MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Vol. 00, No. 0, Xxxxx 0000, pp. 000–000
ISSN 1523-4614 | EISSN 1526-5498 | 00 | 0000 | 0001

INFORMS
DOI 10.1287/xxxx.0000.0000

Forecasting Call Center Arrivals: Fixed-Effects, Mixed-Effects, and Bivariate Models

Rouba Ibrahim
Desautels Faculty of Management, McGill University rouba.ibrahim@mail.mcgill.ca,

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

We consider different statistical models for the call arrival process in telephone call centers. We evaluate the forecasting accuracy of those models by describing results from an empirical study analyzing real-life call center data. We test forecasting accuracy using different lead times, ranging from weeks to hours in advance, to mimic real-life challenges faced by call center managers. The models considered are: (i) a benchmark fixed-effects model which does not exploit any dependence structures in the data; (ii) a mixed-effects model which takes into account both interday (day-to-day) and intraday (within day) correlations; (iii) two new bivariate mixed-effects models, for the joint distribution of the arrival counts to two separate queues, which exploit correlations between different call types. Our study shows the importance of accounting for different correlation structures in the data.

Key words: forecasting; arrival process; dynamic updating; correlation; call centers.
History:

1. Introduction

To increase customer satisfaction, service systems compete in improving the quality of service provided, while maintaining high levels of operational efficiency. As a result, service system managers often need to weigh contradictory objectives. In the context of call centers, quality of service is typically measured by customer delay in the system (i.e., the amount of time that callers spend waiting on hold before being handled by an agent), whereas operational efficiency is measured by the proportion of time that agents are busy handling calls. The quality of service in a call center is usually regulated by a service-level agreement (SLA) which need be respected. The SLA specifies target performance levels, such as the wait-time level or the proportion of abandoning customers. For background on call centers, see Gans et al. (2003) and Aksin et al. (2007).

In order to achieve the right balance between quality of service and operational efficiency, call center managers are faced with multiple challenges. First, there is the problem of determining appropriate staffing levels, weeks or even months in advance, based on long-term forecasts of future incoming demand which is typically both time-varying and stochastic. In the words of Aksin et

al. (2007), that is a problem of "resource acquisition". Second, there is the problem of scheduling (and re-scheduling) the available pool of agents based on updated forecasts, typically made several days or weeks in advance. That is a problem of "resource deployment"; see Avramidis et al. (2010). Finally, there are short-term decisions that need be made, such as routing incoming calls in real time to available agents, or mobilizing agents on short notice due to unforseen fluctuations in incoming demand. Those decisions are based on short-term forecasts, updated one day or even a few hours in advance. As an initial step, pending the analysis of effective scheduling and routing designs, it is crucial to develop accurate forecasts of future call volumes, and to study ways of updating those forecasts at different points in time.

1.1. Main Contributions

In this paper, we consider different statistical models for the call arrival process. Specifically, we consider Gaussian linear mixed-effects models for the square-root transformed call arrival counts. For background on linear mixed-effects models, see Muller and Stewart (2006). We conduct an empirical study using real-life call center data, and generate point and confidence interval forecasts of future arrival counts. We test the accuracy of our forecasts using lead times ranging from weeks to hours in advance: We do so to mimic real-life challenges faced by call center managers. Our study shows the importance of accounting for different correlation structures in the data. For example, our mixed-effects models take into account both interday (day-to-day) and intraday (within-day) correlations in the time series of arrival counts. This paper was motivated by an industry research project with a major telecommunications company in Canada; see §3.

The main novelty of this work lies in jointly modeling the arrival counts for different call types handled at the call center. In particular, we use bivariate linear mixed-effects model for the joint distribution of arrival counts to two separate queues. We exploit inter-(call)type correlations and show that bivariate mixed-effects models can lead to more accurate forecasts than univariate mixed-effects models. Bivariate mixed-effects models are traditionally used in the field of biostatistics, e.g., when analyzing longitudinal data of two associated markers; see Barry and Bowman (2007), and Thiébaut et al. (2007). To the best of our knowledge, ours is the first work that proposes using those models in the context of call center applications.

1.2. Fixed-Effects, Mixed-Effects, and Bivariate Models

We now briefly describe the statistical models considered in this paper; for details, see §4. We first consider a simple fixed-effects (FE) model with day-of-week and period-of-day covariates, and independent residuals. This model also includes cross terms to capture the interaction between the

day-of-week and period-of-day effects. The FE model is equivalent to a historical average approach since it essentially uses past averages as forecasts of future call volumes; it was used as a benchmark model in both Weinberg et al. (2007) and Shen and Huang (2008b). The FE model serves as a useful reference point because it does not incorporate any dependence structures in the data. Since there is strong evidence for correlations in the time series of arrival counts (e.g., see §3.2), we anticipate that forecasts based on the FE model will not be very accurate. In §6 and §7, we show that the FE model is useful with relatively long forecasting lead times, but not otherwise.

In §4.2, we extend the FE model and consider a mixed-effects (ME) model incorporating both fixed and random effects. Random effects, which are Gaussian deviates with a specific covariance structure, are used to model interday correlations. Intraday correlations are modeled by imposing a specific covariance structure on the residuals of the model. The resulting ME model also includes the FE model's fixed effects. Consistent with Aldor-Noiman et al. (2009), we show that the ME model generally leads to accurate point and interval forecasts of future call volumes; see §6 and §7. In particular, it is superior to the FE model with relatively short forecasting lead times.

In real-life call centers, there is usually evidence for correlations between the arrival counts of different call types. Intertype correlations may arise in multi-lingual call centers where certain service requests are handled in different languages. More generally, intertype correlations may arise when arrivals to different queues are driven by the same underlying causes, e.g., specific marketing campaigns or special events. In §5, we extend the ME model to bivariate mixed-effects (BME) models which exploit intertype correlations. A BME model jointly models arrivals to two separate queues; it consists of two dependent ME models, each modeling arrivals to one of the two queues.

We propose two ways of modeling correlations across call types. In our first bivariate model, BME1, we specify a correlation structure between the random effects of the two underlying ME models; see §5.1. In our second bivariate model, BME2, we specify a correlation structure between the residuals of the two underlying ME models; see §5.2. The resulting BME models exploit interday, intraday, and intertype correlations. In §6 and §7, we show that BME models generally yield more accurate point and interval forecasts than both FE and univariate ME models.

1.3. Organization

The remainder of this paper is organized as follows. In §2, we review some of the relevant literature. In §3, we describe the data set that motivated this research. In §4, we describe the fixed-effects and univariate mixed-effects models. In §5, we describe the bivariate models. In §6, we compare the forecasting accuracy of our different models using the data set described in §3. In §7, we provide further empirical evidence by considering two additional data sets. In §8, we draw conclusions and describe managerial insights.

2. Literature Review

We now review some of the existing literature on forecasting call center arrivals. Much of the earlier work focuses on applying standard time series methods, such as Autoregressive Integrated Moving Average (ARIMA) models. For example, Andrews and Cunningham (1995) used the ARIMA/ transfer function methodology to forecast arrivals to L. L. Bean's call center, and emphasized the impact of holidays and marketing campaigns on the arrival process. Bianchi et al. (1998) also used ARIMA models and found that they outperform simple Holt-Winters smoothing.

More recent work includes Weinberg et al. (2007) who used a Bayesian approach to forecast incoming calls at a United States bank's call center. They used the same square-root data transformation that we use in this paper, and exploited the resulting normality of data in their model. Taylor (2008) compared the forecasting accuracy of alternative time series models, including a version of Holt-Winters smoothing which accommodates multiple seasonal patterns. He showed that simple forecasting techniques, such as taking historical averages, are difficult to beat with long forecasting lead times. We reach a similar conclusion in this work as well. Shen and Huang (2008a, b) used a Singular Value Decomposition (SVD) approach to create a prediction model which allows for interday forecasting and intraday updating of arrival rates. Aldor-Noiman et al. (2009) proposed an arrival count model which is based on a mixed-effects model including day-of-week, periodic, and exogenous fixed effects. We use a similar mixed-effects model in this paper.

Other empirical studies have shown several important features of the call arrival process. Avramidis et al. (2004) proposed several stochastic models including a doubly stochastic Poisson arrival process with a random arrival rate. Their models reproduce essential characteristics of call center arrivals, such as: (i) a variance considerably higher than with Poisson arrivals, as observed by Jongbloed and Koole (2001), and (ii) strong intraday correlations, as in Tanir and Booth (2001). Then, they tested the goodness-of-fit of their models via an empirical study of real-life data. We also model intraday correlations in this paper. Additionally, we account for interday correlations. Interday correlations were shown to be significant in the seminal paper by Brown et al. (2005). One last feature of the call arrival process, which we also take into account here, is the time variability of arrival rates. Indeed, there is strong empirical evidence suggesting that arrival rates in call centers are usually not stationary; e.g., see Gans et al. (2003), Brown et al. (2005) and Aksin et al. (2007).

