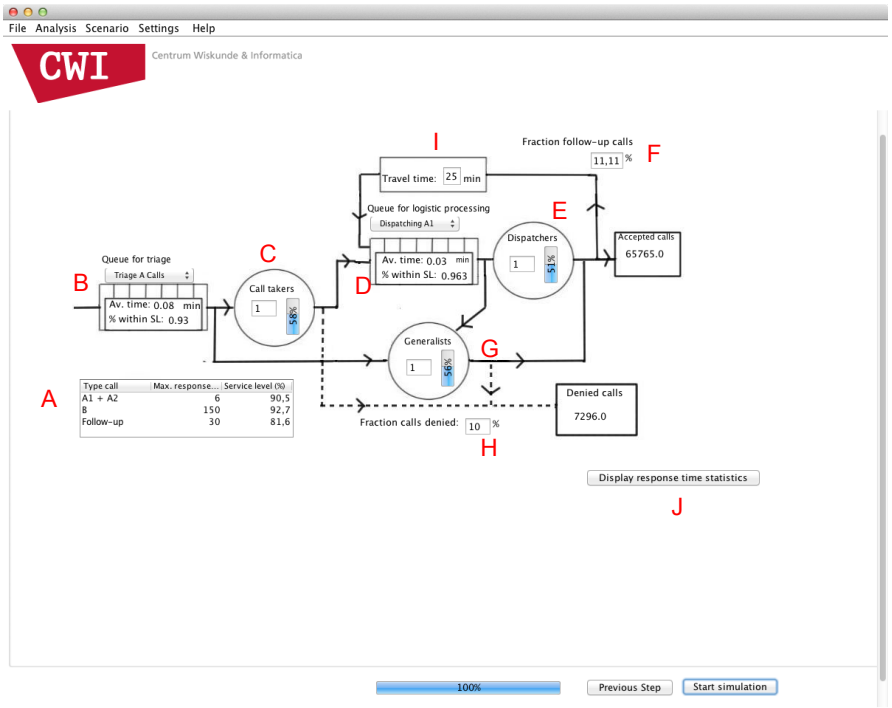


Figure 11: Third page of wizard: results of simulation.

Legend:

A - Output: Table with service levels for each call type.

B - Output: Average waiting time and service level for the triage part.

C - Input/output: Number of call takers and utilization rate of call takers.

D - Output: Average waiting time and service level for the logistic part.

E - Input: Number of dispatchers and the utilization rate of dispatchers.

F - Input: Fraction of follow-up calls.

G - Input: Number of generalists and the utilization rate of generalists.

H - Input: Fraction of triage calls that is denied.

I - Input: Time gap between dispatching and a follow-up call.

J - Output: Button for opening an overview of the waiting time statistics.

4.2.2 Scenario analyzer

Another module of ECCSIM is the scenario analyzer. At the moment there are plans for merging call centers, but the impact of these plans on staffing levels and performance is still unclear. This module allows the user to determine staffing levels for various predefined scenarios or by creating a custom scenario. The following three scenarios are predefined: (as described in Section 8.2).

1. Scenario 1: 10 call centers (the plan of the government)

2. Scenario 2: 2 call centers
3. Scenario 3: 1 call center (nationwide call center)

Figure 12 shows the user-interface for selecting/creating a scenario.

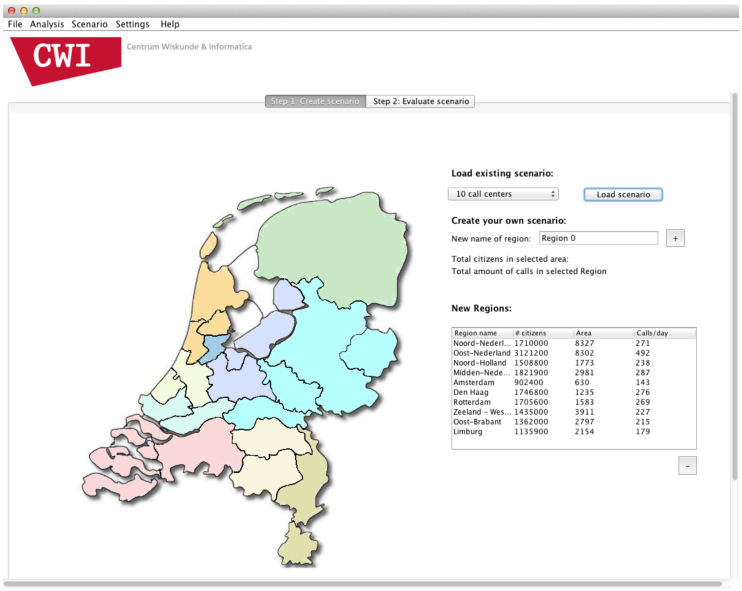


Figure 12: User interface of the scenario analyzer.

Figure 13 shows the output of the selected scenario. The explanation of the letters is explained in the legend on the next page.

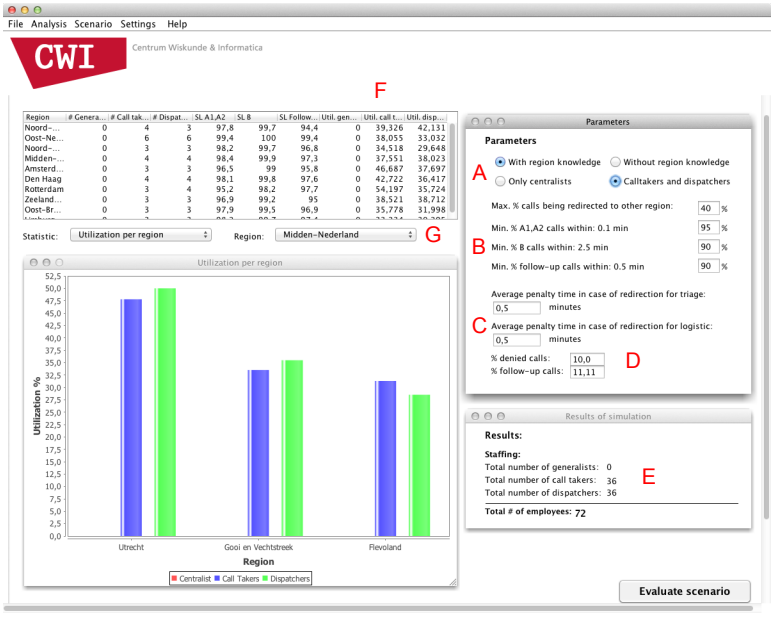


Figure 13: User interface of the scenario analyzer.

Legend:

A - Input: Choose whether regional knowledge should be preserved in the merged call center.

B - Input: Service levels requirements for different call types and the maximum fraction of calls that is allowed to be redirected to an employee of another region.

C - Input: The average penalty time when a triage call/logistic call is redirected to another region.

D - Input: The fraction of triage calls that is denied and the fraction of logistic calls that involves a follow-up call.

E - Output: Minimum required staffing levels.

F - Output: Staffing levels and performance indicators for each region.

G - Output: Plot options.

4.2.3 Sensitivity analyzer

The last module contains a sensitivity analysis which shows the impact of changing arrival rates, service durations and other parameters on the performance indicators. Figure 14 provides an example of the sensitivity analysis and it shows the impact of the arrival rate on the performance indicators for different staff configurations.

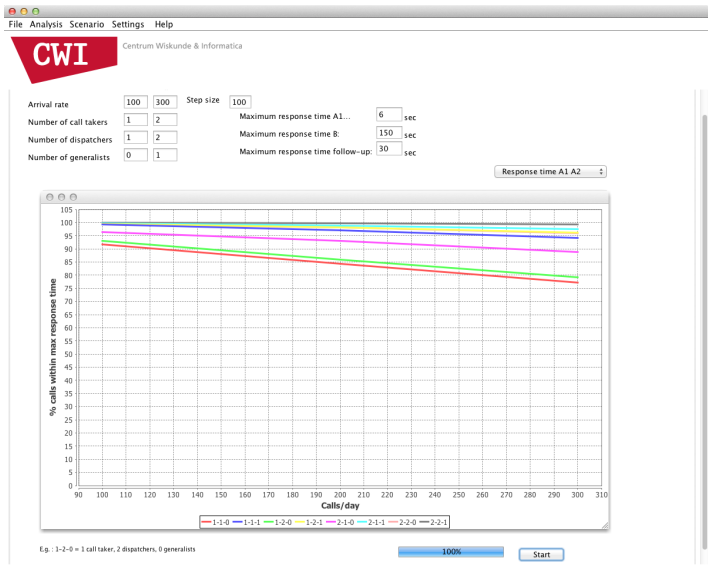


Figure 14: User interface of the sensitivity analysis.

4.3 Accuracy of the simulation

In this section, the impact of the simulation on the performance indicators is analyzed. In order to simulate the emergency call centers, call arrivals and call durations are randomly drawn from a probability distribution. In general, if the simulation duration is set very short, there will be a lot of variation in the output because there are not enough observations. To capture this variance, a 95% confidence interval has been constructed for the performance indicators. The following equations show the mathematical notation for computing the confidence interval. The underlying assumption of this confidence interval is that sample is distributed according to a t-distribution, which is a fair assumption

to make because the performance metrics are continuous variables and the observations are centered around a (unknown) true value. The following expressions show the mathematical notation for a α - confidence interval.

$$\alpha \text{ - confidence interval} = [\bar{X}_j - \tau_{1-(1-\alpha)/2}\frac{\sigma_j}{\sqrt{n}}, \bar{X}_j + \tau_{1-(1-\alpha)/2}\frac{\sigma_j}{\sqrt{n}}]$$

, where:

$\tau_{1-(1-\alpha)/2,n-1}$= $1-(1-\alpha)/2$-quantile of the t-distribution with $n-1$ degrees of freedom.

σ_j = sample standard deviation of the performance metric that has been computed using a simulation of j days.

$\bar{X}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$ for $j \in \{1, 2, \dots\}$, where:

x_{ij} = service level of the i^{th} iteration of a simulation run with a duration of j days.

$i \in \{1, 2, \dots, n\}$, n = sample size

$j \in \{1, 2, \dots\}$

Figure 15 to 17 show the accuracy of fraction of each call type that has been answered within the service level requirement as a function of the simulation length. The results come from simulating a call center with 1 call taker, 1 dispatcher and a call volume of 400 calls/day. The shaded area represents a 95% confidence interval. It confirms that there is a significant variation in the output when the simulation duration is set short. It is worth noticing that just a few days of simulation are needed to get a relatively high accuracy. The reason for this effect is that emergency call centers are in general lightly loaded systems. The fluctuations may seem big at first sight, but the vertical axis is centered at a small range. For the service level of follow-up calls, it takes longer converge because only 11.11% of the calls lead to a follow-up call (so there are less observations).

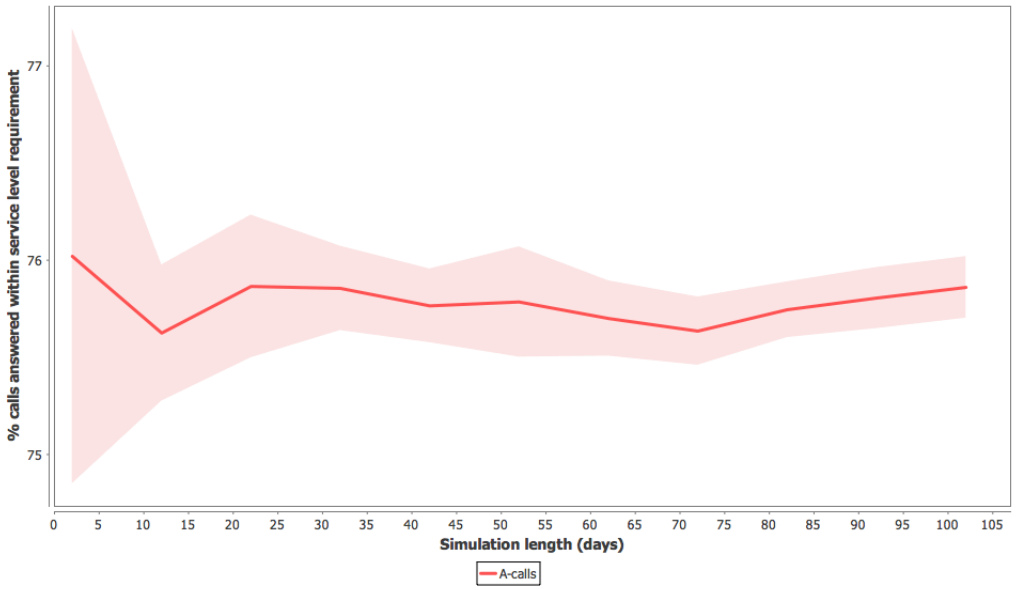


Figure 15: Fraction of A-calls with a waiting time less than 30 seconds as a function of the simulation length. The shaded area shows a 95% confidence interval.

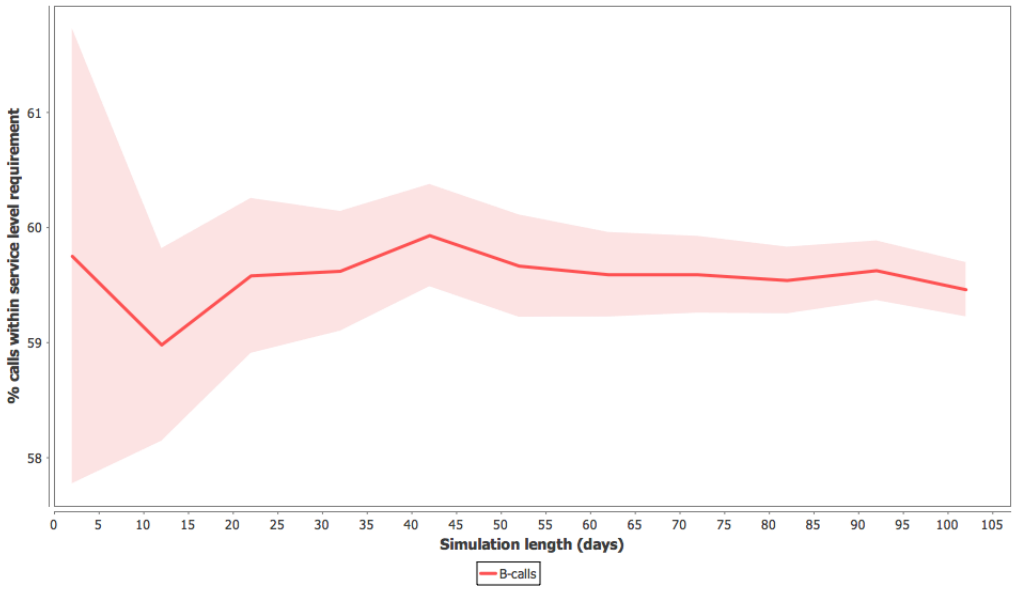


Figure 16: Fraction of B-calls with a waiting time less than 30 seconds as a function of the simulation length. The shaded area shows a 95% confidence interval.

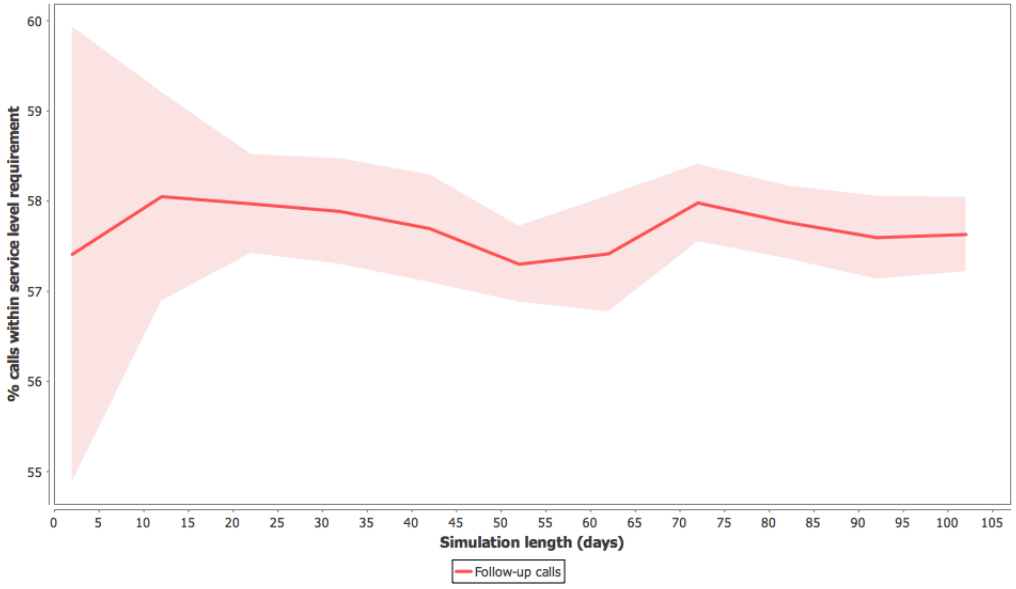
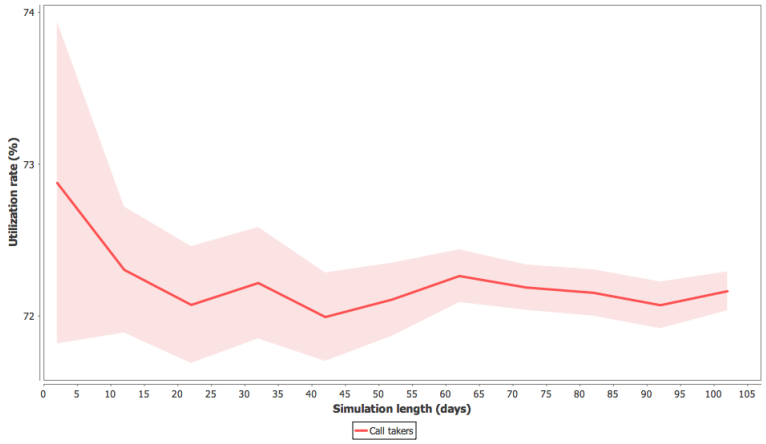
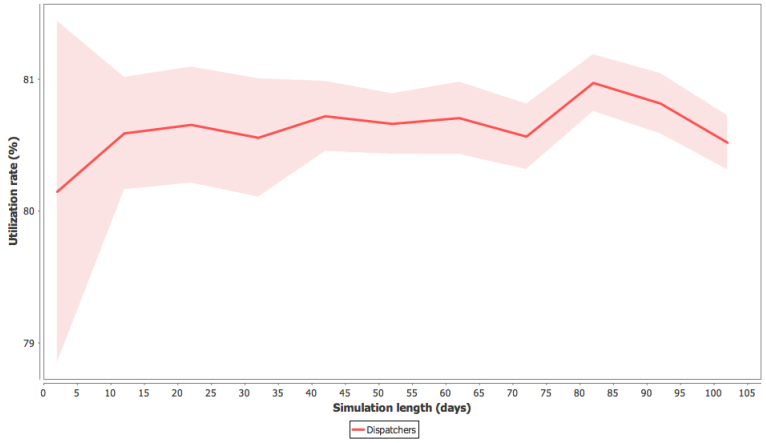


Figure 17: Fraction of follow-up calls with a waiting time less than 30 seconds as a function of the simulation length. The shaded area shows a 95% confidence interval.

Figure 18(b) shows the utilization rate of call takers and dispatchers as a function of the simulation length for a call center with 1 call taker, 1 dispatcher and 400 calls per day. The accuracy of the utilization rate is already quite good for short simulation durations. Also here, the fluctuations may seem big at first sight, but the vertical axis shows is centered on a specific range.



(a) Utilization rate call takers



(b) Utilization rate dispatchers

Figure 18: Utilization rates of call takers and dispatchers for a call center with 1 call taker and 1 dispatcher as a function of the simulation length. The shaded area shows a 95% confidence interval.

Conclusions

The main findings of this chapter are:

- ECCSIM is an user-friendly simulation tool that has been specially designed for supporting decisions on both *tactical* and *strategic* levels at emergency call centers. The tool is an excellent way of measuring the performance and efficiency of emergency call centers.
- ECCSIM has modules for evaluating various scenarios of merging call centers and measuring the sensitivity with respect to its input parameters.
- The simulation is quite accurate after only simulating a few days because emergency call centers are lightly loaded systems in general.

5 Sensitivity Analysis

In this chapter sensitivity of the model with regard to the input parameters has been analyzed. The goal is to get a better understanding of how changes in input parameters impact the performance of the model. Another goal is to determine the usefulness of the model for supporting decisions on tactical and strategic level. The input parameters that have been used are based on estimates from the data analysis and expert opinions. The real values are probably slightly higher or lower than these estimates and it is therefore useful to test the robustness of the model with regard to its input parameters. The sensitivity with regard to the following three input parameters has been analyzed: *the call durations*, *the fraction of denied calls* and *the fraction of follow-up calls*.

5.1 Sensitivity to the call duration

To test the sensitivity of the performance indicators with regard to the call durations, three different simulations have been performed. Three different simulations have been performed: one with 20% underestimated parameters, one with no change in the parameters and one with 20% overestimated parameters. Figures 19 and 20 show the fraction calls with a waiting time less than 6 seconds for a call center with 2 and 4 generalists, respectively. The figures show that the waiting time is quite insensitive to the changes in call duration for low call volumes but the deviation becomes more significant as the arrival rate increases. The response time service level of A1, A2 calls shows the smallest deviation, because these calls yield the highest priority and they are independent of other call types. B calls show the largest deviation because they have the lowest priority and are sensitive to the call durations of both follow-up calls and A1, A2-calls. The fact that A1, A2 calls have the highest priority causes the service level to be quite robust to varying call durations. However, Figure 19 and Figure 20 also show that B-calls and follow-up calls are less robust to changes in the call durations.

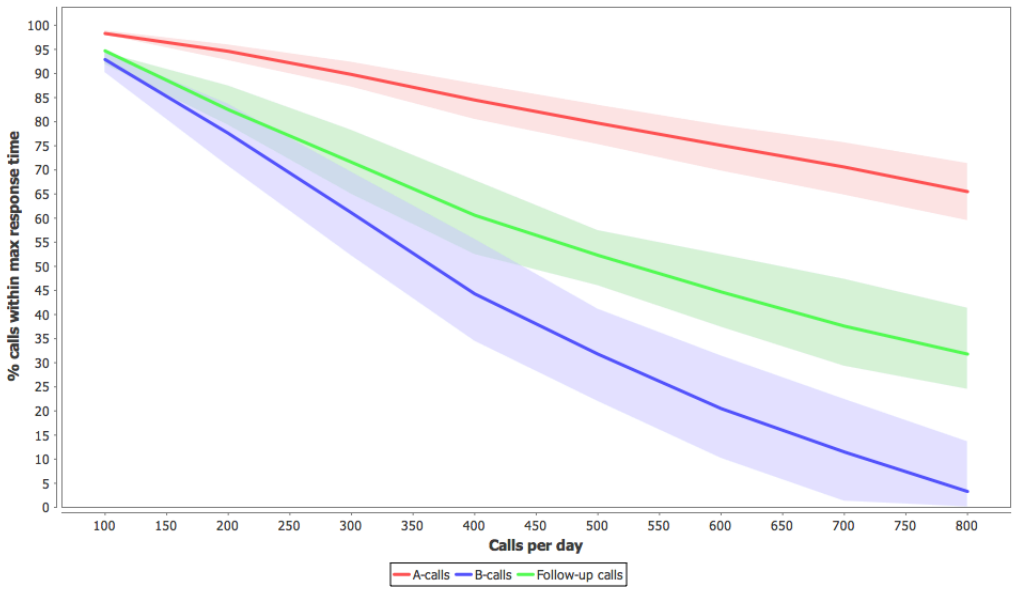


Figure 19: Sensitivity of the waiting time to the call duration for A-calls, B-calls and follow-up calls in a call center with 2 generalists.

