On the modeling and forecasting of call center arrivals

Rouba Ibrahim, Han Ye, Pierre L'Ecuyer, Haipeng Shen

MODELING AND FORECASTING CALL CENTER ARRIVALS: A LITERATURE SURVEY

by

Rouba Ibrahim
Management Science and Innovation
University College London
rouba.ibrahim@ucl.ac.uk

Han Ye
Business Administration
University of Illinois at Urbana Champaign
hanye@illinois.edu

Pierre L'Ecuyer
Computer Science and Operations Research
University of Montreal
lecuyer@iro.umontreal.ca

Haipeng Shen
Statistics and Operations Research
University of North Carolina at Chapel Hill
haipeng@email.unc.edu

Abstract

The effective management of call centers is a challenging task mainly because managers are consistently facing considerable uncertainty. Among important sources of uncertainty are call arrival rates which are typically time-varying, stochastic, dependent across time periods and across call types, and often affected by external events. Accurately modeling and forecasting 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 in building good statistical arrival models. Additionally, we evaluate the forecasting accuracy of selected models in an empirical study with real-life call center data. We conclude by summarizing future research directions in this important field.

1. Introduction

Call center is a large and important service industry with more than 2.7 million agents working in the United States and 2.1 million agents working at Europe, the Middle East, and Africa ([1]). Efficiently managing a call center is a challenging task because managers have to make staffing and scheduling decisions to balance between staffing cost and service quality, which always contradict, in the presence of uncertain arrival demands. Most staffing or scheduling plans start with forecasting customer call arrivals, which is highly stochastic. Accurately forecasting call arrivals is one of the keys to achieve optimal operational efficiency,

since under-forecasting leads to under-staffing and then long customer waiting, while on the other hand over-forecasting results in a waste of money on over-staffing.

The process of customer arrivals is nontrivial. This process can be modeled as a Poisson arrival process and has been shown to possess several features ([1, 13, 18, 20, 50]). One of the most important features is that the arrival rate is time varying, which adds complexity to the forecasting process. Call arrival rates may exhibit intra-day, weekly, monthly, and yearly seasonalities. Given the time-varying arrival rate, the doubly stochastic arrival process can be modeled as a non-homogeneous overdispersed Poisson process. Call center arrivals also show different types of dependencies including intra-day (within-day), inter-day, and inter-type dependence. A reasonable forecasting model needs to appropriately account for some of all types of dependencies that exist in real data.

In the presence of intra-day and inter-day dependence of call arrival rates, standard time series models may be applied to forecast call arrivals, for example Autoregressive Integrated Moving Average Models and Exponential Smoothing ([22]). In addition, some recent papers have proposed Fixed Effect Models and Mixed Effect Models to account for within-day dependence, inter-day dependence, and inter-type dependence of call arrivals. Dimension reduction or Bayesian techniques are also adopted in the existing literature. Detailed review of various existing forecasting approaches is given in Sections 3 and 4.

We then conduct a case study and implement several recently proposed forecasting approaches on a Canadian call center data set, which reveals the practical features of those approaches.

The remainder of the paper is organized as follows. In Section 2, we discuss the key features of call center arrival process. In Section 3, we review some relevant theories that can be applied to forecasting call center arrivals. In Section 4, we review forecasting methods that are proposed in the existing literature. We then conduct a case study to compare several models proposed in the recent literature in Section 5.

2. Key Properties of Call Center Arrival Processes

A natural model for call arrivals is the Poisson arrival process ([1, 13, 18, 20, 50]). This model is theoretically justified by assuming a large population of potential customers where each customer independently makes a call with a very small probability; the total number of calls made is then approximately Poisson. As mentioned in [27], the so-called Poisson superposition theorem is a supporting limit theorem, e.g., see [8].

Recent empirical studies have shown multiple important properties of the call arrival process, many of which are not consistent with the Poisson modeling assumption. In this section, we describe those properties in detail; for a more abridged description, see §2 in [24].

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, weekly, monthly, and yearly seasonalities. We illustrate this time-dependence property in Figures 1, 2, and 3 (taken from [23]), which show arrival patterns that are commonly observed in call centers.

In Figure 1, we plot the number of calls per day arriving to the call center of a Canadian company between October 19, 2009 and September 30, 2010. Figure 1 shows that there exist 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 to 93 in the plot.

In Figure 2, we illustrate weekly seasonality by plotting daily arrival counts, of the same call type as in Figure 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. Figure 2 clearly shows that there is a strong weekly seasonality in the data. Such weekly patterns are very commonly observed in practice, e.g., see Figure 1 in [47] and Figure 2 in [48].

For a more microscopic view of arrivals, we plot half-hourly average arrival counts, per weekday, in Figure 3. These intraday averages constitute the *daily profile* of call arrivals.