3. Preliminary Data Analysis

The present data were gathered at the call center of a major telecommunications company in Canada. They were collected over 329 days (excluding days when the call center is closed, such

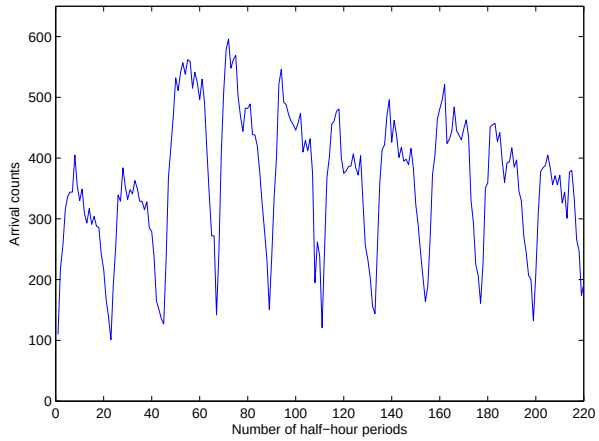

Figure 1 **Type A arrivals for two weeks starting August 19, 2010.**

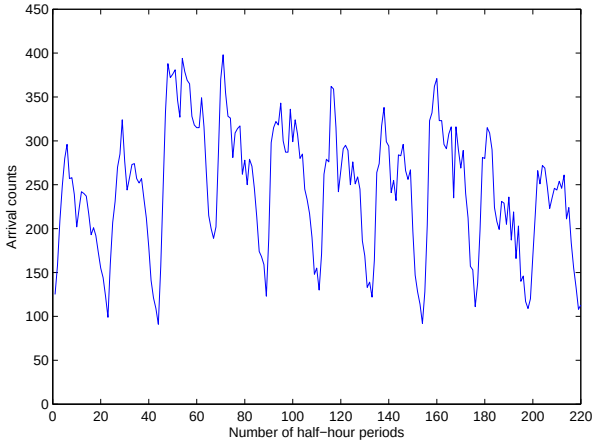

Figure 2 **Type B arrivals for two weeks starting August 19, 2010.**

as holidays and Sundays) ranging from October 19, 2009 to November 11, 2010. The data consist of arrival counts for two call types, Type A and Type B, whose incoming calls originate in the Canadian provinces of Ontario and Quebec, respectively. In §7, we briefly describe two additional data sets, each consisting of arrival counts for two call types as well.

The call center operates from 8:00 AM to 7:00 PM on weekdays (Monday to Friday), and from 8:00 AM to 6:00 PM on Saturdays. Because the call arrival pattern is very different between weekdays and Saturdays, we focus solely on weekdays in this paper. We thus remove a total of 47 Saturdays from the data set. There are "special" days in the data, such as days with missing values or irregular days (i.e., days on which arrival volumes are unusually high or low). In particular, there is a total of 15 special days (including 9 outlier days) which we remove from the set. This leaves us with $D = 329 - 47 - 15 = 267$ remaining days. Arrival counts for each day are aggregated in consecutive time periods of length thirty minutes each. There are $P = 22$ consecutive thirty-minute periods on a weekday, and a total of $D \times P = 267 \times 22 = 5874$ observations in our data set.

3.1. Overview

In Figures 1 and 2, we present plots for the time series of Type A and Type B call counts, respectively, over two weeks ranging from August 19, 2010 to September 1, 2010. In Figures 3 and 4, we plot the average number of arrivals per half-hour period (for each weekday) for Type A and Type B, respectively. Otherwise, both call types have similar service requests. Figures 3 and 4 show that Type A has higher average arrival counts than Type B. Moreover, the daily profiles for the two call types are different. Figure 3 shows that half-hourly averages for Type A do not fluctuate substantially between the hours of 11:00 AM and 5:00 PM, for each weekday. In contrast, Figure

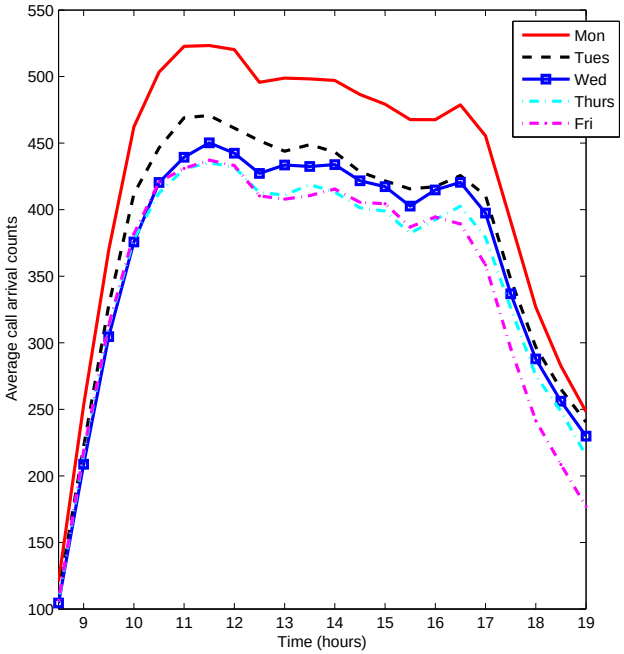

Figure 3 **Average arrival counts per half-hour period for Type A.**

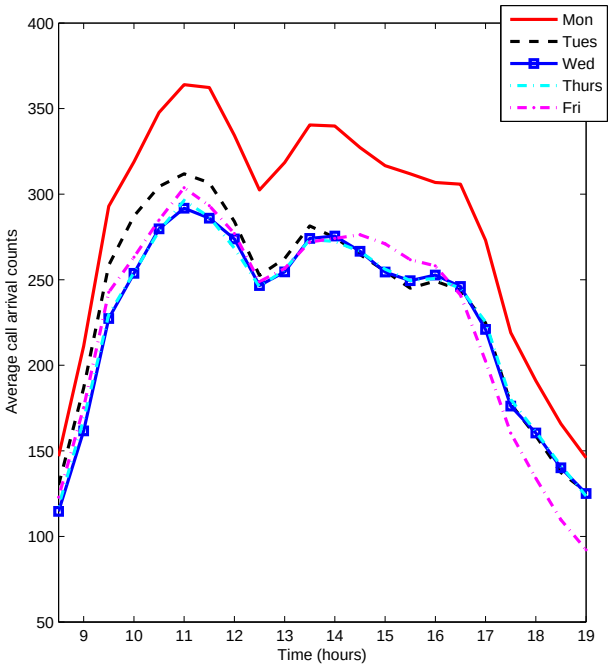

Figure 4 **Average arrival counts per half-hour period for Type B.**

4 shows that there are two major daily peaks for Type B 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 commonly observed in call centers; e.g., see Gans et al. (2003).

3.2. Correlation Structures

Exploratory analysis of our data shows evidence of: (i) strong positive interday correlations between the arrival counts over successive days; (ii) strong positive intraday correlations between the arrival counts over successive half-hour periods of the same day; and (iii) strong positive intertype correlations between the arrival counts of different call types. In Tables 1-3, we present point estimates of those correlations in our data set. Hypothesis tests show that all correlations are strongly statistically significant at the 0.95 confidence level (corresponding p-values are uniformly very small).

3.2.1. Interday correlations. In Table 1, we present estimates of correlations between daily arrival counts over successive weekdays for Type B calls. Table 1 shows that there are substantial positive correlations between days of the same week. In particular, correlations are strong between successive weekdays, and are slightly weaker with longer lags; e.g., the correlation between (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 Type B arrival counts on successive weekdays.

total call volume on) Tuesday and (the total call volume on) Wednesday is 0.68, whereas the correlation between Tuesday and Friday is 0.62. Additionally, Table 1 shows that Mondays are less correlated with the remaining weekdays; e.g., the correlation between Monday and Tuesday is 0.48. Results for Type A calls are largely similar, and are therefore not reported separately.

3.2.2. Intraday correlations. There are positive intraday correlations in the data set. In Table 2, we present estimates of intraday correlations for Type A calls on a given day. In particular, we present correlation estimates for the five consecutive half-hourly periods between 10:00 AM and 12:30 PM on Wednesdays. Table 2 shows that correlations between counts on successive half-hour periods are uniformly strong and positive. Moreover, correlations are weaker with longer lags.