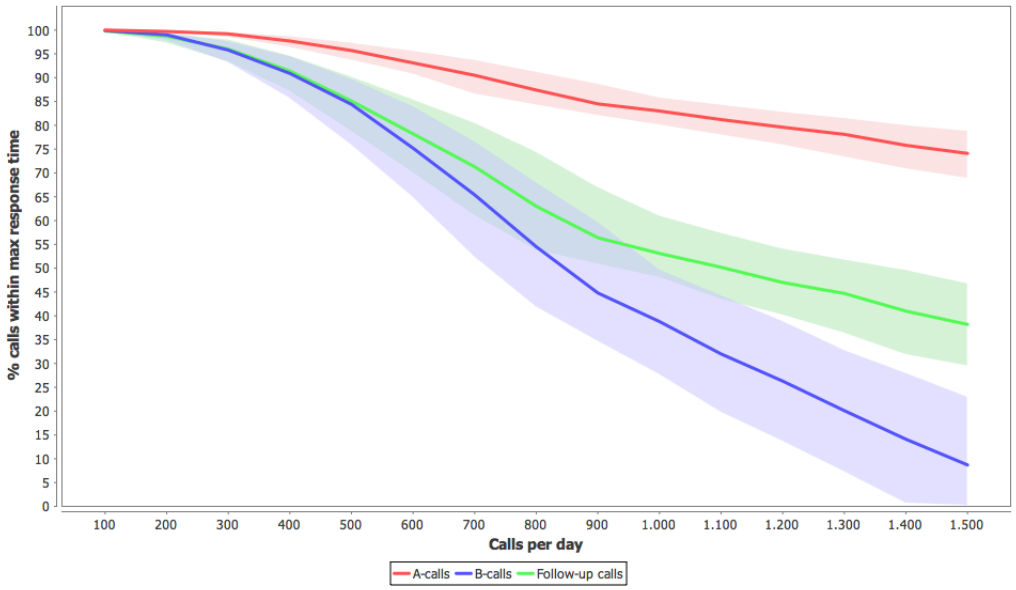


Figure 20: Sensitivity of the waiting time to the call duration for A-calls, B-calls and follow-up calls in a call center with 4 generalists.

Figure 21 shows the sensitivity of the *utilization rate* with respect to the call duration. Three different simulations have been performed: one with 20% underestimated parameters, one with no change in the parameters and one with 20% overestimated parameters. As the call volume increases, the sensitivity of the utilization rate increases for call takers, dispatchers and generalists. The sensitivity of the utilization rate appears to be roughly equal for call takers, dispatchers and generalists. As the call volume increases, the overflow coming from the call takers and dispatchers becomes bigger, causing the utilization rate of the generalist to increase.

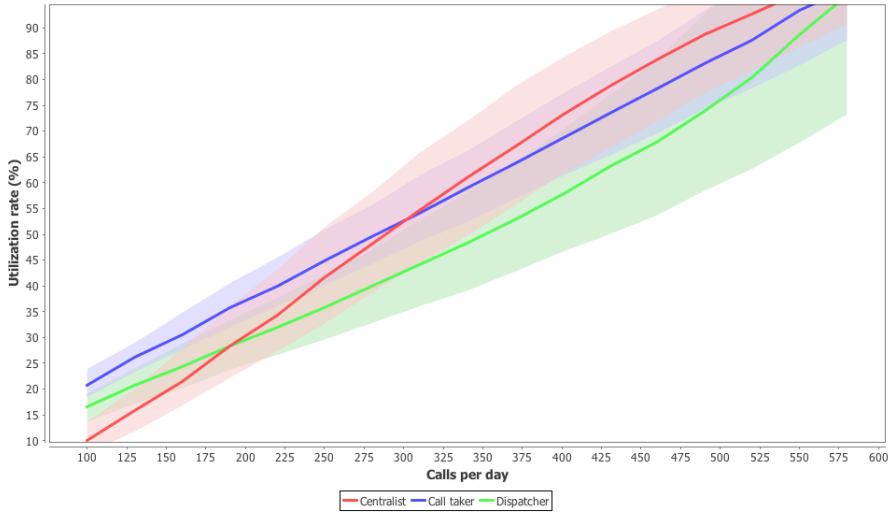


Figure 21: Sensitivity of the utilization rate to the call duration for a call center with 1 call taker, 1 dispatcher and 1 generalist.

5.2 Sensitivity to the fraction of denied calls

Figure 22 shows the sensitivity of the service level with respect to the fraction of denied calls. The A1, A2 calls (A-calls) are insensitive to the fraction of denied calls, because these calls have the highest priority. On the other hand, B-calls show a weak sensitivity to the fraction of follow-up calls, because the dispatcher has a higher availability when more calls are denied. The same reasoning applies for follow-up calls, but now the effect is even more significant because these calls are *only* dependent on the availability of dispatchers. The green line in Figure 22 shows a sudden drop at 100% and it is caused by the fact that when every call is denied, no follow-up calls occur either.

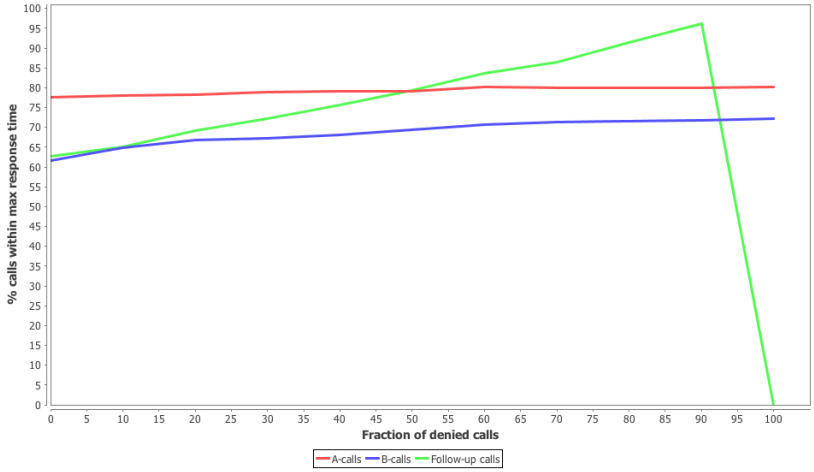


Figure 22: Sensitivity of fraction calls that has been answered within 6 seconds with regard to the fraction of denied calls for a call center with 1 call taker, 1 dispatcher and 300 calls/day.

Figure 23 shows the sensitivity of the utilization rate for call takers and dispatchers with regard to the fraction of denied calls. The utilization rate of call takers is insensitive to the fraction of denied calls calls, because the calls are denied *after* the triage has been performed. However, the utilization rate of dispatchers is affected, because less calls come to dispatcher when the fraction of denied calls increases.

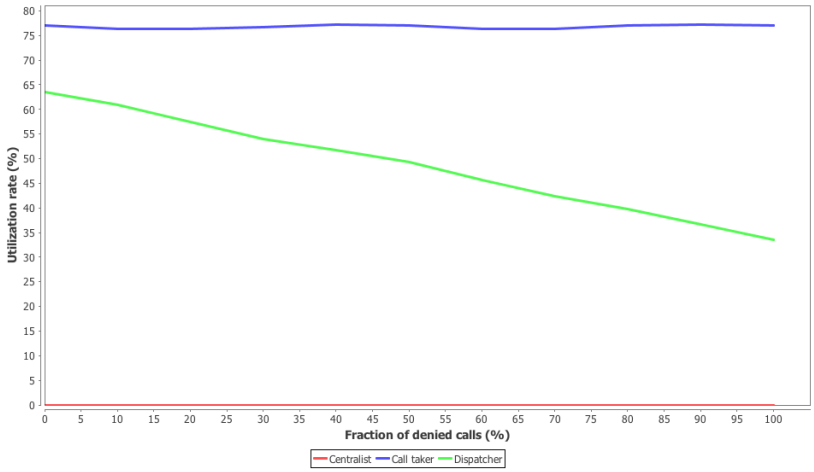


Figure 23: Sensitivity of utilization rate with respect to the fraction of denied calls. Results are from a call center with 1 call taker, 1 dispatchers and 300 calls/day.

5.3 Sensitivity to the fraction of follow-up calls

Figure 24 shows the sensitivity of the waiting time with respect to the fraction of follow-up calls for a 1-1-1 call center with 400 calls/day. As the fraction of follow-up calls increases, the A1, A2 calls remain insensitive because they have higher priority than follow-up calls. However, B-calls have a lower priority compared to follow-up calls and therefore their service level is dependent on the number of follow-up calls.

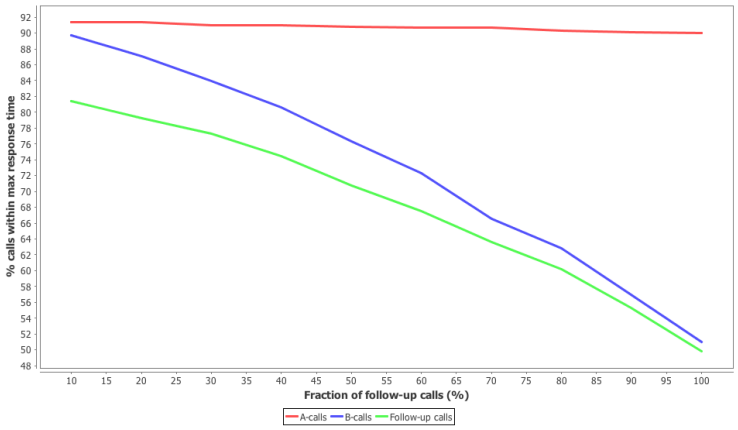


Figure 24: Sensitivity of fraction calls that has been answered within 6 seconds with respect to the fraction of follow-up calls for a call center with 1 call taker, 1 dispatcher, 1 generalist and 400 calls/day.

Figure 25 shows the sensitivity of the utilization rate with regard to the fraction of follow-up calls for a call center with one call taker, one dispatcher, one generalist and 400 calls/day. The utilization rate of the dispatcher is the most sensitive because this employee handles most of the follow-up traffic. The utilization rate of the generalist is also sensitive to the fraction of follow-up traffic because the generalist handles them when the dispatcher is busy. A remarkable result is that utilization of the *call taker* is also sensitive to the fraction of follow-up calls, while the call taker does not handle follow-up calls. It can be explained as follows: as the fraction of follow-up calls is increasing, the generalist becomes more and more busy with handling follow-up traffic, such that the generalist is less able to handle the overflow coming from the call taker. Therefore, the utilization rate of call takers also increases to some extent.

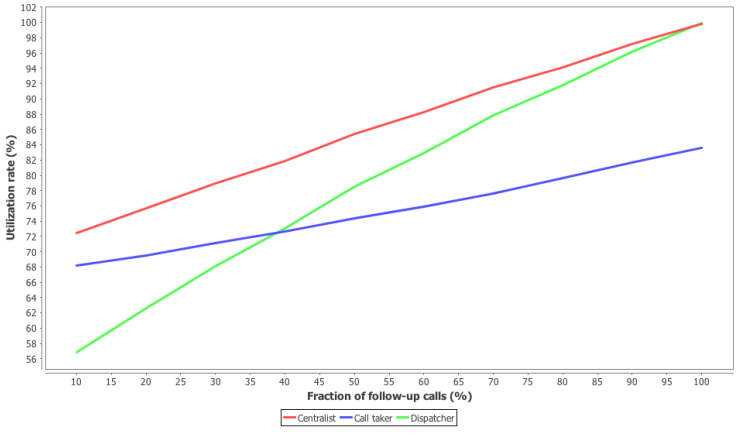


Figure 25: Sensitivity of the utilization rate with regard to the fraction of follow-up calls. Results are from a call center with 1 call taker, 1 dispatcher, 1 generalist and 400 calls/day.

Conclusions

The main findings of this chapter are:

- A-calls are the least sensitive to over-estimated or under-estimated call durations because they have *all capacity* available. On the other hand, B-calls and follow-up calls are more sensitive because they are dependent on calls with higher priority.
- A higher fraction of follow-up calls only affects the waiting time for B-calls and follow-up calls, but not A-calls.

6 Analytical approximation of the performance indicators

This chapter provides several analytical approximation methods of the performance indicators for emergency call centers. Some knowledge of queueing theory is required for reading this chapter, but it can be skipped without any consequences for reading this report. As mentioned earlier, the performance of emergency call centers is measured using the fraction calls that is answered within X seconds. To compute this indicator, the waiting time distribution needs to be known. Computing this distribution is not straightforward, because an emergency call center is not a standard queueing system. For this reason, the main focus will mainly be on approximating the waiting time distribution.

Simulation is suitable and powerful method to analyze the emergency call centers because less assumptions have to be made. However, building a simulation program can be quite time consuming and requires quite some computational power. Using analytical formulas can therefore be useful to provide approximations for performance indicators of a call center. This chapter is divided in two parts: approximations for a call center *with* function differentiation and approximations for a call center *without* function differentiation. In each part, the results of the approximation methods are compared to the simulation.

6.1 Call center with function differentiation

In a call center with function differentiation the triage and logistic processing (i.e., dispatching and processing follow-up calls) is done separately by call takers and dispatchers, respectively. Figure 26 illustrates the process of triage and logistic processing at the emergency call center with function differentiation. In Section 6.1.1, a description of the various approximation methods is provided for triage and in Section 6.1.4 for the logistic processing.

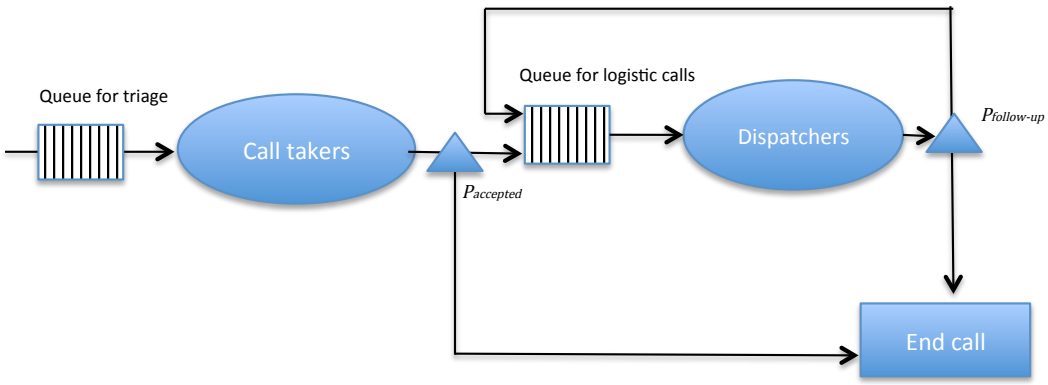


Figure 26: Overview of main processes in an emergency call center.

6.1.1 Approximations for the triage part

The triage part can be modeled as an $M/LN_N/c_1$ queue, where M stands for Markovian arrivals, LN_N for log-normal distributed call durations of N caller groups and c_1 for the number of triage employees. The model described in Chapter 2 is quite extensive and takes many aspects of an emergency call center into account. In order to provide an analytical approximation of the waiting time distribution, this model is somewhat simplified. For instance, there is only looked at the caller group with the highest priority (i.e., A1 and A2-calls). No distinction is made between A1 and A2, because the exact urgency (A1 or A2) is unknown in advance. Another assumption is that each call is accepted (i.e., each call leads to the dispatching of an ambulance).

In queuing theory there are no exact expressions that can calculate the tail probabilities of the waiting time distribution of an M/G/C queue. Some articles have been written by O.J Boxma et al. [8],G.P Cosmetatos [4], A. Al Hanbali et al. [9] on approximations for the expected waiting time. Only Tijms [11] and Tijms-Van Hoorn [10] provide approximations for the tail probabilities of the waiting time distribution and these methods will be analyzed for approximating an $M/LN_N/C$ queue.

Two-moment approximation method by Tijms

This method is introduced by Tijms [11] and uses the first two moments of the call duration. Numerical experiments by Tijms show that the waiting time probabilities are rather insensitive to more than the first two moments (i.e., E(S) and $E(S^2)$) respectively, in which S is the duration of the call. The approximation method uses the squared coefficient of variation, which is denoted as $C_S^2 = \frac{Var(S)}{E(S)^2}$ and measures the normalized spread citeTijms. His technique is based on the Pollaczek-Khintchine formula for the average delay in the queue: $E(W_q) = \frac{1}{2}(1 + c_S^2)\frac{E(S)}{1-\rho}$, in which W_q represents the waiting time in the queue. If $C_S^2 = 1$, then the expected waiting time is the same as the one of an M/M/C system. If $C_s^2 = 0$, then the expected waiting time is the same as the one of an M/D/C system. This expression can be written as $(1 - C_S^2)E(W_{q,M/D/C}) + C_S^2 E(W_{q,M/M/C})$. Tijms extended this expression to the waiting time quantiles and conducted some numerical experiments which showed good results. Tijms remarks that the variation coefficient of the call durations must be not too large ($0 \leq C_S^2 \leq 2$) and that the traffic load on the system cannot be very small. The following expressions describe his approximation technique:

$\xi(p) = p$-quantile of the waiting time distribution, $p \in [0, 1]$.
$\xi_{MD/C}(p) = p$-quantile of the waiting time distribution of an M/D/C queue.
$\xi_{M/M/C}(p) = p$-quantile of the waiting time distribution of an M/M/C queue.
$\xi(p) = (1 - C_S^2)\xi_{M/D/C}(p) + C_S^2 \xi_{M/M/C}(p)$
$C_S = \dfrac{S.Dev(S)}{E(S)}$ (variation coefficient of the call duration S)

The p-quantile of a waiting time distribution for a call center with c centralists can be calculated using the inverse function of the cumulative waiting time distribution: $P(W_q < t)$. The p-quantile of the Erlang C (M/M/C) system is given by:

$$P(W_{q,M/M/C} \leq t) = 1 - C(c, a)e^{-(c\mu-\lambda)t}, t \geq 0$$
$$\xi_{M/M/C}(p) = \frac{\ln(1-p) - \ln(C(c,a))}{-(c\mu - \lambda)}, p \in [0, 1]$$
$$\text{in which: } C(c, a) = \frac{a^c}{(c-1)!(c-a)}\left(\sum_{j=0}^{c-1}\frac{a^j}{j!} + \frac{a^c}{(c-1)!(c-a)}\right)^{-1}$$
$$\text{and } a = \frac{\lambda}{\mu}, \rho = \frac{\lambda}{c\mu},\ \rho < 1$$

An exact expression for calculating the waiting time distribution of an M/D/C system has been given by Crommelin [6], but this method is only practical for systems with a few employees. A better method is provided by Franx [5] and it is displayed in (5). Compared to the M/M/C, this expression is a bit more difficult and taking the inverse function of it is cumbersome. Therefore, the quantiles have been calculated numerically from this expression. The following expressions display the method of Franx.

$$P(W_q < t) = \sum_{j=0}^{kc-1} Q_{kc-1-j} e^{-\lambda(kE(S)-t)} \frac{(\lambda(kE(S)-t))^j}{j!}, \qquad (k-1)E(S) \leq t < kE(S) \qquad (5)$$

$Q_j = \sum_{i=0}^{c+j} p_i$, where p_j denotes the time average probability of having j customers in the system.

and it can be calculated by solving the following linear equations:

$$p_j = e^{-\lambda E(S)} \frac{(\lambda E(S))^j}{j!} \sum_{k=0}^{c} p_k + \sum_{k=c+1}^{c+j} p_k e^{-\lambda E(S)} \frac{(\lambda E(S))^{j-k+c}}{(j-k+c)!} \quad , j \in \{1, 2, \dots\}$$

$$\sum_{i=0}^{\infty} p_i = 1$$

Approximation method using renewal theory by Tijms-Van Hoorn

Van Hoorn and Tijms provide an approximation formula of the waiting time distribution based on renewal theory [10]. They provide the waiting time distribution in a so-called defective renewal equation form. W_q is defined as the waiting time distribution for an arbitrary call. The performance indicator for the response time is defined as the fraction calls that is answered within t seconds. Therefore, this probability is expressed as $P(W_q < t)$. Van Hoorn and Tijms provide an approximation to calculate $P(W_q > t \mid W_q > 0)$. From this expression, $P(W_q < t)$ can be easily calculated using Bayes' theorem: $P(W_q < t) = 1 - P(W_q \geq t \mid W_q > 0)P(W_q > 0)$.

A call center with c employees, an expected call duration $E(S)$ and an arrival rate of λ calls a day is considered. In this case, S represents the distribution of the call duration. The parameter ρ (also called the load)is defined as $\frac{\lambda E(S)}{c}$. Also, the cumulative distribution of the call duration is given by $F(x)$ and the excess call time distribution is given by $F_e(t) = \frac{1}{ES} \int_0^t (1 - F(x))\, dx$. In order to calculate $P(W_q > t \mid W_q > 0)$, the following expressions have been used:

$$P(W_q > t \mid W_q > 0) = (1-\rho)(1 - ((1 - F_e(t))^c) + \lambda \int_0^t P(W_q > x \mid W_q > 0)(1 - F(c(t-x)))dx$$

$$P(W_q > 0) = \frac{(\lambda E(S))^c}{c!(1-\rho)} p_0$$

$$p_0 = \left(\sum_{j=0}^{c-1} \frac{(\lambda E(S))^j}{j!} + \frac{(\lambda E(S))^c}{(c-1)!(c - \lambda E(S))}\right)^{-1}$$

This integral cannot be solved with standard techniques, but Van Hoorn and Tijms provide a numerical procedure that can solve this integral.

M/M/c as approximation method

The third approximation is the so-called Erlang C system (also referred as M/M/c, in which M refers to Markovian and c as the number of employees). For call centers with light traffic (such as an emergency call center), the Erlang C turns out to be a good approximation in general. The Erlang C system is a well-known queuing system for which an expression of the waiting time distribution is available. This system assumes Poisson arrivals, exponentially distributed call durations and an infinite buffer size. Also, let $E(S)$ be the expected call duration and λ the arrival frequency of incoming calls per hour.

Then, the following expressions describe the waiting time distribution of the M/M/c queueing system:

$$P(W_q \leq t) = 1 - C(c,a)e^{-(\frac{c}{E(S)}-\lambda)t)}$$

$$\text{in which: } C(c,a) = \frac{a^c}{(c-1)!(c-a)}(\sum_{j=0}^{c-1}\frac{a^j}{j!} + \frac{a^c}{(c-1)!(c-a)})^{-1}$$

$$\text{and } a = \lambda E(S), \rho = \frac{\lambda E(S)}{c}$$

6.1.2 Parameter estimates

The following paragraphs describe how the input parameters of each approximation methods have been computed.

Two moment approach The data analysis in Chapter 3 showed that the log-normal distribution is a good fit for the duration of the triage. However, there is a complication that occurs with the distribution of the call duration. For each caller group, the parameters for the log-normal distribution are different. For sake of simplicity, the distribution of the call duration is modeled as *one* lognormal distribution with parameters μ and σ. Therefore, the following parameters have been used for applying this approximation method to an emergency call center:

S_{iA} = call duration distribution of caller group i with urgency A

λ_{iA} = arrival rate of caller group i with urgency $A, i \in B$

$\lambda = \sum_{i \in B} \lambda_{iA}$ (total arrival rate per hour)

$E(S) = \sum_{i \in B} p_{iA} E(S_{iA})$

$Var(S) = \sum_{i \in B} p_{iA} E(S_{iA}^2) - (\sum_{i \in B} p_{iA} E(S_{iA}))^2$

$S_{iA} \sim lognormal(\mu_{iA}, \sigma_{iA})$

$p_{iA} = \frac{\lambda_{iA}}{\sum_{j \in B} \lambda_{jA}}, i \in B$

$B = \{112, \text{politie, brandweer, thuiszorg, psychiatrie, Huisarts/HAP, Overig/verloskundig/GHOR}\}$

The following figures have been found for a call center with 100 calls/day:

$\lambda = 2.467$ A-calls/hour

$E(S) = 1.73$ minutes

$Var(S) = 2.03$ minutes^2

$C_S^2 = 0.649$

Method based on renewal theory Also for this method, the call duration has been modeled as *one* lognormal distribution with parameters μ and σ.

$\mu = \sum_{i \in B} p_{iA} \hat{\mu}_{iA}$ (maximum likelihood estimator of $\hat{\mu}_{iA}$ can be found in section 3.2)

$\sigma = \sqrt{\sum_{i \in B} (p_{iA} E(\ln(S_{iA})^2) - \hat{\mu}_{iA}^2)}$

λ_{iA} = arrival rate of caller group i with urgency $A, i \in B$

$p_{iA} = \frac{\lambda_{iA}}{\sum_{j \in B} \lambda_{jA}}$

$B = \{112$, politie, brandweer, thuiszorg, psychiatrie, Huisarts/HAP, Overig/verloskundig/GHOR$\}$

The following formulas describe the density function, the cumulative service duration distribution function and the excess distribution, which are needed for this approximation method.

