The Annals of Applied Statistics
2011, Vol. 5, No. 2B, 1379–1406
DOI: 10.1214/10-AOAS442
© Institute of Mathematical Statistics, 2011

FORECASTING EMERGENCY MEDICAL SERVICE CALL ARRIVAL RATES[1]

BY DAVID S. MATTESON, MATHEW W. MCLEAN, DAWN B. WOODARD AND SHANE G. HENDERSON

Cornell University

We introduce a new method for forecasting emergency call arrival rates that combines integer-valued time series models with a dynamic latent factor structure. Covariate information is captured via simple constraints on the factor loadings. We directly model the count-valued arrivals per hour, rather than using an artificial assumption of normality. This is crucial for the emergency medical service context, in which the volume of calls may be very low. Smoothing splines are used in estimating the factor levels and loadings to improve long-term forecasts. We impose time series structure at the hourly level, rather than at the daily level, capturing the fine-scale dependence in addition to the long-term structure.

Our analysis considers all emergency priority calls received by Toronto EMS between January 2007 and December 2008 for which an ambulance was dispatched. Empirical results demonstrate significantly reduced error in forecasting call arrival volume. To quantify the impact of reduced forecast errors, we design a queueing model simulation that approximates the dynamics of an ambulance system. The results show better performance as the forecasting method improves. This notion of quantifying the operational impact of improved statistical procedures may be of independent interest.

1. Introduction. Considerable attention has been paid to the problem of how to best deploy ambulances within a municipality to minimize their response times to emergency calls while keeping costs low. Sophisticated operations research models have been developed to address issues such as the optimal number of ambulances, where to place bases, and how to move ambulances in real time via system-status management [Swersey (1994); Goldberg (2004); Henderson (2009)]. However, methods for estimating the inputs to these models, such as travel times on road networks and call arrival rates, are ad hoc. Use of inaccurate parameter estimates in these models can result in poor deployment decisions, leading to low performance and diminished user confidence in the software. We introduce methods for estimating the demand for ambulances, that is, the total number of emergency calls per period, that are highly accurate, straightforward to implement, and have

Received July 2010; revised November 2010.

[1] Supported in part by NSF Grant CMMI-0926814.

Key words and phrases. Ambulance planning, dynamic factor model, nonhomogeneous Poisson process, integer-valued time series, smoothing splines.

the potential to simultaneously lower operating costs while improving response times.

Current practice for forecasting call arrivals is often rudimentary. For instance, to estimate the call arrival rate in a small region over a specific time period, for example, next Monday from 8–9 a.m., simple estimators have been constructed by averaging the number of calls received in the corresponding period in four previous weeks: the immediately previous two weeks and the current and previous weeks of the previous year. Averages of so few data points can produce highly noisy estimates, with resultant cost and efficiency implications. Excessively large estimates lead to over-staffing and unnecessarily high costs, while low estimates lead to under-staffing and slow response times. Setzler, Saydam and Park (2009) document an emergency medical service (EMS) agency which extends this simple moving average to twenty previous observations: the previous four weeks from the previous five years. A more formal time series approach is able to account for possible differences from week to week and allows inclusion of neighboring hours in the estimate.

We generate improved forecasts of the call-arrival volume by introducing an integer-valued time series model with a dynamic latent factor structure for the hourly call arrival rate. Day-of-week and week-of-year effects are included via simple constraints on the factor loadings. The factor structure allows for a significant reduction in the number of model parameters. Further, it provides a systematic approach to modeling the diurnal pattern observed in intraday counts. Smoothing splines are used in estimating the factor levels and loadings. This may introduce a small bias in some periods, but it offers a significant reduction in long-horizon out-of-sample forecast-error variance. This is combined with integer-valued time series models to capture residual dependence and to provide adaptive short-term forecasts. Our empirical results demonstrate significantly reduced error in forecasting hourly call-arrival volume.

Few studies have focused specifically on EMS call arrival rates, and of those that have proposed methods for time series modeling, most have been based on Gaussian linear models. Even with a continuity correction, this method is highly inaccurate when the call arrival rate is low, which is typical of EMS calls at the hourly level. Further, it conflicts with the Poisson distribution assumption used in operations research methods for optimizing staffing levels. For example, Channouf et al. (2007) forecast EMS demand by modeling the daily call arrival rate as Gaussian, with fixed day-of-week, month-of-year, special day effects and fixed day-month interactions. They also consider a Gaussian autoregressive moving-average (ARMA) model with seasonality and special day effects. Hourly rates are later estimated either by adding hour-of-day effects or assigning a multinomial distribution to the hourly volumes, conditional on the daily call volume estimates.

Setzler, Saydam and Park (2009) provide a comparative study of EMS call volume predictions using an artificial neural network (ANN). They forecast at various temporal and spatial granularities with mixed results. Their approach offered a significant improvement at low spatial granularity, even at the hourly level. At a high spatial granularity, the mean square forecast error (MSFE) of their approach did not improve over simple binning methods at a temporal granularity of three hours or less.

Methods for the closely related problem of forecasting call center demand have received much more study. Bianchi, Jarrett and Choudary Hanumara (1998) and Andrews and Cunningham (1995) use ARMA models to improve forecasts for daily call volumes in a retail company call center and a telemarketing center, respectively. A dynamic harmonic regression model for hourly call center demand is shown in Tych et al. (2002) to outperform seasonal ARMA models. Their approach accounts for possible nonstationary periodicity in a time series. The major drawback common to these studies is that the integer-valued observations are assumed to have a continuous distribution, which is problematic during periods with low arrival rates.

The standard industry assumption is that hourly call-arrival volume has a Poisson distribution. The Palm–Khintchine theorem—stating that the superposition of a number of independent point processes is approximately Poisson—provides a theoretical basis for this assumption [see, e.g., Whitt (2002)]. Brown et al. (2005) provide a comprehensive analysis of operational data from a bank call center and thoroughly discuss validating classical queueing theory, including this theorem. Henderson (2005) states that we can expect the theorem to hold for typical EMS data because there are a large number of callers who can call at any time and each has a very low probability of doing so.

Weinberg, Brown and Stroud (2007) use Bayesian techniques to fit a nonhomogeneous Poisson process model for call arrivals to a commercial bank's call center. This approach has the advantage that forecast distributions for the rates and counts may be easily obtained. They incorporate smoothness in the within-day pattern. They implement a variance stabilizing transformation to obtain approximate normality. This approximation is most appropriate for a Poisson process with high arrival rates, and would not be appropriate for our application in which very low counts are observed in many time periods.

Shen and Huang (2008b) apply the same variance stabilizing transformation and achieve better performance than Weinberg, Brown and Stroud (2007). They use a singular value decomposition (SVD) to reduce the number of parameters in modeling arrival rates. Their approach is used for intraday updating and forecasts up to one day ahead.

Shen and Huang (2008a) propose a dynamic factor model for 15-minute call arrivals to a bank call center. They assume that call arrivals are a Cox process. A Cox process [cf. Cox and Isham (1980)] is a Poisson process with a stochastic intensity, that is, a doubly stochastic Poisson process. The factor structure reduces the number of parameters by explaining the variability in the call arrival rate with a small number of unobserved variables. Estimation proceeds by iterating between an SVD and fitting Poisson generalized linear models to successively estimate the factors and their respective loadings. The intensity functions are assumed to be serially dependent. Forecasts are made by fitting a simple autoregressive time series model to the factor score series.

We assume that the hourly EMS call-arrival volume has a Poisson distribution. This allows parsimonious modeling of periods with small counts, conforms with the standard industry assumption, and avoids use of variance stabilizing transformations. We assume the intensity function is a random process and that it can be forecast using previous observations. This has an interpretation very similar to a Cox process, but is not equivalent since the random intensity is allowed to depend on not only its own history, but also on previous observations. We partition the random intensity function into stationary and nonstationary components.

Section 2 describes the general problem and our data set. Section 3 presents the proposed methodology. We consider a dynamic latent factor structure to model the nonstationary pattern in intraday call arrivals and greatly reduce the number of parameters. We include day-of-week and week-of-year covariates via simple constraints on the factor loadings of the nonstationary pattern. Smoothing splines are easily incorporated into estimation of the proposed model to impose a smooth evolution in the factor levels and loadings, leading to improved long-horizon forecast performance. We combine the factor model with stationary integer-valued time series models to capture the remaining serial dependence in the intensity process. This is shown to further improve short-term forecast performance of our approach. A simple iterative algorithm for estimating the proposed model is presented. It can be implemented largely through existing software. Section 4 assesses the performance of our approach using statistical metrics and a queueing model simulation. Section 5 gives our concluding remarks.

2. Notation and data description.

We assume that over short, equal-length time intervals, for example, one hour periods, the latent call arrival intensity function can be well approximated as being constant, and that all data have been aggregated in time accordingly. We suppose aggregated call arrivals follow a nonhomogeneous counting process $\{Y_t : t \in \mathbb{Z}\}$, with discrete time index t. Underlying this is a latent, real-valued, nonnegative intensity process $\{\lambda_t : t \in \mathbb{Z}\}$. We further assume that conditional on λ_t, Y_t has a Poisson distribution with mean λ_t.

As shown in Figure 1, the pattern of call arrivals over the course of a typical day has a distinct shape. After quickly increasing in the late morning, it peaks in the early afternoon, then slowly falls until it troughs between 5 and 6 a.m. See Section 4 for more discussion. In our analysis, we consider an arrival process that has been repeatedly observed over a particular time span, specifically, a 24 hour day. Let

$$\{y_t : t = 1, \dots, n\} \equiv \{y_{ij} : i = 1, \dots, d;\ j = 1, \dots, m\}$$

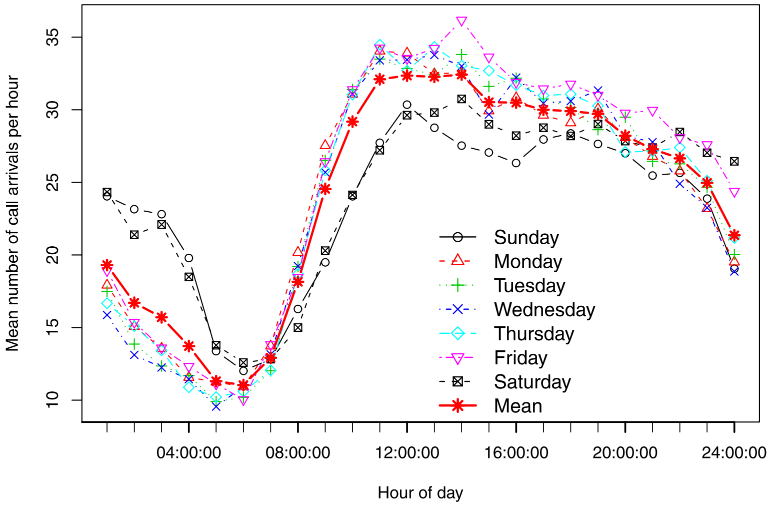

FIG. 1. *Mean number of calls per hour by day of the week.*

denote the sequence of call arrival counts, observed over time period t, which corresponds one-to-one with the jth sub-period of the ith day, so that $n = dm$. Our baseline approach is to model the arrival intensity λ_t for the distinct shape of intraday call arrivals using a small number of smooth curves.

