European Journal of Operational Research 250 (2016) 480–492

Contents lists available at ScienceDirect

European Journal of Operational Research

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

Stochastics and Statistics

Inter-dependent, heterogeneous, and time-varying service-time distributions in call centers

Rouba Ibrahim [a,*], Pierre L'Ecuyer [b], Haipeng Shen [c], Mamadou Thiongane [b]

[a] School of Management, University College London, London WC1E 6BT, UK
[b] Department of Computer Science and Operations Research, University of Montreal, Montreal, QC H3C 3J7, Canada
[c] Innovation and Information Management, School of Business, University of Hong Kong, Hong Kong, China

ARTICLE INFO

Article history:
Received 21 April 2014
Accepted 8 October 2015
Available online 23 October 2015

Keywords:
Applied probability
Call centers
Service times
Agent heterogeneity
Correlation

ABSTRACT

Traditionally, both researchers and practitioners rely on standard Erlang queueing models to analyze call center operations. Going beyond such simple models has strong implications, as is evidenced by theoretical advances in the recent literature. However, there is very little empirical research to support that body of theoretical work. In this paper, we carry out a large-scale data-based investigation of service times in a call center with many heterogeneous agents and multiple call types. We observe that, for a given call type: (a) the service-time distribution depends strongly on the individual agent, (b) that it changes with time, and (c) that average service times are correlated across successive days or weeks. We develop stochastic models that account for these facts. We compare our models to simpler ones, commonly used in practice, and find that our proposed models have a better goodness-of-fit, both in-sample and out-of-sample. We also perform simulation experiments to show that the choice of model can have a significant impact on the estimates of common measures of quality of service in the call center.

1. Introduction

The effective management of call centers is a challenging task mainly because managers are consistently facing considerable uncertainty; see Gans, Koole, and Mandelbaum (2003) and Aksin, Armony, and Mehrotra (2007) for background on call centers. Among important sources of uncertainty are call arrival rates which are typically both time-varying and stochastic, service times which are random and whose distribution may depend on the call type and the agent who handles it, and agents who may not show up or may not follow their planned schedules; see Bhulai and Koole (2003), Avramidis, Deslauriers, and L'Ecuyer (2004), Avramidis and L'Ecuyer (2005), Aldor-Noiman, Feigin, and Mandelbaum (2009), Gans, Liu, Mandelbaum, Shen, and Ye (2010), Ibrahim, L'Ecuyer, Régnard, and Shen (2012), Oreshkin, L'Ecuyer, and Régnard (2015), and references therein.

In this paper, we focus on the effective modelling of service times in call centers. In particular, we carry out a large-scale, in-depth, empirical investigation of service times in call centers. We analyze data gathered at the call center of Hydro-Québec (HQ), which is a government-owned public utility overseeing the generation, transmission, and distribution of electricity for the province of Quebec, in Canada. This real call center setting is complex, consisting of many heterogeneous agents and multiple distinct call types. Our data show that service times differ greatly across such agents, vary in time, and exhibit strong serial and cross correlations. We propose new service-time models which account for those features, and which are a good fit to real-life data, both in-sample and out-of-sample. We are interested in out-of-sample predictions because it is important to verify that our models are reliable tools to predict the future mean service times of agents, which have a considerable impact on future system performance.

Proposing new and more realistic service-time models, as we do in this paper, is important for the effective simulation of call centers. Simulation is an important tool that can be used to evaluate performance measures such as service levels and average waiting times, and to construct work schedules for agents and call routing rules by stochastic optimization algorithms (Avramidis, Chan, Gendreau, L'Ecuyer, & Pisacane, 2010; Chan, Koole, & L'Ecuyer, 2014). We use simulation to show that the choice of service-time model can have a significant impact on the performance measures in call centers, and formulate valuable insights on the practical usefulness of those findings.

* Corresponding author. Tel.: +447786162382.
E-mail addresses: rouba.ibrahim@ucl.ac.uk (R. Ibrahim), lecuyer@iro.umontreal.ca (P. L'Ecuyer), haipeng@email.unc.edu (H. Shen), mdthiongane@yahoo.fr (M. Thiongane).

http://dx.doi.org/10.1016/j.ejor.2015.10.017
0377-2217/

1.1. Background, positioning in the literature, and contributions

Traditionally, both researchers and practitioners relied on standard Erlang queueing models to analyze call center operations. In Erlang queueing models, agent service times are modelled as independent and identically distributed exponential random variables with a constant mean. Going beyond this standard modelling assumption has important operational consequences, as is evidenced by multiple advances in the recent literature.

1.1.1. Agent heterogeneity

There is a stream of papers which studies queueing models with heterogeneous servers, with applications to call center management. One central question which arises in this context is how to route incoming calls to heterogeneous agents so as to minimize a given performance measure, such as the mean waiting time. Given the complexity of this problem, most papers resort to finding optimal routing policies in large-scale systems under heavy-traffic conditions; e.g., see Armony (2005), Gurvich and Whitt (2009), Armony and Ward (2010), Armony and Mandelbaum (2011), and references therein. Mehrotra, Ross, Ryder, and Zhou (2012) resort to a numerical study to characterize overall performance in terms of customer waiting time and overall resolution rate. In general, these papers show that control decisions can actually benefit from agent heterogeneity, e.g., routing incoming calls to the fastest idle agents reduces customer waiting.

