Performance Indicators for Call Centers with Impatience

Oualid Jouini[1], Ger Koole[2] & Alex Roubos[2]

[1]Ecole Centrale Paris, Laboratoire Génie Industriel,

Grande Voie des Vignes, 92290 Châtenay-Malabry, France

[2]VU University Amsterdam, Department of Mathematics,

De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands

oualid.jouini@ecp.fr, ger.koole@vu.nl, a.roubos@vu.nl

July 7, 2012

Abstract

An important feature of call center modeling is the presence of impatient customers. In this paper, we consider single-skill call centers including customer abandonments. We study a number of different service level definitions, including all those used in practice, and show how to explicitly compute their performance measures. Based on data from different call centers, new models are defined that extend the common Erlang A model. We show that the new models fit reality very well.

Keywords: call centers, Erlang A, abandonments, metrics, queueing delays.

1 Introduction

Context and motivation: The Erlang C model is still the most widely-used performance model in call center practice. An important property of this model is that all delayed customers wait until they get service. In reality some calls abandon, and therefore there is a discrepancy between the Erlang C predictions and the call center reports. In the scientific community the crucial role of customer abandonments has been recognized, and a number of models have been proposed.

The simplest model including abandonments is the so-called Erlang A model, which extends the Erlang C model and assigns to each customer an exponentially distributed patience (or abandonment) time. Brown et al. (2005) show for a number of cases that the patience is distinctly non-exponential, and Whitt (2005) shows that the patience is the variable that is most sensitive to higher moments. It can therefore be expected that Erlang A predictions have considerable errors, and this has been confirmed in practice. One of our contributions is that we propose an extension of the Erlang A model in which we allow for the possibility of balking. This simple extension makes the performance prediction by the queueing model much more accurate. The reason is that a relatively large proportion of the calls that get delayed abandon, and that the conditional patience distribution is approximately exponentially distributed from a certain point in time onward.

Another issue is the performance measure that is used. In the Erlang C model there is an unambiguous definition of the service level: the percentage of customers that get connected before a certain acceptable waiting time (AWT). In the case of abandonments it is not immediately clear how to account for these abandonments. Different service-level definitions are used in practice, often in parallel to the abandonment percentage. In scientific work the service-level definition is often based on the virtual waiting time, i.e., the waiting time that a customer with infinite patience would experience. Note that a performance measure is only of practical use if it can also be measured in practice. For definitions based on the virtual waiting time this is not the case and therefore they are of less practical value. An important contribution is that we derive explicit expressions for several performance measures, including all measures that we have encountered in call center practice.

Related literature: The importance of modeling abandonments in call centers is emphasized by Garnett et al. (2002), Gans et al. (2003), and Mandelbaum and Zeltyn (2009). Empirical evidence regarding abandonments in call centers can be found in Brown et al. (2005) and Feigin (2005). We refer the reader to Garnett et al. (2002), and references therein, for simple models assuming exponential patience. Garnett et al. (2002) suggest an asymptotic analysis of their Markovian abandonment model under the heavy-traffic regime. Their main result is to characterize the relation between the number of servers, the offered load, and system performance measures such as the probability of delay and the probability to abandon. This can be seen as an extension of the results of Halfin and Whitt (1981) by adding abandonments.

A number of approximations for the probability to abandon are developed by Boxma and de Waal (1994), who consider a multi-server queue with generally distributed service times and patience times. Brandt and Brandt (1999, 2002) consider a state-dependent Markovian multi-server queue with generally distributed patience times, in which the arrival rate depends on the number of customers in the system and in which the service rate depends on the number of busy servers. They derive the steady-state distribution of the number of customers in the system and various waiting-time distributions. The impact of the patience distribution on the performance is studied by Mandelbaum and Zeltyn (2004). They observe an approximate linearity between the abandonment probability and the average waiting time. To analyze multi-server queues with generally distributed service times and patience times, Whitt (2005) develops an algorithm to compute approximations for standard steady-state performance measures. One of his conclusions is that the behavior of the patience distribution near the origin primarily affects the performance. More recently, Dai and He (2011) and Ward (2012) confirm that conclusion in the context of an $M/M/s+G$ queue. Under the heavy-traffic regime, they propose to approximate the performance of an $M/M/s+G$ queue by that of an $M/M/s+M$ queue with patience parameter equal to the probability density function of the general distribution evaluated at $t=0$. For the real data sets considered in this paper, where patience times are generally distributed, we evaluate the quality of this approximation. Due to the irregularities of the data, it appears that the computation of the mass at zero may be not representative, and may then lead to inappropriate results.

Iravani and Balcıoğlu (2008) propose two approximations that are based on scaling the single-server queue to obtain estimates for the waiting-time distributions. Other papers have treated the impatience phenomenon under various assumptions. Related studies include those by Baccelli and Hebuterne (1981), Altman and Borovkov (1997), Ward and Glynn (2003), and Zeltyn and Mandelbaum (2005). The pioneer paper of Baccelli and Hebuterne (1981) followed by that of Zeltyn and Mandelbaum (2005) have led to very useful results about the performance analysis of queueing systems with impatient customers. They specifically characterize the virtual waiting time. This paper contributes to the literature related to the analysis of $M/M/s+G$ models, in the sense that it proposes new performance measures and explicitly compute them using the existing results on virtual waiting times.

Concerning the estimation of the patience distribution out of real call center data, published resources are scarce. Baccelli and Hebuterne (1981) show that an Erlang distribution with three phases works well. Kort (1983) proposes to model the patience distribution while waiting for a dial tone by the Weibull distribution. In Brown et al. (2005), it was observed that the patience distribution is not exponential as usually assumed for the call center models in the literature. However, an approximately exponential distribution was observed from a certain point in time onward. In this paper, we conduct a statistical analysis that confirms the non-exponentially of the patience distribution. Following the work of Brown et al. (2005), we consider the special distribution they observed and also propose another one that gives better results.

Main contributions: The main contributions of this paper can be summarized as follows.

- We propose to model the patience time using two different distributions. The first one is a discrete mass at zero corresponding to very impatient customers, and a remaining exponential distribution. The second one is a hyperexponential distribution with two phases. Using various sets of real call center data, we show that both models are appropriate. For a wide range of parameters, the hyperexponential model works in particular very well when compared to the real (empirical) model.

- We focus on the important feature of abandonments by providing a comprehensive list of metrics including abandonments. Many of these metrics have been already analyzed in the call center and queueing literature. However, what is new here, is that we propose new metrics and develop an approach (based on the results of Zeltyn and Mandelbaum (2005)) to explicitly derive their expressions.

- We establish for call center managers the importance of choosing the right metrics. For instance, we show the negative effect on staffing levels of choosing the widely-used metrics including short abandonments instead of those excluding short abandonments. To the contrary to regular abandonments, short abandonments may not be considered as a sign of bad service.

The remainder of this paper is organized as follows. In Section 2, we motivate our work and give the research objectives. In Section 3, we conduct a statistical analysis of abandonments on real call center data. Based on this analysis, we develop in Section 4 a call center model and give a list of various metrics including abandonments. We then show how to explicitly compute these metrics in a convenient way. In Section 5, we conduct a numerical analysis in which we draw comparisons between the performance indicators. We also extend our modeling by including the important feature of retrials, and by investigating its impact on the optimal staffing level. Finally in Section 6, we provide some concluding remarks and directions for future research.

2 Context and Research Objectives

Key performance indicators (KPIs) are critical for the successful management of call centers. The right metrics identify the causes of problems and generate solutions that change the results. It is almost impossible to develop a universal set of KPIs that will work equally well in every situation in every call center. Every business unit is different, with its unique structure and problems. Still, it is possible to formulate a set of KPIs useful for most call centers. Correct measurement of such KPIs will offer call center managers valuable information.

KPIs can be classified into two families: those that are product related and those that are process related. Product-related metrics are performance indicators mostly related to the content of the call, while process-related metrics are performance indicators that are related to call center operations.

2.1 Product-related metrics

The following is a list of the most well-known call center product-related metrics that managers can use to improve customer experience.

First Call Resolution (FCR) FCR measures the percentage of customer issues resolved the first time. A call held waiting in the queue that ended with solving the customer issue is better than a call that got instantly connected to an agent who could not properly help the customer. A call center maintaining a good FCR rate receives a small amount of calls coming from customers who have to call back because their issue was not resolved the first time. The call center avoids therefore a significant cost due to higher call volume, increased operating expenses, and dissatisfied customers.