3.2.3. Intertype correlations. In Table 3, we present estimates of correlations between half-hourly arrival counts for Type A and Type B calls. We consider the same consecutive half-hour periods as in Table 2. Table 3 shows that intertype correlations are uniformly strong and positive. Consistent with intuition, intertype correlations are slightly smaller for longer lags. Intertype correlations are relatively easy to interpret in this data set. Indeed, Type A calls originate in the province of Ontario, and are handled in English, whereas Type B calls originate in the province of Quebec, and are mainly handled in French. Otherwise, arrivals to both queues have similar service requests. Thus, we anticipate that there exist correlations between their respective arrival processes. In §5, we propose two bivariate mixed-effects models which exploit such intertype correlations.

3.3. Data Transformation

Let $N_{i,j}$ be the number of arrivals in the j^{th} period of day i, where $i = 1, 2, \ldots, D$ and $j = 1, 2, \ldots, P$. As in Whitt (1999) and Avramidis et al. (2004), we model the arrival process as a doubly stochastic Poisson process with a random arrival rate $\Lambda_{i,j}$. In particular, conditional on $\Lambda_{i,j} = \lambda_{i,j}$ where $\lambda_{i,j} > 0$ is a deterministic value, we assume that $N_{i,j}$ follows a Poisson distribution with arrival rate $\lambda_{i,j}$. As in Jongbloed and Koole (2001), our data possesses overdispersion relative to the Poisson

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 Type A arrivals in consecutive half-hour periods on Wednesday morning.**

Type A / **Type B**	(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.**

distribution, e.g., the variance of the arrival counts is roughly equal to ten times the mean. To stabilize the variance, we use the "root-unroot" method which is commonly used in the literature; e.g, see Brown et al. (2005). In particular, letting $y_{i,j} = \sqrt{N_{i,j} + 1/4}$, it was shown in Brown et al. (2001) that for large values of $\lambda_{i,j}$, $y_{i,j}$ is approximately normally distributed, conditional on $\lambda_{i,j}$, with a mean value of $\sqrt{\lambda_{i,j}}$ and a variance equal to 1/4. Since there are hundreds of calls per period on average in a given weekday, it is reasonable to assume that our square-root transformed counts are roughly normally distributed with the above mean and variance. In §4.2, we exploit normality to fit Gaussian linear mixed-effects models to the transformed data. (We assume that $\sqrt{\lambda_{i,j}}$ is a linear function of random and fixed effects.) In Figures 5 and 6, we consider Type B calls and present Q-Q plots for the residuals of the ME and BME1 models, respectively. (We also include in the plots corresponding envelopes at the 95% confidence level.) We use a forecasting lead time of half a day, and make predictions from August 19, 2010 to November 11, 2010. Figure 6 shows that the normal distribution is a slightly better fit (at both the upper and lower tails) for BME1 residuals than for ME residuals.

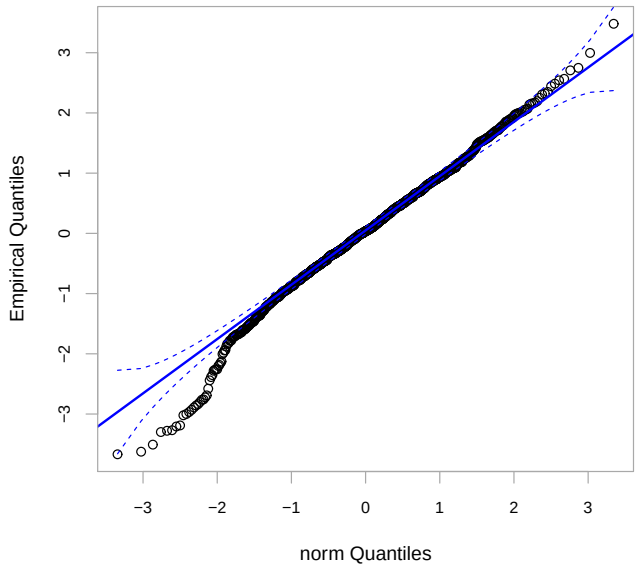

Figure 5 **Q-Q plot for Type B residuals of the ME model with a forecasting lead time of half a day.**

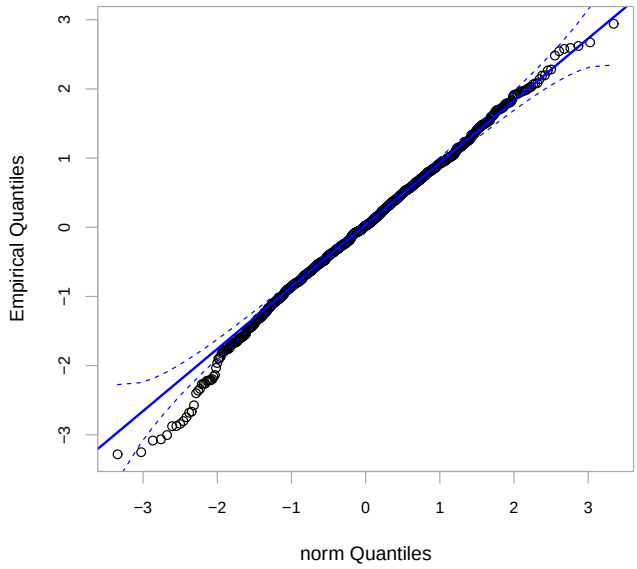

Figure 6 **Q-Q plot for Type B residuals of the BME1 model with a forecasting lead time of half a day.**

4. Fixed-Effects and Mixed-Effects Models

In this section, we describe the fixed-effects and mixed-effects models for the call arrival process. In §6 and §7, we compare our alternative models based on forecasting performance.

4.1. Fixed-Effects (FE) Model with Independent Residuals

The preliminary data analysis of §3 showed that the five weekdays have different expected daily total call volumes. Moreover, the expected number of calls per period for each weekday varies depending on the period; see Figures 3 and 4. We capture those two properties in our first model which is a simple linear additive model incorporating both day-of-week and period-of-day covariates. This model also includes cross terms to capture the interaction between the day-of-week and period-of-day effects. The additional cross terms allow for a different intraday profile for each weekday. We consider the FE model because similar models are often used for forecasting future demand in real-life call centers. As pointed out in §1.2, the FE model is equivalent to a historical average approach since it essentially uses past averages as forecasts of future call volumes. It is a useful reference point because it does not incorporate any correlation structures in the data.

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, $\ldots$, and $d_i = 5$ denotes a Friday. Let j denote the

half-hour period index in day i, where $j = 1, 2, \ldots, P$. We model $y_{i,j}$, the square-root transformed call volume in period j of day i, as:

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

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. The normality assumption enables us to obtain prediction intervals for future observations; see §6 and §7. Equation (1) simplifies to

$$y_{i,j} = \alpha_{d_i} + \beta_j + \theta_{d_i,j} + \mu_{i,j} \ . \tag{2}$$

We estimate model parameters using the method of least squares. Least squares estimates are equivalent to maximum likelihood estimates with normal i.i.d. residuals, as in (2).

4.2. Gaussian Linear Mixed-Effects (ME) Model

As discussed in §3.2, there is evidence of strong correlations in the data at both the interday and intraday levels. In this subsection, we extend the FE model of §4.1 and consider an ME model incorporating both fixed and random effects. We consider the same fixed effects as in (1). Random effects, which are Gaussian deviates with a pre-specified covariance structure, are used to model the interday correlations. Intraday correlations are modeled by imposing a specific covariance structure on the residuals of the model. We fit the ME model to data by computing maximum likelihood estimates of model parameters. Mixed-effects models have been previously considered for call center arrivals, e.g., the ME model described here has been proposed in Aldor-Noiman et al. (2009).

4.2.1. Random effects. Let γ_i denote the daily volume deviation from the fixed weekday effect on day i, where $i = 1, 2, \ldots, D$. 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 $\mathrm{Var}[\gamma] = \sigma_G^2$. (We omit the subscript from a random variable when the specific index is not important.) We assume that random effects follow an autoregressive structure of order 1, AR(1). That is,

$$\gamma_i = \rho_G \gamma_{i-1} + \psi_i \ , \tag{3}$$

where ψ_i are i.i.d. normally distributed random variables with $E[\psi] = 0$ and $\text{Var}[\psi] = \sigma_G^2(1-\rho_G^2)$. Consequently, G has an AR(1) covariance structure. That is, the covariance between γ_i and γ_j is given by

$$\text{cov}(\gamma_i, \gamma_j) = g_{i,j} = \sigma_G^2 \rho_G^{|i-j|} \quad \text{for} \quad 1 \le i, j \le D \ , \tag{4}$$

where ρ_G is the autocorrelation parameter.