$f(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}}$

$F(x) = \frac{1}{2} + \frac{1}{2}\text{erf}(\frac{\ln x - \mu}{\sqrt{2\sigma^2}})$, in which erf(*) is the error function.

$F_e(t) = \frac{1}{E(S)} \int_0^t \frac{1}{2} - \frac{1}{2}\text{erf}(\frac{\ln x - \mu}{\sqrt{2\sigma^2}}) dx$

The following figures have been found for a call center with 100 calls/day:

$\lambda = 2.467$ A-calls/hour

$\mu = 0.295$

$\sigma = 0.717$

M/M/C The M/M/C system assumes an exponentially distributed call durations. The exponential distribution has only one parameter, ν, which represents the service rate. This The estimate of ν is calculated using:

$\lambda_{i,A}$ = arrival rate of caller group i with urgency $A, i \in B$

$\lambda = \sum_{i \in B} \lambda_{iA}$ (total arrival rate per hour)

ν_{iA} = average service rate of caller group i with urgency A1/A2

$\nu = \frac{1}{\sum_{i \in B} p_{iA} \frac{1}{\nu_{iA}}}$

$p_{iA} = \frac{\lambda_{iA}}{\sum_{j \in B} \lambda_{jA}}$

$B = \{112$, politie, brandweer, thuiszorg, psychiatrie, Huisarts/HAP, Overig/verloskundig/GHOR$\}$

The following figures have been found for a call center with 100 calls/day:

$\nu = 35.4$ calls/hour

$\lambda = 2.467$ A-calls/hour

6.1.3 Results

In order to test which approximation method performs best, the results of each method are compared with the simulation output using the absolute error. It is defined as:

$$\epsilon_{absolute\ error} = |x_{approx.method} - x_{simulation}|$$

in which $x_{approx.method}$ is the output of the approximation method and $x_{simulation}$ is the output of the simulation. The following graphs show the results of the approximation methods as a function of the load. The load is defined as the fraction time that a centralist is busy.

Figure 27 shows the probability of waiting $P(W_q \leq 0)$ in the triage queue as a function of the load. The load is defined as the fraction of the timFor small loads, all approximation methods perform well because most of the time the employees are idle. However, as the load increases, the employee gets busier and the error of the approximation method increases. This effect can also be seen in Figure 28, which shows the fraction calls that is handled within 60 seconds and the error appears to be bigger in this figure.

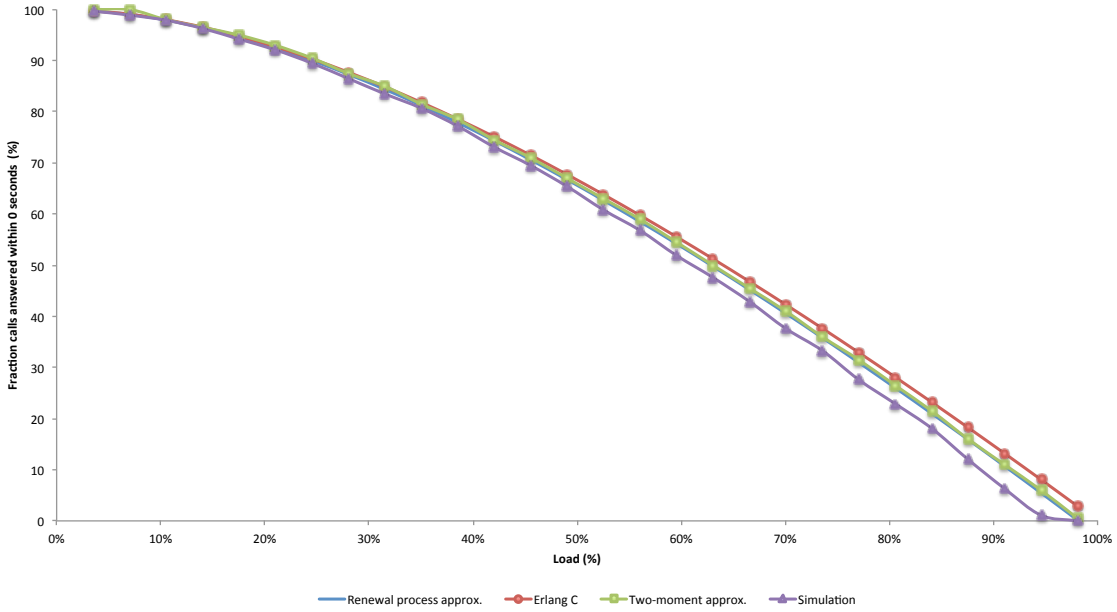


Figure 27: Fraction calls answered within 0 seconds for a call center with 2 call takers.

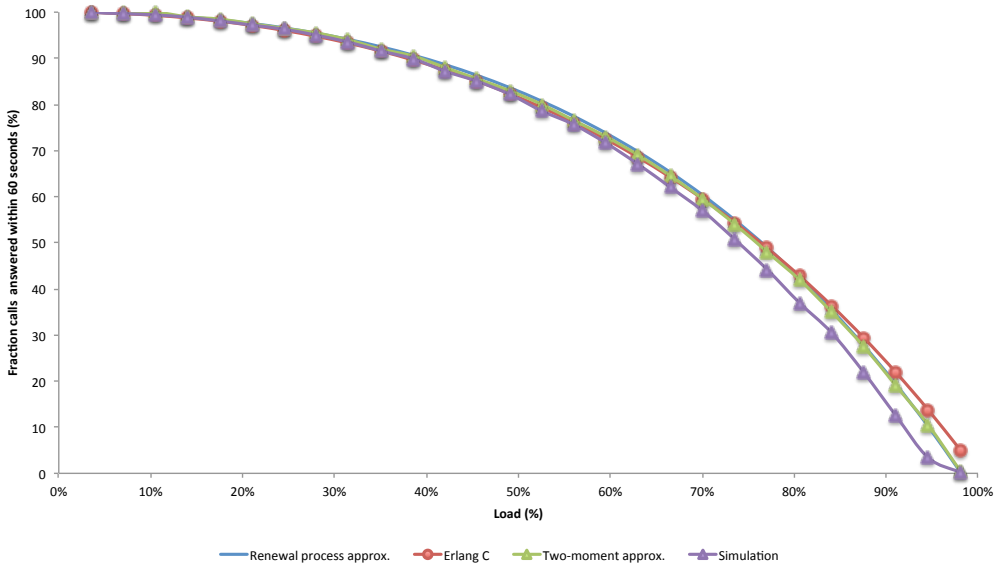


Figure 28: Fraction calls answered within 60 seconds for a call center with 2 call takers.

Figure 29 shows the absolute error of the fraction calls answered instantaneously and it shows that the absolute error increases for all approximation methods as the load gets higher. For low loads, the absolute error is relatively small for all methods, but the error becomes bigger as the load increases. The Erlang C approximation method shows a bigger error compared to the other two methods. A remarkable effect is that all lines in Figure 29 are going up and down at the same time. The reason for this effect is that some error is also included in simulation results, because stochastic processes are simulated.

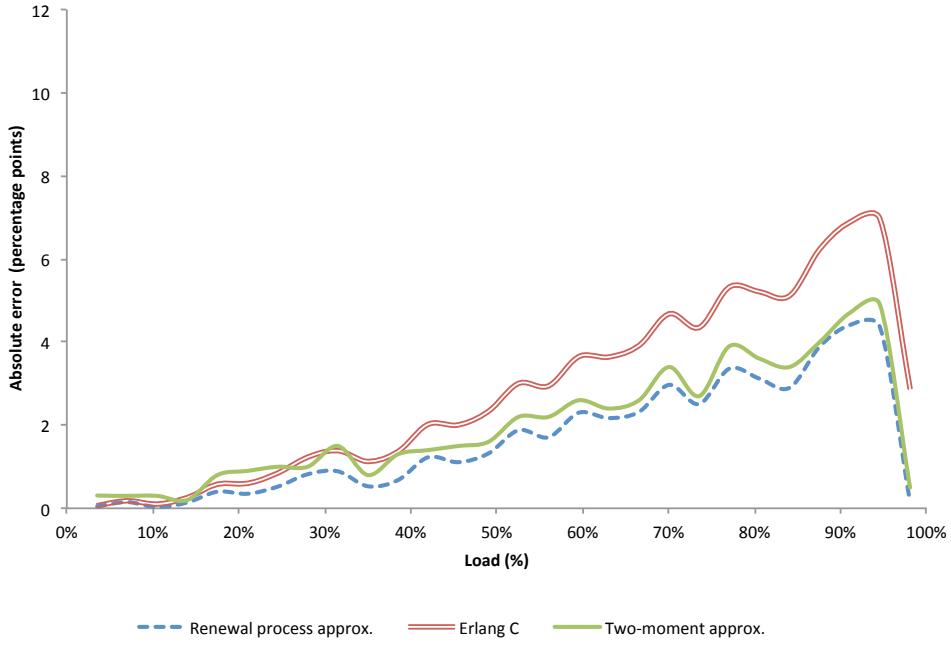


Figure 29: Absolute error of the fraction calls answered instantaneously.

Figure 30 shows the fraction calls that has been answered within 60 seconds. Compared to Figure 29, the absolute error is bigger for all approximation methods.

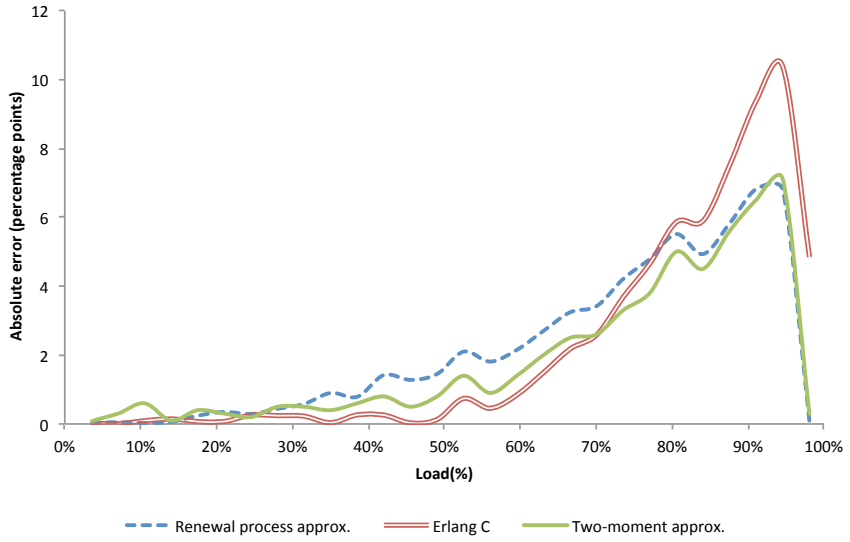


Figure 30: Absolute error for the three approximation methods of the fraction calls answered within 60 seconds.

Figures 29 and 30 show that the approximation methods based on renewal theory and the one based on the two-moments of the call duration distribution, have almost similar performance for all loads. There are two possible explanations for this effect. Either both methods have similar performance or one of the assumptions is causing an error for both methods. As mentioned in Section 6.1.1, the call durations of different caller groups is modeled as *one* log-normal distribution with parameters μ and σ. In the simulation however, each caller group has its own μ_{iA} and σ_{iA}. In order to measure the effect of this assumption, the simulation has also been performed with *one* log-normal distribution for all groups. Figures 31 and 32 show the absolute error of each approximation method with the adapted simulation. Both figures show that the absolute error is significantly lower than in Figures 29 and 30. Also, the approximation method based on renewal theory and the one based on the two moments of the call duration, show different results. This confirms that modeling the call duration with only one lognormal distribution causes a higher absolute error. Another remarkable result is that the Erlang C approximation performs quite well, even for high loads.

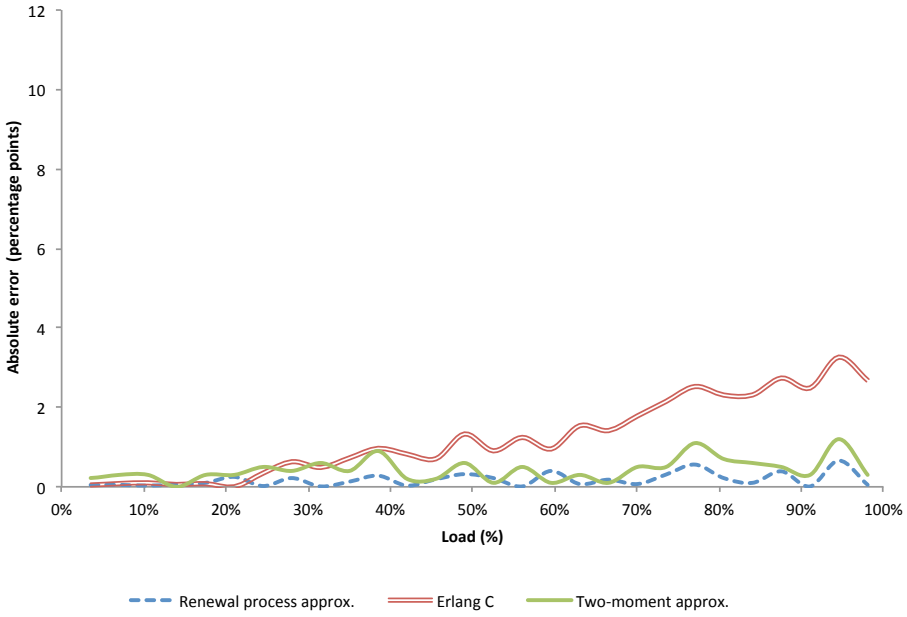


Figure 31: Absolute error of the fraction calls answered within 60 seconds.

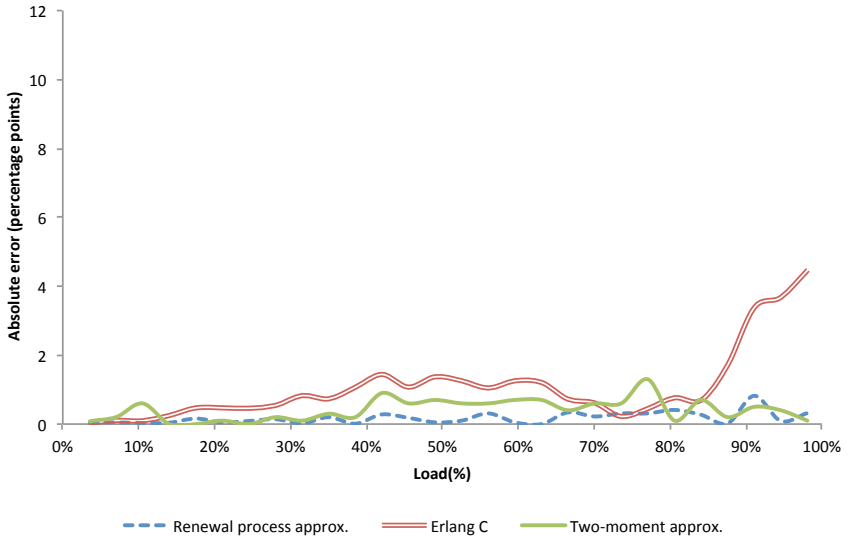


Figure 32: Absolute error of the fraction calls answered within 60 seconds for each approximation method.

6.1.4 Approximations for the logistic process

Figure 33 shows an elaborate overview of the logistic processing in an emergency call center. The incoming call frequency coming from the call takers is denoted with λ_{triage}. There are two call types: calls that need dispatching and follow-up calls. The former group has a higher priority than the latter.

The input process to the logistic processing part is not Poisson anymore, because the call duration distribution in the triage part is not exponential. However, results obtained in Section 6.1.3 showed that the triage part can be well approximated by Erlang C for relatively low loads. Based on this result, Burke's output theorem could be used [12]. This theorem says that the departure distribution of a M/M/C queue is also Poisson [12]. Therefore, the arrival process of calls that require dispatching can be approximated by a Poisson process.

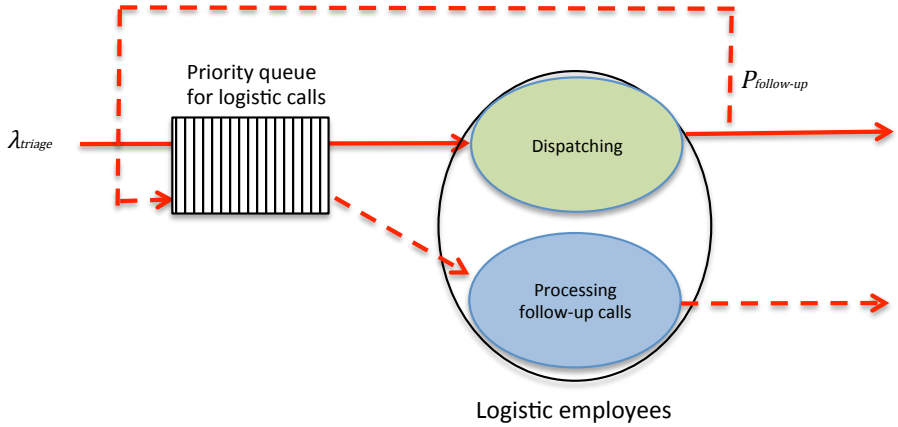


Figure 33: Approximation of the logistic processes in an emergency call center.

The following two paragraphs provide approximation methods for measuring the waiting time distribution of calls that require dispatching and follow-up calls.

Calls that require dispatching The performance of the calls that require dispatching can be approximated using M/M/C (Erlang C), because these calls are insensitive to the presence of other call types due to their high priority. Also, the call durations of these calls have been modeled with an exponential distribution. However, M/M/C is not applicable if the utilization rate of the call takers is lower than 100% because then the arrival rate is bounded by the maximum throughput of the call takers. Figure 34 shows the probability of waiting as a function of the arrival rate. From 2800 calls/day both call takers have a utilization rate of more of 100% and the probability of waiting stays constant. The maximum throughput for dispatchers is higher, because the logistic calls have a shorter call duration and therefore the probability of waiting keeps sticking at $\sim 70\%$.

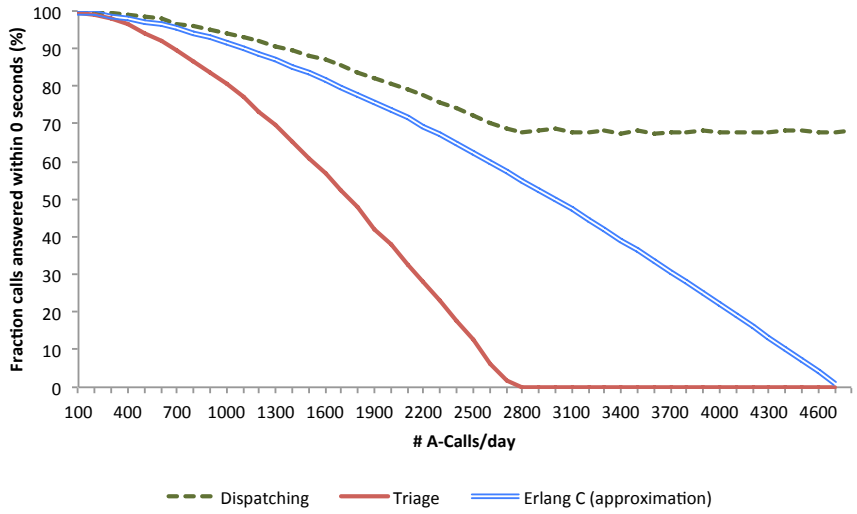


Figure 34: Approximation of the logistic processes in a call center with 2 call takers and 2 dispatchers.

This problem can be solved by expanding the Erlang-C expression for the waiting time distribution. The parameter λ is defined as the incoming arrival rate per hour and $c_{calltakers}$ and c_{disp} are defined as the number of call takers and dispatchers respectively. Furthermore, μ_{triage} and μ_{disp} represent the service rate of triage and dispatching calls. The expressions below show the adapted waiting time distribution and it can be seen that the waiting time distribution becomes insensitive to the arrival rate when $\rho_{triage} \geq 1$ (i.e. $\lambda \geq c_{call\ takers}\mu_{triage}$).

$$P(W_{q,logistic} < t) = \begin{cases} 1 - C(c_{disp}, a_{disp})e^{-(c_{disp}\mu_{disp} - \lambda)t} & \text{if } \lambda < c_{call\ takers}\mu_{triage} \\ 1 - C(c_{disp}, a_{triage})e^{-(c_{disp}\mu_{disp} - c_{call\ takers}\mu_{triage})t} & \text{if } \lambda \geq c_{call\ takers}\mu_{triage} \end{cases}$$

$$a_{triage} = \frac{\lambda}{\mu_{triage}},\ a_{disp} = \frac{\lambda}{\mu_{disp}}$$

$$C(x, y) = \frac{y^x}{(x-1)!(x-y)}\left(\sum_{j=0}^{x-1}\frac{y^j}{j!} + \frac{y^x}{(x-1)!(x-y)}\right)^{-1}$$

Figure 34 also shows that the error between the simulated line (green dashed line) and the Erlang-C approximation (double blue line) appears to be quite big and it appears that the Erlang-C is not a very good approximation. The reason for this deviation is the assumption of having a Poisson arrival process for the logistic part. This is not a accurate assumption, because the triage duration has a lognormal distribution in reality and then Burke's output theorem is not valid anymore. Figure 35 displays the probability of waiting for a call center with 222 call takers and 2 dispatchers. In this situation there is (almost) no error between Erlang C and the simulated results because there is ample capacity in the triage part, which causes that the effect of the lognormal distributed triage duration is small.

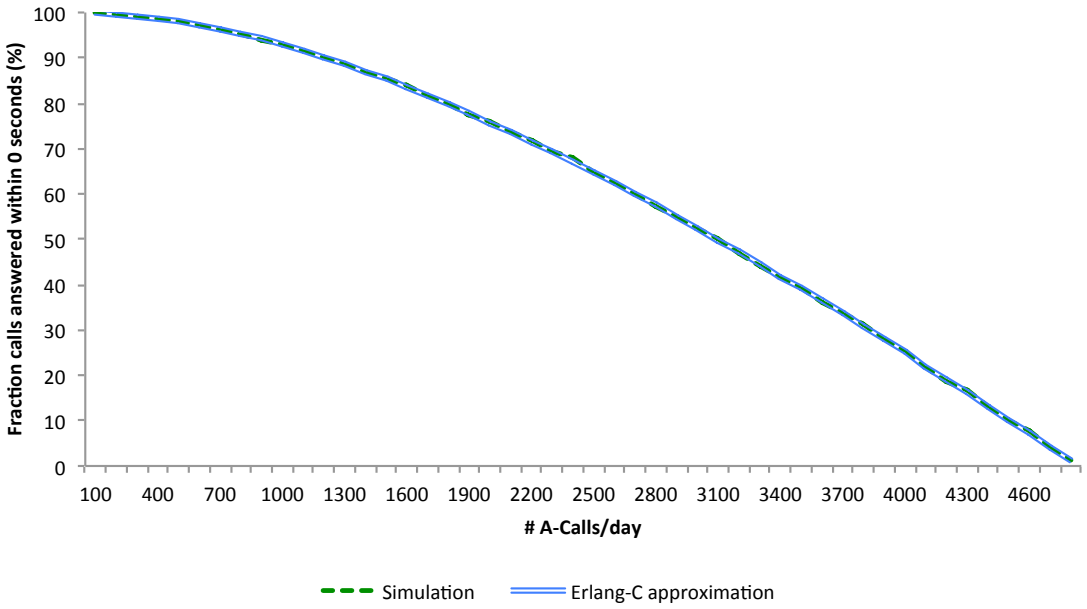


Figure 35: Approximation of the logistic processes in an call center with 222 call takers and 2 dispatchers.

Follow-up calls Approximating the waiting time distribution for follow-up calls is harder, because these calls get interrupted by calls that require dispatching. Articles have been written by Buzen et al. [20], Osogami et al. [14] and Harchol-Balter et al. which provide approximations of the expected waiting time for a multi-server, multi-type and multi-priority queueing system with either pre-emptive resume and non-pre-emptive resume. However, there is no literature available that provide approximations for the waiting time distribution of each priority level because it's a tedious task. For this reason, only the expected waiting time for follow-up calls is considered. This performance metric is less informative compared to the whole waiting time distribution but it still provides some indication of the performance of follow-up calls. In this section, the method for computing the expected waiting time of two priority classes by Osogami [14] is used, which involves a two-dimensional Markov chain.

