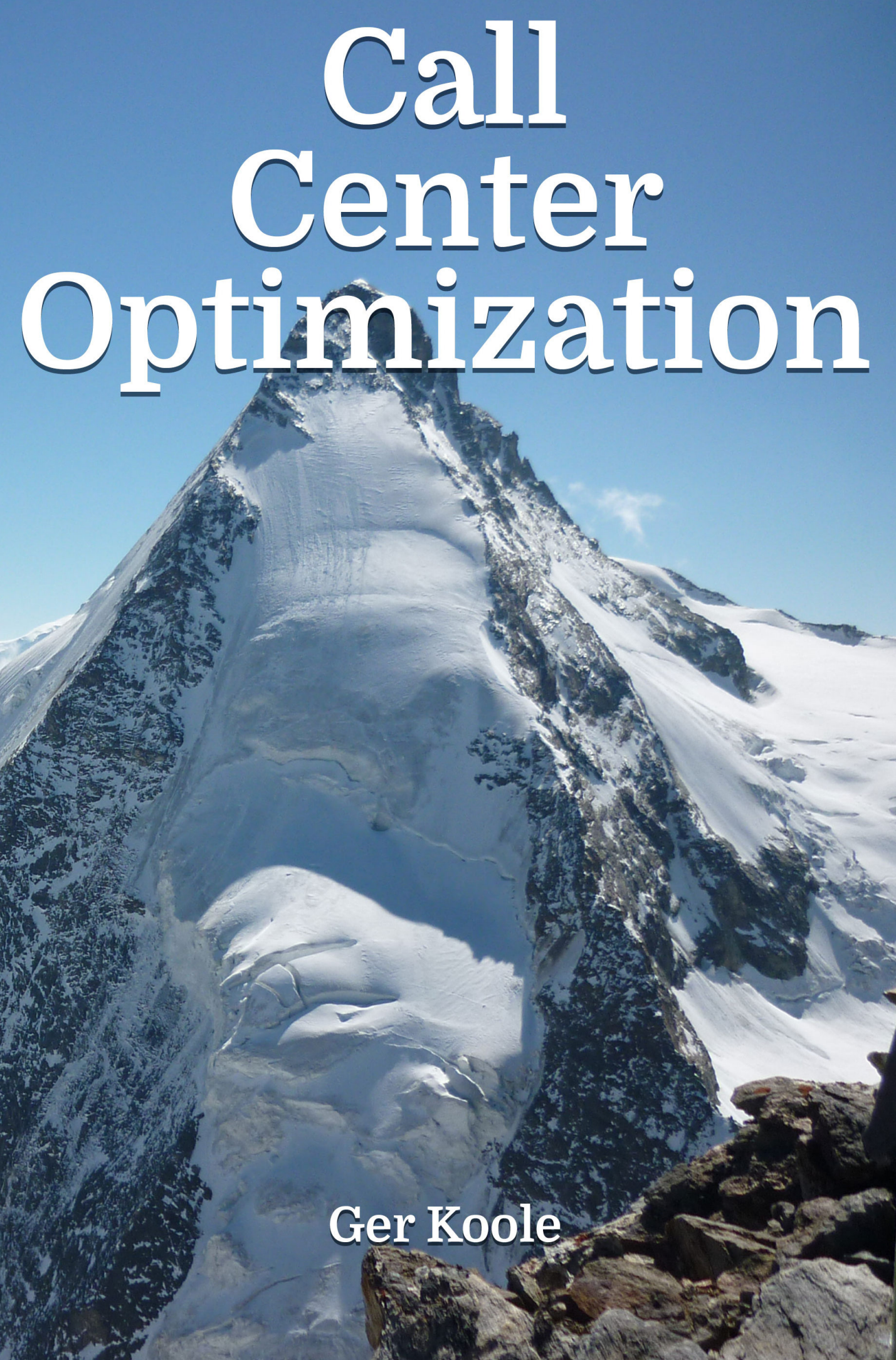
Call Center Optimization



Ger Koole

Call Center Optimization

MG books, Amsterdam
ISBN 978 90 820179 0 8
Cover design: Thom van Hooijdonk
Cover photo: Dent d'Hérens and Matterhorn seen from the Tête de Valpelline

Call Center Optimization

Ger Koole

MG books
Amsterdam

Preface

This book is written for everybody who is dedicated to improving call center performance. It offers a rational, scientific method to the understanding and optimization of call centers. It explains all generic aspects of call and contact centers, from the basic Erlang formula to advanced topics such as skill-based routing and multi-channel environments. It does this without going into technical details, but by showing the outcomes of many calculations. Moreover, there is a companion web site where these calculations can be executed for different input values. Next to understanding call center phenomena we show how to use these insights to improve call center performance in a systematic way. Keywords are data collection, scenario analysis, and simulation.

This book is a bridge between call center management and those parts of mathematics that are useful for call centers. It shows the manager and consultant the benefits of an analytical approach, without having to go into the technical details of it. It also shows the mathematically educated reader an interesting application area of queueing theory and other fields of mathematics. As such, this book can well be used as additional material in an applied course for mathematics and industrial engineering students. Basic knowledge of call centers is assumed, although a glossary is added in case of omissions.

There are many people who helped me writing this book. I would like to thank in particular Karin van Eeden, Theo Peek, Roger Rutherford, and Arnout Wattel for their suggestions, and Marco Bijvank, Joeri van Hoeve, Auke Pot and Alex Roubos for their help with the online tools. I would also like to thank all organizations that allowed me to use their data.

I started this project in 2001. It was hard to find time next to my regular obligations, so I found myself writing during holidays and scientific visits at various locations. At the same time, my understanding of call centers progressed and I changed several times the way the text is set up. Christ-

mas 2012 was the ideal moment to finish. Altogether, it was an extremely interesting experience to write this book. I hope that you will find it equally rewarding to read it.

Ger Koole
Amsterdam/Amstelveen/Sophia Antipolis/le Croisic/Courdemanges
2001–2013

Contents

Chapter 1

Introduction

This chapter introduces the ideas and concepts that we will use throughout this book. It does so by concentrating on one specific part of call center optimization, workforce management (WFM), and more specifically on the determination of the number of agents that need to be scheduled to satisfy a certain service level. Next to introducing the ideas and concepts, it serves as an introduction to the following chapters in which the different parts of WFM are discussed in full detail. In Chapter 9 call center optimization outside of WFM is discussed.

There is no need to worry if this chapter raises more questions than it anwers. All questions will (hopefully) be answered in the subsequent chapters.

1.1 Workforce management

The activity where analytical techniques are used most often is WFM. WFM is the common name of the planning cycle that results in the schedules of the call center agents, usually a few weeks before the period (often a week) for which the schedule is made. As input it uses historic call center data on traffic loads and information on agent availability; the output consists of agent schedules.

WFM can be split into several more or less separate steps. The first is forecasting the traffic load. This needs to be done at the level of time intervals of usually 15 or 30 minutes. A forecast is easily made, but producing accurate forecasts is a complicated task that requires considerable skills and knowledge. Good forecasting is crucial for good WFM, following the GIGO

principle: if you start the WFM with bad input ("garbage in"), then irrevocably the resulting schedule will be of bad quality ("garbage out"). What makes good forecasting difficult are the many factors that influence arriving call volume, and the fact that due to the random nature of the call arrival process one can never be completely sure what the causes were of observed fluctuations.

The second step in WFM is determining the required staffing levels for each interval. Sometimes this is considered to be part of forecasting, but it is at this point in the WFM process that demand and supply are matched: here it is determined how many agents (and possibly which types of agents) are needed at every interval to obtain the required service level. In simple single-skill call centers with only inbound calls the so-called *Erlang C formula* is often used for this. For more complicated operations with multiple skills and other communication channels such as email more advanced techniques such as *simulation* have to be used to evaluate staffing levels accurately. Often however rough approximations are used that give unreliable results. In the next sections we will introduce in more detail simulation and the Erlang formula.

