Bayesian Forecasting of an Inhomogeneous Poisson Process with Applications to Call Center Data

Jonathan Weinberg, Lawrence D. Brown, and Jonathan R. Stroud *

June 26, 2006

Abstract

A call center is a centralized hub where customer and other telephone calls are dealt with by an organization. In today's economy, they have become the primary point of contact between customers and businesses. Accurate prediction of the call arrival rate is therefore indispensable for call center practitioners to staff their call center efficiently and cost effectively. This article proposes a multiplicative model for modeling and forecasting within-day arrival rates to a US commercial bank's call center. Markov chain Monte Carlo sampling methods are used to estimate both latent states and model parameters. One-day-ahead density forecasts for the rates and counts are provided. The calibration of these predictive distributions is evaluated through probability integral transforms. Furthermore, we provide one-day-ahead forecasts comparisons with classical statistical models. Our predictions show significant improvements of up to 25% over these standards. A sequential Monte Carlo algorithm is also proposed for sequential estimation and forecasts of the model parameters and rates.

Keywords: Autoregressive models; Bayesian forecasting; call centers; cubic smoothing spline; inhomogeneous Poisson process; Markov chain Monte Carlo; multiplicative models; sequential Monte Carlo; state space models.

* *Jonathan Weinberg is a Doctoral Student of Statistics, Lawrence D. Brown is the Miers Busch Professor of Statistics and Jonathan R. Stroud is Assistant Professor of Statistics, The Wharton School, University of Pennsylvania, 3730 Walnut Street, 400 Huntsman Hall, Philadelphia, PA 19104-6340, weinber2@wharton.upenn.edu, lbrown@wharton.upenn.edu, stroud@wharton.upenn.edu, 215-563-4347.*

1 Introduction

A call center is a centralized hub that exists solely for the purpose of making or attending calls to or from customers or prospective customers. In today's economy, call centers have not only become the primary point of contact between customers and businesses but also a major investment for many organizations. According to some recent industry estimates, there are approximately 2.86 million operator positions in over 50,000 call centers in the US with some locations employing over 1000 agents. Due to the magnitude of these operations, call center supervisors need to staff their organization efficiently in order to provide a satisfactory level of service at reasonable costs (see Gans *et al.* 2003). Proper management of such a center requires estimation of several operational "primitives", which combined with queuing theory considerations, determine appropriate staffing levels. Accurate prediction of the level of customer arrivals is the most difficult of the primitives to assess.

In the past 15 years, major advances have been made in modeling and predicting arrivals to a telephone call center. Early models describing the arrival process include ARIMA processes and transfer functions (see Bianchi *et al.* 1993 and Andrews and Cunningham 1995). Recent empirical studies have revealed stylized dynamics of the arrival process. Jongbloed and Koole (2001) remark that the process follows an inhomogeneous Poisson process where the arrival rate is stochastic. Furthermore, Avramides *et al.* (2004) remark that the arrival rate varies considerably throughout the day and within their model it appears that call volumes within short periods exhibit strong serial autocorrelation. Finally, Brown *et al.* (2005) observe persistence in the day-to-day dynamics of the arrival rate.

In this paper, we extend on the work of Avramides and co-authors (2004) and Brown *et al.* (2005) and consider an inhomogeneous Poisson model where the arrival rate incorporates both strong within-day patterns and day-to-day dynamics to forecast future arrival rates. The temporal structure of the arrival rate has therefore a two way character, since there is day-to-day variation and also intraday variation. We model these two types of variation separately before combining them in a multiplicative model. Furthermore we provide a fast and efficient Bayesian Markov chain Monte Carlo (MCMC) estimation algorithm.

Recent MCMC methods proposed for inhomogeneous Poisson processes include Shephard and Pitt (1997) and Gamerman (1998) who consider the class of exponential families and generalized linear models, respectively. Soyer and Tarimcilar (2004) provide a Bayesian analysis of an inhomogeneous Cox process to model call center data but focus on the impact of advertising campaigns on arrival rates. A rather different approach is considered here which relies on taking a slightly modified square root of the binned point-process counts and then treating these via variations of multiplicative Gaussian time series models.

The outline of this paper is as follows. In Section 2, we describe the call center analyzed in the

study and the data provided to us. In Section 3, we propose a model for predicting the call arrival rate which is essential in predicting the workload and consequently the staffing of the call center. In addition, we discuss the choice of priors and present the Markov chain Monte Carlo algorithm used to estimate the parameters and latent states in the model. Section 4 gives result from the fitted model to the call center data. A comparative study of competing models based on one day ahead forecasting performance is also presented. A sequential Monte Carlo algorithm, proposed to sequentially estimate the current rate, is presented in Section 4.3. In Section 5, we discuss the advantages of the method developed in this paper, and outline possible extensions to our work.

2 The Data

In this paper we consider data provided by a large North American commercial bank. A detailed summary of each call handled by the call center is provided from March to October 2003. The data was extracted using DATA-MOCCA (see Trofimov *et al.* 2005) and can be found on the authors website at www-stat.wharton.upenn.edu/∼weinber2/data.txt. The path followed by a call through the call center is as follows. A customer places a call to the bank's call center. The customer dials a specific phone number according to the service he desires. Once the call reaches the call center, the customer is greeted by a *voice response unit* (VRU), a computer automated machine which offers various types of information such as the banks opening hours and the user's account information. The user is prompted by the machine to select one of a variety of options. Approximately 80% of customers receive the required information through the VRU and hang up. The remaining 20% of the customers advance to the service queue. There the customer waits until the next available agent becomes available. The service queue does not adhere to the usual *first in, first out* (FIFO) principle as there are premium customers who are prioritized in the queue. If more than one agent is available, the call is routed to the appropriate agent who has been idle for the longest time. While waiting in the queue, a small fraction of customers run out of patience and hang up before actually speaking to an agent.

The bank has various branches of operations such as retail banking, consumer lending and private banking for premium clients, just to mention a few. The call arrival pattern is distinctively different for each type of service. For this reason, we restrict our attention to calls handled by the retail banking division which accounts for approximately 68% of the calls. Although the call center is open 24 hours a day, we concentrate our effort on calls received between 7am and 9:05pm as this is when the call center is most active. Not surprisingly, the pattern of call arrivals is very different for weekdays and for weekends. Furthermore, the weekend opening hours are very different from the weekday hours of operation. For these reasons and to simplify notation, we focus our attention on weekdays in this paper. We do however give results of the analysis that includes weekends in section 4.2.4 The salary of the telephone service operators accounts for almost all the operational costs of a

call center (about 70% of their budget). For this reason, accurate predictions of the arrival rate to the service queue are necessary for supervisors to staff their centers effectively. These predictions coupled with a precise estimate of service times lead to a good forecast of the load using Little's Law (Little 1961).

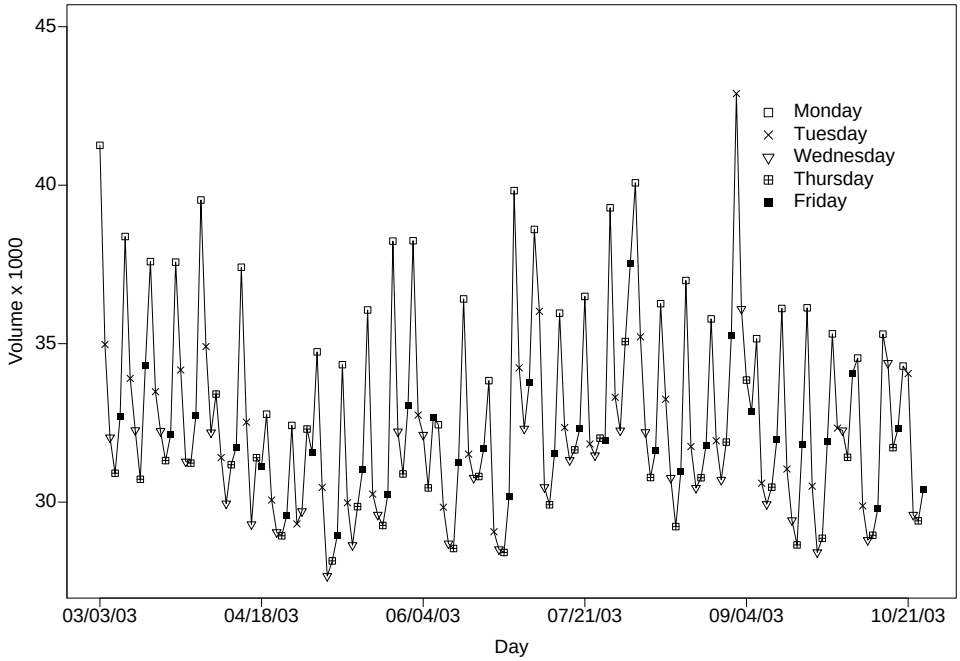

Figure 1: Time series of daily retail banking call volumes handled on non-holiday weekdays between 7am and 9:05pm from March 3 to October 24, 2003.

The data used in the study are summarized in Figures 1 and 2. Figure 1 shows a time series of the daily volumes of calls that reach the call center's service queue from March 3 to October 24, 2003. A number of patterns emerge from this plot: Monday is the busiest day of the week; there is then a gradual decrease in calls from Tuesday to Thursday; the volume increases again on Friday. Furthermore, the plot reveals one significant outlier which lands quite strangely on a Tuesday. After closer inspection, we realize that this corresponds to the day after Labor day. The closure of the call center on that holiday explains the unusually high volume on this particular day. Since this is in effect the first day of the week, we will proceed by modeling this day as a Monday. This is the only such day in our data set hence we don't have enough data to handle this issue in a completely satisfactory manner. Our best guess is that if we did have extensive additional data that it would suggest handling days after Monday holidays as a Monday, as we have decided to do, or (if we had experience from enough such days) that it would suggest making this a new, additional category of day. This is thus an issue about which our data is not sufficient to supply a definitive answer.

