


Application of queueing models with abandonment for Call Center congestion analysis

Aplicação de modelos de filas com abandono para análise de congestão em call centers

Sidney Carlos Ferrari¹, Reinaldo Morabito¹ 

¹ Universidade Federal de São Carlos – UFSCar, Departamento de Engenharia de Produção, São Carlos, SP, Brazil, E-mail. ferrarisc@gmail.com; morabito@ufscar.br

How to cite: Ferrari, S. C., & Morabito, R. (2020). Application of queueing models with abandonment for Call Center congestion analysis. *Gestão & Produção*, 27(1), e3765. <https://doi.org/10.1590/0104-530X3765-20>

Abstract: This paper studies and applies queueing systems to Call Centers regarding the possibility of customer abandonment from the system before being served due to their impatience in waiting for a service. Call Centers are service organizations that predominantly serve customers via phone calls. One of the main concerns in managing them is to provide quality service at a minimum cost. Noticing the quality of services offered is expressed by customers, for example by abandonment from the queue. This paper shows that the M/M/c+G analytical queueing models with abandonment, with patience time represented by generic distributions (particularly mixed distributions), are more effective than the M/M/c+M analytical queueing models with abandonment, with Exponential patience, commonly used to evaluate congestion problems in Call Centers and support sizing and operational decisions in these systems. We conducted a study using data extracted from a Bank Call Center located in Israel and the parameters and some performance measures are determined based on this data. These sampling measures are compared with the same measures achieved by the M/M/c+M and M/M/c+G analytical queueing models considered in this research, which use parameters obtained empirically and the mixed and non-mixed distributions based on Exponential and Lognormal to represent user patience. An experimental discrete simulation model was also used to explore an alternative scenario, showing the potential of using the approaches based on analytical models with abandonment for Call Center analysis.

Keywords: Call Center; Contact Center; Impatient customers; Queueing models with abandonment; Mixed distributions; Congestion Analysis; Simulation.

Resumo: Este artigo estuda e aplica sistemas de filas para *Call Centers* considerando a possibilidade de o cliente abandonar o sistema antes de ser servido, devido a sua impaciência na espera do atendimento. Os *Call Centers* são organizações de serviço que predominantemente servem os clientes via chamada telefônica e uma das principais preocupações nas suas gestões é oferecer serviço de qualidade com mínimo custo. A percepção da qualidade dos serviços oferecidos é manifestada pelo cliente, por exemplo, por meio do abandono da fila de espera. Neste trabalho mostra-se que modelos analíticos de fila com abandono M/M/c+G, com tempo de paciência representado por distribuições genéricas (particularmente distribuições mistas), são mais efetivos do que os modelos analíticos de fila

Received Jan. 24, 2017 - Accepted July 20, 2017

Financial support: The authors thank CNPq and CAPES for the financial support.



This is an Open Access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

com abandono $M/M/c+M$, com paciência Exponencial, comumente empregados para avaliar o problema de congestão em *Call Centers* e apoiar decisões de dimensionamento e operação nesses sistemas. Um estudo foi conduzido com dados extraídos do *Call Center* de um Banco localizado em Israel e os parâmetros e algumas medidas de desempenho são determinadas com esses dados. Essas medidas amostrais são comparadas com as mesmas medidas obtidas por meio dos modelos analíticos de fila $M/M/c+M$, e $M/M/c+G$ considerados nesse estudo, que utilizam os parâmetros obtidos empiricamente e as distribuições mistas e não mistas baseadas na Exponencial e Lognormal para representar a paciência dos usuários. Utilizou-se, também, um modelo experimental de simulação discreta para explorar um cenário alternativo, mostrando o potencial do uso de abordagens baseadas em modelos analíticos com abandono para análise de *Call Centers*.

Palavras-chave: Call Center; Contact Center; Clientes impacientes; Modelos de fila com abandono; Distribuições mistas; Análise de congestão; Simulação.

1 Introduction

Call Centers, also called Contact Centers, are now one of the main communication channels between companies and their customers in various sectors. Knowing customers and offering what they need is a basic requirement for operations to be increasingly customized, meeting their expectations. Call Centers are service organizations that predominantly attend to customers via telephone calls and are considered a particular type of Contact Center that serve their customers by telephone, fax, email, chat, mobile devices and other communication channels.

The Call Center industry has expanded rapidly and, due to this growth, Call Center management has become a complex activity. According to Lima (2012), Call Center managers must select a set of more appropriate metrics that will support strategic and operational decision making.

As the purpose of this article is to analyze Call Centers from a quantitative point of view, we decided to adopt user abandonment measured by the probability of abandonment as metrics for the analyses to be carried out, among those considered important by managers. It is an indicator in contracts between Call Centers and companies that occurs when a user hangs up before being served by an agent. In addition to this metric, it is essential to know the waiting time, measured by its average, and by the probability of waiting, which indicates the time the user waits to be served (Lima, 2012). The values of these metrics extracted from the observed data from a Call Center are compared with those obtained from mathematical and simulation models.

We chose these metrics considering that the main concern in Call Center management is to achieve a level of quality service subject to a specified budget. This quality can be measured in two ways: qualitatively and quantitatively (Mandelbaum & Zeltyn, 2005). From the quantitative point of view, the quality of services is perceived by users mainly by the time of waiting in the queue until being attended to by the agents. When this waiting time is long, users become tired and may abandon the queue, without having received an answer to their request. Abandonment becomes a measure of important operational performance and, through it, users express their subjective perception about the services offered. Those who abandon the queue state that it is not worth waiting for the services offered. On the other hand, users who do not leave the system only express their perception of the services offered, if they are asked after the end of the service.

The abandonment phenomenon can be caused by incorrect sizing of the number of agents, which increases queueing time. Therefore, ignoring the abandonment leads to distortions in the information system, which are important for managers to make decisions.

Thus, abandonment, measured by user patience, can be considered as one of the most important, if not most important, operational measures to evaluate the performance of a Call Center. Certainly, if there are shorter waiting times, fewer customers abandon the queue due to being impatient, and overall, this shorter wait means better service. Abandonment differs from other performance measures because customers inform the system of their subjective perception of the service offered, or the one they are receiving. Other measures such as the rate of return after abandonment, or after being served, do not provide this information. The same is true of measures most commonly used to assess performance, such as waiting times, that are objective at the system level but do not report customer service experiences.

Some attempts have been made to adjust the behavior of abandonment by different classical models of probability distribution (Bacelli & Hebuterne, 1981; Mandelbaum & Zeltyn, 2005, 2012). However, due to the different characteristics in the user population, it may be interesting to use mixed probability distributions to better absorb these population nuances. According to Forbes et al. (2011), mixed distributions (or mixtures) are understood as a probability distribution that is a combination of two or more specific probability distributions, such as a normal distribution and an exponential distribution. These distributions are called mixed components.

These mixed models can represent the presence of subpopulations within a general population (Mandelbaum & Zeltyn, 2012; Oliveira, 2009). In this study, subpopulations are represented by groups of customers who have their own standard for patience time, forming a population with a mixture of these groups with different characteristics. These considerations justify the concern to consider abandonment in the mathematical modeling of the Call Centers and to analyze the standards of customers' abandonment time and the implication in the Call Center system's performance measures.

The operations and performance management of a Call Center are based on scientific principles. It is possible to measure the quality of services by making use of stochastic models in the queueing theory. In a queueing model of a Call Center, users are the people who call to obtain information or solve a particular problem, the servers are the agents of the system that serve the users, or they are the communication equipment and the queues are the users waiting to be served.

A widely used queueing model in Call Centers to quantify the level of operational service in terms of some performance measures and congestion is the well-known M/M/c model. In this model, the users' arrival process into the system is considered as a Poisson process (indicated by the first M in the notation) and the length of the service time is supposed to be distributed exponentially (shown by the second M in the notation), with c identical servers operating in parallel (c in the notation), serving a single user queue. This queueing system does not recognize the behavior of user abandonment, as well as the various classes of users, and is restricted to absorbing the characteristics of the arrival process of Call Center users (Whitt, 1993).

In order to overcome this shortcoming, several models of queues incorporating abandonment have been studied in the literature (Bacelli & Hebuterne, 1981; Mandelbaum et al., 2001; Brown et al., 2002; Mandelbaum & Zeltyn, 2005). These

models consider a single family class of parametric distributions to model abandonment; the most common being exponential distribution, Weibull and Erlang. However, it is common to find a heterogeneous customer population in Call Centers, usually consisting of a mixture of groups with different characteristics. Each of them has their own standard for the patience time, some being more patient than others. For this reason, it is generally not reasonable to model abandonment through a single distribution class, but to adjust it to a mixed probability distribution (mix of distributions).

Based on these considerations, investigating and using mixed probability distributions that best adhere to abandonment times is an interesting research topic, for example, to improve the performance of $M/M/c + G$ queueing models used in Call Center sizing. In the $M/M/c+G$ queueing notation, the first three letters have the same meaning as the $M/M/c$ model and G indicates that customer patience times obey any or a generic distribution. Following the same notation, in particular in the $M/M/c+M$ queueing model, the first three letters also have the same meaning as the $M/M/c$ and the third M indicates that user patience times obey an Exponential distribution.

