Contents lists available at ScienceDirect

International Journal of Forecasting

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

Modeling and forecasting call center arrivals: A literature survey and a case study

Rouba Ibrahim [a,*], Han Ye [b], Pierre L'Ecuyer [c], Haipeng Shen [d]

[a] School of Management, University College London, United Kingdom
[b] Business Administration, University of Illinois at Urbana Champaign, United States
[c] Computer Science and Operations Research, University of Montreal, Canada
[d] Faculty of Business and Economics, The University of Hong Kong, Pokfulam, Hong Kong

ARTICLE INFO

Keywords:
Call center arrivals
Forecasting
Time series
Doubly stochastic Poisson
Fixed-effects
Mixed-effects
ARIMA
Exponential smoothing
Bayesian
Dimension reduction
Dependence
Seasonality
Marketing events

ABSTRACT

The effective management of call centers is a challenging task, mainly because managers consistently face considerable uncertainty. One important source of this uncertainty is the call arrival rate, which is typically time-varying, stochastic, dependent across time periods and call types, and often affected by external events. The accurate modeling and forecasting of future call arrival volumes is a complicated issue which is critical for making important operational decisions, such as staffing and scheduling, in the call center. In this paper, we review the existing literature on modeling and forecasting call arrivals. We also discuss the key issues for the building of good statistical arrival models. In addition, we evaluate the forecasting accuracy of selected models in an empirical study with real-life call center data. We conclude with a summary of possible future research directions in this important field.

1. Introduction

The call center services industry is large and important, with more than 2.7 million agents working in the United States and 2.1 million agents working in Europe, the Middle East, and Africa (Akşin, Armony, & Mehrotra, 2007). Managing a call center efficiently is a challenging task, because managers have to make staffing and scheduling decisions in order to balance staffing costs and service quality, which always conflict, in the presence of uncertainty as to arriving demand. Most staffing or scheduling plans start with the forecasting of customer call arrivals, which are highly stochastic. Accurate forecasts of call arrivals are key for the achievement of optimal operational efficiency, since under-forecasting leads to under-staffing and therefore long customer waits, while over-forecasting results in money being wasted on over-staffing.

The customer arrivals process is nontrivial. This process can be modeled as a Poisson arrival process, and has been shown to possess several features (Akşin et al., 2007; Ceẓik & L'Ecuyer, 2008; Gans, Koole, & Mandelbaum, 2003; Garnett, Mandelbaum, & Reiman, 2002; Wallace & Whitt, 2005). One of the most important of these features is the fact that the arrival rate is time-varying, which adds to the complexity of the forecasting process. Call arrival rates may exhibit intraday, weekly, monthly, and yearly seasonalities. While a time-inhomogeneous Poisson arrival process can easily capture time dependence in call arrival data, it often fails to capture other characteristics. For one thing, call center arrivals typically exhibit a significant dispersion relative to the Poisson distribution. Thus, a doubly stochastic Poisson arrival process may be more appropriate, e.g., see Aldor-Noiman, Feigin, and Mandelbaum (2009);

* Corresponding author.
E-mail addresses: rouba.ibrahim@ucl.ac.uk (R. Ibrahim), hanye@illinois.edu (H. Ye), lecuyer@iro.umontreal.ca (P. L'Ecuyer), haipeng@email.unc.edu (H. Shen).

http://dx.doi.org/10.1016/j.ijforecast.2015.11.012

Avramidis, Deslauriers, and L'Ecuyer (2004); Ding and Koole (2015) and Ibrahim and L'Ecuyer (2013). For another, call center arrivals also exhibit different types of dependencies, including intraday (within-day), interday, and inter-type dependence, e.g., see Aldor-Noiman et al. (2009); Avramidis et al. (2004); Channouf and L'Ecuyer (2012); Shen and Huang (2008b); Tanir and Booth (1999) and Whitt (1999b). A reasonable forecasting model needs to account appropriately for some or all of the types of dependencies that exist in real data.

In the presence of intraday and interday dependence in call arrival rates, standard time series models may be applied for forecasting call arrivals, for example *autoregressive integrated moving average* (ARIMA) models and *exponential smoothing* (Hyndman, Koehler, Ord, & Snyder, 2008). In addition, some recent papers have proposed *fixed-effects* models (Ibrahim & L'Ecuyer, 2013; Shen & Huang, 2008b; Taylor, 2008; Weinberg, Brown, & Stroud, 2007) and *mixed-effects* models (Aldor-Noiman et al., 2009; Ibrahim & L'Ecuyer, 2013) to account for the within-day dependence, interday dependence, and inter-type dependence of call arrivals. Dimension reduction (Shen & Huang, 2005, 2008a,b) and Bayesian techniques (Aktekin & Soyer, 2011; Soyer & Tarimcilar, 2008; Weinberg et al., 2007) have also been adopted in the literature.

The remainder of the paper is organized as follows. The key features of call center arrival processes are discussed in Section 2, and various forecasting methods that have been proposed in the literature are examined in Section 3. A case study in which several methods from the recent literature are compared is reported in Section 4, using a Canadian call center data set, which reveals the practical features of those methods. Discussions of future research directions are provided in Section 5 to conclude the paper. The conference paper by Ibrahim, L'Ecuyer, Régnard, and Shen (2012) served as a starting point for this survey.

2. Key properties of call center arrival processes

A natural model to use for call arrivals is the Poisson arrival process (Akşin et al., 2007; Cezik & L'Ecuyer, 2008; Gans et al., 2003; Garnett et al., 2002; Wallace & Whitt, 2005). This model is justified theoretically by assuming a large population of potential customers where each customer makes calls independently with a very small probability; the total number of calls made in a given time interval is then approximately Poisson. As was mentioned by Kim and Whitt (2014a), the so-called Poisson superposition theorem is a supporting limit theorem, e.g., see Barbour, Holst, and Janson (1992).

Recent empirical studies have shown multiple important properties of the call arrival process, many of which are not consistent with the Poisson modeling assumption. This section describes these properties in detail; for a more abridged description, see Section 2 of Ibrahim et al. (2012).

2.1. Time dependence of call arrival rates

One of the most important properties of call arrival rates is that they vary with time. In particular, call arrival rates typically exhibit intraday (within-day), daily,

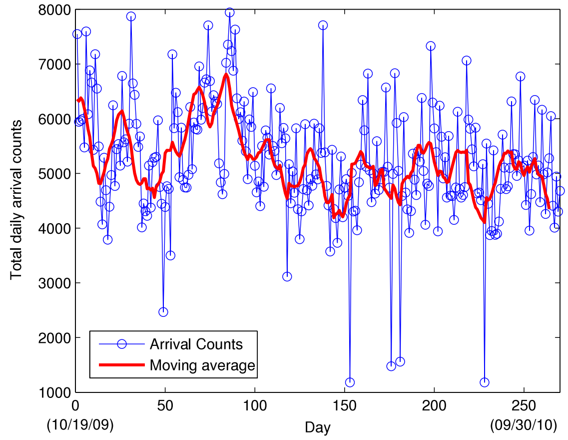

Fig. 1. Daily call arrival counts over successive months in a Canadian call center.