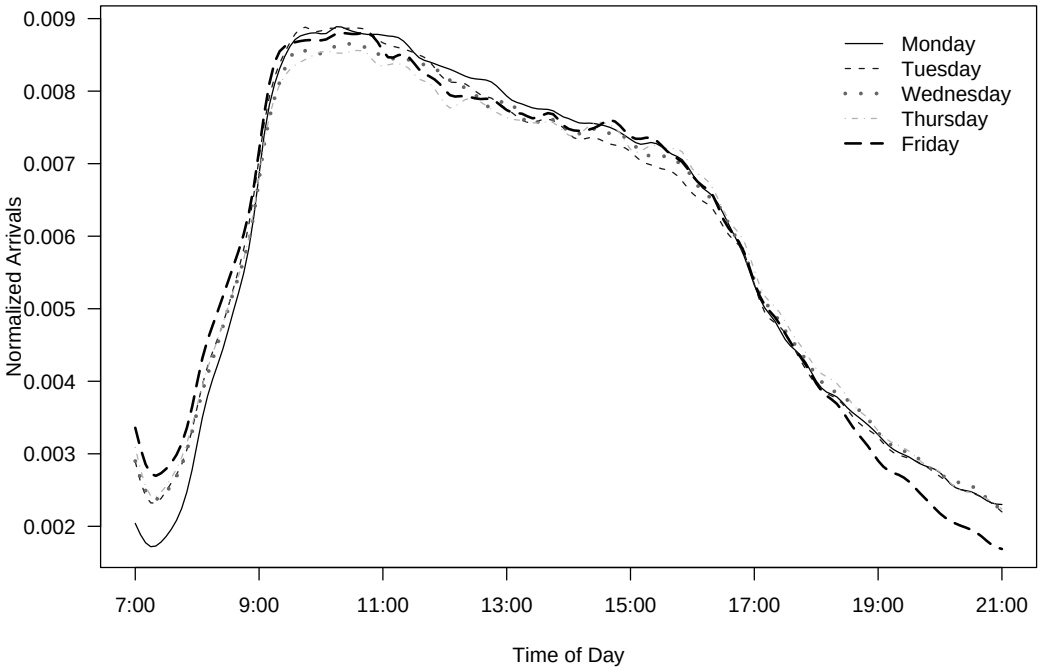

Figure 2: A smoothing spline fit of the Retail banking call arrivals to the service queue normalized by the daily volumes: March 3rd - October 24th 2003

Figure 2 displays the average within-day call arrival patterns for each weekday from March 3 through October 24, 2003. The five curves are obtained by fitting the `smooth.spline` function in S-plus to the volumes of calls that reach the service queue every 5 minutes, normalized by the daily volumes. From this plot, we can infer three distinct within-day features. First, the call arrivals increase sharply between 7am and 10am, at which time the call center is at its busiest. The volume then decreases linearly as the day progresses. At around 5pm, as the work day finishes, we notice a sudden downward slope in the number of calls. Although the within-day patterns share these common trends, there are sufficient differences across weekdays that we model the within-day pattern for each day separately. In particular, we observe that Monday morning starts off slower than the four other weekdays. In addition, there is an unusually quick dropoff in the number of calls on Friday afternoon.

3 The Model

Classical queuing theory assumes that the arrivals to the service queue follow a Poisson process with constant rate. Figure 2 clearly illustrates that the standard theory cannot be applied in this context as there are distinct differences in call arrival rates within a day. To overcome this modeling issue, call center practitioners assume a constant rate for short time intervals and consequently apply the

standard queueing methodology on these shorter time intervals. Jongbloed and Koole (2001) argue that the rate is not constant over these short time frames and proceed by modeling the rate of arrivals in these periods using a Poisson mixture to account for the overdispersion in the data. In their recent work, Brown *et al.* (2005) remark that the arrival pattern follows a time inhomogeneous Poisson process where the rate evolves smoothly through the day. Furthermore, they state that the Poisson arrival rates, which we will refer to as $\lambda(t)$, should be modeled as a stochastic process due to the overdispersed data, rather than a deterministic function of time of day and day of week. Extending on this work, we construct a model to estimate and predict future $\lambda(t)$'s.

Let N_{jk} denote the number of arrivals to the queue on day $j = 1, \ldots, J$ during the time interval $[t_{k-1}, t_k]$ where $k = 1, \ldots, K$ is the kth period in the day. In our study, we have $J = 164$ days and $K = 169$ time periods since we are using five minute intervals and $t_k = k/169$. In addition, let the weekday corresponding to day j be denoted by d_j (for example, $d_j = 1$ signifies that day j is a Monday).

We first consider the following model

$$N_{jk} \sim Poiss(\lambda_{jk}), \quad \lambda_{jk} = R_{d_j}(t_k)\ v_j + \epsilon_{jk} \tag{1a}$$

where λ_{jk} is the arrival rate for day j and period k, $R_{d_j}(t_k)$ is the proportion of daily calls on day j that are handled during the time interval $[t_{k-1}, t_k)$, v_j is a proxy for the daily volume on day j, and ϵ_{jk} is a random error. The assumption here is that every weekday has a different within-day pattern. Furthermore R_i is a density and therefore

$$\sum_{k=1}^{K} R_{d_j}(t_k) = 1 \quad \text{for} \quad d_j = 1, \ldots, 5. \tag{1b}$$

Next, we follow the work of Brown *et al.* (2005), and use the variance stabilizing transformation for Poisson data, which is based on the following result (Brown *et al.*, 2001): If N is Poiss(λ), then $Y = \sqrt{N + \frac{1}{4}}$ has approximately a mean $\sqrt{\lambda}$ and variance $\frac{1}{4}$. In addition, as $\lambda \to \infty$, Y is approximately normal. This approximation is very accurate for the dataset used in our analysis since the arrival counts are fairly large (~ 300).

Using the approximation discussed above and defining $y_{jk} = \sqrt{N_{jk} + \frac{1}{4}}$, we obtain the model

$$y_{jk} = g_{d_j}(t_k) x_j + \epsilon_{jk}, \quad \epsilon_{jk} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2). \tag{2}$$

The connection between (1a) and (2) is established by realizing that $g_{d_j}(t_k) = \sqrt{R_{d_j}(t_k)}$ and $x_j = \sqrt{v_j}$. We take into account the correlation between daily volumes by assuming an autoregressive structure for the daily random effect x_j, adjusting for the type of day. The model for this can be

written as

$$x_j - \alpha_{d_j} = \beta(x_{j-1} - \alpha_{d_{j-1}}) + \eta_j, \quad \eta_j \overset{iid}{\sim} \mathcal{N}(0, \psi^2) \tag{3}$$

where α_{d_j} denote the intercept for day d_j. It should be noted that other models that incorporated linear and quadratic trend or higher order autocorrelation were considered but did not substantially improve the model fit.

Following remarks made on Figure 2, a different daily pattern of call arrivals for each day of the week is assumed and is denoted by g_{d_j}. In addition, we incorporate smoothness in the within-day pattern through the following model

$$\frac{d^2 g_{d_j}(t_k)}{dt_k^2} = \tau_{d_j} \frac{dW_{d_j}(t_k)}{dt_k} \tag{4a}$$

where

$$\sum_{k=1}^{K} g_{d_j}(t_k)^2 = 1, \quad \text{for} \quad d_j = 1, \ldots, 5. \tag{4b}$$

Here $W_{d_j}(t)$ are independent Wiener processes with $W_{d_j}(0) = 0$ and var $\{W_{d_j}(t)\} = t$. Since g refers to the within-day pattern, the restriction (4b) is equivalent to the constraint (1b) on the R's. This assumption is also needed for identifiability purposes. The unconstrained prior on the g's, given by (4a), has close connections to cubic smoothing splines. For more information on the equivalence between this prior distribution and the cubic smoothing spline, see Wahba (1983).

Following the work of Kohn and Ansley (1987), we can define $z_{d_j}(t_k) = \{g_{d_j}(t_k), dg_{d_j}(t_k)/dt_k\}$, and rewrite the prior on the g's as a vector autoregressive process on the z_{d_j}'s. This leads to the following model

$$\begin{aligned}
y_{jk} &= h' \; z_{d_j}(t_k) x_j + \epsilon_{jk}, & \epsilon_{jk} &\sim \mathcal{N}(0, \sigma^2) && (5)\\
x_j - \alpha_{d_j} &= \beta(x_{j-1} - \alpha_{d_{j-1}}) + \eta_j, & \eta_j &\sim \mathcal{N}(0, \psi^2) && (6)\\
z_{d_j}(t_k) &= F(\delta) \; z_{d_j}(t_{k-1}) + u_k, & u_k &\sim \mathcal{N}(0, \tau_{d_j}^2 U(\delta)) && (7)
\end{aligned}$$

where ϵ_{jk}, η_j and u_k are mutually independent, $\delta = t_k - t_{k-1}$ and $h' = [1, 0]$. The matrices F(δ) and U(δ) are defined as

$$F(\delta) = \begin{pmatrix} 1 & \delta \\ 0 & 1 \end{pmatrix}, \quad U(\delta) = \begin{pmatrix} \delta^3/3 & \delta^2/2 \\ \delta^2/2 & \delta \end{pmatrix}.$$

We assume diffuse distributions for the initial states x_1 and $z_{d_j}(t_1)$ for $d_j = 1, \ldots, 5$. Equations (5)-(7) correspond to a multiplicative model with two latent states which evolve on different time

scales. The formulation above also facilitates computation, as conditional on each latent state variable, the model can be cast into linear state space form. The efficient forward-filtering backward-sampling (FFBS) algorithm proposed by Carter and Kohn (1994) and Frühwirth-Schnatter (1994) can consequently be implemented to carry out Bayesian inference on the model.

It should be noted that the Bayesian methodology developed by Shephard and Pitt (1997) and Gamerman (1998) for dynamic generalized linear models could have been applied to the original Poisson model. However, we believe that the root-unroot methodology has several distinct advantages here. First, the normal approximation is very accurate in this study. Furthermore, a conjugate multivariate normal prior with a wide variety of covariance structures such as a moving average or autoregressive process can be imposed on x and g. An equivalent conjugate prior on the Poisson rates is much harder to derive and is the subject of current research. The methodology developed by Chen and Fomby (1999) for stable seasonal pattern models could also have been used here. Incorporating the within-day dependencies through smoothness would however be quite complicated with their methods.

To complete the Bayesian specification of the model, we need to specify the prior distributions on the parameters.