The combination (L, H) indicates that there are L low priority customers and H high priority calls in the call center. Let μ_L be the service rate of low priority calls and μ_H the service rate of high priority calls. Also, let λ_L and λ_H be the arrival rates of the low priority and high priority calls respectively. Then, Figure 36 shows the Markov chain that can be constructed for this situation. This is a Markov chain with an infinite state space, because there is an infinite buffer for low and high priority calls. In order to estimate the steady state probabilities, the Markov chain has to be cut off at a certain level $k \geq 0$ for which the steady state probabilities are near zero. Essentially, the M/M/C priority queue is approximated by an M/M/C/$(k-C)$ queue, where $k-C$ represents the maximum number of high/low priority calls in the queue. For the results in this section, k is set to 50. This is a fair assumption, because the probability of having more than 50 (high or low priority) calls in the system is very small for a light traffic system. However, for high loads this will under estimate the expected waiting time.

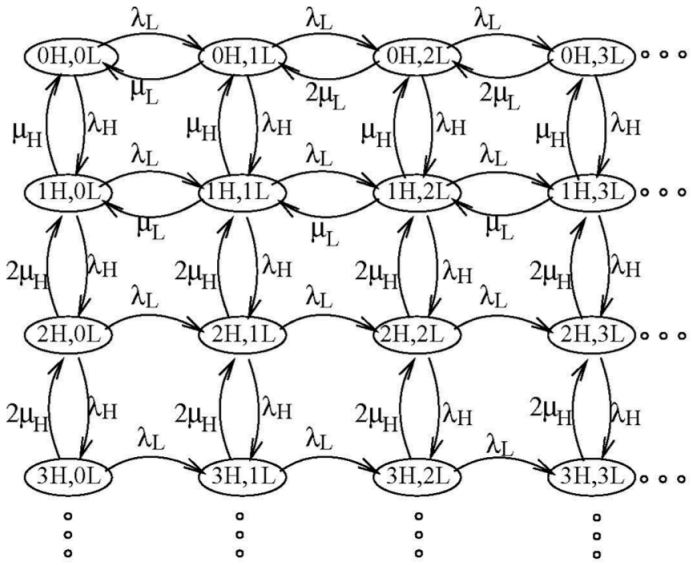


Figure 36: Markov chain for low and high priority calls for a call center with two employees [14]

.

Let p_{ij} be the long term fraction of time of having i high priority calls and j low priority calls in the system. The values of $p_{i,j}$ are obtained by solving the Markov chain displayed in Figure 36. Furthermore, let C denote the number of employees. Then, the expected waiting time of high priority calls, $E(W_{q,high})$, can be calculated using: $\sum_{j=0}^{k} \sum_{i=C+1}^{k} \frac{i}{c\mu_H} \cdot p_{ij}$. This expression is equivalent to computing the expected waiting time of an M/M/C system with service rate μ_H and arrival rate λ_H, because the high priority calls are oblivious to low priority calls.

Calculating the expected waiting time for low priority calls *directly* is a bit more complicated because they are dependent on high priority calls. Therefore, it is derived implicitly from the expected queue length. Let's denote $L_{q,L}$ as the number of low priority calls in the queue. Then, E($L_{q,L}$) can be computed using:

$$E(L_{q,L}) = \sum_{i+j>c} j p_{i,j} \qquad i,j \in \{0,\ldots,k\}, c \geq 1$$

Then, by using Little's law:

$$E(W_{q,L}) = \frac{E(L_{q,L})}{\lambda_L}$$

Parameter estimates

The following parameters have been used to compare the approximation methods with the simulation results. In order to vary the load, the arrival rate has been scaled linearly.

Parameters for a call center with 100 calls/day:

$\lambda_{triage} = 2.47$ A-calls/hour
$p_{follow-up} = 0.11$
$\lambda_{follow-up} = 0.27$ follow-up calls/hour
$\mu_{dispatching} = 60$ dispatch calls/hour (dispatching capacity of one employee)
$\mu_{follow-up} = 15$ follow-up calls/hour (follow-up capacity of one employee)

Numerical results

Figure 37 displays the average waiting time in the queue for an emergency call center with 2 call takers and 2 dispatchers as a function of de load. The load is defined as the fraction time that a centralists is busy. Around a load of 55%, the two call takers have reached their maximum throughput, such that the expected waiting time of follow-up calls stays constant from that point. Before this point, the error is caused by the fact that the arrival process is not exactly a Poisson process.

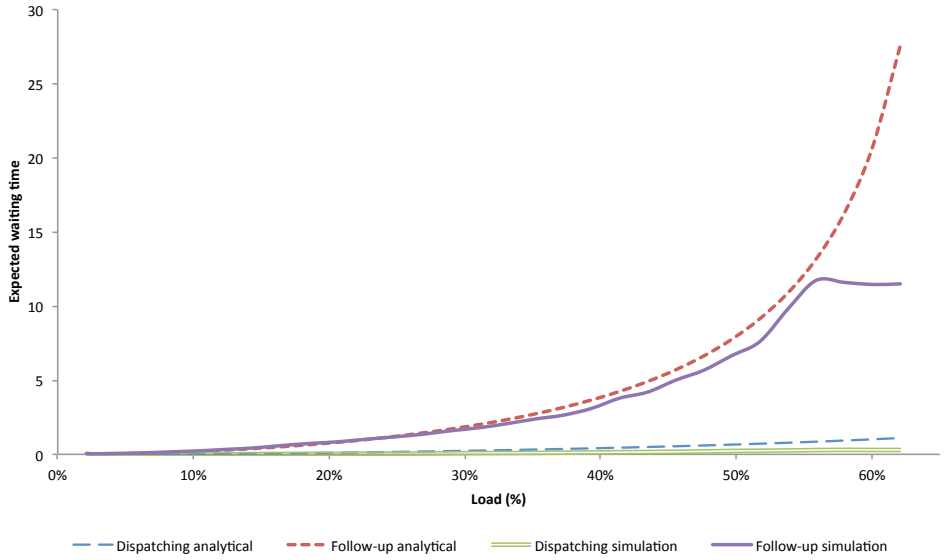


Figure 37: Expected waiting time for a call center with 2 call takers and 2 dispatchers.

Figure 38 shows the expected waiting time for a call center with 222 call takers and 2 dispatchers. The number of call takers is chosen arbitrarily, but is set high enough such that a bottleneck at the call takers is avoided. Compared to Figure 37, the error is less because the impact of not having an exact Poisson process is now less significant.

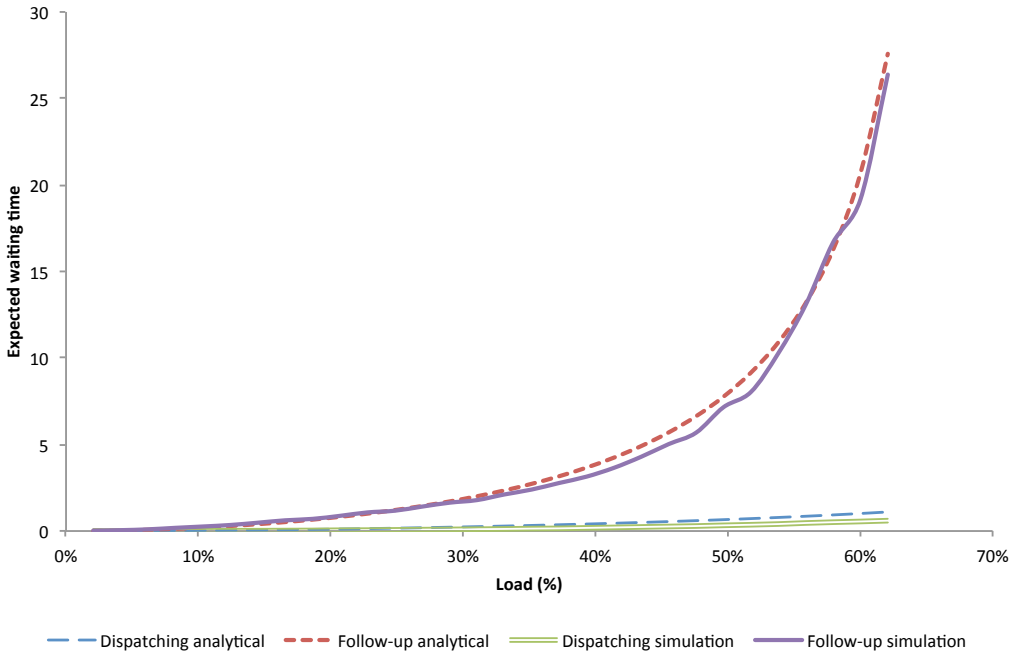


Figure 38: Expected waiting time for a call center with 222 call takers and 2 dispatchers.

6.2 Call center without function differentiation

Figure 39 shows the route of a call through an emergency call center without function differentiation. In this model, the generalist is able to handle both triage- and logistic calls. Now, only two types of calls are considered: incoming calls (that require triage) and follow-up calls. The former group has a higher priority compared to the latter group. The triage and dispatching is now considered as one task and if a low-priority call is interrupted during this task, the service has to start all over once an employee becomes available. The call durations of the triage, dispatching and processing of follow-up calls are kept the same as in the model with function differentiation. In order to provide an approximation of the waiting time distribution for both call types, some assumptions have been made to simplify the model. First, all calls lead to the dispatching of *one* ambulance, meaning that no calls are denied. Second, the elaborate priority system introduced in Section 2.2 is abandoned, but three priorities have been used instead: A-calls have the highest priority, follow-up calls have medium priority and B-calls have a low priority.

This section is constructed as follows: first, approximations of the waiting distribution for incoming calls (only A1, A2-calls) will be given, followed by approximations for follow-up calls.

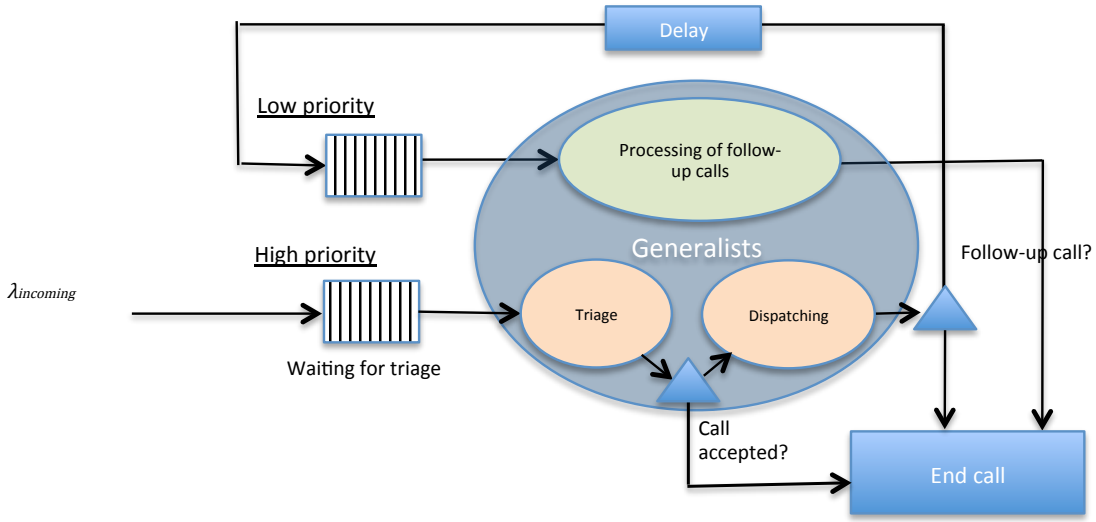


Figure 39: Process overview of a call center without function differentation.

6.2.1 Approximations for incoming calls

Incoming calls first undergo triage, followed by the dispatching of an ambulance. Both tasks are done sequentially by the same centralist. In Chapter 3, the data analysis showed that the log-normal distribution provides a good fit to the duration of the triage. Little is known about the exact distribution of the duration of dispatching, but it is reasonable to assume an exponential distribution. This makes the total duration of triage and dispatching a mixture of two distributions.

The lognormal distribution has two parameters: μ_{triage} (location) and σ_{triage} (scale). Furthermore, $\nu_{dispatching}$ is denoted as the service rate per hour for dispatching. Then:

$$S_{total} = S_{triage} + S_{dispatching}$$

in which: $S_{triage} \sim$ log-normal$(\mu_{triage}, \sigma_{triage})$ and $S_{dispatching} \sim exp(\nu_{dispatching})$. Also, S_{triage} and $S_{dispatching}$ are assumed to be independent.

The incoming calls consist of two groups: A-calls (of which the exact urgency (A1 or A2) is unknown) and B-calls. As mentioned earlier, the B-calls are (usually) requests for the transportation of patients to and from healthcare institutions and these calls have a lower priority compared to A-calls. Approximating the waiting time distribution of B-calls and follow-up calls is tedious, because they are dependent on calls that have higher priority.

In Section 6.1.1, three approximation methods have been presented for approximating the waiting time distribution of a M/G/C queue in a call center with function differentiation. These methods will also be applied to a call center without function differentiation and the following paragraphs display the required parameters.

Two-moment approximation method An elaborate overview of this method can be found in Section 6.1. The two moment approach is based on the variation coefficient of the call duration distribution. The variation coefficient, C^2_{total}, is defined as $\frac{Var(S_{total})}{E(S_{total})^2}$ and is a normalized measure for spread. $E(S_{triage})$ represents the expected duration of the triage and $E(S_{dispatching})$ represents the expected duration of dispatching. Then, $E(S_{total})$ and $Var(S_{total})$ can be computed as follows:

$$\begin{aligned}
E(S_{total}) &= E(S_{triage} + S_{dispatching}) = E(S_{triage}) + E(S_{dispatching}) \\
&= e^{\mu_{triage} + \frac{\sigma^2_{triage}}{2} + \frac{1}{\nu_{dispatching}}} \\
Var(S_{total}) &= Var(S_{triage} + S_{dispatching}) \\
&=^{!} Var(S_{triage}) + Var(S_{dispatching}) \qquad \text{(because } S_{triage} \text{ and } S_{dispatching} \text{ are independent)} \\
&= (e^{\sigma^2_{triage}} - 1)e^{2\mu_{triage} + \sigma^2_{triage}} + \frac{1}{\nu^2_{dispatching}}
\end{aligned}$$

Based on the data, the following estimates for $E(S_{total})$, $Var(S_{total})$, $C^2_{S_{total}}$ have been found for a call center with 100 calls a day.

$\lambda = 2.467$ calls/hour
$E(S_{triage}) = 1.73$ minute
$Var(S_{triage}) = 1.93$ minute^2
$E(S_{dispatching}) = 1$ minute
$Var(S_{dispatching}) = 1$ minute^2
$E(S_{total}) = 1.83$ minute
$Var(S_{total}) = 2.93$
$C_{S_{total}} = \dfrac{2.93 + 1}{2.73^2} = 0.3658$

Approximation method using renewal theory This method has been introduced in Section 6.1 and relies on the cumulative distribution and the excess time distribution of the call duration. In a call center with function differentiation, the log-normal distribution has been used for the duration of triage. In a call center without function differentiation, the total call duration (i.e., triage and dispatching) is a sum of a log-normal and an exponential stochastic variable. Still, the cumulative distribution of this distribution can be derived. Let's define $F_{S_{total}}$ as the cumulative distribution function of the total duration and $f_{S_{triage}}(x)$ and $f_{S_{disp}}(x)$ as the density functions of the triage and dispatching duration respectively (where $x \geq 0$). Also: $S_{triage} \sim$ log-normal$(\mu_{triage}, \sigma_{triage})$ and $S_{disp.} \sim \exp(\nu_{disp.})$. Then, the following expressions formulas have been used for computing $F_{S_{total}}$:

$F_{S_{triage}}(x) = \frac{1}{2} + \frac{1}{2}\text{erf}(\frac{\ln x - \mu_{triage}}{\sqrt{2\sigma^2_{triage}}})$, in which erf(*) is the error function.

$f_{S_{disp.}}(x) = \nu_{disp.} e^{-\nu_{disp.} x}$

$F_{S_{total}}(t) = P(S_{triage} + S_{disp.} < t) = \int_0^t P(S_{Triage} + x < t | S_{disp.} = x) f_{S_{disp.}}(x) dx,\ t \geq 0$

$= \int_0^t F_{S_{triage}}(t-x) f_{S_{disp.}}(x) dx$ (by using Bayes and assuming independency of S_{triage} and $S_{disp.}$)

$= \int_0^t \frac{1}{(t-x)\sqrt{2\pi\sigma^2_{triage}}} e^{-\frac{(\ln(t-x)-\mu_{triage})^2}{2\sigma^2_{triage}}} \nu_{disp.} e^{-\nu_{disp.} x} dx$

$= \int_0^t \frac{\nu_{disp.}}{(t-x)\sqrt{2\pi\sigma^2_{triage}}} e^{-\frac{(\ln(t-x)-\mu_{triage})^2}{2\sigma^2_{triage}} - \nu_{disp.} x} dx$

This last integral is hard to solve analytically and therefore a numerical procedure has been used. Based on the data, the following estimates for λ, μ, σ have been found for a call center with 100 calls a day:

$\lambda = 2.467$ calls/hour
$\mu_{triage} = 0.428$
$\sigma_{triage} = 0.902$
$\nu_{disp.} = 60$ calls/hour

M/M/C as approximation method The last approximation method that has been discussed in Section 6.1 is the M/M/C queueing system. An exact expression of the waiting time distribution exists for this queuing system, but it assumes exponential distributed call durations, while the total call duration of incoming calls is a sum of a log-normal and exponential distributed variables. Despite this difference, it is still interesting how the M/M/C queue performs relative to the other approximation methods. Let's define γ_{total} as the service rate of incoming A-calls per hour.

The exponential distribution has the parameter μ and is defined as the average service rate per hour in case total duration of triage and dispatching is considered to be exponential distributed. Then:

λ_{iA} = arrival rate of caller group i with urgency $A, i \in B$
$\lambda = \sum_{i \in B} \lambda_{iA}$ (total arrival rate per hour)
ν = average service rate per hour for dispatching calls
μ_{iA} = average service rate per hour for triage calls of caller group i with urgency A
$\mu = \frac{1}{\sum_{i \in B} p_{iA} \frac{1}{\mu_{iA}} + \frac{1}{\nu}}$
$p_{iA} = \frac{\lambda_{iA}}{\sum_{j \in B} \lambda_{jA}}$
$B = \{112, \text{politie, brandweer, thuiszorg, psychiatrie, Huisarts/HAP, Overig/verloskundig/GHOR}\}$

Based on the data, the following estimates of μ, ν and λ have been found:

$$\nu = 60 \text{ calls/hour}$$
$$\sum_{i \in B} p_{iA} \frac{1}{\mu_{iA}} = 35.44 \text{ calls/hour}$$
$$\lambda = 2.47 \text{ calls/hour}$$
$$\mu = 22.28 \text{ calls/hour}$$

Numerical results

Figure 40 shows the absolute error of the probability of waiting. It shows that the method based on renewal theory works best, followed by the 'two moment' method. A surprising result is that Erlang C performs quite well too, while this method has not been designed for approximating an M/G/C system. Another observation is that the absolute error is increasing for all methods as the load gets higher. As mentioned in Section 6.1.3, this is caused by the fact that modeling the triage duration is modeled with only *one* lognormal distribution, while in the simulation every caller group has its own log-normal distribution with parameters μ_{iA} and σ_{iA}.

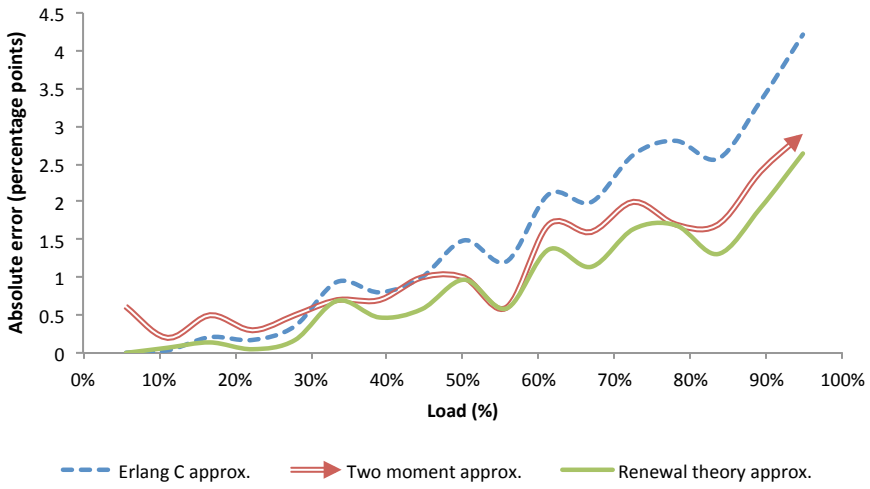


Figure 40: Absolute error for the probability of waiting for a call center with 2 generalists.

Figure 41 shows the absolute error for the probability that a call has to wait less than 1 minute. Surprisingly, the Erlang C approximation outperforms the other methods for loads between 45% and 80%.

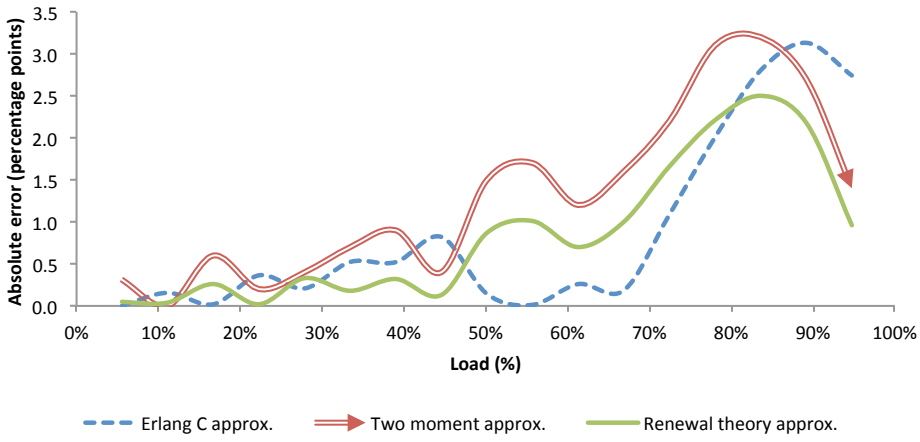


Figure 41: Absolute error of the probability that the waiting time is less than 1 minute for a call center with 2 generalists.

6.2.2 Approximations for follow-up calls

As mentioned earlier, follow-up calls have a lower priority than incoming calls. Approximating the waiting time distributions of these calls is difficult and therefore only the expected waiting time is considered. In Section 6.1.4, a Markov chain was introduced in order to compute the expected waiting time distribution for follow-up calls. This method assumes exponentially distributed call durations for high and low priority calls. However, in a call center without function differentiation, the triage + dispatching duration is not exponentially distributed. Therefore, exponentially distributed triage durations have to be assumed for these calls.

$$\begin{aligned} \lambda_{incoming} &= 2.47 \text{ calls/hour} \\ \lambda_{follow-up} &= 0.2717 \text{ calls/hour} \\ \mu_{incoming} &= 22.28 \text{ calls/hour} \\ \mu_{follow-up} &= 15 \text{ calls/hour} \end{aligned}$$

Figure 42 shows the expected waiting time for incoming calls and follow-up calls for a call center with 2 generalists. The error is lower for follow-up calls than for incoming calls. The reason for this effect is that the triage + dispatching duration is not exponentially distributed. This assumption also causes errors for the follow-up calls, because their performance is dependent on the incoming calls (because they have lower priority).