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. We preserve the true numerical distance between days by fitting the power transformation covariance structure to G, using the actual duration between days; e.g., the lag between Monday and Tuesday of the same week is equal to 1, whereas the lag between Friday and the following Monday is equal to 3.

4.2.2. Model residuals. Let $\epsilon_{i,j}$ denote the residual effect on period j of day i, where $i = 1, 2, .., D$ and $j = 1, 2, \ldots, P$. That is, $\epsilon_{i,j}$ is the normally distributed periodic deviation from the sum of fixed and random effects. Let R denote the within-day $P \times P$ covariance matrix of the residual effects. We assume that R has an AR(1) covariance structure with variance parameter σ_R^2 and autocorrelation parameter ρ_R. Thus, paralleling (3), we have that

$$\epsilon_{i,j} = \rho_R \epsilon_{i,j-1} + \tau_{i,j} \ , \tag{5}$$

where $\tau_{i,j}$ are i.i.d. normally distributed random variables with $E[\tau] = 0$ and $\text{Var}[\tau] = \sigma_R^2(1-\rho_R^2)$. We also assume that residual effects are independent across different days.

4.2.3. Mixed model formulation. We assume that γ and ϵ are independent. In our ME model, we also include the fixed effects of the FE model in (1). The resulting model for $y_{i,j}$ is

$$y_{i,j} = \alpha_{d_i} + \beta_j + \theta_{d_i,j} + \gamma_i + \epsilon_{i,j} \ , \tag{6}$$

where γ_i and $\epsilon_{i,j}$ satisfy equations (3) and (5), respectively, and α_{d_i}, β_j, and $\theta_{d_i,j}$ are the fixed effects of (2). In Table 4, we present maximum likelihood estimates of the covariance parameters σ_G^2, ρ_G, σ_R^2, and ρ_R, for both Type A and Type B arrivals. We use a learning set consisting of 42 days, as in §6 and §7. Positive interday and intraday correlations are indicated by the values of ρ_G and ρ_R. For details on the estimation of linear mixed models, see Henderson (1975).

4.3. Distributional forecasts

Call center managers need both point and distributional forecasts of future arrival counts. Distributional forecasts are important because they quantify variability around point predictions. That is essential for effective decision making, particularly in highly utilized call centers. Gans et al. (2012)

	FE	ME	BME1	BME2
$\sigma^2_{G,A}$	–	0.33	0.20	0.16
$\rho_{G,A}$	–	0.83	0.83	0.81
$\sigma^2_{G,B}$	–	0.58	0.75	0.60
$\rho_{G,B}$	–	0.73	0.81	0.64
$\sigma^2_{R,A}$	0.67	0.32	0.64	0.67
$\rho_{R,A}$	–	0.15	0.10	0.13
$\sigma^2_{R,B}$	0.84	0.34	0.56	0.66
$\rho_{R,B}$	–	0.24	0.25	0.13
$\rho_{R,AB}$	–	–	–	0.46
$\rho_{G,AB}$	–	–	0.25	–

Table 4 Estimates of covariance parameters for Type A and Type B queues with a learning period of 42 days.

described how to obtain distributional forecasts of future arrival rates based on a statistical model for the realized arrival counts (their model is a multiplicative version of our ME model). They integrated those distributional forecasts in a stochastic programming framework to address issues in call-center workforce management. Shen and Huang (2008b) also described the importance of using distributional forecasts for future arrival rates (see §3.2.2 of that paper). Distributional forecasts are needed, for example, in simulation-based algorithms that optimize the staffing and scheduling of agents, as in Cezik and L'Ecuyer (2008) and Avramidis et al. (2010). If a distributional forecast is available for the arrival rate, then we can simulate by first generating the rate, then the arrivals from a Poisson process with that same rate. Alternatively, if a distributional forecast in available only for the counts, we can generate the arrival counts in each period (from their joint distribution), then spread the arrivals of each period uniformly and independently in that period (this is correct if we assume that the arrival rate is constant in each period), e.g., see Avramidis et al (2004).

Distributional forecasts may consist of confidence interval estimates and densities. In this paper, we assume that the square-root transformed counts, $y_{i,j}$, are normally distributed for all $i = 1, 2, \ldots, D$ and $j = 1, 2, \ldots, P$. Therefore, we rely on conditional multivariate Gaussian theory to obtain interval forecasts for future counts; see Henderson (1975). The conditioning set for a single prediction is the corresponding learning set: With additional information about recent arrivals, the learning set is updated and, consequently, so are the predictions. In §6, we assess the forecasting accuracy of our models by computing both point and interval predictions for future arrival counts.

5. Bivariate Mixed-Effects Models

In this section, we describe bivariate mixed-effects (BME) models which extend the univariate ME model of §4.2 by exploiting the dependence structure between arrivals to separate queues.

Correlations between the arrival counts of different call types are commonly observed in practice; e.g., see Table 3. For example, they may arise in multi-lingual call centers where certain service

requests are handled in different languages. More generally, they may arise when arrivals to different queues are driven by the same underlying causes, e.g., specific marketing campaigns or special events. Thus, it is important to propose and study alternative statistical models which exploit intertype correlations. Here, we describe two such models, BME1 and BME2. The BME1 and BME2 models exploit intertype correlations at the daily and half-hourly levels, respectively. In §6 and §7, we study the forecasting accuracy of the BME1 and BME2 models, and show that they uniformly lead to more accurate point and interval predictions.

5.1. The BME1 Model

Let $y^A_{i,j}$ ($y^B_{i,j}$) denote the square-root transformed arrival count for Type A (Type B) in period j of day i, where $i = 1, 2, \ldots, D$ and $j = 1, 2, \ldots, P$. As in §4.1 and §4.2, we assume that both $y^A_{i,j}$ and $y^B_{i,j}$ are normally distributed random variables. In this subsection and the next, we propose different ways of modeling correlations between $y^A_{i,j}$ and $y^B_{i,j}$.

In our first bivariate model, BME1, we propose modeling intertype correlations at the daily level. We begin by modeling the marginal distribution of the arrival counts for each call type. In particular, we assume that $y^A_{i,j}$ and $y^B_{i,j}$ are each (separately) modeled by a univariate ME model, as in §4.2. That is, paralleling (6), we assume that,

$$y^A_{i,j} = \alpha^A_{d_i} + \beta^A_j + \theta^A_{d_i,j} + \gamma^A_i + \epsilon^A_{i,j} \ , \text{ and,} \tag{7}$$

$$y^B_{i,j} = \alpha^B_{d_i} + \beta^B_j + \theta^B_{d_i,j} + \gamma^B_i + \epsilon^B_{i,j} \ , \tag{8}$$

for $i = 1, 2, \ldots, D$ and $j = 1, 2, \ldots, P$. We assume that the random effects, γ^A_i, follow an AR(1) structure with parameters $\sigma_{G,A}$ and $\rho_{G,A}$, as in (3). We assume that the residuals $\epsilon^A_{i,j}$ follow an AR(1) structure with parameters $\sigma_{R,A}$ and $\rho_{R,A}$, as in (5). We make the same assumptions for Type B. That is, we assume that γ^B_i follow an AR(1) structure with parameters $\sigma_{G,B}$ and $\rho_{G,B}$, and $\epsilon^B_{i,j}$ follow an AR(1) structure with parameters $\sigma_{R,B}$ and $\rho_{R,B}$. In (7) and (8), we also use the same fixed effects, α, β, and θ, as in (2).

In order to model the dependence between $y^A_{i,j}$ and $y^B_{i,j}$, we assume that the random effects γ^A_i and γ^B_i are correlated for every $i = 1, 2, \ldots, D$. In particular, we assume that they satisfy the following set of equations

$$\gamma^A_i = \rho_{G,A}\gamma^A_{i-1} + \nu^A_i \ , \text{ and,} \tag{9}$$

$$\gamma^B_i = \rho_{G,B}\gamma^B_{i-1} + \nu^B_i \ , \tag{10}$$

where ν^A_i are i.i.d. normally distributed random variables with $E[\nu^A] = 0$ and $\text{Var}[\nu^A] = \sigma^2_{G,A}(1 - \rho^2_{G,A})$, and ν^B_i are i.i.d. normally distributed random variables with $E[\nu^B] = 0$ and $\text{Var}[\nu^B] =$

$\sigma^2_{G,B}(1-\rho^2_{G,B})$. We assume that (ν_i^A, ν_i^B) follows a bivariate normal distribution, for every i, and that ν_i^A and ν_i^B are correlated. We denote the correlation between them by $\text{cor}(\nu_i^A, \nu_i^B) = \rho_{G,AB}$. As a result, $y_{i,j}^A$ and $y_{i,j}^B$ are correlated for every i and j, i.e., they are correlated across different days, and across the periods of the same day. In Table 4, we present point estimates for the covariance parameters of the BME1 model. We use a learning set consisting of 42 days. The value of $\rho_{G,AB}$ (which is roughly equal to 0.25) indicates that the two call types are positively correlated at the interday level.