Turnover Turnover measures the percentage of agents who leave a call center in say a year. Turnover can be voluntary (an agent chooses to leave) or involuntary (an agent is asked to leave). A high turnover rate is usually indicative of poor performance. It leads to high costs because of the investments in training and to problems related to agent availability.

Attendance and Punctuality Attendance is defined as an agent showing up for work on the scheduled day. Punctuality is defined as an agent showing up on time for the shift as well as being on time after breaks and lunch. One of the biggest challenges most call centers face is the control over attendance and punctuality. Low attendance and punctuality statistics can be very costly to a call center. In practice, it is common to offer incentives for good attendance and punctuality.

Contact Quality This is a common and critical customer-centric performance metric in all call centers, regardless of industry, function, and size. Top centers track contact quality as a high-level, center-wide metric, as well as an individual agent performance measure. Contact quality is typically assessed via a comprehensive evaluation form. Common quality criteria include the use of appropriate greetings and other call scripts, courtesy and professionalism, and grammar and spelling in text communication (e-mail and chat).

Customer Satisfaction Measuring customer satisfaction can be done through mail surveys and phone interviews days after the customer's interaction. Some call centers, with an advanced interactive voice response unit (IVR), survey callers immediately after the interaction occurs: customers are asked a series of questions about their interaction with the agent, their feelings about the organization, and their plans to continue doing business with the company.

2.2 Process-related metrics

In what follows, we give a list of the most familiar process-related metrics used in call centers.

Probability of Blocking It measures the percentage of customers that are not able to access the center at a given time due to insufficient network facilities in place. Failure to include a blocking target allows a call center to always meet its speed-of-answer goal simply by blocking the excess calls. This damages customer accessibility and satisfaction, even though the call center appears to be doing a great job of managing the queue.

Probability of Abandonment It measures the percentage of customers that abandon the queue while waiting, i.e., leave the system without service. Time before abandonment is customer specific. However, a call center can control abandonments by controlling the waiting time in the queue (which in turn affects abandonments). Abandonment is a measure associated with interactive channels, especially calls and web chat.

Short Abandonments The short-abandonment statistic represents the total number of customers that abandon before a specified short-abandonment time. Callers may abandon quickly for many reasons, for example, selecting an incorrect option in the IVR. Short abandonments, in contrast to regular abandonments, are not considered a sign of bad service. Therefore, many call centers count short abandonments differently, for example by not counting them at all.

Service Level, Average Speed of Answer, Longest Delay in Queue The service level measures the percentage of calls answered within a specified time limit. It is the most common process-related metric used in call centers. It is typically stated as X percent of calls handled in Y seconds or less. The average speed of answer (ASA) represents the average waiting time in the queue of all calls in a given period. Another delay measure is how long the oldest call in the queue has been waiting: the longest delay in queue (LDQ). A number of call centers use real-time LDQ to indicate when more staff need to be made immediately available. Historical data of LDQ can be used to indicate the "worst-case" experience of a customer over a period of time.

Agent Occupancy Agent occupancy is the measure of the actual time an agent is busy on customer contacts compared with the available or idle time, calculated by dividing workload hours by staff hours. Occupancy is an important measure of how well the call center has scheduled its staff and how efficiently it is using its resources.

2.3 Motivation

There is no single perfect or complete list of performance metrics for all call centers. On the one hand, basing an entire service strategy on the number of calls handled per hour or on the average speed of answer will inevitably lead to shortcomings in the quality. On the other hand, focusing too strongly on quality metrics while disregarding process-related measurements can still have an adverse effect on customer experience.

We focus on metrics related to queueing delays. These are classic process-related metrics that lie at the heart of effective call center and customer relations management. They are the clearest indication of what customers experience when they attempt to reach the call center. We refer the reader to Cleveland and Mayben (1997) for more details and discussions on process-related metrics. In this paper, we give particular attention to process-related metrics that include abandonments. The feature of customer abandonment is a crucial point in call centers.

One important point has to be clarified before impatience can be included in queueing models. That is, we need additional information concerning the patience, the willingness to wait until service commences. Similarly as for the input of the Erlang C model, the patience has to be determined from historical data. However, a number such as the average patience cannot be determined by simply averaging over the abandonment times. Indeed, the time at which other calls got connected tells us something about their patience, which should be taken into account. Statistical techniques exist to deal with these so-called censored data. Not using these methods can lead to a significant underestimation of patience, because the abandonments occur mostly among the very impatient customers. In Section 3, we conduct a statistical analysis on real call center data in order to characterize the statistical distribution of times before abandonments.

An advantage of Erlang C is that service-level expressions are relatively simple and easy to calculate. By taking abandonments into account, the computations become more difficult. Moreover, even when patience times are assumed to have an exponential distribution (the Erlang A model), there exist only expressions for some metrics, such as the conditional waiting time given service. In this paper, we give a comprehensive list of the metrics including abandonments, and explicitly derive the expressions for the probability distributions of these metrics. By doing so, we obtain existing results and derive new results, such as the conditional waiting time given service of the customers who do not have short patience times.

3 Statistical Analysis and Modeling of Abandonments

Call center data usually consist of very detailed records about the flow of calls through the call center. It is necessary to analyze these data in order to obtain estimates for the model parameters. Moreover, it is also important to validate the modeling assumptions, a step that is often overlooked. To analyze the patience, we need to know how long customers have spent waiting, and whether an abandonment occurred at the end of the waiting time. From customers that have abandoned, we know exactly what their patience is. However, from customers that did not abandon (but received service), we only know that their patience is greater than the time they have waited. To be more precise, we observe the minimum of the patience and the virtual waiting time, and we also know which one we observe. This is called right-censored data. Techniques exist to deal with censored data, one of which is the Kaplan-Meier estimator (see Kaplan and Meier 1958).

In our statistical analysis, we use data obtained from several real call centers. The data originate from a large banking call center located in the US, from a bank located in the Netherlands, and from a Dutch university medical center. Furthermore, we make use of the Anonymous Bank call center data available at *http://iew3.technion.ac.il/serveng/callcenterdata*.

To show the significance of uncensoring the data, consider the following example. On average 20 calls per minute arrive to a call center, and the average handling time is 5 minutes. To reach the target of 80% of the calls answered within 20 seconds, 108 agents should be scheduled according to the Erlang C formula. On a particular data set, the uncensored average patience turns out to be 780 seconds. When we apply the Erlang A model, we need 106 agents to reach the target. However, the censored average patience was 100 seconds. Using this number as the average patience, the Erlang A formula suggests an insufficient number of 95 agents. With 95 agents the real service level will only be 30%. This example also demonstrates that the Erlang C model can lead to erroneous results, since it does not predict any abandonment.

The result of the Kaplan-Meier estimator is the empirical cumulative distribution function $F(t)$ of the patience. By taking the derivative we can obtain the probability density function $f(t)$, and the hazard rate $h(t) = f(t)/(1 - F(t))$. In Figure 1 several empirical hazard rates are displayed. This figure also shows the hazard rates of a hyperexponential distribution, which will be discussed shortly, and of an exponential distribution. The mean of the exponential distribution corresponds to the average of the empirical patience times. The empirical hazard rates are smoothed three times using a moving-average filter with a span of five, to produce better-looking lines. The patience on all four data sets can, for the most part, be characterized in the same way. In the first couple of seconds the hazard rate is high, indicating very impatient customers who are not willing to wait at all. The hazard rate quickly becomes constant thereafter, which suggests that the patience from then on is exponential. Data sets 1 and 3 show additionally several peaks in the neighborhood of sixty seconds. This is because of delay announcements in the call handling system, that actually increase the likelihood of abandoning. We do not directly model delay announcements because this property is not shared by all call centers. We refer to Jouini et al. (2011), who study the impact of announcing delays in a setting of a single customer class with Markovian abandonments. Comparing the empirical hazard rate with the hazard rate of the exponential distribution reveals indeed that patience times are not exponential, and that the exponential distribution severely underestimates the abandonments.