We consider two disjoint information sets for predictive conditioning. Let $\mathcal{F}_t = \sigma(Y_1, \dots, Y_t)$ denote the σ-field generated by $Y_1, \dots, Y_t$, and let $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ denote any available deterministic covariate information about each observation. We incorporate calendar information such as day-of-week and week-of-year in our analysis. We define λ_t as the conditional expectation of Y_t given $\mathcal{F}_{t-1}$ *and* $\mathbf{X}$. We defined this above as the mean of Y_t. In our *model* these coincide; however, this mean may not be the same as the conditional expectation since λ_t may depend on other unobserved random variables. Let $\mu_t = E(Y_t|\mathbf{X}) > 0$ denote the conditional mean of Y_t given only the covariates $\mathbf{X}$. Let

$$\lambda_t = E(Y_t|\mathcal{F}_{t-1}, \mathbf{X}) = \mu_t E(Y_t/\mu_t|\mathcal{F}_{t-1}, \mathbf{X}) = \mu_t \eta_t, \tag{1}$$

in which $\eta_t > 0$ is referred to as the conditional intensity inflation rate (CIIR). By construction,

$$E(\eta_t|\mathbf{X}) = E(E(Y_t|\mathcal{F}_{t-1}, \mathbf{X})|\mathbf{X})/\mu_t = E(Y_t|\mathbf{X})/\mu_t = 1.$$

The CIIR process is intended to model any remaining serial dependence in call arrival counts after accounting for available covariates. In the EMS context we hypothesize that this dependence is due to sporadic events such as inclement weather or unusual traffic patterns. Since information regarding these events may not be available or predictable in general, we argue that an approach such as ours which explicitly models the remaining serial dependence will lead to improved short-term forecast accuracy. In Section 3 we consider a dynamic latent factor model estimated with smoothing splines for modeling μ_t, various time series models for modeling η_t, and finally a conditional likelihood algorithm for estimating the latent intensity process λ_t via estimation of η_t given μ_t.

The call arrival data used consists of all emergency priority calls received by Toronto EMS between January 1, 2007 and December 31, 2008 for which an ambulance was dispatched. This includes some calls not requiring lights-and-sirens response, but does not include scheduled patient transfers. We include only the first call arrival time in our analysis when multiple calls are received for the same event. The data were processed to exclude calls with no reported location. These removals totaled less than 1% of the data.

Many calls resulted in multiple ambulances being dispatched. Exploratory analysis revealed that the number of ambulances deployed for a single emergency did not depend on the day of the week, the week of the year, or exhibit any serial dependence. However, such instances were slightly more prevalent in the morning hours. Our analysis of hourly ambulance demand defines an event as a call arrival if *one or more* ambulances are deployed.

We removed seven days from the analysis because there were large gaps, over at least two consecutive hours, in which no emergency calls were received. These days most likely resulted from malfunctions in the computer-aided dispatch system which led to failures in recording calls for extended periods. Strictly speaking, it is not necessary to remove the entire days; however, we did so since it had a negligible impact on our results and it greatly simplified out-of-sample forecast comparisons and implementation of the simulation studies in Section 4.

Finally, we gave special consideration to holidays. We found that the intraday pattern on New Year's Eve and Day was fundamentally different from the rest of the year and removed these days from our analysis. This finding is similar to the conclusions of Channouf et al. (2007) who found that New Year's Day and the dates of the Calgary Stampede were the only days requiring special consideration in their methodology when applied to the city of Calgary. In practice, staffing decisions for holidays require special planning and consideration of many additional variables.

3. Modeling.

Factor models provide a parsimonious representation of high dimensional data in many applied sciences, for example, econometrics [cf. Stock and Watson (2002)]. We combine a dynamic latent factor model with integer-valued time series models. We include covariates via simple constraints on the factor loadings. We estimate the model using smoothing splines to impose smooth evolution in the factor levels and loadings. The factor model provides a parsimonious representation of the nonstationary pattern in intraday call arrivals, while the time series models capture the remaining serial dependence in the arrival rate process.

3.1. *Dynamic latent factor model.* For notational simplicity, assume m consecutive observations per day are available for d consecutive days with no omissions in the record. Let $\mathbf{Y} = (y_{ij})$ denote the $d \times m$ matrix of observed counts for each day i over each sub-period j. Let $\mu_{ij} = E(Y_{ij}|\mathbf{X})$, and let $\mathbf{M} = (\mu_{ij})$ denote the corresponding $d \times m$ latent nonstationary intensity matrix. To reduce the dimension of the intensity matrix $\mathbf{M}$, we introduce a K-factor model.

We assume that the intraday pattern of expected hourly call arrivals on the log scale can be well approximated by a linear combination of (a small number) K factors or functions, denoted by $\mathbf{f}_k$ for $k = 1, \ldots, K$. The factors are orthogonal length-m vectors. The intraday arrival rate model $\boldsymbol{\mu}_i$ over a particular day i is given by

$$\log \boldsymbol{\mu}_i = L_{i1}\mathbf{f}_1 + \cdots + L_{iK}\mathbf{f}_K. \tag{2}$$

Each of the factors $\mathbf{f}_k$ varies as a function over the periods within a day, but they are constant from one day to the next. Day-to-day changes are modeled by allowing the various factor loadings L_{ik} to vary across days. When K is much smaller than either m or d, the dimensionality of the general problem is greatly reduced. In practice, K must be chosen by the practitioner; we provide some discussion on choosing K in Section 4.

In matrix form we have

$$\log \mathbf{M} = \mathbf{L}\mathbf{F}^{\mathrm{T}}, \tag{3}$$

in which $\mathbf{F} = (\mathbf{f}_1, \ldots, \mathbf{f}_K)$ denotes the $m \times K$ matrix of underlying factors and $\mathbf{L}$ denotes the corresponding $d \times K$ matrix of factor loadings, both of which are assumed to have full column rank. Although other link functions are available, the component-wise log transformation implies a multiplicative structure among the K common factors and ensures a positive estimate of each hourly intensity μ_{ij}. Since neither $\mathbf{F}$ nor $\mathbf{L}$ are observable, the expression (3) is not identifiable. We further require $\mathbf{F}^{\mathrm{T}}\mathbf{F} = \mathbf{I}$ to alleviate this ambiguity and we iteratively estimate $\mathbf{F}$ and $\mathbf{L}$.

3.2. *Factor modeling with covariates via constraints.* To further reduce the dimensionality, we impose a set of constraints on the factor loading matrix $\mathbf{L}$. Let $\mathbf{H}$ denote a $d \times r$ full rank matrix ($r < d$) of given constraints (we discuss later what these should be for EMS). Let $\mathbf{B}$ denote an $r \times K$ matrix of unconstrained factor loadings. These unconstrained loadings $\mathbf{B}$ linearly combine to constitute the constrained factor loadings $\mathbf{L}$, such that $\mathbf{L} = \mathbf{HB}$. Our factor model may now be written as

$$\log \mathbf{M} = \mathbf{L}\mathbf{F}^{\mathrm{T}} = \mathbf{H}\mathbf{B}\mathbf{F}^{\mathrm{T}}.$$

A considerable reduction in dimensionality occurs when r is much smaller than d.

Constraints to assure identifiability are standard in factor analysis. The constraints we now consider incorporate auxiliary information about the rows and columns of the observation matrix $\mathbf{Y}$ to simplify estimation and to improve out-of-sample predictions. Similar constraints have been used in Takane and Hunter (2001), Tsai and Tsay (2010) and Matteson and Tsay (2011).

For example, the rows of $\mathbf{H}$ might consist of incidence vectors for particular days of the week, or special days which might require unique loadings on the common factors. We may choose to constrain all weekdays to have identical factor loadings and similarly constrain weekend days. However, this approach is much more general than simple equality constraints, as demonstrated below.

The intraday pattern of hourly call arrivals varies from one day to the next, although the same general shape is maintained. As seen in Figure 1, different days of the week exhibit distinct patterns. We do not observe large changes from one week to the next, but there are significant changes over the course of the year. We allow loadings to slowly vary from week to week. Both of these features are incorporated into the factor loadings $\mathbf{L}$ by specifying appropriate constraints $\mathbf{H}$. Let

$$\log \mathbf{M} = \mathbf{L}\mathbf{F}^{\mathrm{T}} = \mathbf{H}\mathbf{B}\mathbf{F}^{\mathrm{T}} = (\mathbf{H}^{(1)} \quad \mathbf{H}^{(2)}) \begin{pmatrix} \mathbf{B}^{(1)} \\ \mathbf{B}^{(2)} \end{pmatrix} \mathbf{F}^{\mathrm{T}}, \tag{4}$$

in which the first term corresponds to day-of-week effects and the second to smoothly varying week-of-year effects. $\mathbf{H}^{(1)}$ is a $d \times 7$ matrix in which each row $\mathbf{H}_i^{(1)}$ is an incidence vector for the day-of-week. Similarly, $\mathbf{H}^{(2)}$ is a $d \times 53$ matrix in which each row $\mathbf{H}_i^{(2)}$ is an incidence vector for the week-of-year. (We use a 53 week year since the first and last weeks may have fewer than 7 days.) The $7 \times K$ matrix $\mathbf{B}^{(1)} = (\mathbf{b}_1^{(1)}, \dots, \mathbf{b}_K^{(1)})$ contains unconstrained factor loadings for the day-of-week and $\mathbf{B}^{(2)} = (\mathbf{b}_1^{(2)}, \dots, \mathbf{b}_K^{(2)})$ is a $53 \times K$ matrix of factor loadings for the week-of-year.

3.3. *Factor model estimation via smoothing splines.* We assume that as the nonstationary intensity process μ_{ij} varies over the hours j of each day i, it does so smoothly. If each of the common factors $\mathbf{f}_k \in \mathbb{R}^m$ varies smoothly over sub-periods j, then the smoothness of μ_{ij} is guaranteed for each day. Increasing the number of factors reduces possible discontinuities between the end of one day and the beginning of the next. To incorporate smoothness into the model (2), we use Generalized Additive Models (GAMs) in the estimation of the common factors $\mathbf{f}_k$. GAMs extend generalized linear models, allowing for more complicated relationships between the response and predictors by modeling some predictors nonparametrically [see, e.g., Hastie and Tibshirani (1990); Wood (2006)]. GAMs have been successfully used for count-valued data in the study of fish populations [cf. Borchers et al. (1997); Daskalov (1999)]. The factors $\mathbf{f}_k = f_k(j)$ are a smooth function of the intraday time index covariate j. The loadings $\mathbf{L}$ are defined as before. If the loadings $\mathbf{L}$ were known covariates, equation (2) would be a varying coefficient model [cf. Hastie and Tibshirani (1993)].