Next we have to turn the staffing levels into agent schedules or rosters. This is the scheduling step. It can be done in two different ways. In the first agents specify their preferences beforehand and an advanced algorithm does the assignment taking all constraints (as much as possible) into account. The second consists of agents choosing the shift that they prefer on the basis of some auction system. The latter method is called *shift bidding*, and works best when the number of different types of shifts is limited.

Integrating staffing and scheduling

It often occurs, due to the relative inflexibilty of shifts, that the requirement to satisfy the service levels in each interval leads to considerable overstaffing at certain moments. This happens for example if there is a short spike in call volume or if all shifts overlap at some moment. The overall service level is then considerably higher than necessary. This can be avoided by integrating the staffing and scheduling steps.

Unfortunately WFM does not end with making shifts. Things not always go as planned, and adaptations have to be made. Both changes in call volume and scheduled agents can occur. An important operational task is monitoring *schedule adherence* and reacting accordingly. When call volume is different than expected, then also changes have to be made. This activity is

often called *traffic managament*, although in many cases it is the traffic that is monitored but the workforce that is managed. The whole is called *real-time performance management*.

Management, planning and scheduling
Management is a very broad term including all aspects of accomplishing organizational goals. WFM, as just described, is a much more narrow activity. To cite an expert: "Workforce management is the codeword for forecasting and scheduling software in the contact center industry" [12]. In fact, workforce *scheduling* would be a much better term instead of WFM: scheduling is about assigning tasks to resources, in this case agents. When considering also the other subjects treated in this book then the term *planning* fits best. For example, setting up a rational data-based long-term policy concerning the hiring and training of new agents is clearly planning, but not scheduling.

Agent scheduling is a crucial activity in any call center, without agent schedules the call center cannot operate. Without adequate software it is a laborious activity, especially in bigger call centers. This is the main reason why agent scheduling is automized in many centers, and why many software tools exist for this task. The core consists of algorithms supporting the different steps of WFM, but to make it work efficiently many more modules are necessary: a database filled with historical call volumes, data with agent information, connections with many other systems such as the ACD to get traffic information, and so forth. Next to that, a number of the larger WFM tools are part of software suites that offer other functionality such as email handling.

The functionality of these WFM tools varies enormously, and so does the quality of the proposed solutions. In practice we see that many tools are only partly used, and that users have their own, often Excel-based solutions, for, for example, forecasting and staffing. When selecting a new WFM tool organizations usually mostly look only at the functionality, little or not at the way certain methods are implemented. Every WFM tools has the possibility to forecast call volume, but the quality of the forecasts depends on the tool. Another example is that most WFM tools support multi-skill call centers, but the way in which it is implemented in the scheduling module varies also from tool to tool, and with that the quality of the resulting schedules. One of the objectives of this book is to develop a more critical look at WFM and thereby help the reader make better use of WFM tools.

1.2 The Erlang C formula

In this section and the following ones we consider the staffing of a single-skill inbound call center. We are interested in computing the optimal staffing level, let's say defined as the minimal number of agents required to answer 80% of the calls within 20 seconds. To do so we have to be able to predict the service level (SL) for a fixed number of agents, and then by varying the number of agents we can find the right staffing level. Our prediction of the service level will evidently depend on a number of variables: the forecast (FC) of the call volume, the number of agents, and also the average call handling time (AHT). Probably we are interested in staffing levels for a whole day or longer, but because forecasts and therefore also staffing levels vary over the day we concentrate on an interval of 15 or 30 minutes.

The FC is usually given per interval. To have everything in the same unit we divide by the length of the interval to get the FC per minute. Let us use the greek letter λ for this number (following a perhaps seemingly strange mathematical habit). Similarly, we denote by β the AHT, also in minutes. Then $\lambda \times \beta$ is called the offered load. This is equal to the number of agents needed to be able to handle all incoming calls. However, calls arrive at random moments, and therefore somewhat clustered, and handling times vary. Thus, if you have no or hardly any overstaffing with respect to the offered load, then the service level will suffer from these short random periods of high load. It is the Erlang C formula that gives the relation between FC, AHT, number of agents and service level, taking the randomness into account. The bad news is that this is not a simple relation that anybody can learn. The good news is that the Erlang formula has already been implemented in different spreadsheet add-ins, WFM solutions and other tools.

Example *Consider a call center with a FC of 100 for the 10:00-10:15 interval. The AHT is 3:30. Then* $\lambda = 100/15 = 6.66$ *and* $\beta = 3.5$*. The load is thus* $6.66 \times 3.5 = 23.33$*, and the minimum number of agents required to handle all calls is 24. The Erlang C formula predicts that in that situation only around 21% of all calls wait less than 20 seconds before getting connected. By increasing one by one the number of agents we find that 28 agents are needed to have a SL of at least 80%.*

In Chapter 4 we will study the Erlang C formula in more detail. However, there are also certain disadvantages to using the Erlang formula. The reason is that for the Erlang C calculation reality has been simplified. Certain statistical assumptions are made and certain features of call centers are left out. Without these assumptions and simplications it is not possible to

Do it yourself
There are several Erlang C calculators that can be found on the web. There is also one that is especially designed to accompany this book at www.gerkoole.com/CCO. You can go there and try to reproduce the numbers of the example. We will make extensive use of this and other tools in this book.

compute the formula, but they can lead to considerable discreprencies between prediction and reality. For example, an important feauture that is not part of the Erlang C model is that some calls, while waiting for service in queue, abandon. In situations where little calls abandon this is not necessarily a problem, but especially in underload situations this might result in big prediction errors. Under certain statistical assumptions the model including abandonments, called the Erlang X model, can be solved (see Chapter 4).

Next we consider a different feature of the Erlang formula that can lead to considerable errors. To understand this, we should realize that the performance of call centers is not completely predictable. For example, consider the handling times. We know the average handling time (AHT) and probably some other statistical properties. However, we do not know the exact duration of the call that is to arrive next. That means that any two intervals, even if all parameters such as the FC and number of agents are equal, will have a different performance. This has important consequences for call center management: we always have to deal with unpredictable fluctuations. It also means that we have to take into account the unpredictability when making service level predictions. However, the Erlang C formula gives as output a single number, and no indication of the size of the error. This is because the Erlang C fomula gives the performance as if the call center would run with the same parameters for a very long time. In reality this is not the case, usually we consider 15 or 30 minute intervals at a time. For this reason we should expect variability in the SL. A method that allows to quantify the SL variability, and also many other features, is *simulation*.

1.3 Simulation

The central idea of computer simulation is that we mimic reality in the computer. That is, we generate, on the basis of the forecast, arrival moments. These calls are assigned to virtual agents, or queued if no agents are available. Agents finish serving when the handling times are over, and they start with a new call or become idle, depending on the situation. Possible other

features include calls abandoning. Time progesses as events happen until the simulation time is over. All the while, statistics are assembled, for example on the number of calls that are answered within 20 seconds. Finally the required performance measures are calculated.

Because of the unpredictability, any two runs of the simulation are different, even when all input values are the same. In fact, if we take for example SL, any outcome is always possible. In a highly understaffed call center it might be that all calls need incidentally very little time leading to a high SL; conversely, in a well-dimensioned call center, it might occur that the first calls have very long handling times leading to congestion and a low SL. However, these situations are less likely to occur. Thus, when repeating the simulation often enough, we will see that the outcomes are concentrated around a certain level, apart from a number of outliers. This level is close to the level that is predicted by the Erlang model. In Figure 1.1 we plotted a histogram of 100 runs for the system of Example 1.2 with 28 agents. Note that the Erlang C formula predicts a SL of 83%, which falls within the 80-85% interval with the highest number of occurences, 34. (In Chapter 4 we will sharpen our understanding and see that the long run times, 8 hours, is crucial for this result.)

