FORECASTING CALL CENTER ARRIVALS: A COMPARATIVE STUDY

by

Rouba Ibrahim and Pierre L'Ecuyer

Department of Computer Science and Operations Research

University of Montreal

{ibrahiro, lecuyer}@iro.umontreal.ca

Abstract

We evaluate alternative time series methods for forecasting future call volumes in call centers. Our methods take into account both interday (day-to-day) and intraday (within day) dependence structures, and allow for real-time dynamic updates. We also propose a new model which exploits correlations between the arrival processes of two separate queues, leading to more accurate forecasts. We describe results from an empirical study analyzing real-life call center data. We test the forecasting accuracy of the proposed models using forecasting lead times ranging from weeks to hours in advance.

Keywords: forecasting; arrival process; dynamic updating; time series; correlation; call centers.

1. Introduction

The service sector currently dominates the economic landscape of both emerging and developed economies. For example, the CIA World Factbook (2010) shows that, as of 2008, the service sector in Canada employs over three quarters of Canadians and accounts for over two thirds of the Gross Domestic Product (GDP). In broad terms, the service sector comprises businesses (systems) that produce a service instead of just an end product; e.g., call centers are service systems that provide support or sales services to callers. For background on call centers, see Mandelbaum (2002), Gans

et al. (2003), and Aksin et al. (2007).

Unlike tangible products, services are experienced and not consumed. 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.

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; see Gans et al. (2003), Avramidis et al. (2004), Brown et al. (2005), Shen and Huang (2008b), Aldor-Noiman et al. (2009), and references therein. 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 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 incoming demand (future call volumes), and to study ways of updating those forecasts at different points in time.

In this paper, we estimate and compare alternative time series models for forecasting future call volumes in a call center. We conduct an empirical study using real-life call center data, and test the forecasting accuracy of the proposed models using lead times ranging from weeks to

hours in advance. We do so to mimic the challenges faced by call center managers, as explained above. Our study shows the importance of accounting for correlations in the data when forecasting future arrivals. Here, we focus on three types of correlations: (i) interday (day-to-day), (ii) intraday (within day), and (iii) between the arrival processes of different call types. This work was motivated by an industry research project with a major company in Canada. In §1.1, we provide a brief description of that project.

1.1. Motivation

The company, hereafter referred to as Company X, operates several large call centers with thousands of agents serving over one hundred different call types and handling hundreds of thousands of incoming calls per day. The forecasting team of Company X uses time series methods (linear regression models) to generate forecasts for future *daily total* call volumes. In particular, the team generates forecasts for future daily totals two weeks in advance ("scheduling forecasts"), then updates those forecasts one day in advance ("last intraday forecast").

In order to obtain forecasts for specific half-hour intervals in the future (needed to make detailed agent schedules), the forecasting team uses a call center management software. The major concern of Company X is that the half-hourly forecasts thus generated are often unreliable. Indeed, one specific problem is that the software used does not take into account either interday or intraday dependence structures in the arrival process, which are typically significant; see §3. Instead, it performs a simple exponential smoothing of the data to generate the needed forecasts.

We were asked by Company X to: (i) develop accurate ways of forecasting future *half-hourly* call volumes; (ii) take into account both interday and intraday dependence structures, and allow for dynamically updating the forecasts; and (iii) develop ways of splitting the existing forecasts of daily totals into corresponding half-hourly forecasts. The reason behind using the existing forecasts is that they incorporate important information which impacts the arrival process (and is not in the data set), such as major marketing campaigns or recent price increases. For work on the impact of marketing campaigns on call center arrivals, see Soyer and Tarimcilar (2008). For ease of exposition, we focus in this work on forecasting arrivals for a single call type, say Type A, but our methods

can be easily applied to forecasting any call type handled at the call centers of Company X, or elsewhere.

1.2. Time Series Approach

The time series models that we consider in this paper address the concerns of Company X. Thus, they are appealing from a practical perspective. To capture interday and intraday correlations, we use a Gaussian linear mixed model. Mixed models were shown to generate accurate forecasts of future call volumes in Aldor-Noiman et al. (2009). For background on linear mixed models, see Muller and Stewart (2006). In the first part of this paper (§4 and §5), we compare the mixed model to three other models which do not account for any correlations in the data. Our results show the importance of accounting for interday and intraday correlations, especially when forecasting lead times are not too long. Here is a brief description of the alternative time series methods used; see §4 for details.

First, we consider a simple linear regression model with independent residuals. This model is equivalent to a historical average approach since it essentially uses past averages as forecasts of future call volumes; see Weinberg et al. (2007) and Shen and Huang (2008b). Moreover, it serves as a useful reference point because it does not incorporate any dependence structures in the data.

Second, we consider a Holt-Winters exponential smoothing technique, which is popular in forecasting seasonal time series. We take Holt-Winters smoothing to represent the current way of forecasting half-hourly arrival counts at Company X. It would have been ideal to use the company's actual half-hourly forecasts as reference, but those were unfortunately unavailable.

Third, we consider a method for breaking down the existing daily forecasts of Company X into half-hourly forecasts. In particular, we use the "Top-Down" approach, described in Gans et al. (2003), which splits the daily forecasts based on historical records for the proportion of calls in specific half-hour intervals of a given day.

In the second part of this paper (§6), we extend the mixed model into a bivariate mixed model which exploits the dependence structure between the arrival processes of two call types: Type A and Type B. Arrivals to the Type A queue originate in the province of Quebec, and are mainly

handled in French, whereas arrivals to the Type B queue originate in the province of Ontario, and are mainly handled in English. Otherwise, arrivals to both queues have similar service requests. Initial data analysis indicates a strong dependence between the arrival processes of Type A and Type B queues. In Figure 1, we present a scatter plot of the half-hourly arrival counts to each queue. (We first subtract from each arrival count the average count for the corresponding half-hour period.) Figure 1 shows that there is a significant positive correlation between the arrival processes of Type A and Type B. Indeed, a point estimate of this correlation is equal to 0.71. In §6.2, we show that the bivariate mixed model yields more accurate forecasts than the standard mixed model.

1.3. Main Contributions and Organization

Here are the two main contributions of this paper. First, we present a comparative study of several time series methods using real-life call center data. That is particularly important because there is relatively little empirical work on forecasting call center arrivals; see §2. Moreover, our models are especially appealing from a practical perspective. For example, both the "Top-Down" approach and exponential smoothing are forecasting methods commonly used in practice; therefore, it is important to study their forecasting accuracy.

Second, we show the importance of modeling different types of correlations in the data. In addition to accounting for interday and intraday correlations, we propose a new time series model which exploits correlations between the arrival processes of two separate queues. We show that this new bivariate model leads to more accurate forecasts. To the best of our knowledge, ours is the first work which proposes jointly modeling the arrival processes of different queues to generate forecasts of future call volumes.

The rest 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 candidate time series methods considered, and discuss how model parameters are estimated from data. In §5, we compare the alternative methods based on their forecasting accuracy. In §6, we introduce the bivariate mixed model and show that it leads to more accurate forecasts. In §7, we make concluding remarks.

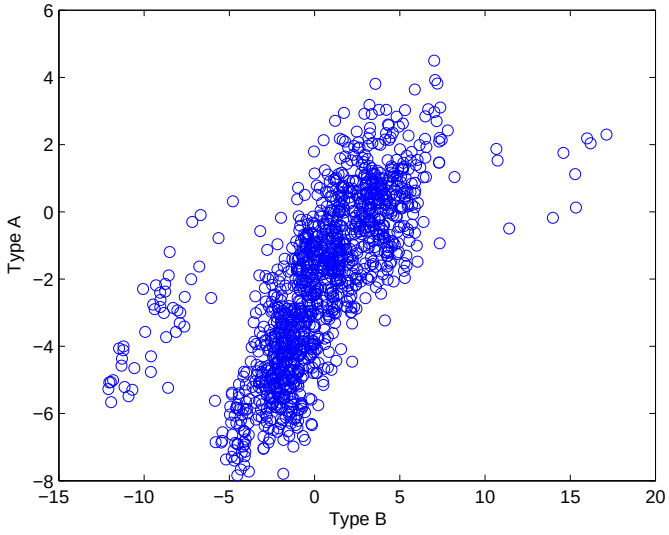

Figure 1: Scatter plot of half-hourly arrival counts (corrected for corresponding means and seasonality) to the Type A and Type B queues.