There are several excellent libraries available in the statistical package *R* [R Development Core Team (2009)] for fitting GAMs, thus making them quite easy to implement. We used the *gam* function from the *mgcv* library [Wood (2008)] extensively. Other popular libraries include the *gam* package [Hastie (2009)] and the *gss* package [Gu (2010)]. See Wood [(2006), Section 5.6] for an introduction to GAM estimation using *R*.

In estimation of the model (2) via the *gam* function, we have used thin plate regression splines with a ten-dimensional basis, the Poisson family, and the log-link function. Thin plate regression splines are a low rank, isotropic smoother with many desirable properties. For example, no decisions on the placement of knots is needed. They are an optimal approximation to thin plate splines and, with the use of Lanczos iteration, they can be fit quickly even for large data sets [cf. Wood (2003)].

When the factors **F** are treated as a fixed covariate, the factor model can again be interpreted as a varying coefficient model. Given the calendar covariates **X**, let

$$\begin{aligned}\log \mu_{ij} &= F_{j1}L_{1i}^{\mathrm{T}} + \cdots + F_{jK}L_{Ki}^{\mathrm{T}} \\ &= \sum_{k=1}^{K} F_{jk}\{\mathbf{H}_i^{(1)\mathrm{T}}\mathbf{b}_k^{(1)} + \mathbf{H}_i^{(2)\mathrm{T}}\mathbf{b}_k^{(2)}\} \qquad (5) \\ &= \sum_{k=1}^{K} F_{jk}\{b_k^{(1)}(\mathbf{x}_i) + b_k^{(2)}(\mathbf{x}_i)\},\end{aligned}$$

in which $b_k^{(1)}(\mathbf{x}_i)$ is a piece-wise constant function of the day-of-week, and $b_k^{(2)}(\mathbf{x}_i)$ is a smoothly varying function over the week-of-year. We may again proceed with estimation via the *gam* function in *R*. Day-of-week covariates are simply added to the linear predictor as indicator variables. These represent a level shift in the daily loadings on each of the factors $\mathbf{f}_k$. In our application it is appropriate to assume a smooth transition between the last week of one year and the first week of the next. To ensure this in estimation of $b_k^{(2)}(\mathbf{x}_i)$, we use a cyclic cubic regression spline for the basis [cf. Wood (2006), Section 5.1]. Iterative estimation of **F**, and **L** via **B**, for a given number of factors K is discussed in Section 3.5.

We allow the degree of smoothness for the factors f_k and the loadings function $b_k^{(2)}(\mathbf{x}_i)$ to be automatically estimated by generalized cross validation (GVC). We expect short term serial dependence in the residuals for our application. For smoothing methods in general, if autocorrelation between the residuals is ignored, automatic smoothing parameter selection may break down [see, e.g., Opsomer, Wang and Yang (2001)]. The proposed factor model may be susceptible to this if the number of days included is not sufficiently large compared to the number of smooth factors and loadings, or if the residuals are long-range dependent. We use what is referred to as a *performance* iteration [cf. Gu (1992)] versus an *outer* iteration strategy which requires repeated estimation for many trial sets of the smoothing parameters. The performance iteration strategy is much more computationally efficient for use in the proposed algorithm, but convergence is not guaranteed, in general. In particular, *cycling* between pairs of smoothing parameters and coefficient estimates may occur [cf. Wood (2006), Section 4.5], especially when the number of factors K is large.

3.4. *Adaptive forecasting with time series models.* Let $\widehat{e}_t = Y_t/\widehat{\mu}_t$ denote the multiplicative residual in period t implied by the fitted values $\widehat{\mu}_t$ from a factor model estimated as described in the previous sections. Time series plots of this residual process appear stationary, but exhibit some serial dependence. In this section we consider time series models for the latent CIIR process $\eta_t = E(Y_t/\mu_t|\mathcal{F}_{t-1}, \mathbf{X})$ to account for this dependence.

To investigate the nature of the serial dependence, we study the bivariate relationship between the $\widehat{e}_t$ process versus several lagged values of the process $\widehat{e}_{t-\ell}$. Scatterplots reveal a roughly linear relationship. Residual autocorrelation and partial autocorrelation plots for one of the factor models fit in Section 4 are given in Figure 5(b) and (c). These quantify the strength of the linear relationship as the lag ℓ increases. It appears to persist for many periods, with an approximately geometric rate of decay as the lag increases.

To explain this serial dependence, we first consider a generalized autoregressive linear model, defined by the recursion

$$\eta_t = \omega + \alpha\widehat{e}_{t-1} + \beta\eta_{t-1}. \tag{6}$$

To ensure positivity, we restrict $\omega > 0$ and $\alpha, \beta \geq 0$. When μ_t is constant, the resulting model for Y_t is an Integer-GARCH(1, 1) (IntGARCH) model [e.g., Ferland, Latour and Oraichi (2006)]. It is worth noting some properties of this model for the constant μ_t case. To ensure the stationarity of η_t, we further require that $\alpha + \beta < 1$. This sum determines the persistence of the process, with larger values of α leading to more adaptability. When this stationarity condition is satisfied, and η_t has reached its stationary distribution, the expectation of η_t given $\mathbf{X}$ is

$$E(\eta_t|\mathbf{X}) = \omega/(1 - \alpha - \beta).$$

To ensure $E(\eta_t|\mathbf{X}) = 1$ for the fitted model, we may parameterize $\omega = 1 - \alpha - \beta$. This constraint is simple enough to enforce for the model (6) and we do so. However, additional parameter constraints such as this may make numerical estimation intractable in more complicated models and they are not enforced by us in the models outlined below.

When μ_t is a nonstationary process, the conditional intensity

$$\lambda_t = \mu_t\eta_t$$

is also nonstationary. Since $E(\eta_t|\mathbf{X}) = 1$, we interpret η_t as the stationary multiplicative deviation, or inflation rate, between λ_t and μ_t. The λ_t process is mean reverting to the μ_t process. Let

$$\widehat{\varepsilon}_t = Y_t/\widehat{\lambda}_t$$

denote the multiplicative *standardized* residual process given an estimated CIIR process $\widehat{\eta}_t$. If a fitted model defined by (6) sufficiently explains the observed linear dependence in $\widehat{e}_t$, then an autocorrelation plot of $\widehat{\varepsilon}_t$ should be statistically insignificant for all lags ℓ. As a preview, the standardized residual autocorrelation plot for one such model fit in Section 4 is given in Figure 5(d). The serial correlation appears to have been adequately removed.

Next, we formulate three different nonlinear generalizations of (6) that may better characterize the serial dependence, and possibly lead to improved forecasts. The first is an exponential autoregressive model defined as

$$\eta_t = \alpha \widehat{e}_{t-1} + [\beta + \delta \exp(-\gamma \eta_{t-1}^2)]\eta_{t-1}, \tag{7}$$

in which $\alpha, \beta, \delta, \gamma > 0$. Exponential autoregressive models are attractive in application because of their threshold-like behavior. For large η_{t-1}, the functional coefficient for η_{t-1} is approximately β, and for small η_{t-1} it is approximately $\beta + \delta$. Additionally, the transition between these regimes remains smooth. As in Fokianos, Rahbek and Tjøstheim (2009), for $\alpha + \beta < 1$ one can verify the η_t process has a stationarity version when μ_t is constant.

We also consider a piecewise linear threshold model

$$\eta_t = \omega + \alpha \widehat{e}_{t-1} + \beta \eta_{t-1} + (\gamma \widehat{e}_{t-1} + \delta \eta_{t-1}) I_{\{\widehat{e}_{t-1} \notin (c_1, c_2)\}}, \tag{8}$$

in which I is an indicator variable and the threshold boundaries satisfy $0 < c_1 < 1 < c_2 < \infty$. To ensure positivity of η_t, we assume $\omega, \alpha, \beta > 0$, $(\alpha + \gamma) > 0$, and $(\beta + \delta) > 0$. Additionally, we take $\delta \leq 0$ and $\gamma \geq 0$, such that when $\widehat{e}_{t-1}$ is outside the range (c_1, c_2) the CIIR process η_t is more adaptive, that is, puts more weight on $\widehat{e}_{t-1}$ and less on η_{t-1}. When μ_t is constant, the η_t process has a stationary version under the restriction $\alpha + \beta + \gamma + \delta < 1$; see Woodard, Matteson and Henderson (2010). In practice, the threshold boundaries c_1 and c_2 are fixed during estimation, and may be adjusted as necessary after further exploratory analysis. We chose $c_1 = 1/1.15$ and $c_2 = 1.15$, that is, thresholds at 15% above and below 1.

Finally, we consider a model with regime switching at deterministic times, letting

$$\eta_t = (\omega_1 + \alpha_1 \widehat{e}_{t-1} + \beta_1 \eta_{t-1}) I_{\{t \in (t_1, t_2)\}} + (\omega_2 + \alpha_2 \widehat{e}_{t-1} + \beta_2 \eta_{t-1}) I_{\{t \notin (t_1, t_2)\}}. \tag{9}$$

This model is appropriate assuming the residual process has two distinct regimes for different periods of the day. For example, one regime could be for normal workday hours with the other regime being for the evening and early morning hours. No stationarity is possible for this model. A drawback of this model is that the process has jumps at t_1 and t_2. As was the case for c_1 and c_2 in (8), t_1 and t_2 are fixed during estimation. After exploratory analysis, we chose $t_1 = 10$ a.m. and $t_2 = 4$ p.m.

3.5. *Estimation algorithm.* The estimation procedure below begins with an iterative algorithm for estimating the factor model from Sections 3.1–3.3 through repeated use of the *gam* function from the *mgcv* library in *R*. Any serial dependence is ignored during estimation of μ_t for simplicity. Given estimates for the factor model $\widehat{\mu}_t$, conditional maximum likelihood is used to estimate the conditional intensity λ_t via one of the time series models given in (6)–(9) for the CIIR process η_t.