3.1 Prior Selection

For notational convenience, let $\alpha = (\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5)$, $\overline{\alpha} = \sum_{i=1}^{4} \alpha_i$, $\tau^2 = (\tau_1^2, \tau_2^2, \tau_3^2, \tau_4^2, \tau_5^2)$ and $\theta = (\alpha, \beta, \psi^2, \tau^2, \sigma^2)$. After some careful sensitivity analysis, the following priors were selected for the parameters of the model described above. Independent priors are imposed on all parameters except on the individual α_i. Hence, $p(\theta)$ can be written as the product of the following priors

$$\begin{aligned}
p(\alpha) &\propto \left(\sum_{i=1}^{4} (\alpha_i - \overline{\alpha})^2 \right)^{-1} \\
p(\beta) &= \mathcal{U}[0,1] \\
p(\tau_i^2) &= \mathcal{IG}(a_{\tau_i}, b_{\tau_i}) \quad i = 1, \ldots, 5 \\
p(\sigma^2) &= \mathcal{IG}(a_\sigma, b_\sigma) \\
p(\psi^2) &= \mathcal{IG}(a_\psi, b_\psi).
\end{aligned}$$

Several remarks should be made on the choice of priors. The traditional multivariate normal prior is not used on α. Instead, a flat prior on the overall mean $\overline{\alpha}$ (Lindley and Smith, 1972) coupled with the harmonic prior on the differences $\alpha_i - \overline{\alpha}$ is our preferred choice due to their good minimax and shrinkage properties (Stein, 1981). The specific construction of the prior leads to shrinkage towards the overall mean rather than the origin. For modeling purposes, we also assume that $x_j - \alpha_{d_j}$ follows a stationary process with positive correlation and therefore impose a uniform

$\mathcal{U}(0,1)$ prior on the autoregressive coefficient. Finally, conjugate inverse gamma priors are used on the variance components. The hyperparameters are specific to the data analyzed and therefore will be presented in the case study section.

3.2 Posterior Inference

The goal of our analysis is to carry out exact inference on the joint posterior distribution

$$p(x, z, \theta | Y_{1:J}) \quad \propto \quad p(Y_{1:J}|x, z, \theta)\ p(x|\theta)\ \prod_{i=1}^{5} p(z_i|\theta)\ p(\theta)$$

where $Y_{1:J} = (y_{11}, \ldots, y_{1K}, \ldots, y_{J1}, \ldots, y_{JK})$, $x = (x_1, \ldots, x_J)$, $z = (z_1, z_2, z_3, z_4, z_5)$ and $z_{d_j} = (z_i(t_1), \ldots, z_{d_j}(t_K))$ for $d_j = 1, \ldots, 5$. We sample the parameters and latent states using a hybrid MCMC algorithm developed in the Appendix that utilizes both Gibbs sampling steps and random walk Metropolis steps (see Robert and Casella 2004 and Metropolis *et al.* 1953). The algorithm recursively cycles between sampling from $p(x|z, \theta, Y_{1:J})$, $p(z|x, \theta, Y_{1:J})$ and $p(\theta|x, z, Y_{1:J})$. We benefit from the forward-filtering backward-sampling algorithm to draw directly the whole path of x's from $p(x|z, \theta, Y_{1:J})$. The same algorithm is also used to draw the whole path of z_{d_j}'s from $p(z_{d_j}|x, \theta, Y_{1:J})$ for each day of the week. This algorithm enables fast mixing of the Markov chain.

We should point out one slight complication in the model. The Gaussian prior selected for the g_{d_j} $(d_j = 1, \ldots, 5)$ corresponds to a cubic smoothing spline constrained to lie on the sphere $\sum_k g_{d_j}^2(t_k) = 1$. This quadratic constraint complicates posterior simulation since the posterior is also a constrained Gaussian. A Metropolis algorithm would be infeasible to simulate from this distribution as we would either have to evaluate the prior on the sphere which would involve a 169-dimensional integral, or use the uninformative constrained prior as the proposal distribution which would lead to rejecting all proposed moves. To avoid this problem, we sample from the unconstrained posterior using FFBS and then renormalize the draws. In our application, the unconstrained samples of $\sum_k g_{d_j}^2(t_k)$ fall between 0.99 and 1.01, indicating that the normalized draws are approximately from the true conditional posterior distribution. If the constraint set were affine, the argument would be exact because in this case conditioning is the same as projecting onto a line. Since the posterior is highly concentrated around a point on the sphere, and the sphere is locally linear, the approximation is very accurate.

4 Case Study

Our case study consists of three steps. First, the MCMC algorithm is applied to the whole data set to estimate latent states and parameters and check model adequacy. A careful assessment of convergence is also conducted. We then perform an out-of-sample forecasting exercise using

the same model. We compare our forecasting performance to seasonal regression models. The competing models and the results will be discussed in Section 4.2. Finally, a sequential Monte Carlo algorithm for within day learning and forecasting is proposed. Consequently, we measure whether incorporating morning data greatly improves the afternoon and evening forecasts.

4.1 Posterior Estimation

As stated above, the dataset used in this study consists of 164 days. Within each day, the number of retail banking call arrivals per 5 minute interval is recorded between 7am and 9:05pm, leaving us with 169 time periods. The algorithm ran according to the following specifications. The chain is run for 49,000 iterations after a burn-in period of 1000. The analysis of the autocorrelation function of each parameter and latent state led us to save every 10th iteration leaving 4899 (approximately) independent samples for posterior inference. In what follows, we will refer to the MCMC sample size as M.

The prior on θ described in Section 3.1 is used with the following hyperparameters: $a_\sigma = 0.05$, $b_\sigma = 0.05$, $a_\psi = 0.05$, $b_\psi = 0.05$, $a_{\tau_i} = 0.05$ and $b_{\tau_i} = 0.05$ for $i = 1, \ldots, 5$, which corresponds to very diffuse priors. Furthermore x_1 and $z_i(t_1)$ are drawn from the following diffuse priors, $x_1 \sim \mathcal{N}(0, 10^5)$ and $z_i(t_1) \sim \mathcal{N}(0, 10^5 I)$. An interesting feature of the model arises when performing a prior sensitivity analysis . After some careful analysis on simulated data, we observe that for small values of β, the variance components σ^2 and ψ^2 become non-identifiable and more peaked priors are needed. However, due to the high persistence of the autoregressive process in our application, we are able to place diffuse priors on the two variances. The marginal posterior distributions of x and z are not sensitive to the choice of hyperparameters due to the vast quantities of data available in the analysis.

We ran the algorithm from multiple starting values concluding that the choice of initial values didn't affect the convergence of the Markov chain. The results described below are based on the following initial parameter values: $\alpha^{(0)} = (190, 180, 175, 175, 180)$, $\beta^{(0)} = 0.65$, $(\sigma^2)^{(0)} = 0.25$, $(\psi^2)^{(0)} = 19.85$ and $(\tau^2)^{(0)} = (25, 25, 25, 25, 25)$.

Careful analysis of the traceplots and autocorrelation function (ACF) of each parameter was performed to check the convergence of the Markov chain to its stationary distribution. Figures 3 and 4 display the marginal posterior distributions $p(\theta|Y_{1:J})$. Several conclusions can be made from these plots. First, we should point out that the posterior mean of σ^2 is 0.347. This is a very encouraging result. If the model accounted for all the variance in the data, σ^2 would be approximately 0.25 but no smaller, according to the root-unroot theory. We initially considered a model with the same within-day pattern g for all days. The estimated σ^2 was 0.432 which is substantially larger than the model considered here. The autoregressive coefficient β is highly significant with a posterior mean of 0.68. The analysis also confirms that there is an obvious difference in daily random effects

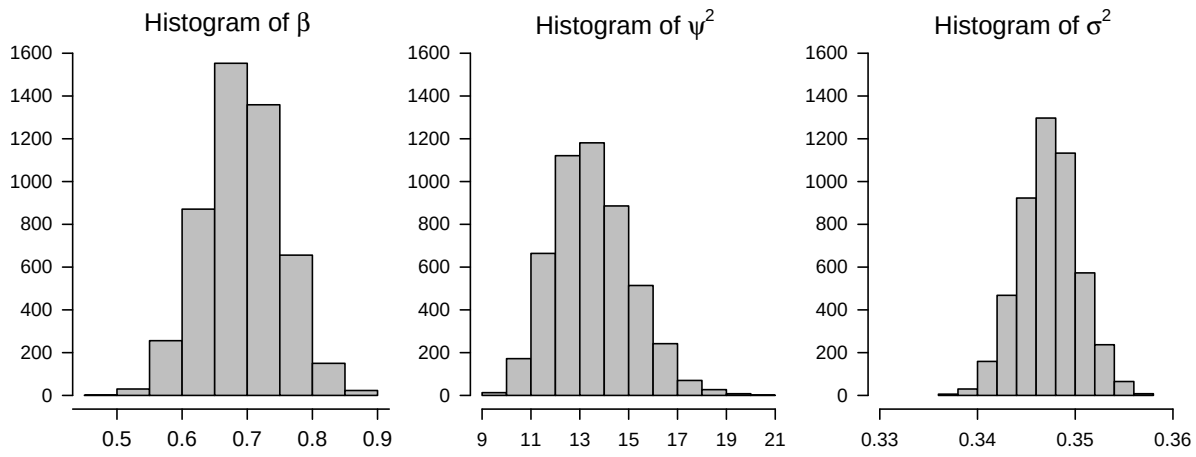

Figure 3: Posterior histograms of β, ψ^2 and σ^2.