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, which we also consider in this paper.

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, with long forecasting lead times, simple forecasting techniques such as taking historical averages are difficult to beat. We reach a similar conclusion in this work as well. Shen and Huang (2008a) used

a Singular Value Decomposition (SVD) approach to forecast the time series of an inhomogeneous Poisson process by first building a factor model for the arrival rates, and then forecasting the time series of factor scores. Shen and Huang (2008b) used the same SVD idea 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 Poisson process approach incorporating day-of-week, periodic, and exogenous effects. We use a similar mixed model in this paper as well.

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 correlations between the arrivals in successive periods of the same day, 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 company in Canada. They were collected over 329 days (excluding days when the call center is closed, such as holidays and Sundays) ranging from October 19, 2009 to November 11, 2010. The data consist of arrival counts for the Type A queue whose incoming calls originate in the province of Quebec. In §6, we also consider call arrival data for a closely related queue, Type B, whose calls originate in the province of Ontario.

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 54 Saturdays from the data set, which leaves us with $D = 329 - 54 = 275$ remaining days. Arrival counts for each day are aggregated in consecutive time periods of length thirty minutes each. There are $P = 22$ thirty-minute periods on a weekday, and a total of $D \times P = 275 \times 22 = 6050$ observations in our 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, see Figures 2 and 3). Standard forecasting techniques generally produce unreliable forecasts for outlier days, since these are very different from the remaining days of the year. In practice, forecasts for outlier days are typically made off-line, and are largely based on the experience of call center managers. In a statistical study, we could: (i) remove outlier days altogether from the data set, or (ii) replace the arrival counts on outlier days by historical averages based on past data. Option (i) is only useful when the forecasting methods considered are able to tolerate gaps in the time series of arrival counts. In this work, we opt for (ii) to allow for more flexibility. As a result, we leave the natural seasonality in the data intact. In §5 and §6, we compare the forecasting accuracy of the alternative models in periods excluding all outlier days.

3.1. Overview

In Figure 2, we plot daily arrival volumes to the call center between October 19, 2009 and October 19, 2010 (one year). Figure 2 shows that there exist some outlier days with unusually low call volumes, e.g., on 05/24/2010 which is day 157 in the plot. Figure 2 also shows that there may exist monthly fluctuations. In particular, the moving average line in the plot (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, which correspond to days 54-93 in the plot. (This was also confirmed by the company.) We do not incorporate a month-of-year effect into our models because of insufficient data. Indeed, if several years of half-hourly data had been available, then we would expect to see a significant monthly seasonality.

In Figure 3, we present a box plot of the daily arrival volume for each day of the week. The

boxes have lines at the lower quartile, median, and upper quartile values. The whiskers are lines extending from each end of the boxes to show the extent of the rest of the data. Outliers are data with values beyond the ends of the whiskers. Figure 3 clearly shows that Mondays have a higher call arrival volume relative to the remaining weekdays. We use Figures 2 and 3 to identify 9 outlier days which we replace with historical averages.

In Figure 4, we plot the average arrival count per period for each weekday, after smoothing the data. Figure 4 shows that all weekdays have a similar intraday profile: There are two major peaks in call arrival volumes during the day. 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). Upon closer inspection, Figure 4 shows that the average number of arrivals per period is nearly identical on Wednesdays and Thursdays. That is confirmed by hypothesis tests for statistical significance, detailed in §4.1. We also see that Tuesdays have a slightly different morning arrival pattern relative to the remaining weekdays, whereas Fridays have a different afternoon arrival pattern (between 2:00 PM and 4:00 PM).

3.2. Interday and Intraday Correlations

Exploratory analysis of our data set shows evidence of: (i) strong positive (interday) correlations between arrival counts over successive days, and (ii) strong positive (intraday) correlations between arrival counts over successive periods of the same day. In §4, we take both interday and intraday correlations into account when making forecasts of future call volumes.

In Table 1, we present estimates of correlations between daily arrival volumes over successive weekdays. (We first subtract from each daily total the average arrival volume for the corresponding weekday.) Table 1 shows that there are significant positive correlations between days of the same week. In particular, correlations are strong between successive weekdays, and are slightly smaller with longer lags; e.g., the correlation between (the 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.

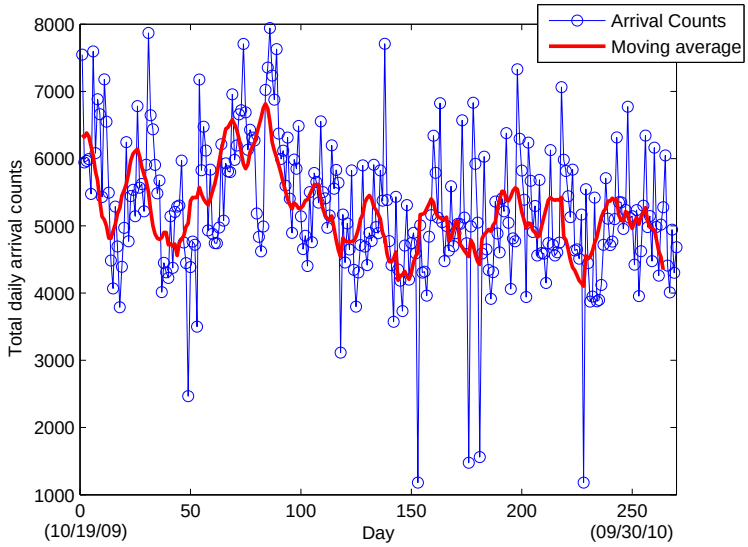

Figure 2: Daily arrivals to the Type A queue between October 19, 2009 and October 19, 2010.

	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 (corrected for seasonal trends) on successive weekdays for the Type A queue.

Additionally, Table 1 shows that Mondays are less correlated with the remaining weekdays; e.g., the correlation between Monday and Tuesday is 0.48.

There are also strong intraday correlations in the data set. Here, we plot estimates of intraday correlations on Wednesday, but similar patterns hold for all weekdays as well. More precisely, in Figure 5, we plot correlations between the fifth period on Wednesday and the remaining periods of that day; e.g., the correlation between periods 5 and 7 on Wednesday is equal to 0.9, and the correlation between periods 5 and 20 is roughly equal to 0.6.

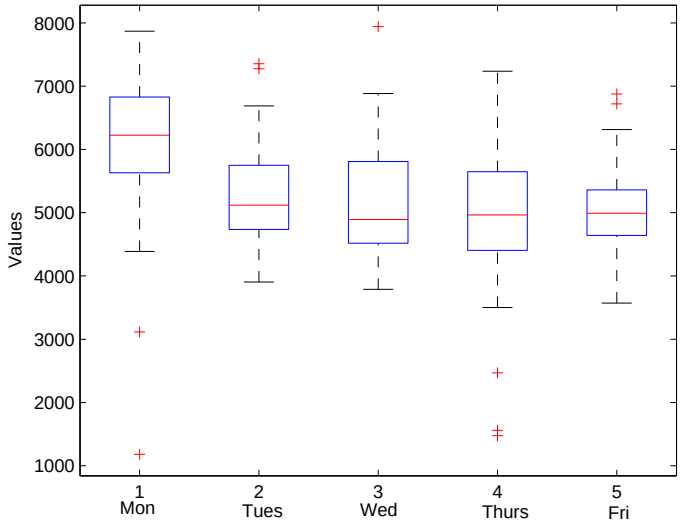

Figure 3: Box plot of daily arrival counts to the Type A queue for each weekday.

3.3. Data Transformation

Let $N_{i,j}$ be the number of arrivals in the j^{th} period of day i, where $1 \leq i \leq D$ and $1 \leq j \leq P$. 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 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 models to the transformed data. In the mixed model, we assume that $\sqrt{\lambda_{i,j}}$ is a linear combination of both fixed and random effects.