1. Initialization:
   - (a) Fix K and $\mathbf{H}$.
   - (b) Choose some $c \in (0, 1)$ and define $\mathbf{Y}_c = (y_{ij} \vee c)$.
   - (c) Apply a singular value decomposition (SVD) to find $\log(\mathbf{Y}_c) = \mathbf{U}_0\mathbf{D}_0\mathbf{V}_0^{\mathrm{T}}$.
     - (i) Let $\mathbf{U}_0^{(1:K)}$ denote the first K columns of the left singular matrix $\mathbf{U}_0$.
     - (ii) Let $\mathbf{V}_0^{(1:K)}$ denote the first K columns of the right singular matrix $\mathbf{V}_0$.
     - (iii) Let $\mathbf{D}_0^{(1:K)}$ denote the upper-left $K \times K$ sub-matrix of $\mathbf{D}_0$, the diagonal matrix of singular values.
   - (d) Assign $\mathbf{L}_0 = \mathbf{U}_0^{(1:K)}\mathbf{D}_0^{(1:K)}$ and $\mathbf{F}_0 = \mathbf{V}_0^{(1:K)}$.
     No smoothing is performed and the constraints $\mathbf{H}$ are omitted in initialization.
2. Update:
   - (a) Fit the Poisson GAM model described in Section 3.3 with $\mathbf{F} = \mathbf{F}_n$ and $\mathbf{H}$ as fixed covariates.
     - Assign $\mathbf{B}_{n^*}$ as the estimated parameter values from this fit and let $\mathbf{L}_{n^*} = \mathbf{H}\mathbf{B}_{n^*}$.
   - (b) Fit the Poisson GAM model described in Section 3.3 with $\mathbf{L} = \mathbf{L}_{n^*}$ as a fixed covariate.
     - Assign $\mathbf{F}_{n^*}$ as the estimated parameter values from this fit.
   - (c) Apply an SVD to find $\mathbf{B}_{n^*}\mathbf{F}_{n^*}^{\mathrm{T}} = \mathbf{U}_{n+1}\mathbf{D}_{n+1}\mathbf{V}_{n+1}^{\mathrm{T}}$.
     - (i) Assign $\mathbf{B}_{n+1} = \mathbf{U}_{n+1}^{(1:K)}\mathbf{D}_{n+1}^{(1:K)}$.
     - (ii) Assign $\mathbf{F}_{n+1} = \mathbf{V}_{n+1}^{(1:K)}$.
     - (iii) Assign $\mathbf{L}_{n+1} = \mathbf{H}\mathbf{B}_{n+1}$.
   - (d) Let $\log \mathbf{M}_{n+1} = \mathbf{L}_{n+1}\mathbf{F}_{n+1}^{\mathrm{T}}$.
3. Repeat the *update* steps recursively until convergence.

Convergence is reached when the relative change in $\mathbf{M}$ is sufficiently small. After convergence we can recover $\log \widehat{\mu_t}$ from the rows of the final estimate of $\log \mathbf{M}$. These values are then treated as fixed constants during estimation of η_t. We use conditional maximum likelihood to estimate the parameters $(\omega, \alpha, \beta, \ldots)$ associated with a time series model for η_t. The recursion defined by (6)–(9) requires initialization by choosing a value for η_1; the estimates are conditional on the chosen initialization.

We may always specify the joint distribution $P_{\mathbf{Y}}$ of the observations $\mathbf{Y}$ as an iterated product of successive conditional distributions P_{Y_t} for Y_t given $(Y_{t-1}, \dots, Y_1)$ as

$$P_Y(y_T, y_{T-1}, \dots, y_2, y_1) = P_{Y_1}(y_1) \prod_{t=2}^{T} P_{Y_t}(y_t | y_{t-1}, \dots, y_1).$$

We follow the standard convention of fixing $P_{Y_1}(y_1) = 1$ in estimation. For large sample sizes the practical impact of this decision is negligible. We may therefore write the log likelihood function as the sum of iterated conditional log likelihood functions. The conditional distribution for the observations is assumed to be Poisson with mean $\lambda_t = \mu_t \eta_t$.

For uninterrupted observations over periods $1, \dots, T$, we define the log likelihood function as

$$\begin{aligned} \ell(\omega, \alpha, \beta, \dots | \widehat{M}, Y, \eta_1) &= \sum_{t=2}^{T} \ell_t(\omega, \alpha, \beta, \dots | y_t, y_{t-1}, \widehat{\mu}_t, \widehat{\mu}_{t-1}, \eta_{t-1}) \\ &= \sum_{t=2}^{T} (y_t \log \lambda_t - \lambda_t - \log y_t!) \\ &= \sum_{t=2}^{T} \bigl(y_t \log(\widehat{\mu}_t \eta_t) - \widehat{\mu}_t \eta_t - \log y_t!\bigr). \end{aligned} \tag{10}$$

This recursion requires an initial value for η_1. For simplicity, we use its expected value, $\eta_1 = 1$. When there are gaps in the observation record, equation (10) is calculated over every contiguous block of observations. This requires reinitialization of $\eta_t = 1$ at the beginning of each block. The log likelihood for the blocks are then added together to form the entire log likelihood. The maximum likelihood estimate is the argmax of this quantity, subject to the constraints given in Section 3.4. Finally, η_t is estimated by the respective recursion given by equations (6)–(9) with parameters replaced by their estimates, again with reinitialization of $\eta_t = 1$ at the beginning of each contiguous block of observations. Blocks were large enough in our application that the effect of reinitialization was negligible.

4. Empirical analysis.

Using the data described in Section 2, we perform the following analysis: (a) we define various statistical goodness-of-fit metrics suitable for the proposed models; based on in-sample performance, these metrics are used to determine the number of factors K for use in the dynamic factor models. (b) We compare the out-of-sample forecast performance for the factor model in (3), the factor model with constraints in (4), and the factor model with constraints and smoothing splines in (5). These comparisons help ascertain the improvement from each refinement and validate the proposed selection methods for K. (c) For the latter factor model, we compare the out-of-sample forecast performance with the addition of the CIIR process, via use of the various time series models defined in Section 3.4. (d) We quantify the practical impact of these successive statistical improvements with a queueing application constructed to approximate ambulance operations.

4.1. *Interpreting the fitted model.* The mean number of calls was approximately 24 per hour for 2007 and 2008, and no increasing or decreasing linear trend in time was detected during this period. We partition the observations by year into two data sets referred to as *2007* and *2008*, respectively. Each year is first regarded as a *training set*, and each model is fit individually to each year. The opposite year is subsequently used as a *test set* to evaluate the out-of-sample performance of each fitted model. To account for missing days, we reinitialize the CIIR process η_t in the first period following each day of missing data. This was necessary at most five times per year including the first day of the year.

We found the factor model fit with constraints, smoothing splines, and $K = 4$ factors to be the most appropriate of the factor models considered. The estimated factors $\mathbf{f}_k$ for 2008 are shown in Figure 2(a). Each of the four factors varies smoothly over the hours of the day via use of smoothing splines. The first factor $\mathbf{f}_1$ is strictly positive and the least variable. It appears to capture the mean diurnal pattern. The factor $\mathbf{f}_2$ appears to isolate the dominant relative differences between weekdays and weekend days. The defining feature of $\mathbf{f}_3$ and $\mathbf{f}_4$ is the large increase late in the day, corresponding closely to the relative increase observed on Friday and Saturday evenings. However, $\mathbf{f}_3$ decreases in the morning, while $\mathbf{f}_4$ increases in the morning and decreases in the late afternoon. As K increases, additional factors become increasingly more variable over the hours of the day. Too many factors result in overfitting the model, as the extra factors capture noise.

The corresponding daily factor loadings $\mathbf{L}$ for the first four weeks of 2008 are shown in Figure 2(b). The loadings $(\mathbf{L}_1 - 14.5)$ are shown to simplify comparisons. The much higher loadings on $\mathbf{f}_1$ confirm its interpretation as capturing the mean. The peaks on Fridays coincide with Friday having the highest average number of calls, as seen in Figure 1. Weekdays get a positive loading on $\mathbf{f}_2$, while weekend days get negative loading. Loadings on $\mathbf{f}_3$ are lowest on Sundays and Mondays and loadings on $\mathbf{f}_4$ are largest on Fridays and Saturdays. As K increases, the loadings on additional factors become increasingly close to zero. This partially mitigates the overfitting described above. Factors with loadings close to zero have less impact on the fitted values $\widehat{\mu}_t$. Nevertheless, they can still reduce out-of-sample forecast performance.

The daily factor loadings for all of 2008 are shown in Figure 2(c). The relative magnitude of each loading vector with respect to day-of-week is constant. This results from use of the constraint matrix $\mathbf{H}^{(1)}$ in (4). As the loadings vary over the days of the week, they also vary smoothly over the course of the year, via use of the constraint matrix $\mathbf{H}^{(2)}$ and the use of cyclic smoothing splines in estimation of $\mathbf{B}^{(2)}$ in (4). The loadings on $\mathbf{f}_1$ show how the expected number of calls per day varies

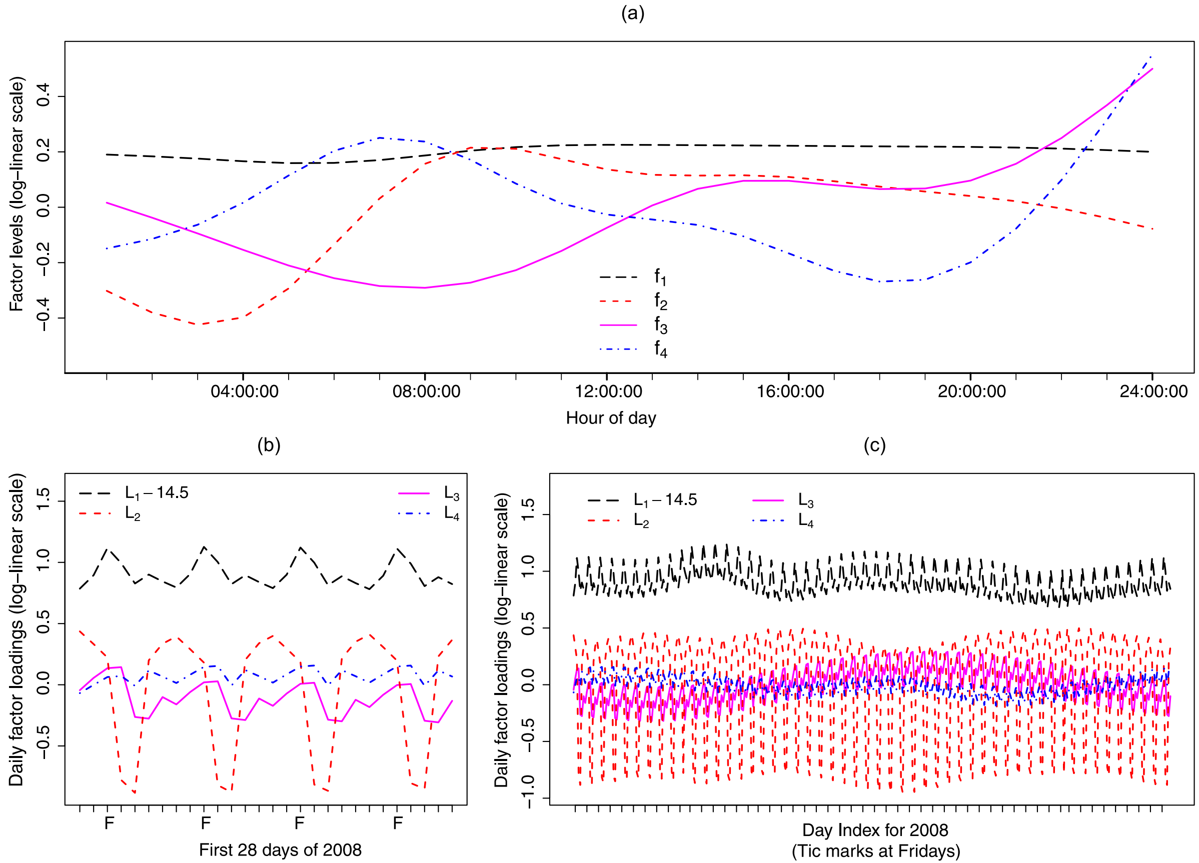

FIG. 2. 2008 *fitted* (a) *factor levels* $\mathbf{f}_k$ (*log-linear scale*) *and* [(b) *and* (c)] *corresponding factor loadings* $\mathbf{L}_{k\cdot}$ (*log-linear scale*) *for a factor model fit with constraints*, *smoothing splines and* $K = 4$ *factors*. ($\mathbf{L}_{1\cdot} - 14.5$) *is shown for easier comparison*.