5.2. The BME2 model

For the BME2 model, we also assume that $y_{i,j}^A$ and $y_{i,j}^B$ are each separately modeled by a univariate ME model. In particular, we assume that (7) and (8) continue to hold. As in §5.1, we assume that γ_i^A and γ_i^B each follow an AR(1) structure with covariance parameters ($\sigma_{G,A}$, $\rho_{G,A}$) and ($\sigma_{G,B}$, $\rho_{G,B}$), respectively, for $i = 1, 2, \ldots, D$. However, in contrast with the BME1 model, we now assume that ν_i^A and ν_i^B in (9) and (10) are independent across call types, for every $i = 1, 2, \ldots, D$.

To model intertype correlations, we assume that the residuals of the two underlying ME models, i.e., $\epsilon_{i,j}^A$ and $\epsilon_{i,j}^B$ in (7) and (8), are correlated for each given i and $j = 1, 2, \ldots, P$. As in the BME1 model, we let the residuals $\epsilon_{i,j}^A$ follow an AR(1) structure with parameters $\sigma_{R,A}$ and $\rho_{R,A}$, for each given i and $j = 1, 2, \ldots, P$. Similarly, we let the residuals $\epsilon_{i,j}^B$ follow an AR(1) structure with parameters $\sigma_{R,B}$ and $\rho_{R,B}$. To model correlations between $\epsilon_{i,j}^A$ and $\epsilon_{i,j}^B$, we assume that

$$\epsilon_{i,j}^A = \rho_{R,A}\epsilon_{i,j-1}^A + \kappa_{i,j}^A \ , \text{ and,} \tag{11}$$

$$\epsilon_{i,j}^B = \rho_{R,B}\epsilon_{i,j-1}^B + \kappa_{i,j}^B \ , \tag{12}$$

for $i = 1, 2, \ldots, D$ and $j = 1, 2, \ldots, P$. In (11), $\kappa_{i,j}^A$ are i.i.d. normally distributed random variables with $E[\kappa^A] = 0$ and $\text{Var}[\kappa^A] = \sigma^2_{R,A}(1-\rho^2_{R,A})$. Similarly, in (12), $\kappa_{i,j}^B$ are i.i.d. normally distributed random variables with $E[\kappa^B] = 0$ and $\text{Var}[\kappa^B] = \sigma^2_{R,B}(1-\rho^2_{R,B})$. We assume that $(\kappa_{i,j}^A, \kappa_{i,j}^B)$ follows a bivariate normal distribution, for every fixed i and $j = 1, 2, \ldots, P$, and that $\kappa_{i,j}^A$ and $\kappa_{i,j}^B$ are correlated. We denote the correlation between them by $\text{cor}(\kappa_{i,j}^A, \kappa_{i,j}^B) = \rho_{R,AB}$. We assume that κ^A and κ^B are independent across different days. As a result, $y_{i,j}^A$ and $y_{i,j}^B$ are correlated within a given day i, for all $j = 1, 2, \ldots, P$. However, they are independent across different days. In Table 4, we present point estimates for the covariance parameters of the BME2 model. The estimated value of $\rho_{R,AB}$ is roughly equal to 0.46, and indicates that the two call types are positively correlated at the intraday level.

5.3. Distributional forecasts

It is relatively easy to generate distributional forecasts for bivariate models. Indeed, the joint distribution of the arrival counts for both call types, Type A and Type B, is assumed to be multivariate normal. Therefore, as with the ME model, we can derive distributional forecasts by relying on conditional multivariate Gaussian theory, where the conditioning set is the learning set for a given prediction. In the next section, we assess the forecasting accuracy of our models by computing both point and interval predictions for future arrival counts.

6. Model Comparison

In this section, we compare the statistical models of §4 and §5 based on their forecasting performance. We conduct an empirical study using the data set described in §3. In particular, we make out-of-sample forecasts for alternative forecasting lead times, and quantify the accuracy of the forecasts generated by the candidate models. In Table 5, we present point and interval predictions for both Type A and Type B arrivals. The best values are highlighted in bold. In §7, we consider additional data sets to further substantiate our results.

6.1. Lead Times and Learning Period

We generate out-of-sample forecasts for the forecasting horizon ranging from August 19, 2010 to November 11, 2010. That is, we make forecasts for a total of 55 days (excluding weekends and removed outlier days) and generate $55 \times 22 = 1210$ predicted values for each call type. We consider four different forecasting lead times to mimic real-life challenges faced by call center managers; see §1. In particular, we consider lead times of 2 weeks, 1 week, 1 day, and half a day (which corresponds to 11 half-hour periods in our context). We let the learning period consist of 42 days, i.e., 6 weeks. When we generate a forecast for all periods of a given day, we roll the learning period forward so as to preserve the length of the forecasting lead time. We re-estimate all model parameters after each forecast. We do the estimation using our own code, written in MATLAB.

6.2. Performance Measures

We quantify the accuracy of a point prediction by computing the *mean squared error* (MSE) per half-hour period, defined by:

$$\mathrm{MSE} \equiv \frac{1}{K} \sum_{i,j} (N_{i,j} - \hat{N}_{i,j})^2 \ , \tag{13}$$

where $N_{i,j}$ is the number of arrivals in the j^{th} period of a given day i, $\hat{N}_{i,j}$ is the predicted value of $N_{i,j}$, and K is the total number of predictions made. Consistent with standard practice, we also consider the square root of the MSE, the *root mean squared error* (RMSE), given by

$$\mathrm{RMSE} \equiv \sqrt{\mathrm{MSE}} = \sqrt{\frac{1}{K} \sum_{i,j} (N_{i,j} - \hat{N}_{i,j})^2} \ . \tag{14}$$

Predictions for a forecast lead time of 14 days

	Type A				Type B			
	MSE	RMSE	MAPE	Cover	MSE	RMSE	MAPE	Cover
FE	**3059**	**55.3**	**13.9**	0.47	1292	36.0	**13.3**	0.50
ME	3301	57.5	14.5	0.86	1290	36.0	13.3	0.92
BME1	3293	57.3	14.3	0.90	**1264**	**35.5**	13.4	**0.93**
BME2	3414	58.4	14.7	**0.90**	1269	35.6	13.4	0.92

Predictions for a forecast lead time of 7 days

	Type A				Type B			
	MSE	RMSE	MAPE	Cover	MSE	RMSE	MAPE	Cover
FE	**3186**	**56.4**	**14.1**	0.45	1330	36.5	13.6	0.47
ME	3663	60.5	15.0	0.85	1351	36.8	13.8	0.92
BME1	3506	59.2	14.5	0.93	**1287**	**35.9**	**13.5**	**0.95**
BME2	3366	58.0	14.5	**0.93**	1288	35.9	13.5	0.95

Predictions for a forecast lead time of 1 day

	Type A				Type B			
	MSE	RMSE	MAPE	Cover	MSE	RMSE	MAPE	Cover
FE	3193	56.5	14.1	0.50	1273	35.7	13.4	0.50
ME	2636	51.3	12.6	0.88	1127	33.6	12.5	0.92
BME1	2463	49.6	**12.2**	**0.95**	**1113**	**33.3**	**12.4**	**0.95**
BME2	**2440**	**49.4**	12.3	0.94	1126	33.5	12.5	0.94

Predictions for a forecast lead time of 0.5 days

	Type A				Type B			
	MSE	RMSE	MAPE	Cover	MSE	RMSE	MAPE	Cover
FE	3194	56.5	14.1	0.50	1273	35.7	13.4	0.50
ME	**1612**	**40.2**	**9.80**	0.91	871	29.5	10.6	**0.91**
BME1	1639	40.5	9.80	**0.93**	832	28.8	10.4	0.89
BME2	1616	40.2	9.80	0.92	**806**	**28.3**	**10.2**	0.90

Table 5 Accuracy of point and interval predictions for Type A and Type B calls for alternative forecasting lead times and a learning period of 42 days.