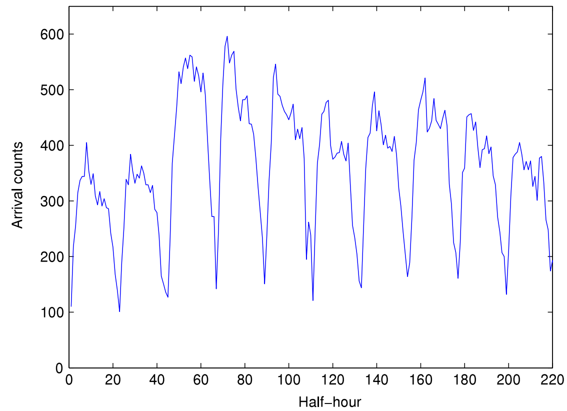

Fig. 2. Half-hourly call arrival counts over two consecutive weeks.

weekly, monthly, and yearly seasonalities. We illustrate this time-dependence property in Figs. 1, 2, and 3 (taken from Ibrahim & L'Ecuyer, 2013), which show arrival patterns that are observed commonly in call centers.

In Fig. 1, we plot the numbers of calls per day arriving at the call center of a Canadian company between October 19, 2009, and September 30, 2010. Fig. 1 shows that there are monthly fluctuations in the data. For example, the moving average line in the plot, which is computed for each day as the average of the past 10 days, suggests that there is an increase in call volume during the months of January and February, i.e., days 54–93 in the plot.

In Fig. 2, we illustrate weekly seasonality by plotting daily arrival counts, of the same call type as in Fig. 1, over two consecutive weeks in the call center. The call center is closed on weekends, so we have a total of 10 workdays in the plot. Fig. 2 clearly shows that there is a strong weekly seasonality in the data. Such weekly patterns are observed very commonly in practice, e.g., see Figure 1 of Taylor (2008) and Figure 2 of Taylor (2012).

For a more microscopic view of arrivals, we plot half-hourly average arrival counts per weekday, in Fig. 3. These intraday averages constitute the *daily profile* of call arrivals. Fig. 3 shows that call volumes are higher, on average, on Mondays than on the remaining weekdays. Fig. 3 also

Please cite this article in press as: Ibrahim, R., et al., Modeling and forecasting call center arrivals: A literature survey and a case study. International Journal of Forecasting (2016), http://dx.doi.org/10.1016/j.ijforecast.2015.11.012

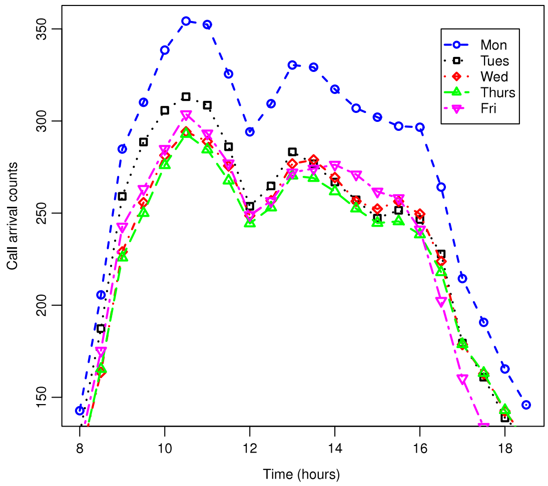

Fig. 3. Intraday profiles of call arrival counts by weekday in a Canadian call center.

shows that all weekdays have similar daily profiles, with two major daily peaks for call arrivals. The first peak occurs in the morning, shortly before 11:00 AM, and the second peak occurs in the early afternoon, around 1:30 PM. (There is also a third, smaller "peak" that occurs shortly before 4:00 PM on Mondays, Tuesdays, and Wednesdays.) Such intraday arrival patterns are also characteristic of call center arrivals; e.g., see Aldor-Noiman et al. (2009); Avramidis et al. (2004); Channouf, L'Ecuyer, Ingolfsson, and Avramidis (2007); Gans et al. (2003) and Tanir and Booth (1999).

Given that arrival rates are time-varying, a feature which is not accounted for in a Poisson arrival process, a natural extension is to consider a nonhomogeneous Poisson process with a deterministic and time-varying arrival-rate function. For simplicity, it is commonly assumed that call arrival rates are constant in consecutive 15 or 30 min intervals during a given day; e.g., see Brown et al. (2005); Green, Kolesar, and Whitt (2007), and Liao, Delft, Koole, and Jouini (2012).

Nevertheless, it is important to perform statistical tests to confirm that it is appropriate to model call center data as a nonhomogeneous Poisson process. Brown et al. (2005) proposed a specific test procedure based on a Kolmogorov–Smirnov test and did not reject the null hypothesis that arrivals of calls are from a nonhomogeneous Poisson process with piecewise constant rates. Kim and Whitt (2014b) examined several alternative test procedures which have greater power than that suggested by Brown et al. (2005). Kim and Whitt (2014a) applied Kolmogorov–Smirnov tests to banking call center and hospital emergency department arrival data and showed that they are consistent with the nonhomogeneous Poisson property, but only if certain common features of the data have been accounted for, including data rounding, interval partition, and overdispersion caused by combining data.

2.2. Overdispersion of arrival counts

One consequence of the Poisson modeling assumption is that the variance of the arrival count in each time period is equal to its expectation during that period. However, there is empirical evidence that invalidates this assumption. Indeed, it has been observed that the variance of an arrival count per time period is usually much larger than its expected value; see Aldor-Noiman et al. (2009); Avramidis et al. (2004); Jongbloed and Koole (2001), and Steckley, Henderson, and Mehrotra (2005). One way of dealing with this overdispersion of count data is to assume that the Poisson arrival process is doubly stochastic, i.e., that the arrival rate itself is a stochastic process; e.g., see Aldor-Noiman et al. (2009); Avramidis et al. (2004); Ibrahim and L'Ecuyer (2013); Jongbloed and Koole (2001); Shen (2010b); Soyer and Tarimcilar (2008); Steckley, Henderson, and Mehrotra (submitted for publication), and Weinberg et al. (2007).