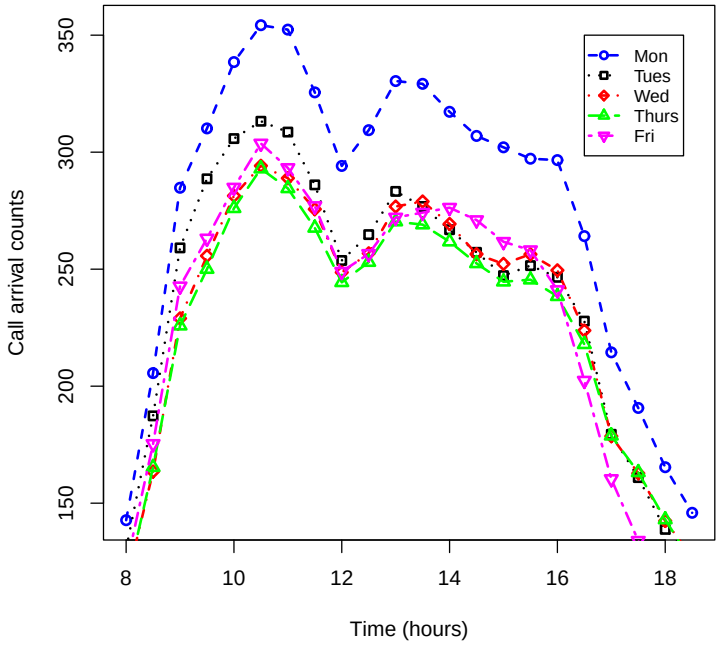

Figure 4: Intraday arrival patterns for each weekday, for the Type A queue, averaged over all arrival counts between October 19, 2009 and November 11, 2010.

The goodness of fit of the Gaussian model, described in §5.4, indicates that the Poisson model is appropriate in our context.

4. Time Series Models

In this section, we describe candidate time series models for the arrival process and discuss how different model parameters are estimated from data. In §5, we compare the alternative models based primarily on forecasting performance.

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

The preliminary data analysis of §3 shows 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 Figure 4. We capture those two properties in our first time series

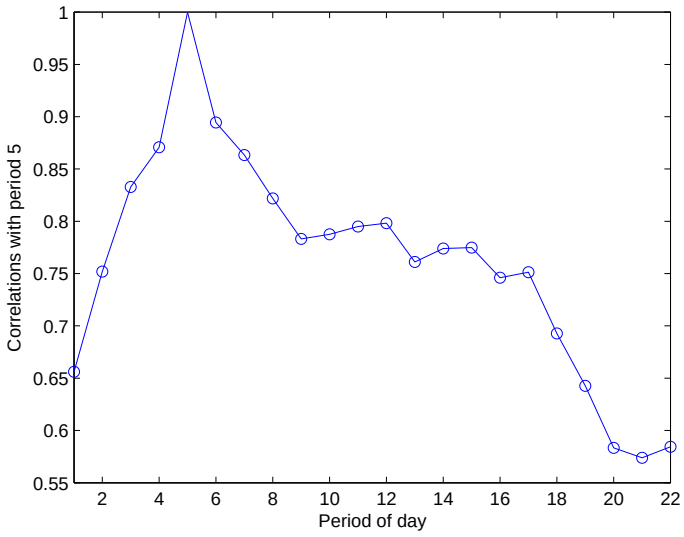

Figure 5: Correlations between period 5 and remaining periods on Wednesday for Type A call arrivals.

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, consistently with Figure 4. We consider the Fixed-Effects (FE) model because similar models are often used in the call center industry; e.g., see Weinberg et al. (2007) and Shen and Huang (2008b). 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. Moreover, the FE model serves as a useful reference point because it does not incorporate any dependence structures, such as interday and intraday correlations; see Table 1 and Figure 5.

Let d_i be the day-of-week of day i, where $1 \leq i \leq D$. (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.) Let j denote the half-hour period index in day i, where $1 \leq j \leq P$. Recall that $D = 275$ is the number of days in our data set, and $P = 22$ is the total number of half-hour periods per day.

We model $y_{i,j}$, the square-root transformed call volume in period j of day i, as:

$$y_{i,j} = \sum_{k=1}^{5} \alpha_k I_{d_i}^k + \sum_{l=1}^{22} \beta_l I_j^l + \sum_{k=1}^{5} \sum_{l=1}^{22} \theta_{k,l} I_{d_i}^k I_j^l + \epsilon_{i,j} \ , \tag{4.1}$$

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 $\epsilon_{i,j}$ are independent and identically distributed (i.i.d.) normal random variables with mean 0 and variance σ_ϵ^2. The normality assumption enables us to obtain prediction intervals for future observations; see §5. Equation (4.1) simplifies to

$$y_{i,j} = \alpha_{d_i} + \beta_j + \theta_{d_i,j} + \epsilon_{i,j} \ . \tag{4.2}$$

We estimate model parameters using the method of least squares. It is well known that least squares estimates are equivalent to maximum likelihood estimates with normal i.i.d. residuals, as in (4.1). We perform t-tests for the significance of the fixed effects in (4.1) at the 90% level, i.e., significant parameters have an associated p-value which is less than 0.10. Motivated by the principle of parsimony in statistical modeling, we exclude from (4.1) some terms with no statistical significance. In Table 2, we summarize some results based on fitting the FE model to the entire data set. Table 2 shows that the 5 weekdays have significantly different effects, i.e., each weekday has a different expected daily call volume. In contrast, not all periods have significant effects. In particular, our results indicate that the effects of 6 out of the 22 periods in a day (namely periods 4, 10, 14, 15, 16, and 17) are not statistically significant. Consistent with intuition, periods 1 (first) and 22 (last) have significant negative effects on the expected number of incoming calls. Since some period indicators are strongly significant, whereas others are not, we retain in our FE model all the period-of-day effects. Our results also show that none of the indicators corresponding to cross terms between Thursday and period-of-day, and Wednesday and period-of-day, is statistically significant. Therefore, those cross terms are removed from the model. We keep all cross terms between a specific weekday and the period-of-day if at least one of those terms is statistically significant. For example,

Table 2 shows that the indicator for period 4 on Tuesday is statistically significant whereas the indicator for period 8 on Tuesday is not. As a result, we do not exclude from our model any cross terms corresponding to the interaction between Tuesday and period-of-day. Similarly, we do not exclude any cross terms corresponding to the interaction between Monday and period-of-day, and Friday and period-of-day.

Category	Coefficient	Std. error	p-value
Monday	17.2	0.210	< 0.0001
Tuesday	15.8	0.205	< 0.0001
Wednesday	15.6	0.205	< 0.0001
Thursday	15.4	0.207	< 0.0001
Friday	15.6	0.207	<0.0001
Period 1	-4.96	0.289	<0.0001
Period 4	0.222	0.289	0.443
Period 10	0.249	0.289	0.390
Period 18	-0.845	0.289	0.00350
Period 22	-4.47	0.289	< 0.0001
Monday $\times$ Period 13	-0.00415	0.415	0.992
Monday $\times$ Period 22	-0.833	0.415	0.0447
Tuesday $\times$ Period 4	0.845	0.409	0.0391
Tuesday $\times$ Period 8	0.142	0.409	0.728
Wednesday $\times$ Period 11	0.119	0.417	0.773
Wednesday $\times$ Period 14	-0.0291	0.417	0.942
Thursday $\times$ Period 1	0.214	0.411	0.603
Thursday $\times$ Period 10	-0.0106	0.411	0.980
Friday $\times$ Period 1	0.0524	0.413	0.899
Friday $\times$ Period 20	-1.37	0.413	< 0.0001

Table 2: Partial results for the Fixed-Effects model specified in (4.1). Point estimates of model coefficients, α, β, and θ, are shown with corresponding standard errors and p-values of t-tests for statistical significance.

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 a linear mixed model incorporating both fixed and random effects. We consider the same fixed effects as in (4.1). 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 use the Mixed Procedure in the SAS® software to implement the linear mixed model, and compute maximum likelihood estimates of all model parameters.