Model 1

A way to model this customer behavior is to extend Erlang A by including the possibility of balking. Let T denote the random variable measuring the patience times. The distribution of T consists of a discrete mass at zero corresponding to very impatient customers, and a remaining exponential distribution for customers with a positive patience. We denote by α the probability that a customer,

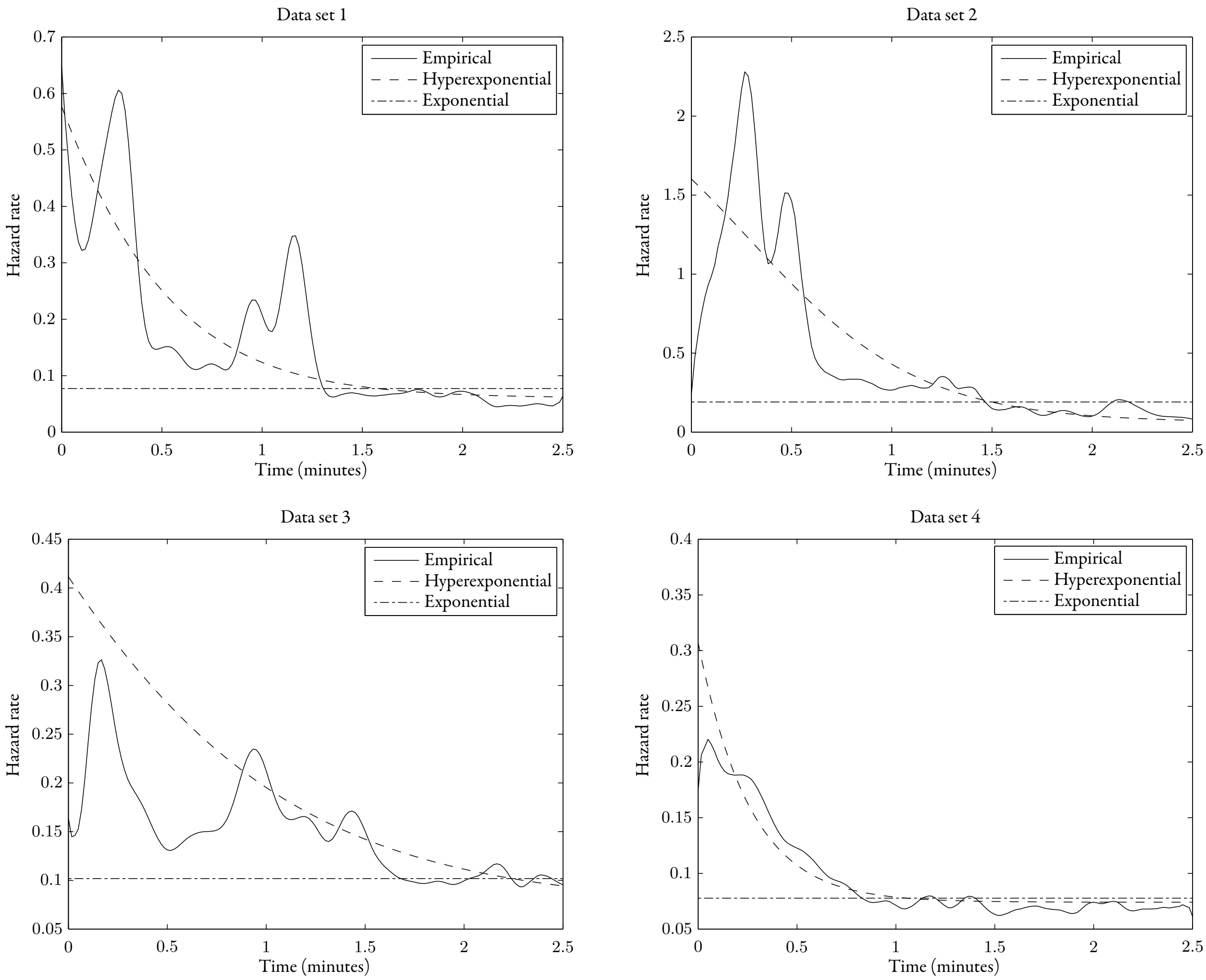

Figure 1. Hazard rates of the patience of four different data sets.

arriving to a busy system, will immediately balk. This feature models a non-negligible portion of the customers who immediately hang up once they know that they have to wait for service. On the other hand, with probability $1-\alpha$, customers who find a busy system will accept to join the queue. For these customers, the patience thresholds are independent and exponentially distributed with rate γ. Hence, the cumulative distribution function is

$$F_T(t) = \alpha + (1-\alpha)(1-e^{-\gamma t}),$$

for $t \geq 0$.

	Model 1		Model 2		
Data set	α	γ	p	γ_1	γ_2
1	0.1866	0.0656	0.2222	2.3843	0.0603
2	0.4626	0.1625	0.6593	2.3986	0.0617
3	0.1968	0.0864	0.2734	1.3100	0.0735
4	0.0506	0.0755	0.0583	4.0780	0.0742

Table 1. Parameters of the two models for the four data sets. The unit of time is minute.

Model 2

Another way to model customers' patience is by the hyperexponential distribution with two phases. The hyperexponential distribution is a mixture of two exponential distributions such that with probability p it is exponential with parameter γ_1 and with probability $1-p$ it is exponential with parameter γ_2. If T is hyperexponential, its cumulative distribution function F_T is given by

$$F_T(t) = p(1 - e^{-\gamma_1 t}) + (1-p)(1 - e^{-\gamma_2 t}),$$

for $t \geq 0$. By inspecting Figure 1, it seems that the hyperexponential distribution fits the empirical patience very well. The parameters of the random variable T are obtained by minimizing the mean squared error between $F(t)$ and $F_T(t)$. Table 1 lists these parameters for both models. From the figure we can deduce the following. Data set 4 is the perfect example of hyperexponential patience. The empirical hazard rate is approximately non-increasing, and the hazard rate of the hyperexponential distribution follows it very closely. The fits on data sets 1 and 2 also look reasonable, even though the empirical hazard rate starts out low for the first 0.25 minutes on data set 2. This could be explained by a welcome message that is played at the start of joining the queue. The empirical hazard rate is overestimated in the first minute on data set 3, but this fit is close afterwards.

The estimated parameters in Table 1 warrant additional attention. Since γ_1 is much higher than γ_2, with probability p a delayed customer will quickly abandon. This is equivalent to the modeling of balking with probability α in Model 1. This also agrees with the fact that γ and γ_2 are close to each other. Furthermore, we see that p is close to α and a bit higher. This is as expected since in Model 2 these very impatient customers do not necessarily abandon immediately. Next, we perform

	Hyperexponential		Balking + Exponential		Weibull		Erlang	
Data set	MSE	p-value	MSE	p-value	MSE	p-value	MSE	p-value
1	7.13e-5	0.747	6.72e-4	2e-12	7.68e-4	0.002	0.031	1e-30
2	2.32e-4	0.018	8.10e-3	4e-43	2.38e-3	1e-11	0.052	5e-35
3	1.40e-4	0.006	7.05e-4	8.6e-5	1.88e-4	0.006	0.031	6e-28
4	2.67e-5	0.974	6.98e-5	0.518	1.52e-4	0.424	0.014	9e-13

Table 2. Comparison of different patience distributions.

a statistical test to assess the fit of our models.

Earlier research by Baccelli and Hebuterne (1981) and Kort (1983) mentioned that the patience distribution could be Erlang with three phases or Weibull. In Table 2 we make a comparison between these distributions, together with the hyperexponential distribution and balking plus exponential, for different statistics. The first statistic is the mean squared error (MSE), which should be as low as possible for a good model. The second statistic is the p-value of the Kolmogorov-Smirnov test (see Massey 1951), which tests the null hypothesis that the empirical distribution and the tested distribution come from the same distribution. Values below the default significance level of 0.05 reject this hypothesis.

From the table it is clear that the statistics support the modeling of customers' patience by the hyperexponential distribution. If we look at the p-values of the Kolmogorov-Smirnov test, we observe that the null hypothesis is actually rejected on data sets 2 and 3 at a significance level of 0.05. However, for a significance level of 0.01, the null hypothesis will not be rejected for data set 2. For the model that includes balking these statistics are a bit misleading, since customers that balk are not always represented with a patience of zero in the data.

In conclusion, we presented two models for the patience based on real call center data. The first model is a simple extension of the Erlang A model by allowing customers to balk. The second model is a slightly more advanced model, where the patience is modeled by the hyperexponential distribution.

