Workforce Management in Call Centers: Forecasting, Staffing and Empirical Studies

Sihan Ding

丁思涵

Workforce Management in Call Centers

Forecasting, Staffing and Empirical Studies

ISBN: 978-94-6328-035-8

The research in this thesis was supported by the Netherlands Organization for Scientific Research (NWO), as part of the research cluster 'Stochastics: Theoretical and Applied Research' (STAR), and by the Dutch Ministry of Defense.

Cover: designed by Siyi Liao
Print: CPI – KONINKLIJKE WÖHRMANN

VRIJE UNIVERSITEIT

Workforce Management in Call Centers

Forecasting, Staffing and Empirical Studies

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor aan
de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. V. Subramaniam,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de Faculteit der Exacte Wetenschappen
op woensdag 25 mei 2016 om 13.45 uur
in de aula van de universiteit,
De Boelelaan 1105

door

Sihan Ding

geboren te Xiangtan, China

promoters: prof.dr. G.M. Koole
prof.dr. R.D. van der Mei

Acknowledgements

The past four years have been a very pleasant and unforgettable period for me. Finishing a PhD degree is a long and heavy project, and it challenges not only the mathematical knowledges and skills, but also other things, such as time management and self-discipline. Besides working, there are also many memorable trips I made to many places around the world. In all, I have enjoyed this journey!

I would not have achieved it without the guidance of my two supervisors, Ger Koole and Rob van der Mei. Ger is a very inspiring researcher, his knowledge and experience in call centers and operations research are profound both in theory and in practice. I enjoyed very much working with him, especially, I cherish those moments when we sat together and asked questions to find the right research pass. Also, he is the most sportive and adventurous mathematician I have ever known. During my PhD, I did also some consulting and research work for the Ministry of Defence. It was quite challenging and difficult to work on those projects next to my PhD research, and I would not be able to handle it well without Rob's support. Besides, his passion for work and how he balances all the workload set a good example for me.

This project would not have been there without the consistent efforts of Jan Kalden. Together, we worked on interesting and successful projects. His enthusiasm for data analytics and his unique life story will keep motivating me in the future.

I would like to thank Vanad Laboratories and the municipality of Rotterdam for sharing data with us, helping us to understand the data sets, and answering related questions. The research is much more applicable and insightful with the existence of the real data. In addition, I thank Benjamin Legros and Oualid Jouini for hosting my short visit in Paris during the beautiful Autumn. It has been great pleasure to work with them.

I also want to thank the reading committee members, Zeynep Akşin, Sandjai Bhulai, Sem Borst, Rommert Dekker, Pierre L'Ecuyer, and Jiheng Zhang for their efforts in reading my thesis and giving useful and interesting feedbacks.

Working at CWI and at the VU have always been enjoyable, firstly because of the relaxed environment (from which I developed a good habit of running), but largely because of all the nice colleagues. Specifically, I thank Thije, Maria, Dirk, Daphne, Jan-Pieter, Asparuh, Caroline, Ewan, Bart, Pieter, Peter, Frank, Marie-Collette and Kacha at CWI; René, Ruben and Marijn at the VU. Moreover, I pay special thanks to Alex, Bert, Sandjai, Raik, Martijn, Chrétien, Joost, Masha, Arnoud and Martin, for many discussions we had together, and their helps in different aspects.

I am grateful to my friends in the Netherlands and abroad, Jilin, Yaxian, Cong, Dai, Julia, Yuehong, Vincent, Pengqi, Yu, Plamen and Maggie, with whom I had many memorable evenings and weekends. Also, many thanks to Siyi Liao, who is a special friend to me and has designed and made the cover of this thesis.

这篇论文的存在离不开我的父母，感谢你们永远在支持和鼓励。我希望我的这个成就能够带给你们快乐和骄傲！This thesis would not exist without the support of my parents, who are always supporting and encouraging me. I hope the accomplishment will bring as much joy and proud to you as it brings to me.

Last but not least, much love and gratitude to my newly-become wife, Siyi. You bring joy and colors to my life in the Netherlands and everywhere I go.

Sihan Ding
November 2015
Amsterdam

Contents

Chapter 1

Introduction

This thesis covers two crucial steps of workforce management in call centers, *forecasting* and *staffing*. It involves knowledge in several fields, such as statistics, operations research, as well as domain knowledge. Many researchers and practitioners across the world have spent countless efforts in developing models and methodology during the last few decades. This thesis contributes in solving and discovering various aspects of forecasting and staffing, with the support of *empirical analysis*. These aspects include: the best error measurement in forecasting, traffic management, how caller behavior influences forecasting accuracy, its corresponding staffing method and the comparison of different staffing methods.

1.1 Call centers and workforce management

A call center is a place where customers or callers are handled by a group of agents, who use telephones or other telecommunication means to address callers' requests or questions. It is nowadays also referred to as a *contact center*, as agents may use other means of communication, such as posts, emails, web chats, etc. Depending on who initiates the call, call centers can be categorized into two types, *inbound call centers* and *outbound call centers*. In an inbound call center, calls are initiated by callers, and agents usually provide support and information for callers. In an outbound call center, agents are the ones who initiate the calls, for example, for the purpose of telemarketing or market research. Sometimes, call centers support a mixture of inbound calls and make outbound calls. Another way of categorizing call centers is based on whether they are single-skill or multi-skill call centers. A multi-skill call center has agents who have different skills, and different agent may specialize in answering different type of calls. A single-skill call center has only one type of agents, who are designated to answering one type or all types of calls.

Call centers are either cost centers or profit centers, with expenditure in many things, such as facilities and equipment. Among those costs, the workforce costs are the main source of costs in call centers; they account for roughly 60% – 70% of the total operational costs (Gans et al. (2003)). Therefore, well-planned and -executed workforce management can lead to a great deal of savings, and optimizing workforce management has received tremendous attention over the past few decades.

Workforce management is essentially balancing between costs and quality. On the one hand, having too little agents can result in system congestion, which then leads to callers having excessive waiting times; on the other hand, having too many agents may generate unnecessary costs, although callers will experience less waiting. The goal of workforce management is to make sure that this trade-off is well balanced. It is usually divided into four steps (Koole (2013)):

Forecasting: estimating future workload;

Staffing: determine the staffing level, i.e., determine the number of agents;

Scheduling and Rostering: making shift schedules and rosters;

Traffic management: adjusting staffing levels and rosters in real-time.

In sections 1.2 and 1.3, we will elaborate on forecasting and staffing.

1.2 Forecasting

Estimating future workload is a main, yet difficult, part of workforce management. Depending on the goal of forecasting, there are usually three types of forecasts with different forecasting horizons. *Long-term* forecasts are often made a few months or even years in advance, for the purposes of making strategic decisions such as financial planning and hiring policies. *Short-term* forecasts are usually made a few weeks in advance, and serve as an input to staffing formulas, which translate them into the number of agents of each day and each interval, which is then used to make scheduling and rostering decisions. Finally, it is common in some call centers to make real-time adjustments to forecasts during the day itself. We will elaborate on such real-time adjustments more in Section 1.5.

In Figure 1.1, we show an example of the actual number of arrivals per interval from real call center data and its forecasts. In this example, the call center operates from Monday till Friday and from 8:00 am in the morning till 20:00 pm in the evening, and in this plot we start with a Monday and end with a Monday two weeks later, thus, 11 consecutive operational days in total. Also, we divide the operations hours into 24 intervals, with each interval having a length 30 of minutes. This is a common practice in call centers, as it makes call center operations easier to manage and to report. Other interval lengths, such as 15 minutes and 1 hour, are also often used. Based on this figure, we see some interesting patterns in the arrival process. The first pattern is the intraday seasonality, e.g., we see many arrivals at around 9:30 am, and the volume goes down at around 12:00 am or 12:30 pm, which is during the period for the lunch breaks, and it follows another small peak shortly after lunch at around 13:00 pm or 13:30 pm. The second pattern is the intraweek seasonality, as we can observe that there are in general more arrivals on Monday than the rest of the week, and it receives less calls on Friday than other days of the weeks. These two patterns are not coincidental, since we see them in many other call center data sets. Therefore, in order to have an accurate forecasting model, one should take these seasonalities into consideration. To generate the forecasts in Figure 1.1, we simply take the average of the same interval of the same weekday over the past four weeks. This is a simple and intuitive method, but the forecasts are quite accurate in this specific example. Besides the two seasonalities we have discussed in this example, there are more issues that forecasters must pay attention to. For example, often the number of arrivals in different months of the year also shows a pattern, as there might be less callers in the summer during the summer holiday period compared to other months of the year. This is referred to as the intrayear seasonality. Another example is the special days or events, such as billing days (a day where all employees receive salary slips), which will trigger many questions and subsequent phone calls to the financial department. We refer to Andrews and Cunningham (1995) for an easy but practical method to deal with such a holiday effect.

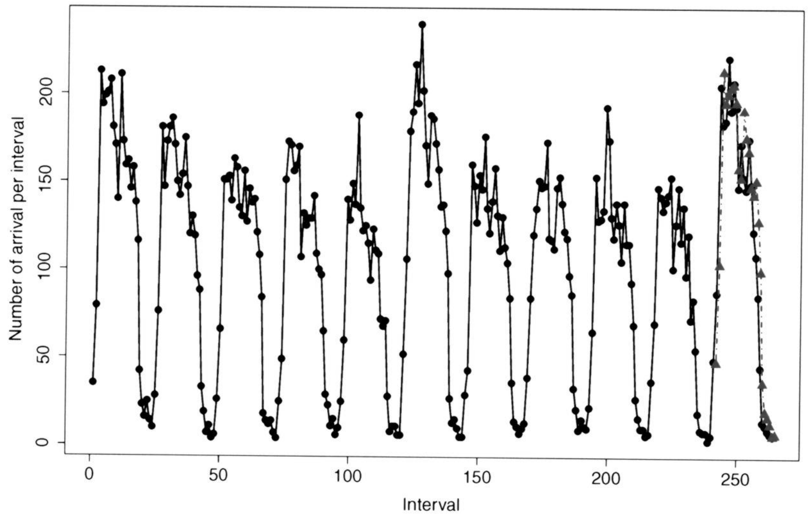


Figure 1.1: Actuals (solid line with dots) and forecasts (dashed line with triangle).

Based on the example we show in Figure 1.1, we see that making the right forecasts is not simply about applying different statistical models and choosing the one that has the least errors, but it also involes understanding the business and the data, which requires domain knowledges. Generally speaking, there are two ways to make call volume forecasts, *qualitative* and *quantitative* ways. The former way is based on human instinct, experience and judgment, while the latter way makes use of statistical methods, or more specifically, time-series analysis. Accurate forecasts or forecasting methods should integrate and combine both statistical methods and field experiences, as the statistical approaches are objective and easily reproducable, while experiences and judgments can provide expertise in choosing the right methods and parameters and identifying outliers such as special days that have large influences in call volume. More detailed discussions on forecasting are given in Chapter 3.

Due to the random nature of call arrivals, forecasting errors are inevitable. Different error measurements are used to make comparison of different forecasts and forecasting models. The commonly used error measurements include the mean absolute percentage error (MAPE), the weighted mean absolute percentage error (WAPE), the sum of absolute deviance (SAD) or the equivalently sum of absolute error (SAE) and the sum of squared error (SSE). If we denote x_i as the actual number of arrivals in interval i, and $\hat{x}_i$ is the forecast of x_i, $i = 1, 2, \ldots, I$, where I is the forecasting horizon, then

$$\begin{aligned}
\text{MAPE} &:= \frac{1}{I}\sum_{i=1}^{I}\frac{|\hat{x}_i - x_i|}{|x_i|},\\
\text{WAPE} &:= \sum_{i=1}^{I}\frac{|\hat{x}_i - x_i|}{|x_i|}\cdot\frac{|x_i|}{\sum_{j=1}^{I}|x_j|} = \frac{\sum_{i=1}^{I}|\hat{x}_i - x_i|}{\sum_{i=1}^{I}|x_i|},\\
\text{SAD} &:= \sum_{i=1}^{I}|\hat{x}_i - x_i|,\\
\text{SSE} &:= \sum_{i=1}^{I}(\hat{x}_i - x_i)^2.
\end{aligned}$$

Note that WAPE is the weighted sum of MAPE, and the weight of each day is based on the call volume of that day. More discussion about error measurements can be found in Chapter 2.

Call volume forecasting should not be the only part of workload forecasting; other things such as the average handling times (AHT), or even the distribution of the handling time (see for example Ibrahim et al. (2016a)), should also be estimated. For example, AHT differs per interval and per weekday, and similar to the number of arrivals, it also shows intraweek and intraday patterns (Ibrahim et al. (2016a)). Furthermore, as we show in Chapter 6, different agents have different AHT, and new agents' AHT decreases over time as they learn. Without taking the AHT fluctuations into consideration, the accuracy of the workload estimation will be strongly undermined.

Agents are not available all the time to answer calls. They take breaks, do trainings, and sometimes take holidays or sick-leave days. These types of activities are called *shrinkage*. Shrinkage is a feature that should not be ignored in workforce management in call centers.

1.3 Staffing

The staffing procedure is in principle translating forecasts and target service levels (SL) into the number of agents. For a single-skill call center, several models have been developed in the literature to assist this decision making process. Different models have different assumptions and include different features to mimic reality. We now describe three commonly used models: the Erlang C model, the Erlang A model and the Erlang

X model. Besides these existing models, we also introduce a new model with extra features of redial and reconnect found in real call center data. This model has not been studied before in the context of call centers.

1.3.1 The Erlang C model

The Erlang C model is a well-known model for single-skill call center staffing. The model is referred to as a $M/M/s$ queueing system, illustrated in Figure 1.2. In this model, it is assumed that calls arrive according to a Poisson process with rate λ. There are s agents who handle inbound calls. An arriving call is handled by an available agent, if there is any. Otherwise, it will wait in a queue with infinite buffer size. The calls are handled in the order of arrival, which is also called First Come First Serve (FCFS). We denote the random variable B as the handling time (HT) of a caller. It is assumed that B has an exponential distribution, denoted $B \sim exp(\mu)$, where $\mu > 0$ is the service rate.

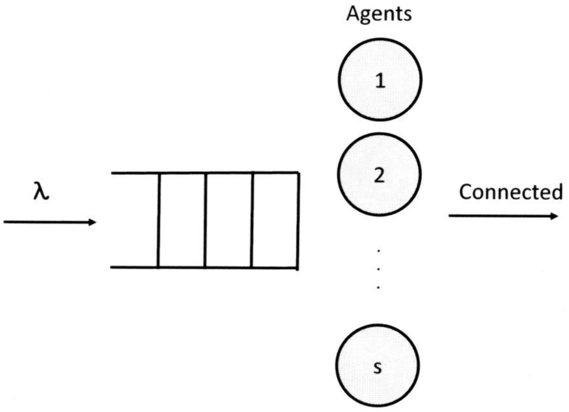


Figure 1.2: The Erlang C model.

The Erlang C formula was derived by Erlang (1917), who analytically solved such a $M/M/s$ system by modeling it as a continuous-time Markov Chain. Let $a = \lambda/\mu$, and assume that $s > a$. Then for some given acceptable waiting time (AWT), the Erlang C formula writes

$$\mathrm{SL} := \mathrm{P}(W \leq \mathrm{AWT}) = 1 - \mathrm{P}_w(s,a) \cdot e^{-(s\mu-\lambda)\mathrm{AWT}}, \tag{1.1}$$

where $\mathrm{P}_w(s,a)$ is the probability that an arbitrary caller has to wait, and it is given by (Gans et al. (2003))

$$\mathrm{P}_w(s,a) = \frac{a^s}{(s-1)!(s-a)} \left(\sum_{i=0}^{s-1} \frac{a^i}{i!} + \frac{a^s}{(s-1)!(s-a)} \right)^{-1}. \tag{1.2}$$

SL can be interpreted as the fraction of callers that wait no longer than AWT time units. By using the Erlang C formula (1.1)-(1.2), one can determine the minimum number of agents needed to satisfy the service level target.

Other Key Performance Indicators (KPIs) are also of interest in practice, such as average speed of answer (ASA), which is the average waiting time of callers. For the Erlang C model, ASA can be derived via the following expression

$$\mathrm{ASA} = \frac{\mathrm{P}_w(s,a)}{s\mu - \lambda}.$$

To further explain SL and ASA, we plot the SL and the ASA of the Erlang C model as a function of λ in Figure 1.3. We let λ range from 44 to 51.5 with increment of 0.1, and we let $1/\mu = 1$ minute, $s = 52$ and $\mathrm{AWT} = 1/3$ minutes.

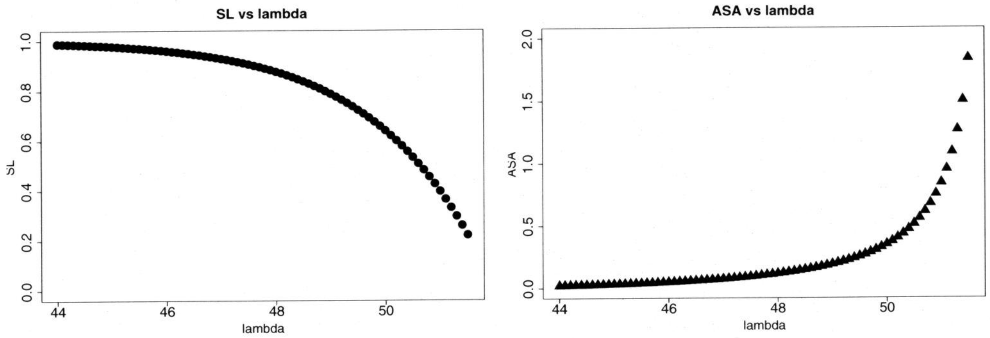


Figure 1.3: SL (left) and ASA (right) in minutes of the Erlang C model of different values of λ.

1.3.2 The Erlang A model

The Erlang C model ignores an important caller feature in call centers, which is *abandonment*, as waiting callers get impatient while waiting, and may choose to abandon. The Erlang A (where "A" stems from abandonment) supplements the Erlang C model with an extension of this feature. A diagram of the Erlang A model is shown in Figure 1.4, which is also denoted by the $M/M/s+M$ queueing sysytem. In the Erlang A model a caller who waits in the queue has limited patience, and the patience, denoted by the random variable H, is assumed to be exponentially distributed. We assume that $\mathrm{E}H = 1/\theta < \infty$, where θ is the abandonment rate.

We denote W as the *actual* waiting time of an arbitrary caller in the Erlang A model, and V as the *virtual* waiting time, which is the waiting time of a caller with infinite patience. With such notations

$$W = \min\{V, H\}.$$

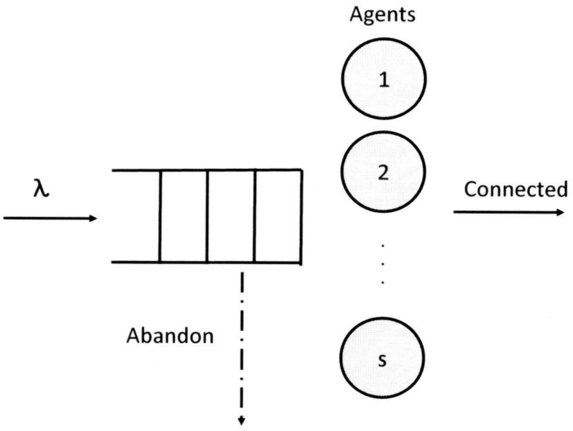


Figure 1.4: The Erlang A model.

The Erlang A model can be solved in a similar fashion as the Erlang C model, i.e., by modelling it as a continuous-time Markov Chain, and solving the global balance equations to obtain the steady state distribution. By doing so, we can derive

$$\begin{aligned} \mathrm{P}(W > \mathrm{AWT}) &= \sum_{i=1}^{\infty} \mathrm{P}(W > \mathrm{AWT} | L = i) \pi_{s+i} \\ &= \sum_{i=1}^{\infty} \pi_{s+i}\, \mathrm{e}^{-(s\mu+\theta)\cdot \mathrm{AWT}} \sum_{j=0}^{i} \frac{\phi_j (1 - \mathrm{e}^{-\theta \cdot \mathrm{AWT}})^j}{j!}, \end{aligned}$$

where $\phi_j := \phi(\phi+1)\cdots(\phi+j-1)$ for $j > 1$ with $\phi = s\mu/\theta$ and $\phi_0 = 1$, L is a random variable representing the number of callers in the queue, and π_i is the steady state probability of having i callers in the queue and in service, which is given by (Riordan (1962))

$$\pi_i = \begin{cases} \dfrac{\lambda^i}{\mu^i i!} \pi_0, & 1 \le i \le s, \\ \dfrac{\lambda^i}{\mu^s s! \prod_{j=1}^{i-s} (s\mu + j\theta)} \pi_0, & i > s, \end{cases}$$

and

$$\pi_0 = \left(\sum_{i=0}^{s} \frac{\lambda^i}{\mu^i i!} + \sum_{i=s+1}^{\infty} \frac{\lambda^i}{\mu^s s! \prod_{j=1}^{i-s} (s\mu + j\theta)} \right)^{-1}.$$

Also, the ASA is then given by

$$\mathrm{ASA} = \frac{\mathrm{E}L}{\lambda} = \frac{\sum_{i=s+1}^{\infty} (i-s)\pi_i}{\lambda}.$$

Moreover, according to Roubos (2012), for the virtual waiting time V, one has

$$\begin{aligned}
\mathrm{P}(V > \mathrm{AWT}) &= \sum_{i=1}^{\infty} \mathrm{P}(V > \mathrm{AWT}|L = i)\pi_{s+i} \\
&= \sum_{i=1}^{\infty} \pi_{s+i}\, \mathrm{e}^{-s\mu\cdot\mathrm{AWT}} \sum_{j=0}^{i} \frac{\phi_j(1-\mathrm{e}^{-\theta\cdot\mathrm{AWT}})^j}{j!}.
\end{aligned}$$

The extra feature of callers abandoning results in new KPIs. For example, in some call centers, managers look at the *abandonment percentage*. They do not count callers who abandoned before AWT for violating SL. To be more precise, we define the following KPIs and re-define SL.

$$\begin{aligned}
r &:= \frac{\#\,\text{abandoned}}{\#\,\text{offered}}, \\
\mathrm{SL}_1 &:= \frac{\#\,\text{answered} \leq \mathrm{AWT}}{\#\,\text{answered}}, \\
\mathrm{SL}_2 &:= \frac{\#\,\text{answered} \leq \mathrm{AWT}}{\#\,\text{offered}}.
\end{aligned}$$

r is the percentage of callers that have abandoned, and SL_1 is the proportion of answered callers that are answered within AWT time units, and SL_2 is the proportion of callers that are answered within AWT time units.

The abandonment percentage r, SL_1 and SL_2 of the Erlang A model can be expressed in the following way.

$$\begin{aligned}
r &= \sum_{i=s+1}^{\infty} \frac{(i-s)\theta\pi_i}{\lambda}, \\
\mathrm{SL}_1 &= \frac{\mathrm{P}(V < \mathrm{AWT}, V < H)}{1-r}, \\
\mathrm{SL}_2 &= \mathrm{P}(V < \mathrm{AWT}, V < H).
\end{aligned}$$

To further explain different KPIs for the Erlang A model, we use the following examples. We let $s = 50$, 5 minutes AHT, the average patience is 2 minutes, and AWT$= 1/3$ minutes, while let the λ range from 8 to 15 with increment of 0.1. $\mathrm{SL}_1, \mathrm{SL}_2, r$ and ASA are shown in Figures 1.5 and 1.6. Note that when $\lambda > 10$, the offered load (i.e., λ times AHT) is larger than the number of agents. This means the incoming workload exceeds the workload that agents can handle. In the Erlang C model, this would lead to infinite queue length and customers waiting infinite time in the long run. However, in the Erlang A model, one may still have reasonably good SL, due to the presence of customer abandonment. For example, when $\lambda = 10.5$, the workload per agent is $10.5 \cdot 5/50 = 1.05$, and we have $\mathrm{SL}_1 = 77.6\%$, $\mathrm{SL}_2 = 70.3\%$, which means 77.6% of answered customers are answered within 20 seconds and 70.3% of all the customers are answered within 20 seconds, respectively, while having 9.5% of the customers abandonment.

The exponential assumption of the patience distribution is a strong assumption of the

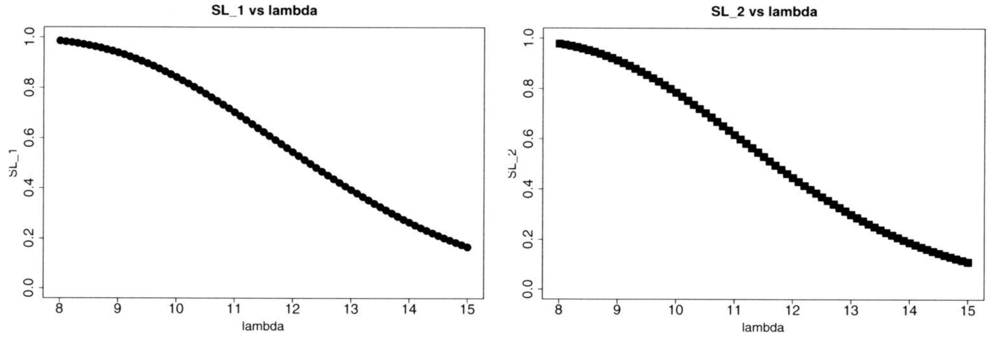


Figure 1.5: SL_1 (dot) and SL_2 (square) of the Erlang A model for different values of λ.

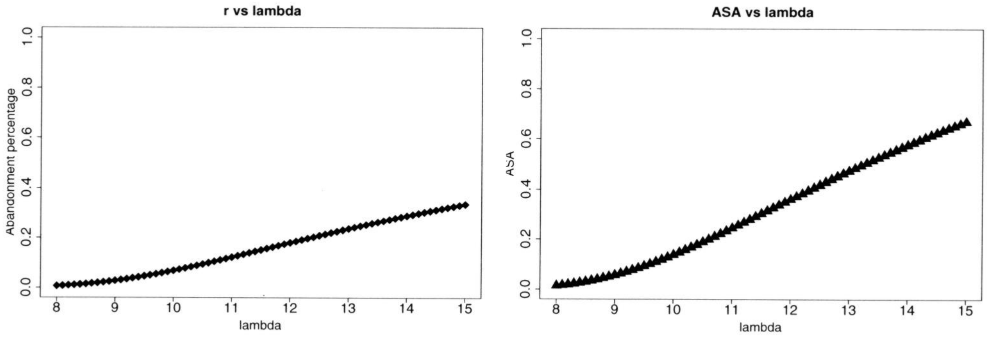


Figure 1.6: r (diamond) and ASA (triangle) in minutes of the Erlang A model for different values of λ.

Erlang A model. However, several empirical studies, such as Roubos and Jouini (2013), Mandelbaum and Zeltyn (2004) and Chapter 6 of this thesis, show that such an assumption is not valid in many cases. With the relaxation of this assumption, this model becomes the $M/M/s+G$ model where G stands for the general patience distribution. For such a model, one could use numerical techniques which involves computation of numerical integration to obtain performance metrics (see for example Zeltyn and Mandelbaum (2005)). As shown in Whitt (2005) that for some queueing models the patience distribution has a larger influence compared to the HT distribution.

1.3.3 The Erlang X model

The Erlang X model is an extension to the Erlang A model. Specifically, in the Erlang X model, it is assumed that a proportion p of the abandoned callers may *redial* after some amount of time denoted by the random variable Γ_{RD}, and $1-p$ portion of them will not redial thus are considered as "lost" callers. We denote δ_{RD} as the redial rate, i.e., $\mathrm{E}\Gamma_{RD} = 1/\delta_{RD}$. We assume that p does not depend on callers' experiences in the system. These experiences include handling times, waiting times and the number of times that callers have already called. A diagram of the Erlang X model is shown in Figure 1.7.

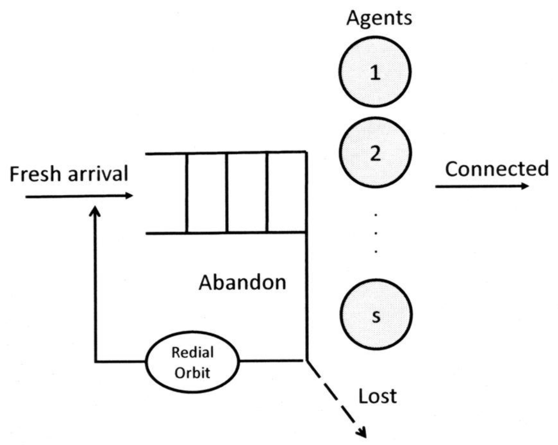


Figure 1.7: The Erlang X model.

Despite the fact that the Erlang X model only adds one feature to the Erlang A model, there is no closed-form expression for the system's performance, even when Γ_{RD} is exponentially distributed. However, Sze (1984) provides approximations for a more general system where the HT can have a general distribution. To be more specific, Sze (1984) uses the following relations

$$\mathrm{P}(W < t) \approx f(r, \lambda'), \tag{1.3}$$

$$r = \int_0^\infty A(t) d\mathrm{P}(W < t), \tag{1.4}$$

$$\lambda' = \lambda + p \cdot r \cdot \lambda', \tag{1.5}$$

where $f(\cdot)$ is the function that calculates the probabiliy that a customer waits less than t time units in the Erlang A model with arrival rate being λ', $A(t)$ is the probabiliy that a customer will abandon before t time units. It is assumed in Sze (1984) that relations (1.3)-(1.5) have a fixed point, and one could solve these relations to obtain the performance metric.

Here we show a numerical example of the Erlang X model. In this example, we let $\lambda = 40, \mu = 1, s = 40, p = 0.5, \mathrm{AWT} = 20$ seconds, patience be 2 minutes, and $\Gamma_{RD} = 10$ minutes. Via simulation, we can obtain the long-term average performance metrics: $\mathrm{SL}_1 = 78.3\%, \mathrm{SL}_2 = 71.1\%, r = 9.2\%$ and ASA= 18.4 seconds.

1.3.4 A call center model with redial and reconnect

In some real call center data, we discovered another feature which is not described in either of the models mentioned above. This feature is called *reconnect*, as connected callers also call back. To make the distinction between initial calls and redials or reconnects, we refer to these initial calls as *fresh* calls. The notation λ is not specific and clear enough for this model, as an arrival could be incurred by a fresh call, a redial or a reconnect. To avoid confusion, in this model, we denote λ_F as the arrival rate of the fresh calls, and λ_T as the total arrival rate including the redials and the reconnects.

In addition to the assumptions made in the Erlang X model, in this model, an answered caller enters the reconnect orbit with probability q, and will reconnect after some generally distributed time Γ_{RC}, with $\mathrm{E}\Gamma_{RC} = \delta_{RC} < \infty$. We refer to these calls as *reconnects*. We assume that q does not depend on callers' experiences in the system. A diagram for a call center model with redials and reconnects is shown in Figure 1.8.

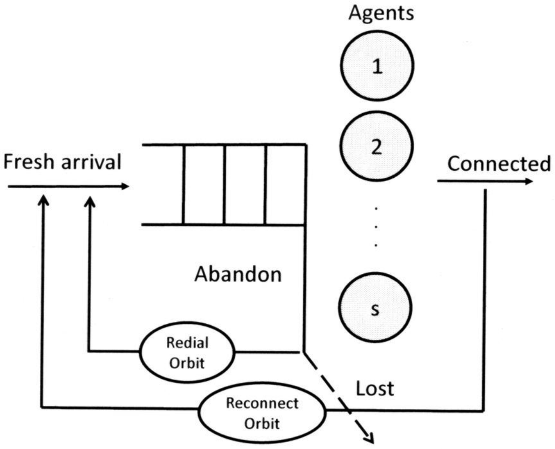


Figure 1.8: A call center model with redial and reconnect.

We now show a numerical example. We let $\lambda_F = 40, \mu = 1, s = 40, p = 0.5$, AWT $=$ 20 seconds, patience be 2 minutes, and $\Gamma_{RD} = 10$ minutes, and $q = 0.1, \Gamma_{RC} = 50$. We calculate the performance metrics via simulation: $\mathrm{SL}_1 = 44.2\%, \mathrm{SL}_2 = 36.3\%, r = 17.8\%$ and ASA$=$ 36.2 seconds. By comparing the performance metrics of this example with those of the example in Subsection 1.3.3, we can see that reconnects have a strong influence on the performance metrics we consider.

1.4 Scheduling and rostering

In the previous section, we introduced some staffing models that translate demand (forecasts) into capacities (agents) with respect to meeting requirements (SL). In this section, we discuss briefly how to obtain agents schedules.

Agents scheduling can be done in two steps: shift scheduling and assigning agents to shifts. The inputs for shift scheduling are: number of agents needed per interval s_i, shifts of agents a_{ij} ($a_{ij} = 1$, if interval i is covered in shift j, $a_{ij} = 0$, otherwise), costs per shift c_j, and the outputs are: the number of agents per shift x_j, while the constraints are: $x_j \geq 0$ and x_j is integer, for interval $i = 1, 2, \ldots, I$ and shift $j = 1, 2, \ldots, J$, with the objective being having minimum costs $\sum_{j=1}^{J} c_j x_j$. We express this problem as an Integer Programming (IP) problem :

$$\begin{aligned}
&\min_{x_j} \sum_{j=1}^{J} c_j x_j \\
&\text{s.t.} \quad \sum_{j=1}^{J} a_{ij} x_j \geq s_i, \quad i = 1, 2, \ldots I, \\
&\qquad\quad x_j \geq 0, x_j \text{ is integer} \quad j = 1, 2, \ldots J.
\end{aligned} \tag{1.6}$$

Such an IP formulation dates back to G.B. Danzig, who has made tremendous contribution both in theory and in practice of this model (Cottle et al. (2007)).

We now use an example to illustrate this step. First, s_i are calculated using the Erlang C formula and based on the forecasts in the example in Section 1.2. We consider six types of different shifts, the first one starts at 8:00 am and ends at 17:00 pm with an half hour break after each two hours of work, the second shift, third shift and the fourth shift are the same as the first one except that the starting times and the ending times are one hour, two hours and three hours later than the first shift, respectively. Each of these four shifts covers 8 working hours per day, and they are meant for full-time agents, and the costs for these shifts are 1. The fifth shift starts at 8:00 am and ends at 12:00 am, with an half hour break at 10:00 am, and the sixth shift starts at 10:00 am and ends at 14:00 with an half hour break at 12:00 am. The fifth and the sixth shifts represent those shifts for part-time agents, and they have costs of 0.6. Solving the IP problem (1.6) for this example, we obtain the number of agents per shift, which is plotted in Figure 1.9. Note that we have also added 10% of shrinkage in the following way: if x_j agents are needed for shift j, then we need to have at least $x_j/(1 - 10\%)$ number of agents with the consideration of the shrinkage. As one can see that a large amount of overstaffing is introduced in the scheduling process. There are some solutions to solve this problem. For instance, one could add more shifts, especially shorter shifts. Also, instead of having SL constraints for each interval, one could have a SL constraint for the whole day, which allows violating constraint $\sum_{j=1}^{J} a_{ij} x_j \geq s_i$ for some i and will enable more efficient planning.

Now we discuss assigning agents to shifts. This can be a quite sophisticated step, as it involves fairness, agents' preferences, labor laws and regulations. One way of making rostering is letting the agents choose the shifts themselves, either based on the duration of their working hours or via a bidding system where each agent has certain amount of points to spend in choosing the preferred shifts.

Besides the two-steps approach we mentioned above, one can integrate these two steps into one by assigning agents to shifts without determining the optimal shifts. Such a procedure can easily get more complex as one needs to take several factors into consideration simultaneously, such as law and regulations on workforce, agents' preferences, shift optimization, etc.

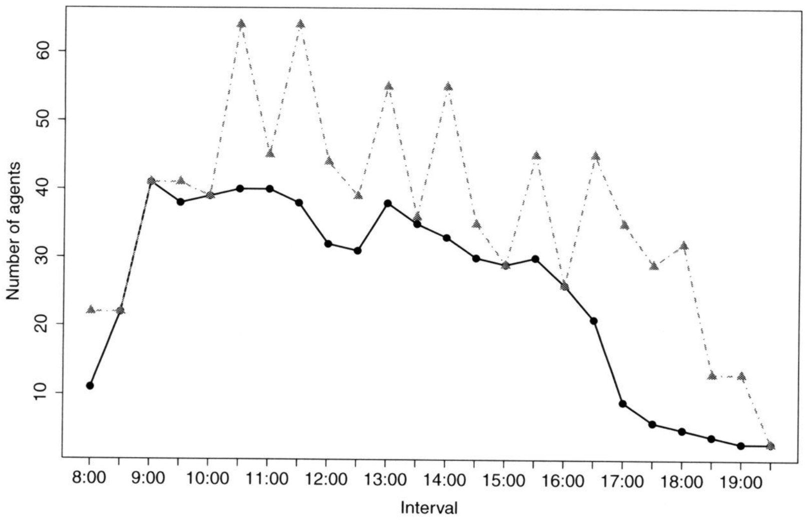


Figure 1.9: Number of agents per interval needed (solid line with dot) and the number of agents per interval scheduled (dashed line with triangle).

In this section, we gave a brief introduction on agent scheduling. However, this procedure can be far more complicated than what we described here. For example, we do not consider the topic of finding the best shifts with breaks, which is mathematically very challenging (Koole (2013)). Also, for multi-skill call centers, staffing and shift scheduling can be combined in one step, and can be optimized via simulation (Koole and Pot (2006)). However, due to the large size of the solution space, finding the optimal agent schedules via simulation can be sometimes challenging. In fact, even when each solution can be evaluated quickly, this problem is a NP-hard problem. Thus, heuristic searching methods have been proposed, such as the neighborhood search method by Avramidis et al. (2009) and the cutting-plane method proposed by Cezik and L'Ecuyer (2008) and later refined by Avramidis et al. (2010). We refer the readers to Aksin et al. (2007) and references therein for more discussion on agent scheduling. Moreover, for multi-channel call centers, there are possibilities to make more efficient planning by having agents work on other channels rather than inbound calls, but it makes the agent scheduling more complex. Due to these difficulties as well as other additional functionalities (for example, integrating scheduling with training and hiring plans), many call center managers choose to use workforce management tools, such as NICE/IEX, and Aspect (DMG Consulting LLC. (2012)).

1.5 Traffic management

Even when all the previous three steps are carefully designed and implemented, workforce management can still go wrong if there is no traffic management (also known as real-time performance management) in place. This is because there are still many uncertainties and fluctuations in the number of arrivals, time of arrivals, shrinkage, HT, SL, etc. These uncertainties and fluctuations may be caused by a sudden event or simply due to pure randomness where we have no control. Traffic management is a way to balance them, and have more control over the outcome (SL in this case) under uncertainties. In this section, we describe some means of traffic management. Some of them come from the literature while others have already been adopted in practice.

Making real-time forecast updates is a common way of doing traffic management. With the acquisition of the most recent data, such as the data within the same day, forecasters can update their initial forecasts. Usually the shorter the forecasting horizon, the more accurate the forecasts become (Taylor (2012)). One way to model it is using a "busyness" factor proposed by Avramidis et al. (2004b) and later refined by Steckley et al. (2009), and by Oreshkin et al. (2014) with the additional feature that such a "busyness" factor is dependent across different periods within a day. The idea of such a "busyness" factor is that it indicates how busy a day is, and by updating it using real-time data, one can on average obtain more accurate forecasts for the rest of the day. Another way of making updates is adding the amount of redials and reconnects for the rest of the day. To be more specific, if a traffic manager observes a busy morning where many callers have abandoned and been answered, then with the consideration of the redial and reconnect behaviors (we will study these two behaviors in more depth in Chapters 3 and 4), she can add the expected amount of redials and reconnects to the initial forecasts for the afternoon, and make staffing adjustment accordingly. Besides these two ideas on making forecasts adjustments, we refer to Shen and Huang (2008) and Gans et al. (2012) for other statistical models on intraday updating.