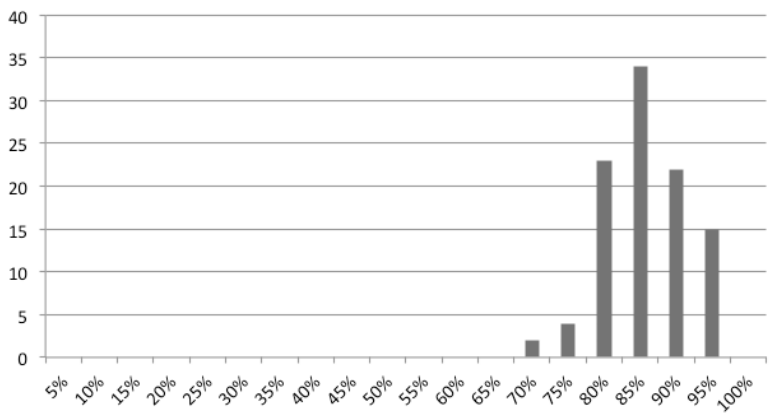


Figure 1.1: SL histogram for 100 runs of 8 hours

The calculations of Figure 1.1 were based on the Erlang C system. Using simulation it is easy to extend the model with features such as abandonments, different statistical assumptions underlying the handling times, different agents working at different speeds, redials, and so forth. When

extending to systems with different skills with skill-based routing (SBR), simulation really becomes essential. No equivalent for the Erlang formula exists in this situation, and except for some unreliable estimations simulation is the only method that can be used, both for staffing decisions (in which we vary the number and skills of the agents) and determining the best call routing parameters. Please note that the last activity is not part of the WFM cycle but can clearly be approached by similar techniques. SBR is discussed in Chapter 6.

Simulation can be used instead of the Erlang formula to obtain staffing levels, which can be used to determine the agent schedules in the scheduling step. Certain tools use an integrated procedure where new solutions are evaluated using simulation, on the basis of which again new solutions are selected, etc. The question remains how to deal with the fact that every simulation run will give a different result. In fact, the question to be answered first is how to deal with the fact that performance in reality will vary from day to day, even if the parameters remain the same. Does this mean that performance prediction is useless, because the outcomes will be different anyway?

Of course, performance prediction is useful, because the outcomes are often close to the prediction (in a statistical sense, which will be precised). Thus, to avoid having to change too much during the day, it is reasonable to base the schedule on some form of average performance. This is the type of average which is calculated by the Erlang C formula and which can also be obtained by simulation, by averaging over many runs. Unfortunately, the precision of simulation outcomes increases slowly with the number of runs, revealing the main problem of simulation: to obtain a high precision for the prediction run times can become very long. This is a big difference with formulas such as the Erlang C, which produce answers within split seconds. Especially in situations where interactively the best solution is sought simulation can lead to very long execution times, to low quality solutions, or to both. Highly skilled mathematicians and software engineers are needed to design and build WFM systems that use this method. But, even if a highly reliable estimate is obtained, we should never forget that real-time performance management will always be necessary, to deal with random fluctuations and other more or less unpredictable events such as agent absence and forecasting errors.

Integrating scheduling and real-time performance management
In theory, the best thing to do would be to integrate the schedule and the real-time performance management steps as well. Then exactly the right amount of flexibility is scheduled, and over or understaffing with respect to the average staffing levels will be chosen optimally. At several research centers around the world scientist are developing these methods.

1.4 Call center optimization

Optimizing a call center is more than doing WFM the best possible way. There are many decisions that can benefit from a rational data-driven optimization approach. Some of these concern repetitive operational problems such as WFM, others are of a more ad hoc nature. The former are therefore executed by dedicated people that count these activities as (one of) their main tasks, the latter require specialists who can apply their analytical skills to many different problems.

A first class of problems are those the enable good WFM. A good long-term hiring policy, skill-based routing, and the right mix of employee contract types is essential for good WFM. In outsourcing contracts the payments depend on the actual traffic, but often also on the difference between the forecast and the actual traffic. Thus knowledge of forecasting, and of forecasting errors, is crucial to call centers who outsource (part of) their traffic. All these issues will be discussed in the relevant chapters related to WFM.

WFM is sometimes criticized that it focuses only on efficiency, not on quality. The promise of *workforce optimization* (WFO) is to remedy this problem. WFO refers to software suites that include, next to WFM, modules for quality monitoring and call recording, (agent) performance management, and eLearning. We discuss the analytics of these activities in Chapter 9.

There are many activities that can be thought of that are not yet part of call center software, but that would give a company a competitive advantage when addressing. Optimizing a call center means seizing also these opportunities. All activities together is what we call call center optimization.

1.5 Further reading

Reynolds [25] is an excellent introduction to WFM. Cleveland & Mayben [10] is less of an overview, but easier to read, and it contains many interest-

ing insights.

For a list with the major WFM tools see Rosenberg [27].

The book [8] contains the historical background of Erlang's work.

Gans et al. [16] and Akşin et al. [2] are overviews giving the state of the art concerning mathematical models relevant to call center management. Both are written for academics, and assume solid mathematical knowledge. Stolletz [34] is also more mathematical.

Chapter 2

Performance measures and customer behavior

This chapter starts by defining the goals of customer contact and the way it is linked to customer behavior. Then we consider call center data. We discuss how to derive useful performance indicators from the data, considering both quality of service and efficiency. The outcomes of this chapter will serve as objectives and input parameters of the chapters on WFM and as a basis for the treatment of call center optimization in general in Chapter 9.

2.1 Call center objectives

Products are characterized by quality and price. The quality of a product might have many different aspects. In a production environment where tangible products are made, many aspects are related to what we might call the product itself. When we consider a TV set, this could be the size of the screen, the quality of the sound, and so forth. But there are also aspects that are not directly related to the product itself, such as the delivery conditions and the shopping experience (for example, whether it is bought online or in a shop). With customer contact a similar situation exists: certain aspects are related to the contact itself, certain are related to the way it is delivered. Examples of the former are the quality of the answer and the politeness of the call center agent, an example of the latter is the time that a customer has to wait before being connected to an agent.

With customer contact it rarely occurs that callers pay directly for the

call. An example of an exception is a directory service. Instead, call centers are mainly used to support certain business functions such as sales or product support. Then the customer pays indirectly for the call center when buying the product for which sales or service calls are done. In the case of outsourcing, calls directly generate the revenue, but it is not the caller who pays. Whatever the situation, high call center costs will translate, directly or indirectly, into a high price for the product concerned. Controlling costs is therefore essential. Whatever the revenue model is for the call center, the bigger part of its costs are personnel costs. They usually account for around 70% of the total costs.

To be able to improve the quality of a product and/or reduce its costs, we have to able to measure the quality and the costs. To do so we use *performance indicators* (PIs). These PIs are used in two ways: internally in the call center, that is, as input to our analytical techniques, and to communicate with stake holders outside of the call center. We have three types of PIs: those that are related to the costs of customer contact, those that are related to the quality of the contact, and those that are related to the way customer contact is delivered. For the latter often *service level agreements* (SLAs) are used.

Service level is an ambiguous term, especially in call centers. In its general sense it refers the quality of a product, especially to the non-product related properties such as waiting time. An example of a service level agreement is that the abandonment rate should not exceed 5%, 80% of the calls should be answered within 20 seconds, and the FCR rate should be at least 90%. To meet or exceed these objectives a certain budget is made available in case it concerns an internal call center. In case of an outsourcer the revenue depends on the extend to which the SLA is met. Usually a penalty is applied when the service level is lower than agreed upon.

The aspects of a product by which its quality is defined should also be chosen such that they can be measured. Some aspects are relatively easy to measure, such as the waiting time of a call. Some are harder to measure, such as the friendliness of the agent or the *First Call Resolution* (FCR) rate. Sometimes they require laborious data analysis or the use of advanced analytical tools, as can be the case for the FCR rate. In other situations customer surveys are needed, for example by automatically asking part of the customers for their opinion about the friendliness of the agent.

