M&SOM MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Vol. 10, No. 3, Summer 2008, pp. 391–410
ISSN 1523-4614 | EISSN 1526-5498 | 08 | 1003 | 0391

informs

DOI 10.1287/msom.1070.0179

Interday Forecasting and Intraday Updating of Call Center Arrivals

Haipeng Shen
Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, haipeng@email.unc.edu

Jianhua Z. Huang
Department of Statistics, Texas A&M University, College Station, Texas 77843, jianhua@stat.tamu.edu

Accurate forecasting of call arrivals is critical for staffing and scheduling of a telephone call center. We develop methods for interday and dynamic intraday forecasting of incoming call volumes. Our approach is to treat the intraday call volume profiles as a high-dimensional vector time series. We propose first to reduce the dimensionality by singular value decomposition of the matrix of historical intraday profiles and then to apply time series and regression techniques. Our approach takes into account both interday (or day-to-day) dynamics and intraday (or within-day) patterns of call arrivals. Distributional forecasts are also developed. The proposed methods are data driven, appear to be robust against model assumptions in our simulation studies, and are shown to be very competitive in out-of-sample forecast comparisons using two real data sets. Our methods are computationally fast; it is therefore feasible to use them for real-time dynamic forecasting.

Key words: dimension reduction; dynamic forecast updating; principal component analysis; penalized least squares; singular value decomposition; vector time series
History: Received: July 17, 2006; accepted: March 15, 2007. Published online in *Articles in Advance* January 4, 2008.

1. Introduction

Call centers have become a primary contact point in modern business between service providers and their customers. For example, it is estimated that in 2002 more than 70% of all customer-business interactions were handled in call centers, and the U.S. call center industry employed more than 3.5 million people, or 2.6% of the workforce (Brown et al. 2005). With their growing presence and importance in various organizations, managing call center operations more efficiently is of significant economic interest.

For most call centers, 60%–70% of operating expenses are capacity costs, in particular human resource costs (Gans et al. 2003). Effective management of a call center requires call center managers to match call center resources with the workload. To accurately forecast the workload, the first and most critical step is to provide an accurate forecast of future call volumes. Usually two kinds of forecasts are needed by a call center manager for staffing and scheduling purposes: (1) forecast the call volumes several days or weeks ahead and (2) on a particular day, dynamically update the forecast using newly available information as additional calls arrive throughout the day.

There are empirical justifications for the within-day updating, as both Avramidis et al. (2004) and Steckley et al. (2004) report evidence of positive correlation among time periods within a given day, i.e., *intraday dependence*. This updating is operationally beneficial and feasible. If the within-day updating can be done appropriately, one can substantially reduce the forecast error for the rest of the day. Furthermore, one can perform the updating as early as two to three hours into the working day to have a significant effect. In turn, operational benefits would follow from the ability to change agent schedules according to the updated forecast. For example, to adjust for a revised forecast, call center managers may send people home early, have them take training sessions, or work on alternative tasks; or they may arrange for agents to work overtime or call in part-time or work-from-home agents. To the best of our knowledge, Weinberg et al. (2007) is the only existing study dealing with forecast updating in the call center context.

Quantitative approaches to call arrival forecasting use historical call arrival data, which record the times at which calls arrive at the center. Typically, call arrival data are aggregated into total numbers of arrivals during short time periods, such as 15-minute or 30-minute intervals; subsequently the target of forecasting is future call volumes over such periods (Jongbloed and Koole 2001). We refer to the collection of such aggregated call volumes for a given day as the *intraday call volume profile* of that day (§3.1).

Early work on forecasting call arrivals usually ignores intraday dependence and focuses on interday dependence among daily total volumes, partly because of lack of relevant data. Time series (TS) autoregressive integrated moving average (ARIMA) models are usually called on to perform the forecasting (Andrews and Cunningham 1995; Bianchi et al. 1993, 1998). Recently, with the availability of call-by-call data, stochastic models of call arrivals that take into account the intraday dependence have been introduced. Avramidis et al. (2004) develop doubly stochastic models that reproduce three empirically observed features of call arrival data, *overdispersion*, *time-varying rate*, and *intraday dependence*. Brown et al. (2005) consider a time-varying Poisson process with random rate and suggest a multiplicative two-way, mixed-effects model to forecast a future rate. Considering both interday forecasting and intraday updating, Weinberg et al. (2007) extend the work of Avramidis et al. and Brown et al. to model both interday and intraday dependences. A two-way multiplicative Bayesian Gaussian model is proposed, and a Markov chain Monte Carlo (MCMC) algorithm is developed for parameter estimation and forecasting; see §4.2 for details of the model. The Bayesian approach is theoretically interesting, but the computation (MCMC) algorithm is rather sophisticated to implement and can take a long time to converge.

Intraday schedule updating has been investigated in the literature under the name of real-time work schedule adjustment for various service systems other than call centers. The real-time schedule adjustment process usually involves two phases (Hur et al. 2004). During the first phase, forecast monitoring systems (Hur 2002) are applied to detect when a forecast update is needed; once an update is determined to be necessary, the forecast can be updated using a partially known demand forecasting model (Kekre et al. 1990, Guerrero and Elizondo 1997).

The second phase involves schedule adjustment to handle the updated forecast, which has received very little attention until recently. For example, Thompson (1996) investigates various options for adjusting intraday schedules in the context of hospitality workforce management. Hur et al. (2004) propose schedule updating techniques, which are illustrated in the context of quick service restaurants. Easton and Goodale (2005) model absence recovery problems for service operations. In the context of call centers, Mehrotra et al. (2006) study in detail intraday rescheduling of agent schedules and demonstrate its value.

These forecasting approaches are mostly model driven, where stochastic models are carefully designed to fit real-world problems. In this paper, we introduce a system of statistical methods that can generate both interday (day-to-day) and *especially* dynamic intraday (within-day) forecasts. Unlike the model-driven approaches, our approach is data driven, which starts with a dimension-reduction technique to let the data suggest parsimonious *underlying* important features, and then uses them to build TS forecasting models. We illustrate that our approach can work as a competitive and complementary alternative to the model-driven approaches.

The data-driven nature suggests that our approach is robust (§4.3) and can work for a wide range of applications, for example, electricity demand forecasting (Cottet and Smith 2003) and bond yield curve forecasting (Diebold and Li 2006). We are interested in applying our approach to these applications to see how it performs. Furthermore, our method can be easily implemented in widely available softwares such as MATLAB and SPLUS/R, and it computes faster than the Bayesian method of Weinberg et al. (2007), which makes it more suitable for real-time forecast updating. In spite of its simplicity, our method is still very competitive in generating accurate forecasts (§4.5).

Our approach treats the intraday call volume profiles as a vector TS across days, where each intraday profile is considered one observation of the vector TS. Thus, the dimensionality of the TS is high. We propose to first reduce the dimensionality by applying

singular value decomposition (SVD) to the matrix of the intraday profiles. As a result, the high dimensional intraday profile series is summarized by a few pairs of interday and intraday features (see (1) in §3.1). Then the intraday features are kept fixed, while the interday feature series are subject to TS modeling.

Because the interday feature series are orthogonal, we can apply univariate TS forecasting techniques to forecast them separately. The resulting forecasts are subsequently combined with the intraday features to produce day-to-day forecasts of future intraday profiles. Distributional forecasts are also developed using a bootstrap procedure. Usually only a few interday and intraday feature pairs are needed; two or three such pairs are sufficient to achieve very good forecasting performance for the two case studies in §4.

Our procedure for intraday dynamic updating is based on a technique known as penalized least squares (PLS). For a particular day, suppose we have obtained some day-to-day TS forecasts for the interday features. As calls arrive as the day progresses, one direct approach to updating the TS forecasts is via a least squares (LS) regression using the newly available call volume profile as the response and the corresponding part of the intraday feature vectors as the independent variables (see (6) in §3.3.1). The corresponding regression coefficients then give the updated interday feature forecasts. However, this approach relies solely on the new arrival information of the present day and makes no use of the TS forecasts. Our PLS method adjusts the direct approach by pushing the LS updates toward the TS forecasts and finding a balance point between them (see (8) in §3.3.2). This method effectively combines the newly available information with the information up through the end of the previous day. The empirical results in §4 show that such a combination of information is very useful in reducing forecasting errors.

The rest of the paper is structured as follows. Section 2 introduces the call center arrival data that motivate the current research. Our forecasting approach is described in detail in §3. Section 3.1 introduces dimensionality reduction through SVD, and is followed by §3.2 on the method for one- or multi-day-ahead intraday profile forecasting; §3.3 focuses on dynamic intraday updating. Both point and distributional forecasts are considered. Section 4 presents comparative studies of several competing methods based on their out-of-sample forecasting performance. Section 4.1 gives the forecasting performance measures we use, and §4.2 describes the competing methods. Two simulated examples are presented in §4.3 to illustrate the robustness of our method. The data described in §2 are used in §4.4 to compare our method with several alternatives. In §4.5, a comparison between our method and the Bayesian approach of Weinberg et al. (2007) is provided, using their call center data. We conclude in §5. Some technical details are provided in the appendix.

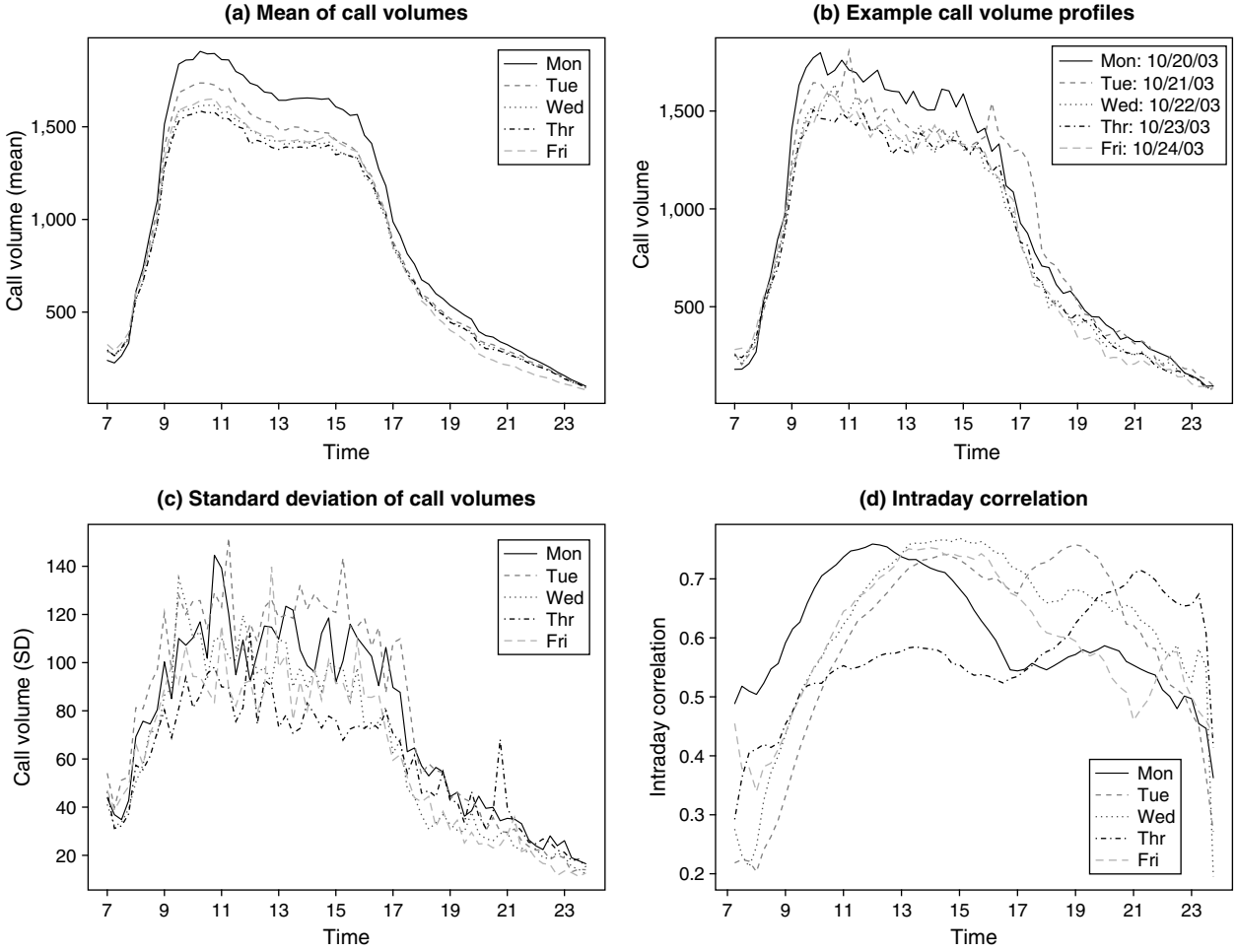
2. Data

The motivating data were gathered at an inbound call center of a major northeastern U.S. financial firm between January 1 and October 26, 2003. The original database has detailed information about every call that was connected to this call center during this period (except three days when the data-collecting equipment went out of order). The center is normally open between 7:00 AM and noon. For the current study, we are interested in understanding the arrival pattern of calls to the service queue and generating out-of-sample forecasts of future call volumes. As a result, the portion of the data of interest to us involves information about the time every call arrives at the service queue.

