European Journal of Operational Research 246 (2015) 250–262

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Innovative Applications of O.R.

On the estimation of the true demand in call centers with redials and reconnects

S. Ding[a,b,*], G. Koole[b], R. D. van der Mei[a,b]

[a] Center for Mathematics and Computer Science (CWI), Amsterdam, The Netherlands
[b] VU University Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Received 4 February 2015
Accepted 11 April 2015
Available online 18 April 2015

Keywords:
Queueing
Forecasting
Redials
Reconnects
Call centers

ABSTRACT

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 paper 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 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 estimations of redial and reconnect probabilities and the number of fresh calls are close to the real ones, both via real data analysis and simulation.

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, Henderson, & Mehrotra, 2010). There is extensive literature on different forecasting methods applied to call centers. Andrews and Cunningham (1995) used the Autoregressive Integrated Moving Average (ARIMA) method to forecast the inbound call volume of the L. L. Bean's call center. Taylor (2012) adjusted the traditional Holt-Winters exponential smoothing method to the Poisson count model with gamma-distributed arrival rate, and took both intraweek and intraday patterns into account in his model. Taylor (2008) compared the accuracy of a few forecasting models for a British retail bank call center. He concluded 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) used the Singular Value Decomposition (SVD) method to reduce the dimension of square-root-transformed call center data. Then they applied time series and regression analysis techniques to make distributional forecasts. Besides the forecasts, they 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 was then further developed in Avramidis, Deslauriers, and L'Ecuyer (2004), where three variants of doubly stochastic model were analyzed and compared. Ibrahim and L'Ecuyer (2013) added 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 was used by Gans, Shen, Zhou, Korolev, McCord, and Ristock (2009).

Call center forecasting models aim to achieve the minimum error in the forecasts, where total inbound volumes are used. In this paper, we show that the true inbound volume (we refer to it as the *fresh*

* Corresponding author. Tel.: +31684120642.
E-mail addresses: s.ding@cwi.nl, dingsihan@hotmail.com (S. Ding), ger.koole@vu.nl (G. Koole), mei@cwi.nl (R.D. van der Mei).

http://dx.doi.org/10.1016/j.ejor.2015.04.018
0377-2217/

volume from now on) 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. We define a *redial* as a reattempt of the abandoned calls, and reconnect as re-attempt of answered calls. 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 what is 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 more stable over time, since they represent the customer behaviors. In this paper, 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, Massey, Reiman, Stolyar, and Rider (2002), Aguir, Karaesmen, Akşin, and Chauvet (2004), Aguir, Akşin, Karaesmen, and Dallery (2008). The reconnect behavior (also referred as feedback or re-entrant in some papers) in service industry has been studied by Yom-Tov and Mandelbaum (2014). They 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. In Yom-Tov and Mandelbaum (2014), customer abandonment is not included in their model. In all the existing works mentioned above, it is assumed that the retrial/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 certain intraday pattern, see (Gans et al., 2009; Ibrahim & L'Ecuyer, 2013; Shen & Huang, 2008). In this paper, we propose a queueing model that has two extra orbits than 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 in accurate 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 estimations are close to the real values. A shorter version of this paper has appeared in Ding, Koole, and van der Mei (2013).

The remainder of the paper is organized as follows. In Section 2, we describe the queueing model. We also show simulation examples of such a model to understand the influence of redials and reconnects on the total volume, as well as the necessity of distinguishing the fresh calls, redials and reconnects. In Section 3, we present our estimation models both for constant arrival rate and arrival rate with intraweek seasonal patterns. These estimation models are also validated via simulation as well as real call center data sets.

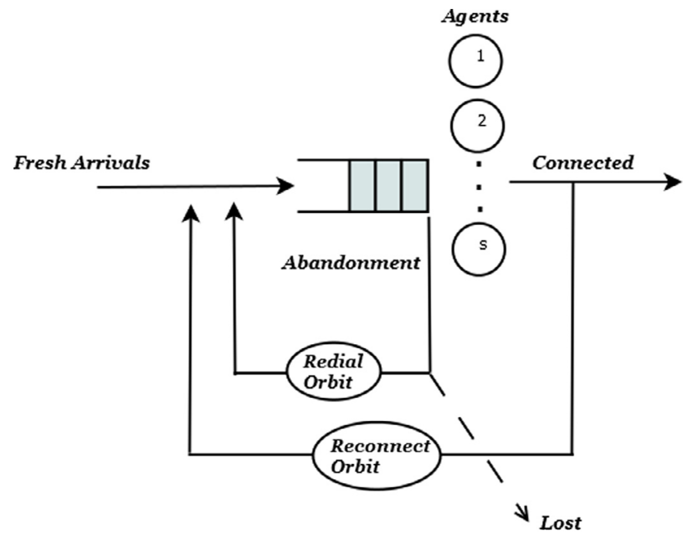

Fig. 1. Queueing diagram of a call center.

2. Model description

Consider the queueing model illustrated in Fig. 1. We assume that calls arrive according to a Poisson process. We refer to these calls as *fresh calls*. There are s agents who handle inbound calls. An arriving call will be 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. After a generally distributed amount of time H, a waiting customer who did not get connected to an agent will lose his patience and abandon. We assume $\mathbb{E}H = \theta < \infty$. With probability p, an abandoned customer will enter the redial orbit, and he will redial after some generally distributed amount of time Γ_{RD}, with $\mathbb{E}\Gamma_{RD} = \delta_{RD} < \infty$. We refer to these calls as *redials*. With probability $1 - p$, this customer will not call back, and this call is considered as a 'lost' call. We assume that the service time B of a customer has a general distribution with mean $\mathbb{E}B = 1/\mu < \infty$. After the call has been finished, this customer will enter the reconnect orbit with probability q, and he will reconnect after some generally distributed time Γ_{RC}, with $\mathbb{E}\Gamma_{RC} = \delta_{RC} < \infty$. We refer to these calls as *reconnects*. We assume that p and q do not depend on customers' experiences in the system. These experiences include holding time, waiting time and the number of times that customers have already called. We use this queueing model to represent the situation of a single-skill call center.

According to the model description, the total volume is influenced by the service level in the call center, 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, illustrated by Fig. 1. The fresh arrival rate is set to be 10 calls per minute every day,