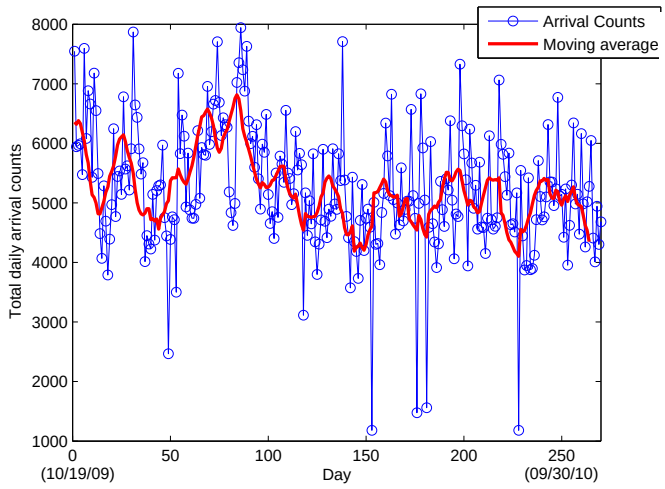

Figure 1: Daily call arrival counts over successive months in a Canadian call center.

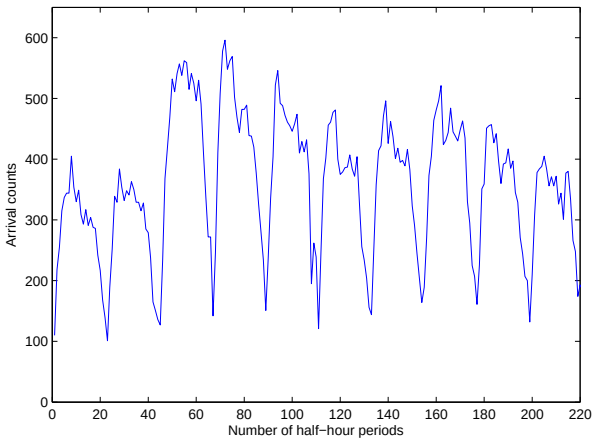

Figure 2: Daily call arrival counts over two consecutive weeks in a Canadian call center.

Figure 3 shows that call volumes are higher, on average, on Mondays than on the remaining weekdays. Figure 3 also shows that all weekdays have a similar daily profile: there are 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 "peak", smaller in magnitude, which 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 [3, 6, 16, 18, 45].

Given that arrival rates are time-varying, 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 minute intervals during a given day; e.g., see [11, 21, 31].

Nevertheless, it is important to perform statistical tests to confirm that it is appropriate to model call center data as a nonhomogeneous Poisson process. [11] proposed a specific test procedure based on Kolmogorov-Smirnov test and did not reject the null hypothesis that arrivals of calls are from a nonhomogeneous Poisson process with piecewise constant rates. [28] examined several alternative test procedures which have greater power compared

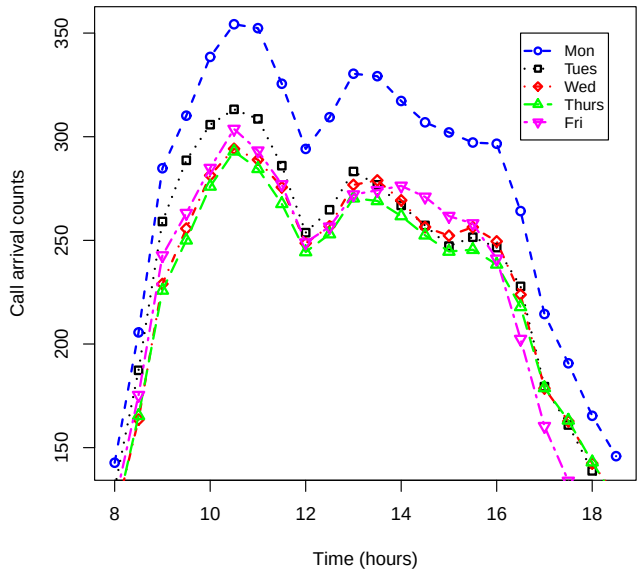

Figure 3: Intraday profile of call arrival counts per weekday in a Canadian call center.

to the one suggested in [11]. [27] 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 data have been accounted for including: data rounding, interval partition and overdispersion caused by combining data.

Overdispersion of arrival counts. A 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 which 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 [3, 6, 26, 43]. 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 [3, 6, 23, 26, 37, 41, 42, 51].

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 acting as 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 by X_j the number of arrivals in a given period 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 notation, we assume in this paper that all periods have the same length and also 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:

$$\mathrm{Var}[X_j] = \mathbb{E}[\mathrm{Var}[X_j|\Lambda_j]] + \mathrm{Var}[\mathbb{E}[X_j|\Lambda_j]] = \mathbb{E}[\Lambda_j] + \mathrm{Var}[\Lambda_j] \ . \tag{2.1}$$

With a random arrival rate function, we have that $\mathrm{Var}[\Lambda_j] > 0$ on the right-hand side of (2.1), which accounts for the additional variance in $\mathrm{Var}[X_j]$.

To model the doubly stochastic Poisson process, [33] proposed a Poisson mixture model with a parametric form of the random Poisson arrival rate. They then incorporated the Poisson mixture model into the M/M/n + G queue and derived asymptotically optimal staffing levels (c-staffing).