In this article, a research question to be investigated is: are analytical queueing models $M/M/c+G$ and $M/M/c+M$, incorporating abandonment and based on generic (particularly mixed) or Exponential distributions to model abandonment, effective in representing the congestion problem in Call Centers?

Other issues are also considered in this research problem: do the $M/M/c+G$ and $M/M/c+M$ queueing models grasp the reality of Call Centers well? Are models that use mixed distributions to represent abandonment better than models that use a single parametric distribution class to model abandonment? Is there a model that behaves better than any other in all measures, using or not mixed distributions to model abandonment?

It can be observed that some mixtures of distributions were used in literature studies to empirically adjust the patience time in Call Centers, as in Mandelbaum & Zeltyn (2012). However, they were not inserted into analytical queueing models to observe their impact on performance measures. In Oliveira (2009), the relationship between the average waiting time in a queue and the probability of user abandonment for mixed patience distributions in the $M/M/c+G$ queueing model applied to the Call Center was analyzed. By using simulation and mixed patience distribution, it was shown that the mixed exponential and the mixed uniform had shorter waiting times in the queue, with simulated Call Center data, when compared to the exponential and uniform distributions of the same parameters.

This article does not particularly deal with formally developing new mathematical models of congestion in stochastic systems, which are already relatively well developed by the queueing theory. The main concern is to study this subject in depth and to make an empirical complement to this theory, an area reserved for the science of queue congestion. We intend to compare the theoretical models of queues $M/M/c+M$ and $M/M/c+G$, the latter with mixed distribution representing patience via scientific analysis based on actual data, using a case studied at a Bank Call Center.

We also used an experimental model of discrete simulation, which appropriately represents the analyzed Call Center system to compare and verify the analytical models used and explore together with the analytical models, an alternative scenario considering possible increases in demand. The performance measures obtained by the simulation and via analytical queueing models are compared with the purpose of also evaluating the sensitivity of the models. These models and results may be useful

for managers to better configure their systems and manage their operations, as well as more effectively plan the future.

The aim of this article is to show that analytical queueing models $M/M/c+G$ with abandonment, with patience time represented by generic (particularly mixed) distributions, are more effective than analytical queueing models $M/M/c+M$ with abandonment, with Exponential patience, commonly used in practice to evaluate congestion problems in Call Centers and support sizing and operational decisions in these systems.

The relevance is to apply analytical queueing models $M/M/c+G$ considering mixed probability distributions to adjust user patience time in Call Center queues, incorporating these distributions into the abandoned queue models. It is shown that the results obtained from these models with those extracted empirically result in good approximations of reality.

This type of analysis has not been well investigated in the literature and it is understood that knowledge can be gained by evaluating and validating queueing models with abandonment for application in Call Centers. Generally, abandonment in the existing literature has been modeled through a single class of probability distribution. It has not been discussed widely from the point of view of mixed probability distributions, usually more adequate to understand the heterogeneity of each customer's patience time.

This article is structured as follows. In Section 2, a literature review is presented on the $M/M/c+M$ and $M/M/c+G$ queueing models for the Call Center. In Section 3, the methods and techniques used to develop this study are presented. In Section 4, we describe the case of the Bank Call Center, presenting in detail the system, statistical analyses, arrival process analysis, service and abandonment of the data collected. Section 5 presents the results obtained. Section 6 presents the alternative scenario assessment. Finally, in Section 7, the conclusions of the study are drawn.

2 Modeling the Call Center as a Queueing System

Some analytical queueing models are reviewed in this section because of their importance in modeling Call Center systems. We review the queueing models with user abandonment, as well as those that consider exponential and non-exponential distribution (generic) for patience times.

Concerned about developing models that consider the effects of users leaving the system, Garnett et al. (2002) proposed an abandonment model, in which user patience (or time to abandon) is exponentially distributed and the waiting capacity of the system is unlimited ($M/M/c+M$). Relying on asymptotic behavior, the author deduced approximations for performance measures and proposed rules for designing large Call Centers. Zohar et al. (2002) proposed a model based on the $M/M/c$ model that incorporates the customers' adaptive behavior, based on the premise that the users adapt their patience (modeled by the distribution of the time to abandonment) in relation to their expectation of service, particularly for their anticipated waiting time.

Motivated by the practice of Call Centers, Mandelbaum & Zeltyn (2009) modeled a Call Center by a more general queueing system than the previous one, with a $M/M/c+G$ queue also characterized by Poisson arrivals, exponential service times and c servers, but with user patience time with generic distribution. They determined a team of minimum attendants n that satisfies a given cost that can incorporate the fraction of abandonment, the average waiting time and the probabilities of waiting.

The times up to abandoning are censored data and can be estimated using Survival Analysis (Cox & Oakes, 1990). The hazard rate of patient customers can be estimated using a competitive hazard model, according to Palm (1953), who postulated that the hazard rate of the time willing to wait is proportional to a customer's irritation due to waiting. Along the same lines, Aalen & Gjessing (2001) comment on the dynamic interpretation of the hazard rate, but warns about the possibility that the hazard rate of the population does not necessarily represent that of individuals.

Analyzing the reported studies, it can be seen that there is a growing body of literature concerning user abandonment in Call Center queues. Whenever abandonment was considered, the performance evaluation of the Call Center was improved. However, abandonment behavior was modeled preferentially by the Exponential distribution, or by a general, deterministic or Weibull distribution. Owing to this fact, in this paper, we propose to use other probability distributions to model patience times in Call Center waiting queues, formed by mixture distributions capable of better capturing possible different characteristics in the customer population, such as different types of calls due to the heterogeneity of the customers.

The abandonment process, described through mixed distributions, has been studied in the literature, proving to be advantageous compared with those that were not described by the mixed distributions. For example, Oliveira (2009), using discrete simulation models, showed that a decrease in the average waiting time occurs when patience is modeled by mixed distributions. Mandelbaum & Zeltyn (2012) used several mixed distributions to model the abandonment process and obtain a better fit for patience time, than non-mixed distributions, but without incorporating them into analytical queueing models.

However, to the best of our knowledge there are no studies in the literature that incorporate these mixed distributions into analytical queueing models with abandonment, evaluating the results obtained by these models with the actual data extracted from Call Centers.

The following is a summary of the important results of queueing systems, containing the characteristics and performance measures of the $M/M/c+M$ and $M/M/c+G$ models that incorporate patience time into their equations, and that are addressed in this article.

3 The $M/M/c+M$ system

The $M/M/c+M$ model with abandonment was first idealized by Palm (1957), who introduced a simple way of modeling abandonment. The author suggested improving the Erlang-C model ($M/M/c$) as follows: an exponentially distributed patience time with mean $1/\theta$ is associated with each incoming call. An arriving customer finds a waiting time offered, which is defined as the time that this customer would have to wait, if his/her patience were infinite. If this offered waiting time exceeds the user patience time, the customer abandons the system. Otherwise, the customer waits until being served. The parameter of patience θ is called the individual abandonment rate. The model is called Erlang-A (A for abandonment) and also because the model is between the Erlang-C and Erlang-B models.

Customers arrive at the $M/M/c + M$ system by a Poisson arrival process with arrival rate λ and times between arrivals exponentially distributed. They have a patience time τ assumed to have an exponential distribution with rate θ , their

individual abandonment rate. They are served by c channels of statistically identical services, operating in parallel and independently of each other, serving a single queue. Service time is considered exponentially distributed with rate μ . The service, arrival and patience processes are considered mutually independent.

In this system, for a given customer, the patience time of τ is defined as the time that a customer can stand waiting to be served and it is considered that a wait that reaches τ results in abandonment. The waiting time offered by the system is also defined, V , as the time that a customer with infinite patience waits until being served. Thus, the waiting time of an impatient customer, or the waiting time in the queue until starting the service or abandoning it, whichever occurs first, is equal to $\min\{V, \tau\}$, where V is the waiting time in the queue until starting the service, if abandonment does not occur beforehand.

