Density Forecasting of Intraday Call Center Arrivals using Models Based on Exponential Smoothing

James W. Taylor

Saïd Business School

University of Oxford

Management Science, 2012, Vol. 58, pp. 534-549.

Address for Correspondence:

James W. Taylor
Saïd Business School
University of Oxford
Park End Street
Oxford OX1 1HP, UK

Tel: +44 (0)1865 288927
Fax: +44 (0)1865 288805
Email: james.taylor@sbs.ox.ac.uk

Density Forecasting of Intraday Call Center Arrivals using Models Based on Exponential Smoothing

Abstract

A key input to the call center staffing process is a forecast for the number of calls arriving. Density forecasts of arrival rates are needed for analytical call center models, which assume Poisson arrivals with a stochastic arrival rate. Density forecasts of call volumes can be used in simulation models, and are also important for the analysis of outsourcing contracts. A forecasting method, that has previously shown strong potential, is Holt-Winters exponential smoothing adapted for modeling the intraday and intraweek cycles in intraday data. To enable density forecasting of the arrival volume and rate, we develop a Poisson count model, with gamma distributed arrival rate, which captures the essential features of this exponential smoothing method. The apparent stationary level in our data leads us to develop versions of the new model for series with stationary level. We evaluate forecast accuracy up to two weeks ahead using data from three organizations. We find that the stationary level models improve prediction beyond approximately two days ahead, and that these models perform well in comparison with sophisticated benchmarks. This is confirmed by the results of a call center simulation model, which demonstrates the use of arrival rate density forecasting to support staffing decisions.

Key words: call centers; arrival rate; density forecasting; exponential smoothing; seasonality

1. Introduction

The number and size of call centers has increased in recent years to the extent that they employ approximately 3% of the working populations in the US and UK. For many organizations, call centers provide the main channel of communication with their customers. As the human resource costs for call centers typically comprise 60 to 80% of their overall operating budget, it is of great importance to establish an optimal level of staffing. Overstaffing obviously incurs unnecessary operating costs, while understaffing causes queuing times considered by customers as unacceptable, which can lead to the abandonment of calls. Indeed, staff deployment is a balance between service quality and operating costs (Akşin et al. 2007). A key input to the scheduling process is a forecast for the number of calls arriving at the call center. Long-term predictions are needed for recruitment purposes. Forecasts with lead times of several weeks are required for staff scheduling, and these forecasts and schedules need to be regularly revised until the target day itself (Gans et al. 2003). This can be continued within the day, with forecasts for several hours ahead used to enable dynamic updating of the agents' deployment (Mehrotra et al. 2010). The usefulness of intraday forecast updating prompts Shen and Huang (2008a) and Weinberg et al. (2007) to extend their forecasting methods for this purpose.

Time series of intraday call center arrivals consist of variability around a seasonal pattern that is made up of both intraweek and intraday seasonal cycles. Anticipating this variability is important, as higher uncertainty requires higher staffing levels in order to meet specified service level objectives. A prediction of the probability density function (i.e. a density forecast) of the call volume can be used for the arrival process in simulation models, which have been growing in popularity due, at least in part, to the increased complexity of call center operations and developments in computational power (Gans et al. 2003; Mehrotra and Fama 2003). Density forecasts of call volumes can also be used to analyze call center outsourcing contracts (see, for example, Akşin et al. 2008). Contracts may be defined so that the call volume above a specified base level is outsourced, or so that only a base load is outsourced.

For analytical call center models, and for many simulation models, the arrival volume is modeled as a Poisson process, and so a forecast is needed for the arrival rate. Empirical evidence has shown that call center arrivals show significant overdispersion relative to a Poisson process (i.e. the variance is greater than the mean), and this has led to the arrival rate being treated as stochastic (see, for example, Jongbloed and Koole 2001; Avramidis et al. 2004; Akşin et al. 2007, Section 2.4; Bassamboo and Zeevi 2009). In view of this, it is a density forecast of the *arrival rate* that is needed for analytical models. Jongbloed and Koole (2001) describe how to use such a density forecast to assess the impact of the arrival rate uncertainty on the service level, defined as the fraction of calls whose delay falls below a specified target. They show how the θ% quantile of the arrival rate distribution can be plugged into the Erlang C formula to get a service level that will be exceeded with θ% probability. Jongbloed and Koole also describe how, for a given service level, the bounds of an arrival rate prediction interval can be used with the Erlang C formula to deliver bounds for the recommended staffing level. They note that, if a flexible workforce is available, the lower bound for the staffing level can be used as the fixed number of agents, and the interval width can be used as the number of agents needed with flexible contracts. In this paper, we develop models for the density forecasting of both the call volume and the arrival rate.

2. Univariate Models for Intraday Arrivals

One approach to forecasting call arrivals is to use a causal model (see, for example, Soyer and Tarimcilar 2008). The alternative is a method based on only the arrivals series, and it is these univariate methods that tend to be used for intraday call arrivals, which is our focus in this paper. Of the univariate approaches, those involving a statistical model have the advantage that they enable convenient generation of density forecasts. To the best of our knowledge, the only studies of intraday call center arrivals that focus on methods for density forecasting are those of Weinberg et al. (2007) and Shen and Huang (2008a, 2008b). Weinberg et al. extend the random effects statistical model of Brown et al. (2005), which involves the product of the daily total and the proportion of the total that occurs in a given period of the day. Weinberg et al. model both this proportion and the daily total using types of autoregressive models. They use a Markov chain Monte Carlo algorithm to estimate the latent states and model parameters, and hence point and density forecasts for the arrival volume and rate.

The novel method of Shen and Huang (2005, 2008a) involves the use of singular value decomposition (SVD), and it proceeds by arranging the intraday data as a (d×m_1) matrix, where d is the number of days in the estimation sample and m_1 is the number of periods in each day. Each column of this matrix constitutes a time series of daily observations for a particular period of the day. SVD is used to extract the main underlying components in these columns, and thus reduce the problem from having to forecast m_1 daily series to forecasting daily series for just the main components. A simple time series forecasting method is applied to each of these main components. The basic method uses information up to the end of a given day to produce forecasts for all periods of a future day. To enable intraday forecast updating, an additional stage is incorporated. A bootstrap procedure is used to produce density forecasts for the call volumes. To generate density forecasts for the arrival rate, Shen and Huang (2008b) develop a count data version of the method, which involves Poisson modeling with a stochastic arrival rate modeled using a similar dimension reduction approach to that used in their other papers.