Interday and Intraday dependencies. In real-life call centers, there is typically evidence for dependencies between the arrival counts, or arrival rates, in different time periods within a single day, or across several days; e.g., see [3, 6, 15, 40, 45, 53]. Those interday (day-to-day) and intraday dependencies typically remain strong even after correcting for detectable seasonalities. Indeed, it is important to do such a correction to avoid erroneously overestimating dependencies in the data.

In Tables 1 and 2 we illustrate interday and intraday correlations in the same call center as in Figures 1-3. Tables 1 and 2 illustrate several properties which are very commonly observed in practice: (i) correlations are strong and positive between successive weekdays; (ii) interday correlations are slightly smaller with longer lags; (iii) Mondays are less correlated with the

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

Table 1: Correlations between arrival counts on successive weekdays 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 2: Correlations between arrivals in consecutive half-hour periods on Wednesday morning in a Canadian call center.

remaining weekdays; (iv) correlations are strong and positive between successive half-hourly periods within a day; and (v) intraday correlations are slightly smaller with longer lags.

There are 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 [3, 6, 23, 40]. However, since dependencies may not be linear, it is also useful to consider alternative measures such as rank correlation coefficients; see [15] 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 ([3, 23]) and, more generally, copulas ([15, 25]) are ideally suited to

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

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

easily capture interday and intraday dependencies in call center arrival data. Models that fail to account for positive interday and intraday dependence in call arrivals may give an overoptimistic view of call center performance measures, and the resulting errors can be very significant; see [6, 7, 43, 44].

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

In Table 3 (taken from [23]), we present estimates of 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 illustrates that intertype correlations can be strong and positive. Here, call arrivals to the Type A queue originate in the province of Ontario, and are mainly handled in English, whereas arrivals to the Type B queue originate in the province of Quebec, and are mainly handled in French. Otherwise,

arrivals to both queues have similar service requests. Thus, it is reasonable that there exist correlations between their respective arrival processes. There has been some recent effort to model intertype dependencies in the data; see [23] and [25].

Using auxiliary information. Auxiliary information is often available in call centers to improve point or distributional forecasts considerably. For example, when a company sends notification letters to customers, or makes advertisements, this may trigger a large volume of calls; see [30]. Also, large sporting events or festivals can bring a significant increase of calls to emergency systems; see [16].

The past service level in the call center may also be a valuable source of information for predicting future arrivals. For example, long previous delays may lead to a high call abandonment rate, 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 the first call that they make, and they may need to reconnect later. Ignoring such redials and reconnects may lead to considerably underestimating call arrival counts; see [17].

Finally, in certain types of call centers, for example where people may call to report power outages or those designated to 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 [29] 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. In §3, we review some relevant theoretical background, and in §4 we review the relevant literature.

3. Theoretical Background

In this section, we review some relevant theory. Let X_t denote observations taken at equally-spaced intervals. Usually, X_t is the number of call arrivals per time period, such as a half-hour

interval or a day. Alternatively, X_t may also be the rate of call arrivals per time period.

3.1. Autoregressive Integrated Moving Average (ARIMA) Models

ARMA Models. Since an ARIMA model is a generalization of an *autoregressive moving average* (ARMA) model, we begin by describing the latter.

We say that the process $\{X_t, t \geq 0\}$ follows an ARMA model of orders p and q, denoted by ARMA(p, q), if it can be written in the following form:

$$\Phi_p(B)(X_t - \mu) = \Theta_q(B)\epsilon_t \ , \tag{3.1}$$

where:

p, q	are	non-negative integers,
B	is	the backshift operator defined by $BX_t \equiv X_{t-1}$,
$\Phi_p(B)$	$=$	$1 - \phi_1 B - ... - \phi_p B^p$,
$\Theta_q(B)$	$=$	$1 - \theta_1 B - ... - \theta_q B^q$,
ϵ_t	are	independently and identically distributed (i.i.d.) as $N(0, \sigma^2)$.

We call ϕ_i, $1 \leq i \leq p$, the autoregressive parameters and $\theta_j, 1 \leq j \leq q$, the moving-average parameters. We call μ the location parameter. We assume that all roots of the polynomials $\Phi_p(\cdot)$ and $\Theta_q(\cdot)$ lie outside the unit circle. This will guarantee that the process X_t is both stationary and invertible; see [10] for additional details. In what follows, we consider special cases of (3.1) by assigning specific values to the parameters p and q.

MA Models. Letting $p = 0$ and $q = 1$ in (3.1) yields:

$$X_t - \mu = (1 - \theta_1 B)\epsilon_t = \epsilon_t - \theta_1 \epsilon_{t-1} \ . \tag{3.2}$$

This is called a *moving-average model of order 1*, and is denoted by MA(1). Intuitively, ϵ_t and ϵ_{t-1} in (3.2) can be interpreted as shocks that disturb X_t and move it away from the

level μ. The model in (3.2) means that ϵ_t and ϵ_{t-1} affect X_t in the proportions 1 and $-\theta_1$, respectively. However, for a fixed s, each ϵ_s will have no effect on the process beyond time $s+1$. That is, the effect of each shock in the model does not persist with time.