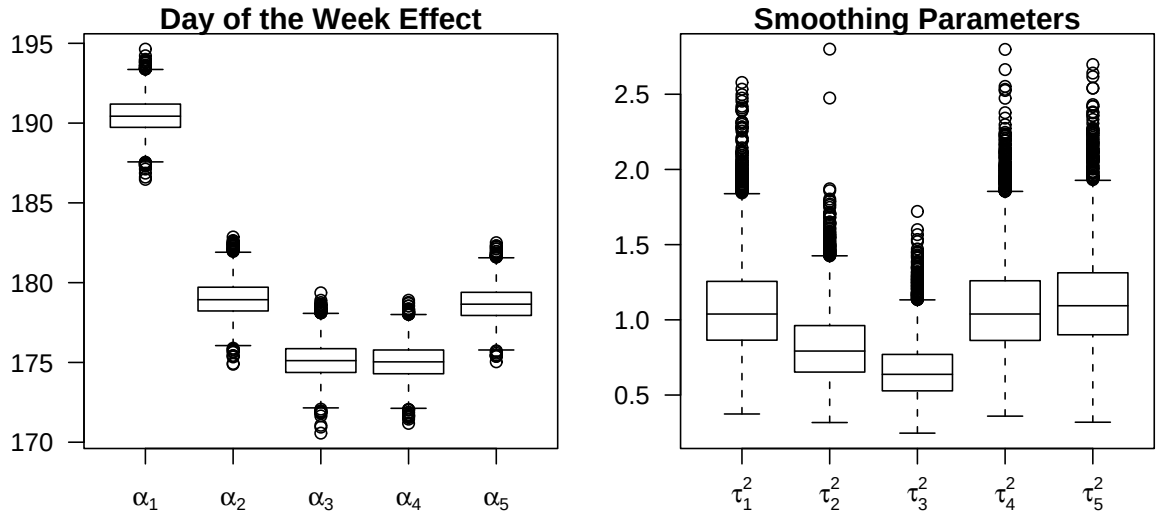

Figure 4: Boxplots displaying the posterior distributions of the weekday effects and the smoothing parameter for the cubic splines for each day of the week.

across weekdays, ranging from a posterior mean of 175 for Thursdays to 190 for Mondays. The posterior mean of τ^2 ranges from 0.66 for Wednesday to 1.08 for Friday. This indicates that the within-day patterns are very smooth as the cubic spline penalty factors for each of the five days $(i = 1, \ldots, 5)$ defined as σ^2 $(\tau_i^2 \sum_{\{j\,:\,d_j=i\}} x_j^2)^{-1}$, are very small.

4.2 One-Day-Ahead Forecasting

The previous section concentrated on the in-sample performance of the model and estimation procedure. This exercise is indispensable to check model adequacy but it is not of primary interest to call center managers who need good forecasts of call arrival rates in order to plan ahead and accurately staff their center. A one-day- ahead prediction exercise was therefore conducted to compare the out-of-sample performance of the model with industry standards. Ideally, call center managers also require forecasts at longer time horizons. Unfortunately, based on our estimated AR(1) coefficient, we are unable to accurately predict call volumes at horizons greater than a week. Additional covariates, not provided in this dataset, would therefore be needed to enhance our predictions.

4.2.1 Forecast Densities for Rates and Volumes

To simplify notation, let $\lambda_{j.} = (\lambda_{j1}, \ldots, \lambda_{jK})$ and $N_{j.} = (N_{j1}, \ldots, N_{jK})$ denote the rates and counts on day j, and define $\Omega = (x_{j-1}, z, \theta)$. The one-day-ahead predictive density for the rates is given by

$$p(\lambda_{j.}|Y_{1:j-1}) \;=\; \iint p(\lambda_{j.}|x_j, z_{d_j})\; p(x_j|x_{j-1}, \theta)\; p(x_{j-1}, z_{d_j}, \theta|Y_{1:j-1})\; dx_{j-1}\; d\theta.$$

Note that the third term in the integrand is the joint posterior distribution of the model parameters and states on day $j-1$, the second term is the AR(1) Gaussian transition density while the first term is a degenerate distribution since λ is a deterministic function of x and z. The one-day ahead predictive density for the counts on day j is given by

$$p(N_{j.}|Y_{1:j-1}) \;=\; \iint p(N_{j.}|\lambda_{j.}, \theta)\; p(\lambda_{j.}, \theta|Y_{1:j-1})\; d\lambda_{j.}\; d\theta.$$

These predictive densities are not available in closed form as the rates and counts are nonlinear functions of x_j and Ω. We therefore approximate these distributions by a large sample drawn using the algorithm outlined below.

Algorithm 1: One-Day-Ahead Forecasts of Rates and Counts

Step 0: Start with an MCMC sample, $\Omega^{(1)}, \ldots, \Omega^{(M)}$, drawn from $p(\Omega|Y_{1:j-1})$.

Step 1: Draw $x_j^{(i)} \sim \mathcal{N}\left(\alpha_{d_j}^{(i)} + \beta^{(i)}(x_{j-1}^{(i)} - \alpha_{d_{j-1}}^{(i)}), (\psi^2)^{(i)}\right)$ for each $i = 1, \ldots, M$.

Step 2: For each period $k = 1, \ldots, K$ and each $i = 1, \ldots, M$.

Step 2a: Set $\lambda_{jk}^{(i)} = \left(x_j^{(i)}\ g_{d_j}(t_k)^{(i)}\right)^2$.

Step 2b: Draw $y_{jk}^{(i)} \sim \mathcal{N}\left(\sqrt{\lambda_{jk}^{(i)}},\ (\sigma^2)^{(i)}\right)$.

Step 2c: Set $N_{jk}^{(i)} = \left(y_{jk}^{(i)}\right)^2 - 0.25$.

Steps 2a and 2c of the algorithm described above provide samples from $p(\lambda_{j.}|Y_{1:j-1})$ and $p(N_{j.}|Y_{1:j-1})$ respectively. Since the arrival rate is an unobservable quantity, we focus our attention on call volumes in order to compare the out-of-sample performance of various models.

4.2.2 Competing Forecasts

In what follows, we will refer to the approach described above as Model 1. For comparative purposes, we will consider two simple alternatives which we refer to as Model 2 and Model 3. The first model is a linear additive model on the transformed data with a day of the week and time of day as covariates. In the second model, we also add an interaction between the day of the week and the time of day effects. The description of the two seasonal linear models follows

$$\begin{aligned} \textbf{Model 2:} \quad y_{jk} &= \mu + \alpha_{d_j} + \beta_k + \epsilon_{jk} \quad \epsilon_{jk} \sim \mathcal{N}(0, \sigma^2) \\ \textbf{Model 3:} \quad y_{jk} &= \mu + \alpha_{d_j} + \beta_k + \gamma_{d_j k} + \epsilon_{jk} \quad \epsilon_{jk} \sim \mathcal{N}(0, \sigma^2) \end{aligned}$$

where $y_{jk} = \sqrt{N_{jk} + \frac{1}{4}}$. We fit the models on historical data using least squares. The normality assumption enables us to also obtain prediction intervals for future observations.

The model proposed by Brown *et al.* (2005) was originally considered as a competing model. Unfortunately, the model in their analysis does not incorporate all the dynamics of the data analyzed here such as day of the week effects and different within-day patterns for each weekday. A reformulation of their model adjusting for all these new features would result in a representation similar to the one presented in this paper and therefore would not be subject to fair comparison.

4.2.3 One-day-ahead forecast analysis

The one-day-ahead forecasting exercise is performed in the following manner. For the 64 days ranging from July 25 to October 24, 2003,

- Consider the 100 preceding days as the historical dataset.
- Estimate the parameters and latent states for the model described in the paper using the same recipe described in Section 4.1. For the competing models, compute the the maximum likelihood estimator for each parameter.
- For all three models, perform a one-day-ahead forecast of call volumes for period k.
- Compute the forecast root mean square error (RMSE) and average percent error (APE), defined for each day j as follows

$$\text{RMSE}_j = \sqrt{\frac{1}{K}\sum_{k=1}^{K}(N_{jk} - \widehat{N_{jk}})^2}, \qquad \text{APE}_j = \frac{100}{K}\sum_{k=1}^{K}\frac{|N_{jk} - \widehat{N_{jk}}|}{N_{jk}}.$$

- Compute the 95% coverage probability (COVER) and the average 95% forecast interval width (WIDTH), defined for each day j as follows

$$\text{COVER}_j = \frac{1}{K}\sum_{k=1}^{K} I(\widehat{N}_{jk}^{2.5} < N_{jk} < \widehat{N}_{jk}^{97.5}), \qquad \text{WIDTH}_j = \frac{1}{K}\sum_{k=1}^{K}(\widehat{N}_{jk}^{97.5} - \widehat{N}_{jk}^{2.5})$$

Here $I(\cdot)$ is the indicator function. $\widehat{N_{jk}}$ and $\widehat{N}_{jk}^{Q}$ are the mean and Qth quantile of the forecast distribution. For Model 1, these quantities are computed using the Monte Carlo sample, for Model 2 and 3, they are based on the maximum likelihood estimates and the assumption of normality.

	RMSE			**APE**		
	M1	M2	M3	M1	M2	M3
Min	11.14	12.94	12.39	5.6	5.8	5.9
25th	14.25	16.22	15.53	7.0	7.8	7.2
50th	15.83	19.12	17.96	7.4	9.3	7.9
Mean	18.28	21.32	20.46	8.4	10.1	9.1
75th	19.83	21.82	22.10	8.5	11.6	9.7
Max	43.42	51.37	45.69	28.6	36.3	29.9

Table 1: Summary of one-day ahead forecasting performance of the three competing models: Model 1 (M1), Model 2 (M2), Model 3 (M3). Summaries are based on data between 7am and 9:05pm.

Table 1 summarizes the distribution of the RMSE and APE for the all three models over the 64-day forecast period. Several remarks can be made based on these results. The model developed in this paper clearly outperforms the two competing models: The median forecast RMSE is 20.8%

and 13.5% lower than those of Model 2 and 3 respectively. We can conclude that the autoregressive feature of the model presented in this paper greatly improves the accuracy of our forecasts.

	Coverage Probability			**Average Width**		
	M1	M2	M3	M1	M2	M3
Min	0.686	0.598	0.609	64.54	76.79	72.38
25th	0.935	0.920	0.938	68.13	79.71	74.98
50th	0.970	0.967	0.976	69.23	80.71	76.01
Mean	0.947	0.937	0.941	70.10	81.48	76.69
75th	0.988	0.990	0.988	72.41	82.40	77.13
Max	1.000	1.000	1.000	79.30	88.10	82.37

Table 2: Summary of 95% one-day-ahead forecast intervals for all three competing models. Summaries are based on calls handled between 7am and 9:05pm.

Table 2 summarizes the distribution of the 64 coverage probabilities and average interval widths for all three models. We note that the empirical coverage of the 95% prediction intervals are extremely accurate with mean coverages of 94.7%, 93.7% and 94.1% respectively. We should note that the prediction intervals for Model 1 are on average 16.2% and 9.4% narrower than Model 2 and 3. While still obtaining coverage probabilities close to the nominal value, Model 1 produces more precise prediction intervals. The rather wide prediction intervals are mainly due to the inherent Poisson variation which is proportional to the arrival rate.