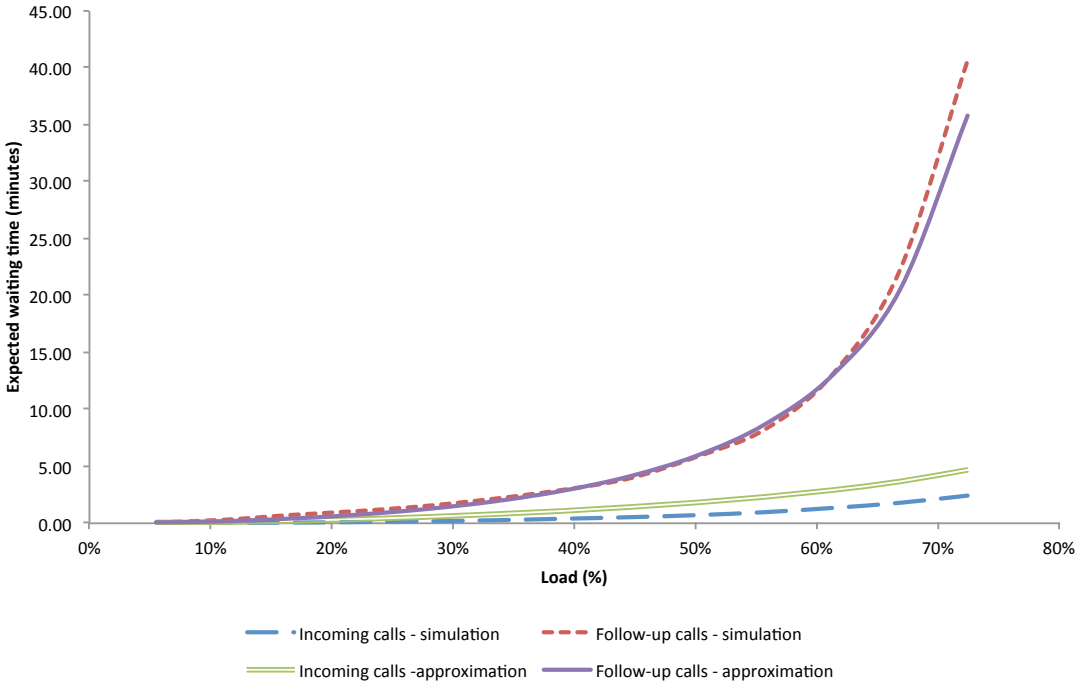


Figure 42: Expected waiting time for incoming calls and follow-up calls.

Conclusions

The main findings of this chapter are:

- Analytical methods are good for calculating the waiting time distribution of A-calls.
- Computing the tail probabilities of the waiting distribution for B-calls and follow-up calls is difficult. Therefore, the expected waiting time is used instead.
- Erlang C (M/M/C) turns out to be a relatively good approximation method for the waiting time distribution of an M/G/C system, while it has not been specially designed for approximating such systems.
- Analytical methods are good for doing *quick capacity calculations*.

7 Impact of staff combinations on the performance indicators

In this chapter, the performance of an emergency call center is analyzed for different staff combinations. As mentioned earlier, there are three different types of employees, each having different skill sets:

- Call taker: only performs triage.
- Dispatcher: does dispatching and processes follow-up calls.
- Generalist: performs triage, does dispatching and processes follow-up calls.

Using only call takers and dispatchers is also called 'function differentiation', because the tasks within an emergency call center are differentiated. Combinations of call takers, dispatchers *and* generalists are referred to as 'hybrid combinations'.

The goal of this chapter is to provide a better understanding of how performance and efficiency are affected by these different staff combinations. This chapter is organized as follows: the first section provides an understanding of how the waiting time is influenced by different staff combinations. The second section analyzes the impact of different staff combinations on the utilization rate of each employee type.

All results in this chapter have been obtained using discrete event simulation. The emergency call center has been simulated for a timespan of 200 days, of which 1 day functioned as a warming-up period.

7.1 Impact of staff combinations on the waiting time

For a call center without function differentiation, the waiting time is the time between the moment a call arrives and the moment a call is answered by a generalist. It suffices as a performance indicator for a call center with *only* generalists, because they do the triage and dispatching sequentially. However, it is an insufficient performance indicator for a call center *with* function differentiation, because the calls that require dispatching can be put in the queue awaiting the next available dispatcher. A better definition for such call centers would be the waiting time for triage *plus* the time a call waits for dispatching. For example, if a call has to wait 20 seconds to be processed by a call taker and the call has to wait 15 seconds to be processed by a dispatcher, the total waiting time will be 35 seconds.

The following paragraphs show the fraction calls that has been answered within time frame for various staff combinations. The time frames vary for each call type and Table 7 shows them for A1, A2-calls, B-calls and follow-up calls respectively. the impact of the staff configuration on the performance for A1/A2 calls, B-calls and follow-up calls.

	Max. waiting time (sec)
A1 + A2	6
B	150
Follow-up	30

Table 7: Service levels for different call types.

Performance of A1 and A2 calls Figure 43 showsthe percentage A1 and A2 calls that had a waiting time less than 6 seconds for multiple staff combinations . The composition of the call center staff is indicated by a code. For example, 01-02-00 means that there is one call taker, two dispatchers and zero generalists. Figure 43 shows that the combinations with only generalists perform better than the ones with function differentiation. The reason for this phenomenon is that generalists are more

flexible than call takers and dispatchers. For example, a 2-1-0 combination performs worse than a 0-0-3 combination, because a bottleneck is created at the logistic processing part such that the waiting time for dispatching increases. In a 0-0-3 combination, the generalists can also handle logistic calls, meaning extra capacity for dispatching. The 1-1-1 combination is the only hybrid configuration in the figure and it performs better than the 2-1-0 and 1-2-0 combinations, but not as good as a call center with only generalists.

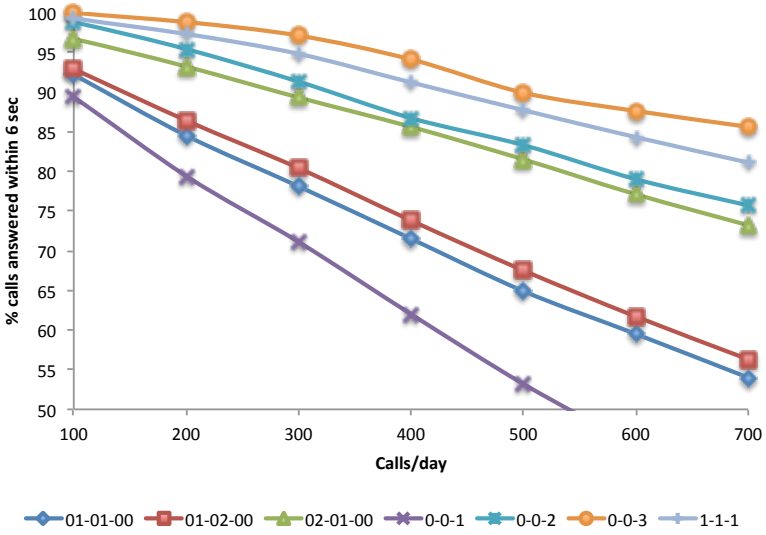


Figure 43: Performance of different staff combinations for A1 and A2 calls.

Performance of B calls Figure 44 shows for multiple call center staff combinations the percentage B-calls that have a waiting time less than 150 seconds. For B-calls, the combinations with only generalists also perform better compared to other combinations with call takers and dispatchers. The 1-1-1 combination is the only hybrid configuration in the figure and it performs better than the 2-1-0 and 1-2-0 combinations, but not as good as a call center with only generalists. Two series stand out in Figure 44: the 0-0-2 combination (light-blue) and the 2-1-0 combination (green). From 100 calls/day to 250 calls/day, the 0-0-2 combination performs slightly better than a 2-1-0 combination. This is a contradicting result because the 2-1-0 combination has an extra employee. The reason for this effect is that there is more capacity for *dispatching* B-calls in a 0-0-2 combination than in a 2-1-0 configuration. However, as the call volume increases, the generalists in the 0-0-2 combination become more occupied with triage *and* dispatching of A-calls, such that the waiting time for B-calls increases.

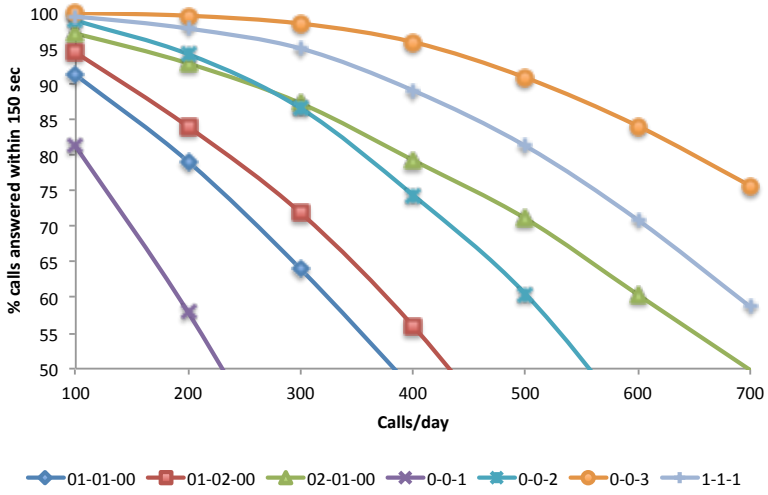


Figure 44: Performance of different staff combinations for B calls.

Performance of follow-up calls Figure 45 shows for multiple call center staff combinations the percentage follow-up calls that had a waiting time less than 30 seconds. In general, emergency call centers with function differentiation perform better for follow-up calls. For example, the 1-2-0 combination performs better than the 0-0-3 combination. The reason for this effect is that there are many dispatchers in a 1-2-0 configuration. These dispatchers only do dispatching and handling follow-up traffic, causing the service level for follow-up traffic to be better. However, the 1-2-0 combination performs not well for A- and B- calls, because the triage capacity is limited in these combinations.

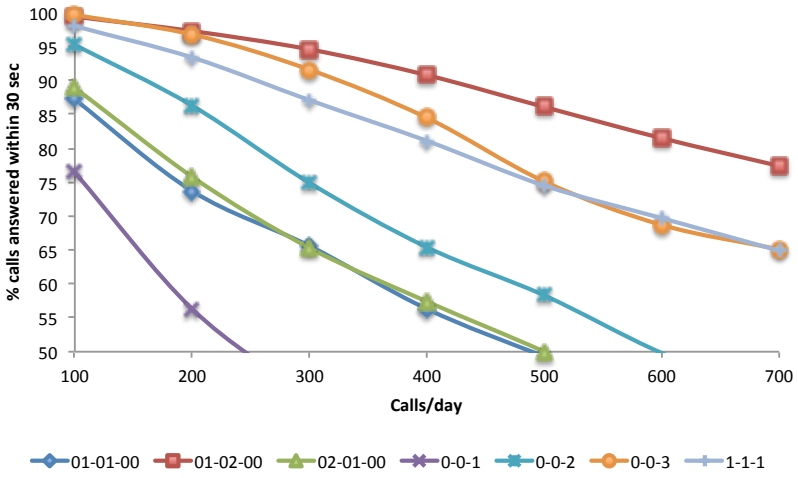


Figure 45: Performance of different staff combinations for follow-up calls.

7.1.1 Distribution of the waiting time

Figures 46 and 47 show the cumulative distribution function of the waiting time for some staff combinations. All results in this subsection have been modeled with a call center that has call volume of 400 calls/day. Figure 46 shows the cumulative distribution of the waiting time for a call center with

1 call taker, 1 dispatcher and zero generalists. This staff configuration provides the best performance for A1, A2 calls, follow-up calls and then B-calls. This is in accordance with the order of priorities for each call type. However, when the composition of the staff is changed, the performance of each call type changes.

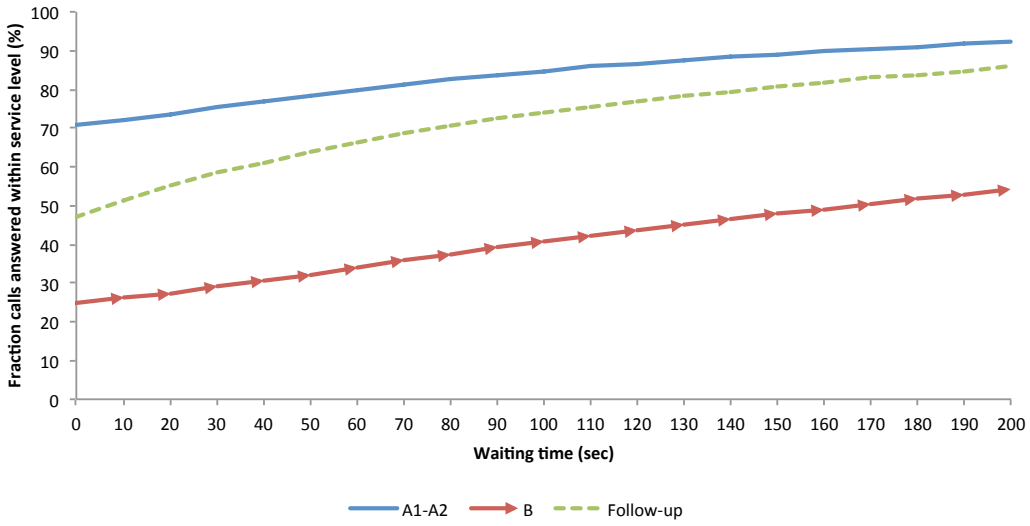


Figure 46: Waiting time distribution for a 1-1-0 staff configuration with 400 calls/day.

Figure 47 shows the performance for a call center with 1 call taker and 2 dispatchers. This configuration has a high capacity for dispatching and follow-up traffic, which results in a significant increase in performance for follow-up calls.

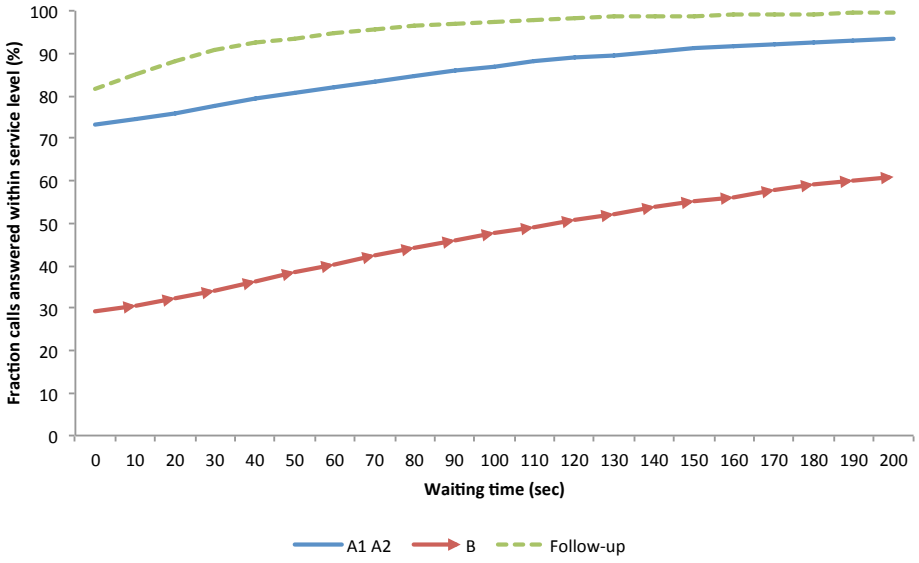


Figure 47: Waiting time distribution for a 1-2-0 staff configuration with 400 calls/day.

Figure 48 shows the performance for a call center with 2 call takers and 1 dispatcher. This configuration has a high capacity for triage, which results in a better performance for A- calls and B-calls. The 2-1-0 configuration has a better performance for all call types than the 1-2-0 configuration (Figure 47), The reason for this effect is that the triage takes more time on average than dispatching and handling follow-up traffic. Therefore, a bottle neck is created at the call taker in a 1-2-0 configuration, causing the performance of A- and B- calls to deteriorate.

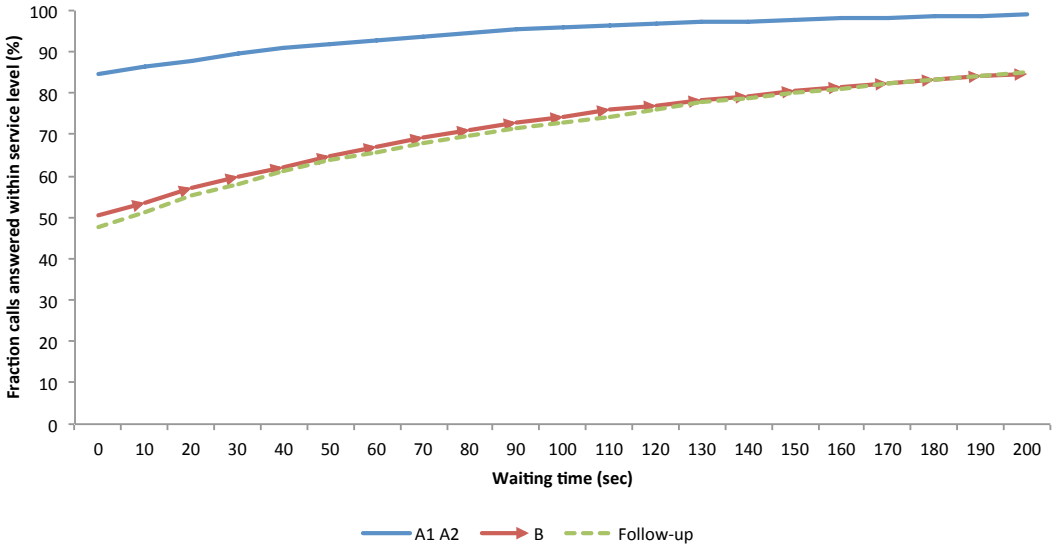


Figure 48: Waiting time distribution for a 2-1-0 staff configuration with 400 calls/day.

7.2 Impact of staff combinations on the utilization rate of employees

The utilization rate indicates to what extent the employees are productive during a time period. The utilization rate is an important performance indicator because it tends to be low for emergency call centers. This low utilization rate is caused by the fact that the current 21 emergency call center regions cover only small areas. In many of these regions there is a significant overcapacity for the actual call volume. In this section, the utilization rate is analyzed for multiple models with different staff combinations. First, an emergency call center with only generalists is considered, followed by a call center with call takers and dispatchers (i.e., function differentiation). This section is concluded by a hybrid version that considers call takers, dispatchers *and* generalists.

7.2.1 A call center without function differentiation

Figure 49 shows the utilization rate for various arrival rates at a call center with only generalists. As the number of employees increases, the utilization rate decreases non-linearly.

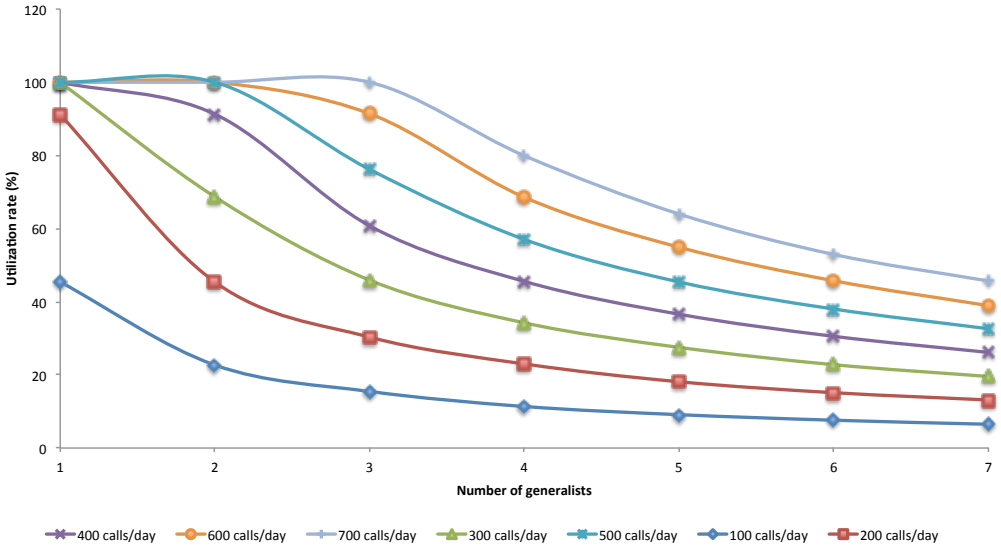


Figure 49: Utilization rate of generalists.

7.2.2 A call center with function differentiation

Figure 50 and 51 show the utilization rate for call takers and dispatchers at a call center with function differentiation. For the dispatchers, it is assumed that there is no bottleneck at the call takers, otherwise the utilization would have been lower for some combinations. Compared to Figure 49, the utilization rate of call takers and dispatchers is much lower compared to the staff combinations with only generalists. Intuitively, this is a logical result, because call takers and dispatchers can handle only specific tasks.

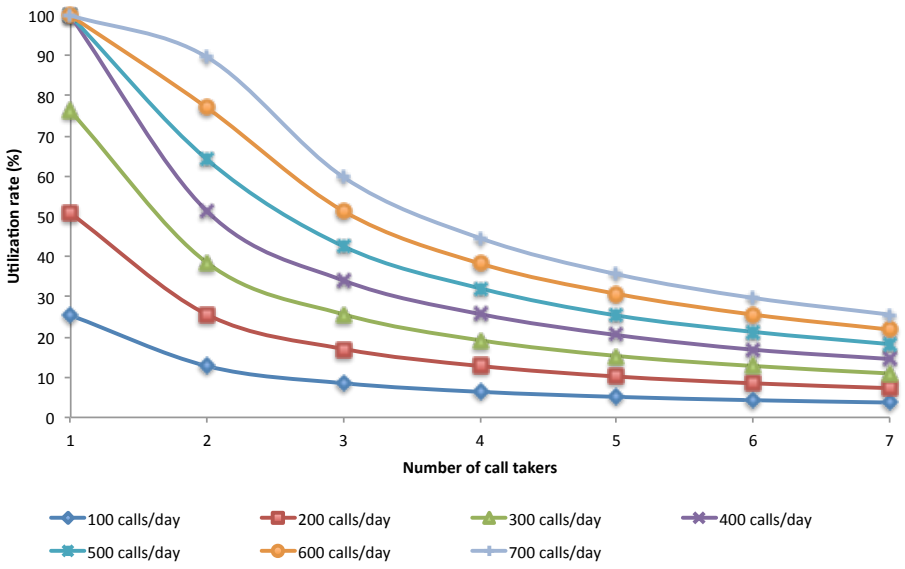


Figure 50: Utilization of call takers.

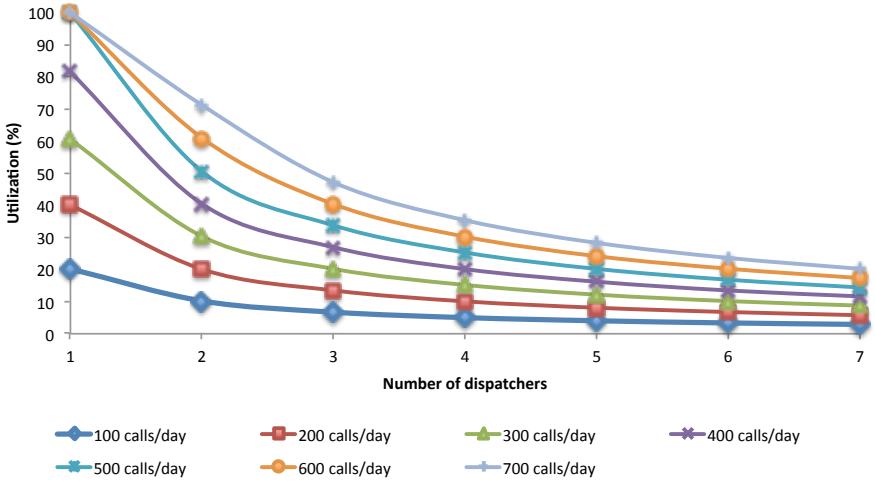


Figure 51: Utilization of dispatchers.

7.2.3 Utilization hybrid model

Figures 52 to 54 show the utilization rate for the hybrid model for several staff combinations. The hybrid model is a combination between a call center with function differentiation model and a call center without function differentiation. This means that there are call takers, dispatchers *and* generalists. The motivation for having this staff configuration is to have higher utilization rate compared to a call center with only call takers and dispatchers, but also to have lower costs. In this setup, the generalists are only used when either the call takers or dispatchers are busy. In other words, the generalist(s) acts like an overflow for triage calls and logistic calls.