In order to obtain MA models of higher orders, simply let $q > 1$ and $p = 0$ in (3.1). For example, the expression for an MA(q) model is given by:

$$X_t = \epsilon_t - \theta_1\epsilon_{t-1} - \theta_2\epsilon_{t-2} - ... - \theta_q\epsilon_{t-q} \ . \tag{3.3}$$

AR Models. Letting $p = 1$ and $q = 0$ in (3.1), along with $|\phi_1| < 1$, yields:

$$(1 - \phi_1 B)(X_t - \mu) = X_t - \mu - \phi_1 X_{t-1} = \epsilon_t \ . \tag{3.4}$$

This is called an *autoregressive model of order 1*, and is denoted by AR(1). Expanding (3.4) by exploiting the relation between X_t and X_{t-1} yields the following expression:

$$X_t - \mu = \epsilon_t + \phi\epsilon_{t-1} + \phi^2\epsilon_{t-2} + ... \ , \tag{3.5}$$

which is a special case of an MA model where the effects of all previous shocks, ϵ_s for $s \leq t$, persist at time t; however, their respective influences decay exponentially since $|\phi_1| < 1$.

In order to obtain AR models of higher orders, simply let $p > 1$ and $q = 0$ in (3.1). For example, the expression for an AR(p) model is given by:

$$X_t = \mu + \theta_1 X_{t-1} + \theta_2 X_{t-2} + ... + \theta_p X_{t-p} + \epsilon_t \ . \tag{3.6}$$

Differencing and ARIMA Models. The models in (3.1), (3.2), and (3.4) are all stationary models. This means that, for every k, the distribution of the vector $(X_t, X_{t+1}, ..., X_{t+k})$ is independent of t and depends only on k. For example, the assumption that $|\phi| < 1$ in (3.4) ensures for the stationarity of the process. For example, to see why it is a necessary

condition, consider the same AR(1) model but let $\phi_1 = 1$. Then, based on (3.5), we obtain that:

$$X_t - \mu = X_{t-1} + \epsilon_t = \epsilon_t + \epsilon_{t-1} + \epsilon_{t-2} + .. \ , \tag{3.7}$$

where the effect of each ϵ_s, $s \leq t$, is permanent. It is not difficult to see that (3.7) implies that the distribution of X_t changes with t (e.g., the variance of X_t increases with t), which means that the process is nonstationary.

Although the process defined in (3.7) is nonstationary, the process Y_t defined as

$$Y_t \equiv X_t - X_{t-1} = \epsilon_t \ ,$$

clearly is. This subtraction operation is called *differencing* and, more generally, it is common to difference a nonstationary time series to transform it into a stationary one.

We are now ready to give the definition of an ARIMA(p, d, q) model, which is given by:

$$\Phi_p(B)(1 - B)^d(X_t - \mu) = \Theta_q(B)\epsilon_t \ , \tag{3.8}$$

using the same notation as in (3.1) and letting d be the differencing degree parameter; for example, with $d = 2$, we first difference the original series X_t and then difference the resulting differenced series Y_t. In general, d is chosen so that the differenced time series is stationary. Time series that can be made stationary by differencing are called *integrated* processes.

It is also possible to accommodate for multiple seasonalities (e.g., daily, weekly, monthly, etc.) in an ARIMA model. The resulting model is a *seasonal* ARIMA model. For example, the general form of an ARIMA$(p, q, d)\times(p_1, q_1, d_1)$ model with one seasonality period, s, is:

$$\Phi_p(B)\Gamma_s(B^s)(1 - B)^d(1 - B^s)^{d_1}(X_t - \mu) = \Theta_q(B)\Omega_s(B^s)\epsilon_t \ , \tag{3.9}$$

where $\Gamma_s(\cdot)$ is a polynomial of some order p_1, $\Omega_s(\cdot)$ is a polynomial of some order q_1, and d_1 is a differencing parameter.

3.2. Exponential Smoothing

Another important special case of ARIMA modelling emerges from letting $p = 0$ and $d = q = 1$ in (3.8). In this case, we obtain the following series:

$$Y_t \equiv X_t - X_{t-1} = \epsilon_t + \theta Y_{t-1} + \theta^2 Y_{t-2} + \theta^3 Y_{t-3} + \ldots ;$$

in other words, Y_t is equal to an exponentially weighted sum of its past observations. This corresponds to the well-known and widely used practice of *exponential smoothing*.

Holt-Winters Smoothing. The *Holt-Winters method* is an extension of exponential smoothing which accommodates both a trend and a seasonal pattern; see [54]. The Holt-Winters method has two versions, additive and multiplicative, the use of which depends on the characteristics of the particular time series at hand. To illustrate, here are the smoothing equations for the additive Holt-Winters method:

$$\begin{aligned} M_t &= \alpha_0(X_t - S_{t-s}) + (1-\alpha_0)(M_{t-1} + B_{t-1}) \ , \\ B_t &= \alpha_1(M_t - M_{t-1}) + (1-\alpha_1)B_{t-1} \ , \\ S_t &= \alpha_2(X_t - M_t) + (1-\alpha_2)S_{t-s} \ , \end{aligned} \tag{3.10}$$

where B_t is the slope component, M_t is the level component, S_t is the seasonal component, and s is the period of seasonality. The constants α_0, α_1, and α_2 are smoothing parameters, whose values are between 0 and 1. In [47], Taylor extended the Holt-Winters method to accommodate multiple seasonalities in the data.