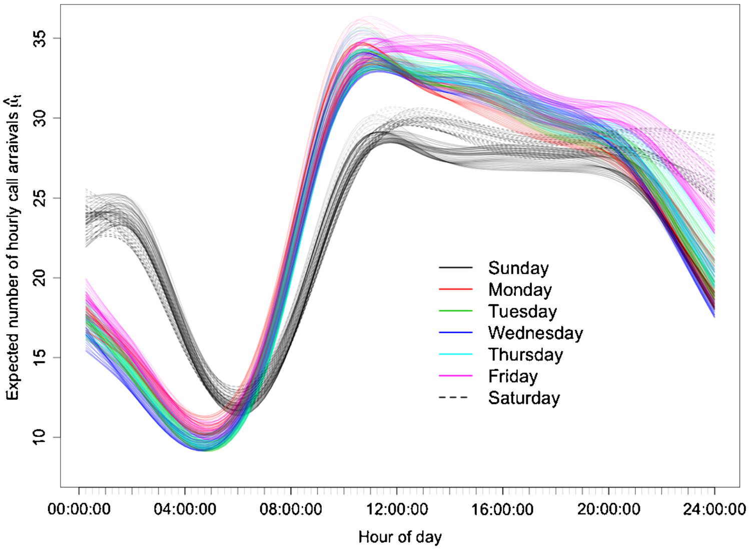

FIG. 3. *The estimated intensity process* $\hat{\boldsymbol{\mu}}_i$, *for every day in* 2008, *for a factor model fit with constraints, smoothing splines and* $K = 4$ *factors, colored by day-of-week, and shaded light to dark by week-of-year.*

over the year. The week to week variability in the other loadings influences how the days of the week change relative to each other over the year. Figure 3 shows

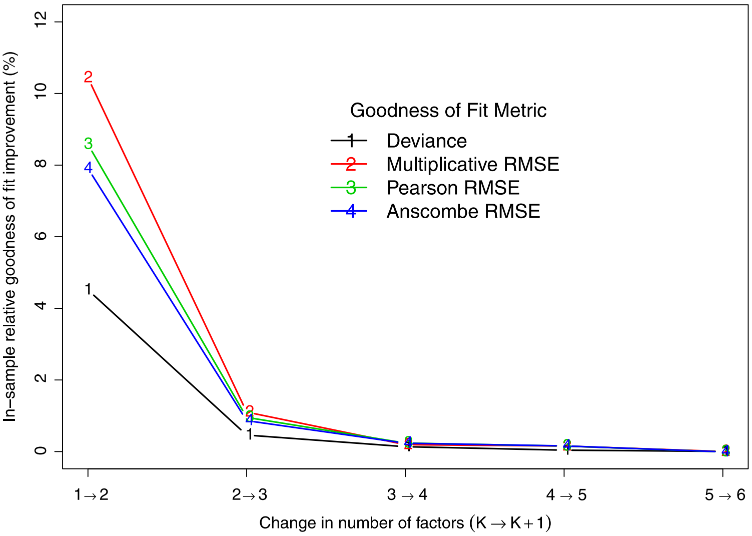

FIG. 4. 2007 *percentage in-sample relative goodness-of-fit improvement by addition of one factor* $(K \to K + 1)$ *for a factor model fit with constraints and smoothing splines.*

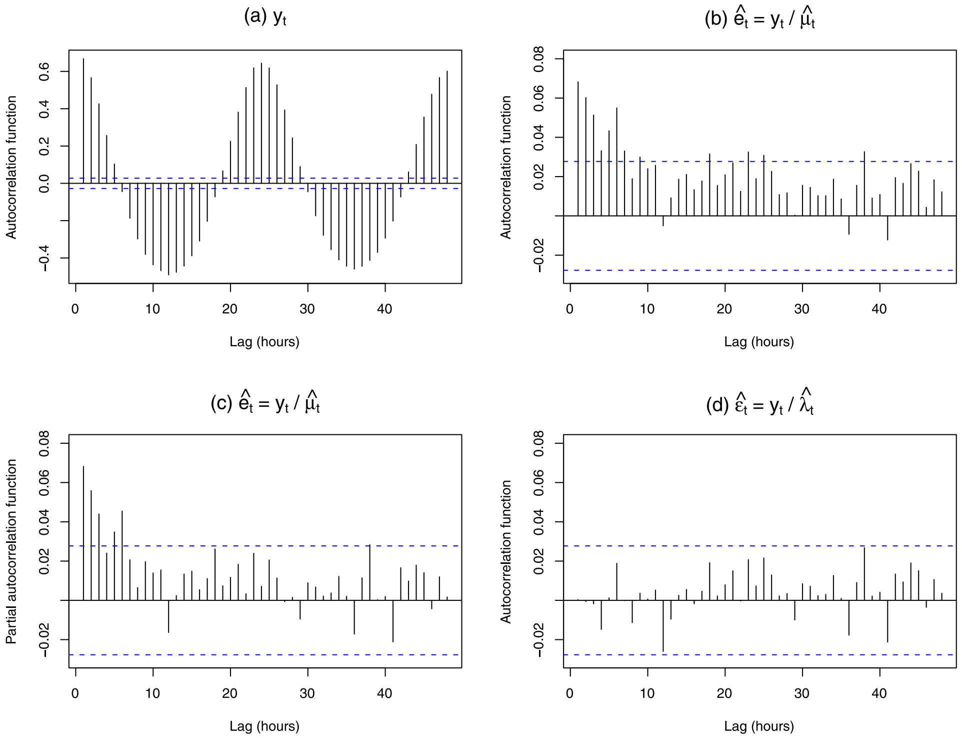

FIG. 5. (a) *Sample autocorrelation function for hourly call arrival counts* y_t. *Residual* $\widehat{e}_t = y_t/\widehat{\mu}_t$ (b) *autocorrelation and* (c) *partial autocorrelation functions for fitted factor model* $\widehat{\mu}_t$ *with* $k = 4$ *factors using factor and loading constraints and smoothing splines.* (d) *Standardized residual* $\widehat{\varepsilon}_t = y_t/\widehat{\lambda}_t = y_t/(\widehat{\mu}_t\widehat{\eta}_t)$ *autocorrelation function for fitted factor model with fitted* IntGARCH(1, 1) *model for* η_t. *Dashed lines give approximate* 95% *confidence levels.*

the estimated intensity process $\widehat{\boldsymbol{\mu}}_i$ for every day in 2008, shaded by day-of-week. The curves vary smoothly over the hours of the day. The fit for each day of the week keeps the same relative shape, but it varies smoothly over the weeks of the year.

Section 3.4 described incorporating time series models to improve the short-term forecasts of a factor model. The models capture the observed serial dependence in the multiplicative residuals from a fitted factor model; see Figure 5. Parameter estimates and approximate standard errors for the IntGARCH model are given in Supplemental material (Table 1). A fitted factor model $\widehat{\mu}_t$ using constraints, smoothing splines and $K = 4$, as well as the factor model including a fitted IntGARCH(1, 1) model $\widehat{\lambda}_t$, are also shown in Figure 6(a), with the observed call arrivals per hour for Weeks 8 and 9 of 2007. The $\widehat{\lambda}_t$ process is mean reverting about the $\widehat{\mu}_t$ process. They are typically close to each other, but when they differ by a larger amount, they tend to differ for several hours at a time. The corresponding fitted CIIR process $\widehat{\eta}_t$ is shown in Figure 6(b). This clearly illustrates the depen-

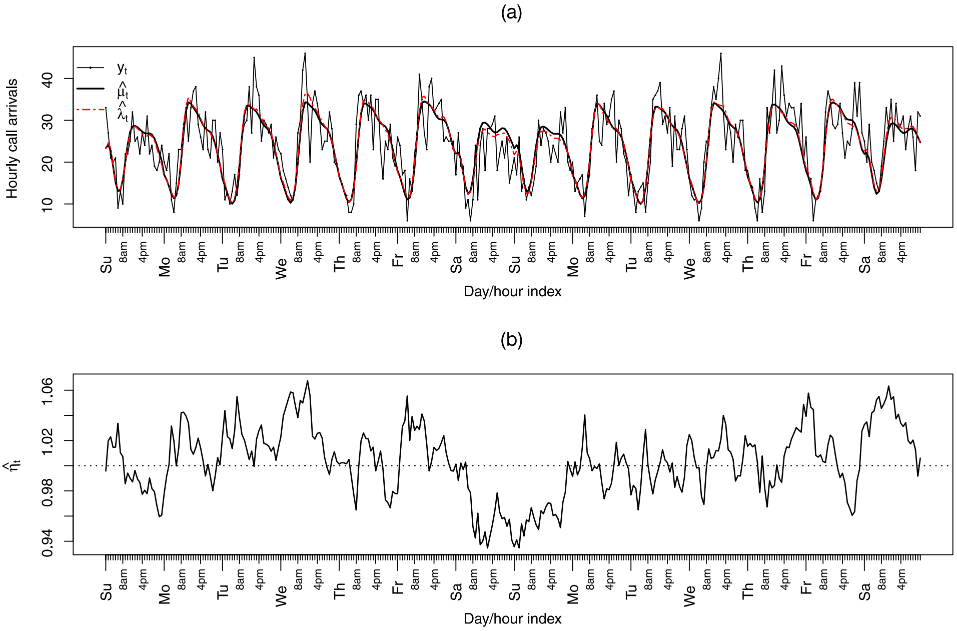

FIG. 6. *Weeks* 8 *and* 9 *of* 2007: (a) *observed call arrivals per hour* y_t, *fitted* $K = 4$ *dynamic factor model* $\widehat{\mu}_t$ *using constraints and smoothing splines, and factor model* $\widehat{\lambda}_t$ *including fitted* IntGARCH(1, 1); (b) *the fitted conditional intensity inflation process* $\widehat{\eta}_t$ *from the* IntGARCH(1, 1) *model.*

dence and persistence exhibited in Figure 6(a). The CIIR process ranges between ±6% during this period. With a mean of 24 calls per hour, this range corresponds to $\widehat{\lambda}_t$ varying about $\widehat{\mu}_t$ by about ±1.5 expected calls per hour.