A doubly stochastic Poisson process can be viewed as a two-step randomization: a stochastic process (for the arrival rate) is used to generate another stochastic process (for the call arrival process) by representing its intensity. We now illustrate why a doubly stochastic Poisson process is a way to deal with a higher variance in the arrival count data. Denote the number of arrivals in a given period j by X_j, and let Λ_j denote the cumulative arrival rate (its integral) over period j. Then, assume that, conditional on Λ_j, X_j has a Poisson distribution with mean Λ_j. To simplify the notation, this paper assumes that all periods have the same length and that the time unit is equal to one period. Then, when the arrival rate is constant over each period, this rate is the same as the cumulative rate Λ_j, and we denote both by Λ_j. By conditioning on Λ_j, the variance of X_j is given by:

$$\begin{aligned}\mathrm{Var}[X_j] &= \mathbb{E}[\mathrm{Var}[X_j|\Lambda_j]] + \mathrm{Var}[\mathbb{E}[X_j|\Lambda_j]]\\ &= \mathbb{E}[\Lambda_j] + \mathrm{Var}[\Lambda_j]. \end{aligned} \tag{1}$$

With a random arrival rate function, we have that $\mathrm{Var}[\Lambda_j] > 0$ on the right-hand side of Eq. (1), which accounts for the additional variance in $\mathrm{Var}[X_j]$.

Maman, Mandelbaum, Whitt, and Zeltyn (2015) study the implications for operational decision making in the system of assuming a doubly stochastic Poisson process. In particular, they proposed a Poisson mixture model with a parametric form of the random Poisson arrival rate for modeling the doubly stochastic Poisson process. They then incorporated the Poisson mixture model into a queuing model and derived asymptotic optimal staffing levels.

Via a statistical analysis, Zhang, Hong, and Zhang (2014) show that the level of stochastic variation of the arrival process is neither as low as in a standard Poisson process, nor as high as in a doubly stochastic process. They therefore propose a model to control for this level of overdispersion. Glynn, Hong, and Zhang (in preparation) show that the timescale is the key to the different conclusions in the literature regarding the Poissonness of the arrivals. In particular, they show that the arrival process is Poisson-like at short timescales (minutes), but not at longer timescales (hours, days, etc.). The effect of the timescale is also examined extensively by Oreshkin, Regnard, and L'Ecuyer (2016).

Table 1
Correlations between arrival counts on successive weekdays in a Canadian call center.

Weekday	Mon.	Tues.	Wed.	Thurs.	Fri.
Mon.	1.0	0.48	0.35	0.35	0.34
Tues.		1.0	0.68	0.62	0.62
Wed.			1.0	0.72	0.67
Thurs.				1.0	0.80
Fri.					1.0

2.3. *Interday and intraday dependencies*

In real-life call centers, there is typically evidence of dependencies between the arrival counts, or arrival rates, in different time periods within a single day, or across several days; e.g., see Aldor-Noiman et al. (2009); Avramidis et al. (2004); Channouf and L'Ecuyer (2012); Shen and Huang (2008b); Tanir and Booth (1999), and Whitt (1999b). These interday (day-to-day) and intraday dependencies typically remain strong even after correcting for detectable seasonalities. Indeed, such corrections are important in order to avoid erroneous overestimates of the dependencies in the data.

In Tables 1 and 2, we illustrate the interday and intraday correlations in the same call center as in Figs. 1–3. Tables 1 and 2 illustrate several properties which are observed very commonly in practice: (i) the correlations between successive weekdays are strong and positive; (ii) interday correlations are slightly smaller, with longer lags; (iii) Mondays are less correlated with the remaining weekdays; (iv) correlations between successive half-hourly periods within a day are strong and positive; and (v) intraday correlations are slightly smaller, with longer lags.

There are various different measures that could be used to capture interday and intraday dependencies in call arrival data. The most commonly used measure is Pearson's correlation coefficient, which captures linear dependence in the data; e.g., see Aldor-Noiman et al. (2009); Avramidis et al. (2004); Ibrahim and L'Ecuyer (2013), and Shen and Huang (2008b). However, since dependencies may not be linear, it is also useful to consider alternative measures such as rank correlation coefficients; see Channouf and L'Ecuyer (2012) and references therein. For example, Spearman's rank correlation coefficient measures how well the relationship between two variables can be described using a monotonic, but not necessarily linear, function.

Mixed-effects models (Aldor-Noiman et al., 2009; Ibrahim & L'Ecuyer, 2013), and copulas (Channouf & L'Ecuyer, 2012; Jaoua, L'Ecuyer, & Delorme, 2013) more generally, are ideally suited to the easy capture of interday and intraday dependencies in call center arrival data. Models that fail to account for positive interday and intraday dependencies in call arrivals may provide an overoptimistic view of call center performance measures, and the resulting errors can be very significant; see Avramidis et al. (2004); Avramidis and L'Ecuyer (2005); Steckley et al. (2005), and Steckley, Henderson, and Mehrotra (2009).

2.4. *Inter-type dependencies*

In multi-skill call centers, there may be positive dependencies between the arrival counts, or arrival rates, corresponding to different call types. As one example, this could occur in multilingual call centers where the same service request is handled in two or more languages. As another example, this may be due to promotions or advertisements which affect several services offered by the same call center. Neglecting the dependencies between different call types can lead to overloads, particularly when a single agent handles several correlated call types.

In Table 3 (taken from Ibrahim & L'Ecuyer, 2013), we present estimates of the correlations between half-hourly arrival counts for two different call types, Type A and Type B. In Table 3, we focus on the same consecutive half-hour periods as in Table 2. Table 3 shows that inter-type correlations can be strong and positive. Here, call arrivals to the Type A queue originate in the province of Ontario, and are handled mainly in English, whereas arrivals to the Type B queue originate in the province of Quebec, and are handled mainly in French. Otherwise, arrivals to the two queues have similar service requests. Thus, it is reasonable to expect there to be correlations between their respective arrival processes. There have been various attempts recently to model inter-type dependencies in the data; see Ibrahim and L'Ecuyer (2013) and Jaoua et al. (2013).

2.5. *Using auxiliary information*

Auxiliary information is often available in call centers, and can improve point or distributional forecasts considerably. For example, when a company sends notification letters to customers, or runs advertisements, this may trigger a large volume of calls; see Landon, Ruggeri, Soyer, and Tarimcilar (2010). Also, large sporting events or festivals can result in a significant increase in calls to emergency systems; see Channouf et al. (2007).

Past service levels in the call center may also be a valuable source of information for predicting future arrivals. For example, long previous delays may lead to high call abandonment rates, which in turn may lead to more redials in the future. Moreover, when the quality of service is poor, callers may not have their problems resolved during their first call, and may need to reconnect later. Ignoring such redials and reconnects may lead to a considerable underestimation of call arrival counts; see Ding, Koole, and Mei (2015).