4.2.4 One-day-ahead forecast analysis with weekends

In this section, we discuss the results of the one-step-ahead forecasting exercise when we include both weekdays and weekends. The number of days J in our case study increases to 231 days ranging from July 25 to October 24, 2003. In this organization, the weekend opening hours are shorter than the weekday operating hours (9am to 5pm) leaving us with only 108 periods within a day for both Saturday and Sunday and 169 periods for each weekday. We perform the same analysis as discussed in the previous section using a 100 day moving window. The main difference, when including the weekends, is that the day-to-day autocorrelation β is smaller with a posterior mean of 0.62. Table 3 summarizes the distribution of the RMSE and APE for the all three models over the 131-day forecast period. Again, we see that Model 1 outperforms the two other models with a median forecast RMSE which is approximately 40% and 8% lower than those of Model 2 and 3 respectively.

	RMSE			APE		
	M1	M2	M3	M1	M2	M3
Min	6.86	11.28	7.31	5.7	6.6	5.7
25th	13.35	17.36	14.06	7.1	9.8	7.6
50th	15.23	21.45	16.47	8.0	12.0	9.2
Mean	16.31	23.02	18.14	9.1	13.4	9.8
75th	18.12	27.11	20.97	9.9	16.2	11.1
Max	40.44	49.00	44.73	28.6	37.9	30.3

Table 3: Summary of one-day ahead forecasting performance of the three competing methods: Model 1 (M1), Model 2 (M2), Model 3 (M3). Summaries are based calls handled between 7am and 9:05pm for Monday through Friday and between 9am and 5pm for Saturday and Sunday

The empirical coverage of the 95% prediction intervals, not shown here, are also very accurate with mean coverages within 1% of the nominal value for all three models. We should also note that the prediction intervals for Model 1 are on average approximately 53% and 5% narrower than Model 2 and 3. These results confirm that our model is robust when including further within-day patterns in our analysis.

4.2.5 Forecast Calibration

One advantage of the Bayesian framework is that it provides the entire forecast density for the rates and counts, fully accounting for uncertainty in the latent states and model parameters. These densities can be used to provide an alternative measure of forecast performance, the probability integral transform (PIT) (Rosenblatt 1952), which is defined for each day j and period k by

$$\text{PIT}_{jk} \quad = \quad \frac{1}{M}\sum_{i=1}^{M} I\left(N_{jk} < N_{jk}^{(i)}\right)$$

where N_{jk} is the actual observation and $N_{jk}^{(i)}$ are Monte Carlo samples from the forecast density. This measure has been used extensively in econometrics (see Shephard 1994 and Diebold *et al.* 1998) and in weather forecasting (see Gel *et al.* 2004). Assuming the predictive densities of the counts are properly calibrated, the marginal distribution of the PIT should be uniform $\mathcal{U}[0,1]$ across all j and k. However, we expect the PIT to be correlated across periods k due to the within-day dependence on the forecast estimate of x_j. Due to this, the within-day distribution of the PIT will tend to show higher variability than an iid sample. On the other hand, when combining all 64 days of data, we would expect the distribution to be more uniform.

Figure 5 summarizes the forecast performance of our model for the week of August 18, 2003. The first column displays the 95% predictive intervals for the rates and counts for each day. Columns

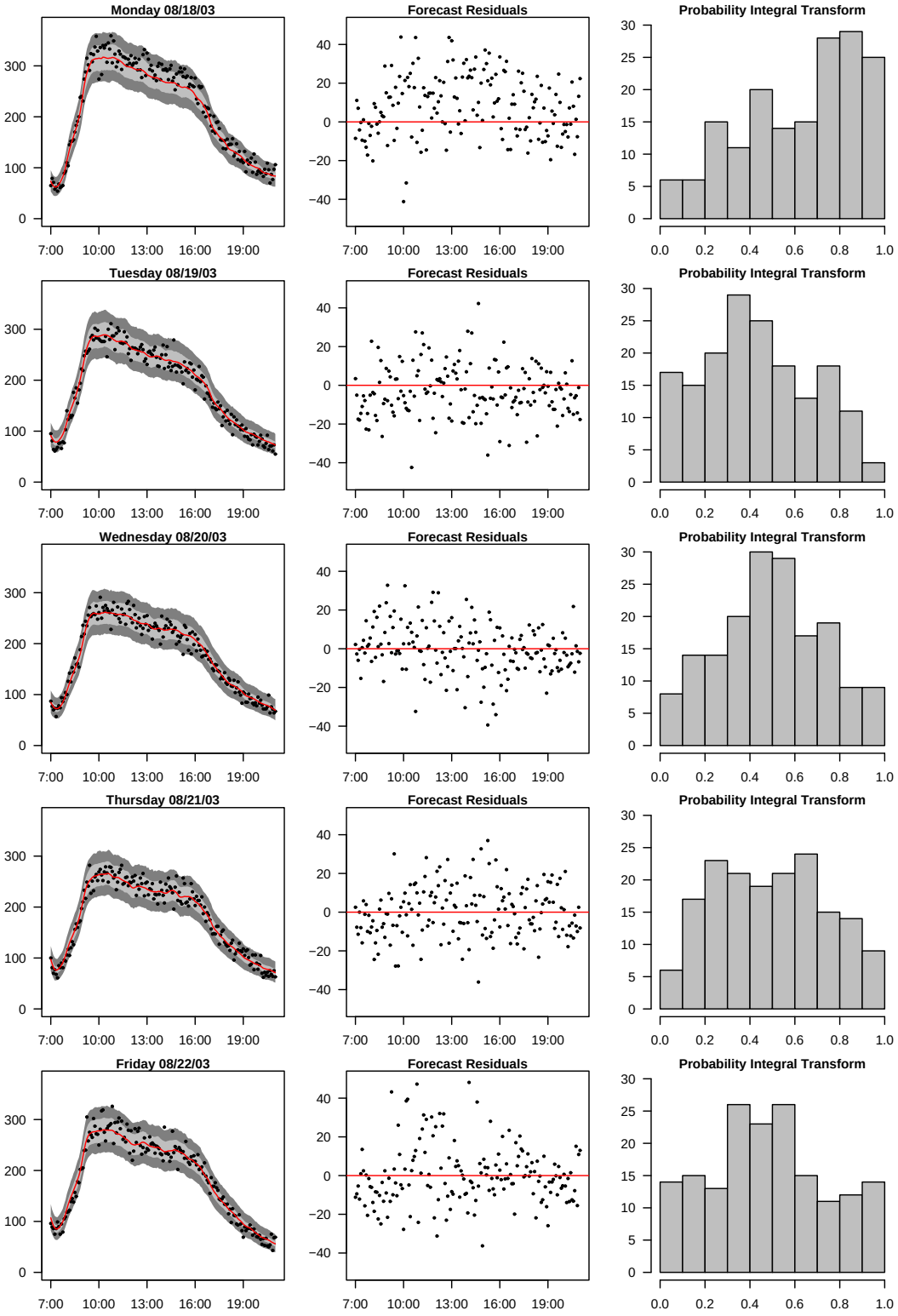

Figure 5: Forecast performance for the week of August 8, 2003. *Left:* One-day-ahead forecast means and 95% intervals for the rates and counts. Points denote the observed counts. *Center:* Forecast residuals (observed counts minus forecast mean). *Right:* Probability integral transform for the observed counts based on the Monte Carlo samples.

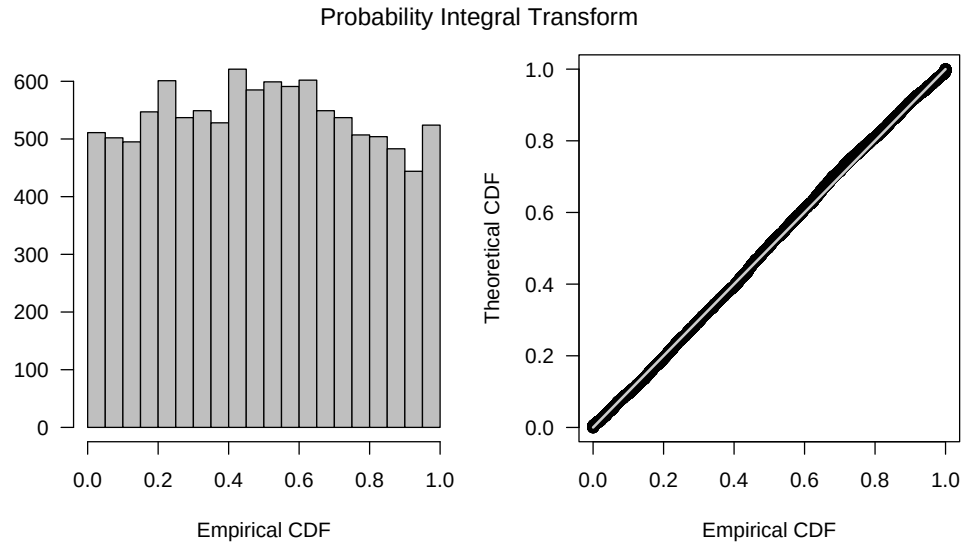

Figure 6: Distribution and QQ-plot of the probability integral transform for the observed counts based on the Monte Carlo samples combining the 64 predicted days.

2 and 3 display the forecast residuals (defined as the observed counts minus the forecast mean), and the histogram of the PIT for each day. The results are very encouraging although the coverage probabilities for that week are slightly high. For the most part, the forecast residual plots show no obvious bias. The histogram of the PIT confirms that our forecasts are well calibrated since the histograms are fairly symmetric and close to uniform. The distribution of the combined PIT over all 64 days is shown in Figure 6. The histogram and the QQ-plot for the PIT exhibit a near-perfect uniform distribution, confirming that the forecasts are very well calibrated.

4.3 Within-Day Learning and Forecasting

The previous section focused on predictions of call volume based on information available at the end of the previous day. However, as new data arrive throughout the day, a manager may want to reevaluate his or her forecast for the remainder of the day. For example, in the bank that provided the data, telephone agents can be called up with as little as three hours notice. Prediction of the afternoon volume based on the morning information would therefore be of use to this organization. Re-estimating the latent states and parameters every few minutes using the algorithm proposed in section 3.2 would be infeasible, as the full MCMC simulation took us approximately 30 minutes to run, even when efficiently implemented in C. Therefore, a fast algorithm is needed to recursively update the posterior density $p(x_j, z, \theta | Y_{1:j-1}, y_{j1}, \ldots, y_{jk})$ for each time period k.