Costs can also be measured in PIs. One could think that total costs are the only relevant factor, but this not account for factors that influence the

Adverse effects of steering on PIs

It is good that performance is measured using PIs, but they should not become a goal by themselves. In that case they can even deteriorate quality of service. For example, if agents are stimulated for having short handling times, then they might be tempted to interrupt conversations before they are ended. In certain call centers it happens that agents are financially rewarded for, essentially, cutting calls after a few seconds. Analyzing call data reveals this type of practice. Another example of adverse effects of purely steering on PIs can be found in outsourcing. If an outsourcer is rewarded on the basis of SL, then there is no reason to answer calls who have waited longer than the service level limit. Thus it is "optimal" not to serve calls that have waited some time and wait until they abandon.

costs such as growth. Next, it does not give insight in the functioning of the call center. Given the fact that personnel costs represent the bigger part of the total costs we should use PIs that are related to the efficient use of the workforce.

When business problems are solved using software systems, as it is the case with WFM, it is very important to define the optimization goals clearly. In the next sections, we focus on the PIs that play an important role in WFM.

2.2 Customer behavior

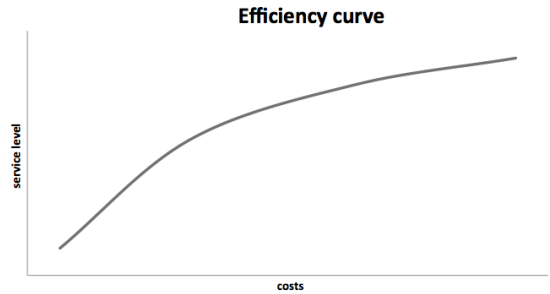
To be able to choose an appriopate SL definition we have to understand customer behavior, as this reflects customer preferences. For example, if customers abandon very quickly, then we conclude that our callers are very impatient. This might lead us to choose a small "time-to-answer". In this section we consider that part of customer behavior that is relevant to WFM: handling times, patience, *redials* and *reconnects*. We make the following difference between redials and reconnects: when a customer calls again after having abandoned then it is a redial; when a customer calls again after having being served then it is a reconnect. To analyze these aspects of customer behavior in all details we need to have access to data at the individual call level. This is not always available in call centers: ACD reports are usually aggregated at the interval level, and it is not always possible to get call-level data.

Let us start by studying the handling times. The most important figure is the average, the AHT. However, it is also interesting to look at the *distribution* by making a histogram as in Figure 2.1. A statistical study of

The cost-quality trade-off

The manager of a call center tries to satisfy the service levels set by higher management, given the call center's budget, and other constraints such as the number of work places (often called seats), the ICT infrastructure, and the available workforce. Of course, the higher the budget, the higher the service level can be, due to better training and more available resources. The main resource is the call center agent or representative, although communication costs can also be high, certainly for toll-free services. This means that the (infra)structure and processes of a call center should be such that the effectiveness and efficiency of the workforce is maximized.

The cost-service level trade-off thus has a central place in quantitative call center management. In general, when costs increase, then the service level (SL) increases. Thus we can draw a graph in which we show the SL as a function of the costs. This is called the *efficiency curve*, see the figure below for the typicaL form. Note that the curve is flattening as the costs increase. We often see these *diminishing returns*. Where the efficiency curve lies depends on the SL definition, but also on the infrastructure and the processes: every call center has its own efficiency curve. Improving the call center infrastructure and processes will shift the efficiency curve up and/or to the left.



In certain situations the profit of each individual call can be measured in terms of money. In such a situation the average profit per handled call can be calculated, and instead of balancing cost and service level, we just maximize profit. We will pay attention to this business model in Chapter 4.

several call centers revealed that the handling time distribution is often well approximated by a so-called *log-normal* distribution. From a practical point of view, it is more relevant to note that the AHT varies in time (time of day, day of week) and with the agent. This has consequences for WFM: in a call center it should certainly be considered to use time-dependent AHTs, and also the longer AHTs of new agents should be accounted for. The differences between AHT are also interesting to study, mainly from a perspective

of WFO. We will do this in Chapter 9.

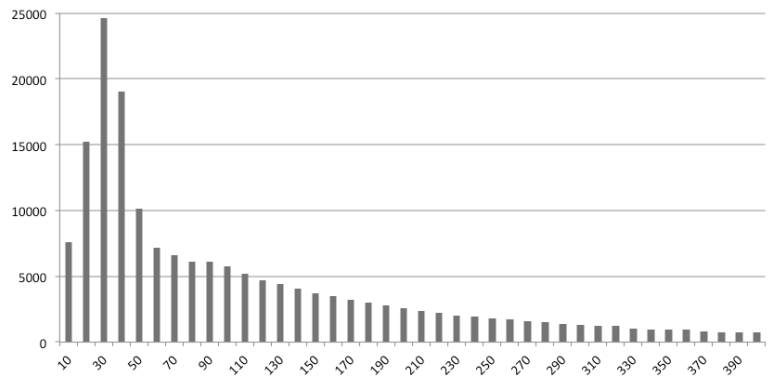
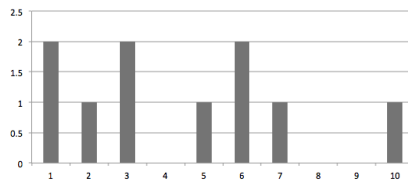


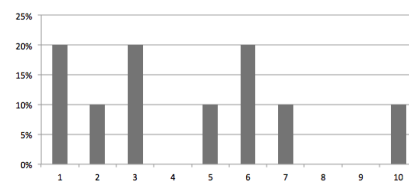
Figure 2.1: Histogram of handling time in a particular call center (in seconds)

Histograms and distributions

A histogram is made directly from a dataset: all outcomes are classified in "buckets" of equal length and the height of a bucket corresponds to the number of data points in the bucket. When the number of points is doubled, then the histogram becomes twice as high. When all numbers in the histogram are divided by the sum of the numbers then we get a distribution. The idea behind that is that these numbers can be used as approximations that the next data point will fall within the corresponding bucket. The validity of these type of questions are studied in statistics, as well as question whether the distribution ressembles some known form (such as the *normal distribution*). In general we can say that the more data points we have the better the approximation is, assuming that the circumstances have not been changed.

As an example, take the data set $\{0.5, 0.8, 1.2, 2.3, 2.5, 4.2, 5.1, 5.8, 6.1, 9.5\}$, and let's use buckets of length 1. Then the histogram and distribution are as in the figures below. The only difference is the verticale scale of the figure.





Next we consider abandonment behavior. Every caller will eventually abandon when not served, but the patience, the time that a customer is willing to wait in queue, differs between customers. Abandonments have both a

Percentages and probabilities

Percentages can used just as probabilities and as fractions, the difference being that a percentage is 100 times higher than the corresponding probability or fraction. For example, the statement "there is a 5% change that the actual is more than 10% higher than the forecast", is equivalent to the statement "the probability that the actual is more than 0.1 higher than the forecast is 0.05". Mathematicians prefer probabilities, because they can be multiplied: if the probability of a high actual is 0.3, and the probability of high absenteeism is 0.2, then the probability of both occuring is $0.3 \times 0.2 = 0.06$. Multiplication of probabilities is only allowed if the events are uncorrelated, that they are not likely to occur together because one is a consequence of the other, or because there is an underlying event that causes both. As an example of the latter, take two lines, each with a probability of 0.1 that the actual is high. Then the probability of high traffic on both lines is often higher that 0.01, because there are underlying causes such as weather conditions that may cause high traffic on both lines.

negative and a positive effect on the call center: negative because a call has not been handled, positive because it reduces congestion. In a call center most calls get served, thus we only know the patience of a small percentage of calls: the others get connected. However, the fact that they got connected gives information on their patience: it was longer than their waiting times. Not taking into account the effect of the connected calls can give a big error when estimating patience.