If a time series consists of high counts that are Poisson-distributed, a square root transformation stabilizes the variance in the series and allows the fitting of Gaussian models. This transformation is employed by Brown et al. (2005), Weinberg et al. (2007) and Shen and Huang (2005, 2008a). For the type of models that we consider in this paper, if such a transformation is used, it is not clear how a density forecast for the arrival rate could be produced. In view of this, we do not transform the data, and instead follow Shen and Huang (2008b) by fitting Poisson count data models to our data. They explain that, if a Poisson assumption is reasonable, then efficiency should improve by modeling the count data directly, rather than modeling the transformed series. A further motivation for developing count data models is that they are needed for call arrivals series that consist of relatively low volumes. An example of this is the quarter-hourly series of total call arrivals at an Israeli bank, analyzed by Brown et al. (2005, Section 6), which has 10% of periods with less than 10 calls. Low volume arrivals often occur at multiskill call centers, which involve different types of calls being served by different groups with specialized skills (see, for example, Cezik and L'Ecuyer 2008; Pot et al. 2008). One of the series considered in this paper consists of relatively low volume arrivals channeled to a specialized group of agents at a multiskill call center.

In the empirical analysis of Taylor (2008), a method that performed well against ARMA and state space alternatives, for lead times up to about four days ahead, was Holt-Winters exponential smoothing adapted by Taylor (2003) for modeling both the intraday and intraweek cycles in intraday data. We term this method HWT exponential smoothing. Exponential smoothing methods are widely used in a variety of business applications (see Gardner 2006). They are based on exponentially weighted averaging, and have the appeal of intuitiveness and transparency, which is important if adoption by practitioners is valued. In this paper, we develop the HWT method by providing the following two main contributions:

- To enable density forecasting of the arrival volume and rate, we develop a Poisson count data model that captures the essential features of the HWT method. We implement a gamma distributed stochastic arrival rate, as in the work of Jongbloed and Koole (2001). We are not aware of previous studies that have considered a univariate seasonal time series count data model for intraday call center arrivals.
- Like all of the standard exponential smoothing methods, the new HWT count data model possesses a nonstationary level. As this can be inappropriate for call arrivals data, we develop new versions of the model that have a stationary level.

In terms of conceptual simplicity and ease of implementation, the HWT models compare well with the existing call arrivals density forecasting methods, which are presented by Weinberg et al. (2007) and Shen and Huang (2008a, 2008b). These authors emphasize the importance of intraday forecast updating, and for the HWT model this is straightforward because the one model is used directly to produce forecasts for all lead times and from all forecast origins. A limitation of the methods of Weinberg et al. and Shen and Huang (2008a) are that they are suitable only for high volume arrivals data.

3. Intraday Call Center Arrivals Data

In this paper, our empirical analysis mainly focuses on one high volume and one low volume series of half-hourly arrivals. The high volume series corresponds to total arrivals at the call centers of NHS Direct, which is the 24-hour telephone helpline provided by the National Health Service in England and Wales. The NHS Direct structure is similar to that described by Shumsky and Pinker (2003), with calls initially handled by a 'gatekeeper' who provides simple information or passes the call on to a nurse or health information staff. The NHS Direct series consists of 35 weeks of observations, which vary from 28 to 2,023 calls. This data is also used in the study of Taylor (2010), which develops new exponentially weighted methods for series with multiple seasonal cycles that might occur in a variety of applications. In contrast to our current paper, Taylor (2010) focuses on only point forecasting and does not consider count data models. It is not clear how one would produce an arrival rate density forecast from the methods presented in that study. In that paper, Figures 1 to 3 show that the fluctuations in the NHS Direct series are dominated by seasonality. The first of these figures shows an intraday seasonal cycle of duration m_1=48 periods, and an intraweek cycle of length m_2=336 periods. Repeating intraday and intraweek seasonal cycles is typical of intraday call arrivals data.

For all series in this paper, exponential smoothing and ARMA count data model parameters were estimated once using an initial sample, and the forecast origin was then rolled forward through a post-sample evaluation period to produce a collection of forecasts for each horizon. We used the first 25 weeks of NHS Direct data for estimation and the remaining 10 for evaluation. The methods considered in this paper are not of use for days on which the pattern of calls is unusual, such as public holidays, and so we replaced observations for these days using simple averaging procedures. These days were not included in our post-sample forecast evaluation.

Our second series comes from the multiskill call center of a UK credit card company. It consists of half-hourly arrivals channeled to a group of agents specializing in a particular type of enquiry. This channel is open from 9 am to 8 pm on Mondays to Saturdays, and is closed on Sundays. The 70-week series is shown in Figure 1. About 10% of the periods had 11 calls or fewer, and about 1% had no calls at all. For such low volume data, count data models are appropriate. We used the first 50 weeks for model estimation, and the final 20 for post-sample evaluation. Figure 2 shows a four-week period of the data. In spite of substantial volatility, there is evidence of a repeating weekly cycle. Given the opening hours, an intraday cycle would consist of m_1=22 half-hourly periods, and an intraweek cycle would be of length m_2=132 periods. Figure 3 shows the average daily call profiles, calculated using the in-sample data.

---------- Figures 1-3 ----------

4. HWT Exponential Smoothing for Continuous Data

In this section, we introduce statistical models based on Taylor's (2003, 2008) HWT method. These models are appropriate for observations from continuous random variables. Although it can be reasonable to treat high volume arrivals data as observations from a continuous random variable, we do not apply the models of this section to arrivals data in this paper. The purpose of the models in this section is simply to provide the foundation for Section 5, where we develop new count data models.

Density forecasts can be generated from an exponential smoothing method by expressing it as a single source of error state space model and then using Monte Carlo simulation (Hyndman et al. 2008). The following single source of error state space 'HWT model' corresponds to the additive HWT method.

HWT model – additive version:

$$y_t = l_{t-1} + d_{t-m_1} + w_{t-m_2} + \phi e_{t-1} + \varepsilon_t \tag{1a}$$

$$e_t = y_t - \left(l_{t-1} + d_{t-m_1} + w_{t-m_2}\right) \tag{1b}$$

$$l_t = l_{t-1} + \alpha\, e_t \tag{1c}$$

$$d_t = d_{t-m_1} + \delta\, e_t \tag{1d}$$

$$w_t = w_{t-m_2} + \omega\, e_t \tag{1e}$$

y_t is the target variable; l_t is a level term; d_t is the seasonal factor for the intraday cycle; w_t is the factor for the intraweek cycle remaining after the intraday cycle is removed; m_1 and m_2 are the respective lengths of these cycles; α, δ, ω and ϕ are parameters; and $\varepsilon_t \overset{iid}{\sim} N(0,\sigma^2)$. The term involving ϕ is an important adjustment for autocorrelation in the residuals, e_t. Note that using expression (1b) to substitute for e_t elsewhere in the model results in a model with just a single random error, ε_t. Having two terms with lags of one in the observation equation of expression (1a) means that the level is modeled by a combination of l_t and e_t, and so the updating of the level is dictated by both α and ϕ. We return to this issue in Section 5.3.