Finally, in certain types of call centers, for example ones where people may call to report power outages or those designated for emergency services, bursts of high arrival rates over short periods of time do occur. In this context, an important accident may trigger several dozen different calls within a few minutes, all related to the same event, resulting in a much larger than expected number of calls during that time frame; e.g., see Kim, Kenkel, and Brorsen (2012) for the modeling of peak periods in a rural electric cooperative call center.

In recent years, there have been a few studies on forecasting call arrivals, and we review the relevant literature in the next section.

Please cite this article in press as: Ibrahim, R., et al., Modeling and forecasting call center arrivals: A literature survey and a case study. International Journal of Forecasting (2016), http://dx.doi.org/10.1016/j.ijforecast.2015.11.012

Table 2
Correlations between arrivals in consecutive half-hour periods on Wednesday mornings in a Canadian call center.

Half-hour periods	(10, 10:30)	(10:30, 11)	(11, 11:30)	(11:30, 12)	(12, 12:30)
(10, 10:30)	1.0	0.87	0.80	0.73	0.66
(10:30, 11)		1.0	0.82	0.74	0.71
(11, 11:30)			1.0	0.83	0.80
(11:30, 12)				1.0	0.81
(12, 12:30)					1.0

Table 3
Correlations between Type A and Type B arrivals in consecutive half-hour periods on Wednesdays in a Canadian call center.

Type B	Type A				
	(10, 10:30)	(10:30, 11)	(11, 11:30)	(11:30, 12)	(12, 12:30)
(10, 10:30)	0.75	0.72	0.67	0.60	0.59
(10:30, 11)	0.76	0.73	0.72	0.64	0.62
(11, 11:30)	0.66	0.65	0.67	0.67	0.63
(11:30, 12)	0.60	0.56	0.63	0.63	0.63
(12, 12:30)	0.58	0.54	0.58	0.65	0.62

3. Call forecasting approaches

In practice, arrival forecasts are needed for a range of purposes, such as long-term capacity planning and short-term scheduling. Therefore, the specifics of the forecasting procedure need to be determined carefully, including the forecasting horizon (for time intervals, for a day, or for multiple days), and whether to combine arrivals from separate queues. Typically, models that incorporate time dependencies are useful for short-term decision making, since such dependencies tend to vanish over longer time scales. Over longer time horizons, simple smoothing or moving average models seem to be sufficient to capture the general trends in the data. Indeed, in practice, arrivals to different call types are often combined when there is not sufficient data for the individual call types. However, typically, this is not done systematically, and is based solely on the experience of the call center managers.

Ideally, we want arrival models that seek to reconcile several objectives. In order for an arrival model to be realistic, it needs to reproduce the properties that we described in Section 2. Simultaneously, in order for an arrival model to be of practical use, it needs to be computationally tractable; that is, it needs to rely on a relatively small number of parameters so as to avoid overfitting. Moreover, these parameters need to be easy to estimate from historical data. Finally, parameter estimates should not be hard to update (e.g., via Bayesian methods) based on newly available information, e.g., throughout the course of a day. These updated estimates would then be used to update operational decisions in the call center.

In this section, we review alternative models that have been proposed in the literature with the aim of reconciling these objectives. We begin by reviewing early papers, which relied mostly on standard forecasting methods (Section 3.1). Next, we focus on more recent models for arrivals over several days or months (Section 3.2). Finally, we move to models for arrivals over a single day (Section 3.3).

3.1. *Standard forecasting techniques*

The early work on the forecasting of call arrivals usually focused on modeling daily or even monthly total call volumes, due partly to the lack of relevant data. In addition, only point forecasts of future arrival rates or counts were produced.

One of the earliest papers on the forecasting of call arrivals was that by Thompson and Tiao (1971), who modeled monthly call arrivals for two different call types. Interestingly, they noted that there may be an interdependence between the arrival streams of these two call types, but they did not explore this issue further. They used seasonal *autoregressive integrated moving average* (ARIMA) models to forecast future call volumes, and relied solely on the past history of call arrivals in their models.

Mabert (1985) relied on multiplicative and additive regression models, including covariates for special events and different seasonalities, for forecasting daily call arrivals to an emergency call center. He also considered model adjustments which exploit previous forecasting errors in order to yield more accurate forecasts. He found that such models yielded the most accurate forecasts, and were superior to ARIMA models.

Other early papers also relied on standard time series models. For example, Andrews and Cunningham (1995) modeled daily call arrivals to the call center of a retailer. The authors considered ARIMA models with transfer functions and incorporated covariates for advertising and special-day effects. They showed that using such information improved the accuracy of their forecasts dramatically, and could have a significant impact on operational decision-making in the call center. Similarly, Bianchi, Jarrett, and Hanumara (1998) used ARIMA models with intervention analysis to forecast telemarketing call arrivals, and found that such models are superior to additive and multiplicative Holt–Winters' exponentially weighted moving average models.

More recently, Antipov and Maede (2002) modeled the daily numbers of applications for loans at a financial services telephone call center. The authors also went beyond standard ARIMA models by including advertising responses and special calender effects, through the addition of exogenous variables to a multiplicative model. Channouf et al. (2007) developed simple additive models for the (small) number of ambulance calls each hour in the city of Calgary. Their models capture daily, weekly, and yearly seasonalities, selected second-order interaction effects (e.g., between the time-of-day and day-of-the-week), special-day effects (such as the Calgary Stampede, leading to increased call volumes), and autocorrelation of the residuals between successive hours. Their best model outperformed a doubly-seasonal ARIMA model for the residuals of a model which captures only special-day effects.

The existing literature has used several different measures of forecasting errors to evaluate the accuracy of forecasting models, including the *root mean squared error* (RMSE), *mean squared error* (MSE), *mean absolute percentage error* (MAPE), *weighted absolute percentage error* (WAPE), etc. The answer to the question of which error measure is appropriate depends on the objective of the call center management. For example, Ding and Koole (2015) studied a call center staffing problem where the costs are the initial staffing costs plus the intraday traffic management costs, and concluded that optimal forecast methods should be those that minimize the WAPE.

3.2. Models over several days

To describe more recent arrival modelling approaches, we need some additional notation. Let $X_{i,j}$ denote the number of call arrivals during period j, $1 \le j \le P$, of day i, $1 \le i \le D$. The standard assumption is that call arrivals follow a Poisson process with a (potentially) random arrival rate $\Lambda_{i,j}$, which is taken to be constant over each period j. The cumulative arrival rate over period j is also $\Lambda_{i,j}$ if a unit time period is assumed. Thus, conditional on the event $\Lambda_{i,j} = \lambda_{i,j}$, $X_{i,j}$ is Poisson distributed with rate $\lambda_{i,j}$.