The early work on call forecasting relied mostly on standard techniques such as ARIMA modeling and exponential smoothing; we review these papers in §4.1. More recently, such traditional forecasting methods are often used as benchmarks against more advanced mod-

elling approaches. For a newly proposed call arrival model to be worth serious consideration, it should at least outperform those standard forecasting techniques.

3.3. Fixed-Effects and Mixed-Effects Models

Gaussian linear fixed-effects and mixed-effects models are useful models which usually build on ARMA models. Indeed, the residuals in fixed-effects and mixed-effects models are often modeled using some ARMA(p, q) process with appropriately chosen p and q. Fixed-effects and mixed-effects models are used in several recent papers on modeling call arrivals; see §4.2.

In general, a *linearly additive fixed-effects* (FE) model for X_t is given by:

$$X_t = \sum_{u=1}^{k} \alpha_u Z_u + \epsilon_t \ , \tag{3.11}$$

where ϵ_t are i.i.d. $N(0, \sigma^2)$ variables and Z_u are explanatory variables assumed to be known with certainty. The parameters α_u, $1 \leq u \leq k$, are to be estimated from data.

For a *linearly multiplicative* FE model, simply replace the sums in (3.11) by products. Multiplicative models are appropriate when the variance of the variable X_t increases with its mean. In order to model a series using a multiplicative model, it is common to proceed indirectly by modeling its logarithm using an additive model.

A *linearly additive mixed-effects* (ME) model is an FE model which includes additional random effects: These are normally distributed random variables which quantify random deviations. The general expression for such an ME model is given by:

$$X_t = \sum_{u=1}^{k} \beta_u Z_u + \sum_{v=1}^{l} \beta_v \gamma_v + \epsilon_t \ , \tag{3.12}$$

where γ_v are normally distributed variables which are assumed to be independent from ϵ, and the parameters β_u, $1 \leq u \leq l$, are to be estimated from data. For example, γ_v may represent the daily deviation for X_t, whereas ϵ_t may represent the normal intradaily noise associated

with X_t. Then, imposing specific covariance structures on γ and ϵ would allow modeling interday and intraday dependencies in the data, respectively. For additional background on linear mixed models, see [34]. In §4.2, we illustrate how linear fixed-effects and mixed-effects models are used in the context of modelling call arrivals.

4. Call Forecasting Approaches

Ideally, we want arrival models that seek to reconcile several objectives. For an arrival model to be realistic, it needs to reproduce the properties that we described in section 2. Simultaneously, for an arrival model to be practically useful, 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 proposed in the literature which aim to reconcile those objectives. We first review early papers which rely mostly on standard forecasting methods (§4.1). Then, we focus on more recent models for arrivals over several days or months (§4.2). Finally, we move to models for arrivals over a single day (§4.3).

4.1. Standard Forecasting Techniques

The early work on forecasting call arrivals usually focused on modeling daily or even monthly total call volumes. Part of the reason for this is due 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 forecasting call arrivals is [49], where the authors 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 ARIMA models to forecast future call volumes, and relied solely on the past history of call arrivals in their models.

In [32], Mabert relied on multiplicative and additive regression models, including covariates for special events and different seasonalities, to forecast daily call arrivals to an emergency call center. He also considered model adjustments which exploit previous forecasting errors to yield more accurate forecasts. He found that such models yield the most accurate forecasts, and are superior to ARIMA models.

Other early papers also relied on standard time series models. For example, [4] 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 can dramatically improve the accuracy of their forecasts, and may have a significant impact on the operational decision-making in the call center. Similarly, [9] used ARIMA models with intervention analysis to forecast telemarketing call arrivals. In this paper, the authors found that such models are superior to additive and multiplicative Holt-Winters exponentially weighted moving average models.

More recently, [5] modeled the daily number of applications for loans at a financial services telephone call center. The authors also went beyond standard ARIMA models by including advertising response and special calender effects; they did so by adding exogenous variables in a multiplicative model. In [16], the authors developed simple additive models for the (small) number of ambulance calls during 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-week), special-day effects (such as the Calgary Stampede which leads to increased call volumes), and autocorrelation of the residuals between successive hours. Their best model outperforms a doubly-seasonal ARIMA model for the residuals of a model which captures only special-day effects.