The HWT model of expressions (1a)-(1e) has additive seasonality and an additive error, ε_t. If the seasonality depends on the level of the series, then multiplicative seasonality is appropriate. A multiplicative error term, which is also known as a relative error, is appropriate if the variance of the randomness depends on the level and seasonality. One approach to dealing with multiplicative seasonality and error is to apply a variance stabilizing transformation, such as a square root or logarithmic transformation, and then use an additive model. A more direct approach is to apply a model with multiplicative structure to the original untransformed data. For all of the standard exponential smoothing models, Hyndman et al. (2008, Section 2.5) consider formulations with additive and multiplicative error, and additive and multiplicative seasonality. In expressions (2a)-(2e), we present the HWT model with multiplicative seasonality and multiplicative error. l_t is a level term; d_t and w_t are multiplicative seasonal factors; and $\varepsilon_t \overset{iid}{\sim} N(0,\sigma^2)$. To be consistent with the form of the standard multiplicative error term $(1+\varepsilon_t)$ in expression (2a), we have incorporated the residual autocorrelation as a multiplicative term $(1+\phi e_{t-1})$.

HWT model – multiplicative version:

$$y_t = l_{t-1}\, d_{t-m_1}\, w_{t-m_2}\, (1+\phi e_{t-1})(1+\varepsilon_t) \tag{2a}$$

$$(1+e_t) = y_t \,/\, (l_{t-1}\, d_{t-m_1}\, w_{t-m_2}) \tag{2b}$$

$$l_t = l_{t-1}(1+\alpha e_t) \tag{2c}$$

$$d_t = d_{t-m_1}(1+\delta e_t) \tag{2d}$$

$$w_t = w_{t-m_2}(1+\omega e_t) \tag{2e}$$

In our two call arrivals series, plotted in Figure 2 of Taylor (2010) and in Figure 1 of this paper, the fluctuations seem to depend on the level of the data. This is consistent with Poisson arrivals, although the precise relationship between the mean and variance will be affected by the stochastic nature of the arrival rate. As the fluctuations depend on the level, it seems appropriate to use a model with multiplicative seasonality and multiplicative error. However, the model of expressions (2a)-(2e) is unsuitable because, as we explained in Sections 1 and 2, we require Poisson-based count data models. We develop such models in the next section.

5. Count Data Models Based on Exponential Smoothing

5.1. Review of Models and Distributions for Count Time Series

In this section, we briefly review models for count time series and describe the distributions that we use in the new models that we introduce in Sections 5.2 and 5.3. The Poisson distribution is often used in count data models. Such models involve the specification of a suitable structure for the arrival rate. An example is Poisson regression, which involves modeling the rate as $\lambda_t = h(\boldsymbol{x}_t'\boldsymbol{\beta})$, where h is an exponential function, $\boldsymbol{x}_t$ is a vector of regressors, and $\boldsymbol{\beta}$ is a vector of parameters.

Univariate models for count time series are either observation-driven or parameter-driven (Jung et al. 2006). The dynamics of observation-driven models are captured by lagged values of the counts. Davis et al. (2003) consider a Poisson form of generalized linear ARMA model, which is observation-driven and relates an exponential function of ARMA terms to the rate of the Poisson distribution. With parameter-driven models, autocorrelation in the counts is modeled using a latent dynamic process for the parameters of the conditional distribution. An example is the model of Zeger (1988), which allows for an autocorrelated term, l_t, within Poisson regression, and is written as:

$$y_t \sim \text{Poiss}(\lambda_t) \tag{3a}$$

$$\lambda_t = h(\boldsymbol{x}_t'\boldsymbol{\beta})l_t \tag{3b}$$

$$l_t = l_{t-1}^{\theta} h(\alpha\varepsilon_t) \tag{3c}$$

where y_t is the count for the period ending at time t; α and θ are parameters; h is the exponential function; and $\varepsilon_t \overset{iid}{\sim} N(0,\sigma^2)$. Feigin et al. (2008) explain that parameter-driven count data models can be viewed as multiple source of error state space models, while observation-driven count data models have a single source of error. Hyndman et al. (2008) explain that density forecasting and nonlinear modeling is particularly convenient with single source of error models. In view of this, we elect to develop in this paper new HWT count data models of this type. There is little literature on the time series modeling of seasonal count data, and we are not aware of any exponential smoothing models for such data.

Call center arrivals tend to be modeled as a Poisson process with a time-varying arrival rate. However, as discussed in Section 1, the arrivals often exhibit overdispersion, and this has prompted the assumption of a stochastic arrival rate. Jongbloed and Koole (2001) consider the use of a Poisson distribution with gamma distributed arrival rate, which implies an assumption of a negative binomial distribution for the arrival volume. In this paper, we use this distribution within time series models. In our models, the mean and variance of the negative binomial distribution are time-varying, and can be written as μ_t and μ_t/ψ, respectively, where $0<\psi\leq 1$. The parameter ψ controls the degree of overdispersion. (The corresponding mean and variance of the gamma distributed arrival rate are μ_t and $\mu_t(1-\psi)/\psi$, respectively.) Although not considered by Jongbloed and Koole, setting $\psi=1/(1+\mu_t\varphi)$ (where $\varphi\geq 0$) implies that the variance of the negative binomial distribution is $\mu_t+\varphi\mu_t^2$, a quadratic function of the mean (see Heinen 2003). (This variance specification is natural when the negative binomial is viewed as describing the number of failures in a sequence of trials before the rth success. In this context, $\varphi=1/r$.)

5.2. An HWT Count Data Model

Expressions (4a)-(4g) present a new negative binomial count data model based on the multiplicative HWT model of Section 4. We use a multiplicative structure because it avoids the possibility of a negative arrival rate, and because, as noted in Section 4, a multiplicative formulation is more suitable for our data.

HWT count data model:

$$y_t \sim NegBin(\mu_t, \psi) \tag{4a}$$

$$\mu_t = l_{t-1}\, d_{t-m_1}\, w_{t-m_2}\, h(\phi\, e_{t-1}) \tag{4b}$$

$$(1+e_t) = y_t \,/\, (l_{t-1}\, d_{t-m_1}\, w_{t-m_2}) \tag{4c}$$

$$l_t = l_{t-1} h(\alpha\, e_t) \tag{4d}$$

$$d_t = d_{t-m_1} h(\delta\, e_t) \tag{4e}$$

$$w_t = w_{t-m_2} h(\omega\, e_t) \tag{4f}$$

$$\text{where } h(x) = \frac{(1+\exp(\rho))}{(1+\exp(\rho - x))} \tag{4g}$$

y_t is the count for the half-hour period ending at time t; μ_t is the mean and ψ is the overdispersion parameter of the negative binomial distribution, as defined in Section 5.1; and h is a logistic function with parameter ρ.