Many call centers have both part-time and full-time agents. The working hours of full-time agents are fixed, thus, they have some flexibility, but only to a certain extent. For instance, they can shift their breaks to some time earlier or later depending on the occupancy at that moment. In contrast, the part-time agents usually do not have fixed shifts. The existence of these part-time agents offers possibilities to make real-time adjustments, since they can be asked to answer calls with short notice. Under the assumption of arrival rate uncertainty, it is beneficial to make initial staffing decisions knowing that part-time agents can cover the excessive loads in case needed. We will explain this part in more detail in Chapter 2. Making such flexible workforce planning is a way to balance not only the uncertainty in the arrival process, but also the fluctuations in the SL (see for example Roubos et al. (2012)).

Updating forecasts and updating staffing levels are making adjustments within the call centers. It can also be done in a way that involves callers, such as using a call-back option modeled by Armony and Maglaras (2004a) where callers can choose to either wait online or be called back. This option flattens some burstiness in the arrival process and subsequently flattens the SL fluctuations, while it does not harming the callers' experiences since some callers would rather take this option than wait a long time in the queue. In Chapter 5, we study another call-back option used in practice, and it shows much higher efficiency compared to the model without the call-back option.

Nowadays, besides inbound calls, many call centers also handle other workloads like emails and outbound calls. To handle them efficiently requires good traffic management. In a call center which handles both emails and inbound calls with a hard contraint on the SL of inbound calls, it is much more cost saving if agents start answering emails only when the occupancy is below a certain threshold compared to having dedicated agents separately answering inbound calls and emails (Legros et al. (2015)).

1.6 Structure of the thesis

In Chapters 2 and 3, we mainly focus on call center forecasting. Specifically, we study how caller behavior influences the accuracy of the forecasts, and what the optimal accuracy measurement is. Chapters 4 and 5 discuss call center staffing for two specific models. We make a validation and comparison of several different staffing models in Chapter 6 using real data.

Two call center data sets are used in this thesis. Both data sets are from VANAD Laboratories, a call center in Rotterdam, The Netherlands. The first data set is used in Chapters 3 and 4. The second data set is used in Chapter 6. Detailed descriptions of the first and the second data set can be found in Chapter 3 and Chapter 6, respectively.

Chapter 2: Optimal forecasts and staffing

In this chapter, we investigate the optimal error measurement in call center forecasting. We consider a model where planners make initial staffing decisions in advance, and they make real-time adjustments or do traffic management on staffing on the day itself. The question we address is what is the optimal forecasting error measurement and optimal initial staffing rules, which minimize the sum of inital staffing and traffic management costs. It is shown that the weighted sum of expected quantile errors is the asymptotic optimal error measurement in a call center model under arrival rate uncertainty where staffing costs consist of initial staffing costs and traffic management costs. If the costs are symmetric for over- and understaffing, such an error measurement is equivalent to the sum of absolute forecasting errors or WAPE. Moreover, it is shown that staffing should occur according to a certain quantile of the distributional forecast rather than the mean.

My main contributions in this chapter are giving the proof of theorems and calculating the numerical examples to illustrate the performance of the proposed method.

Chapter 2 is based on Ding and Koole (2015).

Chapter 3: Estimating true demand in call centers

In this chapter, we discuss the redial and the reconnect behaviors in call centers and analyze how both behaviors influence the variability of the total number of arrivals in call centers. We show that without making a distinction between the redials, reconnects and the fresh calls, one might make inaccurate forecasts. However, accurately estimating the number of redials and reconnects is difficult in practice, due to the lack of caller identity information in the data. We therefore propose a method to estimate them. The method is validated via simulation and real data.

My main contributions in this chapter are building the framework of the estimation models, including the ones with and without seasonality, making simulations and using real-data to validate the model.

Chapter 3 is based on Ding et al. (2013) and Ding et al. (2015a).

Chapter 4: Fluid approximation for a model with redial and reconnect

In this chapter, a staffing algorithm is developed for the call center model with redials and reconnects. This algorithm makes use of the fluid approximation method to approximate the redial and the reconnect rates. Then the sum of redial rate, reconnect rate and the fresh arrival rate are used as an input for the Erlang A formula to approximate the service levels.

My main contributions in this chapter are motivating and validating the assumptions we make in the queueing model, assisting on giving the proof of the theorems, and making all numerical calculations and simulations.

Chapter 4 is based on Ding et al. (2015d).

Chapter 5: A call center model with a call-back option

In this chapter, we study a model with a so-called call-back option, where callers that wait longer than a certain threshold will be disconnected and they will call back at a later moment. Because of the particularities of this model (i.e., whether a call disconnects or not depends on the waiting time, rather than the number of callers waiting in the queue), the traditional approach of numerically solving a Markov process is not possible. Therefore, we apply the technique of discretization of the waiting time of the first caller in line, and subsequently, derive the first order performances of the model via value iteration. We show that this technique offers very accurate approximations to the real system. Furthermore, the model with a call-back option is shown to be more efficient in the sense that by having the same number of agents, callers on average experience much shorter waiting compared to those in the $M/M/s$ queueing model.

My main contributions in this chapter are proposing method to derive the performance metrics from the value iteration results (together with other co-authors), and programing the code for value iteration and numerical validations.

Chapter 5 is based on Ding et al. (2015c).

Chapter 6: Validation of call center models

Often in the literature, researchers and practitioners develop models with certain assumptions to represent real situations in call center operations. The validity of these models is rarely verified. In this chapter, we validate some commonly used models and assumptions for multi-skill call centers, by comparing the service levels predicted by models with those from real data. The comparison results suggest that ignoring some features, such as the agent breaks and agent heterogeneity, leads to large errors. Furthermore, data analysis shows significant amounts of fluctuations in the handling times of each day, which is partially explained by agent heterogeneity and agent learning. We build a model to predict the average handling time of each day, and such a model is validated. The model successfully explains up to 54% of the handling time variability.

My main contributions in this chapter are making all the data analysis, making simulations and comparison of all models.

Chapter 6 is based on Ding et al. (2015b).

Chapter 2

Optimal forecasts and staffing

We formulate the staffing problem in call centers as a newsvendor type problem, where the costs are the initial staffing costs, plus the traffic management costs. Under such a cost structure and the arrival rate uncertainty, the optimal forecasts and staffing levels are derived. We show that the optimal staffing should occur according to a quantile of the distributional forecast, rather than the mean. It is also shown that the errors in staffing are approximately linear in the forecasting errors. This leads to the conclusion that the weighted sum of expected quantile errors should be the error measurement in call center forecasting, since minimizing it minimizes the staffing costs. In special cases where the costs are symmetric for over- and understaffing, this is equivalent to minimizing the sum of absolute forecasting errors (or, equivalently, WAPE). We use numerical examples to show that the reduction in staffing costs can be substantial when one makes staffing decisions according to the optimal percentile instead of the mean.

2.1 Introduction

In call centers, short-term call volume forecasts are often made a few weeks or months in advance (Koole (2013)). We refer to these forecasts as the initial forecasts. Once the initial forecasts are made, the initial staffing is determined, often using the Erlang C or Erlang A formula (Aksin et al. (2007), Gans et al. (2003)). However, some factors are still unknown by the time that the forecasts are made, such as the weather and the effect of advertisement campaigns. Those factors will influence the demand. Therefore, the uncertainties in those factors lead to uncertainties in the future (in a few weeks' time) arrival rates. To avoid large deviations from the required performance, real-time adjustments of forecasts are made just before, and mostly within the day itself (Cleveland and Mayben (2000)). Due to the availability of more recent data, these updated forecasts have a higher accuracy (Taylor (2012)). Based on these updated forecasts, staffing levels are adjusted accordingly. Making these adaptations is called traffic management, and it exploits certain types of flexibility in agent schedules and task assignments. Traffic management is often done at multiple moments, most importantly during the day itself, or just before. In this chapter, we study the initial forecast and staffing level that minimizes the sum of staffing and traffic management costs, taking arrival rate uncertainty into account.

Arrival rate uncertainty has been well studied in the call center literature. As shown by Jongbloed and Koole (2001), the arrival processes in call centers show significant overdispersion, i.e., the variance is much higher than the mean of the number of arrivals in each interval, beyond the explanation of the homogeneous Poisson process. They propose a Poisson mixture model, where the arrival rate of the Poisson process is assumed to be a random variable. Different demand estimation or forecasting models have be been proposed to address such arrival rate uncertainty in call centers (see Avramidis et al. (2004a), Aldor-Noiman et al. (2009), Ye et al. (2014)). In the literature as well as in practice, researchers and practitioners usually take two steps to derive the initial staffing level. In the first step, they make distributional forecasts using parametric forecasting models. The models or parameters are selected or computed such that certain measurement of the errors between the mean of the distributional forecasts and the realizations is minimized. Such error measurements include MSE (mean squared error), RMSE (root mean squared error), MAPE (mean absolute percentage error), SAD (sum of absolute deviations), WAPE (weighted absolute percentage error), etc. For example, Shen and Huang (2008) use MSE as the error measurement to update intraday arrivals. Ibrahim and L'Ecuyer (2013) compare the MAPE, RMSE and MSE of different forecasting models. Aldor-Noiman et al. (2009) compare the MAPE and RMSE of four fixed effects models. In the second step, after having obtained the distributional forecasts, staffing decisions are made based on the mean of the distributional forecast. For example, Aldor-Noiman et al. (2009) and Brown et al. (2005) both use the mean arrival load of the system to generate staffing levels by applying staffing formulas, such as the Erlang A, the Erlang C or square-root staffing.

We show in this chapter that both steps are arguable and might lead to sub-optimal decisions. In the first step, the model selection depends on the error measurement one chooses, and it is often the case that if one chooses a different error measurement, it leads to a different model choice (see for example, Ibrahim and L'Ecuyer (2013), Aldor-Noiman et al. (2009), Shen and Huang (2008)), which eventually leads to different forecasts. Therefore, it is important to know which error measurement one should use, and what its corresponding forecasts are. In the second step, we show that under a reasonable cost structure, staffing according to the mean arrival rate is not the optimal staffing decision. This is especially the case when the arrival rate is assumed to have a skewed distribution (Whitt (1999), Jongbloed and Koole (2001), Taylor (2012), Avramidis et al. (2004a), Steckley et al. (2009)) or when over- and understaffing have different costs (Liao et al. (2012)).

The organization of this chapter is as follows. We show, in Section 2.2, that under arrival rate uncertainty and a cost structure that includes traffic management costs, the staffing problem can be modeled as a newsvendor problem, using quantiles of the arrival rate distribution. This contradicts the traditional approach, which suggests staffing according to the mean arrival rate. To prove this we assume monotonicity of the staffing function. In Section 2.3 we prove this monotonicity for the $M/M/s+G$ model. This is our second contribution. Finally, in Section 2.4, we show that the sum of quantile errors is the optimal error measurement. It is optimal in the sense that minimizing the expected sum of quantile errors leads to the asymptotically optimal forecasts. Furthermore, we show that in special cases where over- and understaffing have the same costs, the WAPE (or equivalently, SAD) is the optimal error measurement. We numerically evaluate the impact of our methods in Section 2.5. The impact is shown to be significant, i.e., there is significant amount of costs reduction when one minimizes WAPE instead of minimizing SSE.

2.2 Cost-optimal staffing

In this section, we first describe the staffing costs in call centers, which consist of the initial staffing costs and the traffic management costs. We show that under such a cost structure, the staffing problem can be modeled as a newsvendor problem. We then derive the cost-optimal staffing under the arrival rate uncertainty and this cost structure.

We assume throughout the following simplified call center staffing process. Call center arrivals can be seen as coming from a Poisson process with parameter λ. Staffing has to be done weeks in advance, at which moment the actual value of λ is still unknown. Therefore we have to forecast λ. Forecasts can take different forms; we will use the one that gives full information, i.e., a distributional forecast. Thus our forecast takes the form of a random variable, Λ. On the basis of Λ we decide on a staffing level, s.

There is one moment at which we do traffic management. We assume that at this moment we know the actual arrival rate λ. Based on this arrival rate we adapt our staffing level to the right staffing level $S(\lambda)$, independent of the initial level $s \in \mathbb{N}$, where S is a function of λ that determines the minimal staffing level such that a certain SL requirement is met. Adapting the staffing level is more expensive than staffing the right level initially. We assume that the initial staffing costs are c per agent, and the costs of overstaffing (i.e., more agents scheduled in the initial staffing than needed) are $(c_o - c)$ per agent, and the costs of understaffing (i.e., more agents are needed) are $(c_u + c)$ per agent. Then, for the initial staffing level s and realization λ, the total staffing costs $C(s, \lambda)$ are the sum of the initial staffing costs and the traffic management costs:

$$\begin{aligned} C(s,\lambda) &= cs + (c_o - c)(s - S(\lambda))^+ + (c_u + c)(S(\lambda) - s)^+ \\ &= cS(\lambda) + c_o(s - S(\lambda))^+ + c_u(S(\lambda) - s)^+, \end{aligned}$$

where $y^+ := \max\{0, y\}$.

The total expected costs are

$$\mathrm{E}C(s,\Lambda) = c\mathrm{E}S(\Lambda) + c_o\mathrm{E}(s - S(\Lambda))^+ + c_u\mathrm{E}(S(\Lambda) - s)^+. \tag{2.1}$$

The cost-optimal staffing $s^* := \arg\min_s \mathrm{E}C(s, \Lambda)$ can then be found by

$$s^* = \arg\min_s \left\{c_o\mathrm{E}(s - S(\Lambda))^+ + c_u\mathrm{E}(S(\Lambda) - s)^+\right\}. \tag{2.2}$$

Equation (2.2) has the form of the newsvendor problem, with the demand replaced by $S(\Lambda)$. Note that Equation (2.2) does not necessarily lead to integer solutions, however, this can be easily solved by rounding the result to the nearest integers above and below, then evaluating both solutions. The same technique can be applied to the rest of this chapter, as we neglect the integer constraint to simplify our notation.

Therefore, if we denote with F_S the cdf of the random variable $S(\Lambda)$, then, by applying the results of the newsvendor problem, we solve problem (2.2), and we obtain

$$s^* = F_S^{-1}\Big(\frac{c_u}{c_o + c_u}\Big),$$

where F_S^{-1} is the *quantile function* of $S(\cdot)$. For any cdf F, its quantile function is defined by $F^{-1}(y) := \inf\{x \in R : F(x) \geq y\}, 0 \leq y \leq 1$.

We assume that S is a non-decreasing function. This is a natural assumption: when there are more arrivals, then we need more agents to obtain the required service level. We can show that S is non-decreasing for a number of often-used models and performance measures; see the next section.

Theorem 2.1. *If $S(\cdot)$ is non-decreasing, then*

$$s^* = S\Big(F_\Lambda^{-1}\Big(\frac{c_u}{c_o + c_u}\Big)\Big).$$

Proof. It suffices to show that $H^{-1}(p) = S(F_\Lambda^{-1}(p))$ for any $0 \leq p \leq 1$. To this end, let $\lambda_p := F_\Lambda^{-1}(p)$. Due to the properties of the quantile function, we have

$$F_\Lambda(\lambda_p) \geq p.$$

Furthermore, we have

$$\mathrm{P}(S(\Lambda) \leq S(\lambda_p)) \geq \mathrm{P}(\Lambda \leq \lambda_p) \geq p,$$

which leads to

$$F_S(S(\lambda_p)) \geq p,$$

from which it follows that $S(\lambda_p) \in \{x \in R : H(x) \geq p\}$. Due to the definition of $F_S^{-1}(p)$, we have

$$F_S^{-1}(p) \leq S\left(F_\Lambda^{-1}(p)\right).$$

Assume $F_S^{-1}(p) < S\left(F_\Lambda^{-1}(p)\right)$. Then, we can always find some $\epsilon > 0$, such that $S\left(F_\Lambda^{-1}(p)\right) = F_S^{-1}(p) + \epsilon$. Moreover, we define $B := \{\lambda \in R : S(\lambda) = F_S^{-1}(p)\}$, and $\lambda' := \sup B$. Clearly, $B \neq \emptyset$. Therefore, under such assumptions, $\lambda' < F_\Lambda^{-1}(p)$ would be true, due to the fact that $S\left(F_\Lambda^{-1}(p)\right) = F_S^{-1}(p) + \epsilon > S(\lambda')$ and S being a non-decreasing function.

$H(F_S^{-1}(p)) \geq p$ must hold, because H is a cdf. Now we show that $F_S(F_S^{-1}(p)) \geq p$ contradicts $\lambda' < F_\Lambda^{-1}(p)$. If $F_S(F_S^{-1}(p)) \geq p$, then due to the definition of λ', we must have

$$p \leq \mathrm{P}(S(\Lambda) \leq F_S^{-1}(p)) = \mathrm{P}(\Lambda \leq \lambda'). \tag{2.3}$$

Inequality (2.3) leads to $F_\Lambda(\lambda') \geq p$, which contradicts with the fact that $\lambda' < F_\Lambda^{-1}(p)$. □

Theorem 2.1 proves that staffing according to the $c_u/(c_o + c_u)$ quantile of the arrival rate distribution minimizes the expected staffing costs. In the special case of $c_o = c_u$, staffing according to the median of Λ is optimal. This means that staffing according to the mean, which is often done, is not optimal, not even in the symmetric case, unless the mean is equal to the median.

In practice it is often simpler to scale up than to scale down. Scaling up is often done by hiring flexible workers, who are often available on a short notice, especially when they work at home. Scaling down is sometimes not even possible, in which case $c_o = c$ and at the initial level staffing should be done very conservatively. Next, many call centers have different layers of flexibility. First, flexibility is sought in the task assignment. If this is not sufficient then the number of agents is changed. This leads to increasing costs of up and downscaling, giving a piece-wise linear costs function C in s. In general, there is no closed-form solution for s^*. A solution can be found numerically by calculating $\mathrm{E}C(s, \Lambda)$ for various values of s.

It is worth mentioning that the form of the newsvendor problem also arises in the model studied by Bassamboo et al. (2010). They consider a call center staffing problem where the objective is to minimize the sum of expected waiting costs, abandonment costs and the staffing costs. Different from Bassamboo et al. (2010), in this chapter, we minimize the traffic management costs while the service level constraint is met.

2.3 Monotonicity of the staffing function

For some yet unspecified call center model we write the (expected) performance as $P(s, \lambda)$. We assume there is some maximal allowable performance level τ. Then S can be written as $S(\lambda) = \inf\{s | P(s, \lambda) \leq \tau\}$. Examples are ASA and its tail probability (SL) in the Erlang C model. In the case of ASA we take P equal to the expected waiting time; in the case of SL we take $P = \mathrm{P}(W > t)$ with W the stationary waiting time in the queue and t the AWT. In the case of abandonments we have to decide how abandonments are integrated in the performance measures. The abandonment % or rate now becomes important, but we also have to decide how the abandonments are accounted for in the measures that are functions of the waiting time. Two regular choices are the time in queue W and the *virtual* or *offered waiting time* V (the waiting time of a *test customer* with ∞ patience), but other choices are possible (see Jouini et al. (2013)). The extension of the Erlang C model to general patience distributions is written as $M/M/s + G$ model. An overview of results for this model can be found in Section 9 of Zeltyn and Mandelbaum (2005). The special case $M/M/s+M$ is known as the Erlang A model.

Theorem 2.2. *The staffing function S is non-decreasing for the $M/M/s+G$ model and any τ and P given by $P(W>t)$, $P(V>t)$, EW, EV, or the abandonment rate.*

Proof. If P is non-decreasing in s and non-increasing in λ then S is non-decreasing. That P is non-decreasing in s can be found for all performance measures in Section 2.1 of the online appendix of Zeltyn and Mandelbaum (2005). To show that P is non-increasing in λ for the different performance measures it suffices to show it for $\mathrm{P}(V>t)$: all other results follow directly from that.

We introduce the following notation, slightly adapted from Zeltyn and Mandelbaum (2005):

$$\begin{aligned} G'(x) &= \int_0^x (1-G(u))du, \\ J_\lambda(t) &= \int_t^\infty \mathrm{e}^{\lambda G'(x)-s\mu x}dx, \\ J_\lambda &= J_\lambda(0), \\ \varepsilon_\lambda &= \int_0^\infty \mathrm{e}^{-t}(1+\frac{t\mu}{\lambda})^{s-1}dt, \end{aligned}$$

where μ is the service rate and G is the cdf of the patience time. Then, according to Zeltyn and Mandelbaum (2005),

$$\mathrm{P}_\lambda(V>t) = \frac{\lambda J_\lambda(t)}{\varepsilon+\lambda J_\lambda}.$$

We now show that for fixed s, if $0>\lambda_1>\lambda_2$, then $\mathrm{P}_{\lambda_1}(V>t)\geq \mathrm{P}_{\lambda_2}(V>t)$, i.e.,

$$\frac{J_{\lambda_1}(t)}{\varepsilon_{\lambda_1}/\lambda_1+J_{\lambda_1}} - \frac{J_{\lambda_2}(t)}{\varepsilon_{\lambda_2}/\lambda_2+J_{\lambda_2}} \geq 0,$$

which is equivalent to showing that

$$\frac{J_{\lambda_1}(t)\varepsilon_{\lambda_2}/\lambda_2 - J_{\lambda_2}(t)\varepsilon_{\lambda_1}/\lambda_1 + J_{\lambda_1}(t)J_{\lambda_2} - J_{\lambda_2}(t)J_{\lambda_1}}{(\varepsilon_{\lambda_1}/\lambda_1+J_{\lambda_1})(\varepsilon_{\lambda_2}/\lambda_2+J_{\lambda_2})} \geq 0.$$

Because $J_\lambda(t)$ is increasing and ε_λ is decreasing in λ it is readily seen that $J_{\lambda_1}(t)\varepsilon_{\lambda_2}/\lambda_2 - J_{\lambda_2}(t)\varepsilon_{\lambda_1}/\lambda_1 > 0$. Because all its terms are ≥ 0 also $(\varepsilon_{\lambda_1}/\lambda_1+J_{\lambda_1})(\varepsilon_{\lambda_2}/\lambda_2+J_{\lambda_2}) > 0$. Thus, we only need to show that

$$J_{\lambda_1}(t)J_{\lambda_2} - J_{\lambda_2}(t)J_{\lambda_1} \geq 0.$$

Its proof is equivalent to that of Equation (2.4) in the online appendix of Zeltyn and Mandelbaum (2005). □

Note that the $M/M/s$ model is a special case of the $M/M/s+G$ model (with ∞ patience). This Theorem 2.2 holds also for the often used Erlang C model.

2.4 The optimal error measurement

In Section 2.2 we discussed how to compute the optimal staffing level based on a (distributional) forecast. However, in practice we only observe the realization of the number of in-coming calls. Therefore, the question is 'how can we measure the quality of the rate on which the original staffing was based?' Note that the quality measure is not a goal by itself, it should measure to which extent the objective, low traffic management costs, is met. Thus a high quality forecast according to our error measure should be equivalent to low costs. In this section we show that the weighted sum of errors is the asymptotical optimal forecasting error measurement, where the weighing depends on the sign of the error. In the special case $c_o = c_u$ this is equivalent to the WAPE. Thus minimizing the WAPE asymptotically minimizes the traffic costs.

In this chapter our traffic management is executed the moment we know λ. The realization of the Poisson distribution with rate λ is, at that moment, not known yet. It is shown in Roubos et al. (2012) that these Poisson fluctuations can have a considerable impact on the performance. For this reason, in practice, traffic management is also executed during the day. However, this effect is minor compared to the consequences of forecasting errors. To see this, it is important to realize that most forecasting models are multiplicative, with factors for the different seasonal components. Thus, for example, a 5% error in the day-of-the-week factor results in a 5% error in the daily volume. Because of this, forecasting errors are proportional to the volume and therefore grow linearly with λ; Poisson errors are sub-linear. Thus for larger volumes (in larger call centers and/or at an aggregated level) forecasting errors have bigger consequences.

The results in this section are based on two theorems. In the first we derive a simple approximation for $C(S(\hat{\lambda}),\lambda)$ in Theorem 2.3. This will later allow us to show the optimality of the error measurement. After that, in Theorem 2.4, we consider the Poisson fluctuations.

In Theorem 2.3 we would like to compare the costs for realization λ and forecast $\hat{\lambda}$ in the limit. To have them grow at the same time with a multiplicative error we assume $\hat{\lambda} = h\lambda$. We assume that $\lambda > 0$ and $h > 0$, thus $\hat{\lambda} > 0$. Finally, we use the following notation: $f(x) = o(g(x))$ if $\lim_{x\to\infty} f(x)/g(x) = 0$.

Theorem 2.3. *For the $M/M/s+G$ model, given performance constraints based on SL, ASA or abandonment rate,*

$$C(S(\hat{\lambda}),\lambda) = cS(\lambda) + (1-\gamma)\big(c_o(\hat{\lambda}-\lambda)^+ + c_u(\lambda-\hat{\lambda})^+\big)\beta + o(\lambda), \qquad (2.4)$$

for some $\gamma \geq 0$ which depends on the performance constraint and β the expected HT.

Proof. Consider first the case $h \geq 1$. Then $\hat{\lambda} \geq \lambda$ and, according to Theorem 2.2, $S(\hat{\lambda}) \geq$

$S(\lambda)$. Therefore,

$$C(S(\hat{\lambda}),\lambda) = cS(\lambda) + c_o(S(\hat{\lambda}) - S(\lambda)).$$

From Mandelbaum and Zeltyn (2009), Section 2, we know that $S(\lambda) = (1-\gamma)\lambda\beta + o(\lambda)$ and also $S(\hat{\lambda}) = (1-\gamma)\hat{\lambda}\beta + o(\hat{\lambda}) = (1-\gamma)\hat{\lambda}\beta + o(\lambda)$. Thus,

$$\begin{aligned} C(S(\hat{\lambda}),\lambda) &= cS(\lambda) + c_o((1-\gamma)\hat{\lambda}\beta - (1-\gamma)\lambda\beta + o(\lambda)) \\ &= cS(\lambda) + c_o(1-\gamma)(\hat{\lambda}-\lambda)\beta + o(\lambda). \end{aligned}$$

Similarly for $h < 1$. □

Remark 2.1. Depending on the performance objective, different operational regimes apply, with different limiting behavior. All are $o(\lambda)$, but in some cases stronger results are obtained: for the SL objective the limiting behavior is $O(\sqrt{\lambda})$ (with $f(x) = O(g(x))$ if $\limsup_{x\to\infty} |f(x)/g(x)| < \infty$). See Section 2 of Mandelbaum and Zeltyn (2009) for details.

Observing a single realization gives little evidence about the quality of a forecast. Therefore we consider I measurements. For $i = 1,\dots,I$, let λ_i be the realizations and $\hat{\lambda}_i$ the forecast on which the initial staffing was based. Note that $\hat{\lambda}_i$ might well be the percentile given in Section 2.2. Then the total costs $C_T(S(\hat{\lambda}),\lambda)$, with $\hat{\lambda}$ and λ I-dimensional vectors, are given by

$$C_I(S(\hat{\lambda}),\lambda) \approx \sum_{i=1}^{I} cS(\lambda_i) + (1-\gamma)\beta\sum_{i=1}^{I}\left(c_o(\hat{\lambda}_i - \lambda_i)^+ + c_u(\lambda_i - \hat{\lambda}_i)^+\right).$$

This value is minimized by the forecast that minimizes $\sum_{i=1}^{I}\left(c_o(\hat{\lambda}_i - \lambda_i)^+ + c_u(\lambda_i - \hat{\lambda}_i)^+\right)$, the weighted sum of errors. Note that in the symmetric case $c_o = c_u$ this reduces to minimizing $\sum_{i=1}^{I}|\hat{\lambda}_i - \lambda_i|$, which is equivalent to minimizing the WAPE. In call centers often the MAPE is used. However, this gives too much weight to the intervals with less volume, which are actually harder to predict because of the Poisson variability, which we study next.

In this chapter we focus on the costs relative to forecasting errors. In practice traffic management is also done to counter the variability in the "Poisson noise". The following theorem shows that this effect is minor compared to forecasting errors, especially for larger systems.

Theorem 2.4. *$E|N_\lambda - \lambda| = O(\sqrt{\lambda})$ for $N_\lambda \sim$* Poisson (λ).

Proof. Jensen's inequality states that $\phi(\mathrm{E}X) \leq \mathrm{E}\phi(X)$ for ϕ convex. By taking $\phi(x) = x^2$ and $X = |N_\lambda - \lambda|$ it follows that $\mathrm{E}|N_\lambda - \lambda| \leq \sqrt{\lambda}$. □

2.5 Numerical evaluation

In this section, we illustrate our results numerically. We start by comparing staffing according to Theorem 2.1 and the usual staffing based on the expected forecast. We consider two situations: a very regular case, and a more asymmetric (but far from unrealistic) case. In both we use the Erlang A model based on a regular 80% within 20 seconds SL requirement based on the virtual waiting time. The AHT is 4 minutes, the average patience is 5 minutes, $c = 1$. The other parameters and results can be found in Table 2.1. Define

$$s_a := S(\mathrm{E}\Lambda),$$

and

$$s_n := S\big(F_\Lambda^{-1}(\frac{c_u}{c_o + c_u})\big).$$

The results are obtained by simulating Λ and calculating the Erlang A values (we have non-integer values by interpolating between the integer values). Note that the values for the lognormal distribution are those of Λ, not those of the normal distribution from which the lognormal is constructed. The differences in the examples are relatively small, but this is related to the choice of the parameters and the variability of Λ.

Next we study the relation between $|S(\hat{\lambda}) - S(\lambda)|$ and $\hat{\lambda} - \lambda$, in relation to Theorem 2.3. We consider again the Erlang A model. The plots are shown in Figure 2.1 with the black dots representing the exact results obtained from the Erlang A formula, and the red lines are obtained via Theorem 2.3. We show two graphs, both with 80/20 SL and an AHT of 1 minute. The error in number of agents is very close to the linear approximation given in Theorem 2.3.

Λ	$\mathrm{E}\Lambda, \sigma(\Lambda)$	c_u, c_o	s_a	s_n	$\mathrm{EC}(s_a, \Lambda)$	$\mathrm{EC}(s_n, \Lambda)$
normal	20, 2	0.2, 0.1	82.2	85.5	82.8	82.7
lognormal	20, 4	0.1, 1	82.2	62.8	88.2	83.8
lognormal	20, 4	1, 0.1	82.2	103.8	87.6	84.5

Table 2.1: Staffing based on newsvendor model vs. average forecast in the Erlang A model.

2.6 Conclusion

In this chapter, we considered the problem of finding the optimal staffing level and forecasting error measurement in call centers, where the staffing costs are the initial staffing costs plus the traffic management costs. We showed that the staffing problem can be formulated as a newsvendor problem with the optimal initial staffing level occurring at the $c_u/(c_o + c_u)$ quantile of the arrival rate distribution. Furthermore, we derive that for the standard performance constraints the staffing cost is linear in the arrival rate error.

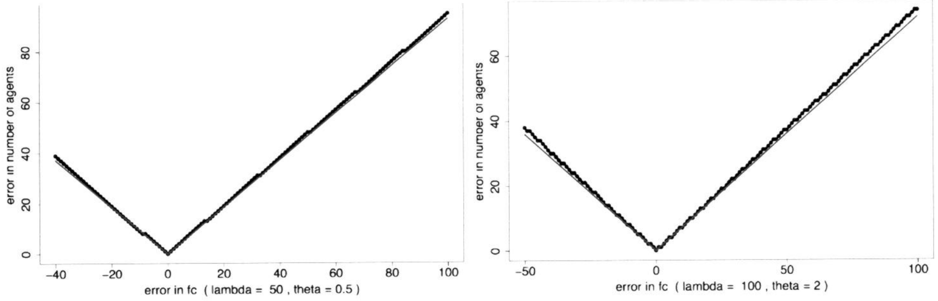


Figure 2.1: Error in number of agents as a function of the error in forecast for two cases.

This led to the conclusion that the sum of expected quantile errors is the asymptotically optimal forecasting error measurement, since minimizing it leads to minimizing costs. This is equivalent to minimizing WAPE in case $c_o = c_u$. Further numerical results show that using the sum of quantile errors as the error measurement works extremely well even for medium sized call centers, and the gap between the estimator and the real optimum is very small.

As for further research, the following possible improvements and extensions can be made. Especially the traffic management model is quite simple. Extensions could include agent shifts and piecewise linear costs representing different layers of flexibility.

Chapter 3

Estimating true demand in call centers

In practice, in many call centers customers often perform *redials* (i.e., reattempt after an abandonment) and *reconnects* (i.e., reattempt after an answered call). In the literature, call center models usually do not cover these features, while real data analysis and simulation results show ignoring them inevitably leads to inaccurate estimation of the total inbound volume. Therefore, in this chapter we propose a performance model that includes both features. In our model, the total volume consists of three types of calls: (1) *fresh calls* (i.e., initial call attempts), (2) *redials*, and (3) *reconnects*. In practice, the total volume is used to make forecasts, while according to the simulation results, this could lead to high forecast errors, and subsequently lead to wrong staffing decisions. However, most of the call center data sets do not have customer-identity information, which makes it difficult to identify how many calls are fresh and what fractions of the calls are redials and reconnects.

Motivated by this, we propose a model to estimate the number of fresh calls, and the redial and reconnect probabilities, using real call center data that has no customer-identity information. We show that these three variables cannot be estimated simultaneously. However, it is empirically shown that if one variable is given, the other two variables can be estimated accurately with relatively small bias. We show that our estimates of redial and reconnect probabilities and the number of fresh calls are close to the real ones, both via real data analysis and simulation.

3.1 Introduction

In an inbound call center, a manager typically uses historical call data sets to forecast the future call volumes. Based on the call volume forecast, one can make staffing decisions. An inaccurate forecast inevitably leads to inaccurate staffing decisions (see Steckley et al. (2009)). There is extensive literature on different forecasting methods applied to call centers. Andrews and Cunningham (1995) use the Autoregressive Integrated Moving Average (ARIMA) method to forecast the inbound call volume of the L. L. Bean's call center. Taylor (2012) adjusts the traditional Holt-Winters exponential smoothing method to the Poisson count model with gamma-distributed arrival rate, and takes both intraweek and intraday patterns into account in his model. Taylor (2008) compares the accuracy of a few forecasting models for a British retail bank call center. He concludes that for forecasting horizons up to two or three days ahead, seasonal ARIMA and Holt-Winters model are more accurate, while for longer lead times, simple historical average is more accurate. Shen and Huang (2008) use the Singular Value Decomposition (SVD) method to reduce the dimension of square-root-transformed call center data. Then they apply time series and regression analysis techniques to make distributional forecasts. Besides the forecasts, they have also developed a method to dynamically update the forecasts when early realizations of the day are given. The doubly stochastic model built by Jongbloed and Koole (2001) addresses the issue of high variability in call arrival volume. This model has been further developed in Avramidis et al. (2004a), where three variants of doubly stochastic model are analyzed and compared. Channouf and L'Ecuyer (2012) proposed a normal Copula model where they show empirically that such a model performs better compared to the models in Avramidis et al. (2004a), in the sense that when trying to fit real data, they achieve better fit of the correlation and the coefficient of variation. Ibrahim and L'Ecuyer (2013) add the correlation between different call types into a model with additive seasonality, interday correlation and intraday correlation. A multiplicative way to model the intraweek and intraday pattern is used by Gans et al. (2012). For a comprehensive literature study on call center forecasting, we refer to Ibrahim et al. (2016b).

Call center forecasting models aim to achieve the minimum error in the forecasts, where total inbound volumes are used. In this chapter, we show that the fresh volume is more appropriate to be used when one makes forecasts, since it is independent of the service levels, the number of agents and other factors in the call center. In contrast, the total inbound volumes are influenced by the service levels and staffing decisions of the call centers, due to the redial and reconnect customer behaviors. Data analysis of a real call center reveals that a significant fraction of the inbound call volume involves redials and reconnects. The reason for customers to redial is clear, since abandoned customers did not get their questions answered in their initial attempts. There are several reasons for customers to reconnect. For example, a customer may check the status of his previous request. Also, solutions offered by agents may not be effective for customers, hence, they may reconnect. Koole (2013) gives more insights on redials and reconnects.

To identify the fresh volume, one would need customer-identity information in the data set, such that redials and reconnects can be filtered out. However, in most of the call center data sets, customer-identity information is either not recorded or not accessible, i.e., we do not know who is the caller of each call. In other words, we do not know whether a call is a fresh call, a redial, or a reconnect. Furthermore, the fresh volume is not stable due to the existence of seasonality and trend. On the other hand, the redial and reconnect probabilities are less influenced by those effects, thus, they are more stable over time, since they represent the customer behaviors. In this chapter, we will show how to estimate the number of fresh calls with the assistance of the redial and reconnect probabilities.

Besides the fact that estimating redial and reconnect probabilities is crucial in estimating the fresh volume, estimating both probabilities themselves is also interesting. Much scientific effort has been spent on analyzing the performance of queueing systems with retrial behaviors (see Artalejo and Pozo (2002), Falin (1995) and the references therein). Some retrial models are developed for call centers, e.g., Stolletz (2008), Mandelbaum et al. (2002), Aguir et al. (2004, 2008). The reconnect customer behavior is first mentioned in Gans et al. (2003) as revisit. In service industry, it is referred to as feedback or re-entrant (Yom-Tov and Mandelbaum (2014)). In Yom-Tov and Mandelbaum (2014), the authors consider a queueing model to represent hospitals where patients might return to service several times, and they apply fluid and diffusion approximations to develop some staffing principles to support healthcare staffing. However, customer abandonment is not included in their model. In all the existing works mentioned above, it is assumed that the retrial or the reconnect probability is known, whereas it can be difficult to calculate in practice.

Hoffman and Harris (1986) are the first ones who address the issue that the total volume does not represent the true demand in call centers. Aiming to have a more accurate forecast for the call volume, they estimate the redial probability for the U.S tax-payer service telephone center. However, Hoffman and Harris (1986) only consider the redial behavior, and they neglect the reconnect behavior. Also, the fresh call arrival rate is assumed to be a constant among certain hours of the day in their model, whereas in most call centers the arrival rate is far from constant over the day, exhibiting a certain intraday pattern, see Shen and Huang (2008), Gans et al. (2012), Ibrahim and L'Ecuyer (2013). In this chapter, we propose a queueing model that has two extra orbits compared to the Erlang C model, where abandoned customers redial via one orbit, and answered customers reconnect via the other orbit. We show that these two extra orbits cannot be ignored, otherwise it will lead to inaccurate estimation of the total arrival volume, and thus inaccurate staffing decisions. Having developed and validated the queueing model, we then estimate the fresh volume, the redial and reconnect probabilities. This estimation problem is formulated as an optimization problem, where the minimum objective value is attained when the actual redial and reconnect probabilities are chosen. We show that these three variables cannot be accurately estimated simultaneously. Nevertheless, if one variable is given, it is verified numerically that the other two variables can be estimated accurately with small relative bias. To allow intraweek seasonality, we adjust our model to a linear programming problem, which is easy to solve. We show both via simulated data and real call center data that our estimates are close to the real values.

The remainder of this chapter is organized as follows. In Section 3.2, we motivate the necessity of making distinction between redials, reconnects and fresh calls. Specifically, we show simulation examples of such a model to understand the influence of redials and reconnects on the total volume. In Section 3.3, we present our estimation models both for constant arrival rate and arrival rate with intraweek seasonal patterns. These estimation models are validated in Section 3.4 via simulation as well as real call center data sets.

3.2 Motivation

According to the description of the call center model with redial and reconnect illustrated by Figure 1.8, the total volume is influenced by the service level, since a bad service level leads to more abandonments, which in turn leads to a larger number of redials. In this way, the total call volumes depend on the staffing decisions. To illustrate this, consider the following example. We set the fresh arrival rate to be 10 calls per minute every day, and the mean HT is set to be four minutes. Since the sum of independent Poisson random variables is again Poisson distributed, F_i (i.e., the fresh volume in day i) is then Poisson distributed with rate $10 \cdot 60 \cdot 24 = 14400$. We take B, H, Γ_{RD} and Γ_{RC} to be exponentially distributed. The total call volume and fresh call volume of each day are plotted in Figure 3.1 for a 100-day time interval. In this example, we set $p = 0.5$ and $q = 0.2$. The number of agents varies per day, and is drawn from a Poisson distribution with mean 43, which is slightly above the fresh arrival load per time unit, i.e., fresh arrival rate times the mean HT. To conduct this simulation, one does not need to assume the number of agents being Poisson distributed; we make this assumption merely to model the fact that the staffing level changes each day in call centers, which is caused by several reasons, such as call centers having different shifts for different weekdays and agents' absenteeism. In the simulation, we generate a call center data set of 100 days.