Figure 52 shows the utilization of staff combination consisting of 1 call taker, 1 dispatcher and 1 generalist. For low call volumes, the generalist is less busy compared to call takers and dispatchers. However, from roughly 250 calls/day, the overflow coming from call takers and dispatchers increases such that the utilization rate of the generalist gets higher.

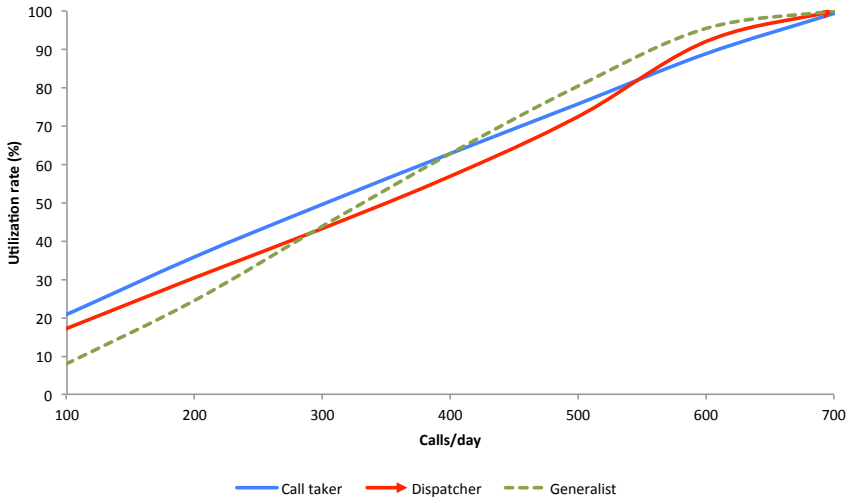


Figure 52: Utilization rate for a 1-1-1 staff configuration.

Figure 53 shows a 1-2-1 combination. In this combination, there is enough capacity to handle logistic tasks (i.e., dispatching and handling follow-up calls), but the capacity to perform triage is now more restricted. The utilization rate of the generalist converges to the one of the call taker, because there is little overflow from the dispatchers.

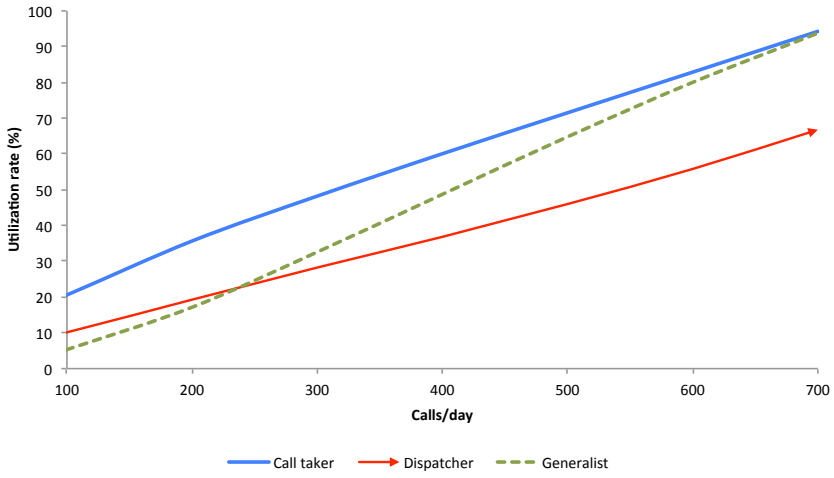


Figure 53: Utilization rate for a 1-2-1 staff configuration.

Figure 54 shows the utilization for a 2-1-1 configuration. It shows the reversed situation of Figure 53, because there is now a overcapacity for triage calls and a bottleneck is created at the logistic part. Therefore, the utilization rate of the generalist now converges to the one of the dispatcher, because there is enough capacity for triage calls.

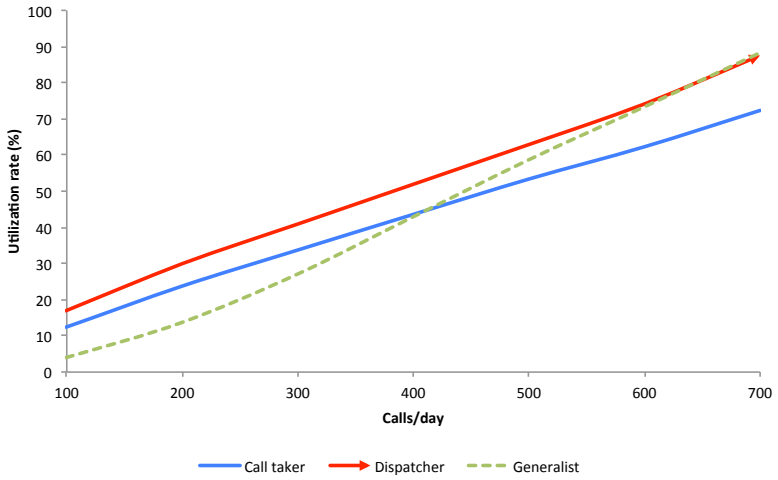


Figure 54: Utilization rate for a 2-1-1 staff configuration.

Conclusions

The main findings of this chapter are:

- Combinations with only generalists have the best overall performance.
- The performance of a staff combination varies per call type. For example, 1-2-0 is better for follow-up calls than for A-calls.
- Function differentiation (i.e., only call takers and dispatchers) is vulnerable for bottle necks. Adding a generalist, can solve these bottlenecks.

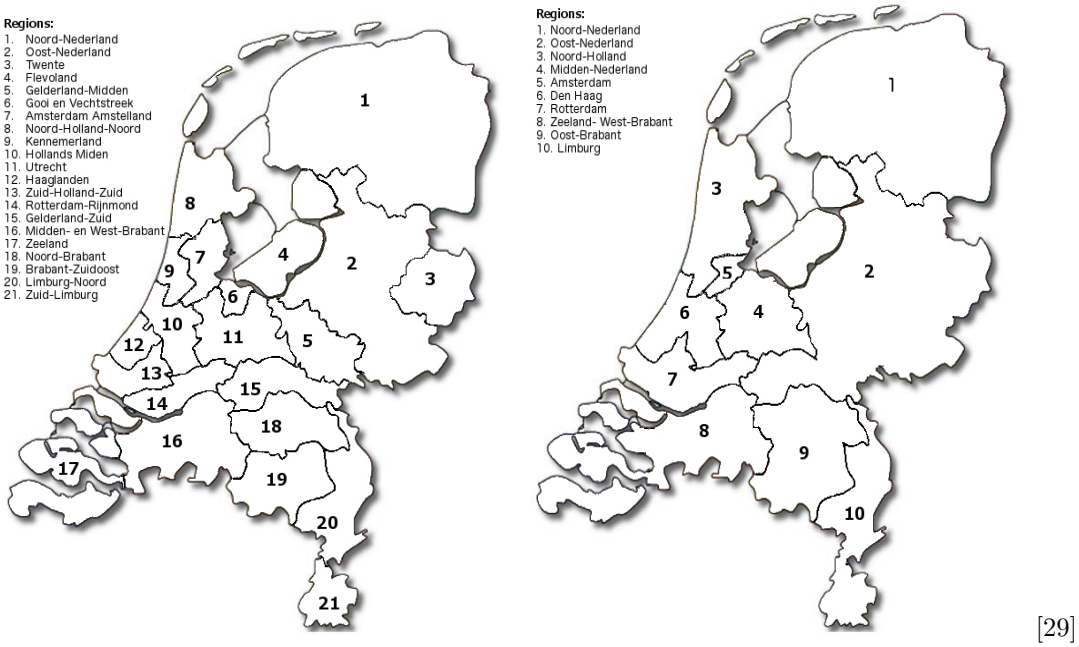
8 Merging emergency call centers

Recently, the government introduced plans for merging emergency call centers. In this chapter, the impact on efficiency of merging call centers is investigated. This chapter is constructed as follows: in Section 8.1, background information is provided on the recent developments of merging emergency call centers. Section 8.2 describes the three different scenarios that have been analyzed. Section 8.3 describes a model for merging call centers *without* taking regional knowledge into account and in Section 8.4 models are provided for including regional knowledge. In these sections, only the emergency call centers of ambulance services are considered, because no data are available of the police and fire department.

8.1 Introduction

Over the past years, a few emergency call centers in the Netherlands have merged. For example, Noord-Nederland is a region covering an area of almost 1.72 million citizens. It is a combined call center, housing the fire department, ambulance department and the police department. Figure 55(a) shows the actual (May 2013) overview of the emergency call centers in the Netherlands. It shows that there are quite many call centers, especially in the west of the Netherlands [24]. The minister of safety and justice, Ivo Opstelten, recently proposed (2012) that there should be one nation-wide organization for all emergency call centers and that the number of call centers should be reduced from 22 to 10 by 2015. These plans originate from 2012 and since then two call centers have merged (IJsselland + Noord- en Oost- Gelderland), which makes the actual number of call centers equal to 21 [30].

The new locations of these call centers will be in Drachten, Apeldoorn, Amsterdam, Den Haag, Rotterdam, Bergen op Zoom, Den Bosch and Maastricht. The locations for Noord-Holland and Midden-Nederland have yet to be determined [24]. Figures 55(a) and 55(b) show the lay-out of the regions in the current and new situation.



[29]

(a) Current regions

(b) New regions

Figure 55: Overview of the current situation and the plan of the government.

Opstelten mentions that the current emergency call centers operate too much independently such there is insufficient coordination among the regions. Moreover, in case of high peak loads, regions have to take care of it by themselves. Another point of his plan is to use multi disciplinary employees who can handle police calls, ambulance calls and fire department calls. They serve as the first point of contact to victims and if they are not able to process a call, it is transferred to a specialized employee.

With this plan, the government wants to achieve a higher efficiency, better cooperation and a higher overall quality of the triage and dispatching process. However, merging call centers means that some local information about the region is lost. For example, if a person calls with a mobile phone to 112 and tells the centralist that he/she is at the Tropen museum in Amsterdam, a centralist coming from Maastricht will probably not know where this is. However, a centralist coming from Amsterdam will probably know this museum and is able to process the call quicker. This issue only plays a role when someone calls with a *mobile phone*, because then the location can not be determined automatically. Nowadays there is software that can support the centralist with determining the exact location, but still the system does not contain all 'soft-knowledge' of the employee. The technique of triangulation could offer the solution to this problem when calling from a mobile phone. With this technique, the location is determined with the use of the signal strength relative to at least three cell phone towers. However, this technique is considered to be too inaccurate for areas with only a few cell phone towers and therefore not practical [31]. A recent development is the integration of GPS into mobile phones. This technique is much more accurate than triangulation, but GPS performs badly in buildings or in tunnels and there are also privacy issues about sharing your position with government institutions. Also for dispatching ambulances and handling follow-up calls, knowledge about the region is important. For example, a dispatcher cannot keep track of all ambulances in more than one region and it takes more time to coordinate with the ambulances.

As it turns out, regional knowledge of the employees is important for handling calls efficiently. It is therefore interesting to explore models that take the regional knowledge aspect into account. When merging call centers, employees could be classified on their knowledge of the region. If a call comes in from region A, the call is preferably directed to an employee with knowledge of region A. However, if all these employees are busy, the call could then be transferred to an available employee with knowledge of another region. However, this leads to inefficient service and therefore a penalty time should be added.

8.2 Analyzing three scenarios

Three scenarios have been analyzed for merging emergency call centers. One of those scenarios is the plan of Opstelten, as mentioned in the introduction. There are no exact data available concerning the arrival frequency of calls for each of the 21 regions. However, the number of A1, A2 and B *dispatched ambulances* per RAV is given in the 'Referentiekader spreiding en beschikbaarheid ambulancezorg 2013' of the RIVM [27]. These numbers underestimate the real call volume at each RAV, because it can happen that multiple calls arrive for dispatching *one* ambulance. Therefore these numbers have been corrected for the number of 'denied calls' (10% of the ambulances that have been dispatched). These corrected numbers are displayed in Table 8 Furthermore, the exact call durations at each RAV are unknown, but it assumed that these durations are the same as at RAVU (Regionale Ambulance Voorziening Utrecht). Further research should be done to what extent the call durations differ per region. Table 8 shows an overview of the estimated call frequencies for the current 21 regions.

		Average number of calls per day			
	Region	A1	A2	B	Total
1	Noord-Nederland	138.8	118.1	109.8	366.7
2	Oost-Nederland	92.7	70.6	62.7	226.0
3	Twente	33.0	41.2	29.7	103.9
4	Gelderland-Midden	47.4	29.6	28.5	105.5
5	Gelderland-Zuid	42.8	27.0	31.8	101.6
6	Utrecht	89.6	67.4	94.9	252.0
7	Noord-Holland Noord	65.5	22.9	28.1	116.5
8	Kennemerland	65.8	18.1	31.8	115.8
9	Amsterdam-Amstelland	179.0	37.9	109.4	326.3
10	Gooi-en Vechtstreek	24.2	6.5	17.6	48.3
11	Haaglanden	118.4	43.6	81.7	243.7
12	Hollands Midden	77.0	28.2	45.3	150.5
13	Rotterdam-Rijnmond	133.1	60.3	98.5	291.9
14	Zuid-Holland Zuid	41.6	21.7	27.7	91.0
15	Zeeland	36.3	20.3	17.3	73.9
16	Midden-en West-Brabant	86.1	60.7	56.1	203.0
17	Brabant-Noord	47.4	35.5	28.3	111.2
18	Brabant-Zuidoost	57.6	33.4	40.2	131.2
19	Limburg-Noord	42.4	31.3	21.5	95.2
20	Zuid-Limburg	53.4	33.3	48.2	134.8
21	Flevoland	37.3	17.2	13.1	67.6

Table 8: Overview of the current 21 regions [27].

8.2.1 Plan of the government: ten call centers

As mentioned in the introduction, the government plans to merge the 21 regions to 10 regions that coincide with the already established police regions. Table 9 shows the call frequencies for these regions. Many of these new regions have more than 300 calls per day and Oost-Nederland even has more than 500 calls per day.

		Average number of calls per day			
	Region	A1	A2	B	Total
1	Noord-Nederland	138.8	118.1	109.8	366.7
2	Oost-Nederland	215.9	168.4	152.7	536.9
3	Noord-Holland	131.3	41.0	60.0	232.3
4	Midden-Nederland	151.1	91.1	125.6	367.9
5	Amsterdam	179.0	37.9	109.4	326.3
6	Den Haag	195.4	71.8	127.0	394.1
7	Rotterdam	174.7	82.0	126.2	382.9
8	Zeeland-West-Brabant	122.4	81.1	73.3	276.8
9	Oost-Brabant	105.0	68.9	68.5	242.4
10	Limburg	95.8	64.6	69.7	230.1

Table 9: Call frequency for the 10 new regions.

8.2.2 Two call centers

Another scenario that has been analyzed is about merging into 2 call centers. Figure 56 shows an overview of the two new regions. The northern region is indicated by blue and the southern region is indicated by red. The composition of the regions has been chosen such that the two region have roughly the same number of calls. In this case, the northern region consists of 1728.5 calls and the southern region consists of 1728.5 calls. Table 10 shows the number of calls per urgency.

Figure 56: North and south region.

		Average number of calls per day			
	Region	A1	A2	B	Total
1	Noord	773.4	429.5	525.7	1728.5
2	Zuid	736.0	395.04	496.5	1627.9

Table 10: Call frequency for the 2 new regions.

8.2.3 A nationwide call center

A nation-wide call center has a few advantages and disadvantages. The primary advantage is that less staff is required than for 21 call centers, due to economies of scale. Also, coordination among centralists is more efficient, because all centralists are housed in the same building. However, as mentioned in the introduction, regional knowledge is lost and this might lead to longer service times due to inefficiency of the centralist. Also, a nationwide call center is vulnerable to equipment failures, which would lead to a total blocking of access to the ambulance services. Table 11 shows the call frequencies of A1, A2 calls and B-calls.

		Average number of calls per day			
	Region	A1	A2	B	Total
1	Nation-wide	1509.4	824.8	1022.2	3356.4

Table 11: Call frequency for a nation-wide call center.

8.3 Merged call centers without regional knowledge

This section starts with a short introduction of the methods used for determining the staffing levels and it is concluded with an overview of the minimum required staffing levels for each scenario.

8.3.1 Using discrete event simulation

Two staff configurations have been considered for each scenario: one with function differentiation (i.e., call takers and dispatchers) and one without function differentiation (i.e., only generalists). It is assumed that all call centers in the Netherlands use the same staff configuration (i.e., all call center are using generalists, or all are using call takers and dispatchers).

As mentioned in Chapter 2, incoming calls can have three different urgencies and each urgency requires different *service level requirements (SLR)*. There are no official guidelines with regard to the SLRs for each urgency level, but a rule for A1- and A2-calls is provided by the RIVM in 'Modellen Referentiekader Ambulancezorg' [17]. This rule says that 95 % of the A1- and A2- calls are required to have a waiting time of less than 6 seconds. For B-calls and follow-up calls there are no SLRs available and therefore reasonable figures have been chosen. These numbers can be seen in Table 12.

	Max. waiting time (sec)	Fraction of calls within max. waiting time.
A1 + A2	6	95%
B	150	90%
Follow-up	30	90%

Table 12: Service levels used for determining staffing levels.

For determining the minimum staffing levels, multiple simulation rounds have been performed. The simulation starts with 1 employee and an employee is added until the service level requirements in Table 12 have been met. This way, the minimum staffing levels have been obtained.

8.3.2 Simulation results

The following figures compare the staffing levels and utilization rates between all scenarios. In appendix A, the staffing levels (for a model with and without function differentiation) and its corresponding performance indicators can be found for each sub region.

Tables 13 and 14 show the minimum number of generalists and the minimum number of call takers and dispatchers for each scenario respectively. When going from 21 regions to 10 regions, significant economies of scale can be gained. These economies of scale tend to fade when the regions get larger. For example, the effect of going from 10 regions to 2 is much less significant compared to going from 21 to 10 call centers. It must be stated that these numbers are based on average call frequencies and that there are no restrictions concerning the shifts of employees. However, useful conclusions can still be drawn by comparing the outcomes of each scenario.

Scenario	Number of generalists	Average utilization rate generalists (%).
21 call centers	50	28.9%
10 call centers	36	41.9%
2 call centers	21	73.2%
1 call center	19	80.8%

Table 13: Utilization rate and the minimum required number of generalists.

	Number of centralists			Average utilization rate %	
Scenario	Call takers	Dispatchers	Total	Call takers	Dispatchers
21 call centers	44	42	**86**	18.8%	15.8%
10 call centers	26	24	**50**	32.4%	28.3 %
2 call centers	14	12	**26**	60.7%	56.8%
1 call center	11	12	**23**	77.7%	56.7%

Table 14: Utilization rate and the minimum required number of call takers and dispatchers.

From Tables 13 and 14 it can be concluded that the difference in minimum required number of employees between function differentiation and without function differentiation decreases when more call centers merge. For example, this difference is 36 employees in the current situation (50 generalists versus 86 call takers and dispatchers), while the difference is only *four* employees in the scenario with one nation-wide call center. The reason for this effect is that the call volumes of the current call centers are too *low* for using call takers and dispatchers, because you need at least 1 call taker and 1 dispatcher and this causes the utilization rate to be low ($< 20\%$). However, as more and more call centers merge, the economies of scale of using call takers and dispatchers increase and function differentiation becomes more effective.

8.4 Merged call centers with regional knowledge

In the introduction, a model was proposed for taking the aspect of regional knowledge into account. In this section, the model is further described and its effect on efficiency is discussed. In the first part, an analytical model is presented for merging call centers of two regions. In the second part, a simulation model is presented for merging more than 2 regions.

8.4.1 Analytical method for two call centers

Including regional knowledge into a emergency call center can be seen as an (financial) optimization problem. A penalty is given for each call that is being redirected to another region. Moreover, each employee also has a certain salary. The goal is to determine the minimum staffing levels such that the total costs are minimized and the blocking probability is under a certain threshold. P_{b1} represents the probability that a call from region 1 is redirected to region 2 and P_{b2} is the probability that a call from region 2 is redirected to region 1. Furthermore, $P_{b1,2}$ is defined as the probability that a call is blocked. Figure 57 shows how a co-operating call center works compared to two stand-alone call centers. In a cooperating call center the expected number of calls that are blocked is equal to $(\lambda_1 + \lambda_2)P_{b1,2}$. For the two stand-alone call centers, the expected number of blocked calls is equal to $P_{b1}\lambda_1$ and $P_{b2}\lambda_2$ respectively.

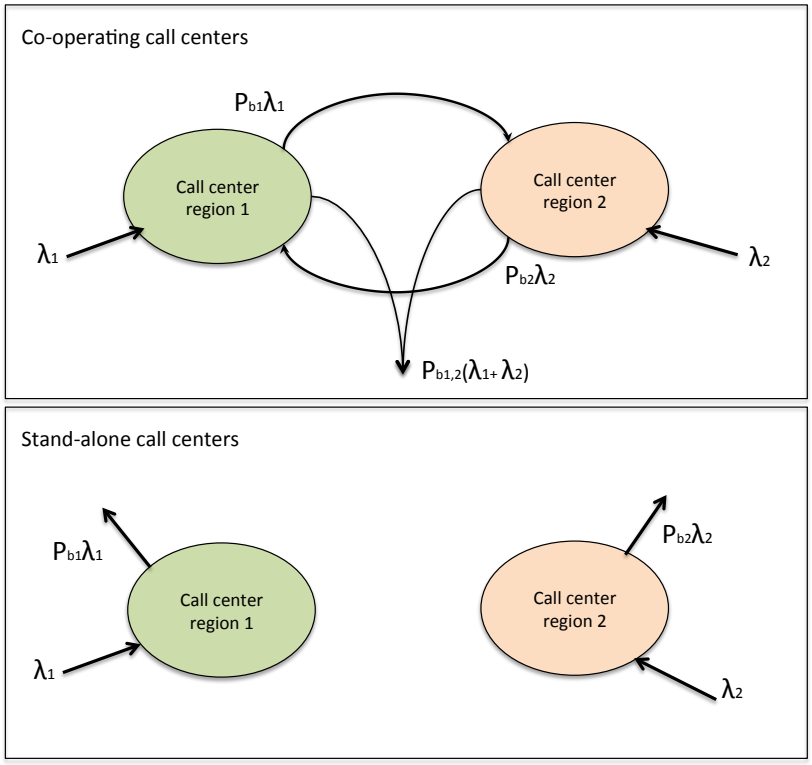


Figure 57: Overview of the different call center types.

In this model only A1, A2 calls are considered, because the staffing levels are mainly influenced by the service level constraint of this group. First, the model is formulated, followed by a few numerical

examples. For the model, the following decision variables and constants are needed:

Decision variables

N = the number of employees with knowledge about region 1.

M = the number of employees with knowledge about region 2.

Constants

S_1 = salary of employee with knowledge about region 1.

S_2 = salary of employee with knowledge about region 2.

B_1 = penalty of handling a call from region 1 by an employee with knowledge about region 2.

B_2 = penalty of handling a call from region 2 by an employee with knowledge about region 1.

λ_1 = arrival frequency per hour of calls from region 1.

λ_2 = arrival frequency per hour of calls from region 2.

μ_1 = service rate of employees with knowledge about region 1.

μ_2 = service rate of employees with knowledge about region 2.

α = maximum blocking probability that is tolerated.

Optimization formulation

$$C^* = \min_{N,M} P_{b1}\lambda_1 B_1 + P_{b2}\lambda_2 B_2 + S_1 N + S_2 M$$

$$S.T. \quad P_{b1,2} < \alpha$$

The blocking probabilities P_{b1}, P_{b2} and $P_{b1,2}$ are a function of λ_1, λ_2, μ_1, μ_2, N, M and can be calculated by solving the following Markov chain displayed in Figure 58. Furthermore, (x_1, x_2) = state in which there are x_1 calls of region 1 and x_2 calls of region 2 in the call center.