Some performance measures of the M/M/c+M queueing model with abandonment are deduced by Mandelbaum & Zeltyn (2005), given by:

Waiting probability in the system: this represents the fraction of customers waiting in the queue and its expression is:

$$P\{W > 0\} = \sum_{j=c}^{\infty} \pi_j = \frac{A(c\mu/\theta, \lambda/\theta)E_{1,c}}{1 + [A(c\mu/\theta, \lambda/\mu) - 1]E_{1,c}} \tag{1}$$

where:

$$A(x, y) = \frac{xe^y}{y^x} \gamma(x, y) = 1 + \sum_{n=1}^{\infty} \frac{y^n}{\prod_{k=1}^n (x+k)}, \quad x > 0, y > 0 \tag{2}$$

and

$$E_{1,c} = \frac{\rho \cdot E_{1,c-1}}{1 + \rho \cdot E_{1,c-1}}, E_{1,0} = 1, \rho = \lambda / c\mu \quad c \geq 1 \tag{3}$$

Probability of abandonment due to impatience: the fraction of abandonment $P\{Ab\}$ can be understood simply as the product between the probability of abandoning due to impatience $P\{Ab | W > 0\}$ and the probability of waiting in the system $P\{W > 0\}$, i.e.:

$$P\{Ab\} = P\{Ab | W > 0\} \times P\{W > 0\} \tag{4}$$

Therefore, it follows that:

$$P\{Ab | W > 0\} = \sum_{n=c}^{\infty} \pi_n P_{n-c}\{Ab\} / P\{W > 0\} = \frac{1}{\rho A(c\mu/\theta, \lambda/\mu)} + 1 - \frac{1}{\rho} \tag{5}$$

Average waiting time of customers who are waiting: exploring the linear relationship between the fraction of abandonment $P\{Ab\}$ and the average waiting

time $E\{W\}$ that occurs not only in the Erlang-A system, but in other models with exponential patience and expressed by $P\{Ab\} = \theta.E(W)$, we find:

$$E[W | W > 0] = \frac{1}{\theta} \left[\frac{1}{\rho_A(c\mu/\theta, \lambda/\theta)} + 1 - \frac{1}{\rho} \right] \tag{6}$$

More details on this queueing system, which incorporates abandonment, are found in Mandelbaum & Zeltyn (2005) and also in Ferrari (2016).

4 The M/M/c+G system

The M/M/c + G abandonment system has a Poisson arrival process with an arrival rate λ and times between exponentially distributed arrivals, as well as the previous system. However, it has a patience time τ assumed to have a generic patience distribution G . It is served by c identical service channels, operating in parallel and independently of each other, serving a single queue. The service time is considered exponentially distributed with rate μ . The service, arrival and patience processes are mutually independent.

It is also assumed that \bar{G} is the function of survival time of patience τ , i.e., $\bar{G}(x) = 1 - G(x) = P\{\tau > x\}, x \geq 0$. As the system is in statistical equilibrium, it is assumed that the arriving customer finds a virtual waiting time V in the queue. This time can be understood to be the time that a customer with infinite patience would have to wait in the queue. Thus, the time that the customer actually waits in the queue is $\min(V, \tau)$.

In this system, Bacelli & Hebuterne (1981) state that arriving customers can calculate the virtual waiting time (or waiting time offered) V at the instant of their arrival, and if this time is greater than their patience time, the customer leaves the system immediately and does not join the queue.

Thus, they consider the Markov process $\{(N(t), \eta(t)), t \geq 0\}$, where $N(t)$ is the number of agents occupied at instant t and $\eta(t)$ is the virtual waiting time of the customer that arrived at instant t .

The exact formulas of the M/M/c+G system performance measures, based on Bacelli & Hebuterne (1981) and presented in Zeltyn (2004), are reproduced below.

Fraction of waiting time in the system:

$$P\{W > 0\} = \frac{\lambda J}{\varepsilon + \lambda J} \bar{G}(0) \tag{7}$$

Probability of abandonment:

$$P\{Ab\} = \frac{1 + (\lambda - c\mu)J}{\varepsilon + \lambda J} \tag{8}$$

Average waiting time in the system:

$$E(W) = \frac{\lambda J_H}{\varepsilon + \lambda J} \tag{9}$$

with:

$$J = \int_0^{\infty} \exp\{\lambda H(x) - c\mu x\} dx \quad (10)$$

$$J_H = \int_0^{\infty} H(x) \exp\{\lambda H(x) - c\mu x\} dx \quad (11)$$

$$H(x) = \int_0^x \bar{G}(u) du \quad (12)$$

$$\varepsilon = \frac{\sum_{n=0}^{c-1} \frac{1}{n!} \left(\frac{\lambda}{\mu}\right)^n}{(c-1)! \left(\frac{\lambda}{\mu}\right)^{c-1}} = \int_0^{\infty} e^{-t} (1 + t\mu/\lambda)^{c-1} dt \quad (13)$$

Other performance measures and other information about the M/M/c+G system can be obtained from Zeltyn (2004) and also Ferrari (2016).

Based on this brief literature review on the use of queueing models in Call Centers, it can be observed that there are few applications that validate the theoretical results of queues through scientific analysis based on data, especially when user patience and the abandonment phenomenon are considered.

To the best of our knowledge, there are not many studies in the literature to verify whether an existing model, using generic distributions and, particularly, mixed distributions to represent patience, is suitable for a particular application. There is great concern about the theoretical development of Call Center queueing models and the use of discrete simulation to prove the suitability of these models.

While developing this research, we used the M/M/c+M and M/M/c+G queueing models to represent the case study of the Bank Call Center (Section 4). The decision for these queueing models can be accounted for by the best fit of non-exponential distributions to represent them and the patience times, obtained using the statistical analysis applied to the data of the object of study considered.

5 Research approach

In this study, we adopted the inductivist conception, based on observing the facts that occurred in the interest phenomenon, which is the customers' patience time in the Call Center queue in the case study proposed (Demo, 2000). The quantitative approach was used, justified by the fact of having a well-defined variable, such as the patience time in the Call Center queue, and making use of mathematical models in the queueing theory. Measurability can be found using data collection in the case studied, the causality in seeking to explain the behavior of patience, generalization by adapting and applying existing queueing models and replication allowing for the opportunity to reproduce the research. Modeling was the method used in the

quantitative approach, because in the case studied (the Call Center) there is a problem to be analyzed, represented by user abandonment in a queue.

The quantitative empirical descriptive research (Bertrand & Fransoo, 2002; Morabito & Pureza, 2010) was the research technique used to analyze two situations not yet explored in the literature: using queueing models with generic distributions (particularly mixed) representing patience and validation of these queueing models analyzing them based on data taken from Call Centers.

The empirical quantitative approach is justified by the studies and analyses carried out in the Call Center, observing that the patience distributions are generic, with a strong potential for mixed distribution applications.

In addition, the analytical queueing models found in the literature are adapted to this reality, based on the observations and data of this case studied, and used to analyze this Call Center (Figure 1).

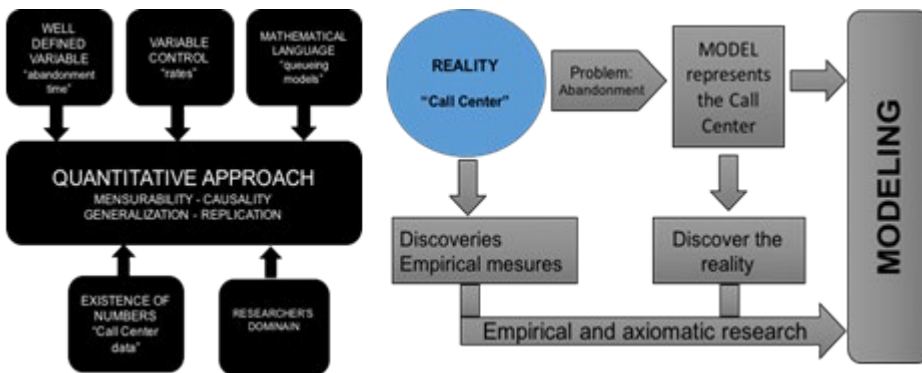


Figure 1. Methods and techniques.

A sample was drawn from the Bank Call Center and different parameters of interest were obtained from this sample, such as arrival, service and abandonment rates and the number of attendants, among others. Some performance measures were also calculated from this sample, such as average waiting times, probabilities of waiting, probabilities of abandonment and traffic intensities. Based on the sample data, the distributions of the arrival, patience and service processes were adjusted using generic (particularly mixed) distributions to represent patience. Data analysis was performed using descriptive statistics, by adherence tests, such as the Kolmogorov-Smirnov test and others, to adjust the probability distributions to the data observed in the studied case, as well as the Kaplan-Meier test, applied to censored data to infer the average patience time. These parameters and obtained distributions were inserted into the analytical queueing models and also into the simulation model considered in this study. The same performance measures were calculated using the analytical and simulation models. Comparing these performance measures of the models (analytical and simulation), the verification was obtained. The analytical queueing and simulation models were validated by comparing their performance measures with those drawn from the sample (Figure 2). By analyzing and comparing the performance measures obtained by the analytical queueing models with the same measurements taken from the actual data sampled, we intend to find out which queueing model is more appropriate to appropriately represent the Call Center.

An alternative scenario was also studied and evaluated in order to analyze the correctness and sensitivity of the analytical queueing models considering changes

caused by demand. The demand for calls (arrival rate) varied and these new rates were inserted into the analytical queueing models and the corresponding simulation model. New values of performance measures (average waiting time, probability of waiting, probability of abandonment and traffic intensity) were calculated by the models considered and compared to each other to validate the analytical queueing model considering this new situation. The simulation was also used to analyze and validate analytical queueing models with generic (particularly mixed) distributions for patience. The diagram in Figure 2 summarizes these discussions.