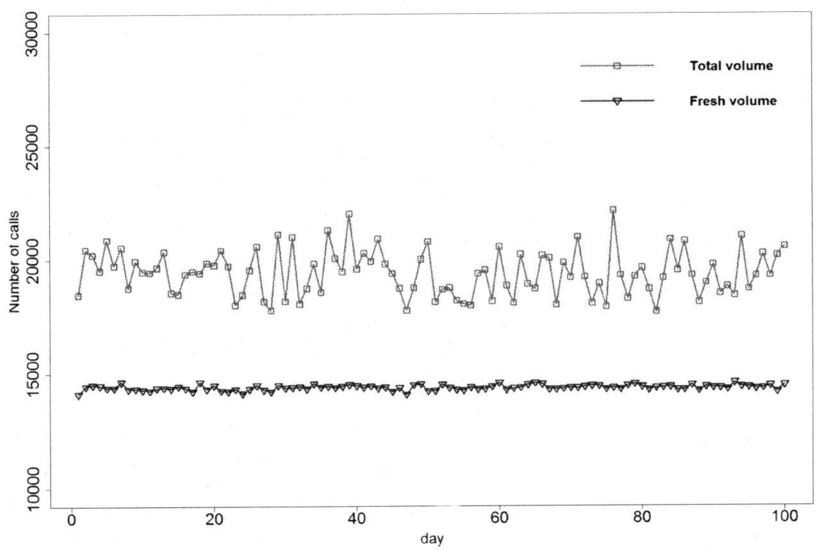


Figure 3.1: Call volumes from a simulation example.

Interestingly, Figure 3.1 shows that not only the total volumes are much higher than the fresh volumes (as they should), but also that they exhibit much higher variability. If a manager were able to see the fresh volume, it would be easy to predict future call volumes, since they are just Poisson realizations with constant rate. However, since the manager cannot identify who the caller is in the data set, he will only see the total volume in the data set. Figure 3.1 shows that the total volume depends on the staffing decisions and thus is highly volatile, which is due to the fact that both the number of redials and reconnects are influenced by the staffing decisions. In contrast, the fresh volume has less variability and is independent of the staffing decisions, which should be used to make forecasts.

In practice, managers usually use the total arrival counts to make forecasts and staffing decisions. Merely having less volatility may not be sufficient to convince them to use the fresh volumes. Naturally, the following questions may arise: 'Why is it important to distinguish between the fresh calls, redials and reconnects?', 'Is knowing the total volume not enough?', 'Are there more reasons to use the fresh volume besides being less volatile?'

To address these questions, we simulate two call centers, called CC1 and CC2, whose parameters are shown in Table 3.1. CC1 represents the case *without* redials and reconnects, while CC2 is its counterpart *with* redials and reconnects, constructed such that the total call volumes are the same. Both call centers have the same service level requirement, with $\text{SL}_2 \leq 80\%$ and AWT being 20 seconds and $r \leq 10\%$. In CC1, we let $p = q = 0$, and it receives 5669 fresh calls in a day. With $s = 40$, it achieves $\text{SL}_2 = 80.1\%$ and $r = 9.6\%$. In CC2, we let $p = 0.5$ and $q = 0.2$, and the fresh arrival rate per day to be 3700. With $s = 20$, it receives also 5669 total calls and achieves $\text{SL}_2 = 21.9\%$ and $r = 4.9\%$, which is far from achieving the service level requirement. Other parameters such as μ and θ are identical in both call centers. Assume that the manager in CC2 wants to add as little agents as possible such that CC2 reaches its service level requirement, which is very close to SL_2 and r in CC1 in this case. Without making a distinction between the fresh calls, redials and reconnects, the manager uses the total volume to make forecasts. We assume that he simply uses the previous observation as the forecasts for the next day, which is also 5669. Consequently, to achieve the same SL_2 and r as in CC1, the manager in CC2 derives that $s = 40$, since all parameters besides p and q are the same for CC1 and CC2. In the third row of Table 3.1, one can see that the SL_2 is far beyond the 80% and r is far less than 10% in CC2 by letting $s = 40$. This means that staffing 40 agents for CC2 causes overstaffing and hence generates unnecessary staffing costs. Moreover, one can see that the realization for the next day is 4628, which is far from the original forecasts. The large forecasting error in this example is caused by not distinguishing the fresh volume from the total volume rather than using the wrong forecasting method. We could easily construct other examples to show that for other forecasting methods, such as ARIMA or exponential smoothing, large forecasting error may still exist. This means that not differentiating the fresh volume from the total volume can lead to large error in forecasts as well as in staffing decisions. In summary, this example emphasizes the necessities of knowing the fresh volumes, as well as using it in making operational decisions in call centers. When we say that the fresh volume represents the true demand, it is actually a subjective claim. People could also choose to claim that the total volume represents the true demand. However, this would make the demand more complicated rather than a simple number. For example, if one uses the total volume to represent the demand, and assumes that the demand is to be 100, then the number 100 is ambiguous, i.e., is the total volume being 100 obtained by staffing 20 agents or 40 agents; is the service level low or high when we receive 100 calls; when we have different HT, would the

	p	q	F	s	Total arrivals	F	SL$_2$	r
CC1	0	0	4	40	5669	5669	80.1%	9.6%
CC2	0.5	0.2	2.6	20	5669	3705	21.9%	49.0%
CC2	0.5	0.2	2.6	40	4628	3689	94.8%	0.7%

Table 3.1: Two simulation results, $\mu = 1/10, \theta = 1/2, \delta_{RD} = 20, \delta_{RC} = 50$.

demand still be 100? These complications or questions will not arise if we use the fresh volume to represent the demand.

3.3 Estimation model

Many call center data sets are similar to the simulated data set we generated in the previous section: customer identity information is not available. Thus, in such call center data sets for N days, we would only know A_i and C_i $(i = 1, \ldots, N)$, which stand for the number of abandoned calls in day i and the number of connected calls in day i, respectively. We denote T_i as the total number of calls in day i, and $r_i := A_i/T_i$, which is the abandonment percentage of day i.

To estimate F_i, p and q, we start with the simple case where $F_i \sim \text{Pois}(\lambda_F)$, i.e., each day has the same arrival rate of fresh calls, but we do not know how big λ_F is. Note that, by this assumption, we ignore the intraweek arrival pattern in the call center data set. We will extend our model to address this pattern in subsection 3.3.2. For the rest of this chapter, we refer to $\hat{p}, \hat{q}, \hat{\lambda}_F$ as estimated values of p, q and λ_F by using our model, respectively, and p^*, q^* and λ_F^* as the true values of p, q and λ_F, respectively.

3.3.1 Basic setup

By definition, we know that an inbound call can either be a fresh call, a redial or a reconnect. Hence, the following equation holds

$$T_i = F_i + RC_i + RD_i, \tag{3.1}$$

where RD_i and RC_i are the number of redials and reconnects in day i, respectively. $RD_i \sim B(A_i, p^*)$, $RC_i \sim B(C_i, q^*)$, where $B(k, p)$ stands for the binomial distribution with parameters k and p. If we let $F_i = \lambda_F + \epsilon_i$, $RD_i := A_i p^* + e_i$, and $RC_i := C_i q^* + \eta_i$, then Equation (3.1) can be rewritten as

$$T_i = \lambda_F + A_i p^* + C_i q^* + \epsilon_i + e_i + \eta_i. \tag{3.2}$$

Also, since a call is either answered or abandoned, we know that $T_i = A_i + C_i$. Inserting this equation into Equation (3.2), we obtain

$$(1 - p^*)A_i + (1 - q^*)C_i - \lambda_F = \epsilon_i + e_i + \eta_i. \tag{3.3}$$

For given data points $A_1, \ldots, A_N$, and $C_1, \ldots, C_N$, we consider the following minimization problem to estimate p, q and λ_F,

$$(\hat{p}, \hat{q}, \hat{\lambda}_F) = \underset{0 \leq p,q<1,\lambda_F}{\operatorname{argmin}} \sum_{i=1}^{N} |(1-p)A_i + (1-q)C_i - \lambda_F|, \tag{3.4}$$

where the objective function is the SAD. Note that in problem (3.4), we use SAD as the estimation error measurement rather than using other error measurements, such as the sum of squared errors. There are two reasons for this. The first reason is that as we showed in Chapter 2, the forecast that minimizes the absolute errors will also minimize the error in the number of agents. Another reason is that minimizing the SAD is more robust against outliers, compared to minimizing the sum of squared errors. In subsection 3.4.2, we emperically verify this claim using two real call center data sets.

In fact, the errors measured by SAD are scaled errors, in the sense that if we choose large numbers for $\hat{p}$ and $\hat{q}$, the error would be smaller. An extreme example that indicates this scaling problem is letting $\hat{p} = 1$ and $\hat{q} = 1$, and SAD would always be 0 by choosing $\hat{\lambda}_F = 0$. Therefore, we introduce the following minimization problem, which uses WAPE instead of SAD as the objective function to remove this scaling problem,

$$(\hat{p}, \hat{q}, \hat{\lambda}_F) = \underset{0 \leq p,q<1,\lambda_F}{\operatorname{argmin}} \frac{\sum_{i=1}^{N} |(1-p)A_i + (1-q)C_i - \lambda_F|}{\sum_{i=1}^{N} \left((1-p)A_i + (1-q)C_i\right)}. \tag{3.5}$$

One can notice that we choose term $\sum_{i=1}^{N} \left((1-p)A_i + (1-q)C_i\right)$ as the denominator of WAPE rather than the term $\sum_{i=1}^{N} \lambda_F$. This is for computational purposes. Because p and q are always bounded between 0 and 1, we can calculate the minimum WAPE on a grid of p and q ranging from 0 to 1. In contrast, we have no information on how big λ_F is, which makes it more difficult to find the minimum WAPE if we let $\sum_{i=1}^{N} \lambda_F$ to be the denominator.

Above, we have shown a regression method for estimating (p^*, q^*, λ_F^*). However, one can notice that we have three degrees of freedom (namely, p, q and λ_F), while only observations for A_is and C_is are being made. This means that in a call center data set without customer identity information, we cannot estimate (p^*, q^*, λ_F^*) simultaneously, and one parameter needs to be given before any regression method can be applied.

In a call center, there are different ways to estimate the reconnect probability. For example, the manager can ask agents to do some polling (e.g. for one whole day), we staff enough agents, so that almost all calls are handled, and we ask each agent to record each connected call's customer name or identity, then by the end of the day, we can calculate how many customers have called back. For the redial probability, this is more difficult to do, since abandoned customer's information is often not recorded. Using polling to determine the number of fresh calls is also difficult, because the number of fresh calls is not stable over time, due to the presence of trend and seasonality (see Shen and Huang (2008) and Ibrahim and L'Ecuyer (2013)).

Assuming $q = q^*$, we present an algorithm to numerically compute $(\hat{p}, \hat{\lambda}_F)$.

```
Algorithm 1:
    Step 0: Let $p = 0$, WAPE $= 1$, and let the grid size to be $\xi$.
    Step 1: Calculate $L_i = (1-p)A_i + (1-q^*)C_i$, for all $i = 1, \ldots, N$, and
            $\lambda_F = \text{median}(L_1, L_2, \ldots, L_N)$, $a = \frac{\sum_i |L_i - \lambda_F|}{\sum_i L_i}$.
    Step 2: If $a <$ WAPE, then let WAPE $= a$, $\hat{p} = p$, $\hat{\lambda}_F = \lambda_F$.
    Step 3: If $p \geq 1$, then stop; else, $p = p + \xi$, go to Step 1.
```

In this estimation model, we only consider redials and reconnects in the same day of the fresh call. We will motivate this assumption when we analyze the redial and reconnect behaviors from a real call center data set.

3.3.2 Intraweek seasonality

In Model (3.5), we made the assumption that each day has the same fresh call arrival rate. Often this is an unrealistic assumption in a real call center. We will show in subsection 3.4.1 that for two real call center data sets, both the total volume and the fresh volume show strong intraweek patterns. Thus, to make our model applicable in call center data with intraweek seasonality, we make adjustments to estimation model (3.5). To this end, we assume that the weekly total fresh calls distributed to each day of the week in a multiplicative way, i.e.,

$$\mathrm{E}\lambda_{F,i} = \mathrm{E}WF_{w_i} \cdot \beta_{d_i},$$

where w_i and d_i are the week number of day i and the weekday of day i, respectively, $d_i \in \{1,2,3,4,5\}$ (since we ignore the weekends), $w_i = 1, 2, \ldots n$, where n stands for the number of weeks. WF_{w_i} is a random variable that stands for the total number of fresh calls of week w_i. Thus, β_{d_i} can be interpreted as the proportion of calls on weekday d_i out of the whole week. A key assumption of this multiplicative model is that β_{d_i} does not depend on the week number. Such a multiplicative model has been applied in several call center forecasting models (see Weinberg et al. (2007), Brown et al. (2005) and Gans et al. (2012)). Therefore, our estimation model changes to

$$\begin{aligned}
(\hat{p}, \hat{q}, \hat{\beta}_{d_i}) = & \operatorname*{argmin}_{0 \leq p, q, \beta_{d_i} < 1} \frac{\sum_{i=1}^{N} |(1-p)A_i + (1-q)C_i - WF'_{w_i} \cdot \beta_{d_i}|}{\sum_{i=1}^{N} ((1-p)A_i + (1-q)C_i)} \\
\text{s.t.} \quad & \sum_{j=1}^{5} \beta_j = 1, \\
& WF'_{w_i} = WA_{w_i}(1-p) + WC_{w_i}(1-q), \quad w_i = 1, 2, \ldots n,
\end{aligned} \tag{3.6}$$

where WF'_{w_i} is the estimated number of fresh calls in week w_i, WA_{w_i} and WC_{w_i} are the total number of abandoned calls and total number of connected calls in week w_i, respectively. Since A_i and C_i are given observations, we can easily obtain their aggregated weekly volumes WA_{w_i} and WC_{w_i}. The intuition behind model (3.6) is that the daily fresh call volume is proportional to the weekly total fresh call volume. Once we have $(\hat{p}, \hat{q}, \hat{\beta}_{d_i})$, $\hat{\lambda}_{F,i}$ can be obtained via

$$\hat{\lambda}_{F,i} = \left(WA_{w_i}(1-\hat{p}) + WC_{w_i}(1-\hat{q})\right) \cdot \hat{\beta}_{d_i}, \quad i = 1, 2, \ldots, N.$$

Similar to the approach we took for solving (3.5), we assume $q = q^*$, and we solve (3.6) on a grid of p. Assuming $q = q^*$, and for a given value of p, problem (3.6) is a quantile regression problem with a linear constraint, which is equivalent to a linear programming (LP) problem. The corresponding LP problem of model (3.6) can be written as

$$\begin{aligned}
&\min_{\beta_{d_i}, Z_i^+, Z_i^-} \sum_{i=1}^{N} (Z_i^+ + Z_i^-) \\
&\text{s.t.} \sum_{i=1}^{N} \beta_{d_i} = 1, \\
&Z_i^+ - Z_i^- = \frac{(1-p)A_i + (1-q)C_i - WF'_{w_i} \cdot \beta_{d_i}}{\sum_{i=1}^{N} \left((1-p)A_i + (1-q)C_i\right)}, \; i = 1, \ldots, N, \\
&WF'_{w_i} = WA_{w_i}(1-p) + WC_{w_i}(1-q), \; i = 1, \ldots, N, \\
&0 \leq p, q, \beta \leq 1, \\
&Z_i^+, Z_i^- \geq 0, \; i = 1, \ldots, N.
\end{aligned} \tag{3.7}$$

In fact, problems (3.6) and (3.7) are equivalent (see Charnes et al. (1955)). We now give the idea of the proof, the equivalence of (3.6) and (3.7) holds if

$$Z_i^+ + Z_i^- = \frac{|(1-p)A_i + (1-q)C_i - WF'_{w_i} \cdot \beta_{d_i}|}{\sum_{i=1}^{N} \left((1-p)A_i + (1-q)C_i\right)}, \; i = 1, \ldots, N,$$

and it suffices to show that at least one of the values Z_i^+ and Z_i^- is zero in the optimal solution; otherwise, assume Z_i^+ and Z_i^- are both non-zero, then one could find a solution which has a smaller objective value by substracting $\min\{Z_i^+, Z_i^-\}$ from Z_i^+ and Z_i^-.

Assuming $q = q^*$, we show the following algorithm to numerically obtain $(\hat{p}, \hat{\lambda}_F)$.

In this chapter, we use *linp* function in package *limSolve* in R to solve the LP problem in Step 2. When choosing how large ξ is, one should bear in mind that when the grid size ξ is big, the precision will be low; when ξ decreases, the computation time will increase. In this chapter, we set the grid size to be 0.01, and for such a grid size, the computation time is small even for $N = 500$, i.e., less than 2 minutes.

```
Algorithm 2:
  Step 0: Let $p = 0$, WAPE $= 1$, and let the grid size to be $\xi$.
  Step 1: Calculate $L_i = (1-p)A_i + (1-q^*)C_i$, for all $i = 1, \ldots, N$,
          and $WF_{w_i} = (1-p)WA_{w_i} + (1-q^*)WC_{w_i}$, for all $w_i = 1, \ldots, n$.
  Step 2: Solve LP problem (3.7) for given $p$ and $q = q^*$.
  Step 3: Let $a$ be the objective value to the optimal solution, and $b_{d_i}$ be the
          optimal value for decision variable $\beta_{d_i}$, $i = 1, 2, \ldots, N$.
  Step 4: If $a <$ WAPE, then WAPE $= a$, $\hat{p} = p$, $\hat{\lambda}_{F,i} = WF_{w_i} \cdot b_{d_i}$, for
          all $i = 1, 2, \ldots, N$.
  Step 5: If $p \geq 1$, then stop; else, $p = p + \xi$, go to Step 1.
```

3.4 Validation

In this section, we test our estimation model (3.5) in the data sets generated by discrete-event simulation. The data generation procedure is the same as described in Section 3.2. Once the data have been generated, we use the model (3.5) for the estimation, then the estimated values are compared with simulation inputs.

Five different parameter settings are tested. These parameters are shown in Table 3.2, where λ_F^* is the fresh arrival rate per minute. For each parameter setting, we also validate our estimation model for different number of days, namely for $N = 20, N = 50$ and $N = 100$. For a given sample size, the estimators are themselves random variables. To understand the bias and the variability of the estimators, we replicate such simulation-estimation procedure fifty times, and then calculate the sample mean and standard deviation of the estimated values. Note that larger numbers of replications lead to more accurate estimates for the means and quantiles, but it will be more computationally expensive.

We used Algorithm 1 to calculate the estimated values. The sample mean, standard deviation, 5% and 95% quantile of the estimated values are shown in Table 3.3.

Example	p	q	λ_F^*	$1/\mu$ min	θ min	δ_{RD} min	δ_{RC} min
1	0.5	0.2	10	4	2	5	10
2	0.5	0.2	10	4	2	15	30
3	0.5	0.2	4	10	2	20	50
4	0.7	0.3	4	9	3	15	50
5	0.7	0	4	9	3	10	N.A.

Table 3.2: Parameters of the simulation experiments.

In Table 3.3, the estimations for p^* and λ_F^* are denoted as $\hat{p}|q^*$ and $\hat{\lambda}_F|q^*$, respectively, and the notation "$|q^*$" stands for the fact that it is an estimator given $q = q^*$. Furthermore, we let SD be the sample standard deviation of the estimators, and $Q_{\alpha,\hat{p}}$ and $Q_{\alpha,\hat{\lambda}_F}$ stand for the sample α quantile ($\alpha = 0.05$ or 0.95) of the estimator $\hat{p}|q^*$ and $\hat{\lambda_F}|q^*$, re-

Example	N	$\mathrm{E}(\hat{p}\vert q^*)$	$SD(\hat{p}\vert q^*)$	$(Q_{0.05,\hat{p}}, Q_{0.95,\hat{p}})$	$\mathrm{E}(\hat{\lambda}_F\vert q^*)$	$SD(\hat{\lambda}_F\vert q^*)$	$(Q_{0.05,\hat{\lambda}_F}, Q_{0.95,\hat{\lambda}_F})$
1	20	0.505	0.027	(0.464, 0.547)	9.960	0.096	(9.810, 10.110)
	50	0.502	0.010	(0.485, 0.516)	9.969	0.039	(9.913, 10.032)
	100	0.501	0.006	(0.493, 0.510)	9.971	0.021	(9.937, 10.000)
2	20	0.500	0.012	(0.480, 0.519)	9.990	0.049	(9.920, 10.058)
	50	0.501	0.009	(0.488, 0.514)	9.987	0.032	(9.936, 10.033)
	100	0.501	0.005	(0.491, 0.508)	9.992	0.021	(9.961, 10.026)
3	20	0.522	0.043	(0.457, 0.595)	3.942	0.070	(3.829, 4.052)
	50	0.512	0.016	(0.482, 0.536)	3.958	0.028	(3.920, 4.015)
	100	0.506	0.010	(0.491, 0.523)	3.969	0.017	(3.946, 3.999)
4	20	0.710	0.022	(0.670, 0.744)	3.930	0.080	(3.809, 4.069)
	50	0.702	0.009	(0.688, 0.719)	3.958	0.036	(3.898, 4.018)
	100	0.702	0.006	(0.693, 0.710)	3.956	0.017	(3.935, 3.983)
5	20	0.708	0.042	(0.642, 0.771)	3.990	0.039	(3.937, 4.066)
	50	0.701	0.014	(0.680, 0.718)	3.996	0.016	(3.971, 4.018)
	100	0.702	0.009	(0.687, 0.715)	3.996	0.011	(3.980, 4.013)

Table 3.3: Estimation results.

spectively. One can see from Table 3.3 that the differences between $\mathrm{E}(\hat{p}|q^*)$ and the p^* are less than 0.03, even for a relatively small sample size such as $N = 20$.

Furthermore, one could see from Table 3.3 that $\hat{\lambda}_F$ is a biased estimator, which underestimates λ_F^*. Here we describe the reason of this bias and argue that it is relatively small compared to λ_F^*. The source of the biases mainly comes from the fact that the median of λ_F would minimize the WAPE, while λ_F^* is the mean of λ_F. In the case that λ_F is a Poisson random variable, the difference between its mean and median is not zero, but some small values that are bounded by 1 (Chen and Rubin (1986)). However, since the estimation method uses daily aggregated volumes, the bias is relatively small compared to λ_F^*, as one can confirm this in the results of all examples in Table 3.3. To illustrate the relation between p, q and the WAPE, we plot the minimum WAPE on a grid of p and q in Figure 3.2. One can see that the true parameters p^* and q^* are on the line where the minimum WAPE is attained. Other simulation examples (not shown here) gave similar graphs as in Figure 3.2. This figure confirms that when $q = q^*$, the minimum WAPE leads to the accurate estimate for p. Moreover, this figure also shows how sensitive $\hat{p}$ is with respect to the choice of q. For example, if one would make a calculation error of ϵ for q^*, then in this example, the estimation error for p would be $\epsilon \cdot 5/8$, since the slope of the line with minimum WAPE is $5/8$ and our estimated point can only be one of the points in this line.

Note that in the simulation, we take B, H, Γ_{RC} and Γ_{RD} to be exponentially distributed. However, since A_i and C_i are realizations which can be obtained from the data, how B, H, Γ_{RC} and Γ_{RD} are distributed becomes irrelevant for estimation model (3.5). We now explain the reason. Assume for simplicity, there is only one fresh call during the day, and this call is connected. However, this customer would like to reconnect today, and this reconnect is answered again. Whether this customer calls back at 2pm or 5pm will not change the fact that there are two answered calls, one of them is a reconnect, and $q = 1/2$. Thus, as long as customers call back within the same day as their corresponding fresh calls, how large δ_{RC} is does not have any influence on estimation results.

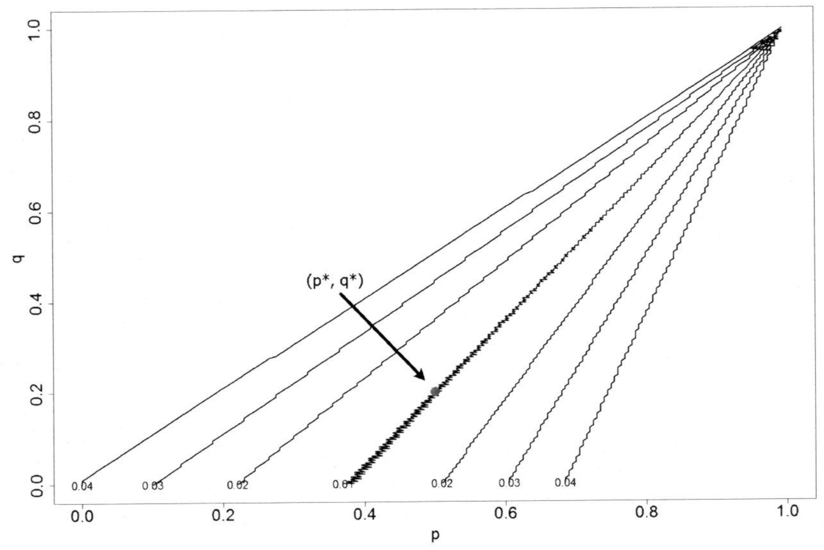


Figure 3.2: Values of WAPE on the grid of p and q for simulated data example 1, with the red points standing for (p^*, q^*).

Consequently, we can extend our estimation model to call centers where these variables have general distributions.

3.4.1 Validation using real data

In this section, we analyze real call center data to understand the redial and reconnect behaviors as well as to validate our estimation model (3.7). This data set consists of call arrival records to the Vanad Laboratory. The calls are recorded from 1st April 2012 to 29th September 2012. There are in total 498508 call records during these periods. On Sundays, the call center is closed. On Saturdays, the arrival volume is quite low, i.e., 5508 total call records for 26 Saturdays. Therefore, we may ignore the weekends call data, and focus only on the weekdays. We also remove the weeks which consist of one or few days of holidays. This leaves us with 22 weeks of data. Each call record consists of seven attributes, i.e., call arrival date, arrival time, caller's phone number, router name, agent number, time that the call is answered and the time that the call is hang up. We assume that each caller is identified by its phone number. Approximately 20% of the caller's phone numbers is unidentified, since some callers set their phone number to be invisible by the call receivers. There are eleven different types of routers that can be selected by a caller. The selection of router is done by customers via Interactive Voice Response (IVR) unit. After the customer has made the selection, his call will be distributed by an Automatic Call Distributor (ACDs). Each router represents one or multiple types of questions that a customer may have. Among those eleven routers, there are four major routers, which consist of approximately 71% of all calls. Among those four routers, we will focus our study on two specific routers which are referred to as router A and router B. The reasons that we choose these two routers are the following; (i) other routers may represent multiple types of questions, and customers who have different types of questions may have different redial and reconnect behaviors; (ii) some routers have been merged or changed their names during the data collection periods.

For this data set, we have the caller-identity information, which allows us to follow each customer and see whether he called back or not. In Figures 3.3 and 3.4, we plot the histograms of realizations of Γ_{RD} and Γ_{RC} for router A. For router B, we obtain similar figures. We can see that both for redial and reconnect, most of the customers call back in the same day as their fresh calls, and they call back shortly after abandonments or connected calls. A small fraction of the customers redial or reconnect one or two days later after the fresh call. Therefore, in our model, it is sufficient to assume that the redials and the reconnects arrive in the same day as the fresh call, i.e., customer who calls again one or more days later will be regarded as another fresh call. Some descriptive statistics are shown in Table 3.4. The total number, the fresh number and the redials and reconnects are plotted in Figures 3.5 and 3.6. In these two graphs, the unidentified calls are removed.

In Figures 3.5 and 3.6, one could still see high variability in the number of fresh calls in contrast to Figure 3.1, where very little variability is observed in the number of fresh calls. This is because besides the redials and reconnects, another source of variability in the total volumes in Figures 3.5 and 3.6 is the intraweek seasonality. In other words, one cannot use the redials and reconnects to explain all the variabilities in real call center data, since intraweek seasonality is also a major cause of variability. This observation would confirm the necessity to include seasonality in the estimation model.

Router	Total volume	Fresh volume	Redials	Reconnects
A	41624	36515	2142	2967
B	28526	23782	1117	3627

Table 3.4: Descriptive statistics

After removing the unidentified calls, RD_i and RC_i become observations in this data set with customer identity information, and we use following formulas to calculate the actual redial and reconnect probabilities,

$$p^* = \frac{\sum_{i=1}^{N} RD_i}{\sum_{i=1}^{N} A_i}, \quad q^* = \frac{\sum_{i=1}^{N} RC_i}{\sum_{i=1}^{N} C_i}.$$

We also calculate the probabilities p^* and q^* of each weekday and show them in Tables 3.5 and 3.6. We see that the redial and reconnect behaviors are quite significant, i.e., the reconnect probability can reach 15%. This further confirms the necessity of including both orbits in the queueing model. Furthermore, we see that both probabilities are different, i.e., the redial probability is usually larger than the reconnect probability. Intuitively, this makes sense, since an abandoned customer has higher urge to call back than an answered customer. For different routers, the redial probability has more fluctu-

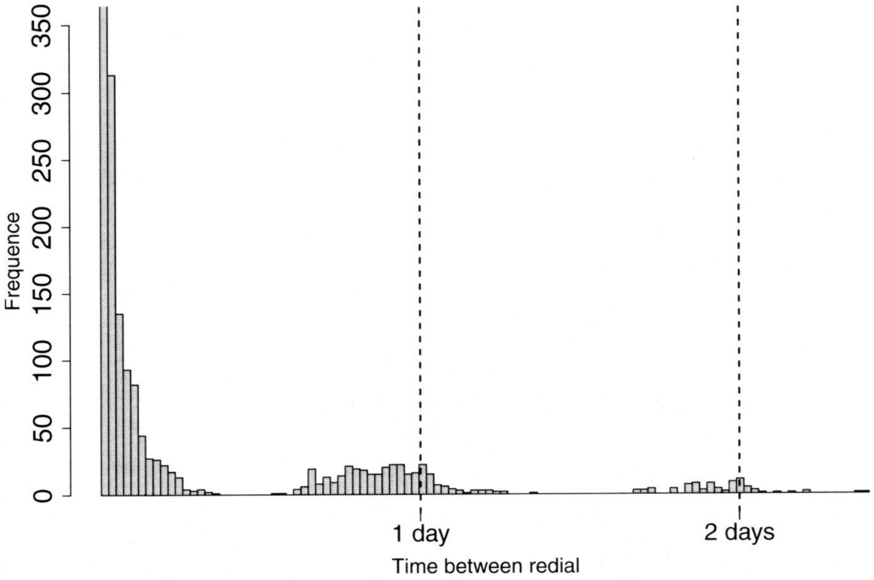


Figure 3.3: The histograms of realizations of Γ_{RD} for Router A.

ations than reconnect probability. However, within the same router for every weekday both probabilities are stable, only except for the redial probability for router A on Friday. Therefore, it is sufficient to have two parameters for all weekdays together for redial and reconnect probabilities of each router.

We apply Algorithm 2 to the Vanad Laboratories data set (with grid size 0.01). The estimation results are shown in Table 3.7, where WAPE_F is used to measure the percentage difference between $\hat{\lambda}_{F,i}$ and $\lambda^*_{F,i}$, and it is defined as

$$\text{WAPE}_F := \frac{\sum_{i=1}^N |\hat{\lambda}_{F,i} - \lambda^*_{F,i}|}{\sum_{i=1}^N |\lambda^*_{F,i}|}.$$

One can see from Table 3.7 that our estimation of redial probability for router A is approximately 0.05 higher than the real redial probability, while for router B, our estimation is about 0.09 higher. For a call center with $r = 20\%$, i.e., 20% of all calls are abandoned, 0.09 error in redial probability would lead to less than 2% error in estimating the number of fresh calls. For these two routers, whose abandonment percentages are much smaller than 20%, 0.09 would lead to even less errors. Therefore, maximum of 0.09 error in our estimate of the redial probability is acceptable. Furthermore, the WAPE_F for both routers are quite small, which are both less than 3%. The real fresh calls and the estimated fresh calls are plotted Figures 3.11 and 3.12. In both figures, our estimates $\hat{\lambda}_{F,i}$ are quite close to the real fresh calls $\lambda^*_{F,i}$.

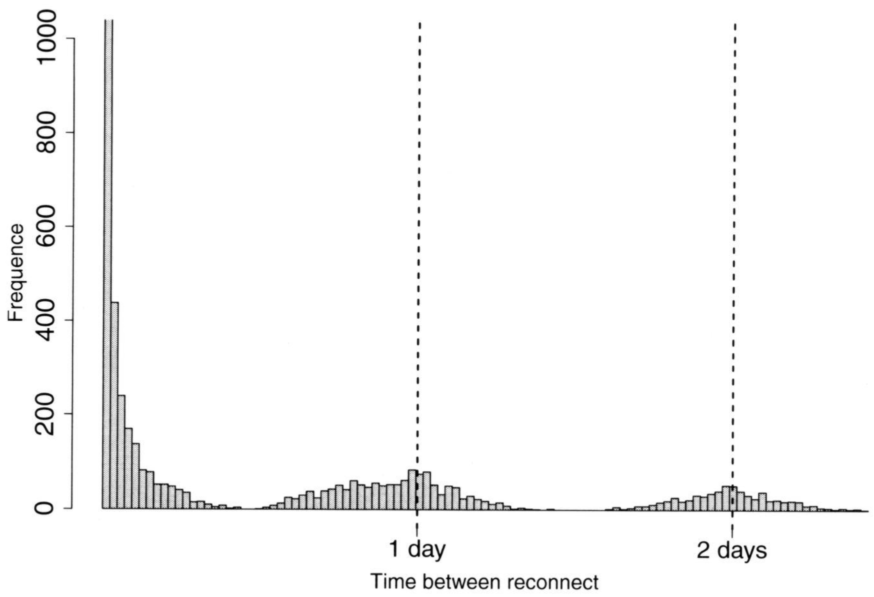


Figure 3.4: The histograms of realizations of Γ_{RC} for Router B.

To understand the relation between p, q and the WAPE, we plotted the minimum WAPE on a grid of p and q in Figure 3.7 and Figure 3.8. One can see that the true parameters p^* and q^* are close to the region where the minimum WAPE is attained. This also suggests that once we know q^*, the minimum WAPE will lead us to a close estimate for p.

3.4.2 Minimizing WAPE vs. minimizing SSE

In this subsection, we compare the estimator that minimizes WAPE with the ordinary least squared (OLS) estimator which minimizes SSE. The WAPE or the absolute errors are more robust against outliers comparing to the squared errors (Narula et al. (1999)). We now emperically validate this claim with our data.

Given q^*, the OLS estimator can be obtained by

$$\begin{aligned}
&(\hat{p}^{(OLS)}, \hat{\beta}_{d_i}^{(OLS)}) = \underset{0 \le p, \beta_{d_i} < 1}{\text{argmin}} \sum_{i=1}^{N} \left((1-p)A_i + (1-q^*)C_i - WF'_{w_i} \cdot \beta_{d_i}\right)^2 \\
&\text{s.t.} \quad \sum_{i=1}^{N} \beta_{d_i} = 1, \qquad\qquad (3.8) \\
&\qquad WF'_{w_i} = WA_{w_i}(1-p) + WC_{w_i}(1-q^*), \quad w_i = 1, 2, \ldots n.
\end{aligned}$$

Minimization problem (3.8) is a standard regression problem with linear constraints.

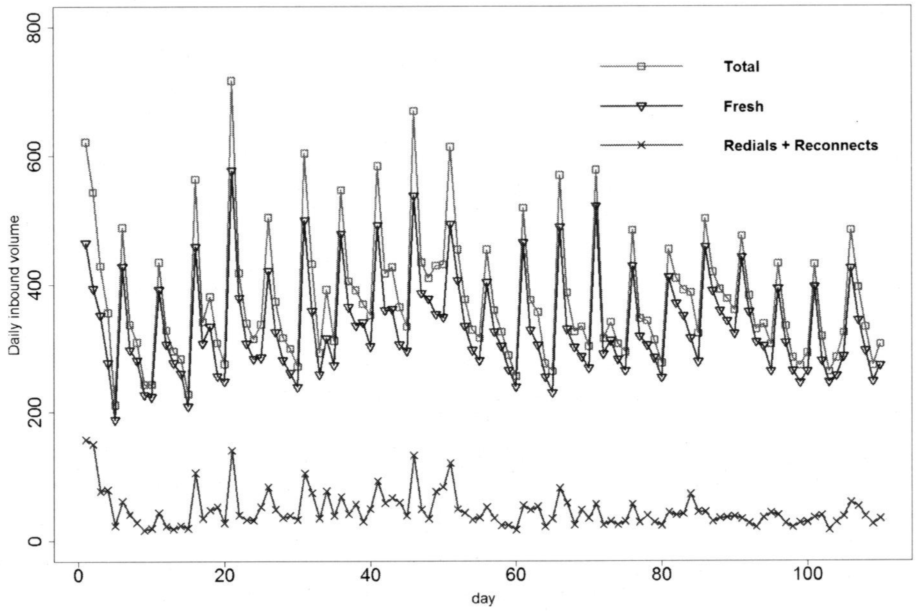


Figure 3.5: The plots of the total number of calls, fresh number of calls, redials plus reconnects for Router A.

The fitted squared errors for routers A and B are plotted in Figures 3.9 and 3.10. Based on these two figures, we can clearly visualize some outliers. These outliers are not holidays or being caused by any special events, thus, one could not identify them in advance. To validate the sensitivity of the OLS estimator with respect to outliers, we removed the whole week data for weeks that contain one or more days of outliers. The WAPE estimator and the OLS estimator based on data *with* and *without* outliers are shown in Tables 3.8 and 3.9. The results show that with the outliers, the OLS estimator leads to larger estimation errors for both routers compared to those for the WAPE estimator; while without the outliers, OLS estimator results in much more accurate estimation. For instance, we see in Table 3.9 that even a single outlier can lead to a very different estimation result for the OLS estimator. On the other hand, it can be seen that the WAPE estimator is much less sensitive to those outliers, in the sense that both with and without outliers, the WAPE estimator leads to accurate estimates. In call centers, the call volume is influenced by a lot of effects, some of which can be easily identified by date, such as holiday effects. However, not all outliers are easily identifiable in call center data. For example, in day 47 of Figure 3.9, it is difficult to judge whether this day is an outlier or not. Therefore, for the advantage of being more robust against outliers, we prefer the WAPE estimator to the OLS estimator.

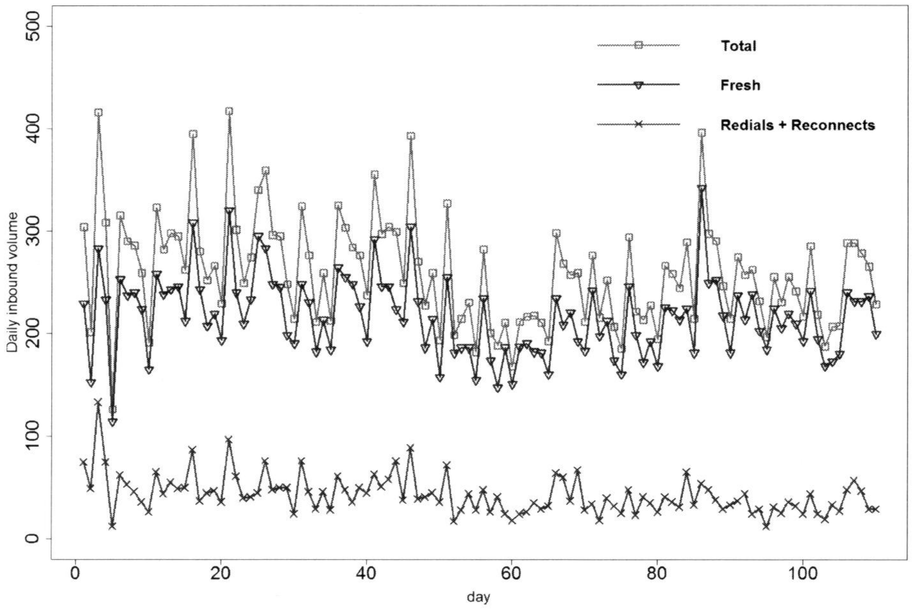


Figure 3.6: The plots of the total number of calls, fresh number of calls, redials plus reconnects for Router B.