Because the call arrival pattern is very different between weekdays and weekends, we restrict our attention to weekdays. In particular, we focus on the 42 five-day weeks between January 6 and October 24. There are some obvious abnormal days in the data—six holidays with very low call volumes and four days with no data; hence we exclude them from our analysis. For a particular day, we divide the 17-hour operating period into 68 15-minute intervals and record the number of calls arriving to the service queue during each interval. The aggregated data form a 200×68 count matrix, with each row corresponding to a particular day in the 42 weeks considered and each column corresponding to one specific 15-minute interval between 7:00 AM and midnight. The data are available from the authors on request.[1]

[1] An online appendix to this paper is available on the *Manufacturing & Service Operations Management* website (http://msom.pubs.informs.org/ecompanion.html).

Figure 1 Exploratory Plots Suggesting that Both Predictive Regularity and Abnormality Are Present in Call Center Arrival Data



Notes. (a) Average call volume profiles for different weekdays; (b) intraday arrival patterns during the final week of our data; (c) standard deviation of call volumes for different weekdays; (d) intraday correlation between past call demand and future demand for different weekdays.

By plotting the data we found that call volumes almost always occur in predictable, repeating patterns. For example, typically there is a peak around 11:00 AM, followed by a second, lower peak around 2:00 PM Figure 1(a) plots the average intraday arrival patterns for different weekdays. Different days of the week have different call arrival patterns. As shown in Figure 1(a), Fridays have the largest volumes in the morning and the smallest volumes after 5:00 PM; the opposite holds for Mondays. In addition, Figure 1(b) shows the intraday arrival patterns for the five days in the final week of the data set. The regularity of arrival patterns serves as the basis of any forecasting method.

Figure 1(b) also exhibits an unusual arrival pattern for the Tuesday, during which there are too many calls arriving after 3:30 PM We keep such abnormal days for the case study to be reported in §4.4. Alternatively, we have tried to identify anomalous days using the technique in Shen and Huang (2005); their call arrival profiles were then replaced by the average of the corresponding periods (i.e., same weekday and time of day) in the two adjacent weeks. This alternative treatment leads to similar forecasting comparison results (not shown) as those reported in §4.4.

Our data appear to possess heteroscedasticity (i.e., nonconstant variance) and overdispersion (i.e., variance greater than the mean) (Avramidis et al. 2004), as shown in Figure 1(c). To stabilize the variance, we

used the root-unroot method (Brown et al. 2005) as follows. Let N denote the call volume for a certain time interval. Set $X = \sqrt{N + 1/4}$. Denote the forecast of X by $\widehat{X}$. Then the forecast of N is given by $\widehat{N} = \widehat{X}^2 - 1/4$.

The motivation of the method is the following. It is a good approximation to assume that the call volume N over a short time interval follows a Poisson distribution, and, if $N \sim \text{Poisson}(\lambda)$, then approximately X has a mean of $\sqrt{\lambda}$ and a variance of $1/4$.

The intraday dependence reported by Avramidis et al. (2004) and Steckley et al. (2004) also exists in our data. For different weekdays, Figure 1(d) plots the intraday correlation between past call demand and future demand as a function of the observation time. The demand is calculated based on the square-root-transformed call volumes as motivated above. The intraday correlation appears to be significantly positive most of the time, which empirically supports the potential effectiveness of intraday updating.

3. Forecasting Methods

3.1. Dimension Reduction via Singular Value Decomposition

Let $\mathbf{X} = (x_{ij})$ be an $n \times m$ matrix that records the call volumes (or a transformation of the volumes as described in §2) for n days, with each day having m time periods. The rows and columns of $\mathbf{X}$ correspond respectively to days and time periods within a day. The ith row of $\mathbf{X}$, denoted as $\mathbf{x}_i^T = (x_{i1}, \ldots, x_{im})$, is referred to as the intraday call volume profile of the ith day. The intraday profiles, $\mathbf{x}_1, \mathbf{x}_2, \ldots$, form a vector-valued TS taking values in $\mathbb{R}^m$. We want to build a TS model for this series and use it for forecasting. However, commonly used multivariate TS models such as vector autoregressive models and more general vector autoregressive and moving average models (Reinsel 2003) are not directly applicable because of the large dimensionality of the TS we consider. For example, the dimensionality m is 68 for our data and 169 for Weinberg's data.

Our approach starts from dimension reduction. We first seek a few basis vectors, denoted as $\mathbf{f}_k$, $k = 1, \ldots, K$, such that all elements in the TS $\{\mathbf{x}_i\}$ can be represented (or approximated well) by these basis vectors. The number of the basis vectors K should be much smaller than the dimensionality m of the TS. Specifically, we consider the following decomposition,

$$\mathbf{x}_i = \beta_{i1}\mathbf{f}_1 + \cdots + \beta_{iK}\mathbf{f}_K + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, n, \tag{1}$$

where $\mathbf{f}_1, \ldots, \mathbf{f}_K \in \mathbb{R}^m$ are the basis vectors and $\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_n \in \mathbb{R}^m$ are the error terms. We expect that the main features of $\mathbf{x}_i$ can be summarized by a linear combination of the basis vectors so that the error terms in (1) would be small in magnitude. This can be achieved by solving the following minimization problem for fixed K:

$$\min_{\substack{\beta_{i1},\ldots,\beta_{iK} \\ \mathbf{f}_1,\ldots,\mathbf{f}_K}} \sum_{i=1}^{n} \|\boldsymbol{\epsilon}_i\|^2 = \min_{\substack{\beta_{i1},\ldots,\beta_{iK} \\ \mathbf{f}_1,\ldots,\mathbf{f}_K}} \sum_{i=1}^{n} \|\mathbf{x}_i - (\beta_{i1}\mathbf{f}_1 + \cdots + \beta_{iK}\mathbf{f}_K)\|^2. \tag{2}$$

For identifiability, we require in (2) that $\mathbf{f}_i^T\mathbf{f}_j = \delta_{ij}$, where δ_{ij} is the Kronecker delta, which equals 1 for $i = j$ and 0 otherwise. The solution to this problem is actually given by the SVD of the matrix $\mathbf{X}$ as shown below.

The SVD of the matrix $\mathbf{X}$ can be expressed as

$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^T, \tag{3}$$

where $\mathbf{U}$ is an $n \times m$ matrix with orthonormal columns, $\mathbf{S}$ is an $m \times m$ diagonal matrix, and $\mathbf{V}$ is an $m \times m$ orthogonal matrix. The columns of $\mathbf{U}$, $\{\mathbf{u}_k = (u_{1k}, \ldots, u_{nk})^T\}$, namely the left singular vectors, satisfy $\mathbf{u}_i^T\mathbf{u}_j = \delta_{ij}$. The columns of $\mathbf{V}$, $\{\mathbf{v}_k = (v_{1k}, \ldots, v_{mk})^T\}$, or the right singular vectors, satisfy $\mathbf{v}_i^T\mathbf{v}_j = \delta_{ij}$. The diagonal elements of $\mathbf{S}$ are the singular values, which are usually ordered decreasingly. Let $\mathbf{S} = \text{diag}(s_1, \ldots, s_m)$ and $r = \text{rank}(\mathbf{X})$. Then $s_1 \geq s_2 \geq \cdots \geq s_r > 0$, and $s_k = 0$ for $r + 1 \leq k \leq m$.

It then follows from (3) that

$$\mathbf{x}_i = s_1 u_{i1}\mathbf{v}_1 + \cdots + s_r u_{ir}\mathbf{v}_r.$$

Keeping only the terms associated with the largest K singular values, we have the following approximation:

$$\mathbf{x}_i \simeq (s_1 u_{i1})\mathbf{v}_1 + \cdots + (s_K u_{iK})\mathbf{v}_K.$$

This K-term approximation is an optimal solution for the minimization problem (2). More precisely, $\beta_{ik} = s_k u_{ik}$ and $\mathbf{f}_k = \mathbf{v}_k$, $i = 1, \ldots, n$, $k = 1, \ldots, K$, solve (2), and the solution is unique up to a sign change to $\mathbf{f}_k$

(Eckart and Young 1936). Thus, we have formally obtained the decomposition (1) using the SVD of $\mathbf{X}$. Note that requiring $\mathbf{f}_i^T\mathbf{f}_j = \delta_{ij}$ is only one way to make the solution to (2) unique. Alternatively, if we require $\sum_{i=1}^n \beta_{ik}\beta_{il} = \delta_{kl}$, then one solution to (2) will be $\beta_{ik} = u_{ik}$ and $\mathbf{f}_k = s_k\mathbf{v}_k$, which is unique up to a sign change to $\mathbf{f}_k$.

Given a historical data matrix $\mathbf{X}$, to estimate the model (1), we apply SVD to $\mathbf{X}$ and extract the first K pairs of singular vectors, along with the corresponding singular values, which then lead to the intraday feature vectors $\mathbf{f}_1, \ldots, \mathbf{f}_K$, and the interday feature series $\{\beta_{i1}\}, \ldots, \{\beta_{iK}\}$. Such features will be used as the basis of our forecasting methods, as described in §§3.2 and 3.3. From the decomposition (1), the high-dimensional intraday profiles $\mathbf{x}_i$ are concisely summarized by a small number of interday feature series, using the same number of intraday feature vectors as the bases.

The decreasing order among the singular values suggests that the feature series/vectors are naturally ordered according to importance in approximating the data matrix. By using a small K, we achieve a substantial dimension reduction—from m to K. In the real data example in §4.4, a K of 2 or 3 already gives good forecasting results, while m is 68. One attractive feature of the decomposition (1) is that it effectively separates the intraday and interday variations, both of which are present in the intraday profile TS. Another benefit is that the variations are decomposed into orthogonal components, which greatly simplifies our forecasting model, as described in §3.2.

Our SVD-based dimension reduction is closely related to principal components analysis (PCA) when principal components (PCs) are calculated from the covariance matrix (Jolliffe 2002). If the data matrix $\mathbf{X}$ is column centered such that the columns (viewed as variables) have a mean of zero, then $\mathbf{X}^T\mathbf{X}$ is proportional to the sample covariance matrix of the columns of $\mathbf{X}$. According to (3), $\mathbf{X}^T\mathbf{X} = \mathbf{V}\mathbf{S}^2\mathbf{V}^T$, which means that the columns $\mathbf{v}_k$ of $\mathbf{V}$, or the intraday feature vectors $\mathbf{f}_k$, are indeed the PC-loading vectors, or $\mathbf{X}\mathbf{v}_k$ are the PCs; and the squared singular values s_k^2 are proportional to the variances of the PCs. In PCA, the quantity $s_k^2/\sum_i s_i^2$ measures the relative importance of the kth PC and can be plotted versus k in a scree plot. To determine the number of PCs, one usually looks for an "elbow" in the scree plot, formed by a steep drop followed by a relatively flat tail of small values. The number of PCs needed then corresponds to the integer prior to the elbow. Although we do not center columns in our context, the connection with PCA suggests the scree plot as a possible way to pick the number of important features K (Shen and Huang 2005). In practice, one can also choose K by comparing out-of-sample forecasting performance using historical data for different choices.

3.2. Interday Forecasting

Consider forecasting the intraday call volume profile $\mathbf{x}_{n+h}$ $(h > 0)$ using the historical data $\mathbf{x}_1, \ldots, \mathbf{x}_n$. By making use of the model (1), forecasting an m-dimensional TS $\{\mathbf{x}_i\}$ reduces to forecasting K one-dimensional interday feature series $\{\beta_{i1}\}, \ldots, \{\beta_{iK}\}$. We consider the decomposition (1) with the $\mathbf{f}_k$s replaced by the ones extracted using SVD from the historical data matrix $\mathbf{X}$.