Several papers (e.g., those of Aldor-Noiman et al., 2009; Brown et al., 2005; Ibrahim & L'Ecuyer, 2013; Weinberg et al., 2007) exploit the following "root-unroot" variance-stabilizing data transformation:

$$Y_{i,j} \equiv (X_{i,j} + 1/4)^{1/2}. \tag{2}$$

Conditional on the event $\Lambda_{i,j} = \lambda_{i,j}$, and for large values of $\lambda_{i,j}$, $Y_{i,j}$ is distributed approximately normally, with mean $\sqrt{\lambda_{i,j}}$ and variance 1/4; see Brown, Cai, Zhang, Zhao, and Zhou (2010). The unconditional distribution, with random $\Lambda_{i,j}$, is then a mixture of such normal distributions, and therefore has a larger variance. Nevertheless, it can be assumed (as an approximation) that the square-root transformed counts $Y_{i,j}$ are distributed normally, particularly if $\mathrm{Var}[\Lambda_{i,j}]$ is not too large. The resulting normality is very useful because it allows linear Gaussian fixed-effects and mixed-effects models to be fitted to the square-root transformed data.

A better alternative than modelling the arrival counts $X_{i,j}$ would be to model the rates $\Lambda_{i,j}$ directly, because it is considerably easier to simulate the system with a distributional forecast for the rates than one for the counts. Indeed, to simulate arrivals based on a distributional forecast for counts, one has to generate the number of arrivals in each period, and then generate the arrival times by splitting the counts uniformly and independently over the given time period. (This is consistent with the Poisson assumption.) In contrast, given a distributional forecast for the rates, one can generate the arrival times directly. Nevertheless, most arrival models in the literature are for the counts $X_{i,j}$, rather than the rates $\Lambda_{i,j}$, because, in practice, we observe, not the arrival rates themselves, but only the counts, which give only partial information as to the rates. This makes the estimation of arrival rates a more complicated task.

Multiple papers, such as those of Ibrahim and L'Ecuyer (2013); Shen and Huang (2008b); Taylor (2008), and Weinberg et al. (2007), consider a *linear fixed-effects* (FE) model as a benchmark for comparison. To illustrate, let d_i be the day-of-the-week of day i, where $i = 1, 2, \ldots, D$. That is, $d_i \in \{1, 2, 3, 4, 5\}$, where $d_i = 1$ denotes a Monday, $d_i = 2$ denotes a Tuesday,..., and $d_i = 5$ denotes a Friday. Ibrahim and L'Ecuyer (2013) considered the following fixed-effects model for the square-root transformed arrival counts:

$$y_{i,j} = \alpha_{d_i} + \beta_j + \theta_{d_i,j} + \mu_{i,j}, \tag{3}$$

where the coefficients α_{d_i}, β_j, and $\theta_{d_i,j}$ are real-valued constants that need to be estimated from the data, and $\mu_{i,j}$ are independent and identically distributed (i.i.d.) normal random variables with a mean of zero. Ibrahim and L'Ecuyer (2013) and Taylor (2008) have shown that it is difficult to beat fixed-effect models in terms of the accuracy of long-term point forecasting (e.g., two weeks or more). Nevertheless, for short-term forecasting, one can exploit interday and intraday dependencies in the data in order to obtain more accurate forecasts.

As an improvement, and based on an analysis of real call center data, Aldor-Noiman et al. (2009) proposed the following *linear mixed-effects* (ME) model:

$$Y_{i,j} = \alpha_{d_i} + \beta_j + \theta_{d_i,j} + \gamma_i + \epsilon_{i,j},$$

where γ_i denotes the daily volume deviation from the fixed weekday effect on day i. Then, γ_i is the random effect on day i. Let G denote the $D \times D$ covariance matrix for the sequence of random effects. The random effects γ_i are identically normally distributed, with expected value $E[\gamma] = 0$ and variance $\mathrm{Var}[\gamma] = \sigma_G^2$. The authors assume that these random effects follow an AR(1) process. Considering an AR(1) covariance structure for G is both useful and computationally effective, because it requires the estimation of only two parameters, σ_G and ρ_G. The residuals $\epsilon_{i,j}$ are also assumed to have an AR(1) structure within each day. As such, this model captures both interday and intraday dependencies in the data. An earlier version of this model, also based on call-center data, but without intraday correlations and without special-day effects, was proposed by Brown et al. (2005).

Ibrahim and L'Ecuyer (2013) extended this ME model to two bivariate ME models for the joint distribution of the arrival counts to two separate queues, which exploit correlations between different call types. These models account for the dependence between the two call types by assuming that either the vectors of random effects or the vectors of residuals across call types are correlated multinormal. This corresponds to using a normal copula;

Please cite this article in press as: Ibrahim, R., et al., Modeling and forecasting call center arrivals: A literature survey and a case study. International Journal of Forecasting (2016), http://dx.doi.org/10.1016/j.ijforecast.2015.11.012

see Kim et al. (2012). The choice of copula can have a significant impact on performance measures in call centers, because of the strong effect of tail dependence on the quality of service (Jaoua et al., 2013). For example, a strong upper tail dependence for certain call types means that very large call volumes tend to arrive together for these call types. When this happens, this produces very large overloads.

To reduce the dimensionality of the vectors $(Y_{i,1}, \dots, Y_{i,P})$, Shen and Huang (2005) proposed the use of singular-value decomposition to define a small number of vectors whose linear transformations capture most of the information that is of relevance for prediction. Based on this, Shen and Huang (2008b) then developed a dynamic updating method for the distributional forecasts of arrival rates. Shen and Huang (2008a) proposed a method for forecasting the latent rate profiles of a time series of inhomogeneous Poisson processes, to enable future arrival rates to be forecast based on a series of observed arrival counts.

Aktekin and Soyer (2011) recently proposed a model based on a Poisson-Gamma process, where $\Lambda_{i,j} = W_{i,j}\lambda_{i,j}$ for fixed values of $\lambda_{i,j}$, and where the multiplicative factors $W_{i,j}$ have a gamma distribution and obey a gamma process. Soyer and Tarimcilar (2008) analyzed the effects of advertisement campaigns on call arrivals, using a Bayesian analysis where the Poisson rate function is modeled using a mixed model approach. This mixed model is shown to be superior to using a fixed-effects model. Weinberg et al. (2007) propose an adaptation of the model of Brown et al. (2005) to enable it to update the forecasts of a day, as defined from the previous days, using observations made available during this day.

Weinberg et al. (2007) also used Bayesian techniques in their forecasts. They exploited the (normal) square-root transformed counts to include conjugate multivariate normal priors, with specific covariance structures. They used Gibbs sampling and the Metropolis–Hastings algorithm to sample from the forecast distributions, which unfortunately involves long computation times. Moreover, it is unclear how exogenous covariates should be incorporated in such a model.