3.5 Conclusion

In this chapter, we propose a queueing model for call centers with redials and reconnects. We use simulation results as well as real data analysis results to show that both features are significant and should not be ignored. We claim that it is more convenient to let the fresh volume represent the demand in call center in the sense that it does not depend on future operational decisions and other parameters such as customer patience and HT. Simulations show that if one does not distinguish between the total volume and the fresh volume, and uses the total volume to make operational decisions, it could lead to unnecessary costs. Thus, knowing the fresh volume is important for call centers. However, direct calculation of the number of fresh calls is difficult in some call centers, since customer identity information is not available in their data. In our model, we try to estimate the redial probability, reconnect probability and the fresh calls simultaneously in call center data without customer identity information by solving a minimization problem. However, we show that these three parameters cannot be estimated simultaneously. It is empirically shown that in order to have an accurate estimation, one variable needs to be given. We propose a polling method in call centers to calculate the reconnect probability. Once the reconnect probability is given, we show via simulation examples that the other two variables can be estimated. We also validate our model via two real call center data sets. Our estimate of the redial probabilities for both data sets are close to the actual redial probabilities, with errors of less than 0.09. Furthermore,

	Mon	Tue	Wed	Thu	Fri
p^*	0.52	0.52	0.46	0.49	0.43
q^*	0.08	0.08	0.07	0.09	0.08

Table 3.5: Real redial and reconnect probabilities of each weekday for Router A.

	Mon	Tue	Wed	Thu	Fri
p^*	0.42	0.40	0.38	0.38	0.39
q^*	0.15	0.13	0.14	0.15	0.12

Table 3.6: Real redial and reconnect probabilities of each weekday for Router B.

Router	p^*	q^*	$\hat{p}\|q^*$	WAPE_F
A	0.49	0.08	0.54	2.6%
B	0.40	0.14	0.49	3.0%

Table 3.7: Estimation results for both routers.

our estimate of the number of fresh calls are very close to the real number of fresh calls, with WAPE_F less than 3%.

In addition to help call center managers to estimate the fresh volumes of the call centers, this chapter also addresses the reconnect customer behavior in call centers. In the data set of Vanad Laboratories, we find out that the number of reconnects is significant. Neglecting it will lead to inaccurate prediction of the call volumes, which will cause inaccurate staffing. Inspired by these findings in this chapter, we propose the following topics for further research. First, for a call center manager, it would be interesting to know what are the consequences of neglecting reconnects in terms of costs or service levels. Second, in order to make the right staffing decisions, it would be useful to evaluate the service levels of a call center with consideration of the reconnect behaviors. Last, the redial and reconnect behaviors will introduce intraday correlation to the call center data. For example, if a manager saw a busy morning, he would expect a busy afternoon, since some “morning customers" will redial or reconnect in the afternoon. This raises an interesting question: 'how should the manager update the agents' schedules dynamically when morning realizations are available?'

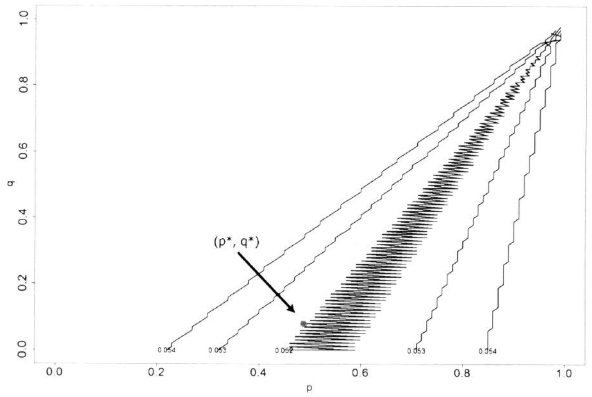


Figure 3.7: Values of WAPE on the grid of p and q for Router A.

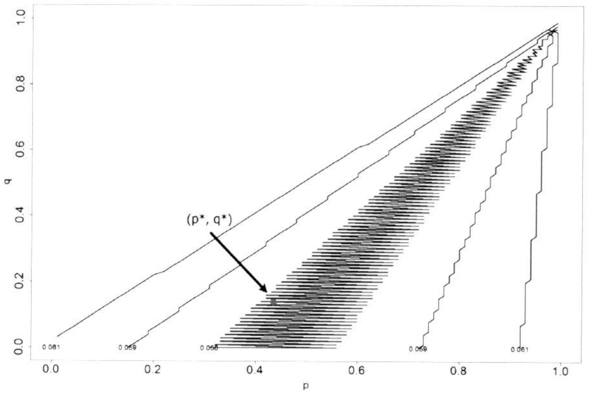


Figure 3.8: Values of WAPE on the grid of p and q for Router B.

	With outliers	Without outliers
$\hat{p}$	0.54	0.57
$\hat{p}^{(OLS)}$	0.70	0.44

Table 3.8: Comparing different estimators for Router A.

	With outliers	Without outliers
$\hat{p}$	0.49	0.52
$\hat{p}^{(OLS)}$	0.81	0.58

Table 3.9: Comparing different estimators for Router B.

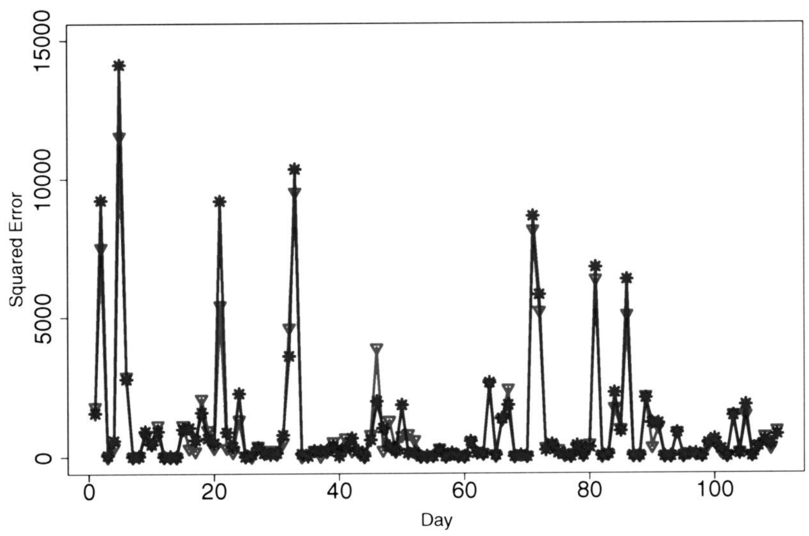


Figure 3.9: Squared errors of the WAPE estimator (star) vs. those of the OLS estimator (triangle) Router A.

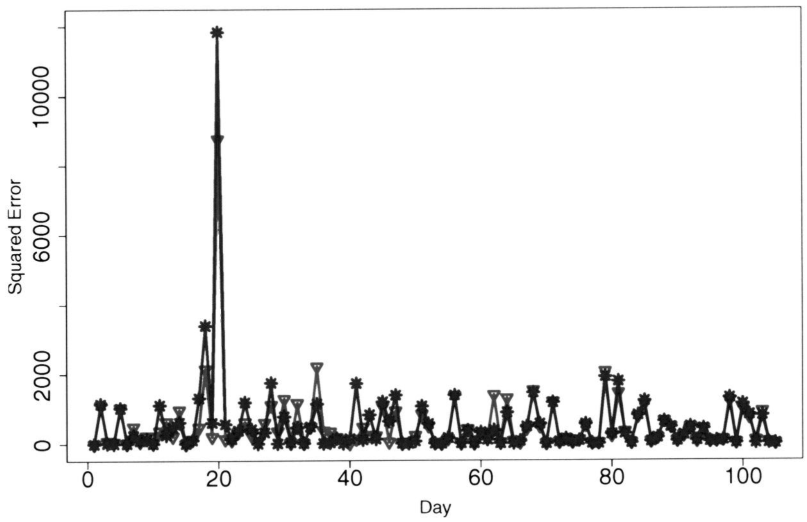


Figure 3.10: Squared errors of the WAPE estimator (star) vs. those of the OLS estimator (triangle) Router B.

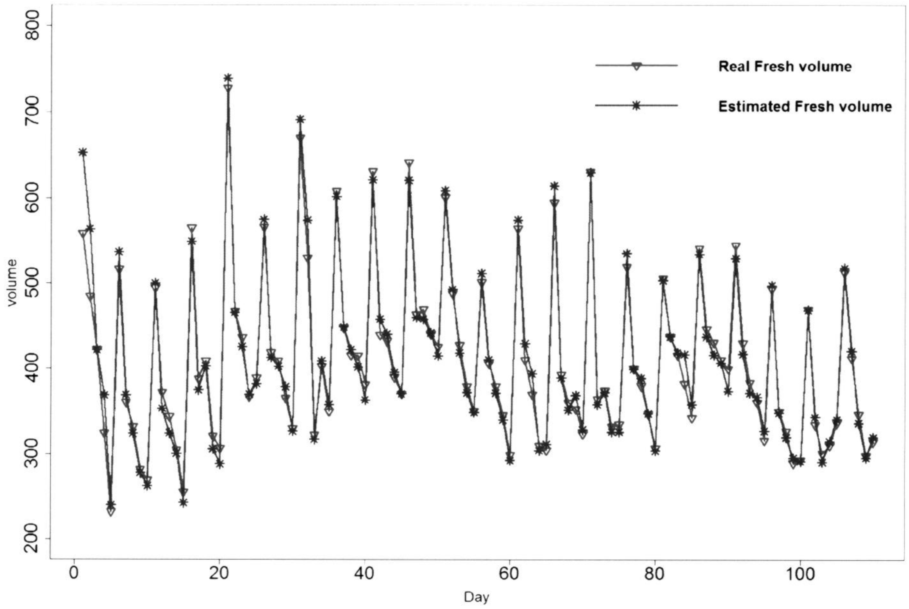


Figure 3.11: Real number of fresh calls vs. estimated number of fresh calls for Router A.

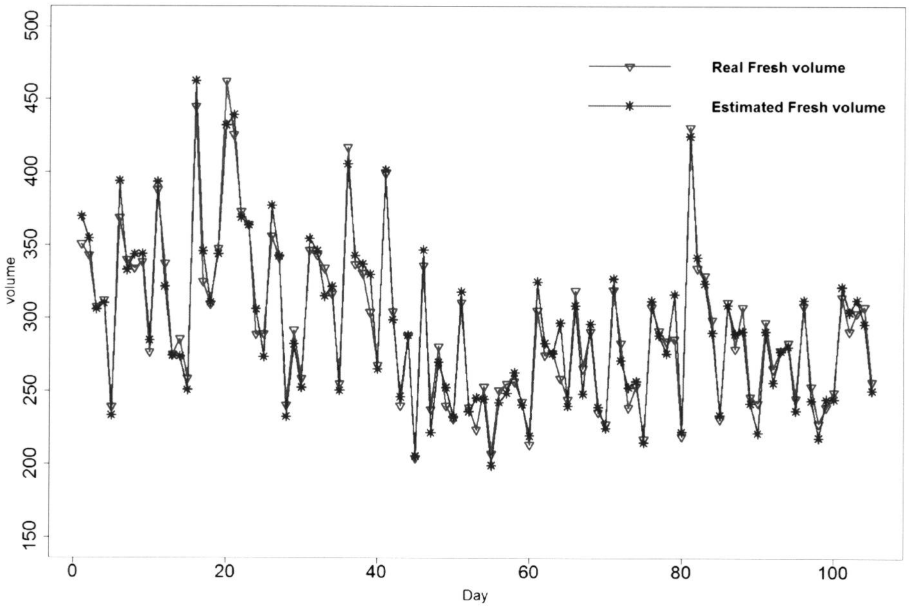


Figure 3.12: Real number of fresh calls vs. estimated number of fresh calls for Router B.

Chapter 4

Fluid approximation of a model with redials and reconnects

Despite the existence and significant amount of redials and reconnects in call centers, these two features together have not been considered when making staffing decisions. As we showed in Chapter 3, ignoring them will inevitably lead to under- or overestimation of call volumes, which results in improper and hence costly staffing decisions.

Motivated by this, in this chapter we study the staffing problems (i.e., for given parameters, determine the number of agents) for the call center model with redials and reconnects. We use a fluid model to derive first order approximations for the number of customers in the redial and reconnect orbits in heavy traffic. We show that the fluid limit of such a model is the unique solution to a system of three differential equations. Furthermore, we use the fluid limit to calculate the expected total arrival rate, which is then given as an input to the Erlang A formula for the purpose of calculating the service levels and abandonment probabilities. The performance of such a procedure is validated numerically in the case of both single intervals with constant parameters and multiple intervals with time-dependent parameters. The results demonstrate that this approximation method leads to accurate estimates for the service levels and the abandonment probabilities.

4.1 Introduction

The response-time performance of call centers is crucial for the customer satisfaction. It is essential to the costs and the performances of call centers that managers make the right staffing decisions (i.e., determine the right number of agents). Various models have been developed in order to support such decision processes. One of the most widely used models is the Erlang C model and there is a lot of literature on it (see Gans et al. (2003) and the references therein). The square-root staffing rule is a simplified and approximated staffing rule for the Erlang C model, which is proposed by Halfin and Whitt (1981). However, the Erlang C model does not include customer abandonments, while the Erlang A model does. Garnett et al. (2002) show that the square-root staffing rule remains valid for the Erlang A model. However, both the Erlang C and the Erlang A model ignore customer redial (a re-attempt after an abandoned call) behaviors in call centers, while this behavior can be quite significant (see Gans et al. (2003) and reference therein). Aguir et al. (2008) discover that ignoring redials can lead to under-staffing or over-staffing, depending on the forecasting assumption being made. This model with reneging was later extended by Phung-Duc and Kawanishi (2014) and Phung-Duc and Kawanishi (2011) with an extra feature of after-call work. Sze (1984) studies a queueing model where abandonments and redials are included, focusing on heavily loaded systems. We refer to Falin and Templeton (1997) for more references on retrial queues.

Motivated by the application in healthcare staffing with reentrant patients, Liu and Whitt (2014), Yom-Tov and Mandelbaum (2014) develop methods to set staffing levels for models with and without Markovian routing. Such methods remain valid for time-varying demand. In Ding et al. (2013), the authors use real call center data to show that an inbound call can either be a fresh call (an initial attempt), a redial or a reconnect. Also, as argued in Ding et al. (2013), redials and reconnects should be considered and modeled, since not distinguishing them from the fresh calls can lead to significantly over- or underestimation of the total inbound volume. As a consequence, neglecting the impact of redials and reconnects will lead to either overstaffing or understaffing. In case of overstaffing, the performance of the call center will be good, but at unnecessarily high costs. In case of understaffing, the performance of the call center will be degraded, which may lead to customer dissatisfaction and possibly customer churn. Despite the economic relevance of including both features in staffing models, to the best of the authors' knowledge no papers have appeared on staffing of call centers where *both* redials and reconnects are included. This chapter aims to fill this gap, that is, we investigate the staffing problem in call centers with the features of both redials and reconnects. We focus on the case of large call centers that operate under heavy load.

In the Erlang C model, if the system is heavily loaded, the expected queue length will go to infinity in stationarity, and arriving customers will on average experience infinitely long waits. However, for large call centers with customer abandonments, especially during the busy hours when the inbound volume is quite large such that the system operates under heavy load, it is possible that most customers will experience relatively short waiting times while having only a small customer abandonment percentage. Further discussions of this effect can be found in Garnett et al. (2002).

In this chapter, we aim to answer the following question: "In large call centers, for given number of agents, what are the SL_2 and abandonment percentage r if both redialing and reconnection of customers are taken into account?" To answer this question, one must first estimate the total number of arrivals into the call center. This is not trivial, since as we have shown in Chapter 3 that the number of total arrivals depends on the number of agents. This dependency becomes more complicated in real life, due to the fact that the rate of fresh calls arriving and the number of agents are often time-dependent. If the number of arrivals cannot be determined, it is impossible to calculate the SL_2. Therefore, in this chapter, we take a two-step approach to calculate SL_2 and r. First, we numerically calculate the expected total arrival rate at any instant time by using a fluid limit approximation. We also show that the fluid limit of this model is a unique solution to a system of three deterministic differential equations. In the second step, under the assumption of the total arrival process being Poisson, we apply the Erlang A formula to obtain the SL_2 and r. This approximation turns out to be quite accurate. In this chapter, we consider only the expected SL_2 and r, for discussions about the SL variability, we refer to the work by Roubos et al. (2012).

Fluid models for call centers have been extensively studied. Whitt (2006) develops a deterministic fluid limit which the authors use to provide first-order performance descriptions for the $G/GI/s + GI$ queueing model under heavy traffic, where the second GI stands for the i.i.d. patience distribution. In Whitt (2006), the redial behavior is not considered, though. The existence and uniqueness of the fluid limit are given as conjectures. Mandelbaum et al. (2002) use the fluid and diffusion approximation for the multiserver system with abandonments and redials. They obtain first order approximations of queue length and expected waiting time as well as their confidence bounds. In Mandelbaum et al. (1999), the authors use a fluid and a diffusion approximation for the time varying multiserver queue with abandonments and retrials. They show that both approximations can be obtained by solving sets of non-linear differential equations, where the diffusion process can provide confidence bounds for the fluid approximation. The work by Mandelbaum et al. (1998) gives more general theoretical results for fluid and diffusion approximations for Markovian service networks. Aguir et al. (2004) extend the model by allowing customer balking behavior, but no formal proof of the fluid limit is given. Besides the applications in staffing call centers, fluid models have also been applied in delay announcement of customers in call centers (see Ibrahim and Whitt (2009, 2011)). Besides the fluid or the diffusion limits, there are other methods that can be used to approximate queueing models, such as the Gaussian Variance Approximation (GVA) method developed by Massey and Pender (2011). Such a GVA approach is generalized by Pender and Massey (2014) to Jackson networks with abandonments, which leads to better approximations compared to approximation results obtained by the corresponding fluid and diffusion limits.

The rest of the chapter is structured as follows. In Section 4.2, we describe the queueing model with the features of the redials and reconnects. In Section 4.3, we propose a fluid model, which is a deterministic analogue of the stochastic model. We prove that the original stochastic model converges to the fluid model under a proper scaling. We numerically compute the fluid approximations to the number of customers in the queue as well as those in two orbits, and simulate the original model, and compare them in the case of single intervals and multiple intervals, where the parameters are time-dependent but remain piece-wise constants within each interval. The Erlang A formula is then used to approximate the waiting time distributions in Section 4.5.

4.2 The redial and reconnect behaviors

To simplify our analysis, in this chapter, we assume that the HT of each customer is independent and exponentially distributed; for the study of dependent HT, please see Pang and Whitt (2012).

The reason to include of reconnects and to consider the exponential assumption of Γ_{RD}

and Γ_{RC} are that in practice the volume of reconnect is significant and Γ_{RD} and Γ_{RC} are approximately exponential. To demonstrate these, we conduct data analysis of real call center data, which has been descibed and used in Chapter 3. For simplicity, we only select one type of call, which accounts for nearly 40% of all call records. The redial and reconnect probabilities of this type of call are 0.40 and 0.15, respectively. Then, in such a case, if all the customers are connected to agents and 15% of the connected customers reconnect exactly once, then more than 13% of the total number of arrivals are reconnects. This further confirms the necessity to include the reconnect customer behavior in call center models. Besides calculating p and q from the data, we also plot the histograms of Γ_{RD} and Γ_{RC} in Figures 4.1 and 4.2. When generating these two figures, if a customer tries to redial or reconnect after one day, we consider that call back as a fresh call. We make this consideration because most of the customers call back in the same day as their corresponding fresh calls (Ding et al. (2013)). The sample averages of Γ_{RD} and Γ_{RC} are $1/\delta_{RD} = 41.46$ minutes and $1/\delta_{RC} = 53.49$ minutes, respectively. As one can see from the shapes in Figures 4.1 and 4.2 that the histograms have longer tails than the exponential distributions. Apart from the longer tails, the exponential distributions seem to be good approximations for Γ_{RD} and Γ_{RC}. This can be demonstrated by the two Q-Q plots in Figures 4.3a and 4.3b. In these two Q-Q plots, we ignore all the samples that are larger than 3 hours for Γ_{RD} and that are larger 4 hours for Γ_{RC}, which account for less than 8% of total sample size. Therefore, we keep the assumption that Γ_{RD} and Γ_{RC} are exponentially distributed.

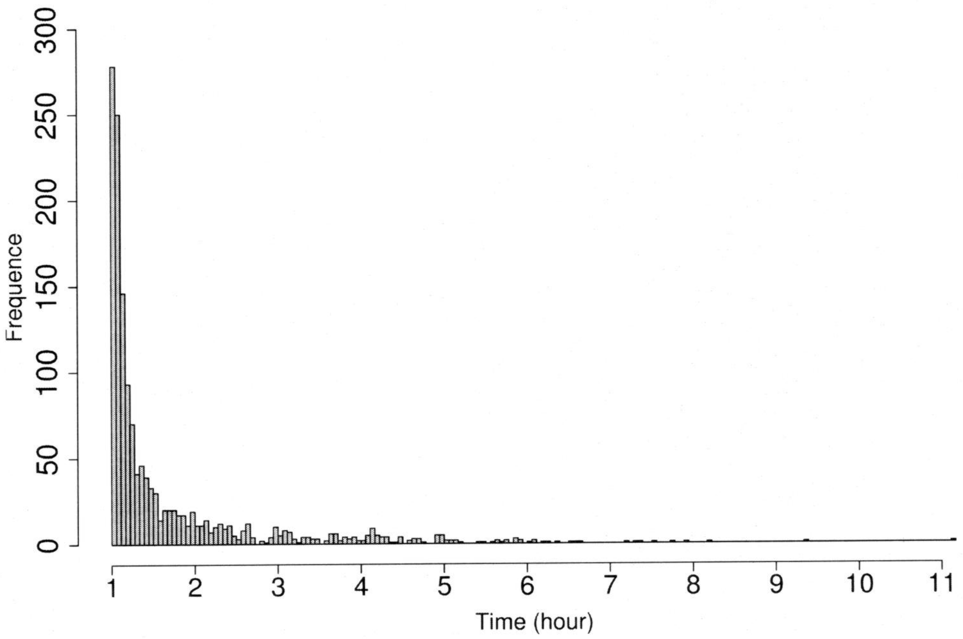


Figure 4.1: Histogram of Γ_{RD}.

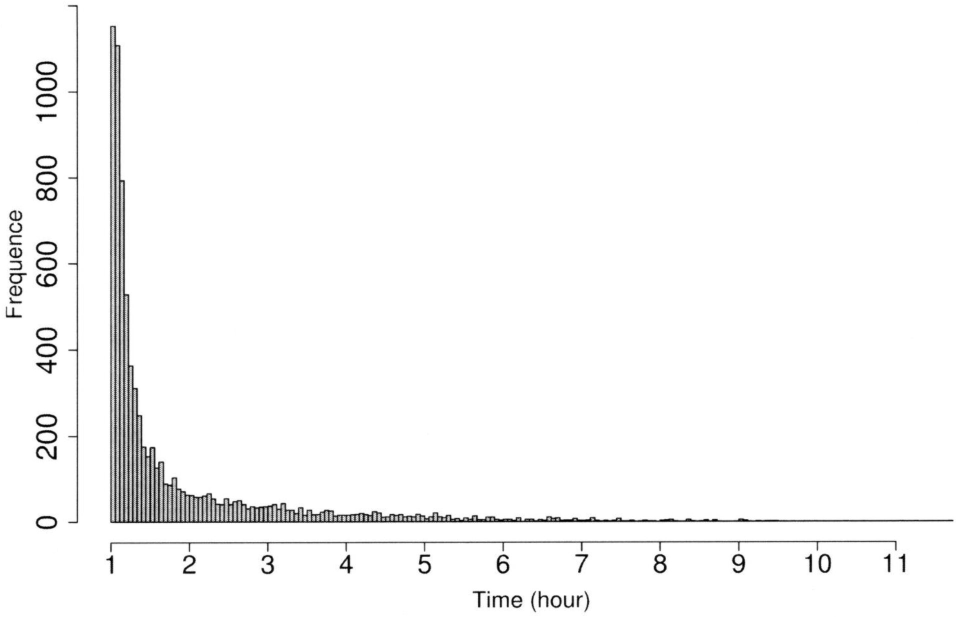
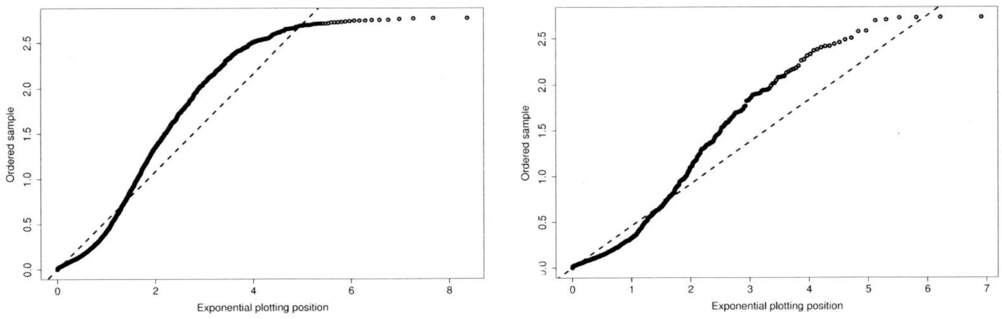


Figure 4.2: Histogram of Γ_{RC}.

In our model, we assume that there is no difference between the handling times of the fresh calls and those of the reconnects. To validate this assumption, we calculate the AHT for the fresh calls and the reconnects, e.g., $\mathrm{E}B = 5.14$ minutes, for the fresh calls, and $\mathrm{E}B = 5.35$ minutes, for the reconnects. Thus, the fresh calls have slightly lower AHT. We think that a possible cause for this is that reconnects might represent more difficult questions of customers than the fresh calls, which means it requires longer efforts to handle the reconnects. However, if we differentiate between the handling times of the fresh calls and the reconnects, this would complicate the model significantly, since instead of knowing the total number of customers in the queue, one would need to know both the number of the reconnects and the number of the fresh calls in the queue, as well as their orders in the queue. Therefore, considering the added complexity and the fact that the difference between them is relatively small, in this chapter, we keep our assumption that the fresh calls are statistically the same as the reconnects in terms of HT.

4.3 Fluid limit approximations

In this section, we first show that the problem of calculating the expected total arrival rate comes down to the problem of calculating $\mathrm{E}Z_Q(t), \mathrm{E}Z_{RD}(t)$ and $\mathrm{E}Z_{RC}(t)$, where $Z_Q(t)$ is the number of customers in the queue plus the number of customers in service



(a) Q-Q plot of the realizations of Γ_{RD}.

(b) Q-Q plot of the realizations of Γ_{RC}.

Figure 4.3: Q-Q plots.

at time t, $Z_{RD}(t)$ is the number of customers in the redial orbit at time t, and $Z_{RC}(t)$ is the number of customers in the reconnect orbit at time t, and $Z_Q(t), Z_{RD}(t)$ and $Z_{RC}(t)$ are random processes. Because an arrival can be a fresh call, a redial or a reconnect, the following equation holds for any t,

$$\begin{aligned} \mathrm{E}\lambda_T(t) &= \lambda_F(t) + \mathrm{E}\lambda_{RD}(t) + \mathrm{E}\lambda_{RC}(t) \\ &= \lambda_F(t) + \delta_{RD}\mathrm{E}Z_{RD}(t) + \delta_{RC}\mathrm{E}Z_{RC}(t), \end{aligned} \tag{4.1}$$

where $\lambda_T(t)$ stands for the total arrival rate at time t, which is a stochastic process, $\lambda_F(t)$ stands for the fresh arrival rate at time t, $\lambda_{RD}(t)$ and $\lambda_{RC}(t)$ stand for the arrival rate due to redials and reconnects at time t, respectively. Therefore, once $\mathrm{E}Z_Q(t), \mathrm{E}Z_{RD}(t)$ and $\mathrm{E}Z_{RC}(t)$ are known, $\mathrm{E}\lambda_T(t)$ can be obtained by Equation (4.1). Note that $Z_Q(t)$ does not appear in Equation (4.1), but we will see later that $Z_{RD}(t)$ and $Z_{RC}(t)$ depend on $Z_Q(t)$.

In fact, the stochastic process $\{\mathbf{Z}(t), t \geq 0\}$, which is defined by

$$\mathbf{Z}(t) := \left(Z_Q(t), Z_{RD}(t), Z_{RC}(t)\right)^T, \tag{4.2}$$

is a three-dimensional Markov process, because the inter-arrival time, HT and other durations are assumed to be exponentially distributed. The state space of this Markov process is $\mathbb{Z}_+^3$. To save space, we will not show the transition diagram here. Since it is a Markov process, we can truncate the system at a certain large state, and numerically obtain the steady state distribution of $\mathbf{Z}(t)$ by solving global balance equations. Theoretically, this method offers almost exact results, in the sense that one can control the error by truncating at some sufficiently large state. However, for the model we consider, it is very difficult to formulate and solve the global balance equations, especially for large systems. Therefore, for the convenience of practical usage, we will not consider solving this Markov process, but some approximation methods.

4.3.1 Fluid limit

In this subsection, we present a fluid model, which we show to arise as the limit under a proper scaling of the stochastic model in Figure 1.8.

Often, the fresh arrival rates are time-dependent in real call centers. The operational hours of call centers are divided into several intervals for the convenience of staffing and making schedules, and it is conventional to assume that the fresh arrival rate differs per interval but remains piece-wise constant within each single interval. Thus, we start our analysis by considering the single interval case, where the fresh arrival rate is assumed to be constant for any t (e.g., $\lambda_F(t) = \lambda_F$, $t \geq 0$). The cases with time-dependent arrival rates will be discussed later. For the single interval case, the following flow conservation equations hold for this stochastic model:

$$Z_Q(t) = Z_Q(0) + \Pi_{\lambda_F}(t) + D_{RD}(t) + D_{RC}(t) - D_s(t) - D_a(t), \tag{4.3}$$

$$Z_{RD}(t) = Z_{RD}(0) + \sum_{j=1}^{D_a(t)} B_j(p) - D_{RD}(t), \tag{4.4}$$

$$Z_{RC}(t) = Z_{RC}(0) + \sum_{j=1}^{D_s(t)} B_j(q) - D_{RC}(t), \tag{4.5}$$

where $\Pi_{\lambda_F}(t)$ is the number of fresh arrivals during time interval $[0,t)$, and $\Pi_{\lambda_F}(\cdot)$ is a Poisson process of rate λ_F. In addition, $D_{RD}(t), D_{RC}(t), D_s(t)$, $D_a(t)$ are the number of redials during $[0,t)$, number of reconnects during $[0,t)$, number of served customers during $[0,t)$ and number of abandoned customers during $[0,t)$, respectively. $B_j(p)$ is a Bernoulli random variable with success probability p, $j = 1, 2, \ldots, D_a(t)$. That is, $B_j(p) = 1$, if the j-th abandoned customer enters the redial orbit; $B_j(p) = 0$, otherwise. Therefore, for given $D_a(t)$, $\sum_{j=1}^{D_a(t)} B_j(p) \sim \text{Bin}(D_a(t), p)$. By the same argument, we have $\sum_{j=1}^{D_s(t)} B_j(q) \sim \text{Bin}(D_s(t), q)$, for given $D_s(t)$.

Let $\Pi_i(\cdot)$, $i = 1,2,3,4$, be independent Poisson processes of rate 1, then we claim the following

$$\begin{aligned}
D_s(t) &\overset{d}{=} \Pi_1\left(\int_0^t \mu \min\{s, Z_Q(u)\} du\right), \\
D_a(t) &\overset{d}{=} \Pi_2\left(\int_0^t \theta\left(Z_Q(u) - s\right)^+ du\right), \\
D_{RD}(t) &\overset{d}{=} \Pi_3\left(\int_0^t \delta_{RD} Z_{RD}(u)\, du\right), \\
D_{RC}(t) &\overset{d}{=} \Pi_4\left(\int_0^t \delta_{RC} Z_{RC}(u)\, du\right),
\end{aligned}$$

where symbol $\overset{d}{=}$ stands for equality in distribution.

Rigorous proof of these four statements can be given along the lines of Pang et al. (2007), see Lemma 2.1.

To introduce the fluid limit, we consider a sequence of models as in Figure 1.8 such that, in the n-th model, the fresh arrival rate is $\lambda_F n$ and the number of servers is ns. We add the superscript " (n) " to all notations in the n-th model. Similarly to (4.3)-(4.5), we then have for the n-th model:

$$Z_Q^{(n)}(t) = Z_Q^{(n)}(0) + \Pi_{\lambda_F n}^{(n)}(t) + D_{RD}^{(n)}(t) + D_{RC}^{(n)}(t) - D_s^{(n)}(t) - D_a^{(n)}(t), \tag{4.6}$$

$$Z_{RD}^{(n)}(t) = Z_{RD}^{(n)}(0) + \sum_{j=1}^{D_a^{(n)}(t)} B_j(p) - D_{RD}^{(n)}(t), \tag{4.7}$$

$$Z_{RC}^{(n)}(t) = Z_{RC}^{(n)}(0) + \sum_{j=1}^{D_s^{(n)}(t)} B_j(q) - D_{RC}^{(n)}(t). \tag{4.8}$$

Now we define the fluid scaled process

$$\bar{\mathbf{Z}}^{(n)}(t) := \left(\bar{Z}_Q^{(n)}(t), \bar{Z}_{RD}^{(n)}(t), \bar{Z}_{RC}^{(n)}(t)\right)^T,$$

where

$$\bar{Z}_Q^{(n)}(t) := \frac{Z_Q^{(n)}(t)}{n}, \quad \bar{Z}_{RD}^{(n)}(t) := \frac{Z_{RD}^{(n)}(t)}{n}, \quad \bar{Z}_{RC}^{(n)}(t) := \frac{Z_{RC}^{(n)}(t)}{n}.$$

Let $D([0,\infty),\mathbb{R}^3)$ be the space of right continuous functions with left limits in $\mathbb{R}^3$ having the domain $[0,\infty)$. We endow $D([0,\infty),\mathbb{R}^3)$ with the usual Skorokhod J_1 topology. Suppose $\{X^{(n)}\}_{n=1}^{\infty}$ is a sequence of stochastic processes, then notation $X^{(n)} \xrightarrow{d} x$ means that $X^{(n)}$ converge weakly to stochastic process x.

Definition 4.1. If there exists a limit in distribution for the scaled process $\{\bar{\mathbf{Z}}^{(n)}(\cdot)\}_{n=1}^{\infty}$, i.e. $\bar{\mathbf{Z}}^{(n)}(\cdot) \xrightarrow{d} \mathbf{z}(\cdot)$, then $\mathbf{z}(\cdot)$ is called the fluid limit of the original stochastic model.

4.3.2 Fluid limit for a single interval

To obtain the fluid limit of the system (i.e., a sequence of stochastic processes specified by Equations (4.6)-(4.8)), we divide both sides of Equations (4.6)-(4.8) by n, then let $n \to \infty$.

Lemma 4.1. *The sequence of scaled processes* $\{\bar{\mathbf{Z}}^{(n)}(\cdot)\}_{n=1}^{\infty}$ *is relatively compact and all weak limits are a.s. continuous.*

Proof. See subsection 4.7.1. □

Theorem 4.1. *If for given deterministic values $(z_Q(0), z_{RD}(0), z_{RC}(0))$, we assume*

$$\left(\bar{Z}_Q^{(n)}(0), \bar{Z}_{RD}^{(n)}(0), \bar{Z}_{RC}^{(n)}(0)\right) \xrightarrow{d} (z_Q(0), z_{RD}(0), z_{RC}(0)), \text{ as } n \to \infty,$$

then the fluid limit of the original stochastic model is the unique solution to the following system of equations

$$\begin{aligned} z_Q(t) &= z_Q(0) + \lambda_F t + \delta_{RD} \int_0^t z_{RD}(u)du + \delta_{RC} \int_0^t z_{RC}(u)du \\ &\quad - \mu \int_0^t \min\{s, z_Q(u)\}du - \theta \int_0^t (z_Q(u) - s)^+ du, \end{aligned} \tag{4.9}$$

$$z_{RD}(t) = z_{RD}(0) + p\theta \int_0^t (z_Q(u) - s)^+ du - \delta_{RD} \int_0^t z_{RD}(u)du, \tag{4.10}$$

$$z_{RC}(t) = z_{RC}(0) + q\mu \int_0^t \min\{s, z_Q(u)\}du - \delta_{RC} \int_0^t z_{RC}(u)du. \tag{4.11}$$

Proof. See subsection 4.7.2. □

Remark. Mandelbaum et al. (1998) suggests an alternative proof of this fluid limit result for a more general model. The approach of Mandelbaum et al. (1998) is based on a Brownian motion approximation of a Poisson process, thus, a second order approximation, which they simultaneously use to derive both the fluid and diffusion limits. Our derivation of the fluid limit is more straightforward and does not use second order approximations. More precisely, we use the recipe of Ethier and Kurtz (1986) which can be considered classic for proving fluid limit results. For us the fluid limit alone is sufficient to obtain an approximation to the waiting time distribution (see the Erlang A approximation in section 4.5).

We could not obtain analytic expressions for $z_Q(t), z_{RD}(t)$ and $z_{RC}(t)$ from Equations (4.9)-(4.11). However, solving them numerically can be done via a standard approach for solving differential equations, and it is relatively fast.

4.3.3 Fluid limit for multiple intervals

We have just shown the fluid limit for a single interval, where the parameters λ_F and s remain the same within the interval. However, in real call centers, parameters can vary during the day, especially the arrival rate $\lambda_F(t)$. As shown by Shen and Huang (2008) and Ibrahim and L'Ecuyer (2013), call volumes normally follow certain intraday patterns. Observing the intraday arrival pattern from the historical data set, managers would schedule different number of agents for each interval to meet the SL requirement. Therefore, we now show the fluid limit for the case of multiple intervals, where λ_F and s vary from interval to interval. We assume that other parameters remain constant.

We divide the operational hours of call centers into m intervals. Each interval starts at t_{i-1} and ends at t_i, $i = 1, 2, \ldots, I$. The fresh arrival rate of interval i is denoted by $\lambda_{F,i}$, and the number of agents in interval i is denoted by s_i, $i = 1, 2, \dots, I$. In the i-th interval, i.e., $t_{i-1} \le t < t_i$, the fluid limit then becomes

$$\begin{aligned} z_Q(t) = {} & z_Q(t_{i-1}) + \lambda_{F,i}\,(t - t_{i-1}) + \delta_{RD} \int_{t_{i-1}}^{t} z_{RD}(u)du \\ & + \delta_{RC} \int_{t_{i-1}}^{t} z_{RC}(u)du - \mu \int_{t_{i-1}}^{t} \min\{s_i, z_Q(u)\}du \\ & - \theta \int_{t_{i-1}}^{t} \left(z_Q(u) - s_i\right)^+ du, \qquad (4.12) \\ z_{RD}(t) = {} & z_{RD}(t_{i-1}) + p\theta \int_{t_{i-1}}^{t} z_Q(u) - s_i)^+ du \\ & - \delta_{RD} \int_{t_{i-1}}^{t} z_{RD}(u)du, \qquad (4.13) \\ z_{RC}(t) = {} & z_{RC}(t_{i-1}) + q\mu \int_{t_{i-1}}^{t} \min\{s_i, z_Q(u)\}du \\ & - \delta_{RC} \int_{t_{i-1}}^{t} z_{RC}(u)du. \qquad (4.14) \end{aligned}$$

Numerically solving Equations (4.12)-(4.14) is similar to solving Equations (4.9)-(4.11), thus, we do not elaborate on the procedure here.

In reality, parameters such as μ, θ, δ_{RD} and δ_{RC} can also be time-dependent, and vary per interval. For example, δ_{RD} may be bigger in the late afternoon than in the morning, since abandoned customers want to have responses by the end of the day. It is possible to extend the model in Equations (4.12)-(4.14) to adapt such situation by simply making the parameters time dependent. In this chapter, for the simplicity of validation, we will not consider such cases.

4.3.4 Model under stationarity

We have just shown that one can numerically solve differential equations (4.9)-(4.11) to obtain the fluid limit $\mathbf{z}(t)$. We now derive the stationary fluid limit, i.e., we develop conditions under which $\mathbf{z}(t)$ is constant.