Note that with $h(x)=(1+x)$, this new model would have similar form to the multiplicative HWT model of Section 4. However, this choice for the function h cannot ensure a non-negative arrival rate in the second of our extensions of the model for stationary level, which we consider in Section 5.3. An alternative for h is an exponential function, which, as we discussed in Section 5.1 has been used in several count data models. Indeed, with this choice of h, the state equations of expressions (4d)-(4f) are similar in structure to expression (3c) of Zeger's (1988) formulation. However, with our data, using an exponential function for h led to occasional extremely large values for the level and seasonal components when performing Monte Carlo simulation to generate density forecasts. The cause of this was low counts y_t, which led to low values for the level term l_t, which in turn resulted in the relative residual error e_t in expression (4c) becoming large, especially when the seasonal factors, d_t and w_t, were at their lowest. The combination of a large value for e_t and the use of an exponential function for h caused expressions (4d)-(4f) to deliver occasional extremely large values for the level and seasonal components. This led to us selecting h to be a function with an upper bound, as well as a lower bound of zero. A natural choice was the logistic function of expression (4g), which has the attractive properties that it is non-negative, it has an upper bound (equal to (1+exp(ρ))), it is monotonic, and it is equal to 1 when x=0. The value of ρ dictates and limits the impact of the previous period's residual on the current period's estimates of the level and seasonal components. We note that upper and lower bounds have previously been considered within a count data model by Fokianos (2001), and that a logistic function is used in logistic regression, which is a form of generalized linear model (see McCullagh and Nelder 1989).

The HWT count data model is an observation-driven (single-source of error) model. This can be seen by using expression (4c) to substitute for e_t in the other expressions to give a model written in terms of y_t. For the HWT count data models in this paper, analytical expressions for point and density forecasts exist for only one step-ahead prediction. For longer lead times, Monte Carlo simulation must be used.

The use of heuristics is common when initializing exponential smoothing models (see Hyndman et al. 2008, Section 5.2). Our heuristic approach was based on the first three weeks of data. We used the mean of these observations to initialize l_t. Our initialization of d_t proceeded by calculating, for a given period, the ratio of the observed value, y_t, to the m_1-point moving average. The initial value of d_t, for each period of the day, was set as the geometric mean of the first seven of these ratios corresponding to the same period of the day. To initialize w_t, we first calculated, for a given period, the ratio of the observed value, y_t, to the m_2-point moving average. The initial value of w_t, for each period of the week, was set as the geometric mean of the first two of these ratios, corresponding to the same period of the week, divided by the initial value of d_t for the same period of the day.

Maximum likelihood was used for parameter estimation. Using the negative binomial density function, we constructed the likelihood function as:

$$\prod_{\substack{t=3m_2+1 \\ t\notin \mathrm{P}}}^{N} \frac{\Gamma\left(y_t+\mu_t\psi/(1-\psi)\right)}{y_t!\,\Gamma\left(\mu_t\psi/(1-\psi)\right)}\psi^{\mu_t\psi/(1-\psi)}(1-\psi)^{y_t}$$

where N is the number of periods in the estimation sample; the set P contains the periods corresponding to public holidays (which were omitted from the parameter optimization); Γ is the gamma function; μ_t is given by expression (4b); and ψ is the overdispersion parameter. To optimize this objective function, we followed a procedure similar to that used by Engle and Manganelli (2004) for a different type of model. Maximization of the likelihood proceeded by sampling 10^4 vectors of parameters using a uniform random number generator with bounds u_L and u_U. Experimentation led to the use of u_L=0 and u_U=1 for α, δ, ω, ψ and φ; u_L=-5 and u_U=5 for ρ; and u_L=0 and u_U=2 for ϕ. For each of the 10^4 randomly sampled vectors, we evaluated the log likelihood function. The three vectors that produced the highest log likelihood values were then used, in turn, as the initial vector in a quasi-Newton algorithm. Of the three resulting vectors, the one producing the highest log likelihood was chosen as the final parameter vector.

In the row labeled “HWT” of Table 1, for the NHS Direct data, we present the estimated parameters for the HWT count data model of expressions (4a)-(4g) with variance modeled as μ_t/ψ. The value of ρ implies that the function, h, has an upper bound of (1+exp(-0.306))=1.736. Substantial overdispersion is indicated by the low value of ψ. In Section 5.1, we described how the variance can be modeled as $\mu_t+\varphi\mu_t^2$, a quadratic function of the mean. In the row labeled “HWT” of Table 2, for the NHS Direct data, we present the estimated parameters for the model with variance modeled in this way. The positive value of φ indicates overdispersion.

--------- Tables 1 and 2 ----------

5.3. HWT Count Data Models with Stationary Level

From Figure 2 of Taylor (2010), the NHS Direct time series appears to be stationary in the level. Although there is more change in the level of the credit card company series in Figure 1, it could be argued that the level of that series is also mean-reverting. If a series has a stationary level, and a model with nonstationary level is used, there would be a misspecification problem and the accuracy of long-term forecasting is likely to be particularly poor. However, in each of the HWT formulations in Sections 4 and 5.2, unless α=0, the level, l_t, is modeled as nonstationary. Indeed, all of the standard forms of exponential smoothing, presented in Gardner's (2006) review, possess a nonstationary structure for the modeling of the level. In this section, we develop new versions of the HWT model with stationary level.

Let us consider setting α=0 in the HWT count data model of expressions (4a)-(4g). This leads to l_t being equal to a constant, which can be interpreted as a constant long-run mean level, $\bar{l}$. We can then rewrite expressions (4b) and (4c) as expressions (5) and (6), respectively.

$$\mu_t = \bar{l}\, d_{t-m_1}\, w_{t-m_2}\, h(\phi e_{t-1}) \tag{5}$$

$$(1+e_t) = y_t \,/\, (\bar{l}\, d_{t-m_1}\, w_{t-m_2}) \tag{6}$$

Let us define a new level as in expression (7).

$$l_t = \bar{l}\, h(\phi e_t) \tag{7}$$

Using this, we can rewrite expression (5) as expression (8b). Let us define a new relative error ε_t in terms of the new level, as in expression (8c). Using expressions (6) and (8c), we obtain $e_t = (l_{t-1} - \bar{l})/\bar{l} + (l_{t-1}/\bar{l})\varepsilon_t$. We can use this to substitute for e_t in expressions (7), (4e) and (4f) to deliver expressions (8d)-(8f), respectively. In summary, the result of setting α=0 in the HWT model of expressions (4a)-(4g) is the new count data model of expressions (8a)-(8g). The interesting feature of this model is that expression (8d) represents a stationary modeling of the level. The level and two seasonal terms are influenced by deviations of the previous period's level from the long-run mean, $\bar{l}$. We optimized $\bar{l}$ in the same procedure as the model parameters. Optimized model parameters for the NHS Direct data are given in Tables 1 and 2, in the rows entitled "HWT with Stationary Level". In each table, the values of δ, ϕ and ρ for this model are generally quite different to the corresponding values for the HWT model of Section 5.2.