4 Analysis of Call Center Metrics

Consider a call center model with a single class of customers and s statistically identical, parallel servers. We assume that arrivals follow a Poisson process with rate λ, and that service times are exponentially distributed with rate μ. The queueing discipline is first-come first-served (FCFS). In addition, we let customers be impatient. As discussed in Section 3, we denote by T the random variable measuring patience times, and we consider two different ways to model T.

The performance measures we analyze next are based on the assumption that the system has reached steady state. Note that our model unconditionally reaches steady state for any random variable $T \neq 0$, see Garnett et al. (2002) for further details. Let τ be the acceptable waiting time and a be the threshold of short abandonments. In practice, reasonable values for τ and a are for example 20 and 5 seconds, respectively. For some managers, customers who immediately balk or those who enter the queue and quickly abandon before a are not really considered as unsatisfied. Therefore, such customers may not be accounted for in the service-level metric of the call center.

In Table 3, we define seven service levels that are useful in practice. We denoted them by SL_i, for $i = 1, \ldots, 7$. We present them, as is customary in call centers, in terms of the numbers of calls that arrive in a certain time period. Later on, we formulate them in terms of the corresponding random variables. The virtual waiting time is defined as the waiting time of customers assuming that they are not abandoning.

What should be the right metric? SL_1 and SL_4 do not give information about abandonments. SL_5 is hard to understand by managers and is also not directly measurable using historical data. For this reason it is, according to our experience, never used in call centers. However, this service-level definition dominates the Erlang A literature. SL_6 does not differentiate between waiting prior to service or to abandonment. SL_7 does not give information about waiting.

SL_2 and SL_3 exclude short abandonments which is a good aspect. The main drawback of these two metrics, similarly to all other metrics that use the parameter τ, is that they do not give any information on how long callers that have exceeded τ still have to wait. They entice managers to give priority to callers who have not yet reached the acceptable waiting time, thereby increasing even more the waiting time of callers that have waited longer than τ. Even though they have perverse

SL_1	$\dfrac{\#\text{ answered} \le \tau}{\#\text{ offered}}$
SL_2	$\dfrac{\#\text{ answered} \le \tau}{\#\text{ offered} - \#\text{ short abandonments}}$
SL_3	$\dfrac{\#\text{ answered} \le \tau}{\#\text{ offered} - \#\text{ abandoned} \le \tau}$
SL_4	$\dfrac{\#\text{ answered} \le \tau}{\#\text{ answered}}$
SL_5	$\dfrac{\#\text{ virtually answered} \le \tau}{\#\text{ offered}}$
SL_6	$\dfrac{\#\text{ sojourn in queue} \le \tau}{\#\text{ offered}}$
SL_7	$\dfrac{\#\text{ abandoned}}{\#\text{ offered}}$

Table 3. Service levels.

effects, these metrics are regularly used in practice. One way to avoid unwanted behavior is to add an objective on the performance of the customers who wait more than τ, or to use a different service-level objective. One possibility is to use the time that waiting exceeds τ. In contrast with the expected waiting time (the average speed of answer) it is sensitive to waiting-time variability. Another intuitive and simple solution is to use FCFS in all cases.

4.1 Computation of Service Levels

In this subsection, we derive the expressions for the service levels defined in Table 3. Let V_Q be the random variable denoting the virtual waiting time of a tagged, infinitely patient customer. In other words if the tagged customer finds a busy system upon arrival, this customer does not balk, neither abandon while waiting in the queue. Note that "answered" means $V_Q \le T$ and "abandoned" means $V_Q > T$. Let W_Q be the random variable measuring the sojourn time of a customer in the queue. This sojourn time will end either as a result of an abandonment or a start of service. Thus

$$W_Q = \min\{V_Q, T\}.$$

In what follows, we first give the expressions for the service levels in Table 3 as a function of the random variables V_Q, W_Q, and T. These expressions will be used later on to fully characterize the service levels. For an event E, $\mathbb{P}(E)$ is defined as the probability that E occurs. We denote by E^c

the complementary event of E, $\mathbb{P}(E^c) = 1 - \mathbb{P}(E)$. We can write the first service level as

$$\mathrm{SL}_1 = \mathbb{P}(V_Q \leq \tau, V_Q \leq T). \tag{1}$$

The second service level is

$$\mathrm{SL}_2 = \frac{\mathbb{P}(V_Q \leq \tau, V_Q \leq T)}{\mathbb{P}((T < V_Q, T < a)^c)}.$$

Since the patience of a customer is independent of all other events, we have $\mathbb{P}(T > a, V_Q > a) = \mathbb{P}(T > a)\mathbb{P}(V_Q > a)$. Observing now that

$$\mathbb{P}((T < V_Q, T < a)^c) = \mathbb{P}(T > a, V_Q > a) + \mathbb{P}(V_Q \leq a, V_Q \leq T),$$

we obtain

$$\mathrm{SL}_2 = \frac{\mathbb{P}(V_Q \leq \tau, V_Q \leq T)}{\mathbb{P}(T > a)\mathbb{P}(V_Q > a) + \mathbb{P}(V_Q \leq a, V_Q \leq T)}. \tag{2}$$

Similarly, SL_3 is given by

$$\mathrm{SL}_3 = \frac{\mathbb{P}(V_Q \leq \tau, V_Q \leq T)}{\mathbb{P}(T > \tau)\mathbb{P}(V_Q > \tau) + \mathbb{P}(V_Q \leq \tau, V_Q \leq T)}. \tag{3}$$

We also have

$$\mathrm{SL}_4 = \frac{\mathbb{P}(V_Q \leq \tau, V_Q \leq T)}{\mathbb{P}(V_Q \leq T)}, \tag{4}$$

$$\mathrm{SL}_5 = \mathbb{P}(V_Q \leq \tau), \tag{5}$$

$$\mathrm{SL}_6 = \mathbb{P}(W_Q \leq \tau), \tag{6}$$

and finally

$$\mathrm{SL}_7 = \mathbb{P}(V_Q > T). \tag{7}$$

In the next subsection, we explicitly derive the previous expressions for the service levels $\mathrm{SL}_1, \ldots, \mathrm{SL}_7$.

4.2 Explicit Expressions for Service Levels

Let the patience be generally distributed with cumulative distribution function $G(x)$, $x \geq 0$. Based on the work of Baccelli and Hebuterne (1981), Zeltyn and Mandelbaum (2005) define the following building blocks for performance analysis. Let $\bar{G}(x) = 1 - G(x)$. Define

$$\begin{aligned} H(x) &= \int_0^x \bar{G}(u)\mathrm{d}u, \\ J(t) &= \int_t^\infty e^{\lambda H(x) - s\mu x}\mathrm{d}x, \\ J_1(t) &= \int_t^\infty x e^{\lambda H(x) - s\mu x}\mathrm{d}x, \\ J_H(t) &= \int_t^\infty H(x) e^{\lambda H(x) - s\mu x}\mathrm{d}x. \end{aligned}$$

Let $J = J(0)$, $J_1 = J_1(0)$, $J_H = J_H(0)$, and define

$$\mathcal{E} = \frac{\displaystyle\sum_{i=0}^{s-1} \frac{(\lambda/\mu)^i}{i!}}{\dfrac{(\lambda/\mu)^{s-1}}{(s-1)!}} = B(s-1, \lambda/\mu)^{-1},$$

where $B(s, \lambda/\mu)$ is the blocking probability in the $M/M/s/s$ queue. Expressed in these building blocks, the probability density function of the virtual waiting time V_Q is, for $x > 0$,

$$v(x) = \frac{\lambda e^{\lambda H(x) - s\mu x}}{\mathcal{E} + \lambda J},$$

with a mass at the origin with value

$$\mathbb{P}(V_Q = 0) = \frac{\mathcal{E}}{\mathcal{E} + \lambda J}.$$

Zeltyn and Mandelbaum (2005) derive a number of performance measures, including

$$\begin{aligned} \mathbb{P}(A) &= \frac{1 + (\lambda - s\mu)J}{\mathcal{E} + \lambda J}, \\ \mathbb{E}V_Q &= \frac{\lambda J_1}{\mathcal{E} + \lambda J}, \\ \mathbb{E}W_Q &= \frac{\lambda J_H}{\mathcal{E} + \lambda J}, \\ \mathbb{P}(V_Q > \tau) &= \frac{\lambda J(\tau)}{\mathcal{E} + \lambda J}, \\ \mathbb{P}(W_Q > \tau) &= \frac{\lambda \bar{G}(\tau) J(\tau)}{\mathcal{E} + \lambda J}, \end{aligned}$$

where $\mathbb{P}(A)$ is the probability to abandon.