Example *Consider a call center with a small group (5%) of calls with short patience (less than 30 seconds). The other callers have a patience that is more than 2 minutes. If the waiting time is usually around 1 minute, then we measure about 5% abandonments with an average patience less than 30 seconds. However, the patience measured over all calls is around 2 minutes or higher.*

On the basis of the patience of abandoned calls and the waiting time of connected calls we can make a statistical estimate of the patience distribution using the so-called *Kaplan-Meier estimate*. The idea behind this method is explained in the box below, but let us first look at some outcomes in Figure 2.2. There are three graphs in this figure, all based on the same data set. The first is the patience distribution based solely on the abandoned calls. The second is the statistically correct distribution, after applying the Kaplan-Meier method to the numbers. As we can see the histogram has shifted to the right, meaning that the patience is longer then we would expect based solely on abandoned calls. The third graph needs some clarification. Next to knowing that "7% of callers have a patience between 3:30 and 4:00", it

might be of interest to know that "11% of callers who have waited for 3:30 are likely to abandon in the next 30 seconds". This so-called *conditional probability* (conditional on the fact that the caller has waited 3:30) is the third graph in the figure. It is surprising to see that after an initial high level the conditional probabilities stabilize. Apparantly there are two types of callers: those who abandon quickly and a larger group with a longer patience of which aboutt he same percentage abandon every time interval. This is of interest to our discussion of service level definitions later on.

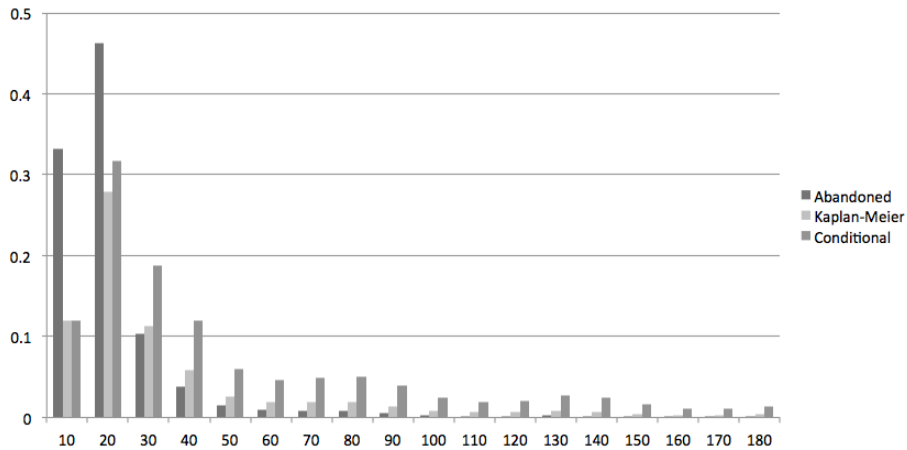


Figure 2.2: Abandonment distributions (in seconds)

When we analyze call-level data then we regularly see the same telephone numbers occuring (assuming we have access to this data, of course), also over shorter periods of time. There can be different reasons for that. The first is that callers who abandoned dial again a little later. This what we call a redial. A caller can also dial again after he or she got connected first. There can be, again, multiple reasons for this: either the initial reason of calling still exists, or the caller calls for a different issue. The former, which is called a reconnect, is not desirable, the latter usually is. Differentiating between the two is difficult, unless we make a detailed analysis of the contents of the call, or if you have a customer satisfaction survey (which always has missing data and is probably therefore *biased*, because the people participating in a survey are not representative for all customers). A practical solution is that we count all calls that got connected twice within the same day (or couple of hours) as reconnects. Analysis of survey data should validate this

The Kaplan-Meier method

The idea behind the Kaplan-Meier method is that connected calls are assumed to have a patience just like the abandoned calls who abandoned after the waiting time of the connected call. For example, if a connected call waited 25 seconds and there are 4 other calls that abandoned after 10, 20, 30 and 40 seconds, then the patience of the connected call could have been 30 or 40, with equal probability. Based on that we derive the overall patience distribution. We assume that a call is equally likely to behave as any one of the calls of which we have data. In the example patience is 10 or 20 with probability 0.2, and 30 or 40 with probability $0.2 + 0.2/2 = 0.3$. See also the table below.

Time	Abandoned/Connected	Kaplan-Meier distribution
10	A	0.2
20	A	0.2
25	C	0
30	A	0.3
40	A	0.3

When there are more connected calls then they should be treated one by one, starting from the ones with the shortest waiting time. Of course 5 calls is by no means enough data to get a reliable patience distribution, usually we need thousands of calls to compute the distribution.

approximation. Both redial and reconnect percentages are important PIs.

2.3 Quality of service

We saw that the goal of call center management is to obtain the right trade-off between costs and quality of service (QoS). We now go into more detail how the QoS can be measured. It consists of several different aspects. Some of these aspects are related to the handling of the calls themselves, such as the way in which the agents attend to the call, and the ratio of calls that need no need further calls, the first-time-fixed or first-call-resolution (FCR) ratio. Others are related to the waiting process, notably the waiting times and the occurrence of abandonments. We focus on waiting times and abandonments, although other aspects of the quality of service can have a large impact on the waiting time and therefore also on the abandonments.

Example *The help desk of an Internet Service Provider had a considerable rate of callers that phoned back after their call because the answer was not sufficiently clear to solve their problems. By improving scripts and documentation and by additional training this rate was reduced considerable. This not only improved the quality of*

service, it also reduced the number of calls. This had a positive effect on the waiting times, and thus again on the service level.

Weighted averages

Often we know the SL for short intervals (often giving by the ACD), and we want to compute the SL for longer intervals, for example in a spreadsheet to make a monthly report. The SL of a long period composed of several shorter of which we know the SL can be calculated be averaging in the right way service levels over the shorter periods. When averaging over a number of intervals the number of calls in these intervals should be taken into account. Consider the table below. At first sight the average service level is 75%, by averaging the four percentages, but now the differences in numbers of calls per week are not taken into account.

Week	Number of calls	Answered within 20 s.	SL
1	2000	1900	95%
2	7000	3850	55%
3	5000	3500	70%
4	3000	2400	80%

The right way of calculating is to compute the *fraction* of calls in each interval first. For example, the fraction of calls in the first interval is $\frac{2000}{17000}$, 17000 being the total number of calls over the four weeks. Using these fractions a *weighted average* is calculated in the following way:

$$\frac{2000}{17000} \times 95 + \frac{7000}{17000} \times 55 + \frac{5000}{17000} \times 70 + \frac{3000}{17000} \times 80\% = 68.5\%.$$

This way of calculating averages corresponds to the answer in case the service level was computed directly for the whole month. Indeed, out of a total of 17000 calls 11650 were answered in time, thus a $\frac{11650}{17000} \times 100 = 68.5\%$ service level.

The difference between 68.5 and 75% is not that dramatic. This is because the number of calls in the different weeks are roughly of the same order of magnitude. If the number of calls in the intervals over which we average are very different, then the way of averaging can have an even bigger impact on the result. These big fluctuations typically occur during days. At peak hours we can easily have ten or twenty times as many calls per hour as during the night. Then the difference between ways of averaging can run into the tens of percents.

Weighted averages can easily be computed in Excel. When, in the table above, "Week" is the contents of cell A1, then the weighted average can be computed by the Excel formula =SUMPRODUCT(B2:B5,D2:D5)/SUM(B2:B5). A simpler calculation using the numbers of calls is =SUM(C2:C5)/SUM(B2:B5).

The common way to define quality of service is by looking at the fraction of calls that exceeds a certain waiting time. We call this fraction the *Service Level* (SL). The waiting time that is considered acceptable is known under different names: *Time-to-Answer* (TTA), *Acceptable Waiting Time* (AWT), and *Service Time* (ST) are all used. The "industry standard" is that 80% of all calls should be answered in 20 seconds, but other numbers are possible as well. The SL can simply be calculated as long as there are no abandonments, by dividing the number of calls handled before the AWT by the total number of calls.