HWT count data model with stationary level:

$$y_t \sim NegBin\left(\mu_t, \psi\right) \tag{8a}$$

$$\mu_t = l_{t-1}\, d_{t-m_1}\, w_{t-m_2} \tag{8b}$$

$$\left(1+\varepsilon_t\right) = y_t \;/\; \left(l_{t-1} d_{t-m_1} w_{t-m_2}\right) \tag{8c}$$

$$l_t = \bar{l}\ h\left(\phi\left(l_{t-1} - \bar{l}\right)/\bar{l} + \phi\left(l_{t-1}/\bar{l}\right)\varepsilon_t\right) \tag{8d}$$

$$d_t = d_{t-m_1}\ h\left(\delta\left(l_{t-1} - \bar{l}\right)/\bar{l} + \delta\left(l_{t-1}/\bar{l}\right)\varepsilon_t\right) \tag{8e}$$

$$w_t = w_{t-m_2}\ h\left(\omega\left(l_{t-1} - \bar{l}\right)/\bar{l} + \omega\left(l_{t-1}/\bar{l}\right)\varepsilon_t\right) \tag{8f}$$

$$\text{where}\quad h(x) = \frac{\left(1+\exp(\rho)\right)}{\left(1+\exp(\rho - x)\right)} \tag{8g}$$

Although formal statistical testing is not the norm when selecting from a set of exponential smoothing methods, we note that, in principle, a unit root test could be performed to establish whether or not a model with stationary level should be used. Such a test would need to be suitable for testing for a unit root in high frequency data in the presence of strong seasonality.

If the model of expressions (8a)-(8g) was introduced with no reference to the earlier HWT count data model, one might ask why, in each state equation, the coefficients of $(l_{t-1} - \bar{l})/\bar{l}$ and $(l_{t-1}/\bar{l})\varepsilon_t$ are specified to be the same. This prompts an unconstrained version of the model, which involves three additional parameters, ϕ', δ' and ω', and which we present in expressions (9a)-(9g). We refer to this as the 'generalized' form of the model. Optimized parameter values for this model are presented in the bottom rows of Tables 1 and 2. In each table, the parameters δ and δ' have quite different values, and so also do ϕ and ϕ'. This gives some justification for considering the generalized version of the model.

HWT count data model with stationary level and generalized:

$$y_t \sim NegBin(\mu_t, \psi) \tag{9a}$$

$$\mu_t = l_{t-1}\, d_{t-m_1}\, w_{t-m_2} \tag{9b}$$

$$(1+\varepsilon_t) = y_t \,/\, \left(l_{t-1} d_{t-m_1} w_{t-m_2}\right) \tag{9c}$$

$$l_t = \bar{l}\ h\left(\phi\left(l_{t-1} - \bar{l}\right)/\bar{l} + \phi'\left(l_{t-1}/\bar{l}\right)\varepsilon_t\right) \tag{9d}$$

$$d_t = d_{t-m_1}\, h\left(\delta\left(l_{t-1} - \bar{l}\right)/\bar{l} + \delta'\left(l_{t-1}/\bar{l}\right)\varepsilon_t\right) \tag{9e}$$

$$w_t = w_{t-m_2}\, h\left(\omega\left(l_{t-1} - \bar{l}\right)/\bar{l} + \omega'\left(l_{t-1}/\bar{l}\right)\varepsilon_t\right) \tag{9f}$$

$$\text{where} \quad h(x) = \frac{(1+\exp(\rho))}{(1+\exp(\rho - x))} \tag{9g}$$

An HWT model can be used directly to produce forecasts for all lead times and from all forecast origins. This relates to intraday updating, and also to how an HWT model would be used within a call center's capacity planning cycle (see Gans et al. 2003, Section 3.4). A forecast would be produced from the HWT model some weeks in advance, and at any point up to and including the scheduled day itself, the staffing plan would be updated based on new forecasts generated from the same HWT model.

The only methods put forward so far in the literature for predicting densities of both call arrival volumes and rates are the methods of Weinberg et al. (2007) and Shen and Huang (2008b). The choice as to whether to use one of these methods or an HWT model will ultimately be decided by forecast accuracy, and by how easily the methods can be understood and implemented.

6. Empirical Analysis

Section 6.1 lists the forecasting methods included in our empirical study. Section 6.2 describes the statistical measures used to evaluate accuracy. Sections 6.3-6.5 report the results for the two series described in Section 3, and the US bank data of Weinberg et al. (2007). Section 6.6 uses a call center simulation model to evaluate the use of the arrival rate density forecasts in setting staffing levels.

6.1. Forecasting Methods

HWT – This is the model presented in expressions (4a)-(4g). We refer to this as the basic HWT model. For all the HWT models and the ARMA model described below, we considered the three distributions discussed in Section 5.1: Poisson, negative binomial, and negative binomial with variance a quadratic function of the mean. For each forecast origin, we used Monte Carlo simulation with 1,000 iterations to generate a density forecast. A point forecast was constructed as the mean of the 1,000 iterations.

HWT with stationary level – This is the model presented in expressions (8a)-(8g).

HWT with stationary level and generalized – This is the model presented in expressions (9a)-(9g).

ARMA – This is an ARMA count data model of the type considered by Davis et al. (2003) with the inclusion of lags corresponding to the intraday and intraweek cycles:

$$\begin{aligned}
& y_t \sim NegBin(\mu_t, \psi) \\
& \mu_t = \exp(z_t) \\
& e_t = (y_t - \exp(z_t))\exp(-\gamma z_t) \\
& z_t = \phi_0 + \phi_1(z_{t-1} + e_{t-1}) + \phi_2(z_{t-2} + e_{t-2}) + \phi_3(z_{t-3} + e_{t-3}) \\
& \qquad + \phi_{m_1}(z_{t-m_1} + e_{t-m_1}) + \phi_{2m_1}(z_{t-2m_1} + e_{t-2m_1}) + \phi_{3m_1}(z_{t-3m_1} + e_{t-3m_1}) \\
& \qquad + \phi_{m_2}(z_{t-m_2} + e_{t-m_2}) + \phi_{2m_2}(z_{t-2m_2} + e_{t-2m_2}) + \phi_{3m_2}(z_{t-3m_2} + e_{t-3m_2}) \\
& \qquad + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \theta_3 e_{t-3} + \theta_{m_1} e_{t-m_1} + \theta_{2m_1} e_{t-2m_1} + \theta_{3m_1} e_{t-3m_1} + \theta_{m_2} e_{t-m_2} + \theta_{2m_2} e_{t-2m_2} + \theta_{3m_2} e_{t-3m_2}
\end{aligned}$$

where ψ, γ, the ϕ_i and the θ_i are constant parameters. We based model selection on the Schwarz Bayesian Criterion with the requirement that all parameters were significant (at the 5% level).