By taking the derivative of Equations (4.9)-(4.11) and assuming that $\frac{d}{dt}\mathbf{z}(t) = 0$ has a constant solution, we can obtain

$$\begin{aligned} 0 &= \lambda_F + \delta_{RD} z_{RD}(\infty) + \delta_{RC} z_{RC}(\infty) - \mu \min\{s, z_Q(\infty)\} - \theta \left(z_Q(\infty) - s\right)^+, && (4.15) \\ 0 &= p\theta \left(z_Q(\infty) - s\right)^+ - \delta_{RD} z_{RD}(\infty), && (4.16) \\ 0 &= q\mu \min\{s, z_Q(\infty)\} - \delta_{RC} z_{RC}(\infty), && (4.17) \end{aligned}$$

where $z_Q(\infty) := \lim_{t\to\infty} z_Q(t), z_{RD}(\infty) := \lim_{t\to\infty} z_{RD}(t), z_{RC}(\infty) := \lim_{t\to\infty} z_{RC}(t)$.

Equations (4.15)-(4.17) can be easily solved with respect to $z_Q(\infty), z_{RD}(\infty)$ and $z_{RC}(\infty)$, yielding

$$z_Q(\infty) = \begin{cases} \dfrac{\lambda_F}{(1-q)\,\mu}, & \text{if } \rho < (1-q) \\ \dfrac{\lambda_F + q\mu s - \mu s}{\theta\,(1-p)} + s, & \text{if } \rho \geq (1-q) \end{cases} \tag{4.18}$$

$$z_{RD}(\infty) = \begin{cases} 0, & \text{if } \rho < (1-q) \\ \dfrac{p\theta\,(z_Q(\infty) - s)}{\delta_{RD}}, & \text{if } \rho \geq (1-q) \end{cases} \tag{4.19}$$

$$z_{RC}(\infty) = \begin{cases} \dfrac{q\mu z_Q(\infty)}{\delta_{RC}}, & \text{if } \rho < (1-q) \\ \dfrac{q\mu s}{\delta_{RC}}, & \text{if } \rho \geq (1-q)\,. \end{cases} \tag{4.20}$$

The results above would offer some insights. $\rho := \frac{\lambda_F}{s\mu}$ is the load of the system due to the fresh arrivals. However, the total load into the system is at least $\hat{\rho} := \frac{\lambda_F}{(1-q)s\mu}$, since $\frac{1}{1-q}$ portion of ρ will reconnect. In the case of $\hat{\rho} < 1$, we have $z_Q(\infty) < s$ and $z_{RD}(\infty) = 0$. This means there is no abandonment at all in the stationary fluid limit when $\hat{\rho} < 1$, and the stationary fluid limit do not depend on δ_{RD} at all. In reality, due to the variabilities in the arrival process, HT and the patience, abandonments would not be 0 though, but very small numbers. One would expect that the fluid approximation has high approximation errors in such a case. If $\hat{\rho} > 1$, by Equation (4.18), $z_Q(\infty) > s$. Therefore, in this case, the stationary fluid limit indicates that there will be $(z_Q(\infty) - s)$ amount of customers waiting in the queue, each abandons the system with rate θ, thus, the total abandonment rate is then $p\theta(z_Q(\infty) - s)$.

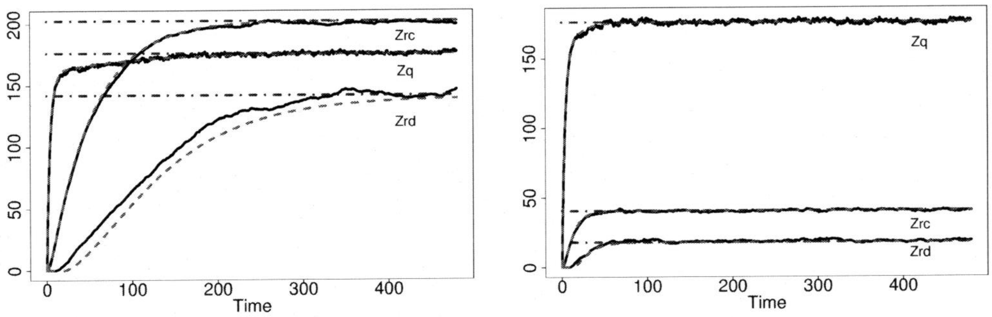
As we mentioned before, we could not obtain an analytical expression for the steady state probability of the original system, thus, the stability condition of the original system is then also difficult to derive. Our fluid limit is not for stability but for approximation of a many-arrivals many-servers system, thus, it cannot derive the exact stability condition. However, Equations (4.18)-(4.20) could give some insight. If we consider $z_Q(\infty), z_{RD}(\infty)$ and $z_{RC}(\infty)$ being less than ∞ as the fluid limit being stable, then following conditions are necessary for the stability of the original system: in the case of $\rho/(1-q) < 1$, one requires $\delta_{RC} > 0$; and in the case of $\rho/(1-q) > 1$, one requires $q < 1, \theta > 0, p < 1, \delta_{RD} > 0$ and $\delta_{RC} > 0$.

4.4 Validation of the fluid limit

In this section, we validate the fluid model via simulation both for a single interval and for multiple intervals. We simulate the system for 480 minutes of time, i.e., 8 hours, which correspond to the busy hours in some call centers. The results obtained via the fluid limit are compared with the simulation results. Since $\mathbf{Z}(t)$ is a stochastic process, it has variability. To reduce the effects of variabilities in the results, we do the simulation 100 times, and then take the average.

4.4.1 Validation of a single interval

We start with the simple case of a single interval, where $\lambda_F(t) = \lambda_F$, for all $t > 0$, and we assume that s, μ as well as other parameters are constants over time. We compare $\mathbf{z}(t)$ (computed via Equations (4.9)-(4.11)) with $\mathbf{Z}(t)$ (simulation results), and with $\mathbf{z}(\infty) := (z_Q(\infty), z_{RD}(\infty), z_{RC}(\infty))^T$ (computed via Equations (4.18)-(4.20)) for different values of $\hat{\rho}$ and λ_F. For each value of $\hat{\rho}$, s changes, while $1/\mu = 4$, $p = 0.5$, $q = 0.1$, $\theta = 0.5$ remain the same. We consider two scenarios; in the first scenario, we let $1/\delta_{RD} = 40$ minutes and $1/\delta_{RC} = 50$ minutes, which correspond to the real values from the data; in the second scenario, we let $1/\delta_{RD} = 5$ minutes and $1/\delta_{RC} = 10$ minutes, which represents the case with "impatient" customers, in the sense that they spend little time in the redial and reconnect orbits. We also consider two different values for λ_F, i.e., $\lambda_F = 10$ for relatively small call centers and $\lambda_F = 40$ for relatively large call centers. Two examples of $\mathbf{z}(t)$ and $\mathbf{Z}(t)$, where $\lambda_F = 40, \hat{\rho} = 1.1$ and $\hat{\rho} = 1.2$, are shown in Figures 4.4 and 4.5, respectively.



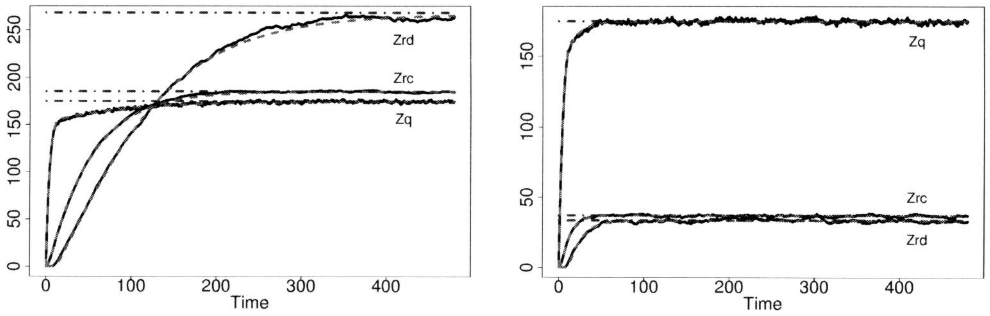
(a) $1/\delta_{RD} = 40$ and $1/\delta_{RC} = 50$.

(b) $1/\delta_{RD} = 5$ and $1/\delta_{RC} = 10$.

Figure 4.4: Simulation results (solid curve), fluid approximations (dashed curve) and stationary fluid limit (dot-dashed curve), $\lambda_F = 40, \hat{\rho} = 1.1$.

One can see from Figures 4.4 and 4.5 that the systems start with zero customers, and as time passes by, $Z_Q(t), Z_{RD}(t)$ and $Z_{RC}(t)$ gradually build up and reach stationarity. These stationarity levels are well approximated by $\mathbf{z}(\infty)$. Furthermore, in both parameter settings, the fluid limits offer close approximations to the original processes, especially for $Z_Q(t)$ and $Z_{RC}(t)$. The approximation error is larger for $Z_{RD}(t)$, especially when $\hat{\rho} = 1.1$. We now explain why. The fluid limits ignore the variability in the number of customers in the queue; when the queue length is not large, such as the period when $Z_{RD}(t)$ does not reach stationarity and is relatively small, ignoring variability can lead to relatively large errors.

Obtaining an approximation to $\mathbf{Z}(t)$ is the intermediate step for calculating $\lambda_{RD}(t)$ and



(a) $1/\delta_{RD} = 40$ and $1/\delta_{RC} = 50$.

(b) $1/\delta_{RD} = 5$ and $1/\delta_{RC} = 10$.

Figure 4.5: Simulation results (solid curve), fluid approximations (dashed curve) and stationary fluid limit (dot-dashed curve), $\lambda_F = 40, \hat{\rho} = 1.2$.

$\lambda_{RC}(t)$. Therefore, for the purpose of testing the errors of the fluid model in number of redials and reconnects, we introduce the error measurements e_{RD} and e_{RC}, which are defined by

$$e_{RD} := \frac{\int_0^T |E\lambda_{RD}(u) - \lambda_{RD}^{fl}(u)|du}{\int_0^T E\lambda_{RD}(u)du} = \frac{\int_0^T |EZ_{RD}(u) - z_{RD}(u)|du}{\int_0^T EZ_{RD}(u)du},$$
$$e_{RC} := \frac{\int_0^T |E\lambda_{RC}(u) - \lambda_{RC}^{fl}(u)|du}{\int_0^T E\lambda_{RC}(u)du} = \frac{\int_0^T |EZ_{RC}(u) - z_{RC}(u)|du}{\int_0^T EZ_{RC}(u)du},$$

where $\lambda_{RD}^{fl}(t)$ and $\lambda_{RC}^{fl}(t)$ are the arrival rate due to redial and reconnect in the fluid approximation, respectively, and $T = 480$, as the same length of the simulation time. The parameters and results are shown in Table 4.1.

One can see from Tables 4.1 and 4.2 that for the number of reconnects, the fluid model offers good approximations for both scenarios with all values of $\hat{\rho}$. However, for the number of redials, the fluid model performs badly when $\hat{\rho} < 1.1$. This corresponds to the lingering condition pointed out by Mandelbaum et al. (2002), which states that the fluid limit leads to significant inaccuracy when the system stays critically loaded (i.e., $\hat{\rho}$ close to 1 in our case) for a long time. In the next section, we will show that the consequences of these bad performances are not severe in terms of SL_2 and r. In addition, by comparing two different fresh arrival rates from Tables 4.1 and 4.2, we could see that the fluid approximation performs better for bigger call centers.

4.4.2 Validation of multiple intervals

Similar to the validation procedure in the case of a single interval, now we validate the performance of the fluid model for multiple intervals. We divide 480 minutes of simulation time into 16 intervals with duration 30 minutes. The fresh arrival rate $\lambda_{F,i}$

$\hat{\rho}$	$\lambda_F = 10$		$\lambda_F = 40$	
	e_{RD}	e_{RC}	e_{RD}	e_{RC}
1.01	91.1%	5.4%	82.9%	2.4%
1.05	45.8%	2.3%	27.0%	1.2%
1.1	19.6%	1.9%	7.8%	0.8%
1.2	7.2%	2.1%	1.3%	0.7%
1.3	1.8%	1.2%	0.8%	0.5%
1.4	1.4%	1.6%	0.4%	0.5%
1.5	1.5%	1.9%	0.6%	0.8%

Table 4.1: Approximation errors in a single interval, $1/\delta_{RD} = 40, 1/\delta_{RC} = 50$.

$\hat{\rho}$	$\lambda_F = 10$		$\lambda_F = 40$	
	e_{RD}	e_{RC}	e_{RD}	e_{RC}
1.01	85.7%	5.6%	74.3%	1.9%
1.05	39.9%	4.0%	21.2%	1.1%
1.1	16.3%	3.1%	5.4%	1.4%
1.2	5.6%	3.0%	2.8%	1.3%
1.3	4.4%	3.6%	1.8%	1.2%
1.4	2.7%	2.5%	1.6%	1.3%
1.5	2.6%	2.6%	1.3%	1.5%

Table 4.2: Approximation errors in a single interval, $1/\delta_{RD} = 5, 1/\delta_{RC} = 10$.

is assumed to be piece-wise constant within each interval, but it varies from interval to interval. The fresh arrival pattern is shown in Figure 4.6. This arrival pattern mimics the situation in reality, where there is a morning peak hour and an afternoon peak hour. We validate our approximation for different values of $\hat{\rho}$, and for given value of $\hat{\rho}$, $s_i = \frac{\hat{\rho}\mu(1-q)}{\lambda_{F,i}}$, for $i = 1, 2, \dots, 16$. Other parameters are taken to be same as in the case of a single interval.

We omit the figure for $\mathbf{Z}(t)$, since they are similar to the graph in Figure 4.5. The results for e_{RD} and e_{RC} are shown in Table 4.3.

$\hat{\rho}$	e_{RD}	e_{RC}	e_{RD}	e_{RC}
1.01	55.5%	1.8%	57.3%	2.1%
1.05	25.8%	1.4%	23.0%	2.0%
1.1	9.8%	1.2%	9.8%	1.2%
1.2	1.3%	0.8%	4.0%	1.6%
1.3	1.1%	0.9%	1.6%	1.5%
1.4	0.7%	0.8%	1.9%	1.7%
1.5	0.7%	0.9%	1.6%	1.6%

Table 4.3: Approximation errors in multiple intervals.

Similar to Tables 4.1 and 4.2, one can see from Table 4.3 that the fluid model gives close

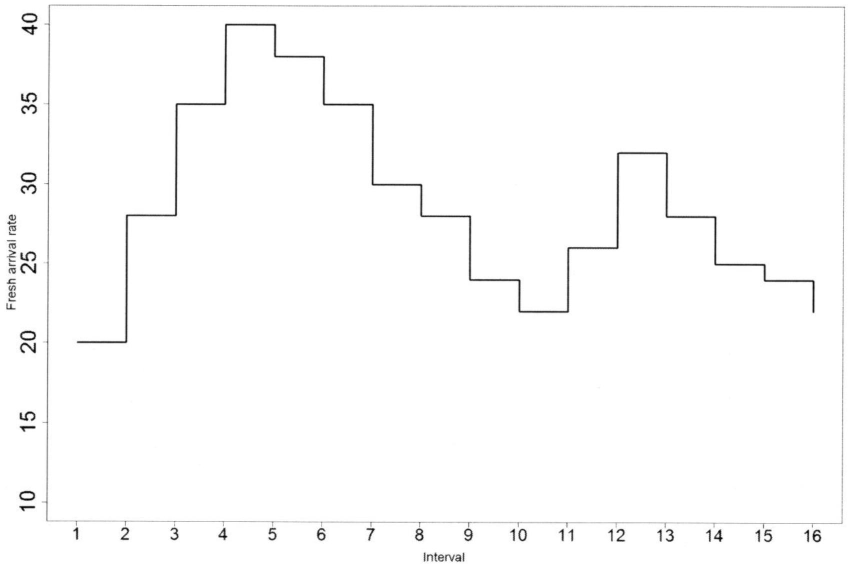


Figure 4.6: Fresh arrival rate per interval.

approximations for the number of reconnects for all values of $\hat{\rho}$, and the approximations for the number of redials gets more accurate when $\hat{\rho} > 1.1$.

4.5 Erlang A approximation

The fluid model gives first order approximations for $Z_Q(t)$, $Z_{RD}(t)$ and $Z_{RC}(t)$. Based on them, we can approximate the expected total arrival rate and expected number of customers in the queue for any time t, from which the expected waiting time can be obtained. However, this is not the ultimate goal, since it gives no information about the waiting time distribution of random customers, which is one of the most used call center performance indicators. Therefore, to this end, we will apply the Erlang A formula to approximate the waiting time distribution. We assume $\lambda_T(t)$ to be the arrival rate of the Erlang A model, whose mean can be obtained via Equation (4.1).

The reason to use the Erlang A model is intuitively clear, since the redial and reconnect behaviors have only direct influence on the total arrival rate, it has no direct influence on the service, such as the HT. Therefore, once the total arrival rate $\lambda_T(t)$ is given, $Z_{RD}(t)$ and $Z_{RC}(t)$ become irrelevant to what happens in the queue, thus, we can treat the system as an Erlang A system by ignoring the redial and reconnect orbits. Note that this is only an approximation of the Erlang A system, since the arrival process is generally not Poisson.

The analytical expressions for the waiting time distribution and the expected r of the

Erlang A model are known. We refer to Deslauriers et al. (2007) and Roubos (2012) for the Erlang A formula and the calculation details.

Now, we evaluate the performance of the Erlang A approximation. To save space, we only evaluate the performances in the case of multiple intervals. The arrival pattern is the same as shown in Figure 4.6. Given all parameters, we compute $\mathbf{z}(t)$ via Equations (4.12)-(4.14). After that, $\lambda_T(t)$ can be obtained via Equation (4.1). $\lambda_T(t)$ will be the input as the arrival rate of the Erlang A formula, from which the SL_2 and r can be obtained. We conduct such a procedure for different values of $\hat{\rho}$, δ_{RD} and δ_{RC}.

We denote SL_2^{sim} and SL_2^{a} as the SL_2 from simulation and from the Erlang A approximation, respectively. We let the acceptable waiting time be 0.5 minute. r from simulation and from the Erlang A approximation are denoted as r^{sim} and r^{a}, respectively. Besides the SL_2 and r, we also compare the probability of waiting from simulation, i.e., P_w^{sim}, with that from the Erlang A approximation, i.e., P_w^{a}. The results are shown in Tables 4.4 and 4.5. In Table 4.4, we let $1/\delta_{RD} = 40$ minutes and $1/\delta_{RC} = 50$ minutes, which are taken from a real call center data. In Table 4.5, we set $\delta_{RD} = 5$ minutes and $\delta_{RC} = 10$ minutes, which represents situations where customers spend short times in the redial and reconnect orbits. In both scenarios, we fix $\mu = 1/4$, $\theta = 0.5$, $p = 0.5$ and $q = 0.1$.

$\hat{\rho}$	SL_2^{sim}	SL_2^{a}	r^{sim}	r^{a}	P_w^{sim}	P_w^{a}
1.01	89.2%	92.3%	6.1%	4.9%	50.7%	46.5%
1.05	81.3%	84.6%	9.4%	8.4%	66.3%	65.2%
1.1	67.7%	69.8%	14.2%	13.7%	81.1%	82.9%
1.2	38.1%	37.4%	23.7%	23.8%	93.9%	96.3%
1.3	17.1%	15.2%	32.2%	32.2%	97.3%	99.1%
1.4	7.6%	5.6%	38.9%	39.1%	98.9%	99.7%
1.5	3.8%	2.2%	44.5%	44.6%	99.1%	99.9%

Table 4.4: Approximation errors of the Erlang A approximation, $1/\delta_{RD} = 40$ and $1/\delta_{RC} = 50$.

$\hat{\rho}$	SL_2^{sim}	SL_2^{a}	r^{sim}	r^{a}	P_w^{sim}	P_w^{a}
1.01	87.8%	91.7%	6.7%	5.3%	54.5%	50.6%
1.05	78.3%	82.7%	10.6%	9.4%	70.9%	71.3%
1.1	63.4%	65.6%	15.6%	15.3%	84.4%	88.3%
1.2	31.4%	29.6%	25.7%	25.8%	95.7%	95.6%
1.3	11.8%	9.0%	34.3%	34.4%	98.3%	99.9%
1.4	4.5%	2.4%	41.0%	41.3%	99.4%	99.9%
1.5	2.1%	0.7%	45.5%	46.8%	99.3%	99.9%

Table 4.5: Approximation errors of the Erlang A approximation, $1/\delta_{RD} = 5$ and $1/\delta_{RC} = 10$.

Based on the results in Tables 4.4 and 4.5, we can see that the Erlang A model offers close approximations both for the SL_2 and r in all values of $\hat{\rho}$. The approximation errors in probability of waiting is larger, but they are bounded by 5% in all scenarios. The Erlang

A approximation performs better when $1/\delta_{RD}$ and $1/\delta_{RC}$ are larger, i.e., with error less than 2% in SL_2, and 1.2% in the r in Table 4.4. However, even for small values of $1/\delta_{RD}$ and $1/\delta_{RC}$, the errors are bounded by 5% in SL_2 and 2% in r, as shown in Table 4.5. The approximation results in Table 4.4 are of special interest, since the parameters are taken to mimic real call centers. One might notice that even though we have large errors in e_{RD} when $\hat{\rho} < 1.1$ in Tables 4.1 and 4.2, the errors in SL_2 and r are small in Tables 4.4 and 4.5. This is caused by the fact that when $\hat{\rho} < 1.1$, the number of redials is small compared to the number of reconnects, thus, the errors in number of redials do not have a big influence on $\lambda_T(t)$.

4.6 Conclusion

In this chapter, we investigate staffing of call centers with redials and reconnects. We consider call centers that operate under heavy load. The model can be described as a three-dimensional Markov process $\{\mathbf{Z}(t), t > 0\}$, defined in (4.2). However, to avoid the complexity of solving the Markov process, we use a fluid model to approximate $\mathbf{Z}(t)$. We show that the fluid limit is the unique solution of a set of three differential equations. Under the same fluid scaling, we derive the fluid limit of the queueing system in the non-stationary case to mimic the real situation in call centers, as the parameters can change before the system reaches stationarity. We also performed simulation experiments to assess the accuracy of the approximations. To apply the results to real call center applications, we take a further step by calculating the expected total arrival rate, and use this as an input to the Erlang A formula to calculate the SL_2 and r. Simulation results show that our approximation to SL_2 is accurate with error less than 2%, and the approximation to r has errors less than 1.5% when $\hat{\rho} \leq 1.05$ and less than 0.5% when $\hat{\rho} > 1.05$, when the parameters are taken from real data.

The results suggest a number of topics for further research. First, the current chapter is focused on the derivation and usage of fluid limits for staffing problems of large call centers featuring both redials and reconnects, with load per server greater than 1. As a next step, it is interesting to supplement the results presented here with the development of staffing methods for the case where the load is strictly less than 1. To this end, the results of the present chapter and the results for staffing large call centers without redials/reconnects (Borst et al. (2004), Roubos (2012), Sze (1984)) will serve as a good starting point. Second, with the presence of the redial and reconnect behaviors, it would be interesting to explicitly quantify the reduction of staffing costs while still meeting the target SL by more efficient planning of call center agents. Third, we use a first order approximation. It would also be interesting to derive the diffusion limit of this model, which suggests a second order approximation. This may lead to a more intuitive and simple staffing formula such as square-root staffing in the spirit of Halfin and Whitt (1981). Moreover, in this chapter, we neglect the slight difference between the holding times of the reconnects and those of the fresh calls. As an extension to this, one could relax this assumption and study the correlation between the holding times of the fresh calls and its corresponding reconnects. Last but not the least, besides the influences in call centers staffing, the analysis of reconnect and redial behaviors can also offer insight to call center management. For example, by looking at the reconnect probability of each agent, managers can have some overview information of the quality of service offered by each agent. Furthermore, often the agents have some control on the holding time of each call, and by looking at the correlation between the reconnect probability and the holding time of each call, manager may find the "right" amount of holding time of each call, such that the holding time and the quality of service is well balanced.

4.7 Proofs

In this section, we show the proof of Lemma 4.1 and Theorem 4.1. First we introduce some notations.

Dividing by n on both sides of Equations (4.6)-(4.8), we have

$$\bar{Z}_Q^{(n)}(t) = \bar{Z}_Q^{(n)}(0) + G_Q^{(n)}\left(\bar{\mathbf{Z}}^{(n)}\right)(t) + \int_0^t H_Q\left(\bar{\mathbf{Z}}^{(n)}\right)(u)\,du, \tag{4.21}$$

$$\bar{Z}_{RD}^{(n)}(t) = \bar{Z}_{RD}^{(n)}(0) + G_{RD}^{(n)}\left(\bar{\mathbf{Z}}^{(n)}\right)(t) + \int_0^t H_{RD}\left(\bar{\mathbf{Z}}^{(n)}\right)(u)\,du, \tag{4.22}$$

$$\bar{Z}_{RC}^{(n)}(t) = \bar{Z}_{RC}^{(n)}(0) + G_{RC}^{(n)}\left(\bar{\mathbf{Z}}^{(n)}\right)(t) + \int_0^t H_{RC}\left(\bar{\mathbf{Z}}^{(n)}\right)(u)\,du, \tag{4.23}$$

where

$$\begin{aligned} G_Q^{(n)}\left(\bar{\mathbf{Z}}^{(n)}\right)(t) := {} & \left(\frac{\Pi_{\lambda_F n}^{(n)}(t)}{n} - \lambda_F t\right) - \left(\bar{D}_s^{(n)}(t) - \int_0^t \mu \min\{s, \bar{Z}_Q^{(n)}(u)\}du\right) \\ & - \left(\bar{D}_a^{(n)}(t) - \int_0^t \theta\left(\bar{Z}_Q^{(n)}(u) - s\right)^+ du\right) \\ & + \left(\bar{D}_{RD}^{(n)}(t) - \int_0^t \delta_{RD}\bar{Z}_{RD}^{(n)}(u)\,du\right) \\ & + \left(\bar{D}_{RC}^{(n)}(t) - \int_0^t \delta_{RC}\bar{Z}_{RC}^{(n)}(u)\,du\right), \end{aligned} \tag{4.24}$$

$$\begin{aligned} G_{RD}^{(n)}\left(\bar{\mathbf{Z}}^{(n)}\right)(t) := {} & \left(\sum_{j=1}^{n\bar{D}_a^{(n)}(t)} B_j(p)/n - \int_0^t p\theta\left(\bar{Z}_Q^{(n)}(u) - s\right)^+ du\right) \\ & - \left(\bar{D}_{RD}^{(n)}(t) - \int_0^t \delta_{RD}\bar{Z}_{RD}^{(n)}(u)\,du\right), \end{aligned} \tag{4.25}$$

$$
\begin{aligned}
G_{RC}^{(n)}\left(\bar{\mathbf{Z}}^{(n)}\right)(t) := {} & \left(\sum_{j=1}^{n\bar{D}_s^{(n)}(t)} B_j\left(q\right)/n - \int_0^t q\mu \min\{s, \bar{Z}_Q^{(n)}(u)\}du\right) \\
& - \left(\bar{D}_{RC}^{(n)}\left(t\right) - \int_0^t \delta_{RC}\bar{Z}_{RC}^{(n)}\left(u\right)du\right),
\end{aligned}
\tag{4.26}
$$

and

$$
\bar{D}_s^{(n)}(t) = \Pi_1\left(n\int_0^t \mu \min\{s, \bar{Z}_Q^{(n)}(u)\}du\right)/n, \tag{4.27}
$$

$$
\bar{D}_a^{(n)}(t) = \Pi_2\left(n\int_0^t \theta\left(\bar{Z}_Q^{(n)}\left(u\right) - s\right)^+ du\right)/n, \tag{4.28}
$$

$$
\bar{D}_{RD}^{(n)}(t) = \Pi_3\left(n\int_0^t \delta_{RD}\bar{Z}_{RD}^{(n)}\left(u\right)du\right)/n, \tag{4.29}
$$

$$
\bar{D}_{RC}^{(n)}(t) = \Pi_4\left(n\int_0^t \delta_{RC}\bar{Z}_{RC}^{(n)}\left(u\right)du\right)/n, \tag{4.30}
$$

and

$$
\begin{aligned}
\int_0^t H_Q\left(\bar{\mathbf{Z}}^{(n)}\right)(u)\,du := {} & \int_0^t \lambda_F + \delta_{RD}\bar{Z}_{RD}^{(n)}\left(u\right) + \delta_{RC}\bar{Z}_{RC}^{(n)}\left(u\right) \\
& - \mu\min\{s, \bar{Z}_Q^{(n)}\left(u\right)\} - \theta\left(\bar{Z}_Q^{(n)}\left(u\right) - s\right)^+ du, \\
\int_0^t H_{RD}\left(\bar{\mathbf{Z}}^{(n)}\right)(u)\,du := {} & \int_0^t p\theta\left(\bar{Z}_Q^{(n)}\left(u\right) - s\right)^+ - \delta_{RD}\bar{Z}_{RD}^{(n)}\left(u\right)du, \\
\int_0^t H_{RC}\left(\bar{\mathbf{Z}}^{(n)}\right)(u)\,du := {} & \int_0^t q\mu\min\{s, \bar{Z}_Q^{(n)}\left(u\right)\} - \delta_{RC}\bar{Z}_{RC}^{(n)}\left(u\right)du.
\end{aligned}
$$

For the convenience of notation, we rewrite Equations (4.21)-(4.23) in the vector form

$$
\bar{\mathbf{Z}}^{(n)}\left(t\right) = \bar{\mathbf{Z}}^{(n)}\left(0\right) + \mathbf{G}^{(n)}\left(\bar{\mathbf{Z}}^{(n)}\right)(t) + \int_0^t \mathbf{H}\left(\bar{\mathbf{Z}}^{(n)}\right)(u)\,du, \tag{4.31}
$$

where

$$
\begin{aligned}
\mathbf{G}^{(n)}\left(\bar{\mathbf{Z}}^{(n)}\right)(t) &:= \left(G_Q^{(n)}\left(\bar{\mathbf{Z}}^{(n)}\right)(t), G_{RD}^{(n)}\left(\bar{\mathbf{Z}}^{(n)}\right)(t), G_{RC}^{(n)}\left(\bar{\mathbf{Z}}^{(n)}\right)(t)\right)^T, \\
\mathbf{H}\left(\bar{\mathbf{Z}}^{(n)}\right)(u) &:= \left(H_Q\left(\bar{\mathbf{Z}}^{(n)}\right)(u), H_{RD}\left(\bar{\mathbf{Z}}^{(n)}\right)(u), H_{RC}\left(\bar{\mathbf{Z}}^{(n)}\right)(u)\right)^T.
\end{aligned}
$$

4.7.1 Proof of Lemma 4.1

In order to show that $\{\bar{\mathbf{Z}}^{(n)}(\cdot)\}_{n=1}^{\infty}$ is relatively compact with continuous limits, it is sufficient to show the following two properties (see Corollary 7.4 and Theorem 10.2 of Ethier and Kurtz (1986)).

1. Compact Containment: for any $T \geq 0, \epsilon > 0$, there exists a compact set $\Gamma_T \subset \mathbb{R}^3$ such that

$$\mathrm{P}\left(\bar{\mathbf{Z}}^{(n)}(t) \in \Gamma_T,\ t \in [0,T]\right) \to 1, \quad \text{as } n \to \infty;$$

2. Oscillation Control: for any $\epsilon > 0$, and $T \geq 0$, there exists a $\delta > 0$, such that

$$\limsup_{n\to\infty} \mathrm{P}\left(\omega\left(\bar{\mathbf{Z}}^{(n)}, \delta, T\right) \geq \epsilon\right) \leq \epsilon, \tag{4.32}$$

where

$$\omega(\mathbf{x}, \delta, T) := \sup_{\substack{v,t\in[0,T] \\ |t-v|<\delta}} \max_{j\in J} |x_j(t) - x_j(v)|,$$

and $J := \{Q, RD, RC\}$.

Proof of Compact Containment property:
The following trivial upper bound holds for the total number of customers in the system (only arrivals are taken into account and no departures): for $t \in [0,T]$,

$$\bar{Z}_Q^{(n)}(t) + \bar{Z}_{RD}^{(n)}(t) + \bar{Z}_{RC}^{(n)}(t) \leq \bar{Z}_Q^{(n)}(0) + \bar{Z}_{RD}^{(n)}(0) + \bar{Z}_{RC}^{(n)}(0) + \Pi_{\lambda_F n}^{(n)}(T)/n.$$

Since $\Pi_{\lambda_F n}^{(n)}(\cdot)$ is a Poisson process of rate $\lambda_F n$, by the Law of Large Numbers (LLN), we have

$$\Pi_{\lambda_F n}^{(n)}(T)/n \xrightarrow{d} \lambda_F T \quad \text{as } n \to \infty.$$

By the assumption of Theorem 4.1, we have

$$\bar{Z}_Q^{(n)}(0) + \bar{Z}_{RD}^{(n)}(0) + \bar{Z}_{RC}^{(n)}(0) \xrightarrow{d} z_Q(0) + z_{RD}(0) + z_{RC}(0).$$

Hence

$$\mathrm{P}(\bar{\mathbf{Z}}^{(n)}(t) \in \Gamma_T,\ t \in [0,T]) \to 1 \quad \text{as } n \to \infty,$$

where $\Gamma_T = \{(x_1, x_2, x_3) \mid x_1 + x_2 + x_3 \leq z_Q(0) + z_{RD}(0) + z_{RC}(0) + \lambda_F T + 1,\ x_1, x_2, x_3 \geq 0\}$, and the compact containment property indeed holds.

Proof of Oscillation Control property:

It follows from Equations (4.6)-(4.8) that, for all $v, t \geq 0$,

$$
\begin{aligned}
&|\bar{Z}_Q^{(n)}(t) - \bar{Z}_Q^{(n)}(v)| \leq |\Pi^{(n)}_{\lambda_F n}(t) - \Pi^{(n)}_{\lambda_F n}(v)|/n + \sum_{j \in \{s,a,RD,RC\}} |\bar{D}_j^{(n)}(t) - \bar{D}_j^{(n)}(v)|, \\
&|\bar{Z}_{RD}^{(n)}(t) - \bar{Z}_Q^{(n)}(v)| \leq \sum_{j \in \{a,RD\}} |\bar{D}_j^{(n)}(t) - \bar{D}_j^{(n)}(v)|, \\
&|\bar{Z}_{RC}^{(n)}(t) - \bar{Z}_Q^{(n)}(v)| \leq \sum_{j \in \{s,RC\}} |\bar{D}_j^{(n)}(t) - \bar{D}_j^{(n)}(v)|,
\end{aligned}
$$

where the processes $\bar{D}_j^{(n)}(\cdot)$ are defined by (4.28)-(4.30).

Also, from the Compact Containment property, we know that there exists a finite constant V such that

$$
\mathrm{P}\left(\underbrace{\bar{Z}_Q^{(n)}(u), \bar{Z}_{RD}^{(n)}(u), \bar{Z}_{RC}^{(n)}(u) \leq V,\ u \in [0,T]}_{=:\ \Omega_n} \right) \to 1 \quad \text{as } n \to \infty.
$$

On the event Ω_n, the following inequalities hold for all $v, t \in [0,T]$ such that $|t - v| \leq \delta$:

$$
\begin{aligned}
&\int_v^t \mu \min\{s, \bar{Z}_Q^{(n)}(u)\} du \leq c_1 \delta, && c_1 := \mu s, \\
&\int_v^t \theta \left(\bar{Z}_Q^{(n)}(u) - s \right)^+ du \leq c_2 \delta, && c_2 := \theta V, \\
&\int_v^t \delta_{RD} \bar{Z}_{RD}^{(n)}(u) du \leq c_3 \delta, && c_3 := \delta_{RD} V, \\
&\int_v^t \delta_{RC} \bar{Z}_{RC}^{(n)}(u) du \leq c_4 \delta, && c_4 := \delta_{RC} V.
\end{aligned}
$$

Employing formulas (4.27)-(4.29), we then get

$$
\begin{aligned}
\mathrm{P}\left(\omega\left(\bar{\mathbf{Z}}^{(n)}, \delta, T \right) \geq \epsilon \right) \leq{}& \mathrm{P}\left(\Omega_n' \right) + \mathrm{P}\left(\omega\left(\Pi^{(n)}_{\lambda_F n}(\cdot)/n, \delta, T \right) \geq \epsilon/5 \right) \\
&+ \sum_{j=1}^{4} \mathrm{P}\left(\omega\left(\Pi_j(n \cdot)/n, c_j \delta, c_j T \right) \geq \epsilon/5 \right),
\end{aligned}
$$

where

$$
\omega\left(\Pi^{(n)}_{\lambda_F n}(\cdot)/n, \delta, T \right) \xrightarrow{d} \lambda_F \delta, \quad \omega\left(\Pi_j(n \cdot)/n, c_j \delta, c_j T \right) \xrightarrow{d} c_j \delta, \quad 1 \leq j \leq 4,
$$

by the LLN for the Poisson processes $\Pi^{(n)}_{\lambda_F n}(\cdot)/n$, $\Pi_j(n \cdot)/n$ and by the continuity of the moduli of continuity $\omega(x(\cdot), \delta, T)$, $\omega(x(\cdot), c_j \delta, c_j T)$ with respect to $x(\cdot)$.

By the last two displays, the oscillation control property (4.32) indeed holds with any δ

such that $\lambda_F \delta < \epsilon/5, c_j \delta < \epsilon/5, 1 \leq j \leq 4$.

4.7.2 Proof of Theorem 4.1

In Lemma 4.1, we have shown that the sequence $\{\bar{\mathbf{Z}}^n(\cdot)\}_{n=1}^{\infty}$ is relatively compact with continuous limits, that is, from any subsequence $\{\bar{\mathbf{Z}}^{n_k}(\cdot)\}_{k=1}^{\infty}$, we can extract another subsequence $\{\bar{\mathbf{Z}}^{n_{k_l}}(\cdot)\}_{l=1}^{\infty}$ that converges weakly in $D([0,\infty),\mathbb{R}^3)$, say to a continuous process $z^*(t)$. We then call $z^*(t)$ a particular limit of the original sequence $\{\bar{\mathbf{Z}}^n(\cdot)\}_{n=1}^{\infty}$.

Consider an arbitrary particular limit $\mathbf{z}^*(\cdot)$ along a subsequence $\{\bar{\mathbf{Z}}^{n_k}(\cdot)\}_{k=1}^{\infty}$. If we can show that $\mathbf{z}^*(\cdot)$ satisfies Equations (4.9)-(4.11), and Equations (4.9)-(4.11) have a unique solution, then, due to the arbitrariness of $\mathbf{z}^*(\cdot)$, there must be a unique fluid limit defined by Equations (4.9)-(4.11).

We have

$$\bar{\mathbf{Z}}^{n_k}(\cdot) - \bar{\mathbf{Z}}^{n_k}(0) - \int_0^{\cdot} \mathbf{H}\left(\bar{\mathbf{Z}}^{n_k}\right) du = \mathbf{G}^{n_k}(\cdot). \tag{4.33}$$

On the one hand, since $\bar{\mathbf{Z}}^{n_k}(\cdot) \xrightarrow{d} \mathbf{z}^*(\cdot)$ as $k \to \infty$ and the limit $\mathbf{z}^*(\cdot)$ is continuous,

$$\bar{\mathbf{Z}}^{n_k}(\cdot) - \bar{\mathbf{Z}}^{n_k}(0) - \int_0^{\cdot} \mathbf{H}\left(\bar{\mathbf{Z}}^{n_k}\right) du \xrightarrow{d} \mathbf{z}^*(\cdot) - \mathbf{z}(0) - \int_0^{\cdot} \mathbf{H}\left(\mathbf{z}^*\right) du$$

by the continuous mapping theorem.

On the other hand, below we show that $\mathbf{G}^{n_k}(\cdot) \xrightarrow{d} 0$, and then (4.33) implies that

$$\bar{\mathbf{Z}}^{n_k}(\cdot) - \bar{\mathbf{Z}}^{n_k}(0) - \int_0^{\cdot} \mathbf{H}\left(\bar{\mathbf{Z}}^{n_k}\right) du \xrightarrow{d} 0.$$

As we combine the last two displays together, it follows that the particular limit $\mathbf{z}^*$ a.s. satisfies Equations (4.9)-(4.11). Also, the mapping $\mathbf{H}$ is Lipschitz continuous and then, by Lemma 1 in Reed and Ward (2004), Equations (4.9)-(4.11) have a unique solution. Hence, all particular fluid limits are the same, namely they coincide with the unique solution to (4.9)-(4.11).