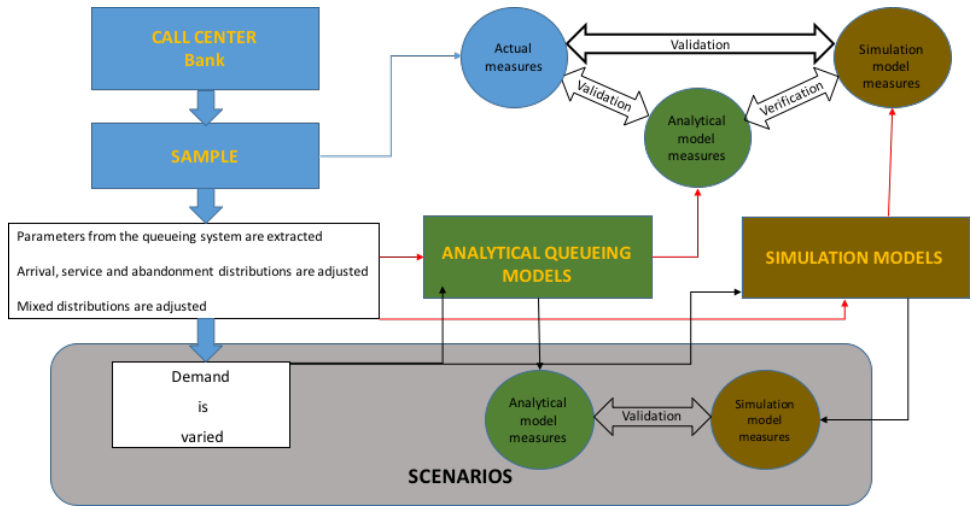


Figure 2. Work structure.

6 The Bank Call Center

The Bank Call Center serves customers from an Israeli bank, which offers various types of services, such as information for current and future customers, current account transactions and savings accounts, stock trading and technical support for Internet users on the bank's website.

A customer can contact the Call Center by telephone, connecting to an Interactive Voice Response (IVR). At this point, the customer identifies him/herself and chooses one of the options available to reach the service he/she wants. From the IVR, the customer can leave the system without solving his/her problem or solving his/her problem and abandoning the system or, even, connecting with an attendant, if there are any available. If the customer does not find any agents available, the customer waits in the queue to be answered. At this stage, he/she may leave the system after a long wait. Those who are served can complete their service after talking with the attendant, or they can be disconnected by the attendant.

This Call Center has eight regular positions for agents and one supervisor doing shiftwork, working from Sunday to Thursday from 7 am to midnight, closing at 2 pm on Friday and reopening at 8 am on Saturday. A basic representation of the call flow in this Call Center can be seen in Figure 3, whose values show the monthly calls.

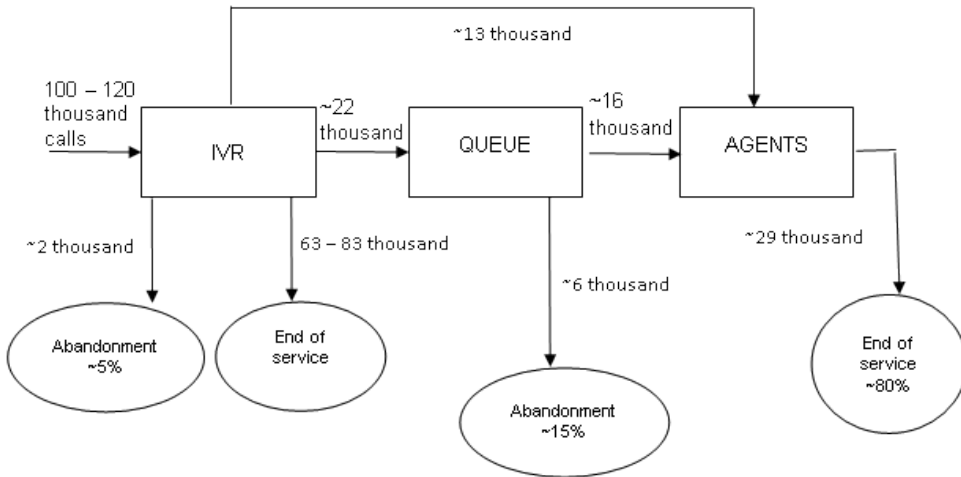


Figure 3. Call flows in the Bank Call Center. Source: Mandelbaum et al. (2001).

In this study, the arrival, abandonment and service processes are analyzed from the instant the user joins the queue, after being screened in the IVR. Therefore, the arrival, abandonment and service processes completed in the IVR are disregarded.

7 Data description

The data source (Call Center Data) (Technion – Israel Institute of Technology, 2002) used in this study is the same kindly provided and also used by Mandelbaum et al. (2001) and Brown et al. (2002). The collected data were organized into a spreadsheet, where each line corresponds to the record of each call, containing its identification, user information and call flow information.

A detailed description of the information contained in each record can be obtained from Mandelbaum et al. (2001). The intention is to initially reproduce the results of these authors using the M/M/c+M analytical queueing model and then also generate new results using the M/M/c+G analytical queueing model and compare them. These data correspond to all the calls answered by the Bank Call Center during the 12 months from January to December, 1999.

There are 20,000 to 30,000 calls per month. We considered the information only on November 9, 1999, considered by Mandelbaum et al. (2001) and Brown et al. (2002) on a typical Bank Call Center day from 12pm to 1pm, classified as high congestion ($\rho = 0.8726$). This same sample of actual data from Mandelbaum et al. (2001) and Brown et al. (2002) was also used to generate the results and comparisons made in this study.

8 Arrival process analysis

The arrival process records the instants when incoming phone calls proceed to the queue after they have visited the IVR. When they leave the IVR, users join the queue and if an attendant happens to be available, the service starts immediately and their queueing time is zero. On the other hand, if there are no available attendants, the user waits for a positive amount of time in the queue, until starting their service, or loses their patience and abandons the Call Center.

This process can be described in two different levels of detail, considering the number of calls per unit of time, or the time between the consecutive arrivals of the incoming calls in the queue.

This study is based on the time between incoming calls, which is evaluated by the descriptive statistics information, as well as its stochastic variability. The statistics of the times between the incoming calls at the Bank Call Center in the time interval from 12pm to 1pm are summarized in Table 1.

9 Service process analysis

Service by the agent is the last step of a user's visit to the Call Center. Therefore, the knowledge of this service reveals information about the level of service being provided. The service can be evaluated quantitatively, through statistical information about the length of the service, or qualitatively, when the user reveals his/her satisfaction about the service received, answering a short questionnaire before leaving the system.

The concern in this study is about the quantitative evaluation of the service provided by the Bank Call Center. Considering the same time interval analyzed previously, from 12pm to 1pm, a statistical analysis of the length of the service times (service time) was made, calculating its mean, median and standard deviation. Table 2 summarizes these service times statistics. Service time records equal to zero were ignored.

Table 1. Time between arrivals (min).

Sample data	12-1pm
Mean	0.577
Median	0.400
Standard deviation	0.495
Coef. of Variation	0.858
Arrival rate (call/min)	1.717

Source: prepared by the author.

Table 2. Service time (min).

Sample data	12-1pm
Mean	3.021
Median	2.050
Standard deviation	3.339
Coef. of Variation	1.105
Service rate (call/min)	0.328

Source: prepared by the author.

10 Abandonment process analysis

This process takes place when users join the queue, wait for a positive amount of time, lose patience and abandon the system. This user attitude shows to Call Center managers if it is worth waiting for the service provided. This level of service offered by the system can be evaluated by the number of abandonments that occur in a given time interval.

An increase in the number of abandonments can indicate a long waiting time and, consequently, a low level of service. Therefore, abandonment behavior and waiting time are closely related, as all users who abandoned the queue had previously been waiting.

Furthermore, the time it takes for a customer to be served is not observed if he/she abandons the queue. Therefore, this time needed to be served and patience time are based on censored data.

To analyze these times better, Brown et al. (2002) makes three distinctions. The first is the difference between the queueing time and the waiting time, neglecting the zero wait. It considers the waiting time (i.e., only with positive waiting times) to be more relevant for Call Center managers, especially when analyzed together with the fraction of those who wait. The second is the distinction between the user waiting times that were served in the system and those who abandoned the system. Brown et al. (2002) state that only the time of those who have abandoned the queue (patience time) does not reflect the patience of all users. The third is the distinction between the time a user needs to wait before being served by an agent (virtual waiting time) and the time a user is willing to wait before abandoning the system (patience).

The virtual waiting time can be understood as the amount of time that a (virtual) user who has patience has to wait until being served, while the patience time refers to the user's operational measure of patience (or impatience). However, none of these measures can be directly observed and should be estimated.

The characterization of patience and virtual waiting time is based on censored data according to Mandelbaum et al. (2001). They consider that for those users who are patient enough to wait for an agent, their waiting times are a sample of the time needed to wait (virtual waiting time). In this case, they affirm that user patience (time willing to wait) is censored by the waiting time for the service (virtual waiting time), in which only the waiting time for the service is observable. On the other hand, for the users who abandon the queue, the patience time (time willing to wait) censors the time needed to wait (virtual waiting time) and, in this case, the patience time is observable.

In order to estimate the times that are censored and, therefore, not observable as user patience, we use the Kaplan-Meier estimator, which is a Survival Analysis tool, a branch of Statistics that is concerned with analyzing censored data (Kalbfleisch & Prentice, 1980). This section analyzes experiences that users had in the queue until abandoning the Bank Call Center system, taking into account the same time interval from 12 to 1pm analyzed earlier. Statistical analysis of the telephone call patience times was carried out by calculating their mean, median and standard deviation, which are summarized in Table 3.