3.2.1. Point Forecast. If we can obtain a forecast of the coefficients $\beta_{n+h,1}, \ldots, \beta_{n+h,K}$, then a forecast of $\mathbf{x}_{n+h}$ is given by

$$\hat{\mathbf{x}}_{n+h} = \hat{\beta}_{n+h,1}\mathbf{f}_1 + \cdots + \hat{\beta}_{n+h,K}\mathbf{f}_K,$$

where $\hat{\beta}_{n+h,k}$ is a forecast of $\beta_{n+h,k}$, $k = 1, \ldots, K$. Denote $\mathbf{f}_k = (f_{1k}, \ldots, f_{mk})^T$. The forecast of the call volume of the jth time period of day $n + h$ is then given by

$$\hat{x}_{n+h,j} = \hat{\beta}_{n+h,1}f_{j1} + \cdots + \hat{\beta}_{n+h,K}f_{jK}, \quad j = 1, \ldots, m.$$

Because of the way the SVD is constructed, $(\beta_{1k}, \ldots, \beta_{nk})$ is orthogonal to $(\beta_{1l}, \ldots, \beta_{nl})$ for $k \neq l$. This lack of contemporaneous correlation suggests that the cross-correlations at nonzero lags are likely to be small. Therefore, it is reasonable to believe that forecasting each series $\{\beta_{ik}\}$ separately using univariate TS methods, for $k = 1, \ldots, K$, is adequate. To forecast each series, one could use univariate TS models such as exponential smoothing, ARIMA models (Box et al. 1994), state space models (Harvey 1990), and other suitable models. Which model to use would depend on the actual situation and can be decided by analyzing historical data.

In our case study of the call center data described in §2, we apply a varying coefficient AR(1) model that

takes into account the day-of-the-week effect. This model works well in our out-of-sample forecasting exercise in §4.4.

3.2.2. Distributional Forecast. For the agent scheduling and updating problem discussed in §1, call center managers also need distributional forecasts of $\mathbf{x}_{n+h}$, such as prediction intervals and densities, in addition to point forecasts. Understanding the variability around point forecasts is essential for several reasons. A threshold on predicted volumes could be determined to trigger appropriate managerial actions (Jongbloed and Koole 2001). The staffing level for a time period depends on both the mean and the variability of the forecast. As suggested by Steckley et al. (2004), the higher the variability, the higher the staffing targets should be to achieve certain service level goals. In addition, the variability in the forecasts may provide motivation for contracting additional contingency resources, routing calls across different sites, and cross-training across different queues (Betts et al. 2000, Wallace and Whitt 2005).

As our approach assumes no probabilistic models, we propose using the following nonparametric bootstrap method to derive the distributional forecasts. Bootstrapping is an increasingly popular method for estimating the sampling distribution of an estimator by sampling with replacement from the original sample (Efron and Tibshirani 1994).

For a fixed k, using the fitted TS model for the series $\{\beta_{ik}\}$ recursively, we can generate B series $\{\hat{\beta}^b_{n+1,k}, \ldots, \hat{\beta}^b_{n+h,k}\}, 1 \le b \le B$, by bootstrapping the model errors from the fitted TS model (see Appendix A.1 for details). In our data examples (§4), we set the number of bootstrap samples B to be 1,000. Similarly, we can bootstrap the approximation error profile $\boldsymbol{\epsilon}$ in the model (1) from the fitted error profiles $\hat{\boldsymbol{\epsilon}}_1, \ldots, \hat{\boldsymbol{\epsilon}}_n$ obtained by fitting the model (1) to the historical n call volume profiles. Let $\hat{\boldsymbol{\epsilon}}^b$, $1 \le b \le B$, denote a random draw with replacement from $\hat{\boldsymbol{\epsilon}}_1, \ldots, \hat{\boldsymbol{\epsilon}}_n$. Then we obtain B forecasts of $\mathbf{x}_{n+h}$,

$$\hat{\mathbf{x}}^b_{n+h} = \hat{\beta}^b_{n+h,1}\mathbf{f}_1 + \cdots + \hat{\beta}^b_{n+h,K}\mathbf{f}_K + \hat{\boldsymbol{\epsilon}}^b, \quad b = 1, \ldots, B, \quad (4)$$

from which interval and density forecasts can be derived. For a $100(1-\alpha)\%$ forecast interval, the two end points are correspondingly the $\alpha/2$ and $(1-\alpha)/2$ empirical percentiles of $\{\hat{\mathbf{x}}^1_{n+h}, \ldots, \hat{\mathbf{x}}^B_{n+h}\}$. The histogram or density estimate of $\{\hat{\mathbf{x}}^1_{n+h}, \ldots, \hat{\mathbf{x}}^B_{n+h}\}$ then gives a density forecast of $\mathbf{x}_{n+h}$.

Our approach makes two implicit assumptions. First, the TS model used should be a reasonably good approximation of reality. Second, for the validity of the nonparametric bootstrap, the TS model errors are assumed to be independent and identically distributed with a mean of zero. This is why the approximation error profile $\boldsymbol{\epsilon}$ is bootstrapped as a whole, so that we can capture any remaining intraday correlation. These implicit assumptions remain applicable when we discuss dynamic intraday updating in §3.3. Note that our approach does not assume a specific parametric distribution for the errors.

3.3. Dynamic Intraday Updating

Using the method in the previous section, a manager can forecast the call volume profile for day $n+h$ at the end of day n. In particular, $h = 1$ corresponds to the next day. As calls arrive during day $n+1$, the manager may want to dynamically update the forecast for the remainder of the day using the information from the early part of that day. This dynamic updating is useful because it adds flexibility to allocation of available resources, which leads to higher efficiency and productivity, as well as better quality of service. By reevaluating the forecast for the remainder of the day, the manager can then schedule meetings or training sessions for agents free from work at short notice, or call in back-up agents.

Suppose we have the call volumes during the first m_0 time periods of day $n+1$. Denote them collectively as $\mathbf{x}^e_{n+1} = (x_{n+1,1}, \ldots, x_{n+1,m_0})^T$, a vector containing the first m_0 elements of $\mathbf{x}_{n+1}$. Denote $\mathbf{x}^l_{n+1} = (x_{n+1,m_0+1}, \ldots, x_{n+1,m})^T$ to be the intraday call volume profile for the later part of day $n+1$. For notational simplicity, we suppress the dependence of $\mathbf{x}^e_{n+1}$ and $\mathbf{x}^l_{n+1}$ on m_0.

Let $\hat{\beta}^{\mathrm{TS}}_{n+1,k}$ be a TS forecast based on $\{\beta_{i,k}\}$, for $k = 1, \ldots, K$, using information up to the end of day n. The TS forecast of $\mathbf{x}^l_{n+1}$ is then given by

$$\hat{x}^{\mathrm{TS}}_{n+1,j} = \hat{\beta}^{\mathrm{TS}}_{n+1,1} f_{j1} + \cdots + \hat{\beta}^{\mathrm{TS}}_{n+1,K} f_{jK}, \quad j = m_0+1, \ldots, m. \quad (5)$$

These forecasts do not use any new information from day $n+1$. We discuss two ways to incorporate the new information in $\mathbf{x}^e_{n+1}$ to obtain an updated forecast of $\mathbf{x}^l_{n+1}$.

3.3.1. Direct LS Updating. When being applied to the intraday profile of day $n+1$, the decomposition (1) can be written as

$$x_{n+1,j} = \beta_{n+1,1} f_{j1} + \cdots + \beta_{n+1,K} f_{jK} + \epsilon_{n+1,j}, \quad j = 1, \ldots, m. \quad (6)$$

Let $\mathbf{F}^e$ be a $m_0 \times K$ matrix whose (j,k)th entry is f_{jk}, $1 \le j \le m_0$, $1 \le k \le K$, $\boldsymbol{\beta}_{n+1} = (\beta_{n+1,1}, \ldots, \beta_{n+1,K})^T$, and $\boldsymbol{\epsilon}^e_{n+1} = (\epsilon_{n+1,1}, \ldots, \epsilon_{n+1,m_0})^T$. Then, with the availability of $\mathbf{x}^e_{n+1}$, we have the following linear regression model:

$$\mathbf{x}^e_{n+1} = \mathbf{F}^e \boldsymbol{\beta}_{n+1} + \boldsymbol{\epsilon}^e_{n+1}.$$

This suggests that we can forecast $\boldsymbol{\beta}_{n+1}$ by the method of LS. Solving

$$\min_{\boldsymbol{\beta}_{n+1}} \|\mathbf{x}^e_{n+1} - \mathbf{F}^e \boldsymbol{\beta}_{n+1}\|^2,$$

we obtain

$$\hat{\boldsymbol{\beta}}^{\mathrm{LS}}_{n+1} = (\mathbf{F}^{eT}\mathbf{F}^e)^{-1}\mathbf{F}^{eT}\mathbf{x}^e_{n+1}.$$

The LS forecast of $\mathbf{x}^l_{n+1}$ is then given by

$$\hat{x}^{\mathrm{LS}}_{n+1,j} = \hat{\beta}^{\mathrm{LS}}_{n+1,1} f_{j1} + \cdots + \hat{\beta}^{\mathrm{LS}}_{n+1,K} f_{jK}, \quad j = m_0 + 1, \ldots, m. \quad (7)$$

To make the LS forecast of $\boldsymbol{\beta}_{n+1}$ uniquely defined, we require that $m_0 \ge K$. This makes sense intuitively, in that one needs more observations than parameters to carry out the LS estimation. Operationally, this suggests a trade-off between having more feature factors (i.e., a larger K) and updating forecasts earlier in the day (a smaller m_0). One achieves a better approximation of the data matrix $\mathbf{X}$ by having more factors; however, the manager then has to wait longer to start the direct LS forecast updating. We propose a penalized LS updating approach that seems to offer a solution to this trade-off.

3.3.2. PLS Updating. Direct LS updating uses the additional information available at the early part of day $n+1$, but it needs a sufficient amount of data (i.e., a large enough m_0) for $\hat{\boldsymbol{\beta}}^{\mathrm{LS}}_{n+1}$ to be reliable. This might create a problem if the manager wants to update the forecast early in the morning, for example, at 8:00 AM with $m_0 = 4$ or at 10:00 AM with $m_0 = 12$ for the data in §2. Another disadvantage of direct LS updating is that it does not make full use of the historical information other than the estimated intraday feature vectors. In particular, it ignores the day-to-day dependence present in the interday feature series.

We propose combining the LS forecast with the TS forecast of $\boldsymbol{\beta}_{n+1}$ by using the idea of penalization. Specifically, with respect to $\beta_{n+1,1}, \ldots, \beta_{n+1,K}$, we minimize the following PLS criterion,

$$\sum_{j=1}^{m_0} |x_{n+1,j} - (\beta_{n+1,1} f_{j1} + \cdots + \beta_{n+1,K} f_{jK})|^2 + \lambda \sum_{k=1}^{K} |\beta_{n+1,k} - \hat{\beta}^{\mathrm{TS}}_{n+1,k}|^2, \quad (8)$$

where $\hat{\beta}^{\mathrm{TS}}_{n+1,k}$ is a TS forecast based on the information up to the end of day n, and $\lambda > 0$ is a penalty parameter. Since only one penalty parameter is used, it makes sense for the K TS $\{\beta_{ik}\}$, $1 \le k \le K$, to be roughly on the same scale. This can be achieved by requiring $\sum_{i=1}^{n} \beta_{ik}\beta_{il} = \delta_{kl}$ or $(1/n)\sum_{i=1}^{n} \beta_{ik}\beta_{il} = \delta_{kl}$ in (2).

The PLS criterion (8) involves two terms: The first term measures how well the model prediction matches the observed call volumes in the early part of the day, and the second term penalizes a large departure from the TS forecast. The $\boldsymbol{\beta}_{n+1}$ obtained as the solution to the minimization problem will be a compromise between the two terms based on the size of λ, the penalty parameter. In practice, λ can be selected based on the forecasting performance on a rolling hold-out sample; see §4.4.2 for a detailed description of one selection procedure. The case study in §4.4.2 also exemplifies the relative importance of the two terms.