It is left to show that $\mathbf{G}^{n_k}(\cdot) \xrightarrow{d} 0$.

By the LLN,

$$\Pi_1(n\cdot)/n - \cdot \xrightarrow{d} 0 \text{ in } D([0,\infty),\mathbb{R}),$$

and also, since $\bar{\mathbf{Z}}^{n_k}(\cdot) \xrightarrow{d} \mathbf{z}^*(\cdot)$ and $\mathbf{z}^*$ is continuous,

$$\int_0^{\cdot} \mu \min\{s, \bar{Z}_Q^{(n_k)}(u)\} du \xrightarrow{d} \int_0^{\cdot} \mu \min\{s, \mathbf{z}^*(u)\} du \text{ in } D([0,\infty),\mathbb{R}).$$

Then, by (4.27) and the Random time change Theorem in Billingsley (2009),

$$\bar{D}_s^{(n_k)}(t) - \int_0^t \mu \min\{s, \bar{Z}_Q^{(n_k)}(u)\} du \xrightarrow{d} 0 \text{ in } D([0,\infty),\mathbb{R}).$$

By the same argument, one can show that the other terms in $G_Q^{(n_k)}(\cdot)$ converge to 0, and that $G_{RC}^{(n_k)}(\cdot)$, $G_{RD}^{(n_k)}(\cdot)$ converge to 0, too. Hence, the proof of Theorem 4.1 is finished.

Chapter 5

A call center model with a call-back option

We study a call center model with a call-back option. In this call center, customers whose waiting time exceeds a given threshold receive a voice message mentioning the option to be served in priority if they call back later. This call-back option differs from the traditional ones found in the literature in that it is based on the experienced waiting time instead of the queue length, priority is given to call-backs instead of fresh calls and call-backs are inbound calls instead of outbound ones. Due to its particularities, the traditional approach of defining the number of customers in the system as the state does not apply here. Instead, we model it as a three-dimensional Markov process, with one dimension being a unit of a discretization of the waiting time. This is an approximation of the original model due to the presence of the discretization. We validate this approach via simulation for small call centers. Furthermore, we show via simulation that the call-back option leads to significant reduction in expected waiting time compared to the $M/M/s$ queueing model without call-back, given the same number of agents.

5.1 Introduction

We consider in this chapter the call-back option proposed and currently being used by an European call center. The idea is that customers who experienced only long waiting times (a waiting time that exceeds a given threshold) receive a voice message mentioning the option to be served in priority if they call back later.

The flexibility of the callback option comes from the willingness of some customers to accept future processing. The call center can then make use of this opportunity to better manage arrival uncertainty, which in turn would improve the system performance. There are a few papers on different call-back options in call centers. Armony and Maglaras (2004a) consider a model in which customers are given a choice of whether to wait online for their call to be answered or to leave a number and be called back within a specified time or to immediately balk. Upon arrival, customers are informed (or know from prior experience) of the expected waiting time if they choose to wait and the delay guarantee for the callback option. Their decision is probabilistic and based on this information. Under the heavy-traffic regime, Armony and Maglaras (2004a) develop an estimation scheme for the anticipated real-time delay that is asymptotically correct. They also propose an asymptotically optimal routing policy that minimizes real-time delay subject to a deadline on the postponed service mode. In Armony and Maglaras (2004b), the authors develop an asymptotically optimal routing rule, characterize the unique equilibrium regime of the system, and propose a staffing rule that picks the minimum number of agents that satisfies a set of operational constraints on the performance of the system.

Kim et al. (2012) consider a call center model with a call-back option where the capacity of the queue for the inbound calls is finite. Customer balking and abandonment are allowed. The authors provide an efficient algorithm for calculating the stationary probabilities of the system. Moreover, they derive the Laplace-Stieltjes transform of the sojourn time distribution of virtual customers. Dudin et al. (2013) consider a slightly different model, where agents make outbound calls to those lost customers. There are two agent teams, one that handles in priority inbound calls, and another that handles in priority outbound calls. Dudin et al. (2013) compute the stationary probabilities, and deduce from that some performance measures. They also numerically address the staffing issue of the two teams.

The call-back option we discuss in this chapter has three particularities compared to the traditional call-back option found in the call center literature. First, the decision to propose the call-back option is based on the experienced waiting time of a given customer and not on her expected waiting time at arrival or the number of customers in the queue. Second, call-backs receive priority over the fresh calls (initial attempts). These two particularities introduce more fairness to the model. For example, if one proposes the option of calling back with priority to an arriving customer with high expected waiting time, the rational decision for this customer would be to call back immediately, then she will receive priority, which is not fair for those fresh callers who arrived earlier than her and have waited longer times; on the other hand, if a call-back does not receive priority over those fresh calls in the queue, then this model is not fair to this call-back, since during her initial attempt, she has already waited certain threshold time units, which is longer than the time any fresh call has waited. The third particularity of this call-back option is that call-backs are also inbound calls, rather than the outbound calls considered in Armony and Maglaras (2004a). This feature has some advantages over having outbound call-backs. For example, it is more suitable for cost-driven call centers, since making outbound calls usually generates extra costs to call centers. In addition, a manual call-back (generated by an agent) may be a waste of time for the agent because of dialing time and the possible non-availability of the customer; on the other hand, an automated call-back is generated by an automated call dialer via predicative dialing, and sometimes call-backs are made to customers when there are no agents available, which is not desirable for the customer that is called back.

In this chapter, we present a call center model with a call-back option with fairness to customers. We model this system as a 3-dimensional Markov process, where the first dimension is the waiting time of the customers in line. For small call centers, we use this approach to show that if the right threshold is chosen, customers in this model would experience much shorter waiting times compared to those in the $M/M/s$ model. Numerical results show that there is an optimum threshold at which the call-back message should be presented to the customers. We numerically find those thresholds for various parameter settings.

The remainder of this chapter is structured as follows. In Section 5.2, we present the model as well as some notations. The formulation and the procedure to numerically solve a 3-dimensional Markov process is shown in Section 5.3. In Section 5.4, we validate our method, and compare the mean customer waiting time of this model to that of the $M/M/s$ model. Conclusions and possible topics for further research are shown in Section 5.5.

5.2 Modeling

In this section we explain the queueing model we use to analyze the call-back system. We consider a call center modeled as a multi-server queueing system with s identical, parallel agents. Calls arrive according to a homogeneous Poisson process with rate λ. The HT of a call is exponentially distributed with rate μ.

The call-back system works as follows. When a customer calls for the first time if at least one agent is available then this customer is directly served, otherwise she is routed to a first-come-first-served queue called queue 1. After having waited K time units, a waiting customer in queue 1 hears a voice message, saying that the call center is congested and that she has to call back later with the benefit of receiving priority. The connection terminates automatically after this customer hears the voice message. Therefore, if a customer is not served before K units of time then this customer is rejected from queue 1.

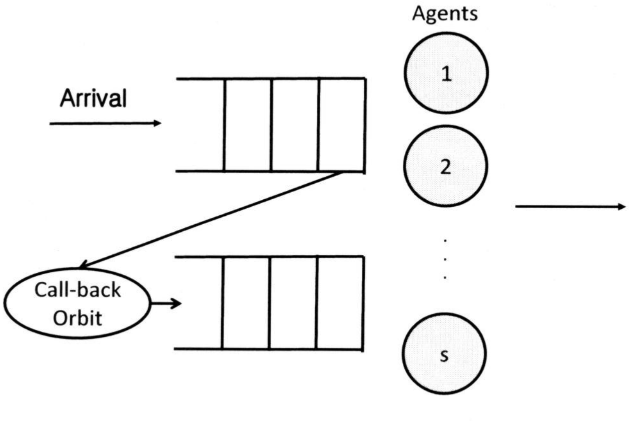


Figure 5.1: Diagram.

We assume that customers who are rejected from queue 1 enter a virtual call-back orbit, and they will stay in this call-back orbit for an exponentially distributed time Γ, with $E\Gamma = \gamma$ before they call back. When a customer calls back, if at least one agent is available then this customer is directly served; otherwise, her phone number is recognized and her call is considered as a call-back, and she is routed to a first-come-first-served queue called queue 2 instead of queue 1. The customers in queue 2 have non-preemptive priority over the customers in queue 1. Customers in both queues are handled by the same agent pool. Because of the urge for customers to obtain a service we ignore abandonment in the modeling, and we assume that all rejected customers will call back. A diagram of this model is shown in Figure 5.1.

We denote by W_1, a random variable, the waiting time of a customer in queue 1. W_2 is a random variable, the waiting time of a customer in queue 2; and $W_1 + W_2$ the waiting time of a customer in both queue 1 and queue 2. P_m is the probability of hearing the message in queue 1. We are interested in $E(W_1 + W_2)$ for given threshold K, since it measures the overall waiting times of customers in the system.

5.3 Performance analysis

In this section we develop a numerical method to compute the performance measures defined in the previous section. This system can be modeled as a Markov process, since the HT, the inter-arrival time and the orbit time Γ are all exponentially distributed. The traditional way of modeling a Markov process for a queueing system is by associating the number of customers in each queue with a dimension in the state space. This suggests that for this system, one could define the number of customers in queue 1 as the first dimension, the number of customers in queue 2 plus the number of customers in the service as the second dimension and the number of customers in the call-back orbit as the third dimension. However, this traditional approach does not apply here, since whether customers in queue 1 hear the message or not is based on their experienced waiting times, rather than the number of customers in queue 1. We thus propose to use a discretization of the waiting time of the first customer in line in queue 1, instead of using the number of customers in this queue. The modeling of the first in line as a tool for analyzing a queueing system was proposed by Koole et al. (2012).

Let us define the stochastic process $\{x(t), y(t), z(t), t \geq 0\}$, where for an instant $t \geq 0$, $x(t)$ denotes the stage of the waiting time of the first in line in queue 1 at time t, $y(t)$ denotes the number of busy agents plus the number of customers in queue 2 at time t and $z(t)$ denotes the number of customers in orbit at time t. We consider an exponential elapsing of time with parameter θ. This means that customers can wait maximum J stages in queue 1, with $J = K\theta$. Assume that the first customer in line leaves queue 1 at a waiting stage k ($k > 0$), then at her departure epoch, the probability of the next customer in queue 1 being in waiting stage $k - h$ is $p_{k,k-h}$, with

$$p_{k,k-h} = \left(\frac{\lambda}{\lambda + \theta}\right)\left(\frac{\theta}{\theta + \lambda}\right)^h,$$

for $k > 0$ and $0 \leq h < k$, and

$$p_{k,0} = \left(\frac{\theta}{\theta + \lambda}\right)^k.$$

We also truncate the number in queue 2 and the number in orbit by some sufficiently large number M. The parameters J, θ and M are the parameters to control the precision of the approximation.

The process $\{x(t), y(t), z(t), t \geq 0\}$ is a three-dimensional Markov process. To obtain the performance measures such as $\mathrm{E}(W_1 + W_2)$, one could formulate and numerically solve the global balance equations of the embeded Markov chain of this process, which results in steady state probability distribution. However, this is very time consuming and burdensome. For example, for a system with $J = 50$ and $M = 200$, it will lead to $2 \cdot 10^6$ equations. Thus, instead of solving the global balance equations, we iteratively compute the value functions for this process. This is a much simpler approach compared to solving the global balance equations, since one only needs to iteratively compute one value function. A disadvantage of this approach is that it only gives first order results. However, this is sufficient in our case, since we are mainly interested in $\mathrm{E}(W_1 + W_2)$.

We denote by $V_n(x, y, z)$ this value function after n iterations at state (x, y, z), and we let $V_0(x, y, z) = 0$ for $0 \leq x \leq J$, $0 \leq y, z \leq M$. The uniformization is done by the maximum event rate $\lambda + s\mu + M\gamma + \theta = \tau$. In the following relations we denote by $\mathbb{I}_A$ an indicator function of a given set A.

Therefore, the value iteration writes, for $0 \leq y, z \leq M$,

$$\begin{aligned}
V_{n+1}(0, y, z) &= \delta_2 \max(y - s, 0) + \delta_3 z \\
&+ \frac{\lambda}{\tau}\left(\mathbb{I}_{y<s} V_n(0, \min(y+1, M), z) + \mathbb{I}_{y \geq s} V_n(1, y, z)\right) \\
&+ z\frac{\gamma}{\tau} V_n(0, \min(y+1, M), \max(z-1, 0)) \\
&+ \frac{\mu}{\tau} \min(y, s) V_n(0, \max(y-1, 0), z) \\
&+ \frac{1}{\tau}(\mu(s - \min(y, s)) + (M - z)\gamma + \theta) V_n(0, y, z);
\end{aligned}$$

for $1 \leq x < J, s \leq y \leq M, 0 \leq z \leq M$,

$$\begin{aligned}
V_{n+1}(x, y, z) = &\frac{s\mu}{\lambda}\frac{\delta_1}{\theta} x + \delta_2 \max(y - s, 0) + \delta_3 z + \frac{\theta}{\tau} V_n(x+1, y, z) \\
&+ z\frac{\gamma}{\tau} V_n(x, \min(y+1, M), \max(z-1, 0)) \\
&+ \frac{\mu}{\tau} s(\mathbb{I}_{y=s} \sum_{h=0}^{x} p_{x,x-h} V_n(x-h, y, z) + \mathbb{I}_{y>s} V_n(x, \max(y-1, 0), z)) \\
&+ \frac{1}{\tau}(\lambda + (M - z)\gamma) V_n(x, y, z);
\end{aligned}$$

and for $s \leq y \leq M, 0 \leq z \leq M$,

$$\begin{aligned}
V_{n+1}(J,y,z) = {} & \frac{s\mu+\theta}{\lambda}\frac{\delta_1}{\theta}J + \delta_2 \max(y-s,0) + \delta_3 z \\
& + \frac{\theta}{\tau}\sum_{h=0}^{J} p_{J,J-h} V_n(J-h, y, \min(z+1, M)) \\
& + z\frac{\gamma}{\tau} V_n(J, \min(y+1, M), \max(z-1, 0)) \\
& + \frac{\mu}{\tau} s(\mathbb{I}_{y=s} \sum_{h=0}^{J} p_{J,J-h} V_n(J-h, y, z) + \mathbb{I}_{y>s} V_n(J, \max(y-1, 0), z)) \\
& + \frac{1}{\tau}(\lambda + (M-z)\gamma) V_n(J, y, z),
\end{aligned}$$

where δ_1, δ_2 and δ_3 are coefficients, and by changing the values of them, one could obtain different performance measures.

The long-term performance measures can be obtained through value iteration, by recursively evaluating V_n, for $n \geq 0$. When n goes to infinity the difference $V_{n+1}(x,y,z) - V_n(x,y,z)$ for $x, y, z \geq 0$ converges to the long-term average performance metrics. Thus, we stop the iteration when the following criterion is met:

$$\max_{x,y,z}\{V_{n+1}(x,y,z) - V_n(x,y,z)\} - \min_{x,y,z}\{V_{n+1}(x,y,z) - V_n(x,y,z)\} < \epsilon,$$

for some given ϵ.

Now we show how to obtain $\mathrm{E}(W_1 + W_2)$ via the aforementioned value iteration. First, note that by conditioning on whether a customer hears the message or not, one can derive that

$$\begin{aligned}
\mathrm{E}(W_1 + W_2) &= \mathrm{E}(W_1 + W_2 | \mathbb{M}) P_m + \mathrm{E}(W_1 + W_2 | \mathbb{M}^c)(1 - P_m) \\
&= K P_m + P_m \cdot \mathrm{E}(W_2 | \mathbb{M}) + (1 - P_m) \mathrm{E}(W_1 | \mathbb{M}^c) \\
&= \mathrm{E}W_1 + P_m \cdot \mathrm{E}(W_2 | \mathbb{M}), \qquad (5.1)
\end{aligned}$$

where $\mathbb{M}$ stands for the event that a customer hears the message, and $\mathbb{M}^c$ stands for the complement of $\mathbb{M}$.

Equation (5.1) suggests that we must first obtain $\mathrm{E}W_1, P_m$ and $\mathrm{E}(W_2|\mathbb{M})$ to compute $\mathrm{E}(W_1 + W_2)$. In fact, one can calculate $\mathrm{E}W_1$ simply by letting $\delta_1 = 1, \delta_2 = 0, \delta_3 = 0$ in the value iteration, and then calculate $V_{n+1}(x,y,z) - V_n(x,y,z)$ for any $x, y, z \geq 0$, since we multiply the variable x ($1 \leq x \leq J$) in the value functions by $\frac{s\mu}{\lambda}$ for $1 \leq x < J$ and by $\frac{s\mu+\theta}{\lambda}$ for $x = J$ so as to consider the state probability in the embedded Markov chain at service initiations or rejection epochs. In addition, by letting $\delta_1 = 0, \delta_2 = 0, \delta_3 = 1$, the value iteration will lead to the expected number of customers in the orbit, denoted by $\mathrm{E}N_o$. Given that the expected time in the orbit is $1/\gamma$ and that the rate going in the orbit is λP_m, with Little's law we compute P_m through the relation

$$P_m = \frac{\gamma}{\lambda} \mathrm{E} N_o. \tag{5.2}$$

Finally, we now show how to obtain $\mathrm{E}(W_2|\mathbb{M})$. By letting $\delta_1 = 0, \delta_2 = 1, \delta_3 = 0$, the value iteration will lead to $\mathrm{E}N_2$, where N_2 is the number of customers in queue 2. Then, using Little's law we can derive that

$$\mathrm{E}(W_2|\mathbb{M}) = \frac{\mathrm{E}N_2}{\lambda P_m}. \tag{5.3}$$

Combining Equations (5.1) to (5.3), we obtain

$$\mathrm{E}(W_1 + W_2) = \mathrm{E}W_1 + \frac{\mathrm{E}N_2}{\lambda}, \tag{5.4}$$

where $\mathrm{E}W_1$ can be obtained by letting $\delta_1 = 1, \delta_2 = 0, \delta_3 = 0$ in the value iterations, and $\mathrm{E}N_2$ can be obtained by letting $\delta_1 = 0, \delta_2 = 1, \delta_3 = 0$ in the value iterations. Note that this method requires a procedure of iterating over each state. Then the computation becomes expensive when the state space gets large. Therefore, this method is mainly useful for small call centers.

Remark 5.1. When $\Gamma \equiv 0$, then in stationarity, the waiting time distribution in this model is the same as the waiting time distribution in the $M/M/s$ queueing model. This result follows from the fact that the orders and times that the customers are served is the same for our model and for the $M/M/s$ queueing model when $\Gamma \equiv 0$.

5.4 Numerical results

In this section, we compare our approximation results with simulation results. We also numerically show that this model with a call-back option is better than the $M/M/s$ model, in the sense that the model with a call-back option leads to significant reduction in customer waiting time for the same number of agents.

We plot P_m and $\mathrm{E}(W_1 + W_2)$ from simulation and from value iteration method for different values of K and γ in Figures 5.2 and 5.3. We let $\epsilon = 10^{-5}$ in the value iterations, and J and M are set to be 200 and 80, respectively.

Figures 5.2 and 5.3 show that the results from value iterations are close to the results from simulations, especially for small values of K and γ. There are two reasons why the results from value iterations and those from simulations are not equivalent; first, the results from value iterations are approximations of the original system, since we truncate the system and make a discretization of the time; second, simulation results have also certain variabilities even when the system reaches stationarity. Also, Figure 5.2 reveals P_m decreases when K increases. Furthermore, one can notice that $\mathrm{E}(W_1 + W_2)$ is smaller when γ is smaller. This can be explained as follows. Customers mostly hear the message during the congested periods. If those customers call back very soon after having heard the message, then with high probability, they will also encounter a congested system

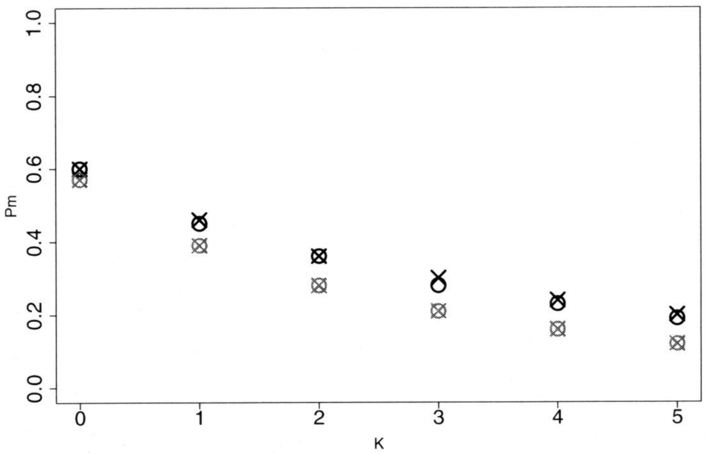


Figure 5.2: Comparing simulation results (cross) with approximation results (circle) for $\gamma = 0.1$ (dots above) and $\gamma = 0.01$ (dots below) for P_m ($s = 5, 1/\mu = 5, \lambda = 0.85$).

when they call back, which leads to long waiting times in queue 1. In addition, it is very interesting to see from both figures that there are optimal Ks, under which $\mathrm{E}(W_1 + W_2)$ is minimized.

To further illustrate the existence of the optimum K such that $\mathrm{E}(W_1 + W_2)$ is minimized, and compare this model with the $M/M/s$ model, we test more examples. In the following numerical study, we vary the load per agent ρ, defined by $\rho := \frac{\lambda}{s\mu}$, the value of γ, as well as the number of agents. We denote by K^* the optimum value of K such that $\mathrm{E}(W_1 + W_2)$ is minimized, and $\mathrm{E}W^*$ the value of $\mathrm{E}(W_1 + W_2)$ when $K = K^*$, and $\mathrm{E}W_{M/M/s}$ the expected waiting time in the $M/M/s$ queue. The results are shown in Tables 5.1-5.4. Note that in Tables 5.2 and 5.4, all the results for $s = 100$ are obtained via simulation, since the computation becomes very expensive for large call centers due to the large number of states. Also, we only look at the integer numbers for the values of K. To search for K^*, we start with $K = 0$, and compute $\mathrm{E}(W_1 + W_2)$, then we increment K by 1, and compute the corresponding $\mathrm{E}(W_1 + W_2)$. This procedure is repeated until a local optimum is reached. This procedure guarantees finding the global optimum K^* when $\mathrm{E}(W_1 + W_2)$ is unimodal in K.

As one can see from Tables 5.1-5.4, K^* increases when ρ increases. We now give an intuitive explanation for this observation. When ρ is relatively small, the system seldom gets congested, thus, when customers in queue 1 start to experience small amount of waiting, it makes sense to send them to the call-back orbit, since when they call back, they will experience relatively small amount of waiting time with high probability.

By comparing $\mathrm{E}W^*$ to $\mathrm{E}W_{M/M/s}$ in Tables 5.1-5.4, we see that the model with the call-back option is much better than the $M/M/s$ queueing model, since with the same number of agents, it leads to much smaller waiting times. The reduction is significant for all

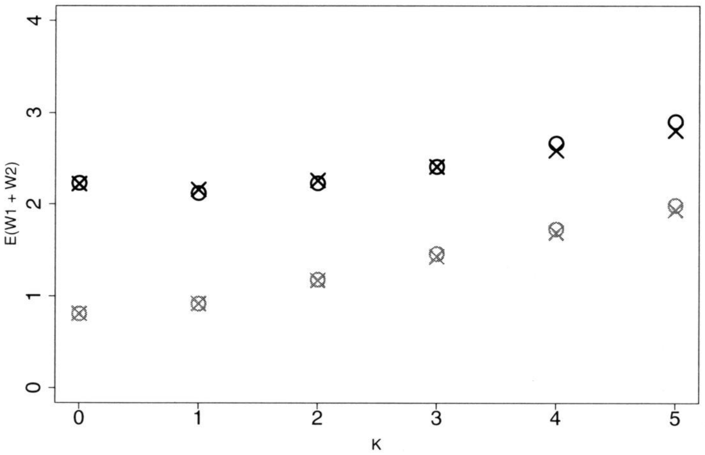


Figure 5.3: Comparing simulation results (cross) with approximation results (circle) for $\gamma = 0.1$ (dots above) and $\gamma = 0.01$ (dots below) for $E(W_1 + W_2)$ ($s = 5, 1/\mu = 5, \lambda = 0.85$).

	$s = 1$			$s = 5$		
ρ	K^*	EW^*	$EW_{M/M/s}$	K^*	EW^*	$EW_{M/M/s}$
0.5	0	3.15	5.00	0	0.04	0.26
0.6	0	5.22	7.50	0	0.12	0.59
0.7	2	8.86	11.67	0	0.35	1.26
0.8	5	16.28	20.00	0	1.14	2.77
0.9	8	39.65	45.00	2	4.47	7.62

Table 5.1: $\gamma = 0.1, 1/\mu = 5$.

cases except for the cases where $EW_{M/M/s}$ is close to 0. The reduction is caused by the fact that the call-back option flattens some variability in the arrival process. To be more specific, the Poisson arrival process has certain variability, and in the $M/M/s$ model, for those customers who arrive during a bursty period, they will experience relatively long waiting times; however, in this model with a call-back option, the arrivals during the bursty periods will delay their arrival time, and with a positive probability, their future arrival or call-backs will happen during a non-bursty period, thus, they will experience shorter waiting times in the model with a call-back option compared to the $M/M/s$ model.

5.5 Conclusion

In this chapter, we study a call center model with a call-back option, where customers that wait longer than a threshold K time units will hear a message mentioning that the

	$s = 20$			$s = 100$		
ρ	K^*	$\mathrm{E}W^*$	$\mathrm{E}W_{M/M/s}$	K^*	$\mathrm{E}W^*$	$\mathrm{E}W_{M/M/s}$
0.7	0	0.01	0.16	0	0.00	0.00
0.8	0	0.08	0.64	0	0.00	0.01
0.9	0	0.85	2.75	0	0.00	0.22
0.95	1	3.84	7.55	0	0.05	1.01

Table 5.2: $\gamma = 0.1, 1/\mu = 10$.

	$s = 1$			$s = 5$		
ρ	K^*	$\mathrm{E}W^*$	$\mathrm{E}W_{M/M/s}$	K^*	$\mathrm{E}W^*$	$\mathrm{E}W_{M/M/s}$
0.5	0	1.91	5.00	0	0.01	0.26
0.6	0	3.28	7.50	0	0.04	0.59
0.7	3	5.57	11.67	0	0.12	1.26
0.8	7	9.83	20.00	0	0.41	2.77
0.9	13	23.55	45.00	1	1.62	7.62

Table 5.3: $\gamma = 0.01, 1/\mu = 5$.

system is congested at the moment, and if they call back later, they will receive priority. We model this system as a 3-dimensional Markov process, with one dimension being the waiting stage of the first customer in queue 1. We then discretize the time and suggest a method to make use of the value iteration to numerically compute the long-term average overall waiting times of customers. We show that this method offers close approximations for small call centers. For large call centers, this method is computationally expensive, due to the large number of states, thus, simulation is preferred. Furthermore, by comparing this model to the $M/M/s$ model, we see that customers experience much shorter waiting times in this model for all scenarios we consider. This is caused by having customers calling back, which reduces the variability of the arrival process to the system. Such effect is so strong that the reduction in mean waiting times can reach 2000% in some extreme cases.

This chapter suggests the following interesting topics for further research. First of all, our numerical results suggest that $\mathrm{E}(W_1 + W_2)$ is unimodal in K, and we use this as an assumption in finding K^*; it would be interesting to rigorously validate this assumption. Second, intuitively, we explained the reason why $\mathrm{E}(W_1 + W_2)$ is a non-decreasing function of γ, and offering a theoretical proof of this property would be very appealing.

	$s = 20$			$s = 100$		
ρ	K^*	$\mathrm{E}W^*$	$\mathrm{E}W_{M/M/s}$	K^*	$\mathrm{E}W^*$	$\mathrm{E}W_{M/M/s}$
0.7	0	0.00	0.16	0	0.00	0.00
0.8	0	0.01	0.64	0	0.00	0.01
0.9	0	0.11	2.75	0	0.00	0.22
0.95	0	0.73	7.55	0	0.00	1.01

Table 5.4: $\gamma = 0.01, 1/\mu = 10$.

Third, we focus on the performance measures of the system under stationarity in this study, which leaves room for extension to non-stationary cases; especially, one could consider comparing the performance of this model with the $M/M/s$ model with a non-homogeneous Poisson arrival process. Last but not the least, we can extend the model by adding penalty to the performance metric for those people who call back.

Chapter 6

The validation of call center models

In call centers, planners need to calculate the staffing levels such that the service level targets are met. To this end, they can make use of different models, which predict the service levels for given number of agents. These models simplify certain processes in real call center operations, and they make assumptions. In this chapter, by comparing the service levels from real data to the service levels from simulation results of different models, we validate these models and assumptions. The results show that ignoring certain features in call centers, such as agent breaks, agent heterogeneity and wrap-up times, can lead to inaccurate prediction of the service levels. Furthermore, we empirically verify the validity of some common assumptions of these models, such as the inhomogeneous Poisson arrival process, exponential assumptions of the handling times and the customer patience. Comparison results reveal that although these assumptions cause errors in the service level predictions, the errors are not significant.

6.1 Introduction

Deciding on the right number of agents is a crucial part in call center workforce planning, since nearly 75% of total cost is the personnel cost (Gans et al. (2003)). Having too few agents will lead to long waiting times of customers or customer churn; having too many agents will results in unnecessary costs. There are many models that can assist with this decision making process. Given certain information about the arrivals, handling times, customer patience, number of agents, etc, these models predict the service levels. For example, for single-skill call centers, the well-known Erlang C and Erlang A models are widely used. For multi-skill call centers, no analytical solutions exist, and a simulation approach is often used to determine the staffing levels (Avramidis et al. (2009)). A common procedure of call center modeling is first empirically analyzing call center data, which reveals certain features of customers, agents or call center operations processes. Then these features are incorporated in models. Finally, these models are analyzed either numerically or analytically, and the results are being compared to simulation results to measure the performances. For an overview of different models, we refer to Gans et al. (2003), Aksin et al. (2007), and the references therein.

All these call center staffing models make certain assumptions or simplifications to mimic reality. This raises questions: do these models work well in practice? Which model(s) give the most accurate service level predictions? What is the impact of the assumptions in these models? Despite the importance of answering these questions, there has been little attention in the literature. Successfully answering these questions would not only help researchers and practitioners in choosing the right model and assumptions, but also give managers confidence in using them. We aim to fill this gap, by validating and comparing different models with different assumptions. Different from the common modeling approach, the order of our validation process is reversed. To be more specific, we first calculate parameters from real data, such as the number of arrivals, handling times, number of agents, etc; these parameters are given as inputs to different multi-skill staffing models, then we compare the service levels in real data with the service levels predicted by these models.

The main contribution of this chapter can be described as follows. We are the first who use real data to validate and compare different multi-skill call center staffing models, and we identify several models that give accurate service level predictions. In addition, we empirically show that ignoring agents' breaks in call center staffing will lead to large errors in service level predictions. It is also shown that the AHT varies per day, and a model that ignores such variability would lead to inaccurate service levels predictions. This variability is partially caused by agent heterogeneity and agent learning effects, and we develop a model to fit the AHT of each day with consideration of both effects. Moreover, comparison results indicate that the assumption of the in-homogeneous Poisson arrival process does not influence the performances of the model. Furthermore, although the mean wrap-up time is short, numerical results suggest that it still makes a difference in the accuracy of the model. Besides, we empirically show that the exponential assumption of the HT does not have much effect on the accuracy of the model, and we validate that the often made exponential assumption of the patience has significant effect on the accuracy of the model. Finally, statistical analysis gives interesting results on AHT, breaks, etc, which gives useful insights.

There are several attempts in the validation and simulation of the staffing models for single-skill call centers. For example, Robbins et al. (2010) compare the Erlang C model with simulation models where several assumptions are relaxed. They find that the Erlang C model is subject to significant error in predicting system performance. However, they do not make use of real data, and the validation is still not in an empirical way. By using real data, Mandelbaum and Zeltyn (2004) empirically show that the abandonment percentage is linear to the mean waiting time. They theoretically verify that this linear relationship remains valid if the patience has some non-exponential distributions such as uniform and hyperexponential. Regarding multi-skill call centers, L'Ecuyer (2006), Avramidis et al. (2004b) discuss some important aspects of simulating multi-skill call centers, such as HT, abandonments, and retrials.

The rest of the chapter is structured as follows. In Section 6.2, we describe the data, and show statistical analysis results of the data. Different staffing models with different features or assumptions are studied and compared in Section 6.3. Finally, we draw conclusions and discuss topics for further research in Section 6.4.

6.2 Statistical analysis

The data set used in this chapter is collected by VANAD Laboratories for the year of 2014. The data consists of two separate data sets, one is the call log data, which gives the following information: customer call arrival times, departure times, agent identity who handled the call, skill type of the call, etc; the other one is referred to as the activity data, which describes agents' activities at any moment (see Table 6.1 for an example of the activity data).

Activity	Start time	End time	Agent ID
Logging in	1/2/2014 8:54:43	1/2/2014 9:00:44	A
Taking calls	1/2/2014 9:00:44	1/2/2014 9:07:43	A
Meeting	1/2/2014 9:07:43	1/2/2014 9:08:47	A
Taking calls	1/2/2014 9:08:47	1/2/2014 9:11:28	A
Wrap up	1/2/2014 9:11:28	1/2/2014 9:11:30	A

Table 6.1: An example of the activity data set.

The call log data has in total 1543164 call records from 27 different skills. There have been 312 agents working in this call center in the year of 2014. This includes part-time agents, full-time agents, agents that worked only for a few months and agents that worked in every month of the year. Each agent has a skill set, which consists of at least one skill. Not every agent has all the skills. In this chapter, we only focus on the top 8 skills that are most chosen by the customers, since they represent nearly 99% of the total call volume. We explain how we deal with the amount of time that agents spend on working on other skills in Section 6.3.

The routing mechanism works as follows. When a customer calls, she will interact with the IVR (interactive voice response unit) by making use of her key pad to choose the call type. If there is any agent available with the skill to handle that type of calls, then she is routed to the longest idle agent of those available agents; otherwise, she will wait in an invisible queue. The calls in this queue are served in the FCFS (first come first served) order.

6.2.1 Handling times

Service times receive less attention than the arrivals in call centers. However, even small differences in AHT can lead to large differences in operations costs (see Gans et al. (2010)). Ibrahim et al. (2016a) observe that HT are agent- and time-dependent, and they develop models that account for those facts, which give accurate prediction on the AHT.

For our data, we plot the empirical histogram of the HT of one skill in Figure 6.1. There is a peak at nearly 0 seconds, which is caused by the fact some calls end immediately due to a loss of signal or call connection errors. In Brown et al. (2005), the authors also find significant amount of short-HT calls in their data, which is caused by the agents who simply hangup the calls to have extra rest time. However, this is not the case in this call center under consideration.

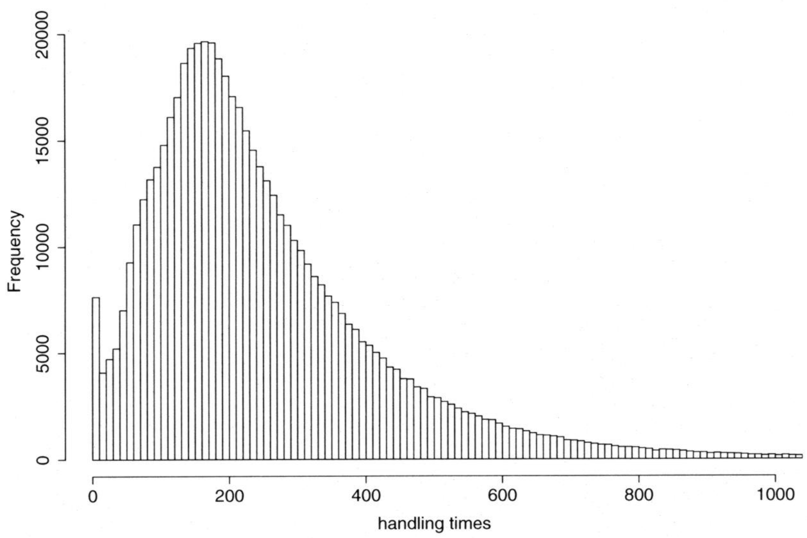
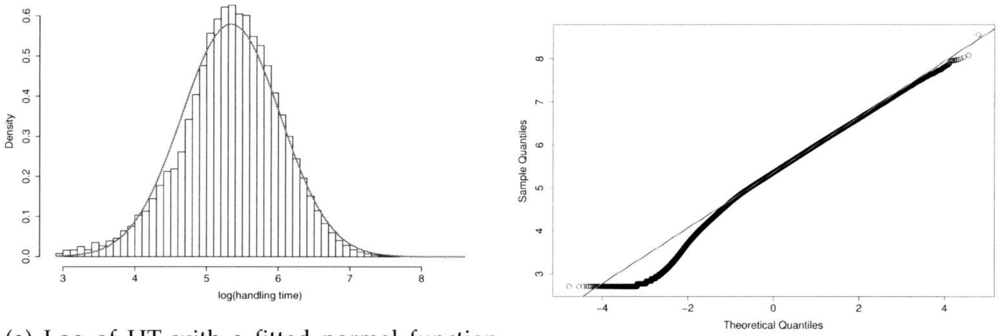


Figure 6.1: Handling times histogram of one skill (units: seconds).

It is common in call center models that the HT are assumed to have an exponential distribution, for the purpose of obtaining analytically tractable results. However, as one can see from Figure 6.1 that exponential distribution will not fit the data well. We remove the calls with short HT (less than 15 seconds), and fit a log-normal distribution to the empirical HT. The histogram with a fitted log-normal Probability Density Function (PDF), as well as the Quantile-Quantile plot are shown in Figure 6.2. Similar to the results found by Bolotin (1994), Brown et al. (2005) and Pichitlamken et al. (2003), Figure 6.2 shows that the log-normal distribution fits our data well.

The AHT per day is shown in Figure 6.3. As one can see, the AHT varies per day, and this variation can be quite significant. For example, the AHT can almost reach 300 seconds around day 140, and it can also be as low as 220 seconds near day 200. There are two reasons for this. The first one is that the SL varies per day which leads to the fluctuation of the AHT per day. To be specific, it is possible that customers who experienced long waiting times would demand longer services, because, they think this call center is difficult to reach and might want to have more questions answered instead of making another call. The second possible reason is that agents are heterogeneous in terms of their AHT, in the sense that some agents handle calls faster than others. To investigate this in more depth, we plot the AHT of each months of experienced agents and new agents in Figure 6.4. As one can see from Figure 6.4a the AHT of each experienced agents is different; furthermore, in Figure 6.4b, the AHT of new agents all exhibit a declining trend, which suggests that new agents learn over time, and their AHT decrease



(a) Log of HT with a fitted normal function (curve).

(b) Q-Q plot.

Figure 6.2: Histogram and Q-Q plot of the log of the HT.

as they learn. The histogram of the AHT of each agent for this specific skill is shown in Figure 6.5, which further confirms the agent heterogeneity. To avoid coincidental high and low values on AHT in Figure 6.5, we only consider those agents who have answered more than 200 calls. This agent-by-agent heterogeneity and learning effect are also shown in Gans et al. (2010). For more empirical results and modeling on HT, we refer to Gans et al. (2010) and Ibrahim et al. (2016a).

To model the learning effect of the new agents and to predict the AHT of each agent of each day, we now develop a model for the AHT of each agent in each month. We assume that

$$\mathrm{EAHT}_{j,m_i} = \alpha_j e^{\omega_j m_i},$$

where m_i is the corresponding month of day i, $i = 1, 2, \ldots, N$, and $m_i = 1, 2, \ldots, 12$, and AHT_{j,m_i} is the AHT of agent j in month m_i, and α_j and ω_j are the parameters of agent j, $j = 1, 2, \ldots, J$, where $J = 312$.