As the interest is to estimate user patience in a Call Center, i.e. how long a user is willing to wait before abandoning the queue, then one should consider the time it takes for a user to reach an agent as a censored observation. In fact, if he/she reaches the agent, it is because the time he/she wanted to wait was longer. In order to estimate user patience in the interval from 12 to 1pm, waiting times in the queue, telephone calls seeking services, abandoning the system or being attended to by an agent were considered. The calls that abandoned the IVR were not considered. Concerning the calls that reached the agent as censored observations, the user patience was estimated, adopting the Kaplan-Meier estimator, using SPSS® software (IBM Corporation, 2011).

Figure 4 shows the Kaplan-Meier estimates of the user patience survival function (time that he/she is willing to wait before he/she leaves) from the time interval from 12pm to 1pm. Table 4 reports the means and standard deviations of the Kaplan-Meier estimate for the distribution of user patience at each hour interval. These results were obtained by conventional SPSS® software procedures when estimating the survival function.

Table 3. Patience time (min).

Measures	12-1pm
Mean (sec)	404.520
Standard deviation (sec)	16.308
Mean (min)	6.742
Abandonment rate (call/min)	0.148

Source: prepared by the author.

Table 4. Patience time estimated by Kaplan-Meier.

Sample data	12-1pm
Mean (sec)	68.390
Median (sec)	45.500
Standard deviation (sec)	77.565
Coef. of Variation	1.134
Mean (min)	1.139
Abandonment rate (call/min)	0.877

Source: prepared by the author.

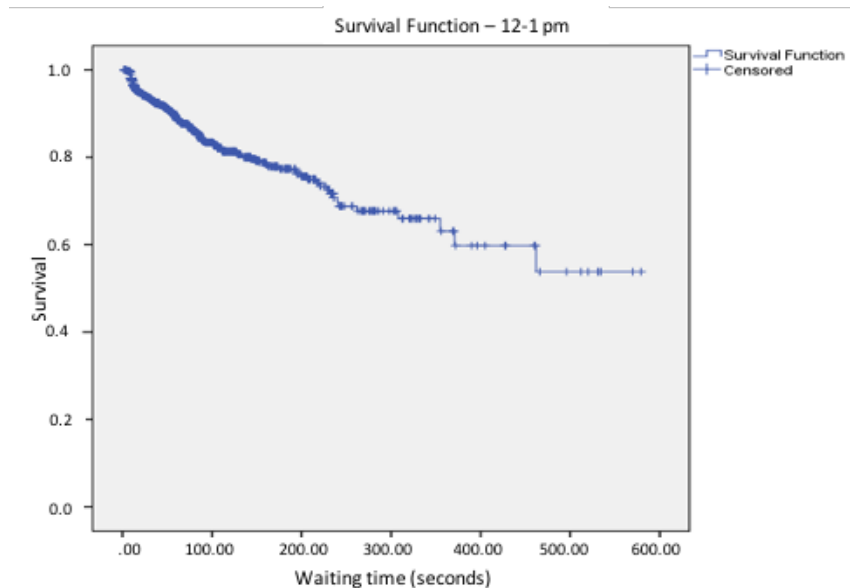


Figure 4. Survival curve. Source: prepared by the author.

The rates of arrival, service and abandonment the processes were also taken from the sample considered from the Bank Call Center, in the same time interval from 12 to 1pm. The performance measures of the average waiting time, probability of

waiting, probability of abandonment and intensity of traffic, reproduced in Table 5, were also obtained.

Table 5. Empirical measures Bank Call Center.

Sample data	12-1pm
Arrival rate (call/min)	1.717
Service rate (call/min)	0.328
Abandonment rate (call/min)	0.148
Average waiting time (min)	0.813
Probability of waiting	58.3%
Probability of abandonment	10.2%
Traffic intensity	82.8%

Source: prepared by the author.

A theoretical probability distribution was also adjusted for user patience times, from 12 to 1pm, in the Bank Call Center. In this analysis, we used the EasyFit® software, version 5.5 (Mathwave Technologies, 2010), which orders the most adherent distributions to the data according to the Kolmogorov-Smirnov test. The results of this analysis are shown in Table 6, with the respective p-value.

Table 6. Patience distributions – Bank Call Center.

Process	Adjusted distribution	p-value 12-1pm
Abandonment	Fatigue Life	0.11255
	Lognormal	0.04254
	Exponential	0.02587

Source: prepared by the author.

11 Results analysis

The M/M/c+M analytical queueing model with Exponential distribution to represent the patience times was used by Mandelbaum et al. (2001) and Brown et al. (2002) in the analysis of the Bank Call Center case. However, they could have used the M/M/c+G analytical queueing model, which allows generic distribution to represent abandonment, to try to obtain better results.

To meet this objective, the actual data of the 12-1pm interval collected in the Bank Call Center was used as the sample. These are the same sample data used by Mandelbaum et al. (2001) and Brown et al. (2002). As discussed in the previous section, these data refer to a relatively long interval (1 hour) of the Bank operation, with intense traffic (82.8%) and no capacity changes in the system; therefore, it was reasonable to assume that the system is in statistical equilibrium and, for the same reason, also to assume that arrival and service rates remain constant, or with a slight variation. The queueing models used here, as well as those of Mandelbaum et al. (2001) and Brown et al. (2002), require these rates to be constant.

The Kolmogorov-Smirnov adhesion test was also used with the EasyFit® software, version 5.5 (Mathwave Technologies, 2010), for the user patience probability distributions (Table 6).

Among the possible distributions, Lognormal and Exponential were chosen to represent the user patience in the theoretical queueing models, giving rise to various

different possibilities of models to represent the Bank Call Center. Analyses with other schedules of this same system were also made, and the results obtained were similar to those presented here and are described in detail in Ferrari (2016).

Mixed distributions were considered when developing these analytical models to model the abandonment process. In the sample data from 12 to 1pm, the mixture of two components was combined, two Lognormal (LogN+LogN) and two Exponentials (Ex+Ex). Although there are methods to determine the amount of components of the mixed distributions, in this study, we chose to use two components only as an initial approach to the research problem.

The parameters of these mixed distributions, as well as the weight of each one of them, were obtained by the maximum likelihood estimation method, calculated using Mathematica software (Wolfram Research, Inc., 2013). These parameters were recalculated using the mean and variance values estimated by the Kaplan-Meier estimator. These estimated values were inserted into the mean and variance expressions of the mixed distribution components to obtain new parameters. This procedure was necessary because the patience times are censored. These mixed distributions with these new parameters were used in the analytical queueing models considered in this study to obtain the performance measures and subsequent analysis of the results.

In these applied queueing models, the arrival, service and abandonment rates and distributions were entered and the following performance measures were obtained by their respective computer algorithms programmed in Mathematica software (Wolfram Research, Inc., 2013): average waiting time, probability of waiting, probability of abandonment and traffic intensity. These performance measures calculated by the queueing models were compared with the same measurements obtained directly from the actual data, described in the statistical analysis of the arrival, abandonment and service processes, obtaining significant differences to validate the analytical queueing models.

In order to understand the analyses more clearly, from this section onwards, the queueing models are represented with a specific notation. We identified the arrival and service process distributions and the generic ones in the abandonment process with a numerical index to distinguish one from the others.

When the distribution is mixed, the letter “*m*” was added next to the numerical index to represent it. For example, G_1 identifies a single parametric distribution ($G_1 \sim \text{LogN}(1,2)$, a Lognormal distribution with parameters 1 and 2) and G_{1m} represents a mixed distribution ($G_{1m} \sim 0.5\text{LogN}(1,2) + 0.5\text{LogN}(3,4)$, a mixed distribution with two Lognormal components and weight $p = 0.5$).

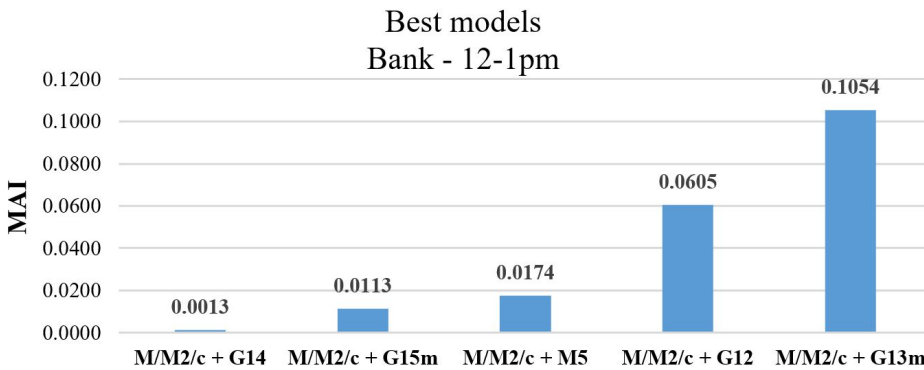
In these comparisons, an analytical queueing model was considered effective to represent, from 12 to 1pm, the Bank Call Center. For this purpose, the deviation between the performance measure values taken from the analytical queueing models and the actual data was calculated. Next, the Model Accuracy Index (MAI) of these deviations was obtained and the least efficient MAI analytical queueing model was considered the most efficient to represent the Bank Call Center because the results obtained from the performance measures of this model have smaller deviations from the same measurements taken from actual data.