Predictions for a forecast lead time of 14 days

	Type C				Type D			
	MSE	RMSE	MAPE	Cover	MSE	RMSE	MAPE	Cover
FE	585	24.2	12.8	0.53	1289	36.0	24.2	0.27
ME	617	24.8	13.1	**0.94**	1198	34.6	23.2	0.89
BME1	538	23.2	12.6	0.91	1186	34.4	23.2	**0.94**
BME2	**531**	**23.0**	**12.5**	0.92	**1117**	**33.4**	**22.5**	0.88

Predictions for a forecast lead time of 7 days

	Type C				Type D			
	MSE	RMSE	MAPE	Cover	MSE	RMSE	MAPE	Cover
FE	546	23.4	12.4	0.52	1132	33.7	22.5	0.37
ME	548	23.4	12.4	0.93	**970**	**31.0**	**20.4**	0.92
BME1	**493**	**22.2**	**12.1**	**0.96**	980	31.2	20.6	0.93
BME2	514	22.6	12.2	0.93	1025	32.0	21.2	**0.95**

Predictions for a forecast lead time of 1 day

	Type C				Type D			
	MSE	RMSE	MAPE	Cover	MSE	RMSE	MAPE	Cover
FE	523	22.9	12.0	0.56	1009	31.8	19.5	0.44
ME	410	20.2	10.6	0.94	592	24.3	14.1	0.94
BME1	**387**	**19.7**	**10.6**	**0.96**	559	23.6	14.0	0.96
BME2	389	19.7	10.7	0.94	**468**	**21.6**	**13.5**	**0.96**

Predictions for a forecast lead time of 0.5 days

	Type C				Type D			
	MSE	RMSE	MAPE	Cover	MSE	RMSE	MAPE	Cover
FE	548	23.4	12.3	0.52	1202	34.6	24.6	0.33
ME	365	19.1	9.9	0.95	471	21.7	13.8	0.95
BME1	336	18.3	9.9	0.96	461	21.5	13.8	0.95
BME2	**335**	**18.3**	**9.8**	**0.96**	**455**	**21.3**	**13.7**	**0.95**

Table 6 **Accuracy of point and interval predictions for Type C and Type D calls for alternative forecasting lead times and a learning period of 42 days.**

Predictions for a forecast lead time of 14 days

	Type E				Type F			
	MSE	RMSE	MAPE	Cover	MSE	RMSE	MAPE	Cover
FE	1781	42.2	**16.2**	0.40	**316**	**17.8**	**22.0**	0.50
ME	**1779**	**42.2**	16.2	**0.87**	319	17.8	22.2	**0.94**
BME1	1860	43.1	16.3	0.84	330	18.2	22.9	0.91
BME2	1979	44.4	16.9	0.85	327	18.1	22.7	0.92

Predictions for a forecast lead time of 7 days

	Type E				Type F			
	MSE	RMSE	MAPE	Cover	MSE	RMSE	MAPE	Cover
FE	2034	45.1	17.6	0.44	**316**	**17.8**	**22.3**	0.52
ME	2056	45.3	17.8	0.86	326	18.0	22.8	0.94
BME1	**2008**	**44.8**	**17.4**	**0.86**	326	18.0	22.9	**0.94**
BME2	2048	45.3	17.6	0.84	324	18.0	22.8	0.94

Predictions for a forecast lead time of 1 day

	Type E				Type F			
	MSE	RMSE	MAPE	Cover	MSE	RMSE	MAPE	Cover
FE	2010	44.8	17.4	0.48	322	17.9	22.8	0.56
ME	1514	38.9	15.0	0.88	319	17.9	22.3	0.94
BME1	**1483**	**38.5**	**14.8**	**0.88**	321	17.9	22.3	0.94
BME2	1484	38.5	14.8	0.86	**318**	**17.8**	**22.2**	**0.94**

Predictions for a forecast lead time of 0.5 days

	Type E				Type F			
	MSE	RMSE	MAPE	Cover	MSE	RMSE	MAPE	Cover
FE	2032	45.6	17.4	0.47	322	17.9	22.7	0.56
ME	1068	32.7	12.0	**0.88**	318	17.9	22.4	0.94
BME1	1026	32.0	11.7	0.82	**312**	**17.7**	**21.8**	0.93
BME2	**1015**	**31.9**	**11.6**	0.82	318	17.9	22.4	**0.95**

Table 7 Accuracy of point and interval predictions for Type E and Type F calls for alternative forecasting lead times and a learning period of 42 days.

In addition to the MSE and RMSE, we compute, for a relative measure of accuracy, the *mean absolute percentage error* (MAPE) defined by:

$$\text{MAPE} \equiv 100 \cdot \frac{1}{K} \sum_{i,j} \frac{|N_{i,j} - \hat{N}_{i,j}|}{N_{i,j}} \ . \tag{15}$$

To evaluate the distributional forecasts generated by the candidate models, we use performance measures to describe the prediction intervals for $N_{i,j}$; see §4.3 and §5.3. In particular, we define the *average cover* (Cover) of a prediction interval for a given model as:

$$\text{Cover} = \frac{1}{K} \sum_{i,j} I(N_{i,j} \in (\hat{L}_{i,j}, \hat{U}_{i,j})) \ , \tag{16}$$

where $I(\cdot)$ denotes the indicator random variable, and $\hat{L}_{i,j}$ and $\hat{U}_{i,j}$ are the lower and upper bounds of the prediction interval, respectively. In this paper, we compute prediction intervals with a confidence level of 95%. If the chosen model adequately captures the correlation structure in the data, then we expect that the average cover be close to 95%. The average cover corresponds to the unconditional coverage of Christoffersen (1998).

6.3. Forecasting Performance

6.3.1. Two-weeks-ahead forecasts.

Predictions for Type A calls. With a forecasting lead time of two weeks, Table 5 shows that the FE model generates the most accurate point forecasts, among all models considered. Consistent with Taylor (2008), this shows that a simple historical average can be difficult to beat with relatively long forecasting lead times. Indeed, Table 5 shows that both MSE(ME) and MSE(BME1) are roughly 8% larger than MSE(FE). Moreover, MSE(BME2) is roughly 12% larger than MSE(FE). Table 5 also shows that MAPE(FE) is roughly 0.6% smaller than MAPE(ME), and roughly 0.4% smaller than MAPE(BME1). However, the predictive power of bivariate models becomes evident when comparing the prediction intervals generated by the alternative models. Indeed, the average cover for both BME1 and BME2 is close to 0.90, whereas Cover(ME) is roughly equal to 0.86 and Cover(FE) lags behind at 0.47. Indeed, the FE model considerably underestimates uncertainty by not capturing any correlation structure between the arrival counts.

Predictions for Type B calls. With a forecasting lead time of two weeks, Table 5 shows that all models perform nearly the same from the MSE and MAPE perspectives. Indeed, MSE(BME1) is the smallest among all models considered, and MSE(FE)/MSE(BME1) is roughly equal to 1.02. Additionally, Table 5 shows that the MAPE's for all models are roughly the same. With Type B arrivals, the average cover of prediction intervals for the ME and BME models is approximately equal to 0.93, whereas the cover for the FE model is much lower, and is roughly equal to 0.5.

6.3.2. One-week-ahead forecasts.

Predictions for Type A calls. With a forecasting lead time of one week, Table 5 shows that the FE model continues to generate the most accurate point forecasts. Indeed, MSE(ME)/MSE(FE) is roughly equal to 1.15, whereas MSE(BME1)/MSE(FE) is roughly equal to 1.10. Both the BME1 and BME2 models generate more accurate point forecasts than the ME model. For example, the difference between MAPE(ME) and MAPE(BME1) is roughly equal to 0.5%. Moreover, the cover of prediction intervals for the BME models is larger than for the ME model; e.g., Cover(BME1) ≈ 0.92 and Cover(BME2) ≈ 0.93, whereas Cover(ME) ≈ 0.85.

Predictions for Type B calls. Table 5 shows that the BME1 model yields the most accurate point predictions. For example, MSE(ME) is roughly 5% larger than MSE(BME1), and MAPE(ME) is roughly 0.3% larger than MAPE(BME1). Moreover, the cover of prediction intervals for the BME models is slightly larger than for the ME model. The FE model continues to yield accurate point forecasts, e.g., MSE(FE)/MSE(BME1) ≈ 1.03. Consistent with previous results, the cover of prediction intervals generated by the FE model is poor. Indeed, Cover(FE) ≈ 0.47.

6.3.3. One-day-ahead forecasts.

Predictions for Type A calls. With a forecasting lead time of one day, Table 5 shows that BME models yield more accurate point and interval forecasts than both the ME and FE models. For example, MSE(ME) is roughly 8% larger than MSE(BME2), and MAPE(ME) is roughly 0.4% larger than MAPE(BME1). Additionally, Cover(BME1) and Cover(BME2) are both roughly equal to 0.95, whereas Cover(ME) is roughly equal to 0.88. The FE model yields considerably less accurate point and interval forecasts than the BME and ME models; e.g., MSE(FE)/MSE(BME2) ≈ 1.31.