The empirical analysis of Taylor (2008) compared several time series models, including autoregressive moving average (ARMA) models and Holt–Winters' exponential smoothing models with multiple seasonal patterns. The latter method was adapted by Taylor (2003) for modeling both the intraday and intraweek cycles in intraday data. Taylor (2012) extended this model and considered the density forecasting of call arrival rates. With this aim, he developed a new Holt–Winters' Poisson count data model with a gamma-distributed stochastic arrival rate, and showed that this new model outperformed the basic Holt–Winters' smoothing model. Shen (2010a) commented on Taylor's work, highlighting the difference between modeling arrivals as a single time series and as a vector time series where each day is modeled as a component of that vector.

3.3. Models over a single day

In this section, we focus on modeling arrivals over a single day. The day is divided into p time periods. We denote the vector of arrival counts in those periods by $\mathbf{X} = (X_1, \dots, X_p)$.

It is commonly assumed that intraday arrivals follow a Poisson process with a random arrival rate. Whitt (1999a) proposed that this be done by starting with a deterministic arrival rate function $\{\lambda(t),\ t_0 \le t \le t_e\}$, where t_0 and t_e are the opening and closing times of the call center for the considered day, and multiplying this function by a random variable W with mean $\mathbb{E}[W] = 1$, called the *busyness factor* for that day. The (random) arrival rate process for that day is then $\Lambda = \{\Lambda(t) = W\lambda(t),\ t_0 \le t \le t_e\}$.

Under this model, the arrival rates at any two given times are perfectly correlated, and $\mathrm{Corr}[\Lambda_j, \Lambda_k] = 1$ for all j, k. We also expect the X_js to be correlated strongly. More specifically, let I_j denote the time interval of period j, let $\bar{\lambda}_j = \int_{I_j} \lambda(t)\mathrm{d}t$, and let X_j be the number of arrivals in I_j. Using the variance and expectation decompositions, one can find that $\mathrm{Var}[X_j] = \bar{\lambda}_j(1 + (1 + \bar{\lambda}_j)\mathrm{Var}[W])$, i.e. Eq. (3.12) of Whitt (1999a), and for $j \neq k$:

$$\mathrm{Corr}[X_j, X_k] = \mathrm{Var}[W]\left[(\mathrm{Var}[W] + 1/\bar{\lambda}_j)(\mathrm{Var}[W] + 1/\bar{\lambda}_k)\right]^{-1/2}.$$

This correlation is zero when $\mathrm{Var}[W] = 0$ (a deterministic rate), and approaches one when $\mathrm{Var}[W] \to \infty$. Avramidis et al. (2004) studied this model in the special situation where W has a gamma distribution, with $\mathbb{E}[W] = 1$ and $\mathrm{Var}[W] = 1/\gamma$. Then, each Λ_j has a gamma distribution, and the X_js have a negative multinomial distribution, with parameters that are easy to estimate. Furthermore, the variance of the arrival counts can be made arbitrarily large by decreasing γ toward zero. The model's flexibility is rather limited, because, given the $\bar{\lambda}_j$s, $\mathrm{Var}[X_j]$ and $\mathrm{Corr}[X_j, X_k]$ for $j \neq k$ are all determined by a single parameter value, namely $\mathrm{Var}[W]$. In an attempt to increase the flexibility of the covariance matrix $\mathrm{Cov}[\mathbf{X}]$, and in particular to enable a reduction of the correlations, Avramidis et al. (2004) introduced two different models for $\mathbf{X}$, based on the multivariate Dirichlet distribution.

Jongbloed and Koole (2001) examined a similar model, but with independent busyness factors, one for each period of the day. Under their model, the Λ_js are independent, as are the X_js, which is inconsistent with an intraday dependence of call center arrivals. Channouf (2008) considered a variant of the model where $\lambda(t)$ is defined by a cubic spline over the day, with a fixed set of knots, and also shows how the model parameters can be estimated. This can provide a smoother (and perhaps more realistic) model of the arrival rate. Channouf (2008) and Channouf and L'Ecuyer (2012) proposed models that account for time dependence, overdispersion, and intraday dependencies with much more flexibility, in order to match the correlations between the X_js, by using a normal copula to specify the dependence structure between these counts. In principle, similar copula models could be developed for the vector of arrival rates, $(\Lambda_1, \dots, \Lambda_p)$, instead of for the vector of counts. Oreshkin et al. (2016) examined the relationship

Table 4
Forecasting comparison among the five methods for call type A.

Type A					
		MU	ME	BME1	BME2
One-day-ahead forecast	RMSE	23.51	21.76	22.59	22.69
	Cover	0.89	0.91	0.93	0.93
One-week-ahead forecast	RMSE	30.63	29.59	31.37	31.10
	Cover	0.88	0.88	0.88	0.86
Two-week-ahead forecast	RMSE	37.64	37.51	38.04	37.07
	Cover	0.84	0.82	0.80	0.79

Table 5
Forecasting comparison among the five methods for call type B.

Type B					
		MU	ME	BME1	BME2
One-day-ahead forecast	RMSE	16.80	16.31	16.46	16.49
	Cover	0.93	0.92	0.95	0.95
One-week-ahead forecast	RMSE	18.55	17.95	17.99	18.12
	Cover	0.95	0.91	0.94	0.94
Two-week-ahead forecast	RMSE	21.05	20.55	20.81	20.75
	Cover	0.93	0.86	0.90	0.90

between modeling for the vector of counts and the vector of rates. In particular, they gave explicit formulas for the relationship between the correlations between the rates and the counts in two given periods, which implied that, for a given correlation between rates, the correlation between counts is much smaller in low traffic than in high traffic.

4. Case studies

In this section, we present empirical results from a case study using real data collected at a Canadian call center, as described in Section 2. We use data based on two call types and 200 consecutive workdays (excluding weekends). For each call type, the study implements four methods (those of Aldor-Noiman et al., 2009; Gans et al., 2015; Ibrahim & L'Ecuyer, 2013) for forecasting arrival counts based on six weeks of historical data:

- MU: the multiplicative univariate forecasting model of Gans et al. (2015);
- ME: the univariate mixed-effects model of Aldor-Noiman et al. (2009);
- BME1: the bivariate mixed-effects model of Ibrahim and L'Ecuyer (2013); and
- BME2: the bivariate mixed-effects model of Ibrahim and L'Ecuyer (2013).