There is very little empirical research supporting that body of theoretical work. To the best of our knowledge, the only exception is Gans et al. (2010) who analyzed call-center data and identified both short-term and long-term factors associated with agent heterogeneity in practice. They also described results from a small simulation study illustrating the operational consequences of ignoring such heterogeneity. (We revisit their example and extend some of their conclusions in Section 6.) Gans et al. indicated that an interesting extension of their paper is to incorporate random effects in service-time models so as to "capture within-agent dependence among the calls handled by the same agent, and enable understanding of a whole agent population" (p. 118). We consider such random-effects models in this paper. In general, random effects represent additional, unobservable and uncontrollable, variability which causes systematic deviations from the average performance of the agent, and due to which successive service times may be dependent. Dependencies between service times are often observed in data, and are therefore important to model, as we do in this paper.

1.1.2. Dependencies among the service times

Service times in practice are often dependent. For one example, an agent may be overworked in given periods (e.g., in weeks where congestion is higher than usual) and this could affect his/her performance in all services that s/he performs during such periods, typically resulting in that agent either slowing down or speeding-up; see Delasay, Ingolfsson, and Kolfal (2015), Dong, Feldman, and Yom-Tov (2015), Feldman, Li, Yom-Tov, and Yom-Tov (2015), and references therein. In this case, agents (servers) may be viewed as *strategic decision makers* that influence their own service rates. As a result of such strategic behavior, successive service times are dependent. For a second example, in a technical call center, there may be a product defect due to which there are multiple related calls, whose durations are all longer than average. In this example too, service times (call durations) are dependent.

There is a well-developed theory studying the performance impact of dependence among service times in single-server queues; e.g., see Chapter 9 of Whitt (2002) for a detailed treatment. However, Pang and Whitt (2012) are among the first to consider the multi-server case, which is more reasonable from a practical perspective. They considered a weakly dependent stationary sequence of service times and demonstrated that, in the heavy-traffic limit, the impact of those dependencies is determined by the bivariate cumulative distribution function of service times. In their numerical study, they considered an EARMA sequence of service times, which is stationary with exponential marginal distributions and the correlation structure of an autoregressive-moving average process. Pang and Whitt demonstrated, via theoretical analysis and computer simulation, how dependencies between service times can significantly alter large system performance. In particular, they showed that those correlations strongly impact the distribution of the number of customers in queue which, in turn, affects staffing decisions. Pang and Whitt concluded their paper by calling for "empirical studies to estimate the strength of dependence among service times in applications" (p. 278). We conduct such a study in this paper.

1.1.3. Time dependence

There are relatively few papers which consider queueing models with time-varying service rates, since this feature substantially complicates the analysis. Some exceptions include Mandelbaum, Massey, Reiman, and Stolyar (1999), Liu and Whitt (2011), and references therein. These papers demonstrate the operational impact of including time-varying service rates; their results apply generally and do not assume a specific form for time dependence in the service rate. Aldor-Noiman et al. (2009) used predictions of future arrival counts and mean service times to estimate future loads in call centers. Aldor-Noiman et al. allowed for mean service times to be time-dependent, and showed how errors in predicting future loads can impact staffing decisions. Their paper assumed homogeneous agents and a single call type. Our service-time models account for time dependence as well, albeit in a much more complex setting, with multiple call types and many heterogeneous agents.

1.1.4. Lognormal distribution

In their seminal paper, Brown et al. (2005) performed a detailed statistical analysis of call center data and showed that service times are not exponentially distributed, as was traditionally assumed, and that the lognormal distribution is a remarkably good fit for the service-time distribution instead. Deslauriers (2003) had also observed the same thing. Motivated by this, Shen and Brown (2006) proposed a new method for inference about non-parametric regression curves when the errors are lognormally distributed, and illustrated their method with both a simulation study and the analysis of real-life call center data. Mandelbaum and Zeltyn (2010) advocated a process-view of service times which are modeled as the evolution of a finite-state continuous-time absorbing Markov process (phase-type distribution). Here, even though we use additional information when modelling service times, such as the time when the call is answered, we continue to assume the lognormality of the individual service times.

In this paper, we supplement the body of theoretical research above with supporting empirical work. As such, we take a step toward filling that gap in the literature. In addition to proposing new service-time models that are a good fit to data, we quantify the performance impact of our alternative service-time models through a simulation study.

1.2. Organization

Here is how the rest of this paper is organized. In Section 2, we describe and do a preliminary analysis of the data set that motivated this research. In Section 3, we describe our candidate models. In Section 4, we compare the in-sample goodness of fit of our models. In Section 5, we compare the out-of-sample predictive accuracy of our models for a large pool of agents. In Section 6, we present the results of simulation experiments which quantify the performance impact of our different models. In Section 7, we make concluding remarks.