At the end of day $j-1$, we have available an MCMC sample from the historical posterior distribution $p(x_{j-1}, z, \theta | Y_{1:j-1})$. One possible way to proceed is to simulate x_j from the transition density $p(x_j | x_{j-1}, \theta)$ and reweight the draws $\{x_j, z, \theta\}$ according to the likelihood $p(y_{jk} | x_j, z, \theta)$

as new data arrive. One problem with this procedure is that, in the presence of outliers, a small number of draws would receive most of the weight, leading to degeneracy of the algorithm. In what follows, we propose a slightly different approach which considerably alleviates this problem. We notice that we can analytically integrate out x_j and use the reweighting scheme described above on $\{z, \theta\}$ alone. Then, conditioning on $\{z, \theta\}$, we can draw x_j directly from its conditional posterior distribution.

In what follows, we will assume dependence on the historical data $Y_{1:(j-1)}$ but suppress it from the notation. Let us also define $Y_{jk} = (y_{j1}, \ldots, y_{jk})$ as the data on day j up to period k. Recalling that $\Omega = \{x_{j-1}, z, \theta\}$, all relevant inference can be obtained from the joint posterior distribution $p(\Omega, x_j|Y_{jk})$, which can be decomposed as follows

$$p(\Omega, x_j|Y_{jk}) \quad = \quad p(\Omega|Y_{jk})\ p(x_j|\Omega, Y_{jk}). \tag{8}$$

The marginal posterior density for Ω is approximated by the discrete distribution

$$p(\Omega|Y_{jk}) \quad \approx \quad \sum_{i=1}^{M} I\left(\Omega = \Omega^{(i)}\right)\ w_k^{(i)}, \tag{9}$$

where $\Omega^{(1)}, \ldots, \Omega^{(M)}$ are the samples from the historical posterior $p(\Omega|Y_{1:(j-1)})$ and $w_k^{(1)}, \ldots, w_k^{(M)}$ are the normalized weights. The conditional posterior for x_j given Ω is normal, given by

$$p\left(x_j|\Omega, Y_{jk}\right) \quad = \quad \mathcal{N}(m_k, v_k). \tag{10}$$

The recursive formulas for the weights, means and variances $\{w_k, m_k, v_k\}$ are defined below.
The normalized weights in Equation (9) are initialized to $w_0^{(i)} = 1/M$, and updated according to

$$w_k^{(i)} \ \propto \ w_{k-1}^{(i)}\ p(y_{jk}|\Omega^{(i)}, Y_{j(k-1)}) \quad \text{where} \quad \sum_{i=1}^{M} w_k^{(i)} \ = \ 1.$$

This follows from the fact that $p(\Omega|Y_{jk}) \ \propto \ p(\Omega|Y_{j(k-1)})\ p(y_{jk}|\Omega, Y_{j(k-1)})$ where

$$p(y_{jk}|\Omega, Y_{j(k-1)}) \ = \ \int p(y_{jk}|\Omega, x_j, Y_{j(k-1)})\ p(x_j|\Omega, Y_{j(k-1)})\ dx_j$$

is Gaussian with moments given in Step 2a of Algorithm 2.

The mean and variance in Equation (10) are initialized based on the Gaussian AR(1) transition density and are updated recursively according to Step 2b in Algorithm 2. The posterior normality and the moment recursions follow from the fact that $p(x_j|\Omega, Y_{jk}) \ \propto \ p(x_j|\Omega, Y_{j(k-1)})\ p(y_{jk}|x_j, \Omega)$, where both densities on the right side are normal.

We are therefore able to express the joint posterior density $p(\Omega, x_j|\ Y_{jk})$ as a mixture of normals.

Hence, our within-day learning algorithm sequentially updates the weights, means and variances of the normal mixture as new data are observed. The algorithm, described below, is equivalent to the mixture Kalman filter proposed by Liu and Chen (2000) where the state variable is static.

Algorithm 2: Within-Day Learning

Step 0: Start with an MCMC sample, $\Omega^{(1)}, \ldots, \Omega^{(M)}$, drawn from $p(\Omega|Y_{1:j-1})$.

Step 1: Initialize the mixture weights, means and variances for each $i = 1, \ldots, M$:

$$w_0^{(i)} \;=\; M^{-1}, \quad m_0^{(i)} \;=\; \alpha_{d_j}^{(i)} + \beta^{(i)} \left(x_{j-1}^{(i)} - \alpha_{d_{j-1}}^{(i)} \right), \quad v_0^{(i)} \;=\; \left(\psi^2\right)^{(i)}.$$

Step 2: For each period $k = 1, \ldots, K$ and each $i = 1, \ldots, M$:

Step 2a: Update the mixture weights:

$$w_k^{(i)} \;\propto\; w_{k-1}^{(i)} \; \phi\left(y_{jk} \mid g_{d_j}(t_k)^{(i)} \; m_{k-1}^{(i)}, \; \left(g_{d_j}(t_k)^{(i)} \right)^2 v_{k-1}^{(i)} + \left(\sigma^2\right)^{(i)} \right).$$

Step 2b: Update the mixture means and variances:

$$m_k^{(i)} \;=\; v_k^{(i)} \left(\frac{m_{k-1}^{(i)}}{v_{k-1}^{(i)}} + \frac{y_{jk} \; g_{d_j}(t_k)^{(i)}}{(\sigma^2)^{(i)}} \right)^{-1}, \quad v_k^{(i)} \;=\; \left(\frac{1}{v_{k-1}^{(i)}} + \frac{\left(g_{d_j}(t_k)^{(i)} \right)^2}{(\sigma^2)^{(i)}} \right)^{-1}.$$

In Step 2a, $\phi(x|m, v)$ is a normal density evaluated at x with mean m and variance v, and the weights are normalized so that $\sum_{i=1}^{M} w_k^{(i)} = 1$.

The algorithm provides a closed form representation of the posterior density at each period k. Given this information, the prediction densities for a future period k', conditioning on the historical data $Y_{1:(j-1)}$, are given by

$$\begin{aligned} p(\lambda_{jk'}|Y_{jk}) &= \iint p(\lambda_{jk'}|x_j, z_{d_j}(t_{k'})) \; p(x_j|x_{j-1}, \theta) \; p(z_{d_j}(t_{k'})|x_j, \theta, Y_{jk}) \; dx_{j-1} \; d\theta \\ p(N_{jk'}|Y_{jk}) &= \iint p(N_{jk'}|\lambda_{jk'}, \theta) \; p(\lambda_{jk'}, \theta|Y_{jk}) \; d\lambda_{jk'} \; d\theta. \end{aligned}$$

As noted previously, these densities are not available in closed form as the rates and counts are nonlinear functions of x_j and Ω. Therefore we propose a resampling algorithm in order to generate draws from these distributions which is given below.

Algorithm 3: Within-Day Forecasting at Period k

Step 0: Start with $\{\Omega^{(1)}, w_k^{(1)}, m_k^{(1)}, v_k^{(1)}\}, \ldots, \{\Omega^{(M)}, w_k^{(M)}, m_k^{(M)}, v_k^{(M)}\}$ from Algorithm 2.

Step 1: Draw the indices $l_1, \ldots, l_M \sim \text{Mult}(M; w_k^{(1)}, \ldots, w_k^{(M)})$.

Step 2: For each index l_i draw $x_j^{(i)} \sim \mathcal{N}\left(m_k^{(l_i)}, v_k^{(l_i)}\right)$.

Step 3: For each period $k' = k+1, \ldots, K$ and each $i = 1, \ldots, M$,

Step 3a: Set $\lambda_{jk'}^{(i)} = \left(x_j^{(i)}\ g_{d_j}(t_{k'})^{(l_i)}\right)^2$.

Step 3b: Draw $y_{jk'}^{(i)} \sim \mathcal{N}\left(\sqrt{\lambda_{jk'}^{(i)}},\ \left(\sigma^2\right)^{(l_i)}\right)$.

Step 3c: Set $N_{jk'}^{(i)} = \left(y_{jk'}^{(i)}\right)^2 - 0.25$.

Steps 3a and 3c of this algorithm provide samples from $p(\lambda_{jk'}|Y_{1:j-1}, Y_{jk})$ and $p(N_{jk'}|Y_{1:j-1}, Y_{jk})$ respectively. Note also that this algorithm is equivalent to Algorithm 1 at period $k = 0$ if we set $l_i = i$ for all i. We now investigate the impact of incorporating morning information on within-day predictions.

4.3.1 The Importance of Within-Day Information for Forecasting

The comparison is performed as follows. For each day j from July 25 to October 24 (64 days), we run the full MCMC algorithm described in Section 4.1 using data from the previous 100 days. Based on this sample, one-day-ahead forecasts of the arrival rates and call volumes are computed using the procedure from Section 4.2.1. Starting at the end of day $j-1$, we then run the within-day learning algorithm described in Section 4.3 through noon of day j. At 10am and 12pm, we produce forecasts for the rest of the day. In order to provide a true out-of-sample comparison, we evaluate the forecasts using only data after 12pm (108 time periods). The analysis presented below compares the predictive densities of the rates and counts on day j conditioning on three different information sets:

- Information at the end of day $j-1$: $p(\cdot|Y_{(j-100):(j-1)})$.
- Information up to 10am on day j: $p(\cdot|Y_{(j-100):(j-1)}, Y_{j(37)})$.
- Information up to 12pm on day j: $p(\cdot|Y_{(j-100):(j-1)}, Y_{j(61)})$.

Table 4 summarizes the results of the forecasting exercise, from which several encouraging points emerge. First, the empirical coverage probabilities of the prediction intervals are very close to the nominal value for all three information sets, as the mean coverage across all 64 days is 92.6%, 93.8% and 95.3%, respectively. As expected, the average width of the forecast interval decreases as

more data are observed, the average width of the forecast interval goes from 67.3 at the end of the previous day to 61.1 to 60.8 for the 10am and 12pm forecasts, respectively. The same improvements are not observed in the RMSE. This is partly due to three days where unexpected peaks in the data between 9:30am and 12pm offset the forecasts and consequently result in an overprediction of the call volume for the whole afternoon and evening.