For each given λ, the PLS problem can be rewritten in a constrained optimization form:

$$\text{minimize} \quad \sum_{k=1}^{K} |\beta_{n+1,k} - \hat{\beta}^{\mathrm{TS}}_{n+1,k}|^2$$

$$\text{subject to} \quad \sum_{j=1}^{m_0} |x_{n+1,j} - (\beta_{n+1,1} f_{j1} + \cdots + \beta_{n+1,K} f_{jK})|^2 \le \xi,$$

where $\xi = \xi(\lambda)$ is chosen appropriately so the solution to this problem and the previous one are the same for the given λ. The optimization problem can then be interpreted as follows: Among all the $\boldsymbol{\beta}_{n+1}$s that yield forecasts that adequately match the observed call volumes of the early part of the day, find the one that is closest to the TS forecast.

We can express the PLS criterion (8) in the following matrix form:

$$(\mathbf{x}^e_{n+1} - \mathbf{F}^e\boldsymbol{\beta}_{n+1})^T(\mathbf{x}^e_{n+1} - \mathbf{F}^e\boldsymbol{\beta}_{n+1}) + \lambda(\boldsymbol{\beta}_{n+1} - \hat{\boldsymbol{\beta}}^{\mathrm{TS}}_{n+1})^T(\boldsymbol{\beta}_{n+1} - \hat{\boldsymbol{\beta}}^{\mathrm{TS}}_{n+1}). \quad (9)$$

Minimizing this criterion gives us the PLS forecast of $\boldsymbol{\beta}_{n+1}$ as

$$\hat{\boldsymbol{\beta}}^{\mathrm{PLS}}_{n+1} = (\mathbf{F}^{eT}\mathbf{F}^e + \lambda\mathbf{I})^{-1}(\mathbf{F}^{eT}\mathbf{x}^e_{n+1} + \lambda\hat{\boldsymbol{\beta}}^{\mathrm{TS}}_{n+1}). \quad (10)$$

The PLS forecast of $\mathbf{x}^l_{n+1}$ is then given by

$$\hat{x}^{\mathrm{PLS}}_{n+1,\,j} = \hat{\beta}^{\mathrm{PLS}}_{n+1,\,1}f_{j1} + \cdots + \hat{\beta}^{\mathrm{PLS}}_{n+1,\,K}f_{jK}, \quad j = m_0 + 1, \ldots, m. \quad (11)$$

The following Theorem 1 suggests that the PLS forecast $\hat{\boldsymbol{\beta}}^{\mathrm{PLS}}_{n+1}$ is actually a weighted average of the LS and the TS forecasts, with the weights controlled by the penalty parameter λ. Note that, if $\lambda = 0$, $\hat{\boldsymbol{\beta}}^{\mathrm{PLS}}_{n+1}$ is simply $\hat{\boldsymbol{\beta}}^{\mathrm{LS}}_{n+1}$; if $\lambda \to \infty$, then $\hat{\boldsymbol{\beta}}^{\mathrm{PLS}}_{n+1}$ reduces to $\hat{\boldsymbol{\beta}}^{\mathrm{TS}}_{n+1}$.

Theorem 1.

$$\hat{\boldsymbol{\beta}}^{\mathrm{PLS}}_{n+1} = A(\lambda)\hat{\boldsymbol{\beta}}^{\mathrm{LS}}_{n+1} + (I - A(\lambda))\hat{\boldsymbol{\beta}}^{\mathrm{TS}}_{n+1},$$

where $A(\lambda)$ is a $K \times K$ matrix satisfying $A(0) = I$ and $A(\lambda) \to 0$ as $\lambda \to \infty$.

We point out that the PLS idea can be derived from a Bayesian perspective. To this end, we treat the TS forecast $\hat{\boldsymbol{\beta}}^{\mathrm{TS}}_{n+1}$ as our prior knowledge of $\boldsymbol{\beta}_{n+1}$, and model $\beta_{n+1,\,k}$ $(1 \le k \le K)$ as random draws from a "prior" normal distribution with mean $\hat{\beta}^{\mathrm{TS}}_{n+1,\,k}$ and variance σ_0^2. The error term $\epsilon_{n+1,\,j}$ in the expression (6) is modeled as normal with a mean of zero and variance σ^2. Then the "posterior" density function of $\boldsymbol{\beta}_{n+1}$ is maximized by the solution of the PLS problem (8) with $\lambda = \sigma^2/\sigma_0^2$.

There is a connection between our PLS approach and the widely used ridge regression (Hoerl and Kennard 1970a, b) in regression analysis. Define $\tilde{\mathbf{x}}^e_{n+1} = \mathbf{x}^e_{n+1} - \mathbf{F}^e\hat{\boldsymbol{\beta}}^{\mathrm{TS}}_{n+1}$ and $\tilde{\boldsymbol{\beta}}_{n+1} = \boldsymbol{\beta}_{n+1} - \hat{\boldsymbol{\beta}}^{\mathrm{TS}}_{n+1}$. Consider a ridge regression with $\tilde{\mathbf{x}}^e_{n+1}$ as the response, $\mathbf{F}^e$ as the design matrix, and $\tilde{\boldsymbol{\beta}}_{n+1}$ as the regression coefficient vector. The ridge estimate of $\tilde{\boldsymbol{\beta}}_{n+1}$ is

$$\hat{\tilde{\boldsymbol{\beta}}}_{n+1} = (\mathbf{F}^{eT}\mathbf{F}^e + \lambda\mathbf{I})^{-1}\mathbf{F}^{eT}\tilde{\mathbf{x}}^e_{n+1},$$

which minimizes the criterion

$$(\tilde{\mathbf{x}}^e_{n+1} - \mathbf{F}^e\tilde{\boldsymbol{\beta}}_{n+1})^T(\tilde{\mathbf{x}}^e_{n+1} - \mathbf{F}^e\tilde{\boldsymbol{\beta}}_{n+1}) + \lambda\tilde{\boldsymbol{\beta}}^T_{n+1}\tilde{\boldsymbol{\beta}}_{n+1}. \quad (12)$$

This is the same as the PLS criterion in (9) up to a reparametrization $\tilde{\boldsymbol{\beta}}_{n+1} = \boldsymbol{\beta}_{n+1} - \hat{\boldsymbol{\beta}}^{\mathrm{TS}}_{n+1}$. Therefore, the minimizers of the two optimization problems are linked through $\hat{\tilde{\boldsymbol{\beta}}}_{n+1} = \hat{\boldsymbol{\beta}}^{\mathrm{PLS}}_{n+1} - \hat{\boldsymbol{\beta}}^{\mathrm{TS}}_{n+1}$. The ridge estimate in (12) shrinks the corresponding LS estimate toward zero. Correspondingly, our PLS forecast modifies the LS forecast $\hat{\boldsymbol{\beta}}^{\mathrm{LS}}$ by shrinking it towards the TS forecast $\hat{\boldsymbol{\beta}}^{\mathrm{TS}}$. As noted by Hoerl and Kennard, the shrinkage of ridge regression has the beneficial effect of reduced variance. Similar variance reduction is expected for our PLS forecast.

3.3.3. Distributional Forecast Updating. The PLS provides a point forecast updating scheme. For the corresponding distributional updating of $\mathbf{x}^l_{n+1}$, we use a nonparametric bootstrap procedure similar to the one described in §3.2.2. First, by bootstrapping the errors from the TS forecasting models for $\boldsymbol{\beta}_{n+1}$, we obtain B TS forecasts $\hat{\boldsymbol{\beta}}^{\mathrm{TS},\,b}_{n+1}$ as discussed in Appendix A.1. These forecasts in turn lead to B PLS forecasts $\hat{\boldsymbol{\beta}}^{\mathrm{PLS},\,b}_{n+1}$ according to (10).

Then we derive B PLS updates of $\mathbf{x}^l_{n+1}$ as follows:

$$\hat{x}^{\mathrm{PLS},\,b}_{n+1,\,j} = \hat{\beta}^{\mathrm{PLS},\,b}_{n+1,\,1}f_{j1} + \cdots + \hat{\beta}^{\mathrm{PLS},\,b}_{n+1,\,K}f_{jK} + \hat{\epsilon}^b_j, \quad j = m_0 + 1, \ldots, m;\ b = 1, \ldots, B,$$

where $\hat{\boldsymbol{\epsilon}}^b = (\hat{\epsilon}^b_1, \ldots, \hat{\epsilon}^b_m)^T$ is sampled with replacement from the fitted error profiles $\hat{\boldsymbol{\epsilon}}_1, \ldots, \hat{\boldsymbol{\epsilon}}_n$ of the historical data. The interval and density forecasts of $x_{n+1,\,j}$ are then obtained using the empirical distribution of $\{\hat{x}^{\mathrm{PLS},\,b}_{n+1,\,j}\colon 1 \le b \le B\}$. For a $100(1-\alpha)\%$ forecast interval, the two end points are the $\alpha/2$ and $(1-\alpha)/2$ empirical percentiles of $\{\hat{x}^{\mathrm{PLS},\,b}_{n+1,\,j}\}$. The histogram or density estimate of $\{\hat{x}^{\mathrm{PLS},\,b}_{n+1,\,j}\}$ then gives a density forecast of $x_{n+1,\,j}$.

4. Forecasting Performance Comparison

We conduct out-of-sample rolling forecasting exercises on simulated and real data sets to evaluate the proposed forecasting method and compare it with competing methods. Both one-day-ahead forecasting and dynamic intraday updating are considered. The root-unroot method (§2) is applied to all forecasting methods to deal with heteroscedasticity. The methods are described on the root-transformed data, and the unroot transformation is needed to derive forecasts of the call volumes.

Section 4.1 presents several forecasting performance measures. Several competing methods are described in §4.2. Two simulated examples are used in §4.3 to compare our data-driven approach with two model-based approaches. The two case studies are reported separately in §§4.4 and 4.5, using the data described in §2 and the data in Weinberg et al. (2007). Showing the forecast-accuracy comparison on two data sets provides evidence of the robustness and general applicability of our approach.

4.1. Performance Measures

We use two measures to assess and compare point forecast performances. Let N_{ij} denote the call volume on day i during period j and $\widehat{N}_{ij}$ be a forecast of N_{ij}. Suppose we are interested in forecasting the whole intraday profile for day i, as, for example, in one-day-ahead forecasting. Then the root mean squared error (RMSE) and mean relative error (MRE) are defined for day i as follows,

$$\mathrm{RMSE}_i = \sqrt{\frac{1}{m}\sum_{j=1}^{m}(\widehat{N}_{ij} - N_{ij})^2} \quad \text{and}$$

$$\mathrm{MRE}_i = \frac{100}{m}\sum_{j=1}^{m}\frac{|\widehat{N}_{ij} - N_{ij}|}{N_{ij}}.$$

We now discuss how to evaluate our bootstrap-based forecast intervals. As described in §3.2, we first obtain a bootstrap sample of B forecasts of N_{ij}. It is proposed to use $[\widehat{N}_{ij}^{100(\alpha/2)}, \widehat{N}_{ij}^{100(1-\alpha/2)}]$ as one $100(1-\alpha)\%$ forecast interval of N_{ij}, where $\widehat{N}_{ij}^{Q}$ is the Qth quantile of the bootstrap sample. In practice, people commonly use 95% prediction intervals, which correspond to $\alpha = 0.05$. For day i, we compute the 95% coverage probability (COVER) and the average forecast interval width (WIDTH) as follows:

$$\mathrm{COVER}_i = \frac{1}{m}\sum_{j=1}^{m}\mathrm{I}\big(\widehat{N}_{ij}^{2.5} < N_{ij} < \widehat{N}_{ij}^{97.5}\big) \quad \text{and}$$

$$\mathrm{WIDTH}_i = \frac{1}{m}\sum_{j=1}^{m}\big(\widehat{N}_{ij}^{97.5} - \widehat{N}_{ij}^{2.5}\big),$$

with I being the indicator function. We expect the empirical coverage probability to be close to the nominal value, and the narrower the forecast interval, the better.

For dynamic updating, suppose we update our forecast after time period j; then the above measures can be calculated by averaging over only those remaining periods after time period j. Summary statistics (over days in the forecasting set) such as mean, median, and quartiles of the measures can be reported to compare the performance of different forecasting methods.