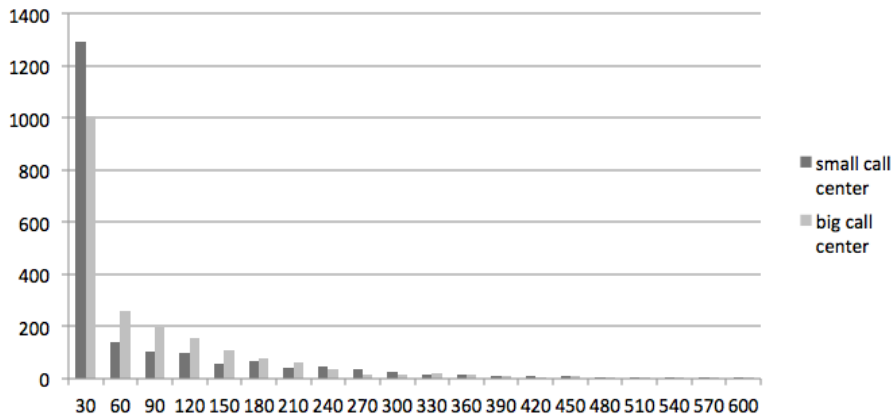


Figure 2.3: Histograms of waiting times (in seconds) for two call centers, with ASA = 60s

The SL is not the only way to measure QoS. Another commonly used waiting time metric is the *average speed of answer* (ASA). This is nothing else than the average over the waiting times. SL and ASA consist of a single number. This is both an advantage and a disadvantage: it is simple, but it gives only limited information. Full information is given by the distribution of the waiting times. Although it is certainly useful to determine this distribution now and then, it does not qualify as PI because of its complexity. The question then becomes: does the SL or the ASA capture the notion of QoS sufficiently? To answer this question we will have a closer look at both. A disadvantage of using the ASA is that the variability is not part of the PI. That is, the ASA does not differentiate between the following two cases: all calls wait exactly 30 seconds, or 90% get connected immediately and 10% waits 300 seconds. In Figure 2.3 we see that the distribution of

waiting times for call centers with the same ASA can be quite different. In fact, there is more variability in the waiiting times of the smaller call center.

Thus we look for a simple PI that is sensitive to variability, especially to calls that wait long. An obvious candidate is the SL: it measures the fraction of calls that wait longer than a certain limit. For the SL a similar figure can be made as for the ASA. In Figure 2.4 we plotted the SL for different AWT's for two call centers. An 80/20 SL implies 91/90 for the small call center and 97/90 for the bigger one.

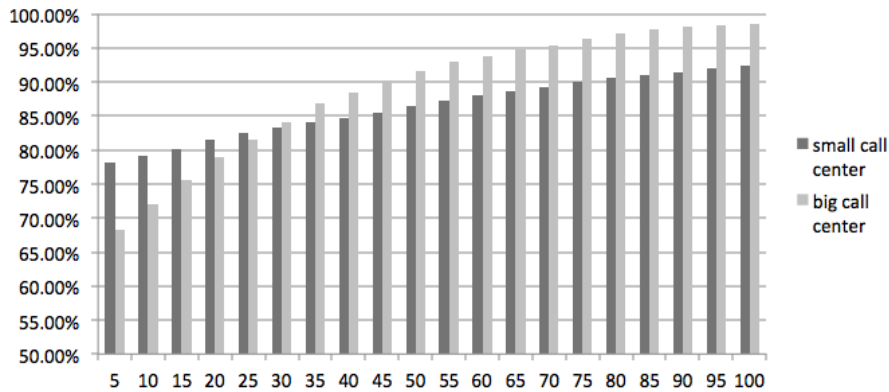


Figure 2.4: SL as function of the AWT (in s) for two call centers with SL = 80%

The disadvantage of the SL is that it not matter how much longer than the AHT a call has waited: it makes no difference between a waiting time of AWT + 1 second and AWT + 100 seconds. A possible solution, that combines ideas from th ASA and the SL, would be the *average excess* (AE): the average time calls have waited beyond the AWT. As an example, if the AWT is 20, and the waiting times are 10, 25, 30 and 25, then the AE is $(0 + 5 + 10 + 5)/4 = 5$ seconds. Despite its disadvantages the SL is the most used PI to represent the waiting time.

The idea of replacing the ASA by the SL is that we avoid long waiting times. However, under an 80/20 SL, 20% of the calls wait longer than the AWT of 20 seconds. To avoid this, we could require a 99/20 SL. However, this requires a much bigger workforce. Is this necessary? This depends on the fact if the caller will accept a long waiting time *now and then*. Consider an individual customer that belongs to the 20% that received bad service. To this customer it is right now irrelevant if the SL was 50/20 or 80/20, in the former case there are just more unsatisfied customers. To the unsatisfied

Service orders

In call centers it is customary to serve calls in order of arrival, that is, longest waiting calls first. However, what is the optimal order if we want to minimize the ASA or maximize the SL? For the ASA the order is irrelevant, for the SL it is even better to serve first calls that have not yet exceeded the AWT. In fact, one can argue that calls that have exceeded the AWT should not be served at all: one can better keep the agents available for new arriving calls. Thus steering on SL only pushes call centers to behave in a customer-unfriendly way. This is a reason to avoid penalty clauses in outsourcing contracts based on SL only. Service order is even more relevant in the context of abandonment, which will be treated later on in this chapter.

customer the SL becomes relevant when he or she tries to call again. If the SL at that moment is again 80/20, then the probability of another bad experience is 0.2, or 20%. 1 out of 25 customers, 4%, have 2 consecutive bad experiences. And how many customers will try a third time after two bad experiences if they have alternatives? If the competition is strong then offering only an 80/20 SL can lead to churn. Thus whether a 80/20 SL or any other choice is the right SL for a call center depends on the behavior of the callers. Will they call back after a bad experience, and is 20 seconds indeed the correct borderline between good and bad service? Things become even more complicated when we take abandonments into account. See the next section on this subject.

Choices related to SL become also more difficult when we consider call centers with multiple types of calls (see Chapter 6 for more on multi-skill call centers). Consider for example two types of calls for which we like to obtain both an 80/20 SL. Now what if we obtain 70/20 on one and 90/20 on the other? And if we have the choice, with the same means, between 70/20 and 90/20 or 75/20 and 80/20? The former has a better average SL (assuming an equal load), the latter has a higher minimum. The answer depends again on the behavior of callers and the nature of the service: will they mostly generate the same type of call, or do they change type? In the former case we should consider the types independently, in the latter case we should perhaps focus on the average SL.

The situations becomes even more complicated when we have different SL constraints for different call types, for example because we want the sales line to have a better SL than the after-sales line. Here we might have 90/20 and 70/20 constraints, and still be more satisfied when we realize 95/20 and 65/20, simply because we value individual sales calls higher than after-

sales calls. A SL definition that corresponds better with our intuitive notion of QoS might consist of a constraint on the high-value calls of 90/20 and an overall constraint over all calls of 80/20.

Example *A call center has two types of calls: calls with a negociated QoS in terms of a SL that has to be met in all situations, and "best effort" traffic where the revenue depends on the QoS. Under high traffic conditions the SL of the first type of calls cannot be met, even when priority is given to these calls. Therefore, the rational decision,* given the contract, *is to give priority to best effort calls in case of high load and to give priority to fixed SL calls when traffic is low to catch up with the SL. This is in complete contradiction with the intentions behind the QoS contract.*

2.4 Abandonments

We saw in Section 2.2 that abandonments are an essential part of customer behavior. In general, abandonments are considered a sign of customer dissatisfaction and should therefore be avoided, even though some calls abandon in less than the AWT. The abandonment rate or percentage is therefore a useful PI for almost any call centers. Usually there is a constraint on the abandonment rate, often in the order of 3 or 5%.

Many call centers combine a constraint on the abandonment rate with a SL constraint. However, we have to decide how abandonments are counted in the SL definition. For this purpose, we classify calls into 4 types. Using the symbol # for the count, this leads to:
- #(connected $\leq$ ATW), the number of calls connected before the AWT;
- #(connected $>$ ATW), the number of calls connected after the AWT;
- #(abandoned $\leq$ ATW), the number of calls abandoned before the AWT;
- #(abandoned $>$ ATW), the number of calls abandoned after the AWT.