4.2. 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, $1 \leq j \leq P$, of day i, $1 \leq i \leq D$. 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 ([3, 11, 23, 51]) exploit the following "root-unroot" variance-stabilizing data transformation:

$$Y_{i,j} \equiv (X_{i,j} + 1/4)^{1/2} \ . \tag{4.1}$$

Conditional on the event $\Lambda_{i,j} = \lambda_{i,j}$, and for large values of $\lambda_{i,j}$, $Y_{i,j}$ is approximately normally distributed with mean $\sqrt{\lambda_{i,j}}$ and variance $1/4$; see [12]. The unconditional distribution, with random $\Lambda_{i,j}$, is then a mixture of such normal distributions; therefore, it has a larger variance. Nevertheless, it can be assumed (as an approximation) that the square-root transformed counts $Y_{i,j}$ are normally distributed, particularly if $\mathrm{Var}[\Lambda_{i,j}]$ is not too large. The resulting normality is very useful because it allows fitting linear Gaussian fixed-effects and mixed-effects models to the square-root transformed data; for background, see §3.3.

A better alternative than modelling the arrival counts $X_{i,j}$ would be to model the rates $\Lambda_{i,j}$ directly. The reason is that it is considerably easier to simulate the system with a distributional forecast for the rates rather 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 $\Lambda_{i,j}$. The reason being that, in practice, we do not observe the arrival rates

themselves but only the counts which give only partial information on the rates. This makes estimating arrival rates a more complicated task.

Multiple papers, such as [23, 40, 47, 51], consider a linear fixed-effects model as a benchmark for comparison. To illustrate, let d_i be the day-of-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. In [23], the authors considered the following fixed-effects model for the square-root transformed arrival counts:

$$Y_{i,j} = \sum_{k=1}^{5} \alpha_k I_{d_i}^k + \sum_{l=1}^{22} \beta_l I_j^l + \sum_{k=1}^{5} \sum_{l=1}^{22} \theta_{k,l} I_{d_i}^k I_j^l + \mu_{i,j} \ , \tag{4.2}$$

where $I_{d_i}^k$ and I_j^l are the indicators for day d_i and period j, respectively. That is, $I_{d_i}^k$ (I_j^l) equals 1 if $d_i = k$ ($j = l$), and 0 otherwise. The products $I_{d_i}^k I_j^l$ are indicators for the cross terms between the day-of-week and period-of-day effects. The coefficients α_k, β_l, and $\theta_{k,l}$ are real-valued constants that need be estimated from data, and $\mu_{i,j}$ are independent and identically distributed (i.i.d.) normal random variables with mean 0. Equation (4.2) simplifies to

$$y_{i,j} = \alpha_{d_i} + \beta_j + \theta_{d_i,j} + \mu_{i,j} \ . \tag{4.3}$$

Fixed-effect models seem hard to beat in terms of accuracy of long-term point forecasting (e.g., 2 weeks or more); see [23] and [47]. Nevertheless, with short forecasting times, one can exploit interday and intraday dependencies in the data to obtain more accurate forecasts.

As an improvement, and based on real call center data analysis, [3] proposed the following linear mixed-effects 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 $\text{Var}[\gamma] = \sigma_G^2$. 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. [11] proposed an earlier version of this model, also based on call-center data, without intraday correlations and without special-day effects.

In [23], the authors 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 the vectors of random effects or the vectors of residuals across call types are correlated multinormal. This corresponds to using a normal copula; see [29]. 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 [25]. A strong upper tail dependence for certain call types, for example, 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 $(Y_{i,1}, \ldots, Y_{i,P})$, [38] proposed the use of singular-value decomposition to define a small number of vectors whose linear transformations capture most of the information relevant for prediction. Based on this, [40] then developed a dynamic updating method for the distributional forecasts of arrival rates. [39] proposed a method to forecast the latent rate profiles of a time series of inhomogeneous Poisson processes to enable forecasting future arrival rates based on a series of observed arrival counts. [51] 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 requires long computational times. Moreover,

it is unclear how to incorporate exogenous covariates in such a model.

[2] recently proposed a model based on a Poisson-Gamma process, where $\Lambda_{i,j} = W_{i,j}\lambda_{i,j}$ for fixed $\lambda_{i,j}$'s, and where the multiplicative factors $W_{i,j}$ have a gamma distribution and obey a gamma process. [41] analyzed the effect of advertisement campaigns on call arrivals. Theirs is a Bayesian analysis where they model the Poisson rate function using a mixed model approach. This mixed model is shown to be superior to using a fixed-effects model instead. [51] propose an adaptation of [11] to enable it to update the forecasts of a day defined from the previous days using newly available observations during this day.

[51] 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 requires long computational times. Moreover, it is unclear how to incorporate exogenous covariates in such a model.

In the empirical analysis of [47], several time-series models are compared including ARMA and Holt-Winters exponential smoothing models with multiple seasonal patterns. The latter method was adapted by [46] for modeling both the intraday and intraweek cycles in intraday data. In [48], Taylor extended his model and considered the density forecasting of call arrival rates. To this aim, he developed a new Holt-Winters Poisson count data model with a gamma distributed stochastic arrival rate. He showed that this new model outperformed the basic Holt-Winters smoothing model. [36] comments about 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.

4.3. 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 by $\mathbf{X} = (X_1, \ldots, X_p)$ the vector of arrival counts in those periods.