	RMSE			**Coverage Probability**			**Average Width**		
	PD	10am	12pm	PD	10am	12pm	PD	10am	12pm
Min	11.33	11.08	11.07	0.593	0.519	0.519	62.45	56.30	55.90
25th	13.17	14.00	13.56	0.958	0.914	0.935	64.96	58.87	58.78
50th	14.60	15.50	14.80	0.972	0.963	0.963	66.31	60.74	60.42
Mean	16.93	17.86	16.59	0.953	0.926	0.938	67.37	61.11	60.80
75th	18.11	19.87	16.58	0.991	0.982	0.982	69.50	63.50	62.34
Max	52.48	57.72	53.66	1.000	1.000	1.000	77.09	70.86	70.01

Table 4: Summary of predicted call volumes handled by the call center between 12:05pm and 9:05pm at three different times of day: at the end of the previous day (PD), on the same day at 10am, on the same day at 12pm.

Figure 7 provides a graphical summary of the predictive distribution on September 2. The top two plots display the mean and 95% equal-tailed interval for the rates and counts after 12pm. As expected, the same-day forecast of the rates have much narrower intervals than those of the previous day. The tightening of the predictive intervals for the counts is not as striking. This is because the inherent Poisson variability dominates the uncertainty about the rates.

We observe a dramatic shift in the forecasts after incorporating same-day observations. The reason for this is apparent as September 2 corresponds to the day after Labor day. Consequently, the call center experienced an unusually high call volume that morning, leading to a significant upward shift in the 10am and 12pm estimates. An even larger shift would have been observed if we had not modeled that day as a Monday. Looking closer at the forecasts, the credible intervals seem nearly parallel across different prediction times. The shift suggests that the estimate of x_j is more accurate having observed the early morning data but that the added information has not modified the estimates of g or θ significantly.

The predictive densities of the rates and counts at 2pm are displayed in the bottom two plots of Figure 7. These were obtained by applying the `density` function in S-plus to the Monte Carlo samples. The upward shift and narrowing of the distribution for the rates are clearly seen. In addition, we note that the predictive densities for the rates and counts are both quite well approximated by a normal distribution.

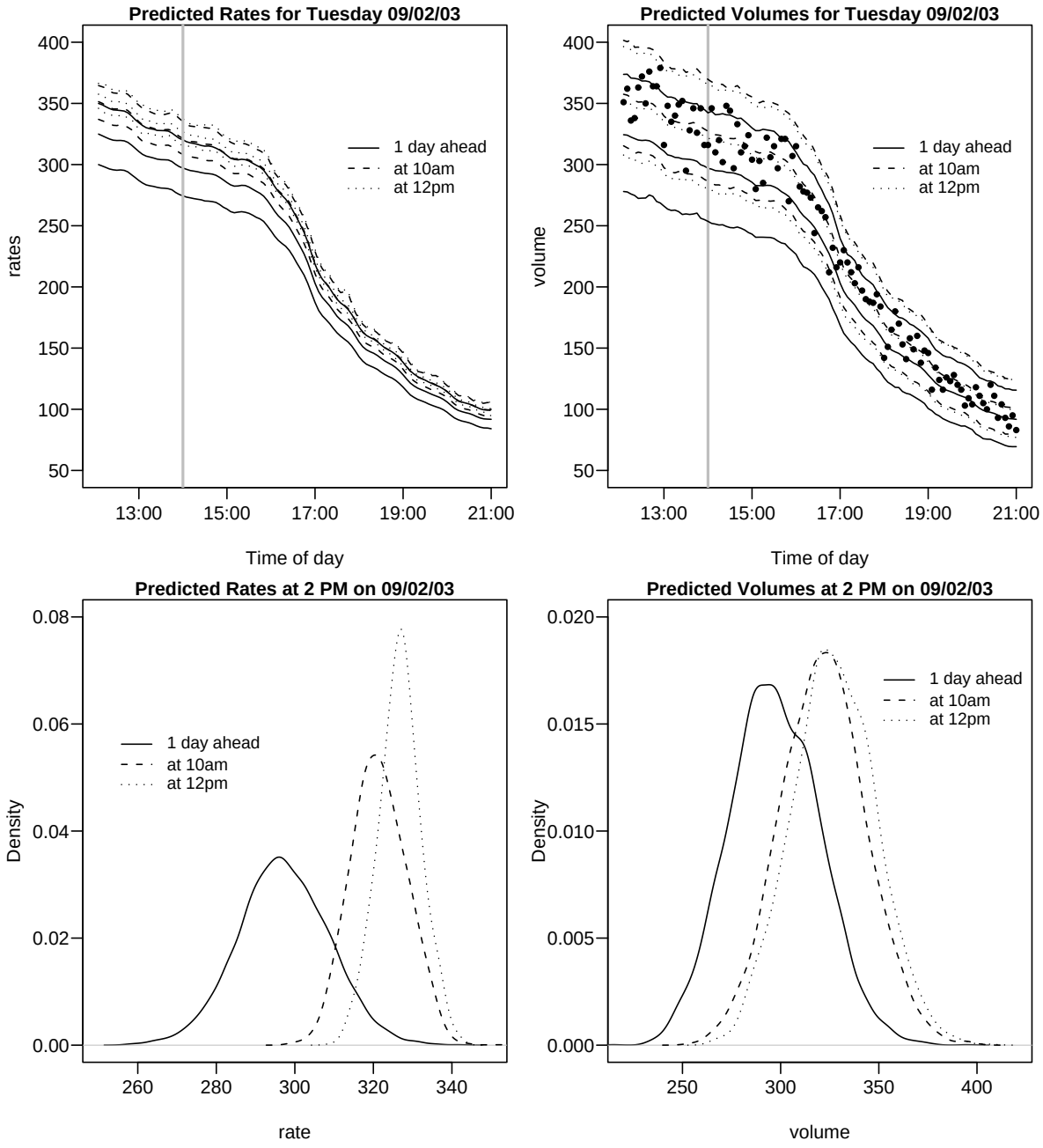

Figure 7: Forecasts of the Poisson rates and call volumes on September 2 using three different information sets. *Top Left:* Forecast mean and 95% intervals for the Poisson rates between 12:05pm and 9:05pm. *Top Right:* Forecast mean and 95% intervals for the call volumes between 12:05pm and 9:05pm. *Bottom Left:* Forecast densities for the Poisson rate at 2:00pm. *Bottom Right:* Forecast densities for the call volume at 2:00pm. Arrow indicates the actual observation.

5 Conclusion

In this article, we provide a multiplicative Gaussian model to measure and predict the arrival rates of an inhomogeneous Poisson process. We fitted this model to call center data provided to us by a North American commercial bank. Based on the results in this study, we find that the fitted model explains most of the variance in the data as the empirical results closely approximate the theory on which this model is based. In addition, the model proposed in this paper clearly outperforms two existing models used by practitioners in the industry when used to predict one-day-ahead arrival rates. With the Bayesian procedures used to fit this model, we are not only able to provide point estimates of the current and future arrival rates but also entire distributions on the parameters, state variables and observables. We believe this is a considerable contribution as call center practitioners need confidence levels on their estimates in order to staff their centers appropriately. Finally, we provide a within-day learning algorithm that enables sequential estimation of the rate as new data reaches the call center. This is particularly useful for managers as they can update their prediction of the rates for the afternoon based on the observed morning pattern of calls and restaff their center accordingly.

Our findings extend the work of Brown *et al.* (2005) who provided a model to predict arrival rates of an inhomogeneous Poisson process. We believe we provide a more statistically sound estimation procedure to the model presented in this paper. In addition, the Bayesian procedure offers several improvements over the iterative least squares procedure. First, as stated above, Bayesian methods automatically provide measures of uncertainty of parameter and latent state estimates. Furthermore, prior knowledge based on the past experience of the manager can be incorporated in the model. In addition, further unobserved components and covariates can be incorporated in the model with little or no complications. Finally, the model presented in this paper can provide k-day-ahead prediction of future arrival rates which is currently unavailable in the work presented by Brown *et al.* (2005).

Although the method proposed in this paper has numerous advantages, we are still left with a small complication that was mentioned previously in the paper. In the model, we impose a constraint on the cubic spline. We do not incorporate this constraint in the prior as the prior would no longer be Gaussian or closed form. We therefore draw from the unconstrained prior and renormalize the posterior so that the constraint is satisfied. Our belief is that the posterior distribution from this procedure is very close to the posterior distribution if we had incorporated the constraint in the prior.

A more general issue arising from the analysis is the problem of distinguishing the two variance components in an AR(1) plus noise model if the autoregressive coefficient is close to zero. To our knowledge, an objective prior for this model has not yet been derived. This is a very interesting problem as this is one of most widely used models in time series analysis due to its relation with

the state space representation. Investigation of this problem is ongoing but still at a very early stage.

Finally, the root-unroot methodology holds several advantages due to its relation with a Gaussian model. As we have seen in this paper, conjugate Gaussian priors with the required covariance structure can be used to model the time series dynamics. Conjugate gamma priors with the equivalent dynamics are currently being investigated in order to model the original counts rather than performing a transformation.

References

[1] Andrews, B. H. and Cunningham S. M. (1995), "L.L. Bean improves call-center forecasting", *Interfaces*, 25, 1-13.

[2] Avramidis, A. N., Deslauriers, A., L'Ecuyer, P. (2004), "Modeling daily arrivals to a telephone call center", *Management Science*, 50, 896-908.

[3] Bianchi, L., Jarrett, J. and Choudary Hanumara, R. (1993), "Forecasting incoming calls to telemarketing centers", *The Journal of Business Forecasting Methods and Systems*, 12, 3-12.

[4] Brown, L. D., Zhang, R., Zhao, L. (2001), "Root un-root methodology for non parametric density estimation", *Technical Report, University of Pennsylvania.*

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

[6] Carter, C. K., Kohn, R. (1994), "On Gibbs sampling for state space models", *Biometrika*, 81, 541-53.

[7] Chen, R., Liu, J.S (2000), "Mixture Kalman Filters", *Journal of the Royal Statistical Society. Series B*, 62, 493-508.

[8] Chen, R., Fomby, T. (1999), "Forecasting with stable seasonal pattern models with an application of Hawaiian tourist data", *Journal of Business and Economic Statistics*, 17, 497-504.

[9] Diebold, F. X., Gunther, T., Tay, A. (1998), "Evaluating density forecasts, with applications to financial risk management", *International Economic Review*, 39, 863-883.

[10] Frühwirth-Schnatter, S. (1994), "Data augmentation and dynamic linear models", *Journal of Time Series Analysis*, 15, 183-202.

[11] Gamerman, D. (1998), "Markov chain Monte Carlo for dynamic generalized linear models", *Biometrika*, 85, 215-227.