The SL is a quotient. The numerator consists of the calls that got good service. In all definitions thus is taken equal to #(connected $\leq$ ATW). More interesting is the denominator. It is clear that all connected calls (#(connected)) should be part of it, but how about the abandoned calls? In practice we find 3 different choices, leading to 3 different SL definitions:

$$\begin{aligned} \mathrm{SL}_1 &= \frac{\#(\text{connected} \leq \text{ATW})}{\#(\text{connected})}; \\ \mathrm{SL}_2 &= \frac{\#(\text{connected} \leq \text{ATW})}{\#(\text{connected}) + \#(\text{abandoned} > \text{ATW})}; \end{aligned}$$

$$\mathrm{SL}_3 \quad = \quad \frac{\#(\text{connected} \leq \text{ATW})}{\#(\text{connected}) + \#(\text{abandoned})}.$$

The first definition, SL_1, is sometimes used in combination with the abandonment rate. The big disadvantage is that by not answering calls that have waited more than the AWT the SL can be improved: if a call gets connected in 30 seconds it counts in the denominator, if it abandons it does not count. Thus SL_1 clearly gives a perverse incentive and should not be used for this reason (see, in this context, also the box on page 22).

SL_2 and SL_3 do not have this disadvantage. Furthermore, it is clear that callers who abandon after the AWT have received bad service, and therefore these calls are added to the number of calls for which the service requirement was not met. For callers that abandon before the AWT it is not that clear. The most reasonable is perhaps not to count these calls at all. This leads to definition SL_2. Counting all abandonments as bad service leads to definition SL_3. Because the numerator increases it is clear that $\mathrm{SL}_1 \geq \mathrm{SL}_2 \geq \mathrm{SL}_3$. That they can be really different is shown in the following example.

Example *A call center receives 510 calls during an hour. The AWT is set equal to 20 seconds. A total of 460 receive service, of which 410 are answered before 20 seconds. Of the 50 abandoned calls 30 abandon before 20 seconds. The different definitions give:* $\mathrm{SL}_1 = 410/460 = 89\%$, $\mathrm{SL}_2 = 410/480 = 85\%$, *and* $\mathrm{SL}_3 = 410/510 = 80\%$, *a considerable difference.*

These ways of calculating the service level are all easily done on the basis of observed waiting times of calls: one needs to remember the numbers of served and abandoned calls and whether that happens before and after the AWT, in total four numbers per interval for which we want to know the SL.

Virtual waiting time

Another way of defining the service level is to compute it from the waiting time of 'test customers' who have infinite patience. In general this leads to numbers very close to the definition in which we ignore customers who abandon before the AWT. This definition is attractive because it is independent of the patience of a caller. On the other hand, it cannot be observed directly and has to be estimated from the observed statistics. One should have all waiting times (not just the counts) and apply the Kaplan-Meier method (which is explained in the box at page 18) to obtain not the patience distribution but the waiting time distribution. From this the SL can be computed.

We should also consider how to incorporate abandonments in the ASA,

in case the ASA is used as service level metric next to or instead of the SL. We can either average over the connected calls or over all calls, or compute the average virtual waiting time using the method explained in the box above.

2.5 Occupancy and shrinkage

In the beginning of this chapter we saw that the service product is characterized by its quality and its costs. In this section we focus on the costs. The majority of the costs in a call center are personnel costs.

Ideally agents should talk to customers 100% of the time they are paid. Unfortunately, this is not the situation, for a number of reasons. Roughly the working time of agents can be divided into two categories: the time that an agent is available to handle calls (or contacts through other channels such as email) and the time an agent cannot take calls. In the former category it can be that the agent is busy with the call (either talking or wrapping up) or that the agent is idle, waiting for a call. In the latter category we find absence because of unforseen situation (such as illness) and holidays, training and coaching, and paid breaks. The fraction of time that the latter category represents is called *shrinkage*. It is an important PI: the lower the shrinkage, the more time agents have for answering calls. On the other hand, a certain amount of shrinkage is unavoidable, because training and coaching are necessary for quality reasons. A shrinkage of 40% is not exceptional.

Next to having a low shrinkage we would like the agents to handle as much contacts as possible while being available for contacts. The indicator for this form of efficiency is the occupancy, measured over a certain period (for example, a week). It is given by:

$$\text{Occupancy} = \frac{\text{Sum of handling times}}{\text{Sum of handling times and total idle time}}.$$

The higher the occupancy, the higher the efficiency. It should be noted however that a occupancy of nearly 100% can only occur for short periods of time, longer periods are too stressful for agents. What a reasonable occupancy target is depends on many factors, including how we count short breaks (are they part of the shrinkage or not?) and the length of the shifts.

Example *An agent has a contract for 36 hours a week. On average she is absent for 4 hours, she spends 3 hours on training and activities outside the call center, she takes breaks during 230 minutes, she is available waiting for calls during 265 minutes, and she is handling calls (talking plus wrap-up) during 1245 minutes.*

Her occupancy is $1485/(1245+265) = 82\%$*, if we do not count brakes as part of shrinkage 72%, and if we count all the time she spends at work 58%.*

Which definition of PIs is best depends on the situation. If agents are free to take breaks whenever they like then it is probably better to include these in the denominator. In any case, all performance indicator should be considered together: a high occupancy is useless if the FCR percentage is low. In fact, decreasing the first-time-resolution percentage decreases the idle time through an increase in calls and thus "improves" the occupancy.

There are other obvious but interesting relations between the performance indicators. If one tries to decrease the number of performance indicators then one probably ends up considering the number of resolved calls. The disadvantage of this criterion however is that it is hard to measure.

2.6 Further reading

Seddon [32] explains clearly which undesirable outcomes strictly thinking in SLAs can have, with a focus on the health care sector. The example on page 23 comes from the scientific paper Milner & Olsen [22].

Chapter 3

Forecasting

Estimating future workloads is an essential but difficult part of WFM. In this chapter we discuss all aspects of forecasting. We will start with a somewhat technical section about the nature of call arrivals.

3.1 The nature of call arrival processes

To really understand forecasting in call centers we have to understand the nature of call arrival processes, and this goes back to the bahavior of the individual caller. Consider a time interval in a call center, a specific half hour, or perhaps a whole day. Let us say that for this interval the forecast is 100. This means that, out of the perhaps millions of (potential) customers, we predict that 100 will call. Who will call exactly we do not know, but if we have say one million customers, then apparantly each has a likelihood of calling of $100/1\text{M} = 1/10000$. Thus a forecast (FC) of 100 is equivalent to a probability of calling of $1/10000$ by 1M people. This is all we know about our callers. It is like flipping coins: we know the expected outcome but we never know how many times heads will come up if we try it once. Thus we can never be sure how many people will call, even if we know exactly the likelihood of calling. That is bad news, but the good news is that we do have some quantitative information about the number of people that will call. In the 19th century the French mathematicain Siméon Poisson found that numbers of arrivals follow a certain law, which we now call the *Poisson distribution*. A histogram with 1000 draws from the Poisson distribution with average 100 is shown in Figure 3.1. Actually, the probabaility that precisely 100 arrival will occur is 4%, the probability that the error is bigger

than 5% (the outcome is lower than 95 or higher than 105) is 58%. The technically interested reader can reproduce these number with the help of Excel (see the box on page 28).