4.2. Competing Methods

4.2.1. Interday Forecasting.

Benchmark. The first alternative is a simple method currently used in the call center industry (Cleveland and Mayben 2004, Weinberg et al. 2007), which incorporates both day-of-the-week and time-of-day effects in a linear additive model. Specifically, let d_i denote the day of the week of day i, and $x_{ij} = \sqrt{N_{ij} + 1/4}$ with N_{ij} being the call volume during period j of day i. Then the model is

$$x_{ij} = \mu + \alpha_{d_i} + \beta_j + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma^2).$$

We refer to this approach as the historical average (HA) approach, because it essentially uses historical averages of the same day of the week as a forecast for the current day's square-root-transformed call volumes.

A Bayesian Gaussian Model. The second alternative is the Bayesian approach proposed in Weinberg et al. (2007), which assumes N_{ij} is a Poisson random variable with a random rate that depends on both day of the week and time of day. Using the root-unroot method, the following Bayesian Gaussian (BG) model is proposed:

$$\begin{cases} x_{ij} = g_{d_i}(t_j) y_i + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma^2), \\ y_i - \alpha_{d_i} = \beta(y_{i-1} - \alpha_{d_{i-1}}) + \eta_i, \quad \eta_i \sim N(0, \phi^2), \\ s\dfrac{d^2 g_{d_i}(t_j)}{dt_j^2} = \tau_{d_i}\dfrac{dW_{d_i}(t_j)}{dt_j}, \end{cases}$$

where t_j is the center of time period j, $g_{d_i}(\cdot)$ models the smooth intraday call arrival pattern that depends on the day of the week d_i $(=1, \ldots, 5)$, y_i is a random effect with an autoregressive structure adjusting for d_i, and $W_{d_i}(t)$ are independent Wiener processes with $W_{d_i}(0) = 0$ and $\mathrm{var}\{W_{d_i}(t)\} = t$. To fit the model,

Weinberg et al. suggest suitable priors on the parameters, $\alpha_1, \ldots, \alpha_5$; $\tau_1^2, \ldots, \tau_5^2$; σ^2; ϕ^2 and β, and propose a MCMC algorithm to estimate them.

We also consider the following two models, which are related to the above existing approaches. The models may appear to be classical, but they have not been explicitly used in the call center forecasting literature. In addition, we use the two models in the simulation study reported in §4.3.

A Multiplicative Model. We assume the following multiplicative (MUL) model:

$$\begin{cases} x_{ij} = y_i \gamma_{d_i j} + \epsilon_{ij}, \quad \sum_j \gamma_{d_i j} = 1, \quad \epsilon_{ij} \sim N(0, \sigma^2), \\ y_i - \alpha_{d_i} = \beta(y_{i-1} - \alpha_{d_{i-1}}) + \eta_i, \quad \eta_i \sim N(0, \phi^2), \\ i = 1, \ldots, n; \quad j = 1, \ldots, m; \quad d_i = 1, 2, 3, 4, 5. \end{cases} \tag{13}$$

In the call center context, this model assumes that, on the square-root-transformed scale, the daily total (y_i) follows an AR(1) model that depends on the day of the week (α_{d_i}); each weekday has its own intraday arrival profile ($\gamma_{d_i j}$). The model is in essence the same as the Bayesian model used in Weinberg et al. (2007), except that the intraday profiles are not required to be smooth. We provide details about one estimation/forecasting procedure for model (13) in Appendix A.3.

An Additive Model. The following additive (ADD) model is assumed on x_{ij}s:

$$\begin{cases} x_{ij} = \mu + \alpha_i + \beta_j + \gamma_{d_i j} + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma^2), \\ \alpha_i - a_{d_i} = b(\alpha_{i-1} - a_{d_{i-1}}) + \eta_i, \quad \eta_i \sim N(0, \phi^2), \\ \sum_i \alpha_i = \sum_j \beta_j = \sum_i \gamma_{d_i j} = \sum_j \gamma_{d_i j} = \sum_{ij} \gamma_{d_i j} = 0, \\ i = 1, \ldots, n; \quad j = 1, \ldots, m; \quad d_i = 1, 2, 3, 4, 5. \end{cases} \tag{14}$$

This essentially corresponds to a linear additive model with time-of-day (β_j) and daily (α_i) effects as well as their interactions ($\gamma_{d_i j}$), and the day-of-the-week main effects are assumed to be an AR(1) process adjusting for the day of the week (a_{d_i}). Such a model is also reasonable for our call center application. One estimation/forecasting procedure of model (14) is detailed in Appendix A.4.

4.2.2. Dynamic Intraday Updating. Weinberg et al. (2007) propose an updating algorithm for the Bayesian approach, while the HA/MUL/ADD approaches do not perform any intraday updating. Besides the Bayesian updating, we compare our proposed dynamic updating approach with the following two updating methods.

Historical Proportion Method. The first method is a simple updating that combines the HA/MUL/ADD forecasts with historical proportions (HP). Suppose the HA forecast for day $n+1$ is $\hat{\mathbf{x}}_{n+1}^{\mathrm{HA}}$. For an updating point m_0, (1) calculate the ratio R between *the* (*square-root-transformed*) *number of calls received up to time period* m_0 and *the number forecasted up to that time*,

$$R = \frac{\sum_{j=1}^{m_0} x_{n+1,j}}{\sum_{j=1}^{m_0} \hat{x}_{n+1,j}^{\mathrm{HA}}};$$

and (2) update the HA forecasts for the remaining of the day as

$$\hat{x}_{n+1,j}^{\mathrm{HP}} = R\, \hat{x}_{n+1,j}^{\mathrm{HA}}, \quad j = m_0 + 1, \ldots, m.$$

We can replace the HA forecast with the MUL/ADD forecast in the above algorithm and refer to the corresponding updates as HPM and HPA, respectively.

A nontrivial observation needs to be pointed out here. Because of the specific forms/estimations of the HA model and the MUL model, for the same forecasting day, the MUL forecast turns out to be the HA forecast multiplied by a constant that takes into account the interday correlation. Hence, MUL and HA produce the same HP updates, because the constant gets canceled out, as one can see from the above algorithm.

Multiple Regression Method. The second method is based on a multiple linear regression (MR). Let $\mathbf{X}^e = (\mathbf{x}_1^e, \ldots, \mathbf{x}_n^e)^T$ be a $n \times m_0$ matrix containing the historical square-root-transformed count up to period m_0 and $\mathbf{X}^l = (\mathbf{x}_1^l, \ldots, \mathbf{x}_n^l)^T$ be the corresponding $n \times (m - m_0)$ matrix after period m_0. The MR method works as follows:

- Regress $\mathbf{X}^l$ on $\mathbf{X}^e$ and obtain the regression coefficient matrix as

$$(\mathbf{X}^{eT}\mathbf{X}^e)^{-1}\mathbf{X}^{eT}\mathbf{X}^l;$$

- Update the forecast of $\mathbf{x}_{n+1}^l$ as the model prediction of $\mathbf{x}_{n+1}^e$,

$$\hat{\mathbf{x}}_{n+1}^{l\mathrm{MR}} = \mathbf{X}^{lT}\mathbf{X}^e(\mathbf{X}^{eT}\mathbf{X}^e)^{-1}\mathbf{x}_{n+1}^e.$$

Note that here we are updating the multivariate response $\mathbf{x}_{n+1}^{l}$ simultaneously, which is equivalent to updating each element of $\mathbf{x}_{n+1}^{l}$ separately.

4.3. Two Simulated Examples

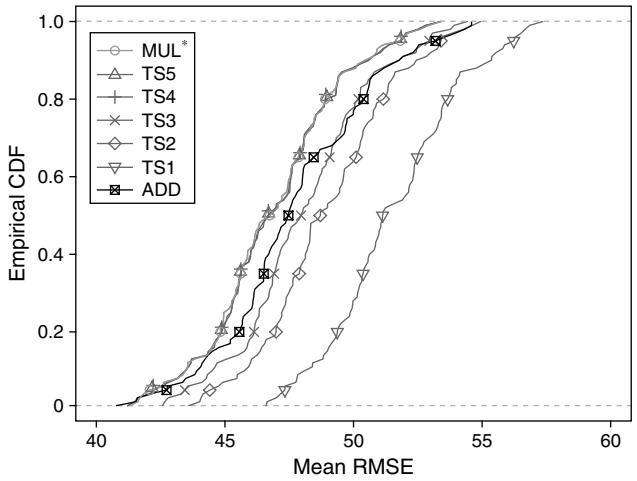
The proposed forecasting approach is data driven, which assumes no probabilistic model on the underlying data-generating process. Here we use two simulation examples to illustrate that our approach is actually very robust to various model assumptions, in the sense that it performs almost as well as the method that uses the true data-generating model. On the other hand, the method based on a specific model gives an inferior forecast when the model assumed is different from the underlying data-generating model. In this case, we refer to the assumed model as a misspecified model and the data-generating model as the true model. Note that such a distinction of true/misspecified models is only feasible in simulation studies. For real data, we never know what the true model is.

For the two examples, the MUL model (13) and the ADD model (14) are respectively considered as the true data-generating model. To choose reasonable model parameters, we fit the models to the data described in §2 (on the square-root-transformed scale). The obtained parameter estimates are rounded for use in the simulations. In particular, some estimated model parameters are reported below:

- For model (13): $\alpha_1 = 2{,}045$; $\alpha_2 = 1{,}935$; $\alpha_3 = 1{,}900$; $\alpha_4 = 1{,}890$; $\alpha_5 = 1{,}885$; $\beta = 0.65$; $\sigma = 0.5$; $\phi = 35$; $n = 200$; and $m = 68$.
- For model (14): $\mu = 28.5$; $b = 0.65$; $\sigma = \phi = 0.5$; $n = 200$; and $m = 68$.

Data sets are generated using each of the two models. Each simulation example is repeated 100 times. During each simulation run, a rolling out-of-sample forecasting exercise is performed with the last 50 days as the forecasting set; and for each day in the forecasting set, the preceding 150 days are used to derive the forecast. Seven forecasting methods are applied to both simulations, including forecasting using the MUL model (13), forecasting using the ADD model (14), and our methods TS1–TS5. The parameters involved in all the methods are reestimated for each update in the rolling forecast exercise for each simulated data set. Note that when the data are generated using the MUL model, the ADD model is a misspecified model, and vice versa. In both cases, our TS methods do not use the model assumptions.

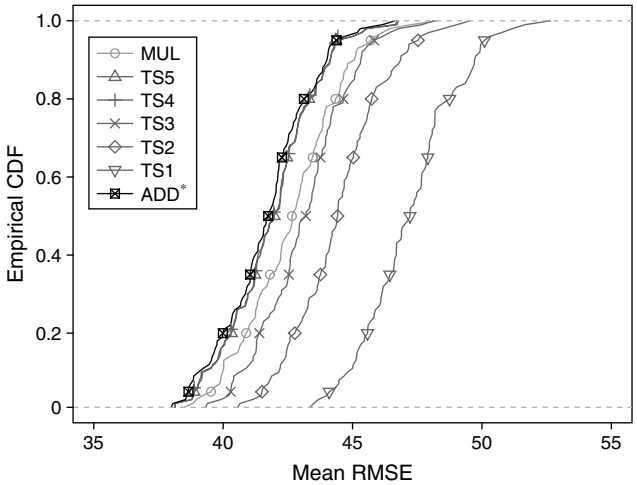
Figure 2 **Comparison of Empirical CDF of Mean RMSE for Several Forecasting Methods Under the Multiplicative Simulation Model (13)**



Note. TS4 and TS5 perform comparably to the true model, while the wrong additive model (14) performs worse.

For each simulation run, we calculate the mean RMSE and mean MRE (%) over the forecasting set for each method. Based on the 100 runs, we calculate the empirical cumulative distribution functions (CDF) of the mean RMSE, which are plotted in Figure 2 when the data are generated from the MUL model and in

Figure 3 **Comparison of Empirical CDF of Mean RMSE for Several Forecasting Methods Under the Additive Simulation Model (14)**



Note. TS4 and TS5 perform comparably to the true model, while the wrong multiplicative model (13) performs worse.

Figure 3 when the data are from the ADD model. The true model is highlighted in the legend using an asterisk. The results for mean MRE (%) are similar and are not shown. To help read these graphs, note that if the CDF of one random variable is above that of the other random variable, then the first variable is stochastically smaller than the second one.