SVD-type approach with AR factor models – This is the method of Shen and Huang (2008b), described briefly in Section 2. Within a Poisson modeling framework, an SVD-type procedure is used to reveal a daily series for each of the main underlying factors in the intraday cycle. Our implementation closely followed Shen and Huang, with use of their alternating maximum likelihood algorithm to extract the factors; implementation of their "varying intercept AR(1) models" to forecast the daily series of each factor; use of their penalized maximum likelihood to update the within-day forecasts of the factors, as the forecast origin moved through each day; and bootstrapping with 1,000 replicates to generate forecasts. Forecasts of the factors for a future day were generated from the varying intercept AR(1) models using, as factors for the current day, the updated within-day forecasts. We used cross-validation to select: the number of factors; the number of iterations to employ within each likelihood maximization; and the weight in the penalized maximum likelihood. The cross-validation used a hold-out sample of the final five weeks of the estimation sample for the NHS Direct data and the US bank data, and for the longer credit card company series, we used the final 10 weeks. This led to the use of five factors for the NHS Direct data, two for the credit card company data, and four for the US bank data.

SVD-type approach with exponential smoothing factor models – In this version of the method of Shen and Huang (2008b), we replace the varying intercept AR(1) models by standard Holt-Winters exponential smoothing models. The introduction of this method was prompted by the SVD-type approach with AR factor models performing relatively poorly for the NHS Direct and credit card company series. Taylor (2010) presents an SVD-based approach that uses just one exponential smoothing model to smooth and update all the factors simultaneously, with the arrival of each new intraday observation. However, it is not a count data model, and so it is unclear how it could be used to deliver density forecasts of the arrival rate.

Seasonal mean – In this point forecasting benchmark method, the forecast for each lead time is the mean of a moving window of in-sample arrivals for the same half-hour of the week as the period to be predicted. We set the moving window to be equal in length to the initial estimation sample. Simplistic methods, such as this, are described by Mehrotra and Fama (2003) as commonly used in practice.

6.2. Forecast Evaluation Measures

This paper is focused on the density forecasting of the arrival rate and of call volumes. However, because the rate is unobservable, we follow Weinberg et al. (2007) by evaluating post-sample accuracy of only the density forecasts of call volumes. Given the ordinal and discrete nature of count data, to evaluate the density forecasts we employed the ranked probability score (RPS) (see Wilks 1995, Section 7.4.8):

$$\mathrm{RPS}(F_t,y_t) = \sum_{j=0}^{J}\left(F_t(j) - I(y_t \le j)\right)^2$$

F_t is the forecast of the cumulative distribution function (cdf), y_t is the actual call volume, I is the indicator function, and J is the maximum possible outcome for y_t. (We set J to be the larger of the following two values: y_t and the largest outcome generated by the Monte Carlo simulation from the model that is being evaluated.) The RPS has been widely used, particularly in the atmospheric sciences, to evaluate discrete density forecasts. It is a squared forecast error measure, where the predicted cdf is evaluated at each possible outcome j. Lower values of the RPS are preferable. The measure captures the two important characteristics of a distributional forecast: its location relative to the observed value and its sharpness around that value. For each lead time, we calculated a mean RPS value by averaging over the RPS values obtained for the different forecast origins.

We also evaluated the 5%, 25%, 75% and 95% quantiles of the density forecasts by calculating the hit percentage, which is the percentage of observations falling below the θ quantile forecast. Ideally, this percentage should be θ. We examined significant difference from this ideal using a test based on the binomial distribution. We note that the hit percentage measures the unconditional coverage in a quantile forecast, and that a thorough evaluation requires a measure of conditional coverage (Christoffersen 1998).

To evaluate point forecast accuracy, we calculated the mean absolute error and root mean squared error. For the NHS Direct and US bank data, we also calculated the mean absolute percentage error, but this could not be calculated for the credit card company series because it contained zeros. The rankings of the methods, for each series, were similar for each error measure used.

For each of the ARMA and stationary HWT models, beyond the very early lead times, the RPS results were similar for the three distributions. Evaluating the upper quantiles using the hit percentage, we found that the results for these models tended to be better for the two negative binomial distributions than for the Poisson. We show this for the NHS Direct data in Section 6.3. For the basic HWT model, a Poisson assumption was slightly preferable, but this is not particularly noteworthy because this model was generally uncompetitive. In the rest of the paper, unless stated to the contrary, we report the results for each model with negative binomial distribution and variance modeled as a quadratic function of the mean.

6.3. Forecasting Results for the NHS Direct Data

Figure 4 presents the mean RPS results for the NHS Direct series. The results are disappointing for the SVD-type approach with AR factor models, although it was more accurate than the ARMA model for the very early lead times. Both of these methods were comfortably outperformed at all lead times by the SVD-type approach with exponential smoothing factor models. This method was also noticeably more accurate than the basic HWT model. Interestingly, both of the HWT models with stationary level delivered a clear improvement on the basic HWT model beyond about one day ahead, and these two methods were, overall, the most accurate. Turning to point forecast evaluation, the rankings of methods were very similar to those shown in Figure 4 for the mean RPS. This indicates that density forecast accuracy is strongly related to the quality of the estimation of the location of the density. The seasonal mean point forecasting method performed very poorly relative to the other methods.

---------- Figures 4-7 ----------

In Figure 5, we show the unconditional coverage results for the 95% quantiles. For the other quantiles, the relative performances of the methods were similar to those for the 95% quantile, except that neither of the SVD-type approaches was competitive for the 5% and 25% quantiles. For the HWT models and the SVD-type approach with AR factor models, the results in Figure 5 are reasonably consistent with the mean RPS results of Figure 4. The best results in Figure 5 are for the ARMA model, which contrasts with the mean RPS results in Figure 4. An explanation for this is that unconditional coverage only conveys the average coverage, and does not evaluate time-variation in the quantile and in the coverage. This has prompted the development of measures of conditional coverage for quantile estimates (see, for example, Engle and Manganelli 2004). An appeal of the RPS measure is that it captures the ability of the density forecast to vary over time with the data, and it assesses the quality of the whole density forecast, rather than just an individual quantile. In Figure 6, we evaluate forecasts of the 95% quantile from the ARMA model and one of the HWT models using the three different distributions. The figure shows that using a Poisson distribution led to poor results, and that the results for the negative binomial were slightly improved with variance modeled as a quadratic function of the mean.

Figure 7 provides an illustration of prediction intervals generated by the generalized form of the HWT model with stationary level, and with distribution specified as negative binomial and variance a quadratic function of the mean. The figure shows the observed calls, forecasts of the mean, and three different prediction intervals. The narrowest interval is a naïve estimate of the prediction interval of call volume. It was produced using the model's prediction for the mean as the deterministic rate of a Poisson distribution. This interval estimate assumes there is no overdispersion, and it is interesting to see how much narrower this interval is than the other interval for volume, which was correctly derived based on the model's assumption of a Poisson distribution with stochastic arrival rate. Turning to the interval for the stochastic arrival rate, we see that it indicates substantial uncertainty. Note that, for a Poisson random variable, y, with stochastic arrival rate, λ, the variance of y is equal to the sum of the mean and variance of λ (Karlis and Xekalaki 2005). Therefore, when the variance of λ is much larger than its mean, the standard deviation of y will be relatively similar to the standard deviation of λ. This implies that large uncertainty in the arrival rate will lead to the interval for the rate being relatively close to the interval for the call volume based on this stochastic rate. We see this in Figure 7. For the SVD-type approach with factor models, the analogous figure showed very similar features for the three types of prediction interval.