Given the fit of AHT of each agent in each month, we can then use the following model to fit the AHT of each day.

$$\mathrm{EAHT}_i = \frac{\sum_{j=1}^{J} \mathrm{AHT}_{j,m_i} n_{j,i}}{\sum_{j=1}^{J} n_{j,i}}, \tag{6.1}$$

where AHT_i stands for the AHT of day i, and $n_{j,i}$ stands for the number of calls that are answered by agent j in day i, $i = 1, 2, \ldots, N$. Equation (6.1) can be interpreted in the following way, the AHT of day i is the weighted sum of the AHT of each agent in month m_i, where the weight of one agent is the proportion of calls answered by this agent.

The actual AHT_i and the fitted AHT of day i, denoted by $\widehat{\mathrm{AHT}}_i$, for $i = 1, 2, \ldots, N$, are plotted in Figure 6.6. Figure 6.6 suggests that the agent heterogeneity can explain a major part of the AHT variability, and our model fits the AHT of each day well. We

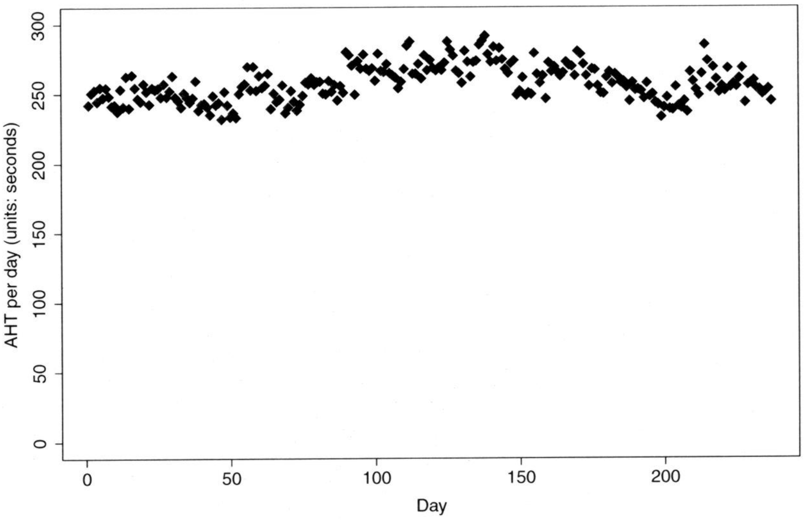


Figure 6.3: AHT per day.

calculate the R^2 values of each router in Table 6.2, which is defined by

$$R^2 := 1 - \frac{\sum_{i=1}^{N}(\mathrm{AHT}_i - \widehat{\mathrm{AHT}}_i)^2}{\sum_{i=1}^{N}(\mathrm{AHT}_i - \overline{\mathrm{AHT}})^2},$$

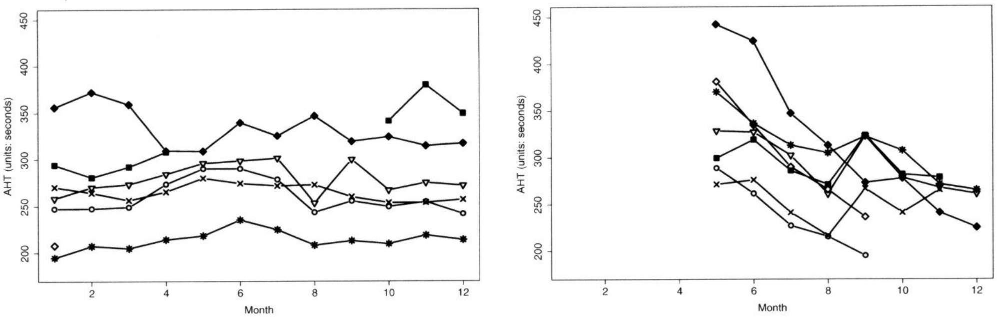
where $\overline{\mathrm{AHT}}$ is the mean AHT over each day of the whole year.

Skill	30175	30560	30172	30181	30179	30066	30518	30214
R^2	0.52	0.54	0.46	0.47	0.27	0.09	0.57	0.24

Table 6.2: R^2 values of $\widehat{\mathrm{AHT}}_i$.

As one can see from Table 6.2, the R^2 are approximately 50% for the first four skills. This suggests that the AHT prediction model in Equation (6.1) can help explain half of the variability in AHT. Also, R^2 fluctuates for the last four skills. The small sample sizes of the last four skills can explain this fluctuation.

For many call centers, the agents' workload does not end at the moment when customers or agents hang up, since sometimes agents still have to do some after-call work. This duration that is spent on after-call work is often referred to as the wrap-up time. We have not yet found any empirical or theoretical results on the wrap-up times in the literature. We think this is mainly because wrap-up times are difficult to measure and they are usually not recorded in call center data. However, in this data set, we can empirically study the wrap-up times, since they are part of the agents' activity, and their



(a) Experienced agents.

(b) New agents.

Figure 6.4: AHT per month of some agents.

start times, end times and durations are all recorded in this data set. In Figure 6.7, we plot the histogram of the wrap-up times. As one can see, the wrap-up times in this data set are in general quite short and often has a duration of 0 seconds, and the mean wrap-up time is approximately 3.28 seconds.

6.2.2 Patience

Customers' patience has a drastic effect on system performance (Gans et al. (2003)). Thus, it is important to have an accurate estimation of the patience. However, call center data are censored data, since we only observe the patience of those customers who have abandoned, and for those customers whose calls have been answered, we do not know what their patience are. To estimate the patience with censored data sets, we use the Kaplan-Meier estimator. The empirical Cumulative Distribution Function (CDF) and the hazard rate function of the patience are shown in Figure 6.8. Figure 6.8 suggests a different customer abandonment behavior than the ones found by Mandelbaum and Zeltyn (2004), Brown et al. (2005) and Roubos and Jouini (2013), where the latter one shows that hyperexponential distributions fit the patience distribution in their data well.

In Figure 6.8a, a sudden increase in the CDF can be seen at around 1300 seconds. This increase is mostly likely to be caused by the lack of uncensored samples, i.e., most customers that waited more than 1200 seconds are answered (censored data points). In Figure 6.8b, we observe peaks at every 25, 26 seconds. The moments of these peaks correspond to the moments when voice messages are announced to all waiting customers. Different from the delay announcement messages studied by Armony et al. (2009), Ibrahim et al. (2015) and Jouini et al. (2011), the voice message in this call center does not give any information on the anticipated amount of delay.

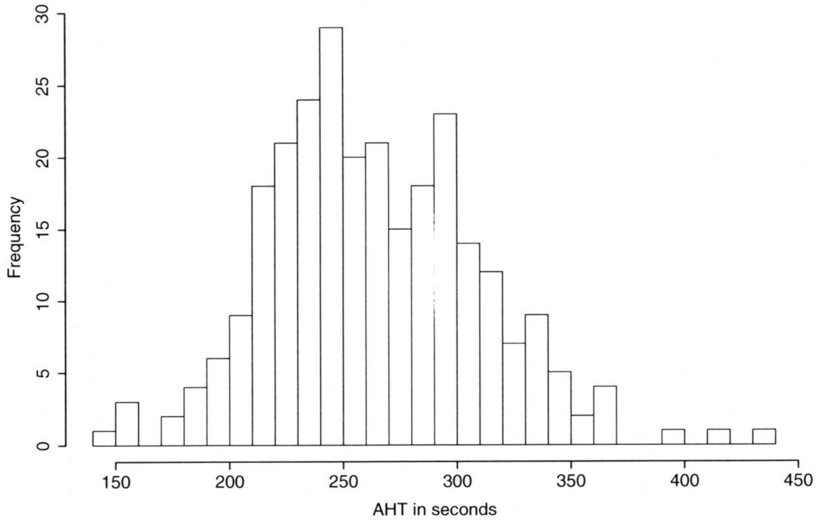


Figure 6.5: Histogram of the AHT of each agent (units: seconds).

6.2.3 Breaks

Paid breaks are an important part of shrinkage in call centers, together with other activities, such as training, illness, and so forth (Koole (2013)). In this data set, we do not have the original schedule of the agents, thus, we can not calculate the amount of shrinkage caused by sickness and training. However, we can observe agent breaks' starting times, ending times and durations, as well as whether it is a paid break or an unpaid break (such as lunch breaks) from the data.

Shrinkage is usually a significant part in call center workforce planning. To illustrate the amount of shrinkage, we plot the histogram of the shrinkage percentage of each agent in Figure 6.9. To generate this plot, we use the following definition

$$\text{Shrinkage\%} := \frac{\text{amount of time in breaks}}{\text{amount of time in working} + \text{amount of time in breaks}}.$$

Note that in Figure 6.9, we remove breaks longer than 30 minutes from the data, since they are mostly caused by mistakes. Also, we only consider the agents whose working hours are longer than 100 hours. The mean Shrinkage% is approximately 6.7%, which is relatively low. In practice, it is uncommon that the Shrinkage% is higher than 10%. Part of the reason that we have low Shrinkage% in this data is that in calculating Shrinkage% we do not include agent absenteeism, such as agents being on holiday or sick, or agents doing training, while in reality, these activities are included.

The histogram of the break durations is shown in Figure 6.10. In Figure 6.10, one can observe some clear peaks at around 300, 600 and 900 seconds, which is 5, 10 and 15 min-

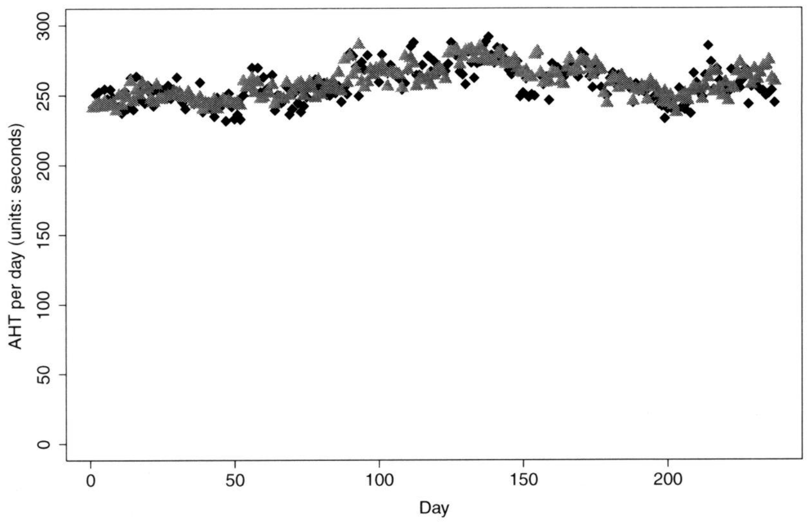


Figure 6.6: AHT_i (diamond) and $\widehat{\mathrm{AHT}}_i$ (triangle), units: seconds.

utes, respectively. The 600 and 900 seconds breaks are pre-specified in the shifts, thus, they are in some sense "plannable". However, they are not completely "plannable", because the starting times of the breaks and the durations can still have some variations depending on the agents preferences and other factors, such as the busyness of the call center at that moment. For example, some agents prefer taking several 300 seconds breaks instead of one 600 or 900 seconds full break; another example could be that agents may take less breaks if the call center is currently busy. Furthermore, besides these "plannable" breaks, there are breaks for other purposes, such as agents going to the toilet, having coffee, etc. These breaks are usually short and they are "unplannable". An ideal model would differentiate between these two types of breaks, and we then can study the consequences of ignoring either the "plannable" or "unplannable" breaks in SL. However, in the data set we have, it is difficult to precisely differentiate between these two types of breaks. For example, if one agent takes a 8-minute break half an hour earlier than the pre-specified break time, then she works for an hour, then takes a 4-minute break, it is not completely clear whether the first and the second breaks are planned or not. Therefore, in this chapter, we do not make a distinction between these two types of breaks in our models.

In order to have more insights in break durations, we plot the break durations histograms of some individual agents in Figure 6.11. One can conclude from these graphs that different agents have different patterns of break durations; for example, agent A prefers more short breaks rather than long breaks (we are told by the manager that this agent works at home and is a part-time agent), while agent D mostly has 10- and 15-minute breaks with some small breaks occasionally.

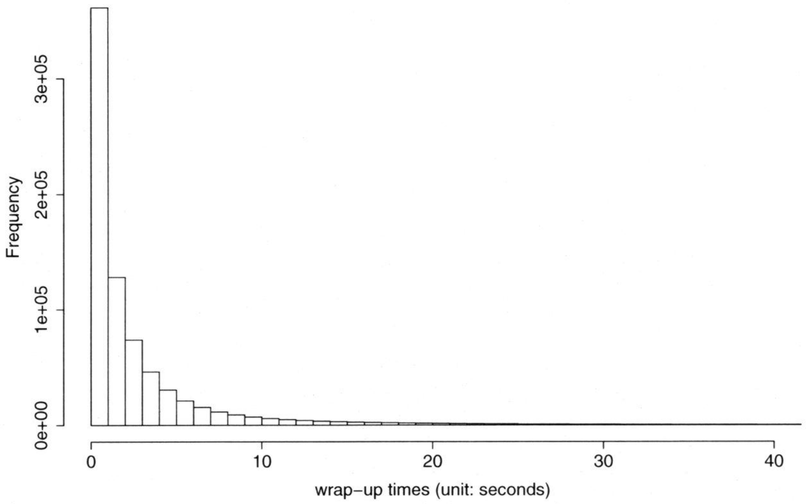


Figure 6.7: Wrap-up times histogram (units: seconds).

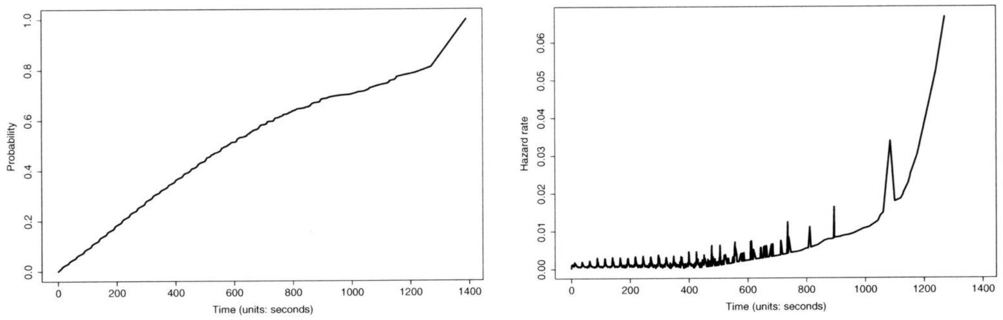
6.3 Comparison of models

In this section, we validate and compare nine different call center models with different assumptions. Now we briefly explain the procedure on how we do the comparison. We first divide the operation hours of this call center into 24 intervals of 30 minutes. Then for each interval, we calculate parameters such as number of arrivals, HT, number of agents and patience from real data; then these parameters are given as inputs into simulations. Then we compare the SL_1 with AWT being 60, r and ASA from simulation with those from real data. The actual SL_1, r and ASA can be computed from the call log data.

The models with their corresponding assumptions are shown in Table 6.3. The assumptions and notations can be interpreted as follows.

Arrival: "Empirical" means that the arrival processes are identical in simulation and in real data, i.e., if there is an arrival at time t in real data, we schedule an arrival at the exact same moment in simulation. "IPP" stands for in-homogeneous Poisson process with piecewise constant rate, which means that if there are A arrivals within certain interval in the data, then we schedule A' arrivals within that interval in simulation, with $A' \sim \mathrm{Poisson(A)}$.

HT, AHT per day: **HT** being "Empirical" means to assign the HT of a customer in simulation, we select a random number from the empirical HT of the whole year if **AHT per day** is "No", or from the empirical HT of that specific day if **HT** is "Yes". **HT** being "Exp" means that we assume that the HT has an exponential distribution with its mean being the mean of the HT over the whole year if **AHT per day**



(a) CDF of the patience.

(b) Hazard rate function.

Figure 6.8: Patience.

is "No", or over the day if **AHT per day** is "Yes". If **AHT per day** is "Fit", then we use $\widehat{\mathrm{AHT}}_i$ as the AHT in simulation.

Wrap-up: "Yes" means that in simulation we add the empirical mean of the wrap-up times in the HT, "No" means we do not consider wrap-up times.

Patience: if it is "Empirical", then for each customer in simulation, we generate a random patience from the empirical CDF estimated by the Kaplan-Meier estimator; if it is "Exp", then we assume the patience has an exponential distribution, with its mean being the empirical mean of the patience from real data, which can also be estimated via Kaplan-Meier estimator.

(paid) Breaks: "Yes" means that if an agent takes a break, then we subtract the proportional staffing levels from the total staffing levels of this interval; otherwise, **Breaks** is "No". For example, assume agents worked (either waiting for a call, answering a call or doing a wrap-up) 180 minutes in total in certain interval, and two agents had breaks during this interval, each break lasted 10 minutes, if **Breaks** is "Yes", then we assume there were in total $(180 - 20)/30 = 5.3$ agents working in this interval, and we round it to 5 in simulation; if **Breaks** is "No", then we ignore the breaks and assume in simulation that there were in total $180/30 = 6$ agents working in this interval.

Note that besides working on these eight skills that we selected and taking breaks, there are other agents' activities, such as working on calls of other skills, making outbound calls, having consultations with managers or other senior agents, etc. The amount of time that spends in these activities are very little compared to the time that are spent in breaks, thus, we exclude these durations in a similar way as we remove the agent breaks.

We now introduce the performance measurements. Note that most parameters are time-dependent, thus, we are simulating non-stationary systems. Consequently, the results such as SL_1, r and ASA are different per simulation (see Roubos et al. (2012)). With consideration of such variability, we repeat the simulation 1000 times. Therefore, for

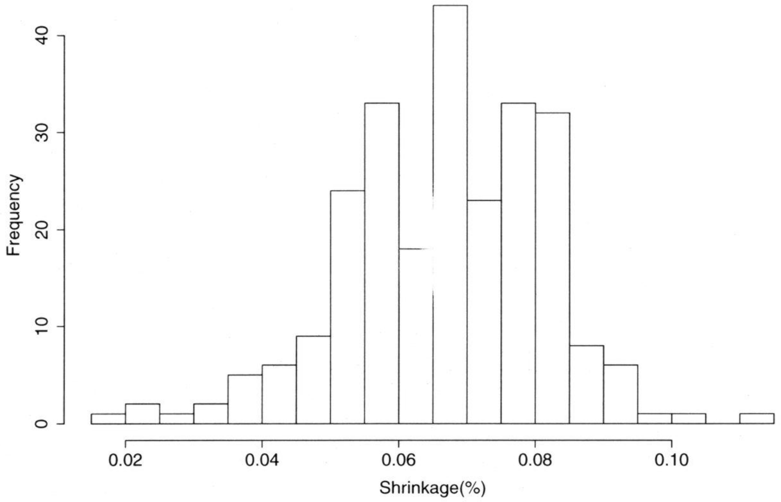


Figure 6.9: Histogram of shrinkage of each agent.

each model and for each day, there are 1000 simulation outcomes from simulation and one realization. To measure the difference between simulated results and the actuals, we use WAE (weighted absolute errors).

$$\mathrm{WAE}_{\mathrm{X}} := \frac{\sum_{i=1}^{n} T_i |\mathrm{E}\hat{\mathrm{X}}_i^{sim} - \mathrm{X}_i^{act}|}{\sum_{i=1}^{n} T_i},$$

where $\hat{\mathrm{X}}_i^{sim}$ is the simulated result of day i, and X_i^{act} is the actual result of day i, and T_i is the number of arrivals in day i. We compute the WAE of SL_1, r and ASA, which is $\mathrm{WAE}_{\mathrm{SL}_1}$, $\mathrm{WAE}_{\mathrm{r}}$ and $\mathrm{WAE}_{\mathrm{ASA}}$, respectively.

WAE measures the difference between the simulation results and the actuals. Ideally, WAE equals 0. However, in all the measures we have, WAE are all positive. One part of the WAE comes from the variability in SL_1, r and ASA; for example, r is different per simulation. The other part of WAE comes from the model; for example, if one simulates a model which does not describe the reality well, then there is a big difference between the simulation results and the actuals. Therefore, we also measure the mean WAE caused by variability, which can be estimated by the following expression

$$\sum_{i=1}^{n} T_i |\mathrm{E}\hat{\mathrm{X}}_i^{sim} - \hat{\mathrm{X}}_i^{sim}| / \sum_{i=1}^{n} T_i.$$

This number can be interpreted as the mean WAE of the variability of a model.

Besides WAE, we also compare $\mathrm{I}_{\alpha,\mathrm{SL}_1}$, $\mathrm{I}_{\alpha,r}$ and $\mathrm{I}_{\alpha,\mathrm{ASA}}$ of each model which are the per-

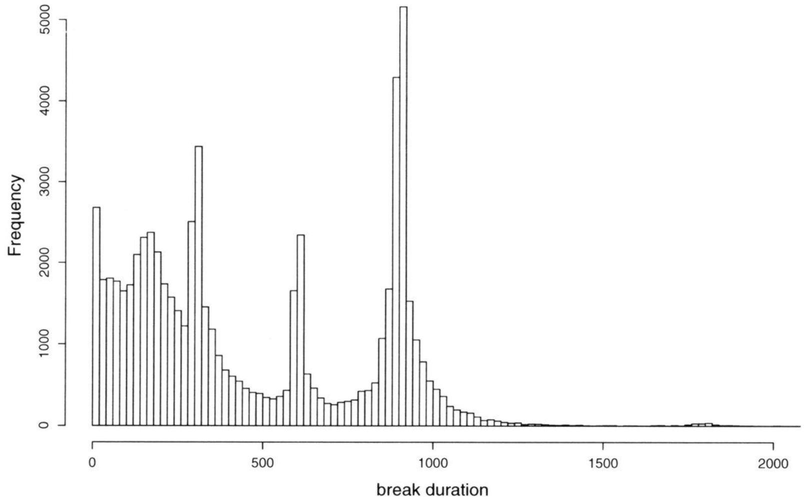


Figure 6.10: Histogram of break durations (units: seconds).

centage of the actuals within the α confidence interval of the simulation outcomes. In this chapter, we let $\alpha = 95\%$. Furthermore, we consider $\mathrm{P}(\mathrm{SL}_1 > Q_{0.5,\mathrm{SL}_1})$, $\mathrm{P}(r > Q_{0.5,r})$ and $\mathrm{P}(\mathrm{ASA} > Q_{0.5,\mathrm{ASA}})$, which stands for the percentage of actual SL_1, r and ASA that is higher than the 50% quantile of the simulation outcomes, respectively.

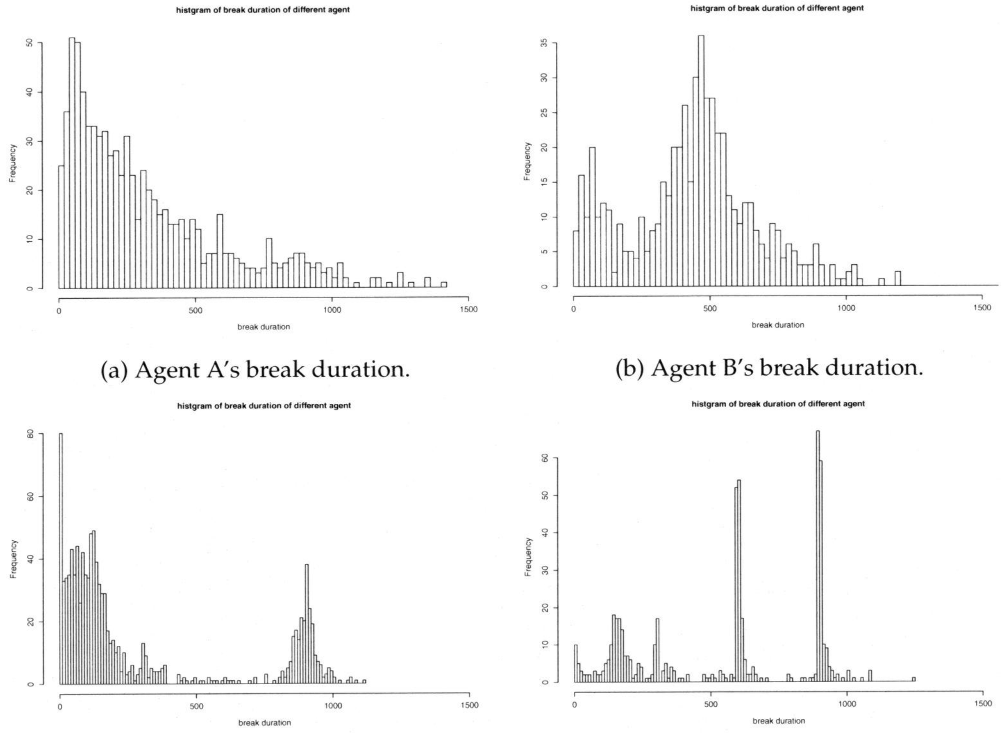
These measurements of each model are shown in Tables 6.4 and 6.5, with the numbers between brackets being the WAE of the simulation variability. Note that in Table 6.4, the units of WAE_{ASA} is seconds. The actual and the predicted SL_1, r and ASA, with their 95% confidence intervals of models 1 to 6 (models 7 to 9 are similar) are shown in Figures 6.14-6.16.

We describe now the observations and conclusions we draw by comparing these models.

In general, the first 3 models are comparably accurate, with WAE being about 3% in SL_1, 0.7% in r, and 6 seconds in ASA, despite the fact that we made certain simplifications, such as rounding the number of agents in each interval, redials and reconnects in call centers (see Ding et al. (2015a)) etc.

Whether the arrival process is "Empirical" or "IPP" does not have a strong influence on the performance of the models. Thus, if the forecasts for the arrival rate are accurate, assuming in-homogeneous Poisson arrival processes will not degrade the model. This can be concluded by comparing models 1 and 2.

The exponential assumption of HT does not have strong influence on performances of the models. Interestingly, by having the exponential assumption of HT, the model performs even slightly better compared to the model with empirical HT. This conclusion



(a) Agent A's break duration.

(b) Agent B's break duration.

(c) Agent C's break duration.

(d) Agent D's break duration.

Figure 6.11: Break duration histogram of some agents.

can be deducted by comparing models 1 and 3. Now we explain the reason. All the models we consider are slightly optimistic compared to reality, in the sense that they predict higher SL_1, lower r and ASA compared to the actuals. This can be confirmed by the fact that the first column of Table 6.5 are all below 50%. This is especially true in the days where the actual SL_1 is low, which could be caused by the fact that customers require longer service if they experienced long waiting. By having the exponential assumption of the HT, it will lead to a more conservative model, since exponential HT has a higher variance than the empirical variance, thus, this assumption will improve the performance slightly.

The model we developed in (6.1) leads to significant improvement of the accuracy in predicting SL_1, r and ASA by incorporating agent heterogeneity and agent learning effect. This can be concluded by comparing models 4 and 9, i.e., model 4 has smaller errors for all performance metrics. However, these two effects do not purely explain the variability of the AHT of each day. As one can see that from Table 6.4 that model 2 performs better compared to model 4.

The exponential assumption of the patience distribution has stronger influence on the accuracy of the models compared to the exponential assumption of the HT. This can be concluded by comparing models 1,3,5 and 6. This finding is also discussed in the

	Arrival	HT	AHT per day	Wrap-up	Patience	Breaks
Model 1	Empirical	Empirical	Yes	Yes	Empirical	Yes
Model 2	IPP	Empirical	Yes	Yes	Empirical	Yes
Model 3	IPP	Exp	Yes	Yes	Empirical	Yes
Model 4	IPP	Exp	Fitting	Yes	Empirical	Yes
Model 5	IPP	Empirical	Yes	Yes	Exp	Yes
Model 6	IPP	Exp	Yes	Yes	Exp	Yes
Model 7	IPP	Empirical	Yes	Yes	Empirical	No
Model 8	IPP	Empirical	Yes	No	Empirical	Yes
Model 9	IPP	Empirical	No	Yes	Empirical	Yes

Table 6.3: Models.

	WAE_{SL_1}	WAE_r	WAE_{ASA}	$\text{I}_{\alpha,\text{SL}_1}$	$\text{I}_{\alpha,r}$	$\text{I}_{\alpha,\text{ASA}}$
Model 1	3.00%(2.80%)	0.748%(0.380%)	6.53(2.83)	76.4%	51.1%	58.6%
Model 2	3.09%(4.63%)	0.786%(0.674%)	6.84 (5.34)	84.8%	75.9%	76.8%
Model 3	2.98%(5.04%)	0.690%(0.737%)	6.24 (6.06)	91.1%	89.0%	88.2%
Model 4	4.21%(6.44%)	1.044%(0.981%)	9.22(8.09)	80.2%	73.4%	75.1%
Model 5	4.81%(4.00%)	0.658%(0.700%)	12.56 (4.57)	63.7%	84.0%	46.0%
Model 6	4.60%(3.71%)	0.577%(0.697%)	11.92 (4.37)	71.3%	89.5%	57.4%
Model 7	8.54%(3.28%)	2.092%(0.498%)	17.40 (3.79)	30.0%	16.0%	16.5%
Model 8	4.30%(4.08%)	1.116%(0.598%)	9.51 (4.71)	72.2%	48.5%	57.0%
Model 9	5.69%(5.48%)	1.524%(0.752%)	12.35 (6.13)	65.8%	55.3%	62.0%

Table 6.4: Performance measures of models.

following remark.

Remark 6.1. Whitt (2005) theoretically shows that the behavior of some queueing models (such as the $M/GI/s/r+GI$) is primarily affected by the HT distribution through its mean while it is primarily affected by the patience distribution by its hazard function near the origin, and not its mean or tail behavior.

Although the wrap-up times are in general very short in this call center, ignoring them will lead to inaccuracy. This can be concluded by comparing models 1 and 8.

Agent breaks have a drastic influence on the accuracy of the models. Without taking them into consideration when making planning decisions, huge errors in predicting SL_1, r and ASA will incur. For example, by comparing models 2 and 7, we see that ignoring agent breaks results in nearly doubled errors in SL_1 and more than doubled errors in r and ASA.

	$P(\mathrm{SL}_1 > Q_{0.5,SL})$	$P(r > Q_{0.5,r})$	$P(\mathrm{ASA} > Q_{0.5,r})$
Model 1	29.1%	83.1%	84.4%
Model 2	24.5%	82.2%	86.1%
Model 3	33.8%	78.9%	78.1%
Model 4	43.5%	64.6%	62.9%
Model 5	12.7%	75.9%	94.5%
Model 6	18.1%	70.9%	90.7%
Model 7	21.1%	98.7%	98.7%
Model 8	13.5%	94.9%	95.4%
Model 9	35.0%	69.6%	71.3%

Table 6.5: Percentage of actuals above the median.

6.3.1 Rates vs. actuals

In models 2 to 9, we make the arrival rates equal to the number of arrivals observed from the data. For example, if the actual is 100 in an interval, then in models 2 to 9, we assume the arrival processes are homogeneous Poisson processes with rates being 100. Then we simulate the models to measure the weighted sum of $|\hat{X}_i^{sim} - X_i^{act}|$, for $i = 1, 2, \ldots, n$, where $\hat{X}_i^{sim}$ is the performance metric, which can be SL_1, r or ASA. Strictly speaking, this is different from what a call center manager would do in practice when he makes SL predictions. There are two differences: the first one is that the number 100 in this example is the realization of a Poisson process with a certain rate, which we do not know from the data; the second difference is that when the manager makes SL predictions, he does not know the real arrival rate nor the realization (100 in this example), but he would make a forecast for the rate. In this subsection, we intend to quantify the difference in SL between what we do in models 2 to 9 and what a call center manager does in practice.

We first introduce some notation. First, for simplicity, we remove the subscript i in $\hat{X}_i^{sim}$ and X_i^{act}, which leads to $\hat{X}^{sim}$ and X^{act}, for $X = \mathrm{SL}_1, r$ and ASA, respectively. In addition, we let X^{sim} be the performance metric obtained from simulation when the real underlying rates are used as inputs in the arrival process, and X^{fc} be the performance metric obtained from simulation when forecasts are used as inputs in the arrival process. Thus, we can never know how large X^{sim} is, since we only observe realizations of X^{sim}, but not X^{sim} itself. In fact, depending on what inputs are for the arrival process, there are three different scenarios: what we **currently** do in models 2 to 9, what a call center manager would do in **practice**, and what the **ideal** situation is. Now we specify the differences between these three scenarios.

Ideal: In an ideal situation, one can predict the arrival rate for the future with full accuracy. In such a case, one can then calculate $|X^{sim} - X^{act}|$, which can be interpreted as the modeling error caused by all sorts of assumptions and simplifications in the modeling phase.

Currently: In Table 6.4, we have shown $|\hat{X}^{sim} - X^{act}|$ for models 2 to 9, and the difference between $|\hat{X}^{sim} - X^{act}|$ and $|X^{sim} - X^{act}|$ can be expressed as follows, using the Triangle Inequality:

$$|X^{sim} - X^{act}| \leq |\hat{X}^{sim} - X^{sim}| + |\hat{X}^{sim} - X^{act}|, \quad X = \mathrm{SL}_1, r \,\text{and ASA}, \tag{6.2}$$

where $|\hat{X}^{sim} - X^{sim}|$ can be understood as the difference in SL between using real rates and using realizations.

Practice: In practice, call center managers make forecasts on the number of arrivals, and then they make SL predictions. They are mainly interested in the difference between predicted SL and actual SL, i.e., $|X^{fc} - X^{act}|$. There are mainly two sources of errors in $|X^{fc} - X^{act}|$: one comes from forecasting errors, and the other one comes from modeling errors.

Now we consider one specific day and a specific model, i.e., model 2, and we use a simple forecasting method to make forecasts for every interval of this specific day, and then we use the results for this specific day to give some indication on how large $|\hat{X}^{sim} - X^{sim}|$ and $|X^{fc} - X^{act}|$ are. To be more specific, we make the forecasts for the first Monday of April for all the skills. The forecasting method is simple: for each skill, we take the average of the number of arrivals in the same interval in the past four Mondays. Then based on the forecasts, we conduct simulation to derive X^{fc} for $X = \mathrm{SL}_1, r$ and ASA. Realizations can be observed from simulation, which is then used as the arrival rates to obtain $\hat{X}^{sim}$ for $X = \mathrm{SL}_1, r$ and ASA. To make it more clear, we give an example: assume that the actual number of arrivals of a certain interval is 100 and the SL is X^{act}, and the forecast of this interval is 105, then we run simulation where the arrival rate is 105 to derive X^{fc}; in the simulation, assume that the number of arrivals is 110, then we make another simulation where the arrival rate is 110 to compute $\hat{X}^{sim}$. In such a way, we can compute $|\hat{X}^{sim} - X^{fc}|$, which can give some indication on how large $|\hat{X}^{sim} - X^{sim}|$ is, since one can never know how large $|\hat{X}^{sim} - X^{sim}|$ exactly is. This is essentially setting up a lab setting, where we know the underlying rates (i.e., the forecasts) and the actuals (i.e., realizations in the simulation), such that we can quantify the difference between the using the rate and the actual. Then, by combining $|\hat{X}^{sim} - X^{fc}|$, Inequality (6.2), and the results we had in Table 6.6 for $|\hat{X}^{sim} - X^{act}|$, we can obtain some indication on how large $|X^{sim} - X^{act}|$ is.

The actual number of arrivals, the forecasts and the realizations based on the forecasts are plotted in Figure 6.12. As one can see from this grpah, this forecasting method leads to relatively accurate forecasts for this specific day. Furthermore, X^{fc} and $\hat{X}^{sim}$ of each interval for $X = \mathrm{SL}_1, r$ and ASA are shown in Figure 6.13. All the overall performance metrics of this day are in Table 6.6.

SL_1^{act}	$\widehat{\mathrm{SL}_1}^{sim}$	SL_1^{fc}	r^{act}	$\hat{r}^{sim}$	r^{fc}	ASA^{act}	$\widehat{\mathrm{ASA}}^{sim}$	ASA^{fc}
88.01%	96.81%	94.63%	2.93%	5.01%	6.39%	24.60	12.50	14.05

Table 6.6: Difference in SL_1, r and ASA (units: seconds).

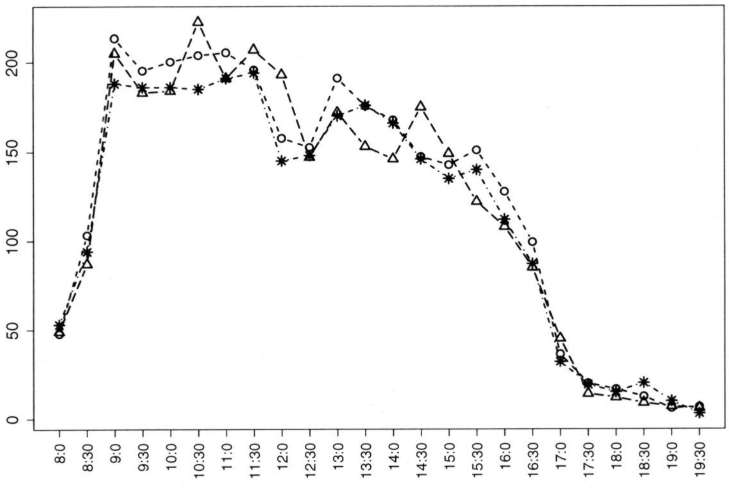


Figure 6.12: Actual number of arrivals (triangle) from the data, forecasts λ (circle) and the realizations N_λ (star) of each interval.

Now we compute $|X^{fc} - X^{act}|$ and $|\hat{X}^{sim} - X^{fc}|$:

$$|\mathrm{SL}_1^{fc} - \mathrm{SL}_1^{act}| = |94.63\% - 88.01\%| = 6.62\%, \tag{6.3}$$

$$|r^{fc} - r^{act}| = |6.39\% - 2.93\%| = 3.46\%, \tag{6.4}$$

$$|\mathrm{ASA}^{fc} - \mathrm{ASA}^{act}| = |24.60 - 14.05| = 10.55 \text{ seconds}, \tag{6.5}$$

and

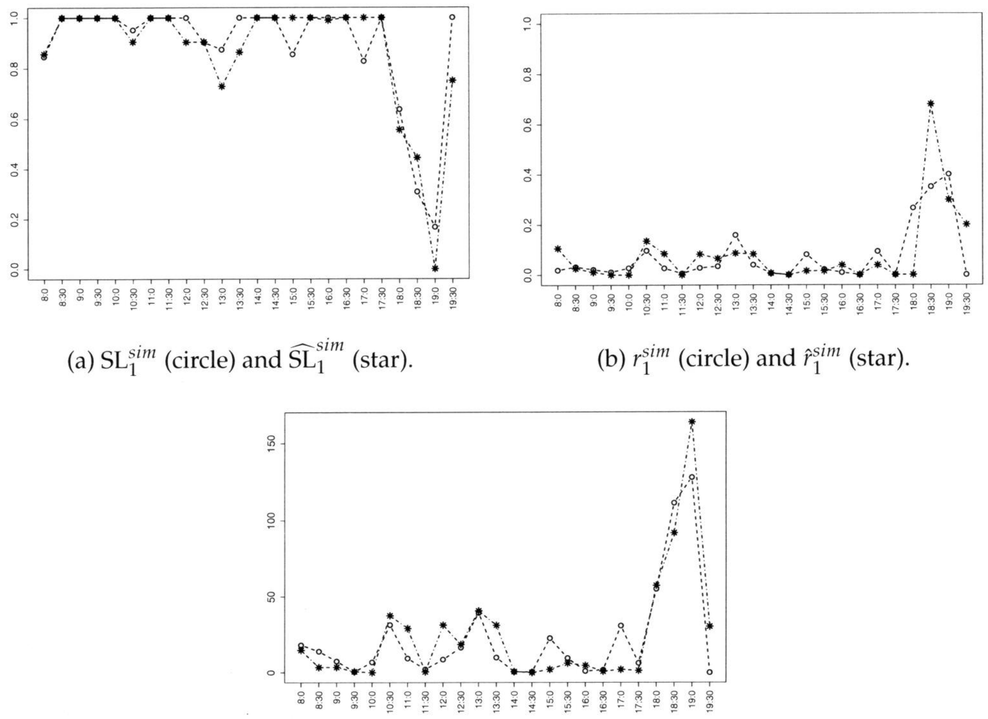
$$|\widehat{\mathrm{SL}_1}^{sim} - \mathrm{SL}_1^{fc}| = 2.18\%, \tag{6.6}$$

$$|\hat{r}^{sim} - r^{fc}| = 0.79\%, \tag{6.7}$$

$$|\widehat{\mathrm{ASA}}^{sim} - \mathrm{ASA}^{fc}| = 1.55 \text{ seconds}. \tag{6.8}$$

Equation (6.6)-(6.8) gives us some indication on how large $|\hat{X}^{sim} - X^{sim}|$ is, and it means that if we use the real underlying rate instead of using the actuals as the rates, then the error in SL will be by 2.18%, 0.79% and 1.55 seconds for SL_1, r and ASA, respectively. In Table 6.6, we know how large $|\hat{X}^{sim} - X^{act}|$ is, combining these with Inequality (6.2), we can calculate $|X_1^{sim} - X_1^{act}|$. Note that $|X_1^{sim} - X_1^{act}|$ also contains errors caused by variability, and by deducting the variability (i.e., the number between brackets in Table 6.6), we derive the modeling errors from models 2 to 9 in Table 6.7.