The first service level can be obtained from Equation (1) and these building blocks in the following way.

$$\begin{aligned}\mathrm{SL}_1 &= \mathbb{P}(V_Q = 0) + \int_0^\tau \bar{G}(x)v(x)\mathrm{d}x \\ &= \frac{\mathcal{E}}{\mathcal{E}+\lambda J} + \int_0^\tau \bar{G}(x)\frac{\lambda e^{\lambda H(x)-s\mu x}}{\mathcal{E}+\lambda J}\mathrm{d}x.\end{aligned}$$

From

$$\int_0^\tau (\lambda\bar{G}(x) - s\mu)e^{\lambda H(x)-s\mu x}\mathrm{d}x = e^{\lambda H(x)-s\mu x}\Big|_0^\tau = e^{\lambda H(\tau)-s\mu\tau} - 1,$$

it follows that

$$\begin{aligned}\int_0^\tau \lambda\bar{G}(x)e^{\lambda H(x)-s\mu x}\mathrm{d}x &= \int_0^\tau (\lambda\bar{G}(x) - s\mu)e^{\lambda H(x)-s\mu x}\mathrm{d}x \\ &\quad + \int_0^\tau s\mu e^{\lambda H(x)-s\mu x}\mathrm{d}x \\ &= e^{\lambda H(\tau)-s\mu\tau} - 1 + s\mu(J - J(\tau)),\end{aligned}$$

and hence

$$\mathrm{SL}_1 = \frac{\mathcal{E} + e^{\lambda H(\tau)-s\mu\tau} - 1 + s\mu(J - J(\tau))}{\mathcal{E}+\lambda J}.$$

Using that $\mathbb{P}(T > a) = \bar{G}(a)$ and

$$\mathbb{P}(V_Q > a) = \frac{\lambda J(a)}{\mathcal{E}+\lambda J},$$

Equation (2) gives

$$\mathrm{SL}_2 = \frac{\mathcal{E} + e^{\lambda H(\tau)-s\mu\tau} - 1 + s\mu(J - J(\tau))}{\bar{G}(a)\lambda J(a) + \mathcal{E} + e^{\lambda H(a)-s\mu a} - 1 + s\mu(J - J(a))}.$$

Similarly, Equation (3) becomes

$$\mathrm{SL}_3 = \frac{\mathcal{E} + e^{\lambda H(\tau)-s\mu\tau} - 1 + s\mu(J - J(\tau))}{\bar{G}(\tau)\lambda J(\tau) + \mathcal{E} + e^{\lambda H(\tau)-s\mu\tau} - 1 + s\mu(J - J(\tau))}.$$

Using that $\mathbb{P}(V_Q < T)$ is the probability of service, Equation (4) leads to

$$\mathrm{SL}_4 = \frac{\mathcal{E} + e^{\lambda H(\tau)-s\mu\tau} - 1 + s\mu(J - J(\tau))}{\mathcal{E} + s\mu J - 1}.$$

Finally, Equations (5)–(7) are

$$\mathrm{SL}_5 = 1 - \frac{\lambda J(\tau)}{\mathcal{E} + \lambda J}.$$

$$\mathrm{SL}_6 = 1 - \frac{\lambda \bar{G}(\tau) J(\tau)}{\mathcal{E} + \lambda J}.$$

$$\mathrm{SL}_7 = \frac{1 + (\lambda - s\mu) J}{\mathcal{E} + \lambda J}.$$

Using the specific functional form of the patience distribution T, we can determine $\bar{G}(x)$ and $H(x)$ in closed form. For Model 1, with exponential patience times, these are

$$\begin{aligned} \bar{G}(x) &= (1-\alpha) e^{-\gamma x}, \\ H(x) &= \frac{1-\alpha}{\gamma}(1 - e^{-\gamma x}), \end{aligned}$$

and for Model 2, with hyperexponential patience times, we get

$$\begin{aligned} \bar{G}(x) &= p e^{-\gamma_1 x} + (1-p) e^{-\gamma_2 x}, \\ H(x) &= \frac{p}{\gamma_1}(1 - e^{-\gamma_1 x}) + \frac{1-p}{\gamma_2}(1 - e^{-\gamma_2 x}). \end{aligned}$$

The function $J(t)$ cannot be given in closed form for these models. On the other hand, there are no difficulties in evaluating $J(t)$ numerically. We have computed all the expressions for the service levels. These expressions will be used next for the numerical illustrations.

5 Numerical Experiments

In this section we first validate the modeling approaches by comparing results of numerical experiments with the empirical results. After that, we show the effect of different service levels on the staffing levels.

5.1 Comparison Between the Models

To illustrate the performance of the two models, we consider the four data sets again and compute the service level given by SL_1. The parameters related to the patience distributions of Model 1 and

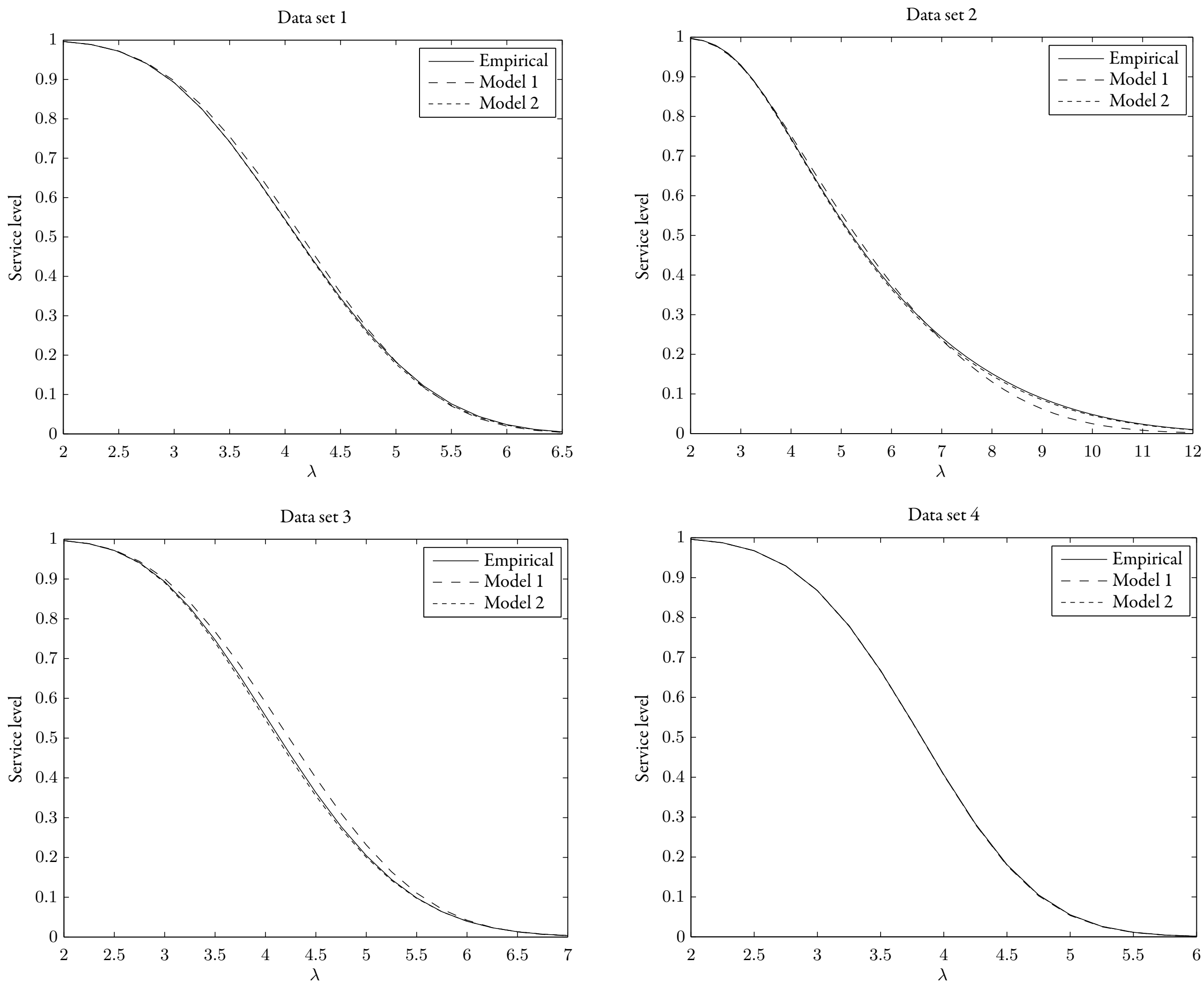

Figure 2. Comparison between the models for a small system.