For ease of representation, we formulate the linear mixed model in matrix form; see (4.7). Mixed models have been previously considered to model call center arrivals. In particular, the model that we describe in this subsection has been proposed by Aldor-Noiman et al. (2009). We consider Gaussian linear mixed models because they are relatively easy to implement and interpret, and they allow for interactions (dependencies) between the different covariates. They have also been previously shown to generate accurate forecasts of future call volumes. The Gaussian linear mixed-effects model, henceforth referred to as the ME model, exploits the normal approximation of the square-root transformed data counts, which is reasonable for large call volumes; see §3.3.

4.2.1. Random Effects.

We use random effects $\gamma = (\gamma_1, \gamma_2, ..., \gamma_D)'$ to model the normal daily volume deviation from the fixed weekday effect. (This means that we add a random effect γ_i to the right-hand side of (4.2).) Let G denote the $D \times D$ covariance matrix of the random effects γ. We assume that G has a first-order autoregressive covariance structure, AR(1), i.e., we assume that the covariance between γ_i and γ_j is given by:

$$\text{cov}(\gamma_i, \gamma_j) = g_{i,j} = \sigma_G^2 \rho_G^{|i-j|} \quad \text{for} \quad 1 \le i, j \le D \ , \tag{4.3}$$

where σ_G^2 is the variance parameter and 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. Some care need be taken to preserve the true numerical distance between the days. We do so 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.

We also considered more complex covariance structures for G, such as an autoregressive-moving-average structure, ARMA(1,1), and a heterogeneous autoregressive structure, ARH(1). For an ARMA(1,1) structure, we assume that:

$$g_{i,j} = \begin{cases} \sigma_G^2 \cdot \delta \cdot \rho_G^{|i-j|-1} & \text{if } i \neq j \ , \\ \sigma_G^2 & \text{if } i = j \ , \end{cases} \tag{4.4}$$

which requires the estimation of three parameters: σ_G, ρ_G, and δ. And, for an ARH(1) structure, we assume that:

$$g_{i,j} = \sigma_i \sigma_j \rho_G^{|i-j|} \quad \text{for} \quad 1 \leq i, j \leq D \ , \tag{4.5}$$

where σ_i and σ_j may not be equal. That is, an ARH(1) structure for G requires the estimation of $D+1$ parameters. Unfortunately, both the ARMA(1,1) and ARH(1) structures are so computationally intensive that the corresponding mixed models could not be fit to data (we ran into computer memory restrictions). Thus, we focus solely in this work on an AR(1) covariance structure for G.

Using the terminology of standard linear models, we denote by Z the design matrix for the random effects, i.e., $Z = I_D \otimes \mathbf{1_P}$ where I_D is the $D-$dimensional identity matrix, $\mathbf{1_P} = (1, 1, ..., 1)' \in \mathbb{R}^P$ is the $P-$dimensional vector of 1's, and $\otimes$ denotes the Kronecker product.

4.2.2. Model Residuals.

Let ϵ denote the $DP-$dimensional vector of residuals of the model, i.e., the Gaussian deviations at the periodic level of the square-root transformed data counts after accounting for both fixed and random effects. Let R denote the within-day $P \times P$ covariance matrix of residuals. We assume that

$$R = R_0 + \sigma^2 I_P \ , \tag{4.6}$$

where R_0 has an AR(1) covariance structure with variance parameter $\sigma_{R_0}^2$ and autocorrelation parameter ρ_{R_0}; see (4.3). We let σ^2 be the residual variance, and I_P be the $P-$dimensional identity matrix. If the mixed model is a good fit for the data, then the estimated value of σ^2 should be roughly equal to the theoretical value of 0.25. In §5.4, we show that this is indeed the case.

In addition to AR(1), we also considered an ARMA(1,1) covariance structure for R_0. We found that the difference in performance between the two models is not statistically significant. Therefore, we focus solely on an AR(1) structure for R_0 in this work. There remains to formulate the ME model, which we do next.

4.2.3. Mixed Model Formulation.

Let W be a $D \times 5$ matrix with $(u, v)^{th}$ entry $w_{u,v} = 1$ if $d_u = v$, and 0 otherwise. Let X_D be the $DP \times 5$ design matrix for the fixed daily effects. Then, $X_D = W \otimes \mathbf{1_P}$. Let X_P be the $DP \times P$ design matrix for the fixed period-by-period effects. Then, $X_P = W \otimes I_P$. Finally, let X_{DP} be the design matrix for the cross terms between the day-of-week and period-of-day effects. Then, X_{DP} has DP rows and $3P$ columns (recall that we removed from the model the cross terms corresponding to both Wednesday and Thursday). The model for the square-root transformed vector of arrival counts $y = (y_{1,1}, ..., y_{D,P})'$ is:

$$y = X_D\alpha + X_P\beta + X_{DP}\theta + Z\gamma + \epsilon \ , \tag{4.7}$$

where the covariance matrix of γ is $\mathrm{Var}(\gamma) = G$, and the covariance matrix of ϵ is $\mathrm{Var}(\epsilon) = I_D \otimes R = I_D \otimes (R_0 + \sigma^2 I_P)$. We assume that both γ and ϵ have expected values $E[\gamma] = E[\epsilon] = 0$, and that they are independent. In Table 3, we present some results for the statistical significance of the daily random effects γ_i for selected days i, obtained when fitting the mixed model in (4.7) to the entire data set. In Table 4, we present corresponding estimates of the covariance parameters σ^2_G, ρ_G, $\sigma^2_{R_0}$, ρ_{R_0}, and σ^2, specified in (4.3) and (4.6). It is interesting to note the strong positive interday and intraday correlations, which are indicated by the high values of ρ_G and ρ_{R_0}. For more details on the estimation of linear mixed models, see Henderson (1975).

4.3. Holt-Winters (HW) Smoothing

Exponential smoothing is a popular forecasting technique, commonly used in call centers, where the forecast is constructed from an exponentially weighted average of past observations. Therefore, it is important to study the accuracy of this method. The Holt-Winters (HW) method is an extension

Category	Coefficient	Std. error	p-value
Day 1	1.63	0.323	< 0.0001
Day 10	2.38	0.318	< 0.0001
Day 20	-1.16	0.322	< 0.0003
Day 30	-0.448	0.318	0.159
Day 40	1.25	0.321	< 0.0001
Day 50	-0.695	0.322	0.0311
Day 60	0.562	0.318	0.0772
Day 71	-2.61	0.322	<0.0001
Day 75	0.0702	0.324	0.828

Table 3: Partial results for the Mixed Model, specified in (4.7), fit to the entire data set. Point estimates of the random effects, γ, are shown along with corresponding standard errors and p-values of t-tests for statistical significance.

σ^2_G	1.59
ρ_G	0.652
$\sigma^2_{R_0}$	0.391
ρ_{R_0}	0.619
σ^2	0.278

Table 4: Covariance parameter estimates for the Mixed Model in (4.7) when fit to the entire data set.

of exponential smoothing which accommodates both a trend and a seasonal pattern; see Winters (1960). The HW method has two versions, additive and multiplicative, the use of which depends on the characteristics of the particular time series. Here, we apply the logarithmic transformation to the original time series of arrival counts, which makes the additive version of HW smoothing appropriate for use. As shown in Figure 4, there is a clear intraday pattern in the arrival process of calls. Therefore, we incorporate a daily seasonality in our model; see (4.8). Since we only have arrival data for about one year, we exclude the presence of a trend.

Taylor (2008) extended the HW method to accommodate two seasonalities; e.g., weekly and daily. Even though this extension may be more appropriate for our call center data, where we found evidence of both daily and weekly seasonal patterns, we restrict attention here to the HW method with a single daily seasonality. (That is, the length of the seasonal cycle is equal to P.) We do so to mimic current practice at Company X. Indeed, we were told that the present way of generating half-hourly forecasts at the company is via a simple smoothing of data, which does

not account for multiple seasonal cycles. We take the forecasts resulting from HW smoothing to represent the current half-hourly forecasts at Company X. We implement the HW method using the corresponding function in the R statistical software.