We observe that the forecasts based on mis-specified models have stochastically larger mean RMSE and mean MRE (%) than the forecasts based on the correctly specified models. Hence, inferior forecasts are obtained when using the misspecified models. For our proposed TS methods, when the number of feature vectors K increases, both performance measures get stochastically smaller and smaller; and TS4 and TS5 have about the same forecasting performance, which are competitively close to the true model. Note that the CDFs of TS4/TS5 and the true model are hard to distinguish in each plot. The robust good performance of the proposed method is important in practice because the proposed method does not require the information of the true underlying model.

4.4. Case Study I: Our Data

In this section, we use our data (§2) to illustrate the proposed methods for interday forecasting as well as intraday updating. To implement the sophisticated Bayesian approach in Weinberg et al. (2007), one needs the expertise with programming MCMC. Hence, we do not apply that approach to our data. Instead, we apply our forecasting methods to their data in §4.5 to compare with their published results.

As described earlier, there are in total 200 intraday call volume profiles (aggregated into quarter hours) in our data. To perform the rolling forecast, we use the final 50 days as the forecasting set; for each day in the forecasting set, we use the 150 preceding days as the historical data to reestimate the model and generate the forecast.

4.4.1. Interday Forecasting. Our approach for one-day-ahead forecasting as described in §3.2 relies on a dimension reduction by SVD. The original high-dimensional vector TS is represented using a small number K of basis vectors. In this study, we consider $K = 1, \ldots, 4$ and denote the methods as TS1, ..., TS4, respectively. The scree plot based on the first 150 days suggests that we need only consider up to five pairs of feature vectors, which makes sense in that the five weekdays appear to have different arrival patterns (Figure 1(a)). We note that TS4 and TS5 turn out to have very similar forecasting performance.

To help interpret the feature factors, Figure 4 plots the first four pairs of intraday feature vectors and interday feature series, all of which are extracted from the first 150 days. The first intraday feature summarizes the average intraday call arrival pattern during a regular weekday; the first interday feature illustrates a strong weekly effect and is highly correlated with total daily call volume. The remaining pairs are second-order effects, as indicated by the much smaller scales of the second to fourth interday feature series. The corresponding intraday features capture certain contrasts between arrival patterns of different within-day segments. For example, the second intraday feature compares morning arrivals with afternoon and evening arrivals; when combined with the second interday feature, they suggest the following specific phenomenon: Fridays usually have above-average call volumes in early mornings and below-average volumes in late afternoons and evenings; the opposite is true for Mondays.

Further analysis of the interday feature series suggests that they can be modeled separately by an AR(1) TS model with a day-of-the-week effect. Specifically, for $k = 1, \ldots, 4$, we model the series $\{\beta_{ik}\}$ by

$$\beta_{ik} = a_k(d_{i-1}) + b_k \beta_{i-1,k} + \epsilon_i,$$

where the varying intercept $a_k(d_{i-1})$ depends on the day of the week of day $i-1$, and the ϵ_is are independent identically distributed errors with mean 0. The same model has been obtained in Shen and Huang (2005) when they analyzed the arrival data from the same call center for a different year. A closely related model is used in Weinberg et al. (2007).

Table 1 compares summary statistics of the RMSE and MRE of the forecasts from the HA/MUL/ADD methods and the four TS methods. The best forecasting method is highlighted in boldface for each summary statistic separately. Based on RMSE, all TS methods show substantial improvement over the benchmark HA, while TS4 gives the best overall performance. As for MRE, TS1 performs comparably with HA, because MRE is more sensitive to forecasting performance at the beginning and the end of the day,

Figure 4 **First Four Pairs of Intraday Feature Vectors and Interday Feature Series**

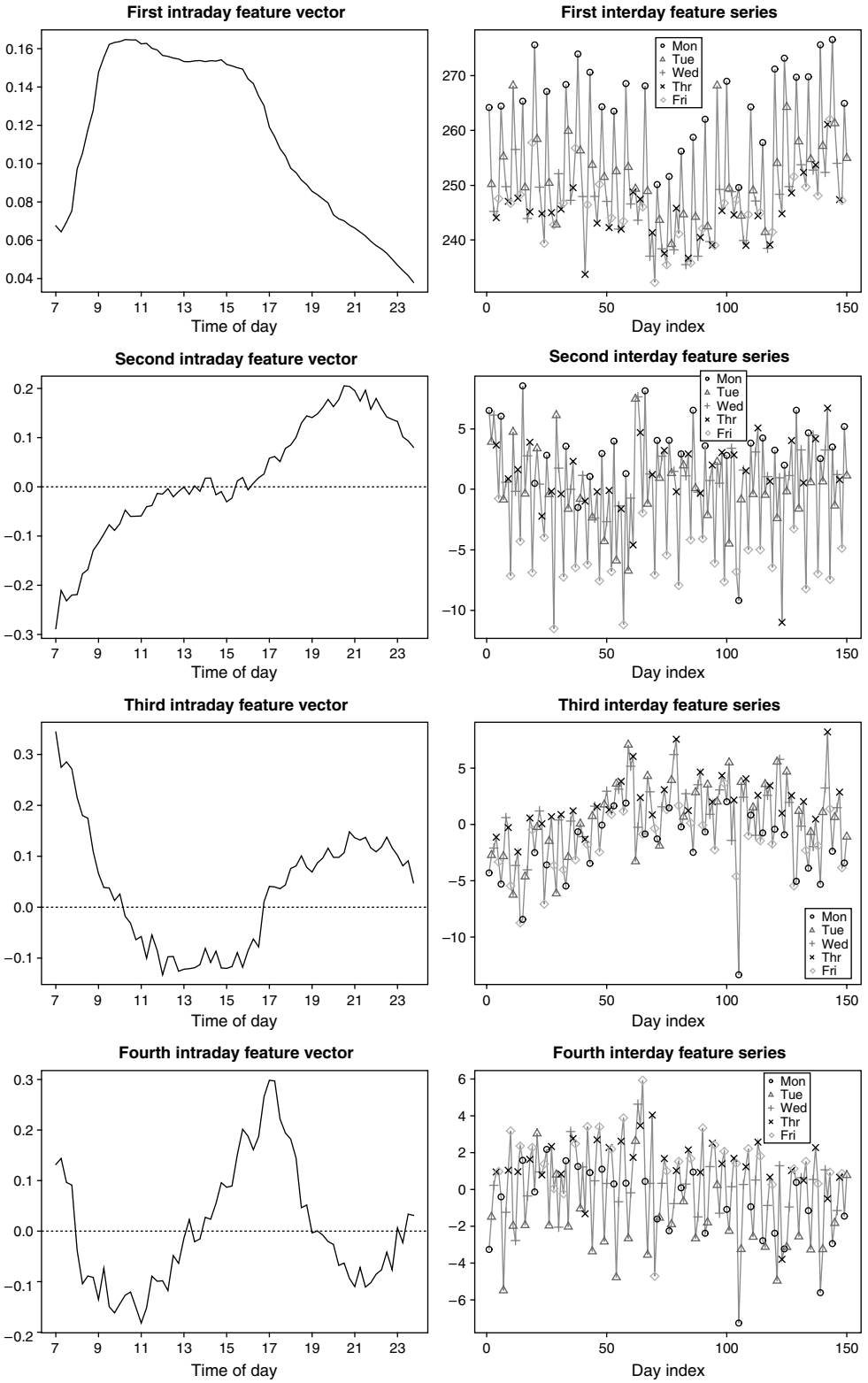


Table 1 **Summary Statistics (Mean, Median, Lower Quartile Q1, Upper Quartile Q3) of RMSE and MRE in a Rolling Forecast Exercise**

	RMSE				MRE (%)			
	Q1	Median	Mean	Q3	Q1	Median	Mean	Q3
HA	43.81	65.57	72.89	90.04	5.2	6.7	7.3	8.7
ADD	41.71	52.56	59.28	63.67	4.8	5.6	6.5	6.9
MUL	41.55	48.54	57.49	59.33	4.6	5.3	**6.1**	**6.6**
TS1	42.91	53.52	60.09	63.76	5.2	6.6	7.4	8.6
TS2	**39.57**	51.29	57.96	61.04	4.9	5.6	6.6	7.1
TS3	42.90	51.96	57.65	**58.01**	4.8	5.5	6.4	6.9
TS4	41.85	**48.23**	**56.88**	59.27	**4.6**	**5.3**	6.2	6.7

Note. The forecasting set contains 50 days.

when call volumes are low. At least two feature vectors are needed to capture the difference among different weekdays in those regions. Both ADD and MUL also improve significantly over HA, with ADD performing similarly to TS2 and MUL similarly to TS4.

We also compare the performance of the 95% bootstrap prediction intervals for our TS methods. The bootstrap procedure described in §3.2 is implemented with $B = 1{,}000$. The average empirical coverage probabilities of the prediction intervals are 94.4%, 94.9%, 94.8%, and 94.6%, respectively, which are quite accurate. Table 2 compares the distribution of the average prediction interval width among the four TS approaches, which suggests that the intervals get stochastically narrower as one increases the number of components, K, in the forecasting model; hence, the forecasting becomes more and more precise.

4.4.2. Dynamic Intraday Updating. For each day in the forecasting set, we consider dynamically updating the forecast every half hour for the rest of the current day, starting from 8:00 AM, using available information up to that time. In addition to the HP/HPM/HPA and MR updating methods (§4.2.2), we consider the three updating methods described in §3.3, namely TS, LS, and PLS; see (5), (7), and (11). We expect the PLS method to give the best forecast performance. This is confirmed empirically below.

Table 2 **Summary Statistics (Mean, Median, Lower Quartile Q1, Upper Quartile Q3) of Average Width in a Rolling Forecast Exercise**

	Q1	Median	Mean	Q3
TS1	234.2	242.0	243.1	249.4
TS2	225.4	232.3	233.3	242.4
TS3	220.3	226.9	227.8	234.0
TS4	**215.1**	**221.8**	**223.1**	**229.7**

Note. The forecasting set contains 50 days. TS4 performs the best. The intervals are stochastically narrower as one increases K.

To implement these three methods, we need to choose K, the number of intraday feature vectors for dimension reduction. In light of the one-day-ahead forecasting performance presented in §4.4.1, we decide to use $K = 4$, and the three methods are therefore denoted as TS4, LS4, and PLS4. For each day in the forecasting set, we use its 150 preceding days to derive the TS forecast TS4. We employ the current day's incoming call volume information in addition to the 150 preceding days to produce the LS4 and PLS4 forecasts. The intraday updating starts at 8:00 AM, corresponding to $m_0 = 4$, so there is enough data to estimate the parameters for LS.

To use the PLS updating approach, we need to decide on the value of the penalty parameter λ at each updating point. To this end, we use the beginning 150 days in our data set as the training set and perform an out-of-sample rolling forecast exercise on this training set. Note that this training set does not overlap with the forecasting set we initially hold out for out-of-sample forecast evaluation. For the purpose of selecting λ, the last one-third (i.e., 50 days) of the training set is used as a rolling hold-out sample. For a given day in this hold-out sample, we use the preceding 100 days to extract the interday feature series to generate the TS forecast TS4. The extracted intraday feature vectors and the obtained TS4 forecast are used for all updating points within that day.

Fix an updating point, for example, 10:00 AM For each given λ, we construct the PLS updating for the given day. Pick a performance measure such as RMSE or MRE. Compute the performance measure for all days in the hold-out sample and calculate the average value. Select λ by minimizing this average performance measure over a grid of candidate values. In this study, we choose λ from $\{0, 10, \ldots, 10^9\}$. When $\lambda = 0$, the PLS approach is the same as the LS approach. As λ goes to infinity, the PLS approach should converge to the TS approach, which is essentially achieved when $\lambda = 10^9$ in our study. Second-round tuning can be carried out over a finer grid around the selected λ to improve performance if needed.

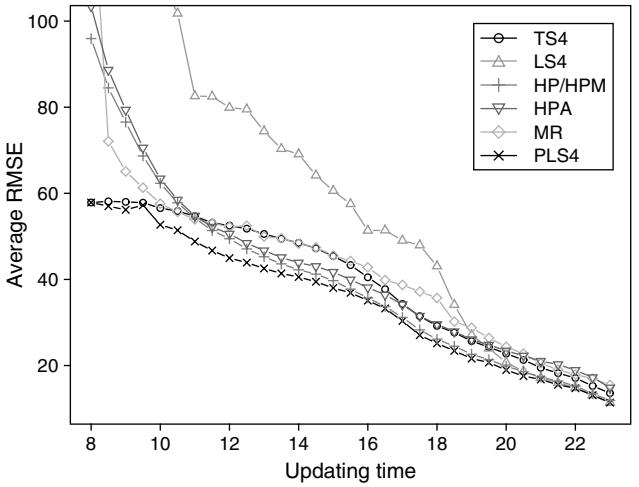
For a given updating point, the selected λ is kept fixed for all days in the forecasting set to generate the PLS forecast for the corresponding updating. We observe that, for every selection criterion (RMSE or MRE), the selected λ in general has a decreasing trend as the updating time progresses into the later part of the day. Note that a smaller λ value would let the PLS forecast put more weight on the LS forecast. Thus, the observed decreasing trend for the selected λ makes intuitive sense because, as time progresses, more information is available about the current day's call arrival pattern, and the forecaster needs less influence from the TS forecast.