Model 2 are given in Table 1. We compare the service-level estimates of these models with the empirical service level. The empirical service level is obtained from the model that directly uses the empirical patience distribution. The analysis is in line with the analysis in Section 4, except that the function $H(x)$ has to be evaluated numerically as well.

As a first example, we consider a relatively small system with the following parameters: $\mu = 0.2$, $s = 19$, $\tau = 1/3$, and a varying λ. The empirical service level and the service-level estimates of both models are depicted in Figure 2. All plots show that both models have an excellent performance. There are only some small differences noticeable on data sets 2 and 3 for Model 1. This gives us confidence in the usefulness of our models.

As a second example, we consider a larger system defined by $\mu = 0.2$, $s = 210$, $\tau = 1/3$, and a

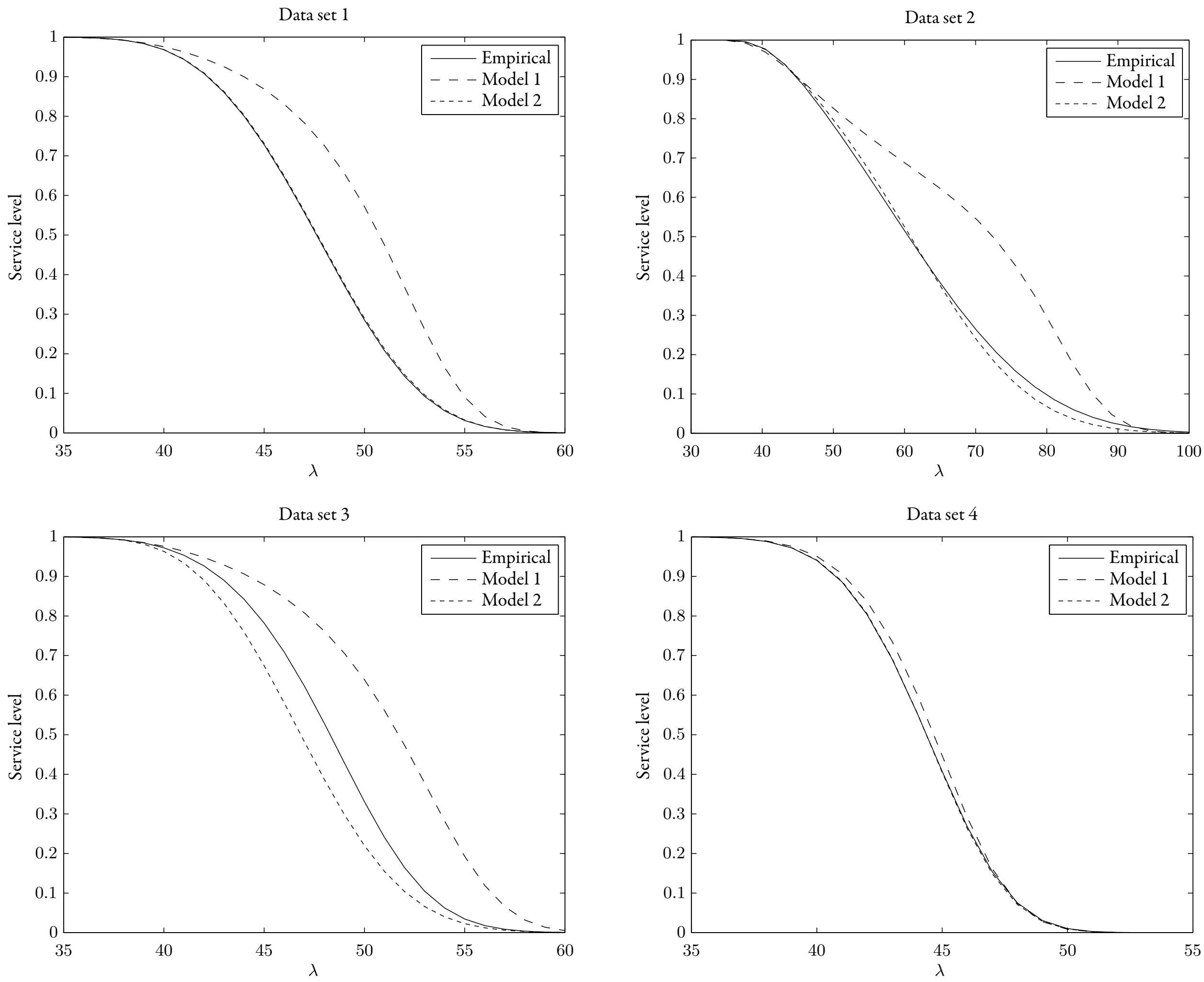

Figure 3. Comparison between the models for a large system.

varying λ. The results of the comparison are shown in Figure 3. Here we observe mixed results. Model 2 has perfect accuracy on data sets 1 and 4, and a very good performance on data set 2. The service level is underestimated on data set 3. This could perhaps be explained by Figure 1, since the fit of the hyperexponential distribution is there not perfect as well. The results for Model 1 are different. The performance on data set 4 is accurate, but the performance on the other data sets is poor. The service level is clearly overestimated. Looking on data set 2, Model 1 is clearly not appropriate.

We also considered SL_7, the abandonment probability. For the sake of conciseness, we only note that both models perform very well with respect to this metric. Only in case of the larger system Model 1 has some minor discrepancies.

All in all, we can conclude from these experiments that both models are useful to model customers' patience for relatively small systems. When the size of the system increases, the model where the patience is modeled by the hyperexponential distribution is preferred. The accuracy of this model compared with empirical results is almost perfect.

5.2 Comparison Between the Metrics

Let us consider the metrics SL_1, SL_2, and SL_3. We choose an acceptable waiting time $\tau = 1/3$ minute. The patience is modeled by the empirical patience of data set 2. Customers who abandon before $a = 5$ seconds are considered as short abandonments, i.e., these are not big issues for the call center manager. The service rate is $\mu = 1$ per minute. We consider different objective values for $\mathrm{SL}_i^* \in \{50\%, 80\%, 95\%, 99\%\}$, for $i = 1, 2, 3$. For each objective SL_i^* and for each $\lambda \in \{3, 5, 7, 10, 15, 20, 30, 50\}$ we compute the optimal staffing level s_i^*. The results are given in Figure 4.

Note that in practice managers usually use SL_1 which is not appropriate since we are penalized with customers who are very impatient. These customers do not really experience frustration. A better metric would be SL_2 which ignores short abandonments. An even better metric could be SL_3 which ignores abandonments within the acceptable waiting time. An additional benefit from using these last two metrics, as shown in Figure 4, is that the required staffing levels are lower than those for SL_1.