It is commonly assumed that intraday arrivals follow a Poisson process with a random

arrival rate. [52] proposed to do that 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 to multiply 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_j's to be strongly correlated. 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 variance and expectation decompositions, one can find that $\mathrm{Var}[X_j] = \bar{\lambda}_j(1 + \bar{\lambda}_j \mathrm{Var}[W])$ and for $j \neq k$:

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

This correlation is zero when $\mathrm{Var}[W] = 0$ (a deterministic rate) and approaches 1 when $\mathrm{Var}[W] \to \infty$. [6] 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, $\mathbf{X}$ has a negative multinomial distribution, the parameters of this distribution are easy to estimate, and 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, [6] introduced two different models for $\mathbf{X}$, based on the multivariate Dirichlet distribution.

[26] examined a similar model, but with independent busyness factors, one for each period of the day. Under their model, the Λ_j's are independent, as are the X_j's, which is inconsistent with intraday dependence of call center arrivals. [14] 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 to estimate model parameters. This can provide a smoother (perhaps more realistic) model of the arrival rate. [14] and [15] proposed models that account for time dependence, overdispersion, and intraday dependencies with much more flexibility to match the correlations between the X_j's, 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, $(\Lambda_1, \ldots, \Lambda_p)$, instead of for the vector of counts. [35] examined the relationship between modeling for the vector of counts and for the vector of rates. In particular, they gave explicit formulas for the relationship between the correlation between rates and that between 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.

5. 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 the data based on two call types and 200 consecutive workdays (excluding weekends). For each call type, we implement the following four methods to forecast the arrival counts based on 6-week historical data. The four methods used in our case study are discussed in [3], [23] and [19].

- MU: the multiplicative univariate forecasting model in [19]
- ME: univariate mixed effect model in [3]
- BME1: bivariate mixed effect model in [23]
- BME2: bivariate mixed effect model in [23].

Each method is applied to the data for an out-of-sample rolling forecasting experiment. To compare different models, we use the Root Mean Squared Error (RMSE) to assess the

point forecast accuracy as defined below:

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

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

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

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 is most accurate in point forecasts in most scenarios. BME1 and BME2 has better coverage probability when the leading period is one day or one week, and MU has better coverage probability when the leading period is 2 weeks.

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

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

6. Conclusions

Forecasting call center arrivals plays a crucial role in call center management such as determining staffing level, scheduling plan and routing policy. Call center arrival process is complex and has to be modeled appropriately to achieve better forecasting accuracy, and as

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

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

a result, more efficient operational decisions. In this survey paper, we reviewed the existing literature on modeling and forecasting call center arrival process. We also conducted a case study to evaluate several recently proposed forecasting methods with real-life call center data.

An interesting future research direction is to extend the existing forecasting models or develop new models 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 better meet the quality of service level and improve operational efficiency.

Another research question worth pursuing is to examine the operational impact of improved forecasts, as most existing literature about call center forecasting evaluate forecasting approaches based on only traditional statistical measures such as RMSE and 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 more insight regarding forecasting model selection and system performance evaluation. [19] has tried to tackle this problem for one call type. More research is needed in this direction.

References

[1] O. Z. Akşin, 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.

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

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

[4] B. Andrews and S. M. Cunningham. L.L. Bean improves call-center forecasting. *Interfaces*, 25:1–13, 1995.

[5] A. Antipov and N. Maede. Forecasting call frequency at a financial services call centre. *Journal of the Operational Research Society*, 53:953–960, 2002.

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

[7] A. N. Avramidis and P. L'Ecuyer. Modeling and simulation of call centers. In M. E. Kuhl, N. M. Steiger, F. B. Armstrong, and J. A. Joines, editors, *Proceedings of the 2005 Winter Simulation Conference*, pages 144–152. IEEE Press, 2005.

[8] A. D. Barbour, L. Holst, and S. Janson. *Poisson Approximation*. Oxford University Press, Oxford, U.K., 1992.

[9] L. Bianchi, J. Jarrett, and R. Ch. Hanumara. Improving forecasting for telemarketing centers by ARIMA modeling with intervention. *International Journal of Forecasting*, 14:497–504, 1998.

[10] G. Box, G. Jenkins, and G. Reinsel. *Time Series Analysis: Forecasting and Control*. Prentice Hall, NJ, 1994.

[11] 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*, 100:36–50, 2005.

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

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

[14] N. Channouf. *Modélisation et optimisation d'un centre d'appels téléphoniques: étude du processus d'arrivée*. PhD thesis, Département d'Informatique et de Recherche Opérationnelle, Université de Montréal, Canada, 2008.

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

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

[17] S. Ding, G. Koole, and R. Van Der Mei. On the estimation of the true demand in call centers with redials and reconnects. In *Proceedings of the 2013 Winter Simulation Conference*, 2013. To appear.

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

[19] N. Gans, H. Shen, Y. P. Zhou, N. Korolev, A. McCord, and H. Ristock. Parametric forecasting and stochastic programming models for call-center workforce scheduling. *Manufacturing and Service Operations Management*, 2015. Under minor revision.