For ease of representation, we now use a slightly different notation for indexing the time series. In particular, let $\{N_1, N_2, ..., N_{DP}\}$ denote the successive half-hourly arrival counts. For example, $N_{P+1} = N_{23}$ is the number of arrivals in the first half-hour interval of the second day in our data set. Let x_t denote the logarithm-transformed data count, i.e., $x_t = \log(N_t)$, for $t = 1, 2, .., DP$. Here are the smoothing equations for the HW method:

$$\begin{aligned} S_t &= \nu(x_t - D_{t-P}) + (1-\nu)S_{t-1} \ , \\ D_t &= \delta(x_t - S_t) + (1-\delta)D_{t-P} \ , \\ \hat{x}_t(k) &= S_t + D_{t-P+k} \ , \end{aligned} \tag{4.8}$$

where S_t is the smoothed level, D_t is the seasonal index for the intraday cycle, and $\hat{x}_t(k)$ is the $k-$step-ahead forecast at time t. For simplicity, the forecast in (4.8) is based on having $k \leq P$, but it is an easy task to rewrite it for longer forecasting lead times. The constants ν and δ are smoothing parameters, whose values are between 0 and 1, which are chosen to minimize the one-step ahead in-sample forecast errors ranging from October 19, 2009 to August 18, 2010 (a total of 4730 data points). Chatfield and Yar (1988) showed the importance of the choice of initial values for the smoothed level, S_t, and seasonal index, D_t. Here, we choose those values by averaging observations from October 19, 2009 to August 18, 2010. In §5, we compare the forecasting accuracy of all models based on out-of-sample forecasts ranging from August 19, 2010 to November 11, 2010.

4.4. The "Top-Down" (TD) Approach

As explained in Gans et al. (2003) and Taylor (2008), the "Top-Down" (TD) approach is very commonly used in call centers to forecast future arrival volumes. Applied to our context, this approach splits forecasts of total daily arrival volumes into forecasts for half-hourly intervals based

on estimates of historical proportions of calls in successive half-hourly intervals of the day. The TD approach is especially useful in our context because it allows us to use the daily forecasts currently generated by Company X. Indeed, those forecasts are based on information which affects the arrival process but is not in the data, such as major marketing campaigns or recent price increases.

In Table 5, we present summary statistics for the historical proportion of calls in different (selected) half-hour intervals on Tuesdays. Results for other periods on Tuesdays, and for the remaining weekdays, are largely similar and are therefore not presented separately. In Table 5, we present point estimates for the mean, the variance, the median, and the first and third quartiles of the historical proportions on Tuesdays. Table 5 shows that those proportions are not highly variable. For example, the ratio between the point estimates of the variance and the square of the mean in period 8 on Tuesday is roughly equal to 0.006. It is worthwhile noting that the TD approach is consistent with modeling the multivariate distribution of arrival counts in successive periods of a day by a multinomial distribution; see Channouf et al. (2007).

Here are the equations for breaking down the daily forecasts, made by Company X, into interval forecasts. Let $\hat{y}_i$ denote the forecast for the number of arrivals on day i, $1 \leq i \leq D$. Let $\hat{y}_{i,j}$ denote the split forecast for period j on day i, $1 \leq j \leq P$. As in §4.1, let d_i denote the day-of-week of day i. Then,

$$\hat{y}_{i,j} = \hat{y}_i \times \hat{p}_{d_i,j} \ , \tag{4.9}$$

where $\hat{p}_{d_i,j}$ is the point estimate of proportion of calls in half-hour period j of day type d_i. The proportions $\hat{p}_{d_i,j}$ are the historical proportions of calls that fall in a given period of a day, relative to the total number of calls on that day. More precisely,

$$\hat{p}_{d_i,j} = \frac{1}{n_{d_i}} \sum_{k=1}^{D} \frac{I_{d_i}^k y_{k,j}}{\sum_{l=1}^{P} y_{k,l}} \ , \tag{4.10}$$

where n_{d_i} is the number of days of type d_i in our data set, and $I_{d_i}^k$ is the indicator for day of type d_i, as in (4.1).

Category	Mean	Variance	Median	1st quartile	3rd quartile
Period 1	0.0249	1.74×10^{-5}	0.0245	0.0222	0.0272
Period 5	0.0580	2.52×10^{-5}	0.0575	0.0540	0.0612
Period 8	0.0543	1.95×10^{-5}	0.0545	0.0516	0.0573
Period 10	0.0502	1.03×10^{-5}	0.0503	0.0484	0.0523
Period 15	0.0470	1.20×10^{-5}	0.0469	0.0451	0.0497
Period 17	0.0467	1.61×10^{-5}	0.0467	0.0443	0.0493
Period 22	0.0243	6.32×10^{-6}	0.0242	0.0230	0.0260

Table 5: Selected historical proportions on Tuesdays, as specified in (4.10), computed based on the entire data set.

5. Model Comparison

In this section, we compare the alternative time series models of §4 based on their forecasting performance. 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.

5.1. 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 85 days, and generate $85 \times 22 = 1320$ predicted values. We consider three forecasting lead times to mimic the challenges faced by real-life call center managers. In particular, we consider lead times of 2 weeks, 1 week, and 1 day. In §6.2, we also consider within-day forecasting updates. In this section, we let the learning period include all days in the data set, up to the beginning of the forecast lag. 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 model parameters after each daily forecast.

5.2. Performance Measures

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

$$\mathrm{SE}_{i,j} \equiv (N_{i,j} - \hat{N}_{i,j})^2 \ , \tag{5.1}$$

where $N_{i,j}$ is the number of arrivals in the j^{th} period of a given day i, and $\hat{N}_{i,j}$ is the predicted value of $N_{i,j}$. We also compute the *relative error* (RE), defined by:

$$\mathrm{RE}_{i,j} = 100 \cdot \frac{|N_{i,j} - \hat{N}_{i,j}|}{N_{i,j}} \ . \tag{5.2}$$

We then average $\mathrm{SE}_{i,j}$ and $\mathrm{RE}_{i,j}$ over all half-hour periods of day i, for each i. In particular, we define the *root-average-squared error* for day i, RASE_i, given by:

$$\mathrm{RASE}_i = \sqrt{\frac{1}{P}\sum_{j=1}^{P} \mathrm{SE}_{i,j}} \ , \tag{5.3}$$

and the *average percentage error* for day i, APE_i, given by:

$$\mathrm{APE}_i = \frac{1}{P}\sum_{j=1}^{P} \mathrm{RE}_{i,j} \ . \tag{5.4}$$

The RASE is an empirical version of the *root mean-squared error*, RMSE. We repeat this procedure for all 85 days in our forecasting horizon. As a result, we have 85 daily values for both RASE_i and APE_i. We then compute point estimates for the mean, median, 1st and 3rd quartiles of the resulting RASE_i and APE_i values.

We also use performance measures to evaluate the prediction intervals for $N_{i,j}$, generated by the FE and ME models. In particular, we define the "Cover" of the prediction interval for $N_{i,j}$ as:

$$\mathrm{Cover}_{i,j} = I(N_{i,j} \in (L_{i,j}, U_{i,j})) \ , \tag{5.5}$$

where $I(\cdot)$ denotes the indicator random variable, and $L_{i,j}$ and $U_{i,j}$ are the lower and upper bounds of the prediction interval, respectively. We also compute the "Width" of the confidence interval, defined as:

$$\mathrm{Width}_{i,j} = U_{i,j} - L_{i,j} \ . \tag{5.6}$$

For each day i in the forecasting horizon, we compute the average coverage probability, $\mathrm{Coverage}_i$, defined as the average over all periods j in day i of the indicators in (5.5). Similarly, we define

Width_i as the average of $\text{Width}_{i,j}$ in (5.6), over all periods j in day i. As a result, we obtain 85 values for both Coverage_i and Width_i, corresponding to all days in our forecasting horizon. Then, we compute the mean, median, 1st and 3rd quartiles of the resulting Coverage_i and Width_i values; see Tables 6-8.

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 coverage probability be close to 95%. Since we are unable to compute prediction intervals for both the TD approach (no confidence intervals generated by Company X), and for the HW method (for which we make no modeling assumptions), we leave the corresponding entries for Coverage and Width empty in Tables 6, 7, and 8.