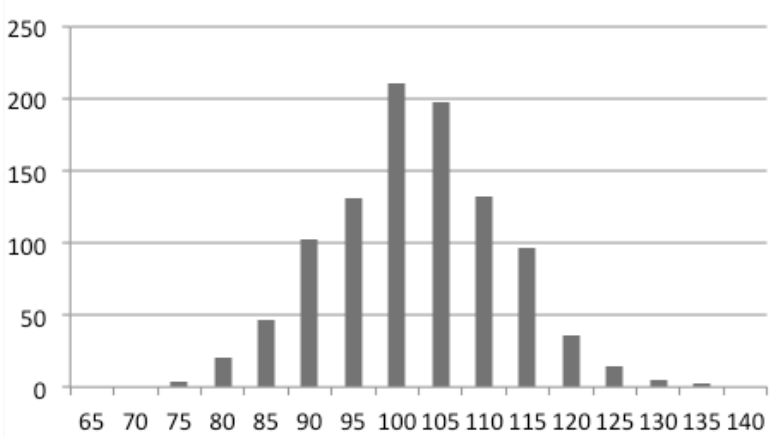


Figure 3.1: Histogram based on the Poisson distribution with average 100

Excel calculations with the Poisson distribution

The Excel Poisson function can be used to calculate Poisson probabilities. It works as follows: if you type $=$ POISSON$(30, 35,$ FALSE$)$ in a cell then you get the probability that there are 30 arrivals when the forecast is 35. Similarly, $=$ POISSON$(100, 100,$ FALSE$)$ gave the 4% mentioned in the text. The "FALSE" refers to the fact that we are only interested in the value 100. To get the probabilities up to a number, we should use "TRUE". Thus $=$ POISSON$(94, 100,$ TRUE$)$ calculates the probability that the outcome is more than 5% lower than the FC 100, around 29%. By calculating $=$ POISSON$(105, 100,$ TRUE$)$ we get the probability that the outcome is lower than or equal to 105, around 71%. Then the probability that the outcome is higher than 105, that is, more than 5% higher than the FC, is 29%. By adding both probabilities we get the probability that the error is more than 5%: 58%.

In case you use a non-English version of Excel you can look up the name of the Poisson function by searching the functions for "poisson" or by searching the internet for a table with translations.

The high variability of the Poisson distribution might come as a surprise: when the FC is 100, then only because of "natural" variability there is an 58% probability that the error is higher than 5%. In many call centers this is not acceptable. Luckily, the percentage error, that is, the error relative to the

Bibliography

[1] How to balance business goals with Avaya Business Advocate, 2011. http://www.avaya.com/uk/resource/assets/brochures/Business%20Advocate%20Gcc0467%20Final.pdf.

[2] O.Z. Akşin, M. Armony, and V. Mehrotra. The modern call-center: A multi-disciplinary perspective on operations management research. *Production and Operations Management*, 16:665–688, 2007.

[3] J. Anton. *Call Center Management by the Numbers*. Purdue University Press, 2007.

[4] J. Anton, V. Bapat, and B. Hall. *Call Center Performance Enhancement Using Simulation and Modeling*. Purdue University Press, 2000.

[5] I. Ayres. *Super Crunchers: Why Thinking-by-Numbers is the New Way to be Smart*. Bantam Books, 2007.

[6] M. Bodin and K. Dawson. *The Call Center Dictionary*. CMP Books, 2002.

[7] D. Brink. *Essentials of Statistics*. bookboon.com, 2010. Free download from http://bookboon.com.

[8] E. Brockmeyer, H.L. Halstrøm, and A. Jensen. The life and works of A.K. Erlang. *Transactions of the Danish Academy of Technical Sciences*, 2, 1948. http://oldwww.com.dtu.dk/teletraffic/Erlang.html.

[9] B. Cleveland. *ICMI's Call Center Management Dictionary*. Call Center Press, 2003.

[10] B. Cleveland and J. Mayben. *Call Center Management on Fast Forward*. Call Center Press, 1997.

[11] T.H. Davenport and J.G. Harris. *Competing on Analytics: The New Science of Winning*. Harvard Business School, 2007.

[12] K. Dawson. Workforce management: The Witness Systems interview. *Call Center Magazine*, May 2006. Interview with Bill Durr, `http://www.icmi.com/Resources/Articles/2006/May/Workforce-Management-The-Witness-Systems-Interview`.

[13] F.X. Diebold. *Elements of Forecasting*. Thomson, 4rd edition, 2007.

[14] A. Fukunaga, E. Hamilton, J. Fama, D. Andre, O. Matan, and I. Nourbakhsh. Staff scheduling for inbound call centers and customer contact centers. *AI Magazine*, 23(4):30–40, 2002.

[15] J. Galbraith. *Designing Complex Organizations*. Addison-Wesley, 1973.

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

[17] P. Harts. *The Relation between Quality and Average Handling Time*. Auditio, 2007. In Dutch. Online available at `http://publications.onlinetouch.nl/5/37/#/0`.

[18] F.S. Hillier and G.J. Lieberman. *Introduction to Operations Research*. McGraw-Hill, 8th edition, 2005.

[19] G.M. Koole and S.A. Pot. A note on profit maximization and monotonicity for inbound call centers. *Operations Research*, 59:1304–1308, 2011.

[20] DMG Consulting LLC. Contact center workforce management market report reprint, 2012. `http://www.nice.com/sites/default/files/nice_2012_wfm_report_reprint_final_june_2012.pdf`.

[21] S.G. Makridakis. *Forecasting, Planning, and Strategies for the 21st Century*. The Free Press, 1990.

[22] J. Milner and T. Olsen. Service-level agreements in call centers: Perils and prescriptions. *Management Science*, 54:238–252, 2008.

[23] C. Palm. Methods of judging the annoyance caused by congestion. *Tele*, 4:189–208, 1953.

[24] S.G. Powell, K.R. Baker, and B. Lawson. Impact of errors in operational spreadsheets. *Decision Support Systems*, 7:126–132, 2009.

[25] P. Reynolds. *Call Center Staffing*. The Call Center School Press, 2003.

[26] P. Reynolds. The power of one in call center staffing. `http://www.callcentrehelper.com/images/penny_webinar_power_of_one.pdf`, 2011.

[27] A. Rosenberg. Best practices in workforce management. *Call Center Magazine*, May 2005. `http://www.icmi.com/Resources/Articles/2005/May/Best-Practices-in-Workforce-Management`.

[28] S.M. Ross. *Introduction to Probability Models*. Academic Press, 7th edition, 1997.

[29] S.M. Ross. *A First Course in Probability*. Prentice Hall, 6th edition, 2002.

[30] D.A. Samuelson. Predictive dialing for outbound telephone call centers. *Interfaces*, 29(5):66–81, 1999.

[31] S. Savage. *The Flaw of Averages: Why We Underestimate Risk in the Face of Uncertainty*. Wiley, 2012.

[32] J. Seddon. *Systems Thinking in the Public Sector*. Triarchy Press, 2008.

[33] A. Smith. *An Inquire into the Nature and Causes of the Wealth of Nations*. Digireads.com, 2009 (first published in 1776).

[34] R. Stolletz. *Performance Analysis and Optimization of Inbound Call Centers*. Springer, 2003.

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

[36] J. Taylor and N. Raden. *Smart (Enough) Systems*. Prentice-Hall, 2007.

[37] H.C. Tijms. *A First Course in Stochastic Models*. Wiley, 2003.

[38] W.L. Winston and S.C. Albright. *Practical Management Science*. Cengage Learning, 4th edition, 2012.

All scientific call center publications by the author can be found at the call center publications page of www.gerkoole.com.

This book gives an accessible overview of the role and potential of mathematical optimization in call centers. It deals extensively with all aspects of workforce management, but also with topics such as call routing and the scheduling of multiple channels. It does so without going into the mathematics, but by focusing on understanding its consequences. This way the reader will get familiar with workload forecasting, the Erlang formulas, simulation, and so forth, and learn how to improve call center performance using it. The book is primarily meant for call center professionals involved in planning and business analytics, but also call center managers and researchers will find it useful. There is an accompanying website which contains several online calculators.

Ger Koole is a mathematician with a background in operations research and applied probability theory, especially queueing models. He applies these techniques to the service sector, especially in call centers and health care. He is full professor at VU University Amsterdam and co-founder of CCmath, a software firm specialized in call center workforce management.

ISBN 978-90-820179-0-8
90000
9 789082 017908