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

[21] L. V. Green, P. J. Kolesar, and W. Whitt. Coping with time-varying demand when setting staffing requirements for service system. *Production and Operations Management*, 2006. In press.

[22] R. J. Hyndman, A. B. Koehler, J. K. Ord, and R. D. Snyder. *Forecasting with Exponential Smoothing: the State Space Approach*. Springer, 2008.

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

[24] R. Ibrahim, P. L'Ecuyer, N. Régnard, and H. Shen. On the modeling and forecasting of call center arrivals. In *Proceedings of the 2012 Winter Simulation Conference*, pages 1 – 12. IEEE Press, 2012.

[25] A. Jaoua, P. L'Ecuyer, and L. Delorme. Call type dependence in multiskill call centers. *Simulation*, 0(0):1–13, 2013.

[26] G. Jongbloed and G. Koole. Managing uncertainty in call centers using Poisson mixtures. *Applied Stochastic Models in Business and Industry*, 17:307–318, 2001.

[27] S. H. Kim and W. Whitt. Are call center and hospital arrivals well modeled by nonhomogeneous poisson processes? 2013.

[28] S. H. Kim and W. Whitt. Choosing arrival process models for service systems: Tests of a nonhomogeneous poisson process. 2013.

[29] T. Kim, P. Kenkel, and B. W. Brorsen. Forecasting hourly peak call volume for a rural electric cooperative call center. *Journal of Forecasting*, 31:314–329, 2012.

[30] J. Landon, F. Ruggeri, R. Soyer, and M. M. Tarimcilar. Modeling latent sources in call center arrival data. *European Journal of Operations Research*, 204(3):597–603, 2010.

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

[32] V. A. Mabert. Short interval forecasting of emergency phone call (911) work loads. *Journal of Operations Management*, 5(3):259–271, 1985.

[33] S. Maman, A. Mandelbaum, W. Whitt, and S. Zeltyn. Queues with random arrival rates: Inference, modelling and asymptotics (c-staffing). 2015. Work in progress.

[34] K. Muller and P. Stewart. *Linear Model Theory: Univariate, Multivariate, and Mixed Models.* Wiley, New York, USA., 2006.

[35] B. N. Oreshkin, N. Regnard, and P. L'Ecuyer. Rate-based daily arrival process models with application to call centers. 2014. Working paper.

[36] H. Shen. Exponentially weighted methods for forecasting intraday time series with multiple seasonal cycles: Comments. *International Journal of Forecasting*, 58:652–654, 2010.

[37] H. Shen. Statistical analysis of call-center operational data: Forecasting call arrivals, and analyzing customer patience and agent service. In J. J. Cochran, editor, *Wiley Encyclopedia of Operations Research and Management Science*. John Wiley, 2010.

[38] H. Shen and J. Z. Huang. Analysis of call centre arrival data using singular value decomposition. *Applied Stochastic Models in Business and Industry*, 21:251–263, 2005.

[39] H. Shen and J. Z. Huang. Forecasting time series of inhomogeneous Poisson processes with application to call center workforce management. *Annals of Applied Statistics*, 2(2):601–623, 2008.

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

[41] R. Soyer and M. M. Tarimcilar. Modeling and analysis of call center arrival data: A Bayesian approach. *Management Science*, 54(2):266–278, 2008.

[42] S. G. Steckley, S. G. Henderson, and V. Mehrotra. Service system planning in the presence of a random arrival rate, 2004. submitted.

[43] S. G. Steckley, S. G. Henderson, and V. Mehrotra. Performance measures for service systems with a random arrival rate. In M. E. Kuhl, N. M. Steiger, F. B. Armstrong, and J. A. Joines, editors, *Proceedings of the 2005 Winter Simulation Conference*, pages 566–575. IEEE Press, 2005.

[44] S. G. Steckley, S. G. Henderson, and V. Mehrotra. Forecast errors in service systems. *Probability in the Engineering and Informational Sciences*, 23(2):305–332, 2009.

[45] O. Tanir and R. J. Booth. Call center simulation in Bell Canada. In *Proceedings of the 1999 Winter Simulation Conference*, pages 1640–1647, Piscataway, New Jersey, 1999. IEEE Press.

[46] J. W. Taylor. Short-term electricity demand forecasting using double seasonal exponential smoothing. *Journal of Operational Research Society*, 54:799–805, 2003.

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

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

[49] H. E. Thompson and G. C. Tiao. Analysis of telephone data: A case study of forecasting seasonal time series. *The Bell Journal of Economics and Management Science*, 2(2), 1971.

[50] R. B. Wallace and W. Whitt. A staffing algorithm for call centers with skill-based routing. *Manufacturing and Service Operations Management*, 7(4):276–294, 2005.

[51] J. Weinberg, L. D. Brown, and J. R. 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.

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

[53] W. Whitt. Improving service by informing customers about anticipated delays. *Management Science*, 45(2):192–207, 1999.

[54] P.R. Winters. Forecasting sales by exponentially weighted moving averages. *Management Science*, 6:324–342, 1960.