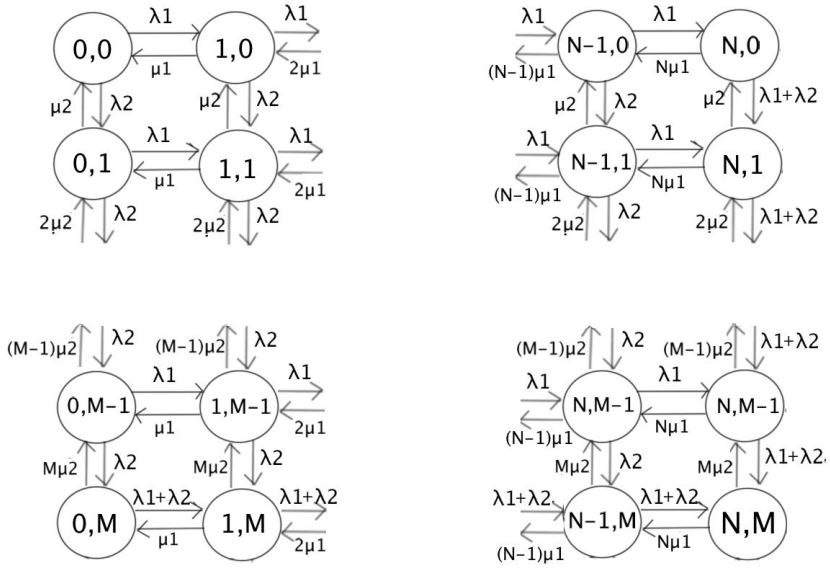


Figure 58: Markov chain for two co-operating call centers.

Assumptions
The following assumptions have been made for this model:

1. A call is blocked when every employee is busy. This assumption does not reflect the real situation, because in practice calls would have entered the queue.
2. If a call is redirected to an employee of an other region, the call duration is just as long as it would have been processed by an employee of the preferred region. It is also assumed that the service rate is equal for all calls, regardless of the urgency and origin of the call.
3. Calls arrive according a Poisson process. As mentioned earlier in Section 2.1, this is a valid assumption for emergency call centers, because there is a big population of potential callers in both regions.
4. Call durations have been modeled with an exponential distribution instead of a lognormal distribution.
5. Both call centers work without function differentiation. In the next section, call centers with function differentiation are considered, using simulation.

Measuring the economies of scale In order to draw conclusions about the economies of scale for cooperating call centers, the costs have to be compared with those of the stand alone call centers. Modeling this type of call centers is straight forward, because it is the well-known Erlang-B model, in which calls are blocked if all employees are busy. The optimization formulas of each stand-alone call center become:

For stand alone call center 1:

$C_1^* = \min_N S_1 N$

$S.T. \;\; P_{b^*1} \leq \alpha$

$P_{b^*1} = \dfrac{\frac{(\frac{\lambda_1}{\mu_1})^N}{N!}}{\sum_{i=1}^{N} \frac{(\frac{\lambda_1}{\mu_1})^i}{i!}}$, the probability that a call is blocked

For standalone call center 2:

$C_2^* = \min_M S_2 M$

$S.T. \;\; P_{b^*2} \leq \alpha$

$P_{b^*2} = \dfrac{\frac{(\frac{\lambda_2}{\mu_2})^M}{M!}}{\sum_{i=1}^{M} \frac{(\frac{\lambda_2}{\mu_2})^i}{i!}}$, the probability that a call is blocked

Then, the economies of scale for coopering call center can be expressed as:

Economies of scale (EOS):

$EOS = (C_1^* + C_2^*) - C^*$

In which:

C^*= the optimal costs for a merged call center.

C_1^*= the optimal costs for stand alone call center 1.

C_2^*= the optimal costs for stand alone call center 2.

Parameters The parameter estimates are based on the data from RAVU. To analyze multiple merging regions, linear scaling is applied as an approximation for the arrival rates. The underlying assumption is that the call frequency is linear with regard to the size of population and independent of special characteristics of the region (i.e., rural area, number of hospitals, etc.).

In order to keep the model compact, there will only be looked at A1 and A2 calls. The weighted average service duration of these calls is 2.93 minutes, including 1 minute of dispatching. This means that a centralist can handle 20.41 A1, A2 calls/hour. The average arrival frequency is 4.78 calls/hour for the Utrecht region. Using linear scaling, the arrival frequency is 3.94 calls/hour/1000.000 citizens. The maximum tolerated blocking probability, α, is set to 0.05.

$\mu_1 = 20.41$ 112 calls/hr/employee
$\lambda_1 = 3.94$ A1,A2 calls/hr/1.000.000 citizens.

Results
The following three scenarios have been analyzed:

1. Merging two call centers with equal call volumes, equal penalty and salary costs.
2. Merging two call centers with different call volumes but equal penalty and salary costs.
3. Merging two call centers with low penalty costs.

The last paragraph contains a comparison between the economies of scale for the different scenarios.

Scenario 1: merging two call centers with equal call volumes In this scenario, two identical call centers co-operate together. The parameters used for this scenario are displayed in Table 15. Figure 15 shows the economies of scale (blue dashed line) and the staffing levels (red lines) for the two call centers with equal arrival rates. The blue axis (on the left) displays the economies of scale and the red axis (on the right) display the number of employees.

	Call center 1	**Call center 2**
Penalty costs/call	20	20
Salary costs/employee	20	20
Arrival rate /hr	3.94	3.94
Service rate /hr	20.41	20.41

Table 15: Parameters for scenario 1.

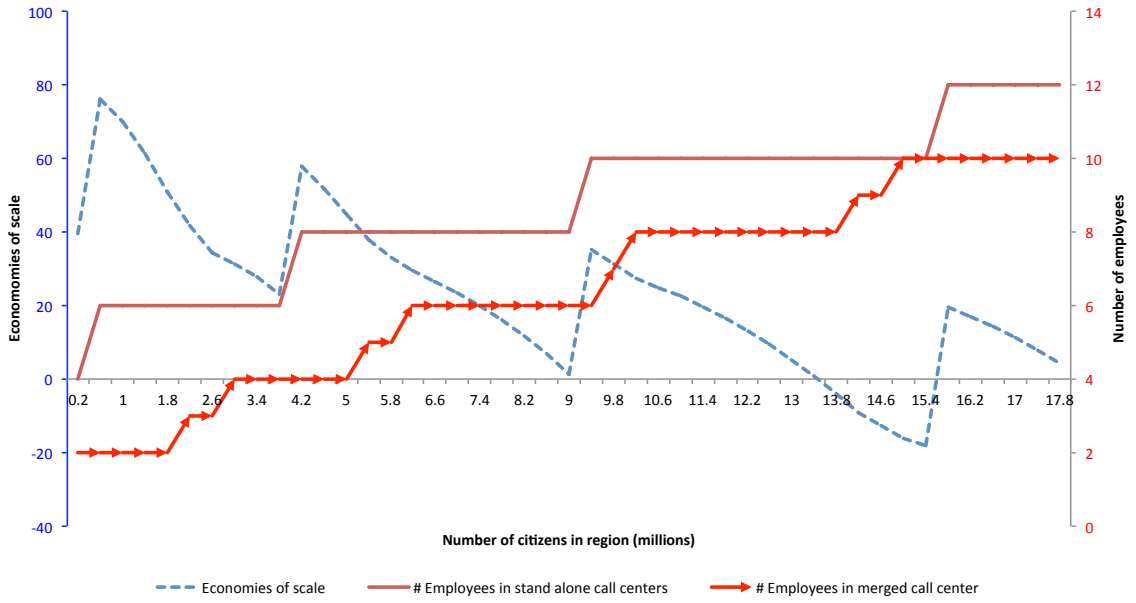


Figure 59: Two call centers with equal arrival rates, equal penalty costs and equal salary costs.

The dashed blue line goes up and down several times. These peaks coincide with increases in the staffing levels for the stand alone call centers. At these peaks, the blocking probability of both call centers falls under the 5% and therefore the stand-alone call centers need to have one extra employee, causing a significant increase in salary costs. After each peak, the economies of scale decrease because the blocking probability increases when the number of citizens in the region increases. Another interesting aspect is that the economies of scale disappear for a merged call center that covers an area of more than 11 million citizens. The reason for this effect is that the stand-alone call centers are now that big, such that the stand-alone call centers work just as efficiently as the merged call center and also penalty costs do not occur for these stand alone call centers. However, the aspect of regional knowledge is probably lost completely for such big regions and therefore this situation is unrealistic.

Scenario 2: merging two call centers with different call volumes In the second scenario, the call volume of call center 2 is 50% smaller compared to the call volume of call center 1. All the other parameters have remained the same. Table 16 shows an overview of the input parameters.

	Call center 1	**Call center 2**
Penalty costs/call	20	20
Salary costs	20	20
Arrival rate /hr	3.94	1.97
Service rate /hr	20.41	20.41

Table 16: Parameters for scenario 2.

Figure 60 shows the economies of scale for merged regions up to 13 million citizens. The shape of the blue dashed line is quite different compared to the line in Figure 59 and it shows more peaks. The reason for this effect is that the staffing levels of the two stand-alone call centers now increase asynchronously, due to the difference in call volumes.

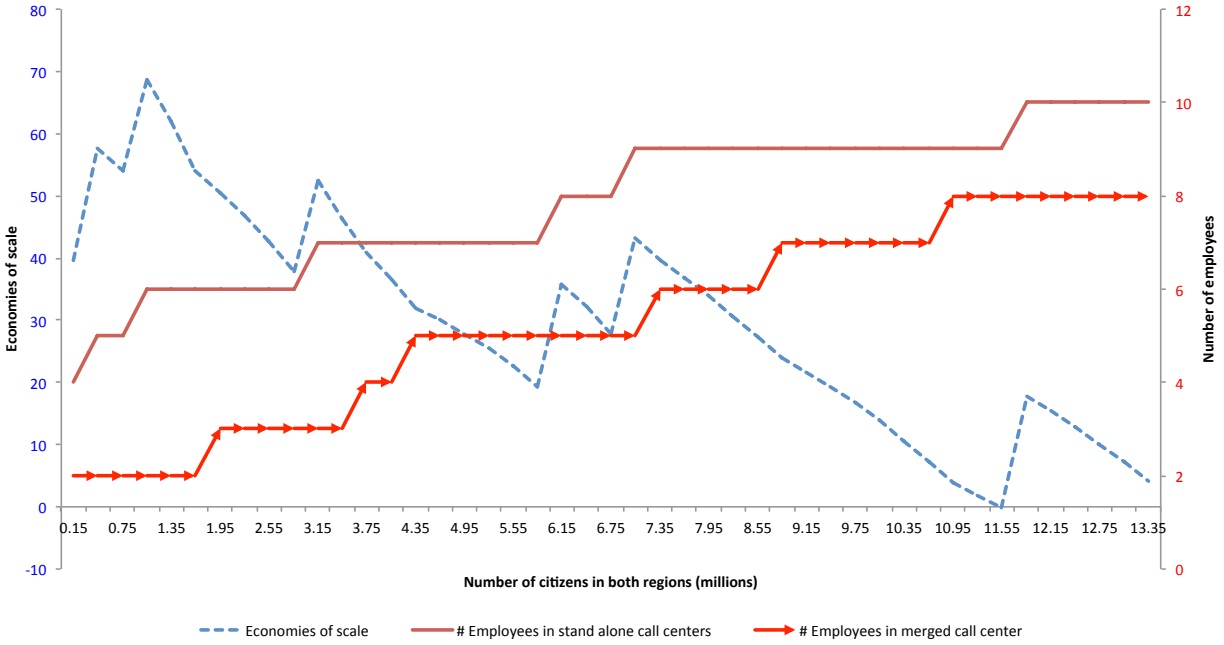


Figure 60: Two call centers with unequal arrival rates, equal penalty costs and equal salary costs.

Scenario 3: merging two call centers with low penalty costs In this scenario, the penalty costs are set 10 times smaller than the salary costs. This scenario might be more realistic compared to the costs of the previous scenarios. Table 17 shows the values of the costs, arrival- and service rates.

	Call center 1	**Call center 2**
Penalty costs/call	2	2
Salary costs	20	20
Arrival rate /hr	3.94	3.94
Service rate /hr	20.41	20.41

Table 17: Parameters for scenario 3.

Figure 61 shows the economies of scale for merged regions up to 18 million citizens. The line of the economies of scale now shows a different behavior compared to Figures 59 and 60. The economies of scale prove to be much higher for this scenario, because the penalty costs are low. The line shows several peaks at moments when the two stand alone call centers need an extra employee. As the number of citizens in the region increases, the economies of scale decrease slowly due to the low penalty costs. However, when the merged call center needs an extra employee, the economies of scale decrease significantly.

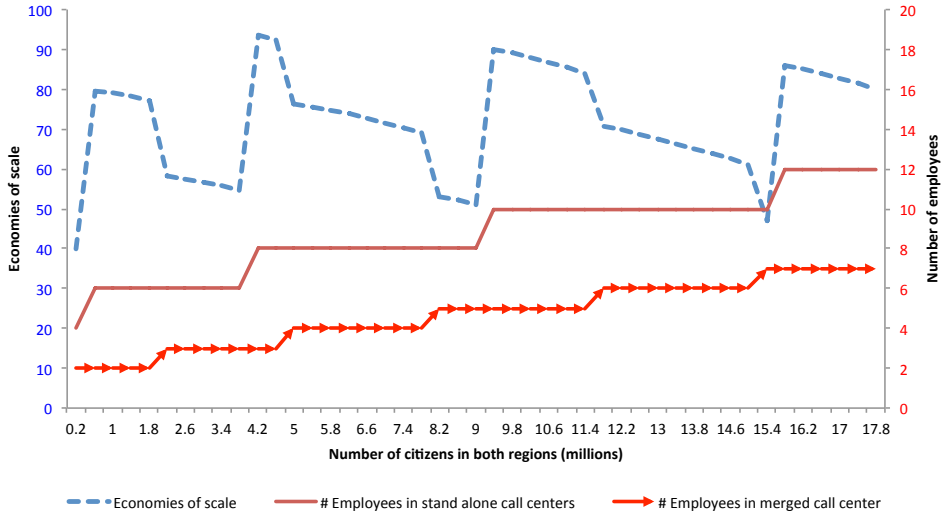


Figure 61: Two call centers with equal arrival rates and low penalty costs.

Analysis of the costs Figure 62 shows the economies of scale for different combinations of salary costs and penalty costs. The blue line displays the scenario in which the penalty costs are equal to the salary costs and this scenario has the worst performance compared to the other scenarios. The green line shows that the economies of scale don't decrease much when penalty costs are set low. The dashed red line shows the economies of scale for a call center with high salary costs and it shows several peaks when the stand alone call centers require an extra employee to meet the blocking probability constraint.

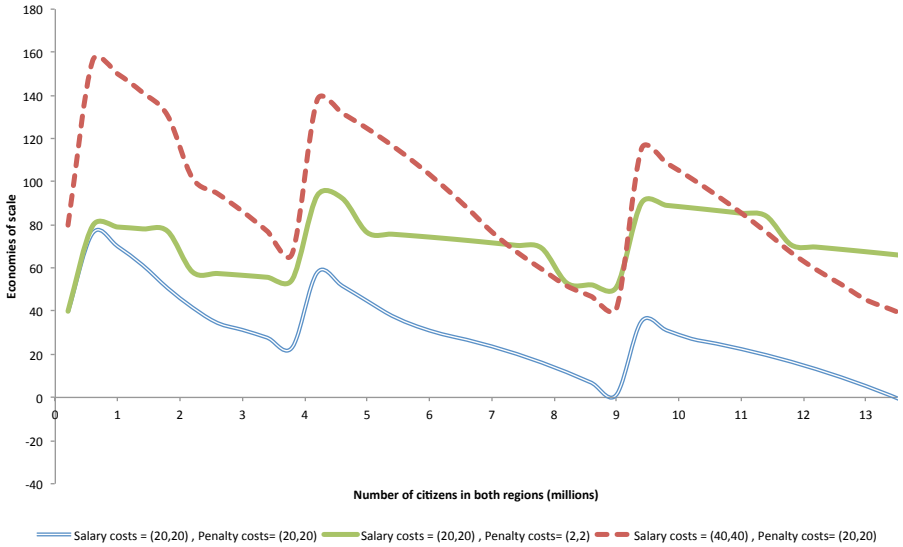


Figure 62: Impact on the ecnomies of scale for different costs.

From this figure it can be concluded that when the salary costs of employees outweigh the penalty costs for redirecting a call, call centers with regional knowledge operate much more efficiently than two stand alone call centers. However, this effect fades out when the stand alone call centers get bigger in

size, because they will also operate more efficiently. Also, it turns out that for specific call volumes the economies of scale are significantly bigger. At these volumes, the stand alone call centers require extra staff because the blocking probability falls below the 5%, while the merged call center does not require any additional employees. In order to make optimal usage of the economies of scale, it is beneficial to merge call centers with these call volumes.

8.4.2 Using discrete event simulation

As mentioned earlier, one of the disadvantages of merging call centers is that knowledge of the region is lost. This may lead to a loss of quality and to a lower efficiency. This section addresses this problem by introducing a call center with region-skilled staff. This means that each employee has a certain geographical knowledge of a region.

First, a framework for a call center with multi-skilled staff is explained. Second, the method of estimating the minimum required staffing levels is explained. This section is concluded with results regarding the staffing levels for the three different scenarios.

Staff with skills based on their region knowledge Suppose there is a merged call center, covering multiple regions. In order to preserve the regional knowledge of the employees, the merged call center must have some employees with knowledge of each sub region. Figure 63 provides an illustration of how such a call center would operate. For example, when a 112 (mobile phone) call comes in from Noord-Nederland, it is directed to the 112 call center in Driebergen. If the call is approved, it is directed to the merged call center of Noord-Nederland, Oost-Nederland and Twente. The call is then allocated to the employee with knowledge of that region. However, when all employees of that region are busy, the call is directed to an employee of another region. This will however lead to a longer service times, because this employee is not familiar with the region. For the triage part, it means that it takes longer to determine the location of a 112 call. For the logistic part it will take more time for the dispatcher to get an overview of the available ambulances in another region. This extra time has been modeled as a stochastic variable with an exponential distribution. For the triage, the penalty is only added to 112 calls and to the logistic calls. The results in the last part of the section are based on an expected penalty time of 0.5 minutes for triage and logistic processing.

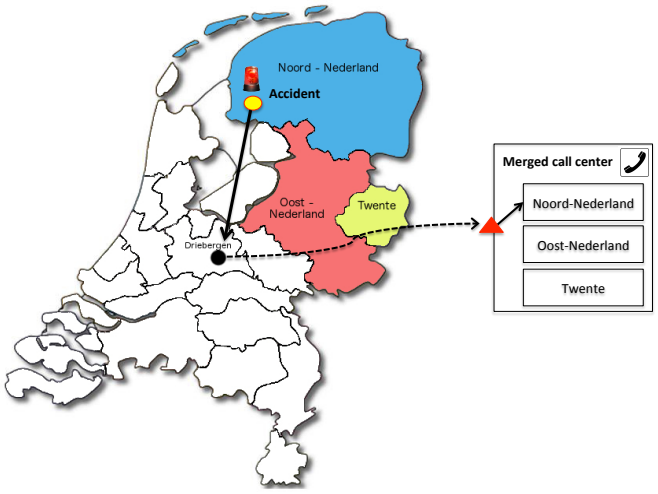


Figure 63: An illustration of how a call center with regional knowledge works.

Estimating the minimum required staffing levels In order to preserve the regional knowledge in a merged call center, an upper bound is imposed on the fraction calls that is routed to other regions. This fraction is computed using the number of triage and logistic redirections, divided by the total number of logistic and triage calls. The minimum number of employees required depends on this fraction. When this fraction is set to 100%, full economies of scale are gained.

The minimum required staffing levels also depend on the service level constraints for each call type. Table 18 shows an overview of these minimum required service levels for each urgency level.

	Max. waiting time (sec)	**Service level requirement**
A1 + A2	6	95%
B	150	90%
Follow-up	30	90%

Table 18: Service levels used for determining staffing levels.

The minimum staffing levels are computed by using simulation. In each iteration of this algorithm, an employee is allocated to the region with the highest fraction of redirections. The effect of this increment on both the service level and the maximum fraction of redirections is evaluated using simulation. Based on these results, an extra employee might be added.

8.4.3 Simulation results

In this section, the staffing levels for the current situation and other three scenarios are provided. Two variants have been studied: a call center *with* function differentiation and one *without* function differentiation. Figure 64 shows the total number of generalists required to make sure that no more than 20%, 40%, 60%, 80% and 100% of the calls are redirected to another region. Figure 65 shows a similar plot, but now for call centers using function differentiation. For all these staffing levels, the service level requirements of A-calls, B-calls and follow-up calls have been met. The current staffing levels (blue line) are insensitive to the maximum fraction of redirected calls because it is assumed that the aspect of regional knowledge is not an issue in the current situation.

Figure 64 shows that when the maximum allowed fraction of redirections is set 20%, the minimum required amount of employees for 10 regions exceeds the amount of the current 21 call centers. This result might seem contradicting at first, but when the fraction of allowed redirections is set too low, the number of employees of each region is increased until (almost) no redirections occur. This automatically means that calls are not put in the queue, because this happens only when every employee (regardless the region) is busy.

Both Figures 64 and 65 show that the economies of scales are fully utilized when the maximum fraction of redirected calls is set to 100%. However, the aspect of region knowledge is then virtually ignored. Compared to the staffing levels without regional knowledge (Section 8.3), the staffing levels are still higher because there is a penalty time involved.

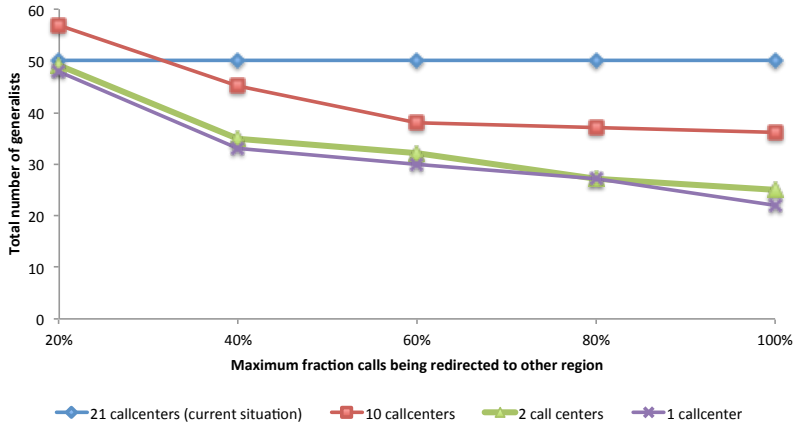


Figure 64: The impact of the fraction of redirections on the staffing levels for a call center without function differentiation.

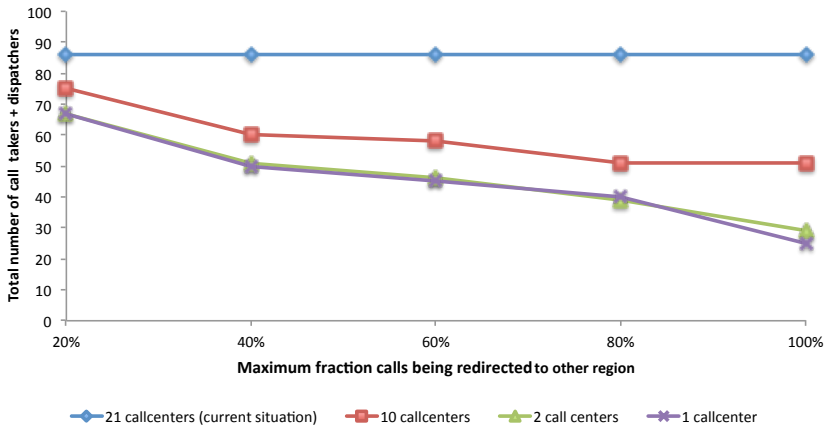


Figure 65: The impact of the fraction of redirections on the staffing levels for a call center with function differentiation.

Conclusions

The main findings of this chapter are:

- The call volumes of the current 21 call centers *too low* to operate with call takers and dispatchers and therefore it is better to use generalists in these call centers. However, as call centers get bigger it becomes more beneficial to use function differentiation
- Merging call centers leads to significant staff reductions and and therefore an increase in efficiency. The plan of the government of going from 21 to 10 call centers will lead to a significant improvement in efficiency, but even more can be achieved by merging into one or two call center(s).
- When regional knowledge is included in the model, efficiency can still be gained when call centers are merged.