Figure 5 plots the average of RMSE as functions of the updating points for the seven approaches, where the averages are taken over the 50 days in the forecasting set. TS4 is treated as the benchmark, as it essentially performs no intraday updating. It just uses the information up to the end of the previous day and does not employ the current day's information. Note that RMSE is calculated at each updating point because we are interested in the performance of the updated forecasts at all time periods after the updating point. For PLS4, RMSE is used to select λ because it is the performance measure in Figure 5.

From the plot, PLS4 is clearly the winner, which has the smallest error measure at all updating points. The improvement over TS4 becomes significant as early as updating at 10:00 AM, which is three hours into the working day. Thus, the significant improvement of PLS4 over TS4 shows the benefit of dynamic updating using the current day's information. One can also see that LS4 performs rather badly at the beginning of the day, which is because of the high collinearity between the initial parts of the higher-order intraday feature vectors. This also justifies the necessity for penalization. The MR updating performs no better than TS4 and much worse prior to 10:00 AM; this suggests that the interday dependence plays a much bigger role in improving forecasting, which is not incorporated in MR. As discussed in §4.2.2, HP and HPM lead to the same updates. The HP/HPM approach behaves rather badly in early mornings and starts to beat TS4 at 11:00 AM; it performs worse than PLS4 until 4:00 PM, when the performance converges to PLS4's. A call center manager needs enough information for the present day for the HP/HPM updating to work well. To the extent that he or she has that information, HP/HPM is a nice candidate to perform updating because of its simplicity. The HPA approach performs uniformly worse than HP/HPM.

Below, we look at the 10:00 AM updating and the noon updating to get some idea about how different approaches perform for individual updating. Table 3 presents summary statistics of the RMSE of the forecasts from TS4, LS4, HP/HPM, HPA, MR, and PLS4. The averages are calculated over the 50 days in the forecasting set. For a fair comparison, we only use data after noon when calculating the RMSE. The superior performance of PLS4 over the other methods is again quite clear. The LS4 method has the worst performance, much worse even than TS4. For every method except TS4, updating later always improves

Figure 5 **Comparison of Mean RMSE as Functions of the Updating Points, Calculated over the 50 Days in the Forecasting Set**



Note. PLS4 performs the best.

Table 3 **Summary Statistics (Mean, Median, Lower Quartile Q1, and Upper Quartile Q3) of RMSE for the 10:00 AM and Noon Updatings**

	10:00 AM updating				Noon updating			
	Q1	Median	Mean	Q3	Q1	Median	Mean	Q3
TS4	36.15	41.74	52.55	54.60	36.15	41.74	52.55	54.60
LS4	83.28	106.90	119.20	128.10	54.92	73.83	79.90	97.46
HP/HPM	37.64	49.76	59.30	70.10	35.68	39.60	49.46	49.74
HPA	37.66	52.17	60.13	71.11	35.84	41.05	50.57	50.38
MR	41.06	48.24	56.50	62.96	40.45	47.04	52.21	52.20
PLS4	37.04	42.08	50.77	58.07	**33.69**	**37.87**	**44.96**	**44.25**

Note. PLS4 outperforms the other methods.

Table 4 **Summary Statistics (Mean, Median, Lower Quartile Q1, and Upper Quartile Q3) of Average Width of 95% Forecast Intervals Among Three Approaches: TS4 (i.e., No Updating), PLS4 (10:00 AM Updating), and PLS4 (Noon Updating)**

	Q1	Median	Mean	Q3
TS4	193.1	199.8	200.1	205.2
PLS4 (10:00 AM)	171.2	175.8	177.1	182.0
PLS4 (noon)	**155.5**	**161.2**	**161.2**	**164.6**

the forecasting accuracy. Note that TS4 performs no intraday updating.

Not only can intraday updating reduce forecasting errors, but it can also narrow the width of forecast intervals while maintaining the same coverage probability. To illustrate this point, Table 4 reports a width comparison of 95% forecast intervals among three forecasting approaches: TS4 where no updating is performed, PLS4 with 10:00 AM updating, and PLS4 with noon updating. The intervals are stochastically narrower as more observations are used for the updating. The two updatings reduce the average interval width by 11.5% and 9.0%, sequentially. The average empirical coverage probability of the three intervals is, respectively, 93.7%, 93.3%, and 94.0%, close to the nominal value.

We recall that the PLS criterion (8) involves two terms. To understand their relative importance, one can evaluate the criterion at the selected λ and $\hat{\boldsymbol{\beta}}_{n+1}^{\mathrm{PLS}}$ and calculate the percentage of each term relative to the total. For the current study, the first term accounts for 89.2% and 90.1% of the total for the 10:00 AM and noon updatings, respectively. The first term becomes more important as one postpones the updating, which makes sense.

4.5. Case Study II: Weinberg's Data

In this section, we compare our method with the BG approach proposed by Weinberg et al. (2007). To compare with their published results, we apply our method to the same cleaned data. Their data consist of five-minute aggregated call volume profiles from 7:00 AM to 9:05 PM for 164 days; thus, the dimensionality is 169. In Weinberg et al. (2007), the data set was cleaned by excluding anomalous days such as holidays. For the rolling forecast, the final 64 days are used as the forecasting set. For each day in the forecasting set, its 100 preceding days are considered as the historical data.

4.5.1. Interday Forecasting.

For our forecasting approach, we consider two methods, one using three intraday feature vectors (TS3) and the other using five (TS5). The BG approach assumes a separate intraday pattern for each weekday, that corresponds to the model with five intraday feature vectors, and the use of three intraday feature vectors has already given good forecast results. Note that Weinberg et al. (2007) constrain the intraday patterns to be smooth, but we do not impose such a constraint. We again use the varying-coefficient AR(1) model to forecast the interday feature series.

Table 5 reports summaries of the RMSE and MRE of the forecasts from various competing approaches. The results for BG are quoted from Weinberg et al. (2007). As expected, all the approaches improve greatly over HA. It is also observed that our approaches perform very competitively with the BG approach. Through personal communication, we find out that the Bayesian approach might take a long time to run because of the many MCMC iterations needed. As a comparison, our approach is computationally fast, which makes it amenable for real-time updating. The results also suggest that a forecasting model with three intraday features (TS3) might be sufficient. Both ADD and MUL methods perform similarly to our TS3 approach, with TS3 using the fewest intraday features. We point out in §4.2.1 that the only major difference between MUL and BG is that BG assumes smooth intraday patterns. The close performance between them suggests that smoothing is not necessary for the one-day-ahead forecasting of the current data.

Table 5 **Summary Statistics (Mean, Median, Lower Quartile Q1, Upper Quartile Q3) of RMSE and MRE for Six Competing Approaches: HA, ADD, MUL, BG, TS3, and TS5**

	RMSE				MRE (%)			
	Q1	Median	Mean	Q3	Q1	Median	Mean	Q3
HA	16.22	19.12	21.32	21.82	7.8	9.3	10.1	11.6
ADD	14.49	15.89	18.46	19.84	7.0	7.6	8.5	8.7
MUL	14.42	15.81	18.28	19.97	6.9	7.3	8.4	8.7
BG	14.25	15.83	18.28	**19.83**	7.0	7.4	8.4	**8.5**
TS3	14.34	**15.76**	18.19	20.22	6.9	7.5	8.5	8.8
TS5	**14.20**	15.82	**18.16**	19.99	**6.8**	**7.3**	**8.3**	8.6

Note. The forecasting exercise is performed on the data used in Weinberg et al. (2007).

Table 6 **Summary Statistics (Mean, Median, Lower Quartile Q1, Upper Quartile Q3) of RMSE and Average Width of 95% Prediction Intervals for BG and PLS3**

	RMSE				Average width			
	Q1	Median	Mean	Q3	Q1	Median	Mean	Q3
10:00 AM								
BG	14.00	15.50	17.86	19.87	58.87	60.74	61.11	63.50
PLS3	**13.31**	14.87	16.48	17.25	59.58	61.17	61.32	62.81
Noon								
BG	13.56	14.80	16.59	16.58	58.78	60.42	60.80	62.34
PLS3	13.33	**14.60**	**16.13**	**16.39**	**58.14**	**59.27**	**59.56**	**61.38**

Note. The forecasting exercise is performed on the data used in Weinberg et al. (2007).

4.5.2. Dynamic Intraday Updating. Weinberg et al. (2007) also propose a Bayesian procedure to perform intraday updating. In this section, we are interested in comparing their updating procedure (BG) with our PLS updating approach (PLS3).

The penalty parameter λ for the PLS updating is chosen using the data-driven procedure described in §4.4.2. We use the last 30 days of the 100-day training set as a rolling hold-out sample. For each day in this sample, its preceding 70 days and the early part of the current day are used to generate the updated forecast for some candidate values of λ. We then select λ as the value that minimizes some performance measure averaged over the hold-out sample.

To compare with the results reported in Weinberg et al. (2007), we look at the 10:00 AM updating and the noon updating, and the forecasting measures are calculated only using data after noon. Table 6 reports the summaries of RMSE for the BG approach (quoted from their paper) and PLS3, which suggest that PLS3 improves slightly over BG in all cases.

The Bayesian approach naturally generates prediction intervals using the MCMC sample. To compare the performance of the intervals, summaries of the average width for 95% prediction intervals are presented in Table 6 for BG and PLS3. We observe that PLS3 with noon updating leads to the narrowest intervals in all categories. All the methods have approximately the right coverage probability, with the empirical averages around 94%.

5. Conclusion

Our approach to forecasting call arrivals is based on the viewpoint that the intraday call volume profiles form a vector TS. This clearly shows that the data have a two-way temporal structure, that is, the interday and intraday variations. Such a structure exists in other applications as well. The SVD, used as a dimensionality reduction tool, effectively separates these two types of variations. The intraday variation is summarized by a few intraday feature vectors that can be extracted from historical data, and the interday variation is modeled by a few interday feature TS (see §3.1). The effectiveness of SVD for dimension reduction is shown in our data examples: Three pairs of interday and intraday features are sufficient to generate good forecasts for a 68 (or 169)-dimensional TS.

Assuming that the intraday features will not change in the near future, the problem of forecasting intraday profiles is reduced to forecasting the interday features. We propose using univariate TS techniques to forecast the interday feature series to generate one- or multiday-ahead forecasts. For dynamic intraday updating, we have considered three ways to update the interday feature forecasts: (1) Apply the interday forecast, using information available up to the end of the previous day (i.e., no updating); (2) Fit a LS regression using the observed portion of the current day's volume profile on the corresponding components of the intraday feature vectors (i.e., LS updating); and (3) Combine the current day's available information with the previous day's information using PLS (i.e., PLS updating). Not surprisingly, the third method performs best and is recommended.

Our methods show improvements over various alternatives in two out-of-sample forecasting comparisons using real data. Despite the good performance, our methods are not much more complicated than the alternatives, and are easy to implement. SVD can be obtained using any software that does matrix computations such as MATLAB and SPLUS/R, and solving PLS is as straightforward as performing a ridge regression. The R codes for performing our forecasting and updating methods, as well as the data, are available from the authors on request. Our methods generate very competitive or slightly better forecasts when compared with the sophisticated Bayesian approach recently proposed by Weinberg et al. (2007). Another advantage of our approach is its fast computation, which makes it a feasible technique for real-time forecasting and updating.

A natural direction for future research is to combine our dimensionality reduction idea with stochastic modeling, such as modeling the arrival process as an inhomogeneous Poisson process (Brown et al. 2005). In this paper, all the methods considered use a square-root transformation to overcome the nonconstant variance problem. Data transformation is not necessary if a distribution-based approach is used in the spirit of stochastic modeling. One possibility is to model the number of arrivals as a Poisson random variable with a random rate having both interday and intraday dependence, which we can model using the dimension reduction idea in this paper. Because the Poisson rate is not observable, such an approach will inevitably be more complex than the current approach. It would be interesting to investigate whether the distributional approach can yield better forecasts.