Predictions for Type B calls. The ME, BME1, and BME2 models perform roughly the same in this case. Indeed, MSE(BME1) is only slightly smaller than MSE(ME), and MAPE(ME) is roughly equal to MAPE(BME1). The average cover of prediction intervals for the BME1 model (approximately equal to 0.95) is slightly larger than for the ME model (approximately equal to 0.92).

6.3.4. Within-day forecasts.

Predictions for Type A calls. Table 5 shows that the ME and BME models perform roughly the same in this case. Indeed, MSE(ME) is only marginally smaller than MSE(BME2). Moreover, the BME and ME models yield prediction intervals with a good cover; e.g., Cover(ME) ≈ 0.91 and Cover(BME1) ≈ 0.93. The FE model performs considerably worse than the ME and BME models, in terms of both point and interval predictions. For example, MSE(FE)/MSE(BME1) is roughly

	FE	ME	BME1	BME2
$\sigma^2_{G,A}$	–	0.42	0.38	0.24
$\rho_{G,A}$	–	0.71	0.62	0.78
$\sigma^2_{G,B}$	–	0.65	0.59	0.56
$\rho_{G,B}$	–	0.66	0.57	0.64
$\sigma^2_{R,A}$	1.22	0.79	0.48	0.45
$\rho_{R,A}$	–	0.60	0.14	0.16
$\sigma^2_{R,B}$	1.31	0.70	0.50	0.53
$\rho_{R,B}$	–	0.56	0.31	0.21
$\rho_{R,AB}$	–	–	–	0.15
$\rho_{G,AB}$	–	–	0.74	–

Table 8 Estimates of covariance parameters for Type C and Type D queues with a learning period of 42 days.

equal to 2. For within-day updates, we also consider a simple adjustment (not shown in Table 5) of the FE model's forecasts. In particular, we implement the historical proportions method described in Mehrotra et al. (2010) and used in Shen and Huang (2008b). For a given updating time point, m, we compute the ratio between the square-root transformed count up to m and the forecasted count up to m, and use this ratio to update forecasts for the remaining half-hour periods after m. We found that this adjustment leads to slightly better predictions than the FE model, but that the difference in performance is not great. Thus, we do not include separate results for the adjusted forecasts here.

Predictions for Type B calls. With a forecasting lead time of half a day, the BME2 model yields more accurate point and interval forecasts than the ME model. For example, MSE(ME) is roughly 8% larger than MSE(BME2) and MAPE(ME) is roughly 0.4% larger than MAPE(BME2). The BME1 model performs slightly worse than the BME2 model, but better than the ME model. Finally, the cover of prediction intervals for the BME and ME models are all roughly equal to 0.9.

The results of this section show that the BME models usually lead to more accurate point and interval forecasts than both the ME and FE models. In §7, we consider two additional data sets and provide further empirical evidence supporting our main conclusions.

7. Additional Data Sets

In this section, we consider two additional data sets, taken from the same call center as in §3, and study the forecasting accuracy of our candidate models using each set. Preliminary data analysis of the additional sets is largely consistent with our previous analysis in §3. Therefore, we omit a detailed description here. In §7.1 and §7.2, we present estimates for the forecasting performance of the FE, ME, and BME models. For our predictions, we continue to use a learning period of 42 days, and forecasting lead times of 2 weeks, 1 week, 1 day, and half a day.

7.1. First Additional Data Set: Call Types C and D

The data from the first set were collected over 305 days (excluding weekends) ranging from June 1, 2010 to July 25, 2011. The data consist of arrival counts for two call types, Type C and Type D, whose incoming calls originate in the Canadian provinces of Ontario and Quebec, respectively. The two types correspond, otherwise, to similar service requests. We remove 8 outlier days from data. Thus, we are left with a total of $D = 305 - 8 = 297$ remaining days. For the BME models in Table 6, we jointly model Type C and Type D arrivals. Table 6 shows that we get consistent results for both call types. Thus, we only discuss results for Type C calls here. In Table 8, we present point estimates for the covariance parameters of all models with a learning period of 42 days.

7.1.1. Two-weeks-ahead forecasts. Table 6 shows that, with a forecasting lead time of two weeks, the BME2 model generates the most accurate point predictions. The BME1 model performs roughly the same as BME2. For example, MSE(ME) is roughly 16% larger than MSE(BME2), and is roughly 15% larger than MSE(BME1). Both MAPE(BME2) and MAPE(BME1) are roughly 0.6% smaller than MAPE(ME). The FE model yields more accurate point predictions than the ME model. Indeed, MSE(FE) is roughly 11% larger than MSE(BME2). The ME and BME models all yield similar prediction intervals of future counts. Indeed, Cover(BME2) is roughly equal to 0.92, whereas Cover(ME) is roughly equal to 0.94. Consistent with §6, the FE model yields prediction intervals which have a relatively small cover. Indeed, Cover(FE) ≈ 0.53.

7.1.2. One-week-ahead forecasts. Table 6 shows that, with a forecasting lead time of one week, the BME1 model yields the most accurate point forecasts. Indeed, MSE(ME) is roughly 11% larger than MSE(BME1). The BME2 model continues to outperform the ME model as well: MSE(ME) is roughly 6% larger than MSE(BME2). The FE model yields similar point predictions as the ME model. The BME2 model yields prediction intervals with the best cover: Cover(BME2) ≈ 0.96.

7.1.3. One-day-ahead forecasts. In this case, the BME1 and BME2 models perform nearly the same, and continue to yield more accurate point and interval forecasts than the ME model. For example, MSE(ME) is about 6% larger than MSE(BME2). The FE model performs considerably worse; e.g., MSE(FE) is 35% larger than MSE(BME2). Prediction intervals for the ME and BME models all have good covers, and are each close to 0.95.

7.1.4. Within-day forecasts. As before, the BME1 and BME2 models perform nearly the same, both outperforming the ME model. (The FE model is no longer competitive in this case.) Indeed, MSE(ME) is roughly 10% larger than both MSE(BME1) and MSE(BME2). The covers of prediction intervals for the ME and BME models are roughly the same; e.g., Cover(ME) ≈ 0.96.

	FE	ME	BME1	BME2
$\sigma^2_{G,A}$	–	0.51	0.36	0.29
$\rho_{G,A}$	–	0.67	0.44	0.87
$\sigma^2_{G,B}$	–	0.17	0.10	0.09
$\rho_{G,B}$	–	0.80	0.82	0.89
$\sigma^2_{R,A}$	0.92	0.37	0.45	0.46
$\rho_{R,A}$	–	0.39	0.45	0.35
$\sigma^2_{R,B}$	0.55	0.24	0.60	0.57
$\rho_{R,B}$	–	0.41	0.34	0.30
$\rho_{R,AB}$	–	–	–	0.1
$\rho_{G,AB}$	–	–	0.53	–

Table 9 Estimates of covariance parameters for Type E and Type F queues with a learning period of 42 days.

7.2. Second Additional Data Set: Call Types E and F

As in §7.1, the data were also collected over 305 days (excluding weekends) ranging from June 1, 2010 to July 25, 2011. The data consist of arrival counts for two call types, Type E and Type F. Both call types originate in the Quebec province, but they correspond to different service requests. Call center managers at the company have informed us that, based on their experience, arrivals to those two call types are dependent in practice. In this subsection, we investigate whether exploiting this dependence leads to more accurate forecasts. We remove 9 outlier days from the data set. Thus, we are left with a total of $D = 305 - 9 = 296$ remaining days. For the BME1 and BME2 models in Table 7, we jointly model Type E and Type F arrivals. In Table 9, we present corresponding point estimates for the covariance parameters of all models with a learning period of 42 days.

Type F calls have a significantly smaller volume than Type E calls; e.g., there are roughly three times more Type E than Type F calls. As a result, the normal approximation of §3.3 is no longer reasonable. With Type F calls, all forecasting methods considered generate similar point and interval predictions. In particular, Table 7 shows that the MSE (and MAPE) for the FE, BME, and ME models are all roughly equal in this case, irrespective of forecasting lead times. We do not explore this issue further here, and leave investigating forecasting methods with small call volumes to future research; see §8. Next, we discuss results for Type E calls.