4.2. *Goodness of fit and model selection.* To evaluate the fitted values and forecasts of the proposed models, three types of residuals are computed: multiplicative, Pearson and Anscombe. Their respective formulas for the Poisson distribution are given by

$$\widehat{r}_{M,t} = \frac{y_t}{\widehat{\lambda}_t} - 1, \qquad \widehat{r}_{P,t} = \frac{y_t - \widehat{\lambda}_t}{\sqrt{\widehat{\lambda}_t}}, \qquad \widehat{r}_{A,t} = \frac{(3/2)(y_t^{2/3} - \widehat{\lambda}_t^{2/3})}{\widehat{\lambda}_t^{1/6}}.$$

We refer to the root mean square error (RMSE) of each metric as RMSME, RMSPE and RMSAE, respectively. The multiplicative residual is defined as before and is a natural choice given the definition for the CIIR. Since the variance of a Poisson random variable is equal to its mean, the Pearson residual is quite standard. However, the Pearson residual can be quite skewed for the Poisson distribution [cf. McCullagh and Nelder (1989), Section 2.4]. The Anscombe residual is derived as a transformation that makes the distribution of the residuals as close to Gaussian as possible while suitably scaling to stabilize the variance. See Pierce and Schafer (1986) for further discussion of residuals for generalized linear models. While the three methods always yielded the same conclusion, we found use of the Anscombe residuals gave a more robust assessment of model accuracy and simplified paired comparisons between the residuals of competing models.

The three RMSE metrics were used for both in- and out-of-sample model comparisons. For in-sample comparisons of the factor models, we also computed the *deviance* of each fitted model $\widehat{\mu}_t$. As a goodness-of-fit metric, deviance is derived from the logarithm of a ratio of likelihoods. For a log likelihood function $\ell(\boldsymbol{\mu}|\mathbf{Y})$, it is defined as

$$-2\{\ell(\boldsymbol{\mu} = \widehat{\boldsymbol{\mu}}|\mathbf{Y}) - \ell(\boldsymbol{\mu} = \mathbf{Y}|\mathbf{Y})\},$$

in general. For a fitted factor model, ignoring serial dependence, the deviance corresponding to a Poisson distribution is

$$2\sum_{t=1}^{n}\{y_t \log(y_t/\widehat{\mu}_t) - (y_t - \widehat{\mu}_t)\},$$

in which the first term is zero if $y_t = 0$.

We compare the fitted models' relative reduction in deviance and RMSE as we increase the number of factors K. Figure 4 shows these results for factor models fit to 2007 data with constraints and smoothing splines. The results for other models and for 2008 were very similar. This plot may be interpreted similarly to a *scree* plot in PCA by identifying the point at which performance tapers off and the marginal improvement from additional factors is negligible. Under each scenario we consistently selected $K = 4$ factors through this graphical criterion. To further justify this as a factor selection strategy, we also consider the impact the number of factors K has on out-of-sample performance for each of the proposed models below. This approach is straightforward, but it does not fully account for the uncertainty on the number of factors. Bayesian estimation would require specialized computation, but it may improve model assessment [see, e.g., Lopes and West (2004)].

4.3. *Out-of-sample forecast performance.* Out-of-sample comparisons were made by fitting models to the 2007 training set and forecasting on the 2008 test set, and vice versa. To make predictions comparable from one year to the next, we align corresponding calendar weeks of the year, not days of the year. This ensures that estimates for Sundays are appropriately compared to Sundays, etc.

The first model considered was the *simple prediction* (SP) method. This simple moving average involving four observations was defined in the Introduction. Next, the forecasts of various factor models (FM) were considered. For $K = 1, \dots, 6$, we evaluated the forecasts from the FM in (3), the FM with constraints in (4), and the FM with constraints and smoothing splines in (5). Finally, for the latter FM, with $K = 4$, we calculate the implied fit from the training set with the inclusion of the CIIR process via the various time series models defined in Section 3.4. We compute the forecast RMSE of each model for the three residual types, for both years.

The forecast results are shown in Table 1. The basic FMs did slightly worse than the SP both years. With only one year of observations, these FMs tend to overfit the training set data, even with a small number of factors. The FMs with constraints give a very significant improvement over the previous models. The forecast RMSE is lowest at $K = 4$ for the 2007 test set, and at $K = 3$ for the 2008 test set. There was also a very large decrease between $K = 1$ and $K = 2$. The FMs with constraints and smoothing splines offered an additional improvement. The forecast RMSE is lowest at $K = 4$ for both test sets. With the addition of the IntGARCH model for the CIIR process to this model, the RMSE improved again. Application of the nonlinear time series models instead offered only a slight improvement over the IntGARCH model.

With only one year of training data, each FM begins to overfit with $K = 5$ factors. Results were largely consistent regardless of the residual used, but the Anscombe residuals were the least skewed and allowed the simplest pairwise comparisons. Although the FMs with constraints had superior in-sample performance, the use of smoothing splines reduced the tendency to over-fit and resulted in improved forecast performance. The CIIR process offered improvements in fit over FMs alone.

We also fit each of the nonlinear time series models discussed in Section 3.4 using a FM with $K = 4$. The regime switching model had the best performance. It had the lowest RMSE for both test sets. The exponential autoregressive and the piecewise linear threshold models performed similarly to the IntGARCH model for both test sets. Although the nonlinear models consistently performed better in-sample, their out-of-sample performance was similar to the IntGARCH model.

4.4. *Queueing model simulation to approximate ambulance operations.* To comprehensively improve ambulance operations, it would be advantageous to simultaneously model the service duration of dispatched ambulances in addition to the demand for ambulance service. Unfortunately, such information was not available. We are currently working with Toronto EMS to use our improved estimates of call arrival rates to improve staffing in their dispatch call center. Extending our approach to a spatial-temporal forecasting model will likely be used to help determine when *and* where to deploy ambulances.

We present a simulation study that uses a simple queueing system to quantify the impact that improved forecasts have on staffing decisions and relative operating costs, for the Toronto data. The queueing model is a simplification of ambulance operations that ignores the spatial component. Similar queueing models have been used frequently in EMS modeling [see Swersey (1994), page 173]. This goodness-of-fit measure facilitates model comparisons and a similar approach may be useful in other contexts.

TABLE 1

Root mean square multiplicative, Pearson, and Anscombe errors for fitting model to 2007 *and forecasting* 2008, *and vice versa*

			2007 model, 2008 residuals			**2008 model, 2007 residuals**		
Model	**Constraint**	**Smoothing**	**RMSME**	**RMSPE**	**RMSAE**	**RMSME**	**RMSPE**	**RMSAE**
Simple prediction	NA	NA	0.2696	1.1955	1.1849	0.2661	1.1902	1.1925
Factor model, $K = 1$	No	No	0.2722	1.2369	1.2237	0.2657	1.2183	1.2263
Factor model, $K = 2$	No	No	0.2721	1.2357	1.2225	0.2661	1.2197	1.2277
Factor model, $K = 3$	No	No	0.2727	1.2374	1.2239	0.2659	1.2182	1.2262
Factor model, $K = 4$	No	No	0.2729	1.2383	1.2249	0.2666	1.2206	1.2283
Factor model, $K = 5$	No	No	0.2732	1.2395	1.2260	0.2670	1.2220	1.2294
Factor model, $K = 6$	No	No	0.2733	1.2401	1.2270	0.2668	1.2217	1.2294
Factor model, $K = 1$	Yes	No	0.2638	1.1863	1.1756	0.2575	1.1633	1.1721
Factor model, $K = 2$	Yes	No	0.2402	1.0938	1.0888	0.2333	1.0722	1.0875
Factor model, $K = 3$	Yes	No	0.2392	1.0877	1.0829	0.2324	1.0688	1.0848
Factor model, $K = 4$	Yes	No	0.2413	1.0945	1.0889	0.2347	1.0761	1.0912
Factor model, $K = 5$	Yes	No	0.2425	1.0994	1.0933	0.2363	1.0817	1.0961
Factor model, $K = 6$	Yes	No	0.2436	1.1051	1.0988	0.2377	1.0858	1.0999
Factor model, $K = 1$	Yes	Yes	0.2633	1.1837	1.1731	0.2573	1.1615	1.1703
Factor model, $K = 2$	Yes	Yes	0.2371	1.0844	1.0805	0.2310	1.0643	1.0803
Factor model, $K = 3$	Yes	Yes	0.2347	1.0744	1.0710	0.2289	1.0561	1.0728
Factor model, $K = 4$	Yes	Yes	0.2344	1.0730	1.0696	0.2288	1.0549	1.0715
Factor model, $K = 5$	Yes	Yes	0.2347	1.0740	1.0706	0.2289	1.0549	1.0714
Factor model, $K = 6$	Yes	Yes	0.2347	1.0739	1.0705	0.2289	1.0551	1.0716
Time series and FM, $K = 4$	Yes	Yes	–	–	–	–	–	–
IntGARCH	–	–	0.2308	1.0571	1.0570	0.2274	1.0442	1.0580
IntExpGARCH	–	–	0.2308	1.0570	1.0569	0.2274	1.0441	1.0579
IntThreshGARCH	–	–	0.2308	1.0571	1.0570	0.2275	1.0443	1.0580
IntRsGARCH	–	–	0.2299	1.0540	1.0554	0.2274	1.0433	1.0565

A Yes in the constraints column implies that the factor model was fit using the constraints outlined in Section 3.2. A Yes in the smoothing column indicates that the model was fit using smoothing splines as described in Section 3.3.

We use the terminology employed in the call center and queueing theory literature throughout the section; for our application, servers are a proxy for ambulances, callers or customers are those requiring EMS, and a server completing service is equated to an ambulance completing transport of a person to a hospital, etc. As before, let y_t denote the observed number of call arrivals during hour t. Our experiment examines the behavior of a simple $M/M/s$ queueing system. The arrival rate in time period t is λ_t. During this period, let s_t denote the number of servers at hand. For simplicity, we assume that the service rate ν for each server is the same, and constant over time. Furthermore, intra-hour arrivals occur according to a Poisson process with rate λ_t, and service times of callers are independent and exponentially distributed with rate ν.