The MAI statistic is the sum of the squares of the differences between the estimated value (\hat{y}_i) and the actual value of the data (y_i), weighted by number of terms (n) and weight α_i (Equation 15).

The weights are assigned according to the importance of the performance measure that it represents. In this study, weights were assigned to 1 ($\alpha_i=1$) for all the deviations obtained, considering that the performance measures have the same level of importance. The model with the lowest MAI was considered the best.

$$MAI = \frac{\sum_{i=1}^n \alpha_i (y_i - \hat{y}_i)^2}{n} \tag{15}$$

Graph 1 shows the MAI of the analytical queueing models used from 12 to 1pm considered in this study.



Distributions:

M ~ Exponential (1.717) M₂ ~ Exponential (0.328) M₅ ~ Exponential (0.148)
 G₁₂ ~ Exponential (0.148) G₁₄ ~ Lognormal (1.7902; 0.4861)
 G_{13m} ~ 0.4785×Exponential (0.3099) + 0.5215×Exponential (0.2844)
 G_{15m} ~ 0.95×Lognormal (1.7334; 0.4973) + 0.05×Lognormal (-2.0101; 1.3585)

Graph 1. Comparison of the M/M/c+M and M/M/c+G model performance.

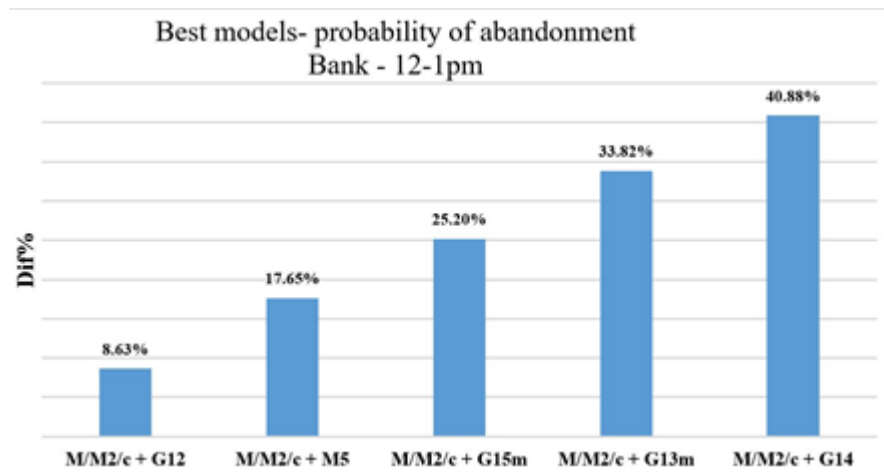
According to Graph 1, the most efficient analytical queueing model to represent the Bank Call Center in the 12-1pm time frame was the M/M₂/c+G₁₄ analytical queueing model, with Exponential service and patience modeled with a single Lognormal distribution of probability. The M/M₂/c+G_{15m} analytical queueing model with Exponential service and patience modeled with a mixed distribution formed by two Lognormal components is also appropriate to represent the Bank Call Center. According to Graph 1, these two models with generic distribution representing the user patience times are better suited to represent the Bank Call Center than the Exponential distribution analytical model (M/M/c+M) representing the user patience times, analyzed by Mandelbaum & Zeltyn (2005). In these cases, the mixed distributions were efficient in terms of representing the patience times.

It was observed in the analyses carried out that there are analytical queueing models that produce better results in some performance measures than in others. This fact motivated us to find the most appropriate analytical queueing model to represent the Call Center, for each of the performance measures, from 12 to 1pm. In this analysis, we considered the “Percentage Difference” metric (*Dif%*) defined by Equation 16, where *V* is the value of the performance measure obtained by the analytical queueing model and *R* is the value of the actual performance measure:

$$\text{Dif\%} = \frac{|V - R|}{R} \quad (16)$$

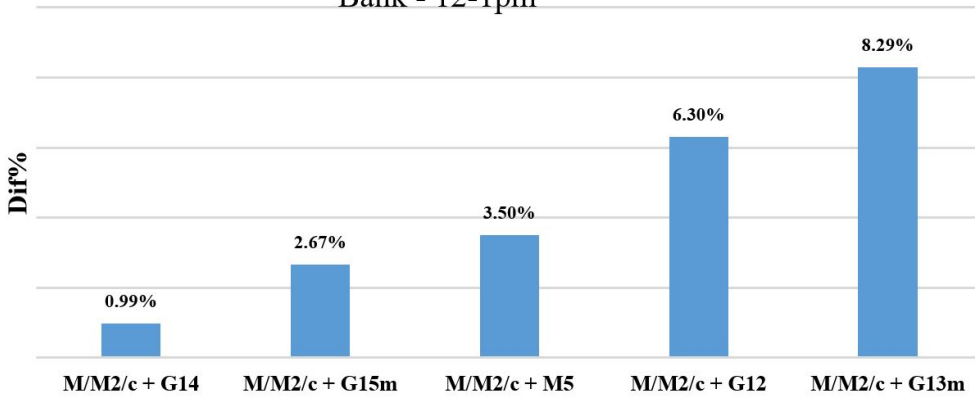
The percentage differences of these performance measures are shown in Graphs 2 to 5, at the time considered in this study. Observing these graphs, it can be observed that the models that have generic distribution for the patience times (M/M/c+G) are more accurate than the model with Exponential distribution to represent the patience times (M/M/c+M) in the four performance measures analyzed. It can also be seen that the models that have mixed distribution to represent the patience times were not the best, but were always competitive, except for the probability of abandoning. Nevertheless, these mixed distribution models for the patience times presented better results, in three performance measures, than the M/M/c+M model, analyzed by Mandelbaum & Zeltyn (2005), with a single Exponential parametric distribution class to represent the patience times.

To sum up, considering the Bank Call Center system, from 12-1pm, the analytical queueing models with generic distributions (mixed or non-mixed) to represent the patience times (M/M/c+G) are more effective than the M/M/c+M analytical queueing model proposed by Mandelbaum & Zeltyn (2005) with a single Exponential parametric distribution class.



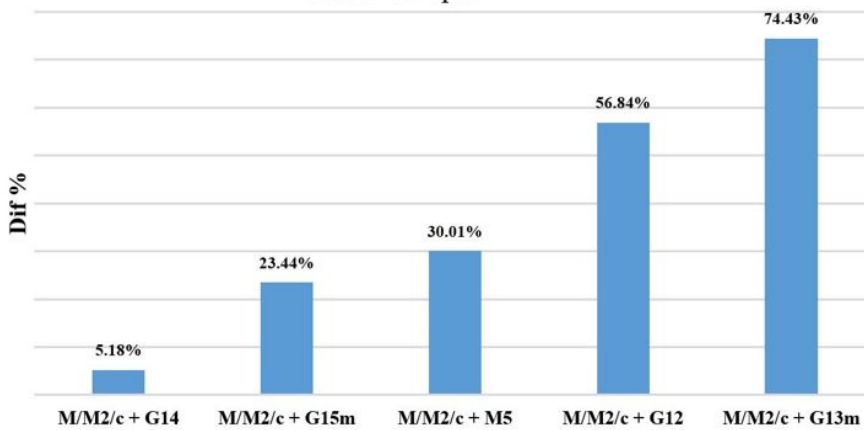
Graph 2. Performance of M/M/c+M and M/M/c+G models – probability of abandonment.

Best models - traffic intensity
Bank - 12-1pm



Graph 3. Performance of M/M/c+M and M/M/c+G models– traffic intensity.

Best models- average waiting time
Bank - 12-1pm

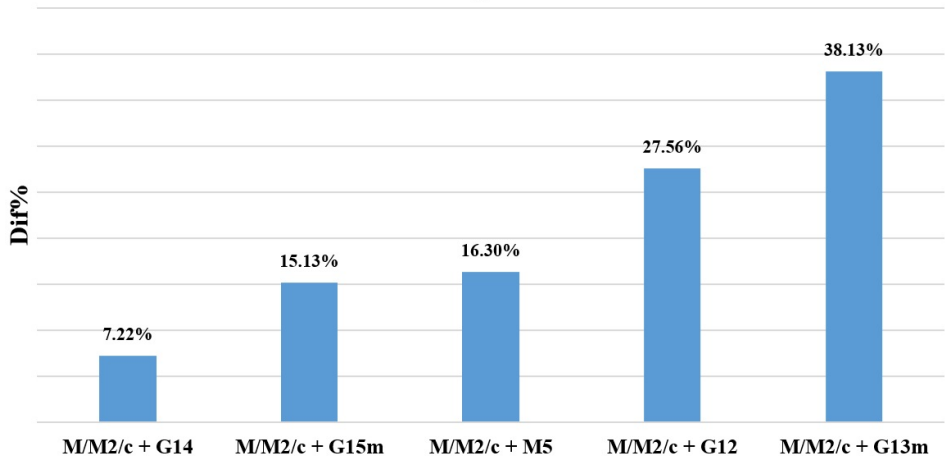


Distributions:

$M \sim \text{Exponential}(1.717)$ $M_2 \sim \text{Exponential}(0.328)$ $M_5 \sim \text{Exponential}(0.148)$
 $G_{12} \sim \text{Exponential}(0.148)$ $G_{14} \sim \text{Lognormal}(1.7902; 0.4861)$
 $G_{13m} \sim 0.4785 \times \text{Exponential}(0.3099) + 0.5215 \times \text{Exponential}(0.2844)$
 $G_{15m} \sim 0.95 \times \text{Lognormal}(1.7334; 0.4973) + 0.05 \times \text{Lognormal}(-2.0101; 1.3585)$

Graph 4. Performance of M/M/c+M and M/M/c+G models– average waiting time.