To go further and confirm the interest of SL_2 and SL_3, it is worth to look on the behavior of the probability of abandonment. To do so, we consider the case $\mathrm{SL}_i^* = 80\%$ with the same optimal staffing levels s_i^* as shown in Figure 4 ($i = 1, 2, 3$). In Table 4, we vary λ and give both the probabilities of abandonment, SL_7, and that of abandoning after τ, say SL_8. The latter can been seen as a "reasonable" probability of abandonment, which does not include customers who abandon before τ. SL_8 can be computed as follows. From Equations (1) and (3), we may write

$$\mathbb{P}(\text{abandon before } \tau) = 1 - \frac{\mathrm{SL}_1}{\mathrm{SL}_3}. \tag{8}$$

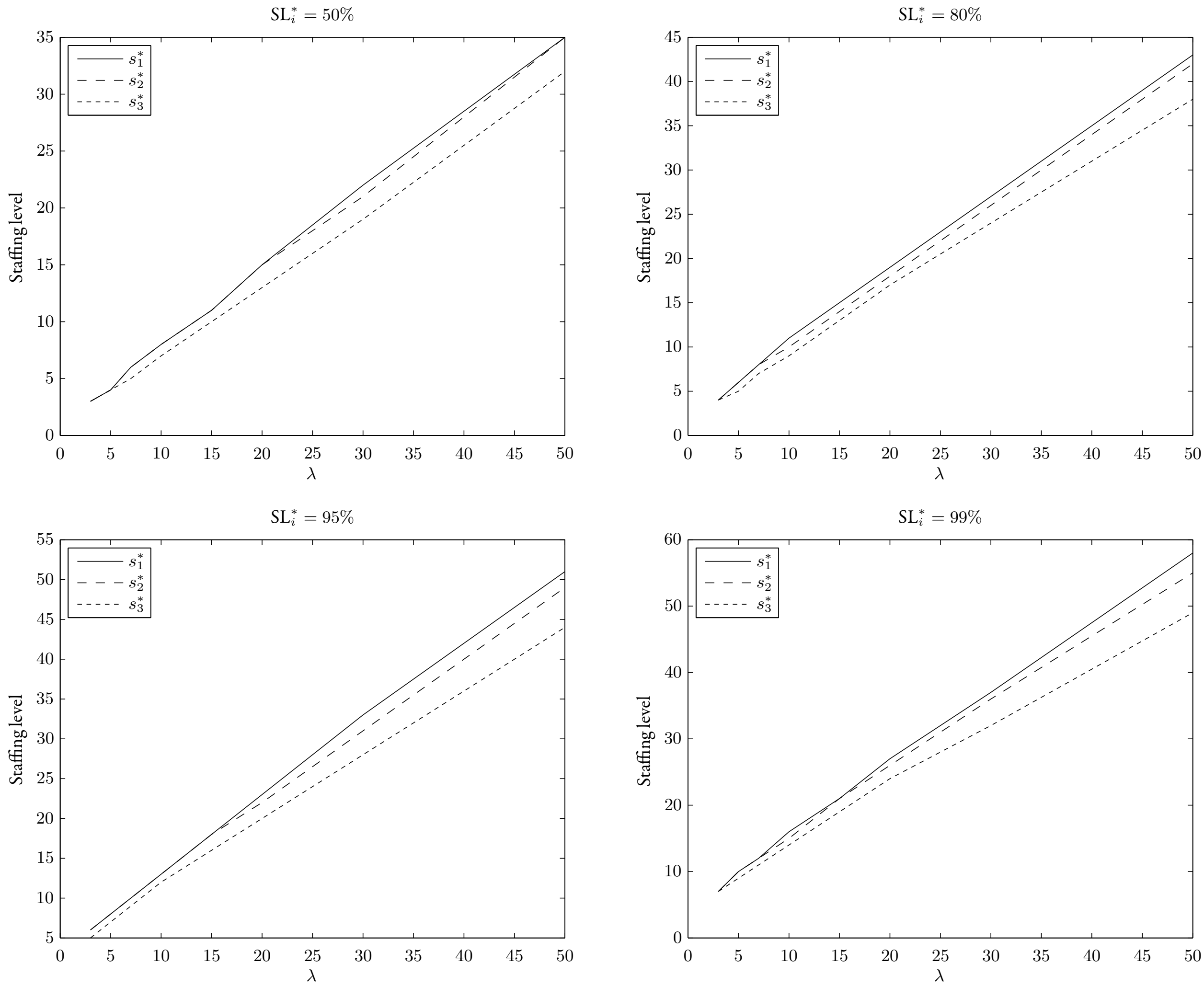

Figure 4. Optimal staffing levels.

Knowing that $\mathrm{SL}_7 = \mathbb{P}(\text{abandon before } \tau) + \mathbb{P}(\text{abandon after } \tau)$, Equation (8) leads to

$$\mathrm{SL}_8 = \mathrm{SL}_7 + \frac{\mathrm{SL}_1}{\mathrm{SL}_3} - 1.$$

From Table 4, we see that the performance in terms of abandonments after τ are acceptable for the metrics SL_2 and SL_3 (while they do need lower staffing levels). This comment is particularly relevant for large call centers, due to the benefit of pooling on performance.

5.3 Comparison with $M/M/s + G$ and $M/M/s + M$ Models

In this section, we further compare Models 1 and 2 with existing models in the literature. The diffusion results from Dai and He (2011) and Ward (2012) show that the performance of an $M/M/s + G$ queue can be approximated by that of an $M/M/s + M$ with an abandonment rate

λ	3	5	7	10	15	20	30	50
s_1^*	4	6	8	11	15	19	27	43
s_2^*	4	6	8	10	14	18	26	42
s_3^*	4	5	7	9	13	17	24	38
SL_7 with s_1^*	0.112	0.104	0.096	0.086	0.105	0.117	0.133	0.151
SL_7 with s_2^*	0.112	0.104	0.096	0.129	0.141	0.148	0.158	0.168
SL_7 with s_3^*	0.112	0.182	0.155	0.184	0.184	0.184	0.212	0.242
SL_8 with s_1^*	0.022	0.016	0.012	0.008	0.007	0.006	0.005	0.003
SL_8 with s_2^*	0.022	0.016	0.012	0.014	0.012	0.010	0.007	0.003
SL_8 with s_3^*	0.022	0.035	0.023	0.025	0.019	0.015	0.013	0.010

Table 4. Probability of abandonment (SL_7), and probability of abandoning after τ (SL_8) for $\mathrm{SL}_i^* = 80\%$.

λ	3	5	7	10	15	20	30	50
Exact (empirical)	5	7	9	12	16	21	30	49
Model 1	5	7	9	11	16	20	29	46
Model 2	5	7	9	12	16	21	30	49
Model 3 ($M/M/s+G$)	5	7	8	11	16	20	29	47
Model 4 ($M/M/s+M$)	5	7	9	12	18	23	33	52

Table 5. Optimal staffing levels for $\mathrm{SL}_1 = 80\%$, data set 1.

equal to $f(0)$, where $f(\cdot)$ is the probability density function of the general patience time (empirical distribution). This leads to a new model, referred to as Model 3. The last model for the comparison is referred to as Model 4, and it simply consists on an $M/M/s+M$ where the mean patience parameter is set to the mean patience of the empirical distribution.

We are interested in evaluating the quality of the modeling of the patience distribution of the 4 models. We then compute the optimal staffing level from each proposed model and compare it with the exact one from the empirical patience distribution. The optimal staffing level is the minimum required number of agents satisfying the service level requirement $\mathrm{SL}_1 = 80\%$. We choose $\mu = 1$ and $\tau = 1/3$, and vary the arrival rate. The results for data sets 1 and 2 are shown in Tables 5 and 6, respectively.

From the experiments we again verify the robustness of Model 2 (hyperexponential patience) which leads to the exact optimal staffing levels. Consistently with the results above, Model 1 can be

λ	3	5	7	10	15	20	30	50
Exact (empirical)	4	6	8	11	15	19	27	43
Model 1	5	6	8	11	15	19	27	43
Model 2	4	6	8	11	15	19	27	43
Model 3 ($M/M/s+G$)	5	7	9	12	17	22	31	50
Model 4 ($M/M/s+M$)	5	7	9	12	17	22	32	51

Table 6. Optimal staffing levels for $\mathrm{SL}_1 = 80\%$, data set 2.

considered as appropriate only for small systems. For large systems, it overestimate the performance and therefore lead to understaffing situations (Table 5). The M/M/s+G model cannot be really considered as reliable. Computing $f(0)$ from real data can be dangerous (not representative) and may lead to wrong results (Table 6). The M/M/s+M model is the worst. The simple exponential patience modeling is a bad approximation.

5.4 Extension to the Case with Retrials

In this section, we extend the analysis by allowing retrials. In practice, some of the customers who balk or abandon will redial and try to access the call center again. For more details on the modeling and analysis of call centers with retrials, we refer the reader to Aguir et al. (2004) and Pustova (2010), and references therein.