6.4. Forecasting Results for the Credit Card Company Data

The mean RPS results for the credit card company series are presented in Figure 8. The figure shows that the ARMA model was relatively inaccurate at all lead times. The relative performances of the three HWT models were similar to those for the NHS Direct data. The basic HWT model was competitive for the early lead times, but it was matched for these early lead times, and comfortably outperformed for longer horizons, by the generalized form of the HWT model for stationary level. The SVD-type approach with AR factor models performed well up to about two days ahead, but was disappointing beyond this. Using exponential smoothing factor models in this approach delivered a clear improvement, and, indeed, led to the most accurate mean RPS results of all the methods up to four days ahead. The quantile unconditional coverage hit percentage results were broadly consistent with the mean RPS results, except that the ARMA model was more competitive. The point forecasting results showed a similar ranking of methods to that for the mean RPS, except that the SVD-type approach with exponential smoothing factor models and the HWT models with stationary level produced similar results. The seasonal mean point forecasting method performed very poorly.

---------- Figure 8 ----------

6.5. Forecasting Results for the US Bank Data

The US bank series of Weinberg et al. (2007) consists of arrivals recorded at a five-minute frequency on weekdays for a 34-week period. We converted it into a series of half-hourly arrivals in order to be consistent with our other two series. The resultant series is shown in Figure 9. It consists of 4,760 observations, corresponding to the weekday half-hourly periods between 7am and 9pm. This implies m_1=28 periods in each day, and m_2=140 periods in each week. As the series possesses only weekday observations, the intraweek cycle is less pronounced than for the other series. The series consists of high volumes, varying from 97 to 2,521 calls, with a median of 1,278. We used the first 24 weeks for estimation of the HWT and ARMA parameters, and the final 10 weeks for post-sample evaluation.

The mean RPS results are presented in Figure 10. As with the other two series, the overall results are relatively poor for the ARMA and basic HWT model. The results are also disappointing for the SVD-type approach with exponential smoothing factor models. The figure shows that the SVD-type approach with AR factor models performed slightly better than the two HWT models with stationary level. Interestingly, beyond about one day ahead, the point forecasting results were very similar for these three methods and for the simplistic seasonal benchmark method. The competitiveness of this simplistic method suggests that there is little change in the strong seasonal patterns over the course of the series. To some degree, this is confirmed by Figure 9, which shows relatively stable behavior when compared with the plots of the other two series in Figure 2 of Taylor (2010) and in Figure 1 in this paper. In terms of the unconditional coverage hit percentage quantile results, the rankings of the six methods varied across the four different quantiles, with no clear superior or inferior method.

---------- Figures 9-11 ----------

6.6. A Simulation Model of NHS Direct

We now consider how an arrival rate density forecast can be used for staffing decisions. The square-root safety staffing principle is often used for reasonably large call centers (see Halfin and Whitt 1981). It is the result of an approximation to the Erlang C formula. The principle states that, with a staffing level of $\left(\lambda/\upsilon + \varphi\sqrt{\lambda/\upsilon}\right)$, the probability of delay equals $\left(1 + \varphi G(\varphi)/g(\varphi)\right)^{-1}$, where λ is the arrival rate; υ is the service rate; G and g are the standard Gaussian cdf and density functions, respectively; and $\varphi > 0$ is a service level parameter. Consider $\varphi = 1$ and $\upsilon = 6$ calls per half-hour period. (The average service time for NHS Direct is about 5 minutes.) With these parameters, the principle implies that 22% of calls will be delayed when the staffing level is:

$$\lambda/6 + \sqrt{\lambda/6} \tag{10}$$

However, the stochastic nature of λ implies uncertainty in this staffing level. Jongbloed and Koole (2001) propose the use of the 95% quantile of the arrival rate as λ in expression (10). For the resulting staffing level, there would be a probability of 95% that less than 22% of calls will be delayed (i.e. a 95% probability of exceeding the service level). In other words, the 95% quantile of the distribution of the proportion of calls delayed will be 22%. In this section, we evaluate how close this quantile is to 22% when the 95% quantile of the arrival rate is estimated by the various density forecasting methods. With this aim, we created a discrete event simulation model of NHS Direct.

We evaluated each density forecasting method for each lead time. For each half-hour in the post-sample period, we generated a density forecast for the arrival rate, and then used its 95% quantile to set the staffing level according to expression (10). We generated arrivals from a Poisson distribution with rate defined as the centered two-point moving average of the series of actual observed arrival volume. As these observed arrival volumes are realizations of Poisson distributions, we feel the averaging more realistically reflects the unobserved arrival rates. For each call, we generated the service time from an exponential distribution with mean of 5 minutes. (Note that results were very similar when a lognormal distribution was used for the service time, and that calls that are still in service at the end of one half-hour will continue to be serviced in the next half-hour.) For each half-hour, we recorded the proportion of calls that have to wait. From the distribution of these proportions from all half-hours in the post-sample period, we recorded the 95% quantile. After repeating the whole simulation exercise 100 times, we calculated the average of this 95% quantile. The results for each lead time and method are shown in Figure 11. The ideal value for the 95% quantile of the distribution of the proportion of calls delayed is 22%.

The results in Figure 11 are reasonably consistent with the NHS Direct mean RPS results in Figure 4, with relatively poor results for the basic HWT model, the ARMA model and the SVD-type approach with AR factor models. The SVD-type approach with exponential smoothing factor models was far more competitive, although its performance depended on the lead time, with a fall in the quantile of the proportion of calls delayed as the lead time lengthened. The results are more consistent for the two HWT models with stationary level. Given that the values plotted in Figure 11 are averages from 100 simulation runs, we can test for statistical significance. The standard errors were very small (typically less than 0.2%), and from these we can conclude that the results for the two HWT models with stationary level were significantly more accurate than the SVD-type approach with exponential smoothing factor models up to about five days-ahead, and beyond this the ranking was reversed.

7. Summary and Concluding Comments

We have considered the density forecasting of call arrival volumes and rates. To enable density forecasting of both volumes and rates for intraday data, we have developed a new HWT Poisson count data model with a gamma distributed stochastic arrival rate. This involved formulating the HWT method as a state space statistical model with multiplicative seasonal and error terms, and then adapting this formulation for count data. The apparent stationary level in our arrivals series led to our development of two versions of the new model designed specifically for stationary level. We are not aware of previous studies that have considered a seasonal time series count data model for intraday data.