5.3. Forecasting Performance

5.3.1. Two Weeks-Ahead Forecasts.

In Table 6, we present point estimates of the performance measures described in §5.2, for each of the candidate models, with a forecasting lead time of 2 weeks. Table 6 shows that the FE model generates the most accurate forecasts in that case. Consistent with Taylor (2008), this shows that with long lead times a simple historical average is difficult to beat. The ME model performs worse than the FE model in this case. Indeed, RASE(ME)/RASE(FE) is roughly equal to 1.04, whereas APE(ME)/APE(FE) is roughly equal to 1.08. Although this may seem counterintuitive at first, since the ME model uses the same fixed effects as the FE model, it can be readily seen that the difference in performance between the two models is due to the fact that the ME model over-fits the data; e.g., the ME model requires estimating a random effect for each day in the learning period, in addition to 4 covariance parameters for G and R, as explained §4.2.1 and §4.2.2.

Table 6 also shows that the TD approach, based on 2-weeks-ahead forecasts made by Company X, is outperformed by both the FE and ME models. For example, RASE(TD)/RASE(FE) is roughly equal to 1.14, and APE(TD)/APE(FE) is roughly equal to 1.19. Finally, the HW smoothing approach yields disappointing results. Indeed, RASE(HW)/RASE(FE) is roughly equal to 1.66 and APE(HW)/APE(FE) is roughly equal to 1.90. That is also consistent with results in Taylor (2008),

where exponential smoothing was found to be ineffective, particularly for long lead times. That is especially interesting because exponential smoothing is the way that half-hour interval forecasts are currently generated at Company X.

It is insightful to compare the Width and Coverage of the prediction intervals for the FE and ME models, respectively. Indeed, both the Width and Coverage of prediction intervals for the FE model are considerably smaller than those for the ME model. For example, Coverage(FE) is roughly equal to 0.22, whereas Coverage(ME) is roughly equal to 0.95, as desired. Additionally, prediction intervals for the ME model are roughly 8 times larger, in an average sense, than those for the FE model. The Width and Coverage reveal that the FE model underestimates uncertainty in the data by not capturing the correlation structure between the arrival counts.

5.3.2. One Week-Ahead Forecasts.

In Table 7, we present results corresponding to a forecasting lead time of 1 week. We do not discuss those results separately because they are largely consistent with those obtained with a forecasting lead time of 2 weeks. However, it is interesting to note that the resulting forecasts are only very slightly more accurate than with 2-week-ahead forecasts. That is consistent with results in Aldor-Noiman et al. (2009), who found that shorter lead times do not always lead to more accurate forecasts. As pointed out by the authors of that paper, this observation may help call center managers decide whether they need to update their forecasts one week in advance. Since Company X does not update its forecasts one week in advance, we do not include forecasts for the TD approach in Table 7.

5.3.3. One Day-Ahead Forecasts.

In Table 8, we present results for a forecasting lead time of one day. As expected, the superiority of the ME model is clearly evident in this case. In particular, it is considerably more accurate than the FE model, which fails to model the interday and intraday correlations in the arrival counts. Indeed, RASE(FE)/RASE(ME) is roughly equal to 1.17 in this case, and APE(FE)/APE(ME) is roughly equal to 1.18. Table 8 shows that the TD approach is competitive. Indeed, this method

generates the second most accurate forecasts, after the ME model; e.g., APE(TD)/APE(ME) is roughly equal to 1.07. This indicates that the daily forecasts of Company X are relatively accurate, and that there is a justification behind adopting a simple splitting technique of those forecasts, as with the TD approach.

With a lead time of one day, HW smoothing generates, once more, the least accurate forecasts. For example, APE(HW)/APE(ME) is roughly equal to 2. With a lead time of one day, all models generate more accurate forecasts compared with the longer lead times of two weeks. For example, APE(ME) decreases from 16.4% to 12.9% (roughly a 30% decrease). As in Tables 6 and 7, the Coverage and Width of the prediction intervals for the ME model show that it correctly captures the correlation structure in the data.

		ME model	TD approach	FE model	HW smoothing
RASE	Mean	41.7	45.8	40.3	67.0
	Median	29.5	33.9	32.0	57.9
	1st quartile	21.4	24.3	22.0	47.2
	3rd quartile	41.8	45.1	43.8	82.1
APE	Mean	16.4	18.2	15.3	29.0
	Median	11.2	13.7	11.9	23.3
	1st quartile	7.87	9.19	8.62	18.5
	3rd quartile	16.3	17.3	17.6	34.1
Coverage Probability	Mean	0.946	-	0.220	-
	Median	1	-	0.182	-
	1st quartile	1	-	0.0909	-
	3rd quartile	1	-	0.363	-
Average Width	Mean	182	-	22.4	-
	Median	180	-	23.3	-
	1st quartile	176	-	18.3	-
	3rd quartile	184	-	25.4	-

Table 6: Comparison of the forecasting accuracy of the alternative time series methods, for a forecasting lead time of two weeks, based on out-of-sample forecasts between August 19, 2010 and November 11, 2010. Point estimates for the performance measures described in §5.2 are shown.

5.4. Goodness of Fit of the Mixed Model

The empirical results of this section showed that the ME model yields the most accurate forecasts, for short forecasting lead times, among all methods considered. We conclude this section by com-

		ME model	FE model	HW smoothing
RASE	Mean	41.5	40.1	65.4
	Median	29.2	31.4	59.7
	1st quartile	21.6	22.6	41.8
	3rd quartile	41.4	43.4	76.6
APE	Mean	16.3	15.2	28.1
	Median	10.9	12.0	25.4
	1st quartile	7.87	8.70	17.0
	3rd quartile	16.4	17.2	32.7
Coverage Probability	Mean	0.946	0.220	-
	Median	1	0.205	-
	1st quartile	1	0.0909	-
	3rd quartile	1	0.330	-
Average Width	Mean	180	22.04	-
	Median	178	23.4	-
	1st quartile	175	18.2	-
	3rd quartile	183	25.5	-

Table 7: Comparison of the forecasting accuracy of the alternative time series methods, for a forecasting lead time of one week, based on out-of-sample forecasts between August 19, 2010 and November 11, 2010. Point estimates for the performance measures described in §5.2 are shown.

menting on the goodness of fit of the ME model. In Figure 6, we present a QQ-plot of the residuals of the ME model which are obtained after fitting the model to the entire data set. Figure 6 shows that the assumption of normality is reasonable.

Consistent with §3.3, we found that the average estimated value of σ^2 is equal to 0.27, which is close to the theoretical values of 0.25. Thus, the ME model adequately captures the predictive structure in the data, and what remains is unpredictable variation. To get a sense of the magnitude of this variance, we need to compare it to the square root of the average number of arrivals per period. Since that number is in the tens of calls, we conclude that the obtained value of σ^2 is small.

6. A Bivariate Mixed (BM) Model

In this section, we describe a bivariate linear mixed (BM) model, which extends the linear mixed model of §4.2, to jointly model the arrival processes of the Type A and Type B queues.

As explained in §1, calls arriving to the Type A queue originate in the province of Quebec,

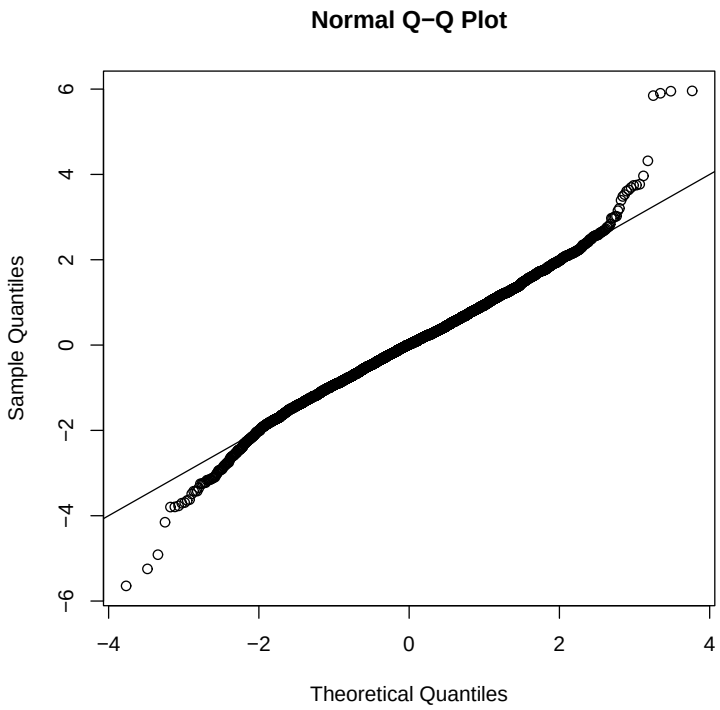

Figure 6: QQ plot for the residuals of the Mixed model specified in (4.7), when fit to the entire data set. Also included is a straight line with slope 1.