As one can see in model 4, the modeling error in SL_1 is negative, which can not be true. Given the fact that the error in r and ASA are both positive, we think that SL_1 being negative is coincidental and caused by variability.



(a) SL_1^{sim} (circle) and $\widehat{\mathrm{SL}_1}^{sim}$ (star).

(b) r_1^{sim} (circle) and $\hat{r}_1^{sim}$ (star).

(c) ASA^{sim} (circle) and $\widehat{\mathrm{ASA}}^{sim}$ (star) (units:seconds).

Figure 6.13: X^{sim} and $\hat{X}^{sim}$ for each interval.

6.4 Conclusion

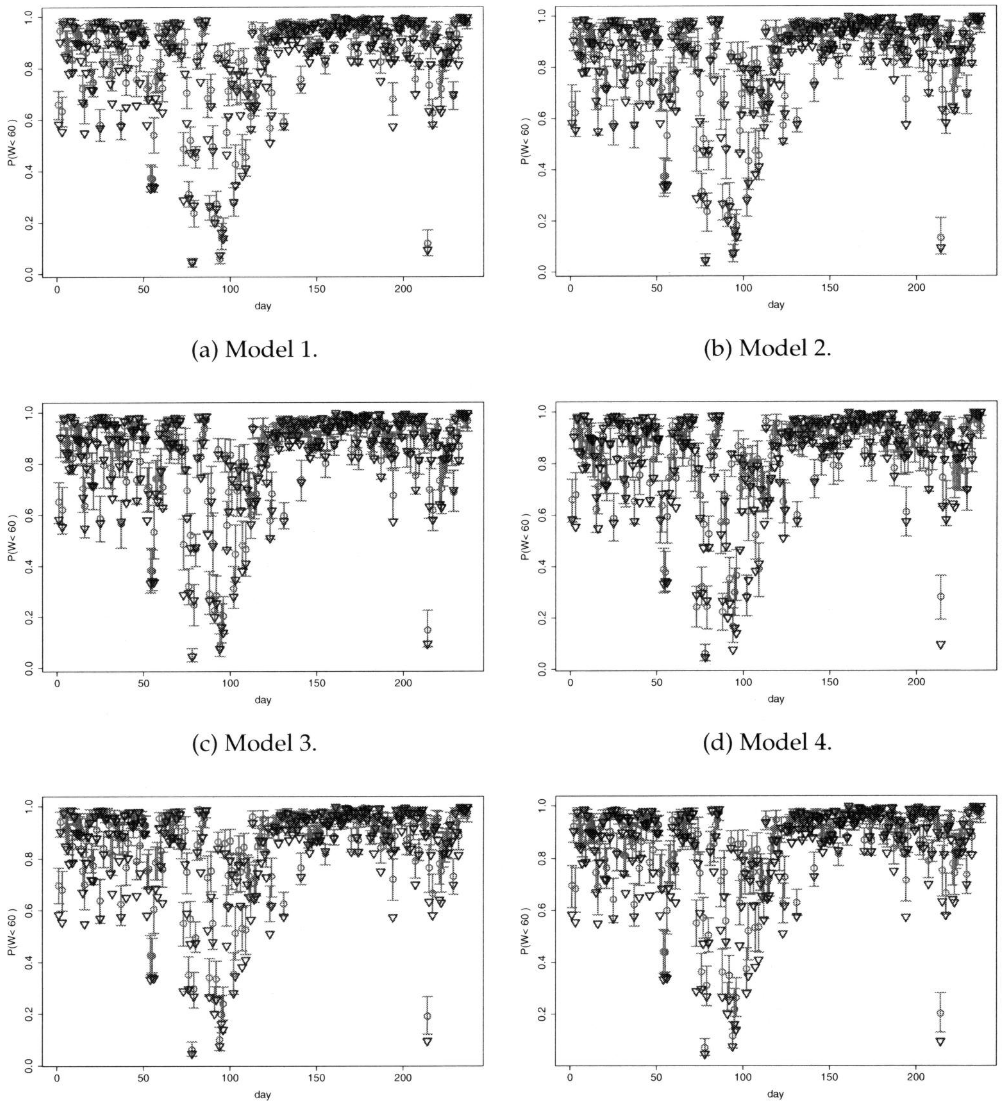
In this chapter, we validate several staffing models for multi-skill call centers. The validation is done by comparing models' predictions in SL_1, r and ASA to the actual SL_1, r and ASA from real data. The comparison results as well as the empirical analysis results suggest several important features in call centers, such as agent breaks, wrap-up times, AHT variability. We show that ignoring these features when making call center models and planning decisions will lead to large errors. Furthermore, we also verify some of the commonly used key assumptions and simplifications in call center models, such as rounding the number of agents per interval, assuming in-homogeneous Poisson arrival processes, exponential assumption of the HT and customer patience. We quantify the influences of these assumptions in terms of prediction errors in SL_1, r and ASA. It turns out that the in-homogeneous Poisson arrival processes assumption and exponential assumption of the HT do not have significant influence on the accuracy of the model, while the exponential assumption of the patience does. Last but not the least, we empirically show that the AHT of each agent differs, and the AHT of new agents decrease as they learn over time. We then develop a model to fit and predict the AHT of

	WAE_{SL_1}	WAE_r	WAE_{ASA}
Model 2	0.64%	0.902%	3.05
Model 3	0.12%	0.743%	1.73
Model 4	−0.05%	0.853%	2.68
Model 5	2.99%	0.748%	9.54
Model 6	3.07%	0.670%	9.10
Model 7	7.44%	2.384%	15.16
Model 8	2.40%	1.308%	6.35
Model 9	2.39%	1.562%	7.77

Table 6.7: Modeling errors of models 2 to 9.

each day. These two effects partially explain the variability in AHT of each day, and a staffing model with fitted AHT leads to large improvement comparing to a model that ignores such variability.

This chapter suggests the following topics for further research. First, the data we have is from a multi-skill call center. It would be interesting to analyze single-skill call center data, and validate single-skill staffing models, such as the Erlang C and Erlang A models. Second, the redial (re-attempt after abandonment) and reconnect (re-attempt after connections) behaviors are not included in the models we studied in this chapter, because we do not have customer identity information. One extension of the current chapter would be to incorporate models where redial and reconnect behaviors are included. Last but not the least, shrinkage has a big impact on call center performance, one example is the agent breaks we studied in this chapter. It would be interesting to study deeper into all causes of shrinkages.



(a) Model 1.

(b) Model 2.

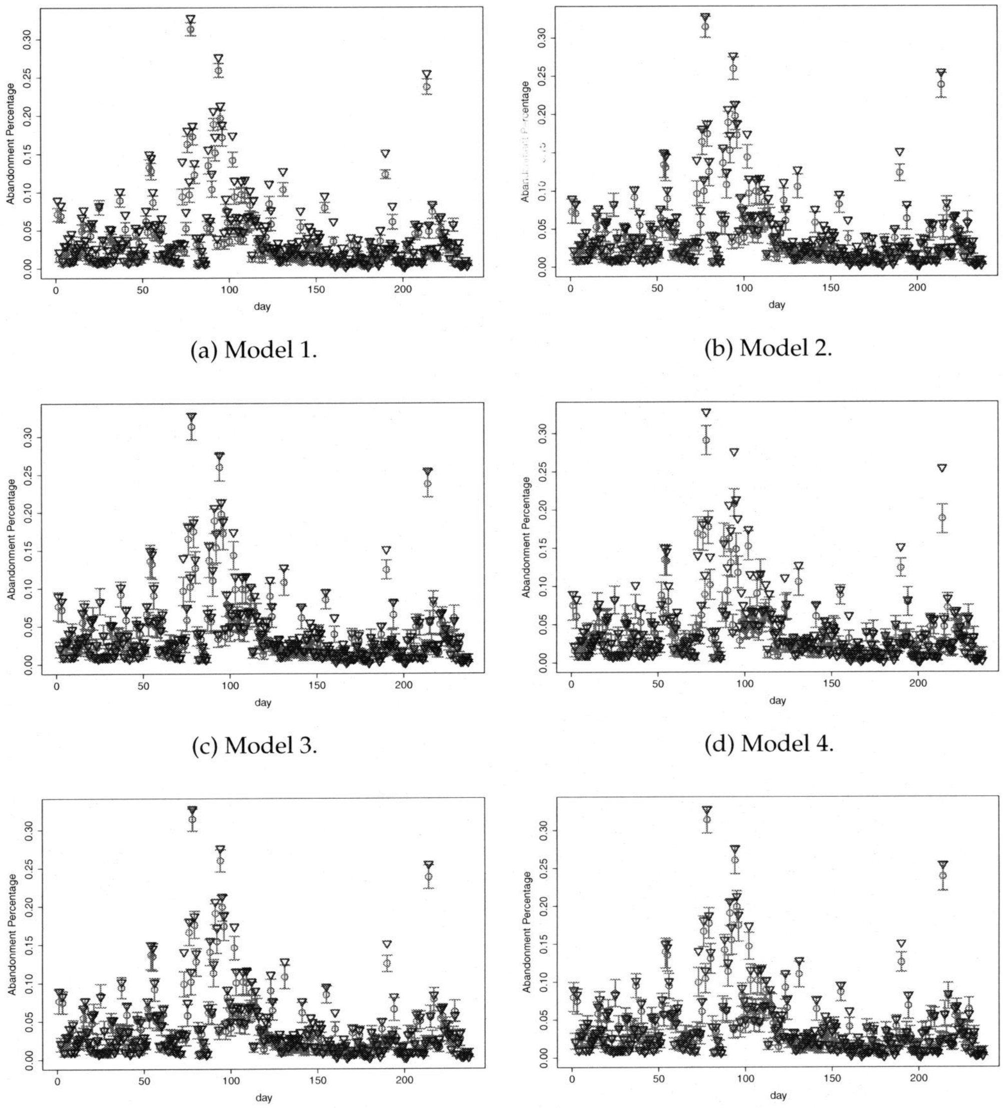
(c) Model 3.

(d) Model 4.

(e) Model 5.

(f) Model 6.

Figure 6.14: Simulation SL_1 (star) with 95% confidence interval (bar) vs. actual SL_1 (triangle).



(a) Model 1.

(b) Model 2.

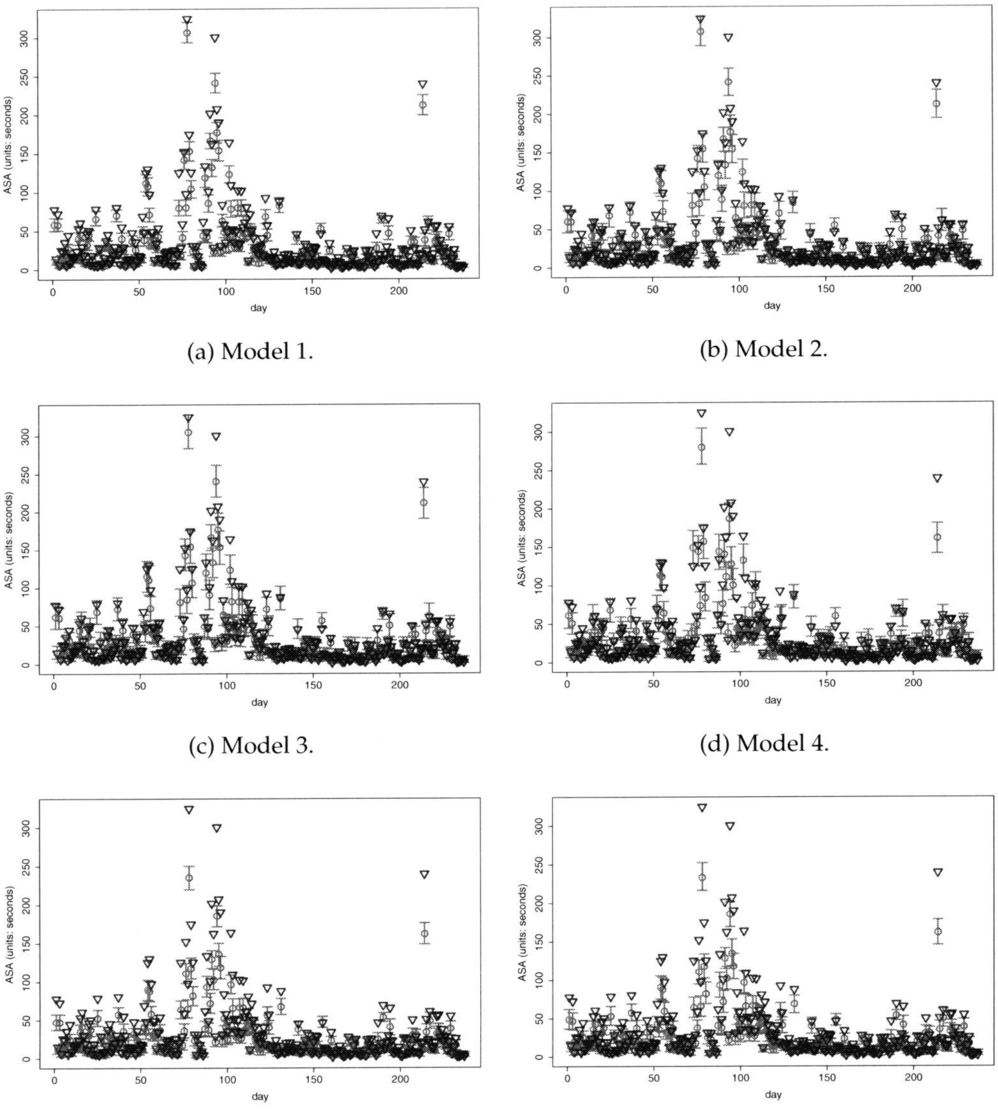
(c) Model 3.

(d) Model 4.

(e) Model 5.

(f) Model 6.

Figure 6.15: Simulation r (star) with 95% confidence interval (bar) vs. actual r (triangle).



(a) Model 1.

(b) Model 2.

(c) Model 3.

(d) Model 4.

(e) Model 5.

(f) Model 6.

Figure 6.16: Simulation ASA (star) with 95% confidence interval (bar) vs. actual ASA (triangle).

Bibliography

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.

S. Aguir, Z. Akşin, F. Karaesmen, and Y. Dallery. On the interaction between retrials and sizing of call centers. *European Journal of Operational Research*, 191(2):398–408, 2008.

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

S. Aldor-Noiman, P.D. Feigin, and A. Mandelbaum. Workload forecasting for a call center: Methodology and a case study. *The Annals of Applied Statistics*, 3:1403–1447, 2009.

B. Andrews and S. Cunningham. L.L. Bean improves call-center forecasting. *Interfaces*, 25(6):1–13, 1995.

M. Armony and C. Maglaras. Contact Centers with a Call-Back Option and Real-Time Delay Information. *Operations Research*, 52:527–545, 2004a.

M. Armony and C. Maglaras. On Customer Contact Centers with a Call-Back Option: Customer Decisions, Routing Rules and System Design. *Operations Research*, 52(2): 271–292, 2004b.

M. Armony, N. Shimkin, and W. Whitt. The impact of delay announcements in many-server queues with abandonment. *Operations Research*, 57(1):66–81, 2009.

J.R. Artalejo and M. Pozo. Numerical calculation of the stationary distribution of the main multiserver retrial queue. *Annals of Operations Research*, 116(1):41–56, 2002.

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

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

A.N. Avramidis, W. Chan, and P. L'Ecuyer. Staffing multi-skill call centers via search methods and a performance approximation. *IIE Transactions*, 41(6):483–497, 2009.

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

A. Bassamboo, R. Randhawa, and A. Zeevi. Capacity sizing under parameter uncertainty: Safety staffing principles revisited. *Management Science*, 56(10):1668–1686, 2010.

P. Billingsley. *Convergence of Probability Measures*, volume 493. Wiley-Interscience, 2009.

V Bolotin. Telephone circuit holding time distributions. In *Proceedings of the ITC*, volume 14, pages 125–134, 1994.

S. Borst, A. Mandelbaum, and M.I. Reiman. Dimensioning large call centers. *Operations Research*, 52(1):17–34, 2004.

L. 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.

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

N. Channouf and P. L'Ecuyer. A normal copula model for the arrival process in a call center. *International Transactions in Operational Research*, 19(6):771–787, 2012.

A. Charnes, W.W. Cooper, and R.O. Ferguson. Optimal estimation of executive compensation by linear programming. *Management science*, 1(2):138–151, 1955.

J. Chen and H. Rubin. Bounds for the difference between median and mean of gamma and poisson distributions. *Statistics & Probability Letters*, 4(6):281–283, 1986.

B. Cleveland and J. Mayben. *Call Center Management on Fast Forward: Succeeding in Today's Dynamic Inbound Environment*. Call Center Press, 2000.

R. Cottle, E. Johnson, and R. Wets. George b. dantzig (1914–2005). *Notices of the AMS*, 54 (3):344–362, 2007.

A. Deslauriers, P. L' Ecuyer, J. Pichitlamken, A. Ingolfsson, and A. Avramidis. Markov chain models of a telephone call center with call blending. *Computers & Operations Research*, 34(6):1616–1645, 2007.

S. Ding and G.M. Koole. Optimal call center forecasting and staffing under arrival rate uncertainty. Technical report, 2015. Under revision. Available via https://drive.google.com/file/d/0B8VA4LgASJE7VnpfTHRsdmNiU3M/edit.

S. Ding, G.M. Koole, and R.D. van der Mei. A method for estimation of redial and reconnect probabilities in call centers. In *Proceedings of the 2013 Winter Simulation Conference*, pages 181–192. Winter Simulation Conference, 2013.

S. Ding, G.M. Koole, and R.D. van der Mei. On the estimation of the true demand in call centers with redials and reconnects. *European Journal of Operational Research*, 2015a. To appear.

S. Ding, G.M. Koole, R.D. van der Mei, and R. Stolletz. The validation of staffing models for multi-skill call centers. Technical report, submitted, 2015b.

S. Ding, B. Legros, R.D. van der Mei, and O. Jouini. A call center model with a call-back option based on customers experienced waiting times. Technical report, submitted, 2015c.

S. Ding, M. Remerova, R.D. van der Mei, and A.P. Zwart. Fluid approximation of a call center model with redials and reconnects. *Performance Evaluation*, 2015d. To appear.

DMG Consulting LLC. Contact center workforce management market report reprint, 2012. Technical report, 2012. Available via http://www.nice.com/sites/default/files/nice_2012_wfm_report_reprint_final _june_2012.pdf.

S. Dudin, C. Kim, O. Dudina, and J. Baek. Queueing system with heterogeneous customers as a model of a call center with a call-back for lost customers. *Mathematical Problems in Engineering*, 2013.

A.K. Erlang. Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges. *Elektrotkeknikeren*, 13:5–13, 1917.

S. Ethier and T. Kurtz. *Markov Processes: Characterization and Convergence*. NY: John Willey and Sons, 1986.

G.I. Falin. Estimation of retrial rate in a retrial queue. *Queueing Systems*, 19(3):231–246, 1995.

G.I. Falin and J.G.C. Templeton. *Retrial Queues*. Springer, 1997.

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.

N. Gans, N. Liu, A. Mandelbaum, H. Shen, and H. Ye. Service times in call centers: Agent heterogeneity and learning with some operational consequences. In *Borrowing Strength: Theory Powering Applications–A Festschrift for Lawrence D. Brown*, volume 6, pages 99–123. Institute of Mathematical Statistics, 2010.

N. Gans, H. Shen, Y.P. Zhou, K. Korolev, A. McCord, and H. Ristock. Parametric stochastic programming models for call-center workforce scheduling. Technical report, 2012. Available via http://faculty.washington.edu/yongpin/Gans-Shen-Zhou-Scheduling.pdf.

O. Garnett, A. Mandelbaum, and M.I. 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.

K. Hoffman and C. Harris. Estimation of a caller retrial rate for a telephone information system. *European Journal of Operational Research*, 27(2):207–214, 1986.

R. Ibrahim and P. L'Ecuyer. Forecasting call center arrivals: Fixed-effects, mixed-effects, and bivariate models. *Manufacturing & Service Operations Management*, 15(1):72–85, 2013.

R. Ibrahim and W. Whitt. Real-time delay estimation based on delay history. *Manufacturing & Service Operations Management*, 11(3):397–415, 2009.

R. Ibrahim and W. Whitt. Wait-time predictors for customer service systems with time-varying demand and capacity. *Operations Research*, 59(5):1106–1118, 2011.

R. Ibrahim, M. Armony, and A. Bassamboo. Does the past predict the future? the case of delay announcements in service systems. Technical report, 2015. Available via http://www.roubaibrahim.com/DelayPred_R1.pdf.

R. Ibrahim, P. L'Ecuyer, H. Shen, and M. Thiongane. Inter-dependent, heterogeneous, and time-varying service-time distributions in call centers. *European Journal of Operational Research*, 250(2):480–492, 2016a.

R. Ibrahim, H. Ye, P. L'Ecuyer, and H. Shen. Modeling and forecasting call center arrivals: A literature survey. *International Journal of Forecasting*, 2016b. To appear.

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

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

O. Jouini, G.M. Koole, and A. Roubos. Performance indicators for call centers with impatient customers. *IIE Transactions*, 45(3):341–354, 2013.

C. Kim, O. Dudina, A. Dudin, and S. Dudin. Queueing system MAP/M/N as a model of call center with call-back option. *Analytical and Stochastic Modeling Techniques and Applications*, pages 1–15, 2012.

G.M. Koole. *Call Center Optimization*. MG books, Amsterdam, 2013.

G.M. Koole and A. Pot. An overview of routing and staffing algorithms in multi-skill customer contact centers. *Submitted*, 2006.

G.M. Koole, B.F. Nielson, and T.B. Nielson. First in line waiting times as a tool for analysing queueing systems. *Operations Research*, 60:1258–1266, 2012.

P. L'Ecuyer. Modeling and optimization problems in contact centers. In *Third International Conference on Quantitative Evaluation of Systems*, pages 145–156. IEEE, 2006.

B. Legros, O. Jouini, and G.M. Koole. Adaptive threshold policies for multi-channel call centers. *IIE Transactions*, 47(4):414–430, 2015.

S. Liao, G.M. Koole, C. Van Delft, and O. Jouini. Staffing a call center with uncertain non-stationary arrival rate and flexibility. *OR spectrum*, 34(3):691–721, 2012.

Y. Liu and W. Whitt. Stabilizing performance in many-server queues with time-varying arrivals and customer feedback. Technical report, Working paper, 2014.

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.

A. Mandelbaum, W. Massey, and M.I. Reiman. Strong approximations for markovian service networks. *Queueing Systems*, 30(1-2):149–201, 1998.

A. Mandelbaum, W. Massey, M.I. Reiman, and B. Rider. Time varying multiserver queues with abandonment and retrials. In *Proceedings of the 16th International Teletraffic Conference*, pages 737–746, 1999.

A. Mandelbaum, W. Massey, M.I. Reiman, A. Stolyar, and B. Rider. Queue lengths and waiting times for multiserver queues with abandonment and retrials. *Telecommunication Systems*, 21(2-4):149–171, 2002.

W.A. Massey and J. Pender. Skewness variance approximation for dynamic rate multiserver queues with abandonment. *ACM SIGMETRICS Performance Evaluation Review*, 39(2):74–74, 2011.

S.C. Narula, P.H.N. Saldiva, C.D.S. Andre, S.N. Elian, A.F. F., and V. Capelozzi. The minimum sum of absolute errors regression: a robust alternative to the least squares regression. *Statistics in Medicine*, 18(11):1401–1417, 1999.

B.N. Oreshkin, N. Regnard, and P. L'Ecuyer. Rate-based daily arrival process models with application to call centers. Technical report, 2014. Available via http://www.iro.umontreal.ca/~lecuyer/myftp/papers/arrival-rates-pg.pdf.

G. Pang and W. Whitt. The impact of dependent service times on large-scale service systems. *Manufacturing & Service Operations Management*, 14(2):262–278, 2012.

G. Pang, R. Talreja, and W. Whitt. Martingale proofs of many-server heavy-traffic limits for markovian queues. *Probability Surveys*, 4(7):193–267, 2007.

J. Pender and W. Massey. Approximating and stabilizing dynamic rate jackson networks with abandonment. Technical report, 2014. Available via http://www.columbia.edu/~jp3404/Jackson_Stabilize_GVA.pdf.

T. Phung-Duc and K. Kawanishi. Multiserver retrial queues with after-call work. *Numerical Algebra, Control and Optimization*, 1(4):639–656, 2011.

T. Phung-Duc and K. Kawanishi. Performance analysis of call centers with abandonment, retrial and after-call work. *Performance Evaluation*, 80:43–62, 2014.

J. Pichitlamken, A. Deslauriers, P. L'Ecuyer, and A.N. Avramidis. Modelling and simulation of a telephone call center. In *Simulation Conference, 2003. Proceedings of the 2003 Winter*, volume 2, pages 1805–1812. IEEE, 2003.

J. Reed and A. Ward. A diffusion approximation for a generalized jackson network with reneging. In *Proceedings of the 42nd Annual Allerton Conference on Communication, Control, and Computing*, 2004.

J. Riordan. *Stochastic service systems*, volume 4. Wiley New York, 1962.

T.R. Robbins, D.J. Medeiros, and T.P. Harrison. Does the erlang c model fit in real call centers? In *Proceedings of the 2010 Winter Simulation Conference*, pages 2853–2864. IEEE, 2010.

A. Roubos. *Service-Level Variability and Impatience in Call Centers*. PhD thesis, VU University Amsterdam, 2012.

A. Roubos and O. Jouini. Call centers with hyperexponential patience modeling. *International Journal of Production Economics*, 141(1):307–315, 2013.

A. Roubos, G.M. Koole, and R. Stolletz. Service-level variability of inbound call centers. *Manufacturing & Service Operations Management*, 14(3):402–413, 2012.

H. Shen and J. Huang. Interday forecasting and intraday updating of call center arrivals. *Manufacturing & Service Operations Management*, 10(3):391–410, 2008.

S. Steckley, S. Henderson, and V. Mehrotra. Forecast errors in service systems. *Probability in the Engineering and Informational Sciences*, 23(2):305, 2009.

R. Stolletz. Approximation of the non-stationary $m(t)/m(t)/c(t)$-queue using stationary queueing models: The stationary backlog-carryover approach. *European Journal of Operational Research*, 190(2):478–493, 2008.

D. Sze. OR practice - a queueing model for telephone operator staffing. *Operations Research*, 32(2):229–249, 1984.

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

J. Taylor. Density forecasting of intraday call center arrivals using models based on exponential smoothing. *Management Science*, 58(3):534–549, 2012.

J. Weinberg, L. Brown, and J. Stroud. Bayesian forecasting of an inhomogeneous poisson process with applications to call center data. *Journal of the American Statistical Association*, 102(480):1185–1198, 2007.

W. Whitt. Dynamic staffing in a telephone call center aiming to immediately answer all calls. *Operations Research Letters*, 24(5):205–212, 1999.

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

W. Whitt. Fluid models for multiserver queues with abandonments. *Operations Research*, 54(1):37–54, 2006.

H. Ye, J. Luedtke, and H. Shen. Forecasting and staffing call centers with multiple interdependent uncertain arrival streams. Technical report, 2014.

G. Yom-Tov and A. Mandelbaum. Erlang-r: A time-varying queue with reentrant customers, in support of healthcare staffing. *Manufacturing & Service Operations Management*, 16(2):283–299, 2014.

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.

Samenvatting

Workforce Management in Call Centers: Forecasting, Staffing and Empirical studies

Veel calcentermanagers en -planners streven naar een betere personeelsplanning, en een goede balans tussen de kwaliteit van de dienstverlening en operationele kosten, wat neerkomt op een balans tussen wachttijden en het aantal in te zetten agents. In het bijzonder moeten callcenters aan service targets voldoen met zo min mogelijk agents. Er is helaas geen simpele oplossing voor dit dilemma. Door de jaren heen hebben onderzoekers en bedrijven modellen ontwikkeld om dit beslissingsproces te ondersteunen. De volledige procedure bestaat in het algemeen uit vier stappen: het voorspellen van het aantal binnenkomende telefoongesprekken, het maken van een personeelsplanning, het maken van roosters, en het managen van telefoonverkeer.

Het voorspellen van binnenkomende gesprekken is ingewikkeld vanwege het feit dat er veel onzekerheid mee gemoeid is. Sterker nog, het is onmogelijk om van te voren het aantal binnenkomende gesprekken vast te stellen. Forecasters streven naar voorspellingen die zo nauwkeurig mogelijk zijn. Dit leidt tot de volgende vraag: "hoe meet men nauwkeurigheid?" Er zijn verschillende maten voor de voorspellingsfout, zoals de mean squared error en de mean percentage error. Verschillende maten leiden tot verschillende keuzes van forecastingmodellen en forecasts. In Hoofdstuk 2 onderzoeken we dit probleem, en komen we erachter dat onder rate uncertainty de weighted mean absolute percentage error de optimale maat is voor de voorspellingsfout. Dit komt doordat de voorspellingen die deze maat minimaliseren ook astymptotisch de initi ́ele personneelsplanningskosten en telefoonverkeerskosten minimaliseren. We laten ook zien dat onder zekere aannames personeelsplanningsbeslissingen gemaakt moeten worden op basis van bepaalde percentielen van de verdeling van de forecast, in plaats van enkel de verwachting van de forecast.

In Hoofdstuk 3 beschrijven we literatuur over forecasting modellen. Naast het kiezen van een geschikt model is het belangrijk om de data die gebruikt wordt bij de forecasts te analyseren. Aan de hand van callcenter datasets laten we zien dat klanten vaak ophangen en terugbellen (redial), of terugbellen nadat ze doorgeschakeld waren (reconnect). Dit gedrag heeft een significante invloed op volumes. Verder laten we zien dat het noodzakelijk is het aantal unieke bellers te gebruiken bij het voorspellen in plaats van het totaal aantal bellers, omdat het aantal unieke bellers niet afhankelijk is van de personeelsplanning, maar het totaal aantal bellers wel. We tonen aan dat als het totaal aantal bellers gebruikt wordt dit kan leiden tot een onnauwkeurige schatting van het aantal bellers.

Hoewel het geobserveerde terugbelgedrag significant is, is er geen personeelsplanningsmodel dat beide aspecten omvat. In Hoofdstuk 4 ontwikkelen we zo'n model en benaderen de performance metrics aan de hand van fluid benaderingen. De fluid benadering geeft een eerste orde benadering van het aantal bellers dat zich in de wachtrij bevindt of in behandeling is, in het redial en reconnect proces. Op basis van deze getallen benaderen we de verwachte totale rate van binnenkomende gesprekken van het systeem, en gebruiken dit als input voor de Erlang A formule om de verdeling van de wachttijd af te leiden.

Callcenters kunnen de redials, reconnects, en nieuwe bellers anders behandelen. In Hoofdstuk 5 presenteren we een voorbeeld, waar we resultaten van een callcenter onderzoeken waar het mogelijk is om terug te bellen. Als het systeem vol zit, worden langwachtende klanten geadviseerd om later terug te bellen, en als ze terugbellen krijgen zij prioriteit over nieuwe bellers. We bespreken dit model en laten zien dat het de gemiddelde wachttijd efficient reduceert en ook dat het fair is ten opzichte van langwachtende bellers.

In de meeste literatuur over callcenters stellen onderzoekers modellen voor om de werkelijkheid in callcenters na te bootsen, met aannames en vereenvoudigingen van bepaalde processen. Er worden analytische oplossingen of benaderingen gegeven, en eventueel worden de methodes numeriek doorgerekend. In Hoofdstuk 6 gebruiken we een andere benadering door sommige annames en vereenvoudigingen te valideren. Dit doen we door de realiteit te vergelijken met simulaties. In het bijzonder vergelijken we de gesimuleerde service levels van enkele personeelsplanningsmodellen met de werkelijke service levels uit de data, voor een multi-skill callcenter. We laten empirisch zien dat modellen in het algemeen nauwkeurig zijn ondanks de aannames; echter, sommige vereenvoudigingen en aannames moeten met zorg behandeld worden. Bijvoorbeeld, pauzes van agents worden vaak niet meegenomen in planningsmodellen, maar zijn wel belangrijk en moeten niet genegeerd worden. Verder laten we wat empirische resultaten zien op het gebied van shrinkage, heterogeniteit van agents, de learning curve van agents, etc., die meer inzicht geven voor managers en planners.

Summary

Workforce Management in Call Centers: Forecasting, Staffing and Empirical Studies

Many call center managers and planners strive to make better workforce planning, and to balance well the quality of service and operational costs, which is essentially balancing customers waiting times and the number of agents. In specific, call centers need to satisfy service level targets with the least number of agents. There is no simple solution to this dilemma. However, researchers and practitioners over the decades have developped experiences and models to assist this decision making process. The whole procedure generally includes four steps: forecasting future call arrival volume, make staffing decisions, make rosters, traffic management.

Predicting future call volume is difficult, as the incoming number of calls involves large uncertainty. In fact, one can never know for sure how many inbound calls in advance. Forecasters aim at forecasts that are as accurate as possible. However, one question naturally arises:"how to measure accuracy?" There are many different measurements for the forecasting errors, such as the mean squared error, the mean percentage error, etc. Choosing different error measurements can sometimes lead to different choices of forecasting models or forecasts. In Chapter 2, we investigate this problem, and we discover that under the rate uncertainty, the weighted mean absolute percentage error is the optimal error measurement. This is because the forecasts that minimize it will also asymptotically minimize the initial staffing costs plus the traffic management costs. Also, we show that under certain assumptions, the staffing decision should be made based on certain percentile of the distributional forecasts, rather than the mean.

In Chapter 3, we show the literature on forecasting models. Besides choosing an appropriate model, another important factor is the data to be used in the forecasts. We show that in call center data sets there are redials and reconnects, as customers call back after abandonments and connected calls. Both behaviors have significant influence on the call volumes. Furthermore, we show that one should use the number of fresh calls (number of unique callers) to make forecasts instead of the total number of calls, since the number of fresh calls do not depend on the staffing decisions, while the total number of calls do. It is shown that by using the total number of calls, it may lead to inaccurate estimation of the call volume.

The redial and the reconnect behaviors are significant, yet, there is no staffing model

that supports both features. In Chapter 4, we develop such a model and approximate the performance metrics by using fluid approximation. The fluid approximation gives a first order approximation on the number of callers in the queue and in service, in the redial and reconnect orbits. Based on those numbers, we approximate the mean total arrival rate to the system, and use it as an input to the Erlang A formula to derive the waiting time distribution.

Call centers may treat differently between the redials, the reconnects and the fresh calls. We show one example in Chapter 5, where we study the performance of a call center model with an call-back option, and the long-waiting callers are suggested to call back some time later if the system is congested, and when they call back, they will receive priorities over the fresh callers. It is discussed and shown that this model is efficient in reducing the mean waiting time and it is fair to those long waiting callers.

In most of literature in call centers, researchers usually propose models to mimic the real situation in call centers, where assumptions and simplifications of certain processes are made. Then analytical solutions or approximation methods are developed, and eventually the methods are evaluated numerically. In Chapter 6, we take a different approach by validating some of the commonly made assumptions and simplifications. This is done by comparing the reality with simulation. To be more specific, we compare the simulated service levels of a few staffing models with the actual service levels from the data for a multi-skill call center. We empirically show that models are in general accurate despite the assumptions made; however, certain simplifications and assumptions should be treated with care. For example, agents' breaks are important and should not be ignored, which usually are not taken into considerations in staffing models. Furthermore, we provide some empirical results in shrinkage, agents heterogeneity, agents learning curve, etc., which give insight to managers and planners.

Summary Chinese

呼叫中心的劳动力管理：预测，坐席数计算和数据分析

呼叫中心管理人员和规划人员无不尽力的做好的劳动力规划来平衡好服务质量和服务成本。这一管理过程其实也就是平衡客户等待时间和坐席数量的过程。具体的来说，这个管理过程中的核心问题就是解决如何用最少的人员来达到预定的服务指标。要解决这一个问题并不简单。在过去的几十年中，国内外许多研究人员和从业人员都在不断总结以往的经验和研发新的模型来帮助呼叫中心的管理。整体的来说，劳动力管理过程包括了四个步骤：话务量预测，决定坐席数，人员排班和实时话务量管理。

预测话务量并不简单，因为未来话务量的多少包含了很多的不确定性。事实上，没有人可以100%准确的预测未来的话务量。预测者要做的就是要做到让预测尽可能的准确。那么，这其中就产生了一个问题：怎么来衡量预测的误差？要知道衡量预测的误差有很多标准，比如，平均方差，平均百分比差，等等。有时，选择不同的误差衡量方式会得到不同的模型和预测。在本文的第二章，我们深入的探讨了这个问题。我们证明当存在话务量的不确定性时，最优的误差衡量方式应为权重平均绝对百分比差。之所以说它是最优的，因为如果一个预测最小化了它，那个根据这个预测得到的排班也将最小化初始排班加上实时排班调整产生的费用。并且，在一定的假设条件下，坐席数应根据百分位数来计算，而不是平均值。

在第三章中，我们学习了很多关于预测方法的文献。但是，除了选对正确的预测方法之外，还有一个不容忽视的重点，那就是数据使用。通多对几个呼叫中心数据的分析，我们发现呼叫中心的话务量中有很多重复呼叫。原因之一是客户在等待接通过程中失去耐心，所以选择挂断，并在挂断一段时间后重新呼叫；另外一个原因是客户在接通并完成通话之后挂断，并在挂断一段时间之后重新呼叫。这两种重新呼叫的客户行为都对呼叫中心得话务量有很大的影响。通过研究，我们发现，在进行话务量预测时，最好采用客户数量来进行建模，而不是用总的呼入量。这是因为客户数量的多少与坐席数并不相关，而总的呼入量却与坐席数相关。我们通过仿真模拟发现，当使用总的呼入量来建模并预测未来的呼入量时，会产生很大的误差。

虽然这两种重新呼叫都是很显著的客户行为，但是很少有排班模型考虑到这两种客户行为。在第四章，我们将研究一个考虑到重新呼叫这两种行为的模型，并对这个模型的进行流体近似。流体近似对这个排队论模型的以下均值进行近似：在队列中的客户和服务中的客户之和，挂断重拨的客户和接通后重拨的客户。根据这些均值，我们进一步的利用Erlang A模型来得到等待时间分布。

有的呼叫中心区分对待重拨的客户和第一次拨入的客户。在第五章中，我们研究了让客

户选择重拨的模型。具体的来说，在这个模型中，当客户等待时间过长时，他们会听到一个语音，提示系统正忙，并建议客户过一段时间之后再重拨。为了让简短重拨的客户的等待时间，重拨的客户将得到接听的优先权。通过研究发现，这个模型对于降低平均等待时间非常有效。

在大多数的呼叫中心研究中，研究人员通常都做一定的假设。然后对简化和假设之后的模型来进行求解和逼近。在第六章中，不同与一般的研究方法，我们验证了一些常用到的假设条件。我们所用的方法是比较实际数据和仿真模拟。更准确的来说，我们对通过仿真模拟得到的服务水平和实际服务水平的进行对比。对比的结果显示，某些假设条件应该在实际运用中小心对待，比如坐席与坐席的不同，坐席的学习曲线，等等，这些都会对模型准确性产生大的影响，并且，通过对他们的研究，管理人员也能增加对呼叫中心的理解。