This paper focuses on one-day-ahead forecasting and intraday dynamic updating. A closely related and interesting problem is two-week-ahead forecasting, which has recently gained some research interest (Taylor 2008, Aldor-Noiman 2006). However, because existing empirical results remain rare, a careful systematic study is needed. We are currently working on extending our framework to this problem. Initial empirical analysis has shown promising results.

Appendix

A.1. Bootstrapping the βs in (4) of §3.2.2

For each k, we describe how to obtain the B bootstrap series $\{\hat{\beta}^b_{n+1,k}, \ldots, \hat{\beta}^b_{n+h,k}\}$, $1 \le b \le B$, as used in (4). Suppose the TS model for $\{\beta_{ik}\}$ can be written as

$$\beta_{ik} = \mathscr{M}(\mathscr{H}_{ik}) + \epsilon_{ik},$$

where $\mathscr{H}_{ik} = \{\beta_{i-1,k}, \beta_{i-2,k}, \ldots, \beta_{1,k}\}$ represents the collection of historical data from the series $\{\beta_{ik}\}$ up to time $i-1$, $\mathscr{M}$ represents the function that gives the prediction of β_{ik} based on the historical data, and ϵ_{ik} is the model error. For example, for an AR(1) model,

$$\mathscr{M}(\mathscr{H}_{ik}) = a_0 + a_1\beta_{i-1,k}.$$

Let $\widehat{\mathscr{M}}$ denote the fitted function using the data and let $\hat{\epsilon}_{ik} = \beta_{ik} - \widehat{\mathscr{M}}(\mathscr{H}_{ik})$ be the fitted model errors. The bootstrapped model errors $\epsilon^b_{n+1,k}$, $1 \le b \le B$, are obtained by sampling with replacement from the set $\{\hat{\epsilon}_{n,k}, \hat{\epsilon}_{n-1,k}, \ldots, \hat{\epsilon}_{1,k}\}$. A bootstrap sample of $\beta_{n+1,k}$ is then obtained using $\beta^b_{n+1,k} = \widehat{\mathscr{M}}(\mathscr{H}_{n+1,k}) + \epsilon^b_{n+1,k}$, $1 \le b \le B$.

We generate the bootstrap sample of $\beta_{n+h,k}$ for $h > 1$ in a sequential manner, assuming availability of $\{\beta^b_{n+1,k}, \ldots, \beta^b_{n+h-1,k}\}$. The historical information set used to generate $\beta_{n+h,k}$ depends on b and is defined as $\mathscr{H}^b_{n+h,k} = \{\beta^b_{n+1,k}, \ldots, \beta^b_{n+h-1,k}\} \cup \mathscr{H}_{n+1,k}$. The bootstrap model errors $\epsilon^b_{n+h,k}$, $1 \le b \le B$, are obtained by sampling with replacement from the set $\{\hat{\epsilon}_{n,k}, \hat{\epsilon}_{n-1,k}, \ldots, \hat{\epsilon}_{1,k}\}$. A bootstrap sample of $\beta_{n+h,k}$ is then generated using $\beta^b_{n+h,k} = \widehat{\mathscr{M}}(\mathscr{H}_{n+h,k}) + \epsilon^b_{n+h,k}$, $1 \le b \le B$.

A.2. Proof of Theorem 1

This can be easily seen as follows:

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{\mathrm{PLS}}_{n+1} &= (\mathbf{F}^{eT}\mathbf{F}^e + \lambda\mathbf{I})^{-1}\mathbf{F}^{eT}\mathbf{F}^e(\mathbf{F}^{eT}\mathbf{F}^e)^{-1}\mathbf{F}^{eT}\mathbf{x}^e_{n+1} \\ &\quad + \lambda(\mathbf{F}^{eT}\mathbf{F}^e + \lambda\mathbf{I})^{-1}\hat{\boldsymbol{\beta}}^{\mathrm{TS}}_{n+1} \\ &= (\mathbf{F}^{eT}\mathbf{F}^e + \lambda\mathbf{I})^{-1}\mathbf{F}^{eT}\mathbf{F}^e\hat{\boldsymbol{\beta}}^{\mathrm{LS}}_{n+1} + \lambda(\mathbf{F}^{eT}\mathbf{F}^e + \lambda\mathbf{I})^{-1}\hat{\boldsymbol{\beta}}^{\mathrm{TS}}_{n+1} \\ &\equiv A(\lambda)\hat{\boldsymbol{\beta}}^{\mathrm{LS}}_{n+1} + (I - A(\lambda))\hat{\boldsymbol{\beta}}^{\mathrm{TS}}_{n+1}. \quad \square\end{aligned}$$

A.3. Estimation and Forecasting of Model (13)

The model is multiplicative, which means that one needs to use nonlinear LS to derive estimates. Below we provide a set of simple estimates, which are close to the true LS estimates (Brown et al. 2005).

$$\begin{cases} \hat{y}_i = x_{i\cdot} = \sum_j x_{ij}; \quad \hat{\alpha}_{d_i} = \dfrac{\sum_{i':\, d_{i'} = d_i} x_{i\cdot}}{\#\{i':\, d_{i'} = d_i\}}; \\ \hat{\gamma}_{d_i j} = \dfrac{\sum_{i':\, d_{i'} = d_i} x_{i'j}}{\#\{i':\, d_{i'} = d_i\}\hat{\alpha}_{d_i}}; \\ \text{The AR(1) coefficient } \beta \text{ can then be estimated} \\ \quad \text{using linear regression.} \end{cases}$$

For day $n+1$, the forecast is then

$$x_{n+1,j} = \hat{y}_{n+1}\hat{\gamma}_{d_{n+1}j},$$

where $\hat{y}_{n+1}$ is forecasted to be $\hat{\alpha}_{d_{n+1}} + \hat{\beta}(\hat{y}_n - \hat{\alpha}_{d_n})$.

A.4. Estimation and Forecasting of Model (14)

Here we provide a set of LS estimates for the model parameters.

$$\begin{cases} \hat{\mu} = \bar{x}_{\cdot\cdot} = \dfrac{\sum_{ij} x_{ij}}{nm}; \\ \hat{\alpha}_i = \bar{x}_{i\cdot} - \bar{x}_{\cdot\cdot} = \dfrac{\sum_j x_{ij}}{m} - \bar{x}_{\cdot\cdot}; \quad \hat{\beta}_j = \bar{x}_{\cdot j} - \bar{x}_{\cdot\cdot} = \dfrac{\sum_i x_{ij}}{n} - \bar{x}_{\cdot\cdot}; \\ \hat{a}_{d_i} = \dfrac{\sum_{i':\, d_{i'} = d_i} \bar{x}_{i'\cdot}}{\#\{i':\, d_{i'} = d_i\}} - \bar{x}_{\cdot\cdot}; \\ \hat{\gamma}_{d_i j} = \dfrac{\sum_{i':\, d_{i'} = d_i} \bar{x}_{i'j}}{\#\{i':\, d_{i'} = d_i\}} - \hat{\mu} - \hat{a}_{d_i} - \hat{\beta}_j, \\ \text{The AR(1) coefficient } b \text{ can then be estimated} \\ \quad \text{using linear regression.} \end{cases}$$

Because of the normality assumption of the model errors, the above estimates should be close to the maximum likelihood estimates.

For day $n+1$, the forecast is then

$$x_{n+1,j} = \hat{\mu} + \hat{\alpha}_{n+1} + \hat{\beta}_j + \hat{\gamma}_{d_{n+1}j},$$

where $\hat{\alpha}_{n+1}$ is forecasted to be $\hat{a}_{d_{n+1}} + \hat{b}(\hat{\alpha}_n - \hat{a}_{d_n})$.

Acknowledgments

The authors thank the editor, the associate editor, and three reviewers whose comments have greatly improved the scope and presentation of the paper. The authors also thank Larry Brown, Noah Gans, Avi Mandelbaum, and Linda Zhao for their valuable suggestions. The first author's work is partially supported by National Science Foundation (NSF) Grant DMS-0606577. The second author's work is partially supported by NSF Grant DMS-0606580.

References

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

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

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

Betts, A., M. Meadows, P. Walley. 2000. Call centre capacity management. *Internat. J. Service Indust. Management* **11** 185–196.

Bianchi, L., J. Jarrett, R. C. Hanumara. 1993. Forecasting incoming calls to telemarketing centers. *J. Bus. Forecasting Methods Systems* **12** 3–12.

Bianchi, L., J. Jarrett, R. C. Hanumara. 1998. Improving forecasting for telemarketing centers by ARIMA modelling with intervention. *Internat. J. Forecasting* **14** 497–504.

Box, G. E. P., G. M. Jenkins, G. C. Reinsel. 1994. *Time Series Analysis: Forecasting and Control*, 3rd ed. Prentice Hall, Upper Saddle River, NJ.

Brown, L. D., N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, L. Zhao. 2005. Statistical analysis of a telephone call center: A queueing-science perspective. *J. Amer. Statist. Assoc.* **100** 36–50.

Cleveland, B., J. Mayben. 2004. *Call Center Management on Fast Forward: Succeeding in Today's Dynamic Inbound Environment*. Call Center Press, Annapolis, MD.

Cottet, R., M. Smith. 2003. Bayesian modeling and forecasting of intraday electricity load. *J. Amer. Statist. Assoc.* **98** 839–849.

Diebold, F. X., C. Li. 2006. Forecasting the term structure of government bond yields. *J. Econometrics* **130** 337–364.

Easton, F. F., J. Goodale. 2005. Schedule recovery: Unplanned absences in service operations. *Decision Sci.* **36** 459–488.

Eckart, C., G. Young. 1936. The approximation of one matrix by another of lower rank. *Psychometrika* **1** 211–218.

Efron, B., R. J. Tibshirani. 1994. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, Toronto, ON, Canada.

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

Guerrero, V. M., J. A. Elizondo. 1997. Forecasting a cumulative variable based on its partially accumulated data. *Management Sci.* **43** 879–889.

Harvey, A. C. 1990. *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press, New York.

Hoerl, A. E., R. W. Kennard. 1970a. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12** 55–67.

Hoerl, A. E., R. W. Kennard. 1970b. Ridge regression: Application to nonorthogonal problems. *Technometrics* **12** 69–82.

Hur, D. 2002. A comparative evaluation of forecast monitoring systems in service organizations. *33rd Annual Meeting Decision Sci. Inst.*, San Diego, CA.

Hur, D., V. A. Mabert, K. M. Bretthauer. 2004. Real-time work schedule adjustment decisions: An investigation and evaluation. *Production Oper. Management* **13** 322–339.

Jolliffe, I. T. 2002. *Principal Component Analysis*, 2nd ed. Springer-Verlag, New York.

Jongbloed, G., G. M. Koole. 2001. Managing uncertainty in call centers using Poisson mixtures. *Appl. Stochastic Models Bus. Indust.* **17** 307–318.

Kekre, S., T. E. Morton, T. L. Smunt. 1990. Forecasting using partially known demands. *Internat. J. Forecasting* **6** 115–125.

Mehrotra, V., O. Ozluk, R. Saltzman. 2006. Intelligent procedures for intra-day updating of call center agent schedules. Technical Report, Department of Decision Sciences, San Francisco State University.

Reinsel, G. C. 2003. *Elements of Multivariate Time Series Analysis*, 2nd ed. Springer-Verlag, New York.

Shen, H., J. Z. Huang. 2005. Analysis of call centre arrival data using singular value decomposition. *Appl. Stochastic Models Bus. Indust.* **21** 251–263.

Steckley, S. G., S. G. Henderson, V. Mehrotra. 2004. Service system planning in the presence of a random arrival rate. Technical report, Cornell University, Ithaca, NY.

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

Thompson, G. M. 1996. Controlling actions in daily workforce schedules. *Cornell Hotel Restaurant Administration Quart.* **37** 82–96.

Wallace, R. B., W. Whitt. 2005. A staffing algorithm for call centers with skill-based routing. *Manufacturing Service Oper. Management* **7** 276–294.

Weinberg, J., L. D. Brown, J. R. Stroud. 2007. Bayesian forecasting of an inhomogeneous Poisson process with applications to call center data. *J. Amer. Statist. Assoc.* Forthcoming.