Best models - probability of waiting Bank - 12-1pm



Distributions:

$M \sim \text{Exponential}(1.717)$ $M_2 \sim \text{Exponential}(0.328)$ $M_5 \sim \text{Exponential}(0.148)$
 $G_{12} \sim \text{Exponential}(0.148)$ $G_{14} \sim \text{Lognormal}(1.7902; 0.4861)$
 $G_{13m} \sim 0.4785 \times \text{Exponential}(0.3099) + 0.5215 \times \text{Exponential}(0.2844)$
 $G_{15m} \sim 0.95 \times \text{Lognormal}(1.7334; 0.4973) + 0.05 \times \text{Lognormal}(-2.0101; 1.3585)$

Graph 5. Performance of M/M/c+M and M/M/c+G models – probability of waiting.

12 Scenario analysis

One of the important applications of the analytical queueing models, which adequately represents a Call Center, is to enable its managers to make decisions based on different scenarios and to make predictions about the operations of these systems. The results obtained from these scenarios are projections made of the performance measures that can be used by managers to analyze Call Center congestion in the future. These projections do not have a corresponding value in the actual data of the Bank Call Center, and it is not possible to verify their accuracy obtained by comparing the results of projections with those observed in reality. This accuracy was verified here by comparing the results of the scenarios with the results obtained by a discrete simulation model that adequately represents the reality of the Bank Call Center, as discussed below.

In this study, a scenario was proposed to know what would happen to the Call Center performance measures if there was an increase in the demand for calls. For this analysis, the arrival rates of 0.5 calls per minute up to 3.0 calls per minute, with a 0.5 call-per-minute interval were substituted in the analytical queueing model and the simulation model to obtain the performance measure values of the average waiting time, probability of waiting, probability of abandonment and traffic intensity.

Once these performance measure values are known, the impact they have on the overall performance of the Call Center can be evaluated and also, until which increase of call demands the Call Center can stand. This scenario was analyzed with the M/M₂/c+G₁₄ analytical queueing model, best suited to represent the Bank Call

Center from 12 to 1pm, with a non-mixed Lognormal probability distribution for the patience times, according to Graph 1.

The simulation model used for comparison in the analysis of this scenario was implemented in the Arena[®] software (Rockwell Automation Technologies, Inc., 2014), through the Arena Contact Center Edition simulation system, specially developed for Call Center managers to construct simulation models and analyze the results produced by these models. This system uses graphical concepts to generate telephone call flows from the time they come in, directing them through the center until they find an attendant. To create a simulation model using this system, it was necessary to describe the sequence of events that occur in the call flow, through the modules: contacts, arrival standard, trunk line capacity, route description and agents.

The queueing problem of this study was simulated with 300 replications, each replication lasting one day, generating at the end of all the rounds an average of 18 thousand calls in the Bank Call Center. Statistics were collected every hour and logged in a report produced by the software. A warm-up period of 144,000 minutes was used that occurred at the 100th replication. From that period onwards, the system began to operate in statistical equilibrium and from then on, the statistics were counted. At the end of the replications, the Arena software generated a report containing various information from each of the rounds, as well as a Siman Summary Report containing the results of all replications. Among these results are the performance measures, such as the average waiting time, the probability of abandonment, the traffic intensity and the probability of waiting, which is calculated by dividing the number of users who were served (excluding those who abandoned the queue) and the number of users created in the simulation. In addition to the outcome of these measures, the half-interval of the confidence interval was also reported. All simulations were performed on a Samsung ultrabook with an Intel[®] Core™ I7 processor, 2.4 GHz and 8.00 GB of RAM, and took an average of 6 minutes to estimate the performance measures for each queueing problem.

Before using the simulation model to analyze the scenario of this section, it was validated by comparing the values obtained by it for each of the performance measures, with the corresponding values extracted from the actual data, from 12 to 1pm, collected from the Bank Call Center. The purpose of this validation was to verify if the simulation model actually represented the reality of the Bank Call Center. In this analysis, we used confidence intervals with $\gamma = 95\%$ of probability, calculated for each of the performance measures.

All the confidence intervals of the simulation model contained the results of the performance measures obtained by the corresponding value extracted from the sample of the actual data from 12 to 1pm (these results are described in Ferrari, 2016). Thus, the simulation model was considered validated and was also used for the scenario analysis to obtain results with values that extrapolate the sample data.

In the 12 to 1pm interval mentioned above, the Bank Call Center operates with a team of six attendants, with an arrival rate of 1,717 calls per minute (actual data). Varying the arrival rate from 0.5 to 3.0 calls per minute, and replacing each of these values in the $M/M_2/c+G_{14}$ analytical queueing model and the simulation model, we obtained the performance measures in Tables 7 and 8.

Table 7. Scenario– Variation of demand – Bank *Call Center* – 12-1pm.

Arrival rate (call/min)	Average waiting time			Probability of waiting		
	Simulation	Confidence Interval	Analytical	Simulation	Confidence Interval	Analytical
0.5	0.0039	(0.0000; 0.0080)	0.0033	0.9998	(0.9977; 1.0019)	0.0056
1.0	0.0854	(0.0702; 0.1006)	0.0915	0.9963	(0.9515; 1.0411)	0.1021
1.5	0.4581	(0.4012; 0.5150)	0.4821	0.9765	(0.9111; 1.0419)	0.3875
2.0	1.4201	(1.3292; 1.5110)	1.2065	0.9098	(0.8131; 1.0065)	0.7201
2.5	2.5984	(2.5085; 2.6883)	1.9127	0.7688	(0.6834; 0.8542)	0.912
3.0	3.4773	(3.4053; 3.5493)	2.3721	0.6654	(0.5240; 0.8068)	0.9781

Source: prepared by the author.

Table 8. Scenario – Demand Variation – Bank *Call Center* – 12-1pm.

Arrival rate (call/min)	Probability of abandonment			Traffic Intensity		
	Simulation	Confidence Interval	Analytical	Simulation	Confidence Interval	Analytical
0.5	0.0002	(0.0000; 0.0023)	0.00008	0.2530	(0.1451; 0.3609)	0.2541
1.0	0.0037	(0.0000; 0.0485)	0.0037	0.5026	(0.3620; 0.6432)	0.5062
1.5	0.0235	(0.0000; 0.089)	0.0318	0.7403	(0.6700; 0.8106)	0.7379
2.0	0.0902	(0.0000; 0.1869)	0.1143	0.9166	(0.7821; 1.0511)	0.9001
2.5	0.2310	(0.1456; 0.3164)	0.2345	0.9993	(0.9189; 1.0797)	0.9724
3.0	0.3346	(0.1932; 0.476)	0.3481	1.0189	(0.9740; 1.0638)	0.9938

Source: prepared by the author.

Observing Tables 7 and 8 and assuming a 16.5% increase in demand for calls from this Bank Call Center, the new arrival rate would be 2.0 calls per minute (from 1.717 calls per minute to 2.0 calls per minute). With this increase, the probability of abandoning would vary from 10.2% (actual data) to 11.4% (Table 8), increasing by 11.8%.

The traffic intensity would vary from 82.8% (actual data) to 90.0% (Table 8) with an increase of 8.7%. With the same increase in demand (16.5%), more significant changes would occur in the average waiting time, which would vary from 0.813 minutes (actual data) to 1.2065 minutes (Table 7), increasing by 48.4%, and the probability of waiting, which would vary from 58.3% (actual data) to 72.0% (Table 7), increasing its value by 13.5%.

The traffic intensity, which measures the congestion of the system, would reach 99.4% when the arrival rate reached 3.0 calls per minute. If there was an increase of 74.7% (from 1.717 calls /min to 3.0 calls/min) in the call demand, this is the limit of increase supported in the Bank Call Center demand, from 12 to 1pm. In all these calculations, the values obtained from the analytical queueing model were considered as the values in Tables 7 or 8, and the value broken down as actual data are shown in Table 5.

The scenario, applied in the 12-1pm time frame in the Bank Call Center, showed that an increase in demand for system calls would also have a strong impact on its overall performance. The analyzed scenario highlighted the potential of using analytical queueing models with abandonment to support managers when making decisions about a possible increase in the number of Call Center calls. The queueing simulation model was also used to produce the results of the scenario and show the consistency and precision of the results of the analytical queueing model by comparing the results obtained by the two types of modeling.