As in Section 4.3, models are calibrated on one year of observations and forecasts for λ_t are made for the other year. Each model's forecasts $\widehat{\lambda}_t$ are then used to determine corresponding staffing levels $\widehat{s}_t$ for the system.

To facilitate comparisons of short-term forecasts, we assume that the number of servers can be changed instantaneously at the beginning of each period. In practice, it is possible to adjust the number of ambulances in real time, but not to the degree that we assume here.

Each call has an associated arrival time and service time. When a call arrives, the caller goes immediately into service if a server is available, otherwise it is added to the end of the queue. A common goal in EMS is to ensure that a certain proportion of calls are reached by an ambulance within a prespecified amount of time. We approximate this goal by instead aiming to answer a proportion, θ, of calls immediately; this is a standard approximation in queueing applications in many areas including EMS [Kolesar and Green (1998)]. For each call arrival, we note whether or not the caller was served immediately. As servers complete service, they immediately begin serving the first caller waiting in the queue, otherwise they await new arrivals if the queue is currently empty. One simulation replication of the queueing system simulates all calls in the test year.

To implement the queueing system simulation, it is first necessary to simulate arrival and service times for each caller in the forecast period. We use the observed number of calls for each hour y_t as the number of arrivals to the system in period t. Since arrivals to the system are assumed to follow a Poisson process, we determine the y_t call arrival times using the well-known result that, conditional on the total number of arrivals in the period $[t, t+1]$, the arrival times have the same distribution as the order statistics of y_t independent Uniform$(t, t+1)$ random variables. We exploit this relationship to generate the intra-hour arrival times given the observed arrival volume y_t. The service times for each call are generated independently with an Exponential(ν) distribution.

The final input is the initial state of the queue within the system. We generate an initial number of callers in the queue as Poisson(y_1), then independently generate corresponding Exponential(ν) residual service times for each of these callers. This initialization is motivated through an infinite-server model; see, for example, Kolesar and Green (1998). Whenever there is a missing day, in either the test set or corresponding training set period, we similarly reinitialize the state of the queue but with y_1 replaced by the number of calls observed in the first period following the missing period. These initializations are common across the different forecasting methods to allow direct comparisons.

To evaluate forecast performance, we define a cost function and an appropriate method for determining server levels from arrival rate estimates. Let n_t denote the number of callers served immediately in period t. The hourly cost function is given by

$$C(n_t, y_t, s_t) = \text{Pen}(n_t, y_t) + s_t,$$

in which

$$\text{Pen}(n_t, y_t) = \begin{cases} 0, & \text{if } n_t \geq \theta y_t, \\ q(y_t - n_t), & \text{otherwise,} \end{cases}$$

$\theta \in (0, 1)$ is the targeted proportion of calls served immediately, and $q \geq 0$ is the cost of not immediately serving a customer, *relative* to the cost of staffing one server for one hour. The total cost, with respect to the hourly server cost, for the entire forecast period is

$$C = \sum_t C(n_t, y_t, s_t) = \sum_t \text{Pen}(n_t, y_t) + \sum_t s_t.$$

This approach, where penalties for poor service are balanced against staffing costs, is frequently used; see, e.g., Andrews and Parsons (1993), Harrison, Zeevi and Shum (2005).

At time $t - 1$, the number of call arrivals and the number served immediately are random variables, denoted as Y_t and N_t, respectively. A natural objective is to choose staffing levels that minimize the hourly expected cost as

$$\widehat{s}_t = \operatorname*{argmin}_{s_t \in \mathbb{N}} E\{C(N_t, Y_t, s_t) | \mathcal{F}_{t-1}, \mathbf{X}\}, \tag{11}$$

in which Y_t is assumed to have a Poisson distribution with mean equal to the arrival rate forecast $\widehat{\lambda}_t$. The staffing levels are then a function of arrival rate forecasts, $\widehat{s}_t(\widehat{\lambda}_t)$. We approximate this expectation numerically by randomly generating J independent realizations as $Y_{t,j} \sim \text{Poisson}(\widehat{\lambda}_t)$. Then, for each $Y_{t,j}$ we simulate one independent realization of N_t. For a fixed value of s_t the expectation is approximated by $J^{-1} \sum_{t=1}^{J} \{\text{Pen}(N_{t,j}, Y_{t,j}) + s_t\}$. We found that $J = 25{,}000$ provided adequate accuracy.

Independent realizations of $N_t | Y_t$ require running the queueing system forward one hour, but this is very computationally intensive. To approximate $N_t | Y_t$, we use a Binomial distribution. Let $N_{t,j} | Y_{t,j} \sim \text{Binomial}\{Y_{t,j}, g(\widehat{\lambda}_t, s_t, \nu)\}$. The function g gives the *steady state* probability that a customer is served immediately for a queueing system with a *constant* arrival rate, server level and service rate, $\widehat{\lambda}_t$, s_t

and ν, respectively. Derivation of this function is available in any standard text on queueing theory [e.g., Gross and Harris (1998), Chapter 2].

Let p_i denote the long run proportion of time such a system contains i customers and let $\rho = \lambda/(\nu s)$. Then

$$g(\lambda, s, \nu) = \begin{cases} 1 - \dfrac{\lambda^s p_0}{s!\nu^s(1-\rho)}, & \text{if } \rho < 1, \\ 0, & \text{if } \rho \geq 1, \end{cases}$$

$$\text{in which } p_0^{-1} = \sum_{u=0}^{c-1} \frac{r^u}{u!} + \frac{r^c}{c!(1-\rho)} \text{ for } \rho < 1.$$

When $\rho \geq 1$, the arrival rate is faster than the net service rate, and the system is unstable; the long run probability that a customer is served immediately is zero. The binomial approximation greatly reduces the computational costs and provides reasonable results, though it tends to underestimate the true variability of $N_t|Y_t$ due to the positive correlation in successive caller delays.

A final deliberation is needed on the removal of servers when $\widehat{s_t}$ decreases. In our implementation, idle servers were removed first, and, if necessary, busy servers were dropped in ascending order with respect to remaining service time. We also considered random selection of servers to be dropped. Doing so produced highly variable results, and is under further study. To further simplify the implementation, if it was necessary to drop a busy server, it was simply discarded, along with any remaining service time for that caller. The effect of this simplification depends on the service rate ν; our results did not appear to be sensitive to this simplification.

Simulation of the queueing system is now rather straightforward. On each iteration i, we note whether each caller was served immediately or not. Forecast performance is assessed by examining the total cost $C^{(i)} = \sum_t C(n_t^{(i)}, y_t, \hat{s}_t)$ over the test period. For both years, we performed 100 simulations over the test year for each forecast method. To demonstrate the robustness of this methodology, we performed the experiment for several different values of the queuing system's parameters. Specifically, all combinations of $q \in \{2, 5, 10\}$, $\nu \in \{1, \frac{2}{3}\}$, and $\theta \in \{0.8, 0.9\}$ were considered, after consultation with EMS experts.

Results for the mean hourly cost over the 100 simulations for each forecasting method, for each test year, are summarized in Figure 7. We see that the mean hourly cost is lowest for the FM w/ IntGARCH, followed by the FM only, and finally by SP. All pairwise differences in mean were highly significant; the smallest t-ratio was 80. In fact, this ordering in performance held for almost every iteration of the queueing system, not just on average.

The mean percentage of callers served immediately can be found in Figure 8. The total number of server hours $\sum \hat{s}_t$ used was also recorded for each model for each set of parameter values. A table containing the values of all these quantities can be found in the online supplemental material. Both mean percentage served immediately and mean hourly cost increase with q. For each test year, for each

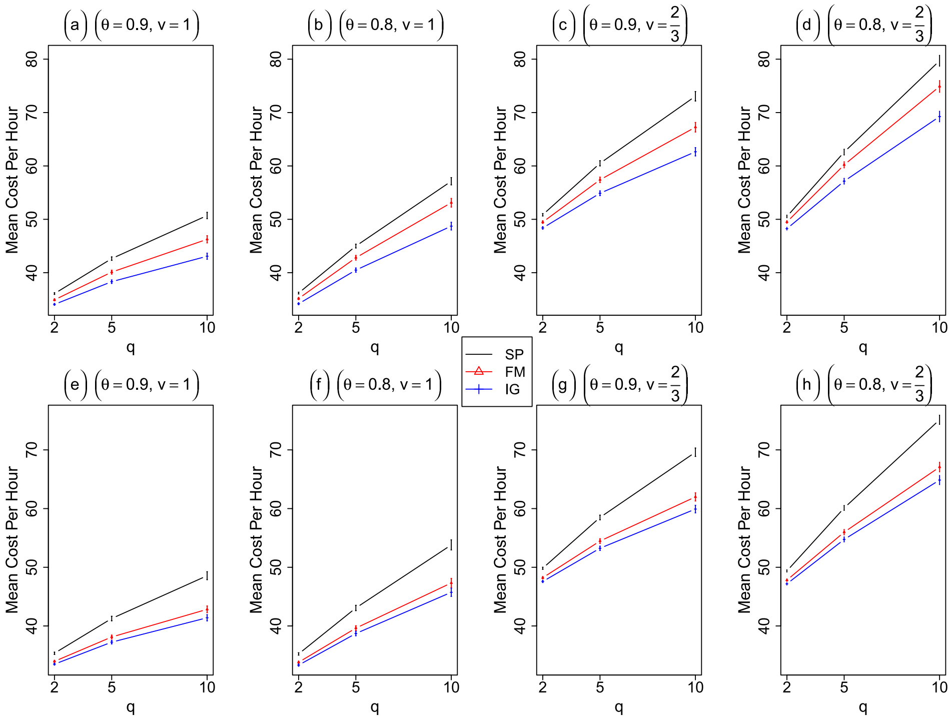

FIG. 7. *Mean total per period cost over* 100 *simulations for different forecasting methods and different values of q, ν and θ. Plots* (a)–(d) *use the* 2008 *test set and plots* (e)–(h) *use* 2007 *as the test set. The vertical lines represent* ±1 *standard deviation.*

level of (q, ν, θ), $\sum_t \hat{s}_t$ differed by between one and three thousand server-hours, for the different models.

5. Conclusions. Our analysis was motivated by a data set provided by Toronto EMS. The proposed forecasting method allows parsimonious modeling of the dependent and nonstationary count-valued EMS call arrival process. Our method is straightforward to implement and demonstrates substantial improvements in forecast performance relative to simpler forecasting methods. We measured the impact of our successive refinements to the model, showing the merit of factor model estimation with covariates and smoothing splines. The factor model was able to capture the nonstationary behavior exhibited in call arrivals. Introduction of the CIIR process allowed adaptive forecasts of deviations from this diurnal pattern.