Each method is applied to the data in turn, and we assess the out-of-sample forecasting accuracy of each. We compare the different models by using the root mean squared error (RMSE) to assess the point forecast accuracy, as defined below:

$$\text{RMSE} = \sqrt{\frac{1}{K}\sum_{i,j}(X_{i,j} - \hat{X}_{i,j})^2},$$

where $\hat{X}_{i,j}$ is the value of $X_{i,j}$ predicted by the model, and K is the total number of predictions. We also evaluate the forecasting distribution by reporting the coverage probability for the 95% prediction interval, defined as:

$$\text{Cover} = \frac{1}{K}\sum_{i,j}\mathbb{I}(X_{i,j} \in (\hat{L}_{i,j}, \hat{U}_{i,j})),$$

where $(\hat{L}_{i,j}, \hat{U}_{i,j})$ is the 95% prediction interval for $X_{i,j}$ given by the model.

Tables 4 and 5 summarize the comparisons among the four methods. For both call types, ME produces the most accurate point forecasts in most scenarios. BME1 and BME2 have better coverage probabilities when the leading period is one day or one week, and MU has a better coverage probability when the leading period is two weeks.

Our numerical results serve to illustrate the complexity of the forecasting problem. Indeed, different models may be most appropriate depending on the lead-time of interest and on the specific criterion being considered, as is shown here.

Ultimately, selecting the "best" forecasting method depends on the specifics of the problem at hand, and an adequate measurement error. Proposing alternative arrival models that capture different features of the data is key to gaining a deeper understanding of the complexities of the call-arrival process.

5. Conclusions and discussion

The forecasting of call center arrivals plays a crucial role in call center management, for example in determining appropriate staffing levels, scheduling plans and routing policies. The call center arrival process is complex and requires appropriate modeling in order to achieve better forecasting accuracies, leading to more efficient operational decisions.

In this survey paper, we have reviewed the existing literature on the modeling and forecasting of call center arrivals. We have also conducted a case study to evaluate several recently proposed forecasting methods using real-life call center data.

An interesting future research direction would be to extend the existing forecasting models or develop new models so as to forecast more than two call types simultaneously. As some stochastic optimization models for staffing and scheduling rely on the joint forecasting distribution of multiple types of arrivals, such multi-type forecasting models with full distributional forecasts have the potential to meet the quality of service levels better, and improve the operational efficiency.

Another research issue that would be worth pursuing is an examination of the operational impact of improved forecasts, since most of the existing studies of call center forecasting evaluate forecasting approaches based only on traditional statistical measures, such as the RMSE and the coverage probability, without looking at how those improved forecasting models affect call center operations. By looking at the operational effect of forecasting models, managers can obtain additional insights regarding forecasting model selection and system performance evaluation. Gans et al. (2015) have tried to tackle this problem for one call type, but further research in this direction is needed.

In addition, different objectives of call center management may require the minimization of different measures of forecasting errors. The question of which is the most appropriate forecasting error measurement to choose and how it relates to call center decision making has not been investigated extensively. Ding and Koole (2015) show the optimal error measurement for minimizing the initial staffing costs plus the traffic management costs. However, the issue of optimal error measurement for other managerial objectives remains unclear and requires study.

Although there has been a lot of progress made in the developmnt of sophisticated forecasting methods, there remains a large gap between academic research and current industry practice, as was discussed by Koole (2013). For example, in practice, most call center forecasting is done using Excel, by implementing simple decomposition-based approaches; most *workforce management* (WFM) tools focus on scheduling, with a limited forecasting functionality. The incorporation of advanced forecasting methods in such WFM tools remains critical for call center practice.

We also want to point out that, besides call arrivals, there are various other factors that require forecasting for efficient call center management, such as the average handling time (AHT), workload, and even absenteeism. The relevant literature on this point is rather sparse. For example, Aldor-Noiman et al. (2009) used mixed effects models to forecast workloads, while Gans, Liu, Mandelbaum, Shen, and Ye (2010) and Ibrahim, L'Ecuyer, Shen, and Thiongane (in press) provided initial attempts to identify factors that affect agent productivity, which can help with predicting the AHT.

Acknowledgments

We thank two anonymous referees and the editor for many helpful comments and suggestions. This work has been supported in part by grants from NSERC-Canada and Hydro-Quebec, and a Canada Research Chair, to P. L'Ecuyer, and from US-NSF (DMS-1106912, DMS-1407655), The Xerox Foundation, and HKU Stanley Ho Alumni Challenge Fund to H. Shen.

References

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

Aktekin, T., & Soyer, R. (2011). Call center arrival modeling: A Bayesian state space approach. *Naval Research Logistics*, *58*(1), 28–42.

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

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

Antipov, A., & Maede, N. (2002). Forecasting call frequency at a financial services call centre. *Journal of the Operational Research Society*, *53*, 953–960.

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

Avramidis, A. N., & L'Ecuyer, P. (2005). Modeling and simulation of call centers. In M. E. Kuhl, N. M. Steiger, F. B. Armstrong, & J. A. Joines (Eds.), *Proceedings of the 2005 winter simulation conference* (pp. 144–152). IEEE Press.

Barbour, A. D., Holst, L., & Janson, S. (1992). *Poisson approximation*. Oxford, U.K: Oxford University Press.

Bianchi, L., Jarrett, J., & Hanumara, R. C. (1998). Improving forecasting for telemarketing centers by ARIMA modeling with intervention. *International Journal of Forecasting*, *14*, 497–504.

Brown, L., Cai, T., Zhang, R., Zhao, L., & Zhou, H. (2010). The root-unroot algorithm for density estimation as implemented via wavelet block thresholding. *Probability Theory and Related Fields*, *146*, 401–433.

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

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

Channouf, N. (2008). *Modélisation et optimisation d'un centre d'appels téléphoniques: étude du processus d'arrivée*. (Ph.D. thesis) Canada: Département d'Informatique et de Recherche Opérationnelle, Université de Montréal.

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

Channouf, N., L'Ecuyer, P., Ingolfsson, A., & Avramidis, A. N. (2007). The application of forecasting techniques to modeling emergency medical system calls in Calgary, Alberta. *Health Care Management Science*, *10*(1), 25–45.

Ding, S., & Koole, G. (2015). Optimal call center forecasting and staffing under arrival rate uncertainty. Working paper.

Ding, S., Koole, G., & Mei, R. V. D. (2015). On the estimation of the true demand in call centers with redials and reconnects. *European Joural of Operational Research*, *246*(1), 250–262.

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

Gans, N., Liu, N., Mandelbaum, A., Shen, H., & Ye, H. (2010). 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. Vol. 6. IMS Collections, Beachwood, OH, pp. 99–123.

Gans, N., Shen, H., Zhou, Y. P., Korolev, N., McCord, A., & Ristock, H. (2015). Parametric forecasting and stochastic programming models for call-center workforce scheduling. *Manufacturing and Service Operations Management*, *17*(4), 571–588.