7.2.1. Two-weeks-ahead forecasts. Table 7 shows that the FE model yields the most accurate point forecasts among all models considered. The ME model performs nearly the same as the FE model. Indeed, MSE(ME) is roughly equal to MSE(FE). The BME2 model yields the least accurate forecasts in this case, with MSE(BME2) roughly 10% larger than MSE(FE). Table 7 also shows that MAPE(ME) and MAPE(BME1) are roughly the same. The BME2 model falls behind, and MAPE(BME2) is roughly 0.7% larger than MAPE(FE). Consistent with prior results, the cover of prediction intervals generated by the FE model is poor (roughly equal to 0.4). The covers

of prediction intervals generated by the BME models (both roughly equal to 0.85) are slightly smaller than the cover of prediction intervals generated by the ME model (roughly equal to 0.87).

7.2.2. One-week-ahead forecasts. Table 7 shows that the most accurate model, under the MSE and MAPE criteria, is the BME1 model. For example, MSE(ME) is roughly 3% larger than MSE(BME1), and MAPE(ME) is nearly 0.4% larger than MAPE(BME1). The BME2 model performs roughly the same as the ME model. Additionally, the average covers of prediction intervals generated by the ME and BME models are all roughly the same (approximately 0.86).

7.2.3. One-day-ahead forecasts. Table 7 shows that the BME1 and BME2 models perform roughly the same in this case. The BME2 model yields the most accurate point predictions, but MSE(BME2) is only 2% smaller than MSE(ME). Additionally, MAPE(ME) is about 0.2% larger than MAPE(BME1). The covers of prediction intervals generated by the ME and BME models are similar. The FE model performs very poorly: MSE(FE)/MSE(BME1) is roughly equal to 1.36.

7.2.4. Within-day forecast. Consistent with previous results, the BME1 model yields the most accurate point predictions. Indeed, MSE(ME) is roughly 5% larger than MSE(BME1). Moreover, MAPE(ME) is roughly 0.4% larger than MAPE(BME1). The BME1 model is only slightly less accurate than the BME2 model: MSE(BME2)/MSE(BME1) is roughly equal to 1.01. The cover of prediction intervals generated by the ME model is better than for prediction intervals generated by the BME models. For example, Cover(ME) ≈ 0.88, whereas Cover(BME1) ≈ 0.82.

8. Conclusions

8.1. Overview

We considered alternative statistical models for the arrival counts to a call center. We evaluated the forecasting accuracy of those models based on three real-life data sets taken from the call center of a major telecommunications company in Canada. We used forecasting lead times ranging from weeks to hours in advance to mimic challenges faced by call center managers.

We investigated the importance of accounting for different correlation structures in the data when forecasting future arrival counts. Exploiting interday and intraday correlation structures was shown to be useful in Aldor-Noiman et al. (2009). Here, we proposed bivariate models exploiting interday, intraday, and intertype correlations, and we fit those models to data consisting of three different pairs of call types. We showed that jointly modeling arrivals to alternative call types is useful, and needs to be studied further. In particular, both point and distributional forecasts for bivariate models are generally more accurate than for other models considered. With the particular data at hand, the difference in performance between the ME and BME models is, admittedly,

not as great as the difference between the FE and BME models. Nevertheless, Tables 5-7 showed that bivariate models are generally a better fit to data than univariate models; e.g., corresponding prediction intervals consistently have a better cover. Thus, our paper showed that bivariate models are of interest and need be investigated further. In particular, our paper motivates a more in-depth study to unveil conditions (e.g., using simulation or other data sets) under which bivariate methods improve both point and distributional forecasts significantly.

8.2. Managerial Insights

8.2.1. Call volumes. In this paper, we used a square-root transformation of the data (§3.3). The transformed counts are roughly normally distributed when the number of arrivals is sufficiently large; see Brown et al. (2005). Throughout, we exploited normality to fit Gaussian mixed-effects models to the transformed data. Figures 5 and 6 showed that a normal approximation is reasonable with many incoming calls, e.g., thousands of calls per day. When the number of arrivals is relatively small, using a normal approximation is no longer reasonable. Table 7 showed that the FE, ME, and BME models perform roughly the same with Type F calls: all forecasts are not very accurate, irrespective of forecasting lead time. In a small real-life call center, traditional forecasting methods, such as Holt-Winters smoothing or seasonal ARIMA modeling, may be better alternatives than using mixed-effects models. To model the joint distribution of arrival counts in successive time periods, we could consider copulas with discrete marginal distributions, as in Channouf et al. (2012). However, we do not explore this issue further here, and leave it as a potential direction for future research. Arguably, there may be less interest in forecasting arrivals to queues with very small call volumes. Indeed, such queues are typically easier to manage in practice.

8.2.2. Forecasting lead times. Consistent with Taylor (2008), Tables 5-7 showed that a simple historical average can be difficult to beat with a relatively long forecasting lead time. Intuitively, modeling correlation structures in the data is not necessary for long-term forecasts. Thus, it is sufficient for real-life call center managers to base their long-term managerial decisions on historical averages, e.g., fixed-effects models. The FE model is appealing because it has the advantage of computational efficiency: It requires fewer parameter estimations than both the ME and BME models. With short forecasting lead times, Tables 5-7 showed the importance of accounting for correlation structures in the data. For example, Table 5 showed that, with a forecasting lead time of one day, the FE model leads to considerably less accurate forecasts than both mixed and bivariate models. Thus, our study shows that real-life call center managers would benefit from updating their long-term forecasts a few days or hours in advance. In doing so, they exploit interday, intraday, and intertype dependencies in their data.

9. References

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

Aldor-Noiman, S. 2006. Forecasting demand for a telephone call center: Analysis of desired versus attainable precision. Unpublished masters thesis, Technion-Israel Institute of Technology, Haifa, Israel.

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

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

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

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

Barry, S. J. E. and A. Bowman. 2007. Linear mixed models for longitudinal shape data with applications to facial modeling. *Journal of Biostatistics*, 9: 555–565.

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

Brown, L. D., Zhang, R. and L. Zhao. 2001. Root un-root methodology for non parametric density estimation. Technical report, The Wharton School of the University of Pennsylvania, Philadelphia, USA.

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

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

Channouf, N., and P. L'Ecuyer. 2012. A Normal copula model for the arrival process in call centers, *International Transactions in Operational Research*, forthcoming.

Cezik, T. and P. L'Ecuyer. 2008. Staffing Multiskill call centers via linear programming and simulation", *Management Science*, 54: 310–323.

Christoffersen, P. 1998. Evaluating interval forecasts. *International Economic Review*, 39: 841–862.

Claeskens, G. and J. Hart. 2009. Goodness-of-fit tests in mixed models. *Test*, 18: 213-239.

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

Gans, N., H. Shen, N. Korolev, A. McCord, and H. Ristock. 2012. Parametric stochastic programming models for call-center workforce scheduling. *Working paper.* Downloadable from http://faculty.washington.edu/yongpin/.

Henderson, C. R. 1975. Best linear unbiased estimation and prediction under a selection model. *Biometrics*, 31: 423–447.

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

Mandelbaum, A. 2002. Call Centers (Centres): Research Bibliography with Abstracts, Version 3. Downloadable from ie.technion.ac.il/ serveng/References/ccbib.pdf.

Mehrotra, V., O. Ozlük, and R. Saltzman. 2010. Intelligent Procedures for Intra-day Updating of Call Center Agent Schedules. *Production and Operations Management*, 19 : 353–367.

Muller, K. and P. Stewart. 2006. Linear Model Theory: Univariate, Multivariate, and Mixed Models. Wiley, New York.

Soyer, R. and M. Tarimcilar. 2008. Modeling and Analysis of call center arrival data: A Bayesian approach. *Management Science*, 54: 266–278.

Shen, H. and J. Z. Huang. 2008a. Forecasting time series of inhomogeneous poisson process with application to call center management software. *Annals of Applied Statistics*, 2: 601–623.

Shen, H. and J. Z. Huang. 2008b. Intraday forecasting and interday updating of call center arrivals. *Manufacturing and Service Operations Management*, 10: 391–410.

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

Tanir, O. and R. J. Booth. 1999. Call center simulation in Bell Canada. *Proceedings of the 1999 Winter Simulation Conference.* P. A. Farrington, H. B. Nemhard, D.T. Sturrock, G.W. Evans, eds, 1640–1647.

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

Thiébaut, R., Jacqmin-Gadda, H., Chne, G., Leport, C. and D. Commenges. 2007. Bivariate linear mixed models using SAS proc MIXED. *Computer Methods and Programs in Biomedicine*, 69: 249-56

Weinberg, J., Brown, L. D. and J. R. Stroud. 2007. Bayesian forecasting of an inhomogeneous Poisson process with applications to call center data. *Journal of American Statistical Association*, 102: 1185–1199.

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