Assessing the impact that different arrival rate forecasts can have on call centers and related applications has received very little attention in the literature. Our data-based simulation approach is straightforward to implement, and was able to clearly distinguish the effectiveness of each forecasting method. The simulation

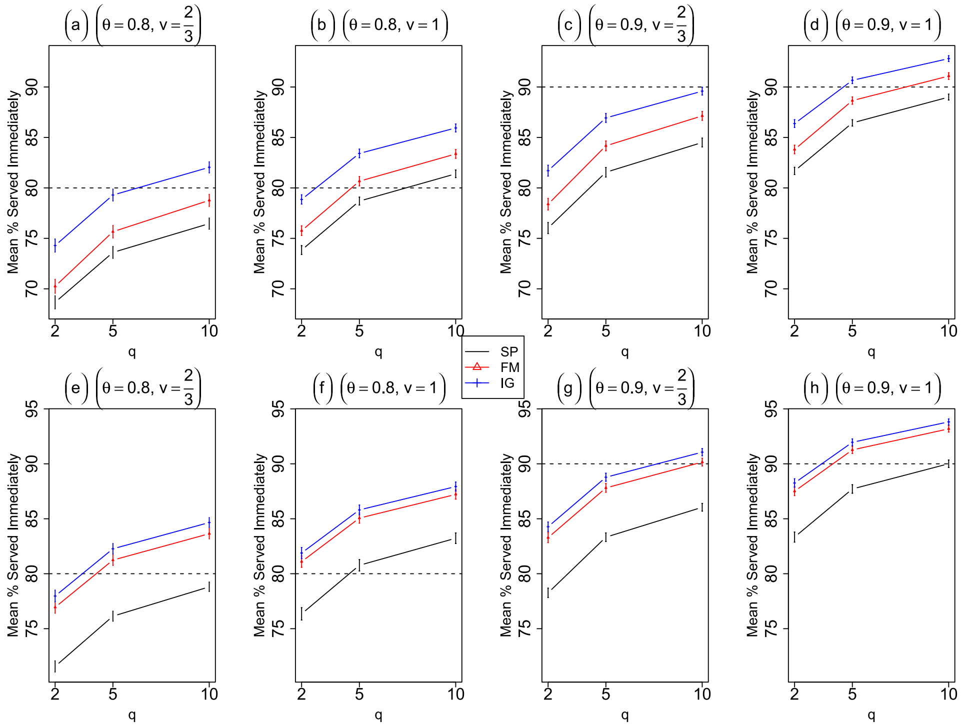

FIG. 8. *Mean percentage served immediately for the entire test set over* 100 *simulations for different forecasting methods and different values of* q, v *and* θ. *Plots* (a)–(d) *use the* 2008 *test set and plots* (e)–(h) *use* 2007 *as the test set. The vertical lines represent* ±1 *standard deviation.*

results coincide with the out-of-sample RMSE analysis in Section 4.3 and provide a practical measure of forecast performance. Relative operating cost is a natural metric for measuring call arrival rate forecasts, and our implementation may easily be extended to many customized cost functions and a wide variety of applications.

Ultimately, we seek to strengthen emergency medical service by improving upon relevant statistical methodology. Future work will consider inclusion of additional covariates and study of other nonlinear time series models. Bayesian methods which directly model count-valued observations have desirable properties for inference and many applications, and are under study. Spatial and spatial–temporal analysis of call arrivals will also offer new benefits to EMS.

Acknowledgments. The authors sincerely thank Toronto EMS for sharing their data, in particular, Mr. Dave Lyons for his comments and support.

SUPPLEMENTARY MATERIAL

Supplement A: Additional tables (DOI: 10.1214/10-AOAS442SUPPA; .pdf). Tables 1 and 2.

Supplement B: Estimation algorithms (DOI: 10.1214/10-AOAS442SUPPB; .R). *R* code for estimating the models in Section 3 and for calculating the RMSE metrics in Section 4.

Supplement C: Simulation algorithms (DOI: 10.1214/10-AOAS442SUPPC; .R). *R* code for implementing the queueing model simulation in Section 4.4.

REFERENCES

ANDREWS, B. and CUNNINGHAM, S. (1995). LL Bean improves call-center forecasting. *Interfaces* **25** 1–13.

ANDREWS, B. and PARSONS, H. (1993). Establishing telephone-agent staffing levels through economic optimization. *Interfaces* **23** 14–20.

BIANCHI, L., JARRETT, J. and CHOUDARY HANUMARA, R. (1998). Improving forecasting for telemarketing centers by ARIMA modeling with intervention. *Internat. J. Forecasting* **14** 497–504.

BORCHERS, D., BUCKLAND, S., PRIEDE, I. and AHMADI, S. (1997). Improving the precision of the daily egg production method using generalized additive models. *Canad. J. Fisheries and Aquatic Sciences* **54** 2727–2742.

BROWN, L., GANS, N., MANDELBAUM, A., SAKOV, A., SHEN, H., ZELTYN, S. and ZHAO, L. (2005). Statistical analysis of a telephone call center: A queueing-science perspective. *J. Amer. Statist. Assoc.* **100** 36–50. MR2166068

CHANNOUF, N., L'ECUYER, P., INGOLFSSON, A. and AVRAMIDIS, A. (2007). The application of forecasting techniques to modeling emergency medical system calls in Calgary, Alberta. *Health Care Management Science* **10** 25–45.

COX, D. R. and ISHAM, V. (1980). *Point Processes*. Chapman & Hall, London. MR0598033

DASKALOV, G. (1999). Relating fish recruitment to stock biomass and physical environment in the Black Sea using generalized additive models. *Fisheries Research* **41** 1–23.

FERLAND, R., LATOUR, A. and ORAICHI, D. (2006). Integer-valued GARCH process. *J. Time Ser. Anal.* **27** 923–942. MR2328548

FOKIANOS, K., RAHBEK, A. and TJØSTHEIM, D. (2009). Poisson autoregression. *J. Amer. Statist. Assoc.* **104** 1430–1439. MR2596998

GOLDBERG, J. B. (2004). Operations research models for the deployment of emergency services vehicles. *EMS Management J.* **1** 20–39.

GROSS, D. and HARRIS, C. M. (1998). *Fundamentals of Queueing Theory*, 3rd ed. Wiley, New York. MR1600527

GU, C. (1992). Cross-validating non-Gaussian data. *J. Comput. Graph. Statist.* **1** 169–179.

GU, C. (2010). gss: General smoothing splines. R Package Version 1.1-3.

HARRISON, J., ZEEVI, A. and SHUM, S. (2005). A method for staffing large call centers based on stochastic fluid models. *Manufacturing & Service Operations Management* **7** 20–36.

HASTIE, T. (2009). gam: Generalized additive models. R Package Version 1.01.

HASTIE, T. J. and TIBSHIRANI, R. J. (1990). *Generalized Additive Models. Monographs on Statistics and Applied Probability* **43**. Chapman & Hall, London. MR1082147

HASTIE, T. and TIBSHIRANI, R. (1993). Varying-coefficient models. *J. Roy. Statist. Soc. Ser. B* **55** 757–796. MR1229881

HENDERSON, S. G. (2005). Should we model dependence and nonstationarity, and if so how? In *Proceedings of the* 37*th Conference on Winter Simulation* (M. Kuhl, N. Steiger, F. Armstrong and J. Joines, eds.) 120–129. IEEE, Piscataway, NJ.

HENDERSON, S. G. (2009). Operations research tools for addressing current challenges in emergency medical services. In *Wiley Encyclopedia of Operations Research and Management Science* (J. J. Cochran, ed.). Wiley, New York.

KOLESAR, P. and GREEN, L. (1998). Insights on service system design from a normal approximation to Erlang's delay formula. *Production and Operations Management* **7** 282–293.

LOPES, H. F. and WEST, M. (2004). Bayesian model assessment in factor analysis. *Statist. Sinica* **14** 41–67. MR2036762

MATTESON, D. S. and TSAY, R. S. (2011). Constrained independent component analysis. Unpublished manuscript.

MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models*, 2nd ed. Chapman & Hall, London.

OPSOMER, J., WANG, Y. and YANG, Y. (2001). Nonparametric regression with correlated errors. *Statist. Sci.* **16** 134–153. MR1861070

PIERCE, D. A. and SCHAFER, D. W. (1986). Residuals in generalized linear models. *J. Amer. Statist. Assoc.* **81** 977–986. MR0867620

R DEVELOPMENT CORE TEAM (2009). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

SETZLER, H., SAYDAM, C. and PARK, S. (2009). EMS call volume predictions: A comparative study. *Comput. Oper. Res.* **36** 1843–1851.

SHEN, H. and HUANG, J. Z. (2008a). Forecasting time series of inhomogeneous Poisson processes with application to call center workforce management. *Ann. Appl. Statist.* **2** 601–623. MR2524348

SHEN, H. and HUANG, J. (2008b). Interday forecasting and intraday updating of call center arrivals. *Manufacturing and Service Operations Management* **10** 391–410.

STOCK, J. H. and WATSON, M. W. (2002). Macroeconomic forecasting using diffusion indexes. *J. Bus. Econom. Statist.* **20** 147–162. MR1963257

SWERSEY, A. (1994). The deployment of police, fire, and emergency medical units. In *Handbooks in Operations Research and Management Science* **6** 151–200. North-Holland, Amsterdam.

TAKANE, Y. and HUNTER, M. A. (2001). Constrained principal component analysis: A comprehensive theory. *Appl. Algebra Engrg. Comm. Comput.* **12** 391–419. MR1864610

TSAI, H. and TSAY, R. S. (2010). Constrained factor models. *J. Amer. Statist. Assoc.* **105** 1593–1605.

TYCH, W., PEDREGAL, D., YOUNG, P. and DAVIES, J. (2002). An unobserved component model for multi-rate forecasting of telephone call demand: The design of a forecasting support system. *Internat. J. Forecasting* **18** 673–695.

WEINBERG, J., BROWN, L. D. and STROUD, J. R. (2007). Bayesian forecasting of an inhomogeneous Poisson process with applications to call center data. *J. Amer. Statist. Assoc.* **102** 1185–1198. MR2412542

WHITT, W. (2002). *Stochastic-Process Limits*. Springer, New York. MR1876437

WOOD, S. N. (2003). Thin plate regression splines. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **65** 95–114. MR1959095

WOOD, S. N. (2006). *Generalized Additive Models: An Introduction with R*. Chapman & Hall, Boca Raton, FL. MR2206355

WOOD, S. (2008). mgcv: GAMs with GCV/AIC/REML smoothness estimation and GAMMs by PQL. R Package Version 1.6-1.

WOODARD, D. B., MATTESON, D. S. and HENDERSON, S. G. (2010). Stationarity of count-valued and nonlinear time series models. Unpublished manuscript.

SCHOOL OF OPERATIONS RESEARCH
AND INFORMATION ENGINEERING
CORNELL UNIVERSITY
282 RHODES HALL
ITHACA, NEW YORK 14853
USA
E-MAIL: dm484@cornell.edu