9 Conclusions

In the introduction, the following research questions have been posed:

'What is the impact of staffing on the performance and efficiency of emergency call centers for ambulance services?'

Sub questions:

- What is the impact of different staff combinations on the performance indicators?
- What is the impact of merging call centers on the efficiency?

In this chapter answers are given to the research questions, using the results of the previous chapters. This chapter is constructed as follows: in the first section, arguments are given about when analytical methods should be used instead of simulation. The second section provides the main results of the sensitivity analysis and the applicability of the model when dealing with uncertain parameters. The third section discusses the impact of staffing combinations on the performance indicators and is followed by a section that describes the impact on efficiency when call centers are merged. This chapter ends with a few recommendations and suggestions for further research.

9.1 Applicability of analytical methods for measuring the performance in emergency call centers

Analytical methods have the advantage that they require less computational power compared to simulation and they are in general easier to implement. However, when a model gets big and complicated, analytical methods are hard to apply. For an emergency call center, the goal is to provide an analytical approximation of the fraction calls that is handled within t seconds. In order to achieve this, the waiting time distribution of a multi- server queue with log-normal distributed call durations and Poisson arrivals has been approximated. In Chapter 6, three approximation methods have been discussed and their performance is evaluated using simulation. The first method is based on the first two moments of the call duration distribution, while the second method is based on renewal theory. The third approximation method is the Erlang C queueing system, which is a well-known system in queueing theory and an exact expression for the waiting time distribution is available. These three methods have been applied to both a call center with function differentiation and a call center without function differentiation.

It turns out that analytical methods provide quite good approximations for measuring the performance of A1 and A2 calls. However, it is much more difficult to approximate the performance of B-calls and follow-up calls because they are dependent on calls with higher priority. For follow-up calls, only the expected waiting time could be computed because no methods are available for approximating the entire waiting time distribution. This expected waiting time calls been calculated using a two dimensional Markov-chain. For low loads, this approximation turns out to be a good approximation but some errors occur as the load increases. This is caused by the fact that the arrival process is not entirely Poisson.

One of the surprising results is that the Erlang-C system also proves to be quite a good approximation for both a call center with and without function differentiation, while it has not been designed for approximating an M/G/C queue. However, compared to the other methods, it shows a higher absolute error.

It remains the question when simulation or analytical methods should be used instead of simulation. The analytical methods show some errors, especially when the load gets high. For measuring the

performance of A1, A2 calls and doing quick capacity calculations, analytical methods are more suited. On the other hand, simulation mimics reality better, because it is also able to analyze the performance of all call types (instead of only A1,A2). Moreover, the approximation methods presented in Chapter 6 are not able to handle a hybrid staffing configuration (i.e., staff consisting of call takers, dispatchers *and* generalists). In general, simulation requires more computational power and consumes more time compared to analytical methods. However, the analysis in Chapter 4 shows that the simulation already converges after only simulating a few days. The reason for this quick convergence is the fact that emergency call centers are light-traffic systems. All in all, simulation is a better option for measuring the performance of emergency call center, because it provides more accurate results and mimics reality better.

9.2 Robustness of the model

In order to test the robustness of the simulation model to its parameters, a sensitivity analysis has been conducted. The purpose of this sensitivity analysis is to provide a better understanding of what would happen when parameters have been under- or over estimated. In Chapter 5, the call durations and the fraction of follow-up calls have been varied and their effect on the performance indicators is displayed in Figures 19 to 25. One of the main conclusions is that A-calls (i.e., A1+A2 calls) are the least sensitive to changes the input parameters, because these calls have a higher priority compared to B-calls and follow-up calls. The sensitivity of B-calls and follow-up calls tends to be higher because these calls are dependent on calls with higher priority. This is a positive result, because the performance of A-calls is more relevant than of B-calls and follow-up calls. Based on these results, it can be said that the model is quite robust, especially for measuring the performance of A-calls. For this analysis it must be noted that only one parameter has been changed at a time. Further research should investigate the impact on the performance indicators when more than one parameter is changed.

9.3 Impact of staffing combinations on the performance indicators

In Chapter 7, three types of call centers have been analyzed: one without function differentiation, one with function differentiation and a hybrid version. In a call center without function differentiation, only generalists are used who are able to perform any task. However, in a call center with function differentiation, two types of employees are used: the call taker and dispatcher. The call taker performs triage and the dispatcher dispatches ambulances and processes follow-up calls. The benefit of splitting the tasks of generalists, is that dispatchers have lower wages compared to call takers and generalists, because dispatchers do not need to have any medical knowledge. The hybrid call center uses call takers, dispatchers *and* generalists. In this report, the generalist is used as an overflow when either the call takers or dispatchers are busy.

The results in Chapter 7 show that the staffing configuration within a call center has a significant impact on the performance of A-calls (i.e., A1+A2 calls), B-calls and follow-up calls. The performance of a staff configuration is measured using the fraction of calls that have a waiting time less than t seconds. In general, all staffing configurations that involve call takers and dispatchers have less overall performance compared to staffing configurations with only generalists, because generalists are more flexible. This overall performance is based on the performance for A-calls, B-calls and follow-up calls. When designing a call center, attention must be paid to bottlenecks. For example, when the number of dispatchers is set low and the number of call takers is set high, a bottleneck is created at the logistic processing part. In this situation, the performance for each call type is poor. In case of a hybrid call center, bottle necks are avoided somehow because the generalist acts like an overflow when either the call taker or dispatcher are busy.

9.4 Impact of merging call centers on efficiency

At the moment, merging emergency call centers is a hot topic in the Netherlands. Recently, the government made a plan of merging 21 regional call centers into 10 call centers. In Chapter 8, this plan and two other scenarios have been evaluated. The other two scenarios that have been analyzed, are about merging 21 call into two call centers and one call center respectively. When merging emergency call centers, a common issue is that knowledge about the regions is lost, causing a decrease in efficiency and quality of service. In Chapter 8, two ways of merging call centers have been introduced: merging without preserving regional knowledge and merging with preserving regional knowledge. In the latter case, incoming calls are preferably directed to an employee of the same region, but when that employee is busy the call is redirected to an employee of another region. In that case, a penalty time is added to the call duration in order to take the loss of efficiency into account. The minimum staffing levels for each scenario have been determined using service level constraints for each call type (i.e., A1, A2-calls, B-calls and follow-up calls) and a constraint regarding the maximum fraction calls that is allowed to be redirected to another region. The following two paragraphs display the staffing levels for both variants.

Staffing levels for a model without preserving regional knowlegde The minimum required number of centralists has been computed using the following service level constraints:

- 95% of the A1,A2 calls that must be answered within 6 seconds.
- 90% of the B-calls that must be answered within 150 seconds.
- 90% of the follow-up calls that must be answered within 30 seconds.

Tables 19 and 20 show the average utilization rate and the minimum required staffing levels for a call center without function differentiation (i.e., only generalists) and a call center with function differentiation (i.e., call takers and dispatchers) respectively.

Scenario	**Number of generalists**	**Average utilization rate generalists (%).**
21 call centers	50	28.9%
10 call centers	36	41.9%
2 call centers	21	73.2%
1 call center	19	80.8%

Table 19: Utilization rate and the minimum required number of generalists.

	Number of centralists			**Average utilization rate %**	
Scenario	Call takers	Dispatchers	Total	Call takers	Dispatchers
21 call centers	44	42	**86**	18.8%	15.8%
10 call centers	26	24	**50**	32.4%	28.3 %
2 call centers	14	12	**26**	60.7%	56.8%
1 call center	11	12	**23**	77.7%	56.7%

Table 20: Utilization rate and the minimum required number of call takers and dispatchers.

Both tables show that much efficiency can be gained when emergency call centers are merged. If the plan of the government would be implemented, a staff reduction of 41.8 % would be achieved for a call center with function differentiation and a 28 % staff reduction for a call center without function differentiation. When every call center in the Netherlands would use function differentiation, 86 employees are needed to meet the service level requirements. This is a lot more than the 50 generalists that are needed in case all call centers would operate without function differentiation. As

the call centers get bigger, this effect diminishes and it becomes more advantageous to use function differentiation. The reason for this effect is that the call frequency for each of the current 21 call centers is *too low* to use function differentiation.

Merged call center in which regional knowledge is preserved Figures 64 and 65 in Section 8.4, show the staffing levels for a call center with and without function differentiation when regional knowledge is taken into account. These minimum staffing levels are based on the service level requirements mentioned earlier, plus an additional constraint for the maximum fraction of calls that is allowed to be redirected. The staffing levels are higher compared to the model when regional knowledge is not taken into account, because a penalty time is added when calls are redirected. In a call center without function differentiation, efficiency is gained when the fraction of redirections is set higher than 20%. For a call center with function differentiation, the threshold is at 40%. In Section 8.4, an analytical method has been introduced for merging two call centers. Although this model has some limitations, it stills provide a good understanding of how the economies of scale are correlated with different call volumes. Figure 62 shows that the economies of scale are maximized at some specific call volumes. In order to make optimal usage of the economies of scale, it is beneficial to merge call centers with these call volumes.

9.5 Recommendations

Based on the results in this research, the following recommendations are suggested:

- Merging call centers leads to staff reductions and and therefore an increase in efficiency. The plan of the government of going from 21 to 10 call centers will lead to a significant improvement in efficiency, but even more can be achieved by merging into one or two call center(s).
- Efficiency can still be gained, when regional knowledge is included in the model.
- The call volumes of the current 21 call centers *too low* to operate with call takers and dispatchers and therefore it is better to use generalists in these call centers. However, as call centers get bigger it becomes more beneficial to use function differentiation.
- Even when parameters have been under- or over estimated, the performance of A1- and A2-calls proves to be quite robust. However, for B-calls and follow-up calls, the performance indicators are less robust.
- Simulation is the most accurate and flexible way to analyze emergency call centers. Analytical approximations are quicker and take less time to implement, but these are less accurate. These methods are more suitable for doing quick capacity calculations.
- A user-friendly simulation tool has been specially developed for supporting decisions on tactical and strategic level at emergency call centers. The tool is an excellent way of measuring the performance and efficiency of an emergency call center.

9.6 Further research

The results presented in this report should be validated furthermore by using data of other regions. The data used in this report only concern the Utrecht region, which suffices for evaluating the performance of the emergency call center in Utrecht. However, when analyzing the impact of merged call centers on performance and efficiency, the data of other regions should also be analyzed.

The results in this report display the *average* staffing levels for emergency call centers. However, these staffing levels are not suitable for determining staffing levels on operational level, because the aspects of time and availability of staff are not taken into account. Further research should include these

aspects for usage on operational level.

In this report a model has been presented that models the processes within an emergency call center for *ambulance services*. Recent plans of the government reveal that the collaboration between the fire department, police department and ambulance services has to intensify. The ultimate goal is to have an all-round skilled employee who is able to handle calls for the ambulance, police and fire department [30]. Calculating the staffing levels for such a call center, involves a whole different model because the processes in each department are different. It is expected that even more economies of scale can be obtained when using a multi-skilled employee for the police, fire services and ambulance service.

In Chapter 8, a model has been introduced that takes regional knowledge into account when emergency call centers are merged. In this model, an employee had only knowledge of *one* region. This might not be realistic, because employees can have knowledge of multiple regions. Further research should investigate multi-skilled employees.

The financial aspects of merging emergency call center have not really been discussed in this report. A cost-benefit analysis could be considered in which is determined how many years it would take until a merged call center becomes profitable. The cost-benefit analysis could then be performed for each of the three scenarios described in Section 8.2.

References

[1] Chin, G., and Coke, J. (2008)
'Simulation Modeling for Staff Optimization of the Toronto Emergency Medical Services Call Centre.'
http://www.mie.utoronto.ca/undergrad/thesis-catalog/files/163.pdf

[2] Aksin, O.Z., Karaesmen, F., and Ormeci, E.L. 'A Review of Workforce Cross-Training in Call Centers from an Operations Management Perspective', Workforce Cross Training Handbook, ed. D. Nembhard, CRC Press, 2007.

[3] Gans, N., Koole, G., Mandelbaum, A. 'Telephone Call Centers: Tutorial, Review and Research Prospects.' Invited review paper by Manufacturing and Service Operations Management (M&SOM), 5 (2), 2003.

[4] G.P. Cosmetatos, 'Some approximate equilibrium results for the multi-server queue (M/G/r)' Oper. Res. Quarterly 27 (1976)

[5] Franx, G.J A. - Simple solution for the M/D/ c waiting time distribution

Operations Research Letters, 2001, Vol.29(5), pp.221-229

[6] Crommelin, C.D. 'Delay probability formulas when the holding times are constant.' P.O. Electr. Engr. J. 25: 4150

[7] Ward Whitt, 'Approximations for the GI/G/m Queue.' Production and Operations Management, vol. 2, No. 2, 1993, pp. 114-161 AT&T Bell Laboratories

[8] Boxma, O.J., Cohen, J.W. & Huffels, N. (1979). 'Approximations of the mean waiting time in an M/G/s queueing system.' Operations Research, 27(6), 1115-1127.

[9] A. Al Hanbali, E.M. Alvarez, M.C. van der Heijden 'Approximations for the waiting time distribution in an M/G/c priority queue' Beta Working Paper series 411

[10] M.H. Van Hoorn & H.C Tijms 'Approximations for the waiting time distribution of the M/G/c queue' 1980 - Stichting mathematisch centrum
http://oai.cwi.nl/oai/asset/9625/9625A.pdf

[11] H. Tijms, 'A First Course in Stochastic Models', Vrije Universiteit, Amsterdam

[12] Burke, P. J. (1956), 'The Output of a Queuing System' Operations Research 4 (6): p 699 - 704

[13] Kolmogorov-Smirnov test
http://en.wikipedia.org/wiki/KolmogorovSmirnov_test

[14] Mor Harchol-Balter, Takayuki Osogami, AlanScheller-Wolf, Adam Wierman
'Multi-server queueing systems with multiple priority classes' Journal Queueing Systems: Theory and Applications archive Volume 51 Issue 3-4, December 2005 Pages 331 - 360

[15] Mahfoud, M., 'Forecasting Models For Ambulance Call Centers', Centrum Wiskunde en Informatica

[16] Bruce G. Lewis, and Ric D. Herbert
'Simulating the Call Streams to an Emergency Services Call Centre.'

[17] G.J. Kommer, S.L.N. Zwakhals 'Modellen referentiekader ambulancezorg pagina 109 Rijksinstituut voor Volksgezondheid en Milieu (RIVM), RIVM rapport 270412002/2013

[18] Regionale Ambulance Voorziening Utrecht (RAVU) - http://www.ravu.nl

[19] Noah Gans, Ger Koole Avishai Mandelbaum, 'Telephone Call Centers- Tutorial, Review and Research Prospects', Manufacturing & Service Operations Management 5:79-141, 2003

[20] Buzen, J., Bondi, A., 'The response times of priority classes under preemptive resume in M/M/m queues.', Oper. Res. 31, 456465 (1983)

[21] J. Puts, 'Emergency Call Center Finding a balance between costs and quality of service when dealing with emergency calls' P.17 -20,

[22] Vijay Mehrotra, Jason Fama, 'Call Centre Simulation Modeling: Methods, Challenges and Opportunities', Proceedings of the 2003 Winter Simulation Conference, pp. 135-143, 2003.

[23] Alysson Barros, Paulo Maciel, Elton Torres, Erica Sousa - 'Evaluation of Performance and Capacity Planning in Call Centers of Emergency Services' Federal University of Pernambuco Center of Informatics

[24] 'Opstelten kiest plekken nationale politie' - Gemeenschappelijkemeldkamer.nl - 15-12-2012 http://www.gemeenschappelijkemeldkamer.nl/forum.asp?id=435

[25] Vladimir A. Bolotin, 'Telephone Circuit Holding Time Distributions' - 'The Fundamental Role of Teletraffic in the Evolution of Telecommunication Networks Volume 1a', pages 125-134 'Teletraffic Science and Engineering', Elsevier

[26] 'Ambulancezorg Kort en bondig', Nationaal Kompas http://www.nationaalkompas.nl/zorg/ambulancezorg/ambulancezorg-samengevat/

[27] G.J. Kommer, S.L.N. Zwakhals 'Referentiekader spreiding en beschikbaarheid ambulancezorg 2013 pagina 33 - Rijksinstituut voor Volksgezondheid en Milieu - RIVM briefrapport 270412003

[28] Calinescu, M. (2009) 'Forecasting and Capacity Planning for Ambulance Services'.

[29] 'Acute zorg - RAV-vorming'
http://www.zorgatlas.nl/zorg/acute-zorg/rav-vorming/

[30] 'Een meldkamerorganisatie voor alle hulpdiensten' - 20-02-2012,
http://www.rijksoverheid.nl/nieuws/2012/02/20/een-meldkamerorganisatie-voor-alle-hulpdiensten.html

[31] Mobile phone tracking - Wikipedia,
http://nl.wikipedia.org/wiki/Mobile_phone_tracking

Appendices

A Staffing levels for a call center without regional knowledge

The tables in this chapter show the staffing levels and the corresponding performance indicators for the scenarios mentioned in Section 8.2.

A.1 Current situation (21 call centers)

Region	**Staffing level**	**Service level**			**Utilization rate (%)**
	Generalists	% A1 A2 within 6 sec	% B within 150 sec	% Follow-up within 30 sec	Generalists
Noord-Nederland	4	98.2	99.7	96.2	41.7
Oost-Nederland	3	97.9	99.4	94.6	34.3
Twente	2	97.8	98.6	95.2	23.6
Gelderland-Midden	2	98	98.7	94.7	24
Gelderland-Zuid	2	98	98.8	95	23.1
Utrecht	3	97.7	99.1	95	38.4
Noord-Holland-Noord	2	97.3	98.3	92.1	26.4
Kennemerland	2	97.3	98.2	94.2	26.3
Amsterdam-Amstelland	3	95.8	98.1	91.2	49.7
Gooi en Vechtstreek	2	99.7	99.8	97.9	11
Haaglanden	3	97.7	99.2	95.2	37.2
Hollands Midden	2	96.1	97.5	90.2	34.1
Rotterdam-Rijnmond	3	96.7	98.6	92.9	44.4
Zuid-Holland-Zuid	2	98.4	99	96.4	21
Zeeland	2	98.6	99.3	96.7	16.8
Midden- en West-Brabant	3	98.4	99.6	96.6	30.9
Brabant-Noord	2	97.5	98.7	93.4	25.3
Brabant-Zuidoost	2	97.1	97.9	92	30.1
Limburg-Noord	2	97.9	98.9	95.4	21.6
Zuid-Limburg	2	96.8	98.1	93.7	30.7
Flevoland	2	99.1	99.7	97.6	15.3
Total	50				

Table 21: Number of generalists for the current 21 regions

Region	**Staffing level**		**Service level**			**Utilization rate (%)**	
	Call takers	Dispatchers	% A1 A2 within 6 sec	% B within 150 sec	% Follow-up within 30 sec	Call takers	Dispatchers
Noord-Nederland	3	3	98.6	99.8	98.2	31	24.9
Oost-Nederland	2	2	97	98.7	95.5	28.6	22.8
Twente	2	2	99.4	99.8	98.3	13.2	10.5
Gelderland-Midden	2	2	99.3	99.8	99	13.4	10.7
Gelderland-Zuid	2	2	99.3	99.8	99.1	13	10.4
Utrecht	2	2	97.1	98.5	95.1	32.1	25.4
Noord-Holland-Noord	2	2	99	99.6	98.8	14.6	11.8
Kennemerland	2	2	99.2	99.7	98.8	14.6	11.6
Amsterdam-Amstelland	3	2	98	99.2	91.9	27.4	32.9
Gooi en Vechtstreek	2	1	98.5	98.9	92.3	6.3	9.7
Haaglanden	2	2	97.2	98.5	95	30.9	24.7
Hollands Midden	2	2	98.8	99.6	98.5	19	15.1
Rotterdam-Rijnmond	2	2	96	97.6	93.8	36.9	29.4
Zuid-Holland-Zuid	2	2	99.5	99.9	99.1	11.4	9.2
Zeeland	2	2	99.7	100	99.4	9.3	7.5
Midden- en West-Brabant	2	2	97.6	99.1	96.4	25.7	20.7
Brabant-Noord	2	2	99.2	99.8	98.7	14	11.4
Brabant-Zuidoost	2	2	98.9	99.7	97.8	16.7	13.2
Limburg-Noord	2	2	99.3	99.9	98.5	11.7	9.6
Zuid-Limburg	2	2	99	99.6	98.4	17.2	13.7
Flevoland	2	2	99.7	99.9	99.6	8.6	6.8
Total	44	42					

Table 22: Number of call takers and dispatchers for the current 21 regions

A.2 Plan of government (10 call centers)

Region	**Staffing level**		**Service level**			**Utilization rate (%)**	
	Call takers	Dispatchers	% A1 A2 within 6 sec	% B within 150 sec	% Follow-up within 30 sec	Call takers	Dispatchers
Noord-Nederland	3	3	98.7	99.8	98.4	30.8	24.8
Oost-Nederland	3	3	96.5	99.1	95.8	45.3	36.3
Noord-Holland	2	2	97	98.6	94.6	29.2	23.7
Midden-Nederland	3	2	97.5	99	90.3	31.1	37.2
Amsterdam	3	2	98.1	99.4	92	27.7	32.9
Den Haag	3	3	98.4	99.7	98	33.4	26.8
Rotterdam	3	3	98.6	99.8	98.4	32.4	25.9
Zeeland West-Brabant	2	2	95.7	97.8	92.6	34.8	28
Oost-Brabant	2	2	96.7	98.5	95.3	30.4	24.5
Limburg	2	2	97.1	98.7	95.1	28.9	23.3
Total	26	24	32.4	28.3			

Table 23: Number of call takers and dispatchers for 10 call centers

Region	**Staffing level**	**Service level**			**Utilization rate (%)**
	Generalists	% A1 A2 within 6 sec	% B within 150 sec	% Follow-up within 30 sec	Generalists
Noord-Nederland	4	98.1	99.8	96.3	41.6
Oost-Nederland	5	97.6	99.8	96.6	48.9
Noord-Holland	3	97.9	99.5	95.4	35.1
Midden-Nederland	4	98.1	99.7	96.6	42
Amsterdam	3	95.8	97.9	91	50
Den Haag	4	97.6	99.7	95.7	44.8
Rotterdam	4	97.9	99.6	96.3	43.7
Zeeland West-Brabant	3	96.7	98.8	92.9	41.8
Oost-Brabant	3	97.7	99.4	94.7	36.8
Limburg	3	97.9	99.4	95.6	34.9
Total	36				

Table 24: Number of generalists for 10 call centers

A.3 Two call centers

Region	**Staffing level**		**Service level**			**Utilization rate (%)**	
	Call takers	Dispatchers	% A1 A2 within 6 sec	% B within 150 sec	% Follow-up within 30 sec	Call takers	Dispatchers
Noord	7	6	96.3	99.9	97.3	62.7	58.5
Zuid	7	6	97.1	99.9	97.9	58.8	55
Total	14	12	60.7	56.8			

Table 25: Number of call takers and dispatchers for 2 call centers

Region	**Staffing level**	**Service level**			**Utilization rate (%)**
	Generalists	% A1 A2 within 6 sec	% B within 150 sec	% Follow-up within 30 sec	Generalists
Noord	11	96.2	100	98.4	71.9
Zuid	10	95	99.9	97.5	74.5
Total	21				

Table 26: Number of generalists for 2 call centers

A.4 A nationwide call center

Region	**Staffing level**		**Response time service level**			**Utilization (%)**	
	Call takers	Dispatchers	% A1 A2 within 6 sec	% B within 150 sec	% Follow-up within 30 sec	Call takers	Dispatchers
Nationwide call center	11	12	95.6	100	99.9	77.7	56.7
Total	11	12					

Table 27: Number of call takers and dispatchers for a nationwide call center

Region	**Staffing level**	**Service level**			**Utilization rate (%)**
	Generalists	% A1 A2 within 6 sec	% B within 150 sec	% Follow-up within 30 sec	Generalists
Nation-wide call center	19	96.7	100	99.7	80.83729
Total	19				

Table 28: Number of generalists for a nationwide call center