[12] Gans, N., Koole, G., and Mandelbaum, A. (2003), "Telephone call centers: tutorial, review and research prospects", *Manufacturing and Service Operations Management*, 5, 79-141.

[13] Gel, Y., Raftery, A. E., Gneiting, T. (2004), "Calibrated probabilistic mesoscale weather field forecasting: The geostatistical output perturbation (GOP) method", *Journal of the American Statistical Association*, 99, 575-583.

[14] Jongbloed G., Koole, G. (2001), "Managing Uncertainty in Call Centers using Poisson Mixtures", *Applied Stochastic Models in Business and Industry*, 17, 307-318.

[15] Kohn, R. and Ansley, C. F. (1987), "A new algorithm for spline smoothing based on smoothing a stochastic process", *SIAM J. Sci. Statist. Comput.*, 8, 33-48.

[16] Lindley D. V., Smith, A.F.M. (1972), "Bayes estimates for the linear model", *Journal of the Royal Statistical Society. Series B*, 34, 1-41.

[17] Little J. (1961), "A proof of the theorem L = λ W', *Operations Research*, 9, 383-387.

[18] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M.N., Teller, A.H, and Teller, E. (1953), "Equation of state calculations by fast computing machines", *Journal of Chemical Physics* 21, 1087-1092.

[19] Robert, C. P., Casella G. (2004),"Monte Carlo statistical methods", *Springer-Verlag, New York*.

[20] Rosenblatt, M. (1952), "Remarks on a multivariate transformation", *Annals of Mathematical Statistics*, 23, 470-472.

[21] Shephard, N. (1994), "Partial non Gaussian state space", *Biometrika*, 81, 115-131.

[22] Shephard, N. and Pitt, M.K. (1997), "Likelihood analysis of non-Gaussian measurement time series", *Biometrika*, 84, 653-67.

[23] Soyer, R. and Tarimcilar, M. (2004), "Modeling and analysis of call center arrival data: a Bayesian approach", Submitted to *Management Science*.

[24] Stein, C. M. (1981), "Estimating the Mean of a Multivariate Normal Distribution", *The Annals of Statistics*, 9, 1135-1151.

[25] Trofimov, V., Feigin, P., Mandelbaum, A. and Ishay E. (2005), "DATA-MOCCA, Data Model for Call Center Analysis".

[26] Wahba, G. (1983), "Bayesian 'confidence intervals' for the cross-validated smoothing spline", *Journal of the Royal Statistical Society, Series B.*, 45, 133-50.

APPENDIX: MCMC algorithm

A combination of the Gibbs sampler and the Metropolis algorithm are used to draw from the posterior distribution $p(x, z, \alpha, \beta, \tau^2, \sigma^2, \psi^2 \mid Y_{1:J})$. The algorithm outline followed by a description of the full conditional posterior distributions are given below.

1: Initialize $x, \alpha, \beta, \psi^2, \tau^2, \sigma^2$.

2: Sample z_i from $z_i|x, \tau_i^2, \sigma^2, Y_{1:J}$, for $i = 1, \ldots, 5$.

3: Sample x from $x|z, \alpha, \beta, \psi^2, \sigma^2, Y_{1:J}$.

4: Sample α from $\alpha|x, \beta, \psi^2$.

5: Sample β from $\beta|x, \alpha, \psi^2$.

6: Sample ψ^2 from $\psi^2|x, \alpha, \beta$.

7: Sample τ_i^2 from $\tau_i^2|z_i$, for $i = 1, \ldots, 5$.

8: Sample σ^2 from $\sigma^2|x, z, Y_{1:J}$.

9: Go to Step 2.

Step 2: Sampling $z_i : i = 1, \ldots, 5$

Conditional on the x's, the model can be cast into state space form:

$$\begin{aligned} w_i(t_k) &= h' \; z_i(t_k) + e_k \qquad e_k \sim \mathcal{N}(0, \nu_k^2) \\ z_i(t_k) &= F(\delta) \; z_i(t_{k-1}) + u_k \qquad u_k \sim \mathcal{N}(0, \tau_i^2 U(\delta)) \quad k = 1, \ldots, K \end{aligned}$$

where $\nu_k^2 = \sigma^2 \; \left(\sum_{\{j \; : \; d_j = i\}} x_j^2 \right)^{-1}$ and $w_i(t_k) = \frac{\nu_k^2}{\sigma^2} \; \sum_{\{j \; : \; d_j = i\}} y_{jk} \; x_j$. The FFBS algorithm can be used to draw from the required posterior density $p(z_i|\tau_i^2, \sigma^2, Y_{1:J})$. Given that $z_i(t_k) = \{g_i(t_k), dg_i(t_k)/dt_k\}$, we consequently obtain a draw of $g_i(t_k)$. We then normalize the posterior draws by setting $g_i(t_k) = g_i(t_k) \left(\sum_{k=1}^K g_i(t_k)^2 \right)^{-\frac{1}{2}}$ so that the restriction $\sum_{k=1}^K g_i(t_k)^2 = 1$ is satisfied.

Step 3: Sampling x

Conditional on z, the model can be cast into state space form:

$$\begin{aligned} v_j &= x_j + \zeta_j \qquad \zeta_j \sim \mathcal{N}(0, \gamma_j^2) \\ x_j &= \alpha_{d_j} + \beta(x_{j-1} - \alpha_{d_{j-1}}) + \eta_j \qquad \eta_j \sim \mathcal{N}(0, \psi^2) \end{aligned}$$

where $\gamma_j^2 = \sigma^2 \; \left(\sum_{k=1}^K g_{d_j}(t_k) \right)^{-2}$ and $v_j = \frac{\gamma_j^2}{\sigma^2} \; \sum_{k=1}^K y_{jk} \; g_{d_j}(t_k)$. The FFBS algorithm is then used to draw from the required posterior density.

Step 4: Sampling α

The full conditional posterior distribution for α is given by

$$\begin{aligned} p(\alpha|x,\beta,\psi^2) &\propto p(x|\alpha,\beta,\psi^2)\ p(\alpha) \\ &\propto \mathcal{N}\left((X'X)^{-1}X'W,\ (X'X)^{-1}\psi^2\right)\ \left(\sum_{i=1}^{4}(\alpha_i-\overline{\alpha})^2\right)^{-1} \end{aligned}$$

where

$$X = \begin{pmatrix} -\beta & 1 & 0 & 0 & 0 \\ 0 & -\beta & 1 & 0 & 0 \\ 0 & 0 & -\beta & 1 & 0 \\ 0 & 0 & 0 & -\beta & 1 \\ 1 & 0 & 0 & 0 & -\beta \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix} \text{ and } W = \begin{pmatrix} x_2-\beta x_1 \\ x_3-\beta x_2 \\ \vdots \\ x_J-\beta x_{J-1} \end{pmatrix}.$$

The random walk Metropolis algorithm is applied to draw α since the prior is not conjugate. Given the current value $\alpha^{(q-1)}$, we draw $\alpha^{(*)}$ from the proposal density $\mathcal{N}(\alpha^{(q-1)}, 0.5\ I)$ and then accept $\alpha^{(*)}$ with probability

$$\min\left\{1,\ \frac{p(\alpha^{(*)}|x,\beta,\psi^2)}{p(\alpha^{(q-1)}|x,\beta,\psi^2)}\right\}.$$

The variance $0.5\ I$ of the proposal density was chosen to provide an acceptance rate of approximately 30%.

Step 5: Sampling β

Under the prior specified in section 3.2, the full conditional density of β is given by

$$\begin{aligned} p(\beta|x,\alpha,\psi^2) &\propto p(x|\alpha,\beta,\psi^2)\ p(\beta) \\ &\propto \mathcal{N}\left(\frac{\sum_{j=1}^{J-1}(x_j-\alpha_{d_j})(x_{j+1}-\alpha_{d_{j+1}})}{\sum_{j=1}^{J-1}(x_j-\alpha_{d_j})^2}, \frac{\psi^2}{\sum_{j=1}^{J-1}\left(x_j-\alpha_{d_j}\right)^2}\right)\ \mathcal{U}[0,1]. \end{aligned}$$

The random walk Metropolis algorithm is employed to sample β since the prior is not conjugate. Given the current value of $\beta^{(q-1)}$, draw $\beta^{(*)}$ from the proposal density $\mathcal{N}(\beta^{(q-1)}, 0.01)$ and accept with probability

$$\min\ \left\{1,\ \frac{p(\beta^{(*)}|x,\alpha,\psi^2)}{p(\beta^{(q-1)}|x,\alpha,\psi^2)}\right\}.$$

The variance 0.01 of the proposal density was chosen to ensure an acceptance rate of approximately 30%.

Step 6: Sampling ψ^2

Assuming the standard conjugate inverse gamma prior $\mathcal{IG}(a_\psi, b_\psi)$, ψ^2 is sampled from

$$p(\psi^2|x,\alpha,\beta) \;=\; \mathcal{IG}\left(a_\psi + \frac{J-1}{2},\; b_\psi + \frac{1}{2}\sum_{j=2}^{J}\left(x_j - \alpha_{d_j} - \beta(x_{j-1} - \alpha_{d_{j-1}})\right)^2\right).$$

Step 7: Sampling $\tau_i^2 : i = 1, \ldots, 5$

Following Ansley and Kohn (1985), if we assume a conjugate inverse gamma prior $\mathcal{IG}(a_{\tau_i}, b_{\tau_i})$, then the full conditional posterior for τ_i^2 is

$$p(\tau_i^2|z_i) \;=\; \mathcal{IG}\left(a_{\tau_i} + K - 2,\; b_{\tau_i} + \frac{1}{2}\sum_{k=2}^{K}\Gamma_i(t_k)'U(\delta)^{-1}\Gamma_i(t_k)\right)$$

where $\Gamma_i(t_k) \;=\; z_i(t_k) - F(\delta)z_i(t_{k-1})$.

Step 8: Sampling σ^2

Assuming the standard conjugate inverse gamma prior $\mathcal{IG}(a_\sigma, b_\sigma)$, σ^2 is sampled from

$$p(\sigma^2|z,x,Y_{1:J}) \;=\; \mathcal{IG}\left(a_\sigma + \frac{JK}{2},\; b_\sigma + \frac{1}{2}\sum_{j=1}^{J}\sum_{k=1}^{K}\left(y_{jk} - g_{d_j}(t_k)x_j\right)^2\right).$$