13 Conclusions

This paper presented and applied analytical queueing models with abandonment, with generic or Exponential distributions, as an effective analysis approach to represent and analyze congestion problems in Call Centers. In particular, it investigated the use of mixed probability distributions that best fit the Call Center user patience times. The greatest interest was to consider the mixed distributions in the M/M/c+G analytical queueing models to be applied to congestion problems in Call Centers, obtaining performance measures that best express and represent reality.

Various queueing models that incorporate abandonment have been studied in the literature, but all of them consider a single class of parametric distributions to model patience, and the most common distributions are Exponential, Weibull and Erlang. However, it is common to find a population of users formed by groups with different characteristics in Call Centers, each of them having different behaviors for the patience time, and some are more patient than others. Due to this, it is generally more reasonable to model abandonment through mixed probability distributions, which are more sensitive to capturing these characteristics of user subpopulations.

Some mixed probability distributions, such as Mixed Lognormal, Mixed Exponential and Mixed Uniform were used in studies to adjust the patience time in Call Centers, but in these studies they were not considered to validate analytical queueing models in practice applied to Call Centers, with data extracted from reality. In the present study, we analyzed the application of these queueing models, considering generic and mixed distributions to model user patience, in a real case of a Call Center, with data from a bank overseas.

The validation was made by comparing the deviations of the performance measures obtained through analytical queueing models and those observed with the actual data and also estimated with the results from discrete simulation models. In order to meet this objective, we studied different analytical queueing models with abandonment found in the literature, which use exponential and generic probability distributions in the abandonment process.

Mixed and non-mixed distributions based on Lognormal and Exponential were used to model the abandonment process, among the possible distributions adjusted by the adherence test applied to these data. The arrival, service and abandonment rates were extracted using data from the Bank Call Center.

In addition, estimates were obtained for performance measures of the system, such as: the average waiting time, the probability of waiting, the probability of abandonment and traffic intensity. These parameters (rates) were inserted into the analytical queueing models considered in this study and these performance measures were calculated. Comparing these measurements with those extracted from the actual data, it can be affirmed that the generic analytical queueing models to represent user patience, considered in this study, produced results very close to those extracted from the actual data and, therefore, captured well the reality of the Call Center in all the analyzed performance measures.

The knowledge generated in this study can help to make improvements in understanding and managing these systems, adjusting them according to the allocated capacities, and allowing managers to interfere in decisions, exploring scenarios that optimize the quality of the service level. It is expected that more calibrated systems will improve customer satisfaction with the services rendered, due to the reduction of waiting times and a consequent reduction in abandonment fractions. The analysis methods proposed in this study are expected to help

companies make decisions about the size of their teams of attendants, improve system performance and reduce costs.

This study presented limitations when developing mixed distributions to represent the abandonment phenomenon. For the sake of simplicity, we chose to use two components of the same parametric family in the mixed distribution composition to represent the user patience times. No method was adopted to determine the amount of components in the mixed distributions used in this study, considering that the objective was to analyze if the application of mixed distributions to adjust the patience times, when inserted in the analytical queueing models makes them more effective to represent the analyzed case of the Call Center. This is an interesting topic for future research.

Other interesting future research would be to carry out a case study in other Call Centers, using this approach with generic distributions both in the service process and in the abandonment process. For example, in Call Centers with other architectures, specifically with those that operate in series. Still following this reasoning, we propose to use these models with parallel queues or hypercube queueing models, which are probably present in spatially distributed Call Centers.

Another interesting line of research would be to develop optimization procedures, based on the proposed analytical queueing models to support optimal system configuration decisions, for example, studying the minimum capacity allocation in the system to achieve performance measure values desired by the system manager, such as the maximum probability of abandonment or the maximum waiting time in the queue.

Acknowledgements

The authors would like to thank the anonymous reviewers for helpful comments and suggestions for revision, and Prof. Avishai Mandelbaum from the Technion-Israel Institute of Technology for kindly providing the data from the Bank case study in Israel. They would also like to thank CNPq and CAPES for the partial financial support for this research.

References

- Aalen, O., & Gjessing, H. (2001). Understanding the shape of the hazard rate: a process point of view. *Statistical Science*, 16(1), 1-22. <http://dx.doi.org/10.1214/ss/998929472>.
- Bacelli, F., & Hebuterne, G. (1981). On queues with impatient customers. In *Proceedings of the International symposium on computer performance - Performance '81* (pp. 159-179). Amsterdam: North-Holland.
- Bertrand, J. W. M., & Fransoo, J. C. (2002). Modelling and simulation: operations management research methodologies using quantitative modeling. *International Journal of Operations & Production Management*, 22(3), 241-264. <http://dx.doi.org/10.1108/01443570210414338>.
- Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S., & Zhao, L. (2002). *Statistical analysis of a telephone call center: a queueing-science perspective* (Working Paper) (pp. 1-61). Retrieved in 2012, July 6, from <http://iew3.technion.ac.il/serveng/References/references.html>
- Cox, D. R., & Oakes, D. (1990). *Analysis of survival data*. London: Chapman and Hall.
- Demo, P. (2000). *Metodologia do conhecimento científico*. São Paulo: Atlas.
- Ferrari, S. C. (2016). *Abordagens de modelos de fila com abandono para análise de congestão em Call Centers* (tese de doutorado). Departamento de Engenharia de Produção, Universidade Federal de São Carlos, São Carlos.

- Forbes, C., Evans, M., Hastings, N., & Peacock, B. (2011). *Statistical distributions*. New Jersey: John Wiley and Sons.
- Garnett, O., Mandelbaum, A., & Reiman, M. (2002). Designing a call-center with impatient customer. *Manufacturing & Service Operations Management: M & SOM*, 3(3), 208-227. <http://dx.doi.org/10.1287/msom.4.3.208.7753>.
- IBM Corporation. (2011). *SPSS statistics* (version 20.0). USA: IBM Corporation.
- Kalbfleisch, J. D., & Prentice, R. L. (1980). *The statistical analysis of failure time data*. Hoboken: John Wiley & Sons.
- Lima, H. (2012). *Call Center em números: métricas e como medir o desempenho*. Retrieved in 2012, July 6, from <http://blog.teclan.com.br/call-center-em-numeros-metricas-e-como-medir-o-desempenho/>
- Mandelbaum, A., & Zeltyn, S. (2005). *The Palm/Erlang-A queue with applications to call centers* (Working paper). Retrieved in 2012, July 6, from <http://iew3.technion.ac.il/serveng/References/references.html>
- Mandelbaum, A., & Zeltyn, S. (2009). Staffing many-server queues with impatient customers: constraint satisfaction in call centers. *Operations Research*, 57(5), 1189-1205. <http://dx.doi.org/10.1287/opre.1080.0651>.
- Mandelbaum, A., & Zeltyn, S. (2012). Data-stories about (im)patient customers in tele-queues. *Queueing Systems*, 75(2), 115-146.
- Mandelbaum, A., Sakov, A., & Zeltyn, S. (2001). *Empirical analysis of a call center* (Technical report). Haifa: Technion - Israel Institute of Technology. Retrieved in 2012, July 6, from <http://iew3.technion.ac.il/serveng/References/references.html>
- Mathwave Technologies. (2010). *EasyFit professional* (version 5.5). USA: Mathwave Technologies.
- Morabito, R., & Pureza, V. (2010). Modelagem e simulação. In P. A. M. Miguel (Org.), *Metodologia de pesquisa em engenharia de produção e gestão de operações* (pp. 165-194). Rio de Janeiro: Elsevier.
- Oliveira, C. C. (2009). *Espera e abandono na fila M/M/n+G e variantes* (Dissertação de mestrado). Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo.
- Palm, C. (1953). Methods of judging the annoyance caused by congestion. *Tele*, 4, 189-208.
- Palm, C. (1957). Research on telephone traffic carried by full availability groups. *Tele*, 1, 107.
- Rockwell Automation Technologies, Inc. (2014). *Arena* (version 14.70). USA: Rockwell Automation Technologies, Inc.
- Technion – Israel Institute of Technology. (2002). *Call center data*. Retrieved in 2010, February 15, from <http://iew3.technion.ac.il/serveng/callcenterdata/index.html>
- Whitt, W. (1993). Approximations for the GI/G/m queue. *Production and Operations Management*, 2(2), 114-161. <http://dx.doi.org/10.1111/j.1937-5956.1993.tb00094.x>.
- Wolfram Research, Inc. (2013). *Mathematica 9* (version 9.0.1.0). USA: Wolfram Research, Inc.
- Zeltyn, S. (2004). *Call Center with impatient customer: exact analysis and many server asymptotics of M/M/n+G queue* (thesis). Technion-Israel Institute of Technology, Haifa.
- Zohar, E., Mandelbaum, A., & Shimkin, N. (2002). Adaptive behavior of impatient customers in tele-queues: theory and empirical support. *Management Science*, 48(4), 556-583. <http://dx.doi.org/10.1287/mnsc.48.4.566.211>.