In what follows, we consider a simple modeling of the retrials and analyze its impact on the optimal staffing level. We want to assess how ignoring retrials may lead to an insufficient staffing level. Let us consider a model with generally distributed patience times. We allow some of the customers who balk or abandon to call back the call center. We denote by θ the probability that one will call back. Delays before customers' call back are assumed to be i.i.d. random variables with a general distribution. For tractability purposes, we assume independence between successive calls in terms the probability to call back. Let $\bar{\lambda}$ be the overall arrival rate to the system, i.e., the sum of the primary calls and the feedback calls. Then,

$$\bar{\lambda} = \frac{\lambda}{1 - \theta \mathbb{P}(A \mid \bar{\lambda})}, \tag{9}$$

where λ is the arrival rate of the primary calls, and the probability to abandon $\mathbb{P}(A \mid \bar{\lambda}) = \mathrm{SL}_7$ is computed based on $\bar{\lambda}$. By varying θ from 0 to 1, we move from our original system with no retrials to a system with high retrials. This simple model falls into the class of product-form networks analyzed by Baskett et al. (1975). As a result, the stationary behavior of this queueing model does not depend on the distribution of the call-back delays. They can thus be ignored. Using this simple modeling, we capture the retrial feature while being able to use the results developed in Section 4. To evaluate the performance of a call center with parameters λ, μ, s, θ, and generally distributed patience times, it suffices to use the results of Section 4 for a call center with parameters $\bar{\lambda}$, μ, s, and the same patience distribution.

Before going further, we should discuss how to compute $\bar{\lambda}$. Denoting the right-hand side in Equation (9) by a continuous function g in $\bar{\lambda}$, we may write $\bar{\lambda} = g(\bar{\lambda})$. Then, $\bar{\lambda}$ is said to be a fixed point of g. To numerically compute $\bar{\lambda}$, we use a fixed-point algorithm (see for example Karamardian and García (1977)).

In what follows, we want to numerically study the impact of retrials on the optimal staffing level. We choose the metric SL_2, and our objective is to compute the optimal staffing level for $\mathrm{SL}_2^* = 80\%$, $a = 5$ seconds, $\tau = 1/3$ minute. We consider different call center sizes. The primary arrival rates are $\lambda \in \{10, 30, 50\}$. We also consider different levels of retrials, $\theta \in [0, 1]$. Similarly to the previous subsection, we choose $\mu = 1$ and model the patience times by the empirical distribution of data set 2. The results are given in Figure 5. As expected, this figure reveals that the optimal staffing level, s_2^*, increases in the probability to call back θ.

An important observation here is that the impact of call backs on s_2^* can be high, although we start from a model with no retrials ($\theta = 0$) that has a high service level ($\mathrm{SL}_2 = 80\%$). For instance, the optimal staffing level can increase by about 20% when moving from a system with no retrials to a system with high retrials ($\theta = 1$). This is particularly due to the high impatience of customers in data set 2. In such cases, ignoring the modeling of retrials may lead to inappropriate results.

6 Conclusion

We have analyzed various process-related call center metrics that include customer abandonment. We showed how to obtain existing results explicitly, and also derived new results for new metrics

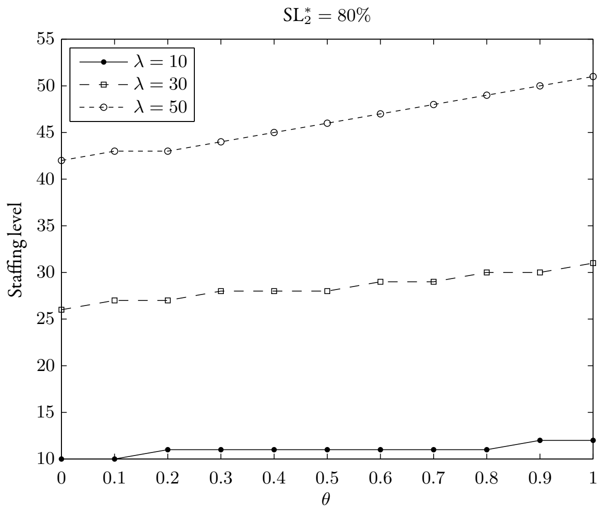

Figure 5. The impact of retrials.

considering short abandonments or abandonments within the acceptable waiting time. In practice, many managers choose not to count short abandonments against the call center performance metrics. Although the models used here are simple (with Markovian assumptions), we have shown their robustness using real call center data. Through numerical analysis, we have also discussed the advantages and disadvantages of the different metrics.

We have presented two models for customers' patience that have a very good agreement with reality. The method to derive the call center metrics works for empirical patience distributions as well. The benefit of using our models is that the Markovian property is preserved. This is especially useful when one wants to consider other service-time distributions.

There are several avenues for future research. It would be useful to extend the analysis to the case of more than one customer type with non-identically distributed patience and service times. Another interesting and challenging extension of the current analysis is to consider a non-stationary arrival process.

Acknowledgements

The authors would like to express their gratitude to the Department Editor, the Associate Editor and to the two anonymous referees for their very interesting and constructive comments, that have considerably improved the content of the paper over its earlier version.

References

M.S. Aguir, F. Karaesmen, O.Z. Akşin, and F. Chauvet. The impact of retrials on call center performance. *OR Spectrum*, 26(3):353–376, 2004.

E. Altman and A.A. Borovkov. On the stability of retrial queues. *Queueing Systems*, 26(3/4): 343–363, 1997.

F. Baccelli and G. Hebuterne. On queues with impatient customers. In *Performance '81*, pages 159–179. North-Holland, 1981.

F. Baskett, K.M. Chandy, R.R. Muntz, and F.G. Palacios. Open, closed, and mixed networks of queues with different classes of customers. *Journal of the American Statistical Association*, 22(2): 248–260, 1975.

O.J. Boxma and P.R. de Waal. Multiserver queues with impatient customers. In J. Labetoulle and J.W. Roberts, editors, *Proceedings of the 14th International Teletraffic Congress*, pages 743–756, 1994.

A. Brandt and M. Brandt. On the $M(n)/M(m)/s$ queue with impatient calls. *Performance Evaluation*, 35(1):1–18, 1999.

A. Brandt and M. Brandt. Asymptotic results and a Markovian approximation for the $M(n)/M(n)/s + GI$ system. *Queueing Systems*, 41(1/2):73–94, 2002.

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

B. Cleveland and J. Mayben. *Call Center Management on Fast Forward: Succeeding in Today's Dynamic Inbound Environment*. Call Center Press, 1st edition, 1997.

J. G. Dai and S. He. Queues in Service Systems: Customer Abandonment and Diffusion Approximations. *Tutorials in Operations Research*, pages 36–59, 2011.

P.D. Feigin. Analysis of customer patience in a bank call center. Working paper, 2005.

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

O. Garnett, A. Mandelbaum, and M. Reiman. Designing a call center with impatient customers. *Manufacturing & Service Operations Management*, 4(3):208–227, 2002.

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

F. Iravani and B. Balcıoǧlu. Approximations for the $M/GI/N + GI$ type call center. *Queueing Systems*, 58(2):137–153, 2008.

O. Jouini, O.Z. Akşin, and Y. Dallery. Call centers with delay information: Models and insights. *Manufacturing & Service Operations Management*, 13(4):534–548, 2011.

E.L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481, 1958.

S. Karamardian and C.B. García. *Fixed Points: Algorithms and Applications*. Academic Press Rapid Manuscript Reproduction. Academic Press, 1977.

B.W. Kort. Models and methods for evaluating customer acceptance of telephone connections. In *GLOBECOM '83*, pages 706–714. IEEE, 1983.

A. Mandelbaum and S. Zeltyn. The impact of customers' patience on delay and abandonment: some empirically-driven experiments with the M/M/n+G queue. *OR Spectrum*, 26(3):377–411, 2004.

A. Mandelbaum and S. Zeltyn. Staffing many-server queues with impatient customers: Constraint satisfaction in call centers. *Operations Research*, 57(5):1189–1205, 2009.

F.J. Massey. The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46(253):68–78, 1951.

S.V. Pustova. Investigation of call centers as retrial queuing systems. *Cybernetics and Systems Analysis*, 46(3):494–499, 2010.

A.R. Ward. Asymptotic Analysis of Queueing Systems with Reneging: A survey of Results for FIFO, Single Class Models. *Surveys in Operations Research and Management Science*, 17:1–14, 2012.

A.R. Ward and P.W. Glynn. A diffusion approximation for a Markovian queue with reneging. *Queueing Systems*, 43(1/2):103–128, 2003.

W. Whitt. Engineering solution of a basic call-center model. *Management Science*, 51(2):221–235, 2005.

S. Zeltyn and A. Mandelbaum. Call centers with impatient customers: Many-server asymptotics of the M/M/n+G queue. *Queueing Systems*, 51(3/4):361–402, 2005.