		ME model	TD approach	FE model	HW smoothing
RASE	Mean	30.4	33.9	35.7	60.7
	Median	27.1	30.3	30.9	48.9
	1st quartile	19.8	22.9	22.9	35.99
	3rd quartile	34.5	41.0	42.8	82.4
APE	Mean	12.9	13.9	15.1	26.4
	Median	10.1	12.5	11.9	20.8
	1st quartile	7.26	8.23	8.67	15.7
	3rd quartile	14.4	15.5	17.1	33.9
Coverage Probability	Mean	0.962	-	0.219	-
	Median	1	-	0.227	-
	1st quartile	1	-	0.0909	-
	3rd quartile	1	-	0.329	-
Average Width	Mean	157	-	21.9	-
	Median	151	-	22.9	-
	1st quartile	147	-	17.8	-
	3rd quartile	160	-	25.4	-

Table 8: Comparison of the forecasting accuracy of the alternative time series methods, for a forecasting lead time of one day, based on out-of-sample forecasts between August 19, 2010 and November 11, 2010. Point estimates for the performance measures described in §5.2 are shown.

and are mainly handled in French, whereas calls for the Type B queue originate in the province of Ontario, and are mainly handled in English. Otherwise, callers to both queues have similar service requests. Therefore, we expect their respective arrival streams to be positively correlated. Here, we propose exploiting this positive correlation to generate more accurate forecasts for Type A arrivals. (For simplicity, we restrict attention to forecasting arrivals for the Type A queue, but similar results also hold for the Type B queue.) Consistent with intuition, we showed in Figure 1 that the arrival streams for Type A and Type B are strongly positively correlated. Indeed, this correlation is estimated at 0.71.

Bivariate linear mixed 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). We fit bivariate mixed models to our data using the Mixed Procedure in SAS®.

6.1. Model Formulation

Let $y_{i,j}^B$ ($y_{i,j}^A$) denote the square-root transformed arrival count of Type B (Type A) in period j of day i, where $1 \leq i \leq D$ and $1 \leq j \leq P$. Let the $P-$dimensional vectors of arrival counts on day i for Type B and Type A, respectively, be given by:

$$y_i^B = (y_{i,1}^B, y_{i,2}^B ..., y_{i,P}^B)' \ , \tag{6.1}$$

and

$$y_i^A = (y_{i,1}^A, y_{i,2}^A ..., y_{i,P}^A)' \ . \tag{6.2}$$

We model the joint distribution of y^B and y^A where $y^B = (y_1^B, ..., y_P^B)'$ and $y^A = (y_1^A, ..., y_P^A)'$. As in the standard mixed model, the bivariate mixed model has both fixed and random effects. We use the same fixed and random effects as in §4.1 and §4.2. In particular, we assume that:

$$y^B = X_D\alpha^B + X_P\beta^B + X_{DP}\theta^B + Z\gamma^B + \epsilon^B \ , \tag{6.3}$$

$$y^A = X_D\alpha^A + X_P\beta^A + X_{DP}\theta^A + Z\gamma^A + \epsilon^A \ ; \tag{6.4}$$

using the same notation as in (4.1) and (4.7). The random effects in (6.3) and (6.4) are chosen as in §4.2.1. That is, γ^B and γ^A denote the $D-$dimensional vector of Gaussian deviates from the daily fixed effects. As in §4.2.1, we assume AR(1) structures for the covariance matrices G^B (of γ^B) and G^A (of γ^A). We also assume that γ^B and γ^A are independent. That is, we assume that:

$$\text{cov}(\gamma_i^B, \gamma_j^B) = g_{i,j}^B = \sigma_{G,B}^2 \rho_B^{|i-j|} \quad \text{for} \quad 1 \leq i, j \leq D \ , \tag{6.5}$$

where $\sigma_{G,B}^2$ is the variance parameter and ρ_B is the autocorrelation parameter. Similarly, for Type A we assume that:

$$\text{cov}(\gamma_i^A, \gamma_j^A) = g_{i,j}^A = \sigma_{G,A}^2 \rho_A^{|i-j|} \quad \text{for} \quad 1 \leq i, j \leq D \ , \tag{6.6}$$

where $\sigma_{G,A}^2$ is the variance parameter and ρ_A is the autocorrelation parameter.

We model the dependence between the arrival streams for Type A and Type B via intraday correlations. In particular, we assume that ϵ^B and ϵ^A are correlated. Let R^{BIV} denote the within-day $2P \times 2P$ covariance matrix of residuals for the bivariate model. We assume that:

$$R^{\text{BIV}} = R^* + \sigma^2_{\text{BIV}} I_{2P} \ , \tag{6.7}$$

where I_{2P} is the $2P-$ dimensional identity matrix, and R^* is the $2P \times 2P$ matrix given by:

$$R^* = \begin{pmatrix} \sigma_B^2 & \sigma_{B,A} \\ \sigma_{A,B} & \sigma_A^2 \end{pmatrix} \otimes \begin{pmatrix} 1 & \rho & \rho^2 & ... & \rho^P \\ \rho & 1 & \rho & ... & \rho^{P-1} \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ \rho^P & \rho^{P-1} & \rho^{P-2} & ... & 1 \end{pmatrix} . \tag{6.8}$$

That is, in addition to intraday correlation between the two arrival streams, each call type has an intraday AR(1) covariance structure, as in §4.2.2. We assume that the intraday AR(1) structures for both regions have the same autocorrelation parameter, ρ.

In addition to the covariance structure specified in (6.8), we also considered other ways to model the dependence between the two call types. For example, we assumed that both types have an AR(1) covariance structure for interday dependence, and that there are correlations between the two call types at the daily level. However, we found that the bivariate model specified in (6.3)-(6.8) yields the most accurate forecasts among all other models considered. That is why we focus solely on that model in this paper. In Table 9, we present some results for the statistical significance of the daily random effects, γ^B and γ^A, obtained when fitting the bivariate mixed model in (6.3)-(6.8) to the first 14 weeks of the data set, i.e., to days ranging from October 19, 2009 to January 29, 2010. In Table 10, we present corresponding estimates of the covariance parameters of the bivariate model. It is interesting to note the strong interday and intraday correlations in the data, and the value of σ^2_{BIV} which is close to 0.25, as desired; see §3.3.

Category	Coefficient	Std. error	p-value
Day 1 (Ontario)	1.34	0.626	0.0323
Day 10 (Ontario)	0.851	0.612	0.164
Day 23 (Ontario)	-1.24	0.609	0.0430
Day 30 (Ontario)	-1.52	0.610	0.0129
Day 59 (Ontario)	-1.072	0.613	0.0802
Day 1 (Quebec)	1.03	0.558	0.0675
Day 10 (Quebec)	0.601	0.548	0.273
Day 23 (Quebec)	0.0447	0.547	0.935
Day 30 (Quebec)	-1.06	0.547	0.0519
Day 59 (Quebec)	-1.36	0.549	0.0136

Table 9: Summary of results for the Bivariate Model specified in (6.3) and (6.4) when fit to the data between October 19, 2009 and January 29, 2010. Point estimates of random effects γ^B and γ^A for selected days are shown along with corresponding standard errors and p-values of t-tests for statistical significance.

$\sigma^2_{G,B}$	1.07
ρ_B	0.934
$\sigma^2_{G,A}$	0.829
ρ_A	0.941
σ^2_B	1.09
$\sigma^2_{B,A}$	0.651
σ^2_A	0.833
ρ	0.924
σ^2_{BIV}	0.214

Table 10: Covariance parameter estimates for the Bivariate Mixed Model (6.3) and (6.4) when fit to the data between October 19, 2009 and January 29, 2010.