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

Glynn, P., Hong, L.J., & Zhang, X. (2014). Modeling call center arrivals: a tale of three timescales. in preparation.

Green, L. V., Kolesar, P. J., & Whitt, W. (2007). Coping with time-varying demand when setting staffing requirements for service system. *Production and Operations Management*, *16*(1), 13–39.

Hyndman, R. J., Koehler, A. B., Ord, J. K., & Snyder, R. D. (2008). *Forecasting with exponential smoothing: the state space approach*. Springer.

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

Ibrahim, R., L'Ecuyer, P., Régnard, N., & Shen, H. (2012). On the modeling and forecasting of call center arrivals. In *Proceedings of the 2012 winter simulation conference* (pp. 1–12). IEEE Press.

Ibrahim, R., L'Ecuyer, P., Shen, H., & Thiongane, M. (2015). Interdependent, heterogeneous, and time-varying service-time distributions in call centers. *European Journal of Operational Research*, in press.
Jaoua, A., L'Ecuyer, P., & Delorme, L. (2013). Call type dependence in multiskill call centers. *Simulation*, *89*(6), 722–734.
Jongbloed, G., & Koole, G. (2001). Managing uncertainty in call centers using Poisson mixtures. *Applied Stochastic Models in Business and Industry*, *17*, 307–318.
Kim, T., Kenkel, P., & Brorsen, B. W. (2012). Forecasting hourly peak call volume for a rural electric cooperative call center. *Journal of Forecasting*, *31*, 314–329.
Kim, S. H., & Whitt, W. (2014a). Are call center and hospital arrivals well modeled by nonhomogeneous Poisson processes? *Manufacturing and Service Operations Management*, *16*(3), 464–480.
Kim, S. H., & Whitt, W. (2014b). Choosing arrival process models for service systems: Tests of a nonhomogeneous Poisson process. *Naval Research Logistics*, *61*(1), 66–90.
Koole, G. (2013). *Call center optimization*. MG Books.
Landon, J., Ruggeri, F., Soyer, R., & Tarimcilar, M. M. (2010). Modeling latent sources in call center arrival data. *European Journal of Operational Research*, *204*(3), 597–603.
Liao, S., Delft, C. V., Koole, G., & Jouini, O. (2012). Staffing a call center with uncertain non-stationary arrival rate and flexibility. *OR Spectrum*, *34*(3), 691–721.
Mabert, V. A. (1985). Short interval forecasting of emergency phone call (911) work loads. *Journal of Operations Management*, *5*(3), 259–271.
Maman, S., Mandelbaum, A., Whitt, W., & Zeltyn, S. (2015). Queues with random arrival rates: Inference, modelling and asymptotics (c-staffing). submitted for publication.
Oreshkin, B. N., Regnard, N., & L'Ecuyer, P. (2016). Rate-based daily arrival process models with application to call centers. *Operations Research*, in press.
Shen, H. (2010a). Exponentially weighted methods for forecasting intraday time series with multiple seasonal cycles: Comments. *International Journal of Forecasting*, *58*, 652–654.
Shen, H. (2010b). Statistical analysis of call-center operational data: Forecasting call arrivals, and analyzing customer patience and agent service. In J. J. Cochran (Ed.), *Wiley encyclopedia of operations research and management science*. John Wiley.
Shen, H., & Huang, J. Z. (2005). Analysis of call centre arrival data using singular value decomposition. *Applied Stochastic Models in Business and Industry*, *21*, 251–263.
Shen, H., & Huang, J. Z. (2008a). Forecasting time series of inhomogeneous Poisson processes with application to call center workforce management. *Annals of Applied Statistics*, *2*(2), 601–623.
Shen, H., & Huang, J. Z. (2008b). Interday forecasting and intraday updating of call center arrivals. *Manufacturing and Service Operations Management*, *10*(3), 391–410.
Soyer, R., & Tarimcilar, M. M. (2008). Modeling and analysis of call center arrival data: A Bayesian approach. *Management Science*, *54*(2), 266–278.
Steckley, S.G., Henderson, S.G., & Mehrotra, V. (2004). Service system planning in the presence of a random arrival rate. Submitted for publication.
Steckley, S. G., Henderson, S. G., & Mehrotra, V. (2005). Performance measures for service systems with a random arrival rate. In M. E. Kuhl, N. M. Steiger, F. B. Armstrong, & J. A. Joines (Eds.), *Proceedings of the 2005 winter simulation conference* (pp. 566–575). IEEE Press.
Steckley, S. G., Henderson, S. G., & Mehrotra, V. (2009). Forecast errors in service systems. *Probability in the Engineering and Informational Sciences*, *23*(2), 305–332.
Tanir, O., & Booth, R. J. (1999). Call center simulation in bell Canada. In *Proceedings of the 1999 winter simulation conference* (pp. 1640–1647). Piscataway, New Jersey: IEEE Press.
Taylor, J. W. (2003). Short-term electricity demand forecasting using double seasonal exponential smoothing. *Journal of Operational Research Society*, *54*, 799–805.
Taylor, J. W. (2008). A comparison of univariate time series methods for forecasting intraday arrivals at a call center. *Management Science*, *54*(2), 253–265.
Taylor, J. W. (2012). Density forecasting of intraday call center arrivals using models based on exponential smoothing. *Management Science*, *58*(3), 534–549.
Thompson, H. E., & Tiao, G. C. (1971). Analysis of telephone data: A case study of forecasting seasonal time series. *Bell Journal of Economics and Management Science*, *2*(2).
Wallace, R. B., & Whitt, W. (2005). A staffing algorithm for call centers with skill-based routing. *Manufacturing and Service Operations Management*, *7*(4), 276–294.
Weinberg, J., Brown, L. D., & Stroud, J. R. (2007). Bayesian forecasting of an inhomogeneous Poisson process with applications to call center data. *Journal of the American Statistical Association*, *102*(480), 1185–1198.
Whitt, W. (1999a). Dynamic staffing in a telephone call center aiming to immediately answer all calls. *Operations Research Letters*, *24*, 205–212.
Whitt, W. (1999b). Improving service by informing customers about anticipated delays. *Management Science*, *45*(2), 192–207.
Zhang, X., Hong, L. J., & Zhang, J. (2014). Scaling and modeling of call center arrivals. In *Proceedings of the 2014 winter simulation conference* (pp. 476–485). IEEE.

Please cite this article in press as: Ibrahim, R., et al., Modeling and forecasting call center arrivals: A literature survey and a case study. International Journal of Forecasting (2016), http://dx.doi.org/10.1016/j.ijforecast.2015.11.012