6.2. Comparison of the Bivariate and Mixed Models

In this subsection, we compare the forecasting accuracies of the BM model of §6.1 and the ME model of §4.2. As noted in §5, the ME model is superior with short forecasting lead times; e.g., see Table 8. Therefore, we focus on short lead times in this subsection.

6.2.1. Forecasting Lead Time and Learning Period.

As in §5, we generate forecasts for the 85 days ranging from August 19, 2010 to November 11, 2010. Here, we consider generating forecasts (i) one day ahead, and (ii) within each day. For (ii),

we consider a lead time of 5.5 hours which corresponds to 11 half-hour periods. Recalling that a day is composed of 22 half-hour periods, we see that option (ii) corresponds to mid-day forecasting updates.

Predictions for a forecast lag of 1 day

	BM model				ME model			
	RASE	APE	Coverage	Width	RASE	APE	Coverage	Width
Mean	33.4	14.2	0.897	128	38.9	16.4	0.858	126
Median	28.1	10.4	0.955	114	33.7	12.8	0.955	111
1st quartile	22.1	8.64	0.909	108	23.7	9.59	0.795	103
3rd quartile	39.5	1	14.4	130	46.9	0.954	0.954	111

Table 11: Comparison of the Bivariate Model described in §6 and the Mixed-Effects Model described in §4.2 for a forecast lag of 1 day. The reported estimates are for forecasts of Type A arrivals.

Predictions for a forecast lag of 1/2 day

	BM model				ME model			
	RASE	APE	Coverage	Width	RASE	APE	Coverage	Width
Mean	28.5	10.9	0.920	102	30.1	11.8	0.898	103
Median	28.4	9.15	1	97.3	28.4	10.1	0.955	97.2
1st quartile	20.1	7.63	0.909	96.5	22.1	8.32	0.864	91.1
3rd quartile	32.9	11.2	1	108	36.6	13.5	1	116

Table 12: Comparison of intraday updating for the Bivariate Model described in §6 and the Mixed Model described in §4.2. The reported estimates are for forecasts of Type A arrivals.

We consider a learning period of 58 days (or $58 \times 2 = 116$ sets of 11 half-hour periods), which corresponds to about 12 weeks. We do so for two main reasons. First, the Bivariate Model is quite computationally intensive. To minimize computational time, we use a smaller learning period than in §5. Second, it is interesting to consider shorter learning periods because those are not uncommon in call center management, where data collected in earlier months may not be available.

6.2.2. Model Comparison.

In Table 11, we present estimates for the average RASE, APE, Coverage, and Width for each model, with a forecasting lead time of one day. In this case, the BM model is superior to the ME model. Indeed, RASE(ME)/RASE(BM) is approximately equal to 1.15, whereas APE(ME)/APE(BM) is

roughly equal to 1.16. The BM model also yields more accurate prediction intervals. Indeed, the coverage probability for the Bivariate Model is roughly equal to 0.9, whereas the coverage probability for the ME model is roughly equal to 0.86. On another note, Table 11 clearly shows the advantage of using longer learning periods when making forecasts of future call volumes. For example, with a forecasting lead time of one day and a learning period of about 10 months, Table 8 showed that RASE(ME) is roughly equal to 13%. In contrast, with the same forecasting lead time and a learning period of only 3 months, Table 11 shows that RASE(ME) is now roughly equal to 16%.

In Table 12, we present estimates for the average RASE, APE, Cover, and Width for each model, with a forecasting lead time of half a day. The Bivariate Model is, once more, superior to the ME model. For example, RASE(ME)/RASE(BM) is roughly equal to 1.05 and APE(ME)/APE(BM) is roughly equal to 1.09. Additionally, Table 11 shows that the Bivariate Model yields better prediction intervals than the ME model. For example, the average coverage probability for the Bivariate Model is 0.92 whereas the average coverage probability of the ME model is close to 0.90. The average width of confidence intervals for the Bivariate Model is also smaller than for the ME model. Finally, consistent with intuition, comparing Tables 11 and 12 shows that the magnitude of errors for both models significantly decreases when updating forecasts within the day. For example, RASE(BM) decreases from about 14% with one-day-ahead forecasts, to about 11% with intraday updating.

7. Conclusions

In this paper, we evaluated alternative time series models for forecasting half-hourly arrivals to a call center. We compared the forecasting accuracy of those models based on real-life call center data of a major company in Canada.

7.1. Summary of Main Contributions

The time series models that we considered in this paper are appealing from a practical perspective. In particular, we evaluated forecasting methods that are commonly used in practice, such as Holt-

Winters smoothing and the "Top-Down" approach; see §4.3 and §4.4. In using the "Top-Down" approach, we broke down actual daily forecasts generated by Company X to produce half-hourly forecasts. It is interesting to consider the forecasts of Company X because they incorporate information about marketing campaigns or recent price increases which affect the arrival process.

We modeled interday and intraday dependence structures, commonly observed in call center arrivals, via a Gaussian linear mixed model (§4.2). The mixed model incorporates fixed effects such as day-of-week, period-of-day, and cross-terms between the two. As a useful reference model, we also considered a fixed model (§4.1) which does not incorporate any dependence structure. We compared the forecasting accuracy of all models based on forecasting lead times ranging from weeks to hours in advance, to mimic the challenges faced by call center managers in real life.

In §5, we compared those four models based on out-of-sample forecasts for 85 days, ranging from August 19, 2010 to November 11, 2010. We quantified the impact of the forecasting lead time on the forecasting accuracy of our models in Tables 6 (2 weeks ahead), 7 (1 week ahead), and 8 (1 day ahead). We found that with sufficiently long lead times, a simple historical average is difficult to beat. But, when lead times are short, such as one day or even hours in advance, it becomes increasingly important to model correlations in the data. Indeed, with short forecasting lead times, the mixed model considerably outperforms the remaining models; see Table 8.

In §6, we extended the mixed model into a bivariate mixed model, thus innovatively exploiting the dependence between two call types to generate more accurate forecasts. Figure 1 showed that the two call types considered are strongly positively correlated. In §6.2, we showed that the bivariate model significantly outperforms the standard mixed model; see Table 11 and 12. As a result, we quantified the advantage of modeling correlations between alternative call types in making forecasts of future call volumes.

7.2. Future Research Directions

One possible direction for future research is to extend the bivariate model into a multivariate model which exploits the dependence structure between multiple related call types. Indeed, experience indicates that the arrival processes of several queues in call centers are often correlated, and exploiting

this correlation promises to improve predictions of future call volumes, as we saw in §6.

As indicated in §1, forecasting call center arrivals is an essential first step towards the better management of call centers. There remains to study the more complicated problem of developing efficient algorithms for scheduling agents, and updating the resulting schedules, based on distributional forecasts of future call volumes. Distributional forecasts consist of densities, in addition to point forecasts and prediction intervals. They are particularly important because the variability of the arrival process greatly impacts the performance measures in the system; e.g., see Steckley et al. (2005). In this paper, we used a square-root transformation of the data (§3.3) and exploited the resulting normality of the transformed counts. However, there remains to characterize the marginal and joint densities of the untransformed arrival counts, as in Avramidis et al. (2004).

Simulation-based methods may be used in complicated systems where there are multiple customer classes and multiple service pools with some form of skill-based routing. Distributional forecasts of future arrivals could potentially be implemented in those simulation models to produce approximate solutions to the agent scheduling problem, see Avramidis et al. (2010).

8. 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, Univ. Pennsylvania.

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.

CIA World Factbook. 2010. Available online at: https://www.cia.gov.

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.

Chatfield . C. and M. Yar. Holt-Winters forecasting: Some practical issues. 1988. *The Statistician*, 37: 129–140.

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

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.

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

SAS. 2009. SAS OnlineDoc® 9.1 SAS Institute Inc., Cary, NC.

Soyer, R. and Tarimcilar, M. M. 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. *Proc 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., H. Jacqmin-Gadda, G. Chne, 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.

Winters PR. 1960. Forecasting sales by exponentially weighted moving averages. *Management Science*, 6 : 324-342.