Our empirical analysis showed that the stationary level HWT models comfortably outperformed the basic HWT model beyond about one day ahead, and that the generalized form of the model was very competitive at earlier lead times. Although the ARMA model performed well, in terms of unconditional coverage, for prediction of the 95% quantile, this method was outperformed at all lead times, in terms of overall density forecast accuracy, by the HWT models with stationary level. We implemented Shen and Huang's (2008b) SVD-type approach with their choice of varying intercept AR factor models. This led to relatively poor results beyond one day ahead for the first two series, while for the third series, the results were similar to those of the HWT models with stationary level. Replacing the AR factor models with exponential smoothing led to noticeably better results for the first two series, but poorer results for the third. By contrast with the other series, the third series contained seasonality that was relatively constant over time, and so it would seem that the exponential smoothing factor models are more suited to series with changing seasonality. The third series also differed from the other two in that it contained observations for only weekdays, and so the intraweek seasonal cycle was less pronounced, and this may have negated the usefulness of the exponential smoothing factor models. In practice, it is not clear how to decide what form of factor models to use with the SVD-based approach. In this respect, the new HWT models with stationary level would seem to have the advantage of robustness, with consistent performance across all three series. Another important practical consideration is that the HWT count data models are relatively simple to implement.

We used a call center simulation model to demonstrate the link between density forecasting and operational decision-making. We compared methods by evaluating performance criteria based on staffing levels that were set using arrival rate density forecasts. Up to five days ahead, the best set of results were produced by the HWT models with stationary level, but beyond this the SVD-type approach with exponential smoothing factor models performed the best.

In the case of a multiskill call center, rather than trying to model arrivals for each channel, it may be preferable to base forecasts for each channel on a forecast for the total calls and predictions for the proportions of the total calls corresponding to each channel. However, this raises the issue of how best to model both the total calls and the proportions for each channel. The hierarchical forecasting literature addresses this problem. It would be interesting to see the development of exponential smoothing methods in this area. Other potential areas for research would be the inclusion of covariates, such as marketing variables, in the HWT models, and the development of quantile models for arrival volumes and rates.

Acknowledgments

We thank NHS Direct and the collaborating credit card company for providing the data. We also thank Jonathan Weinberg for making his data available. We are very grateful to an associate editor and three referees for providing very useful comments that helped greatly to improve the paper.

References

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

Akşin, O.Z., F. de Véricourt, F. Karaesmen. 2008. Call center outsourcing contract analysis and choice. *Management Science* **54** 354-368.

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

Bassamboo, A., A. Zeevi. 2009. On a data-driven method for staffing large call centers. *Operations Research* **57** 714-726.

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

Cezik, M.T., P. L’Ecuyer. 2008. Staffing multiskill call centers via linear programming and simulation. *Management Science* **2** 310-323.

Christoffersen, P.F. 1998. Evaluating interval forecasts. *International Economic Review* **39** 841-862.

Davis, R.A., W.T.M. Dunsmuir, S.B. Streett. 2003. Observation-driven models for Poisson counts. *Biometrika* **90** 777–790.

Engle, R.F., S. Manganelli. 2004. CAViaR: Conditional autoregressive value at risk by regression quantiles. *Journal of Business and Economic Statistics* **22** 367-381.

Feigin, P.D., P. Gould, G.M. Martin, R.D. Snyder. 2008. Feasible parameter regions for alternative discrete state space models. *Statistics and Probability Letters* **78** 2963-2970.

Fokianos, K. 2001. Truncated Poisson regression for time series of counts. *Scandinavian Journal of Statistics* **28** 645-659.

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

Gardner, E.S., Jr. 2006. Exponential smoothing: The state of the art - Part II. *International Journal of Forecasting* **22** 637-666.

Halfin, S., W. Whitt. 1981. Heavy-traffic limits for queues with many exponential servers. *Operations Research* **29** 567-587.

Heinen, A. 2003. Modelling time series count data: An autoregressive conditional Poisson model. Discussion Paper 2003/62, CORE, Catholic University of Louvain, Belgium.

Hyndman R.J., A.B. Koehler, J.K. Ord, R.D. Snyder. 2008. *Forecasting with exponential smoothing: The state space approach*. Springer-Verlag, Berlin, Heidelberg, Germany.

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

Jung, R.C., M. Kukuk, R. Liesenfeld. 2006. Time series of count data: Modeling, estimation and diagnostics. *Computational Statistics and Data Analysis* **51** 2350-2364.

Karlis, D., E. Xekalaki. 2005. Mixed Poisson distributions. *International Statistical Review* **73** 35-58.

McCullagh, P., J.A. Nelder. 1989. *Generalized Linear Models*. Chapman and Hall, London, 2nd edition.

Mehrotra, V., J. Fama. 2003. Call center simulation modeling: Methods, challenges, and opportunities. *Proceedings of the 2003 Winter Simulation Conference* **1** 135- 143.

Mehrotra, V., O. Ozluk, R. Saltzman. 2010. Intelligent procedures for intra-day updating of call center agent schedules. *Production and Operations Management* **19** 353-367.

Pot, A., S. Bhulai, G. Koole. 2008. A simple staffing method for multiskill call centers. *Manufacturing and Services Operations Management* **3** 421-428.

Shen, H., J.Z. Huang. 2005. Analysis of call center arrival data using singular value decomposition. *Applied Stochastic Models in Business and Industry* **21** 251-263.

Shen, H., J.Z. Huang. 2008a. Interday forecasting and intraday updating of call center arrivals. *Manufacturing and Services Operations Management* **10** 391-410.

Shen, H., J.Z. Huang. 2008b. Forecasting time series of inhomogeneous Poisson processes with application to call center workforce management. *The Annals of Applied Statistics* **2** 601-623.

Shumsky, R.A., E.J. Pinker. 2003. Gatekeepers and referrals in services. *Management Science* **49** 839-856.

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

Taylor, J.W. 2003. Short-term electricity demand forecasting using double seasonal exponential smoothing. *Journal of Operational Research Society* **54** 799-805.

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

Taylor, J.W. 2010. Exponentially weighted methods for forecasting intraday time series with multiple seasonal cycles. *International Journal of Forecasting* **26** 627–646.

Weinberg, J., L.D. Brown, J.R. Stroud. 2007. Bayesian forecasting of an inhomogeneous Poisson process with applications to call center data. *Journal of the American Statistical Association* **102** 1185-1198.

Wilks, D.S. 1995. *Statistical Methods in the Atmospheric Sciences*. Academic Press, London, UK.

Zeger, S.L. 1988. A regression model for time series of counts. *Biometrika* **75** 621-629.

Table 1: For the NHS Direct Data, Parameters of the HWT Count Data Models with Negative Binomial Distribution and Variance Modeled as Mean Divided by ψ.

	α	δ	δ'	ω	ω'	ϕ	ϕ'	ρ	ψ
HWT	0.092	0.082		0.339		1.199		-0.306	0.366
HWT with Stationary Level		0.169		0.373		1.717		-0.437	0.337
HWT with Stationary Level and Generalized		0.306	0.127	0.278	0.328	1.950	1.028	-0.204	0.369

Table 2: For the NHS Direct Data, Parameters of the HWT Count Data Models with Negative Binomial Distribution and Variance Modeled as Quadratic Function of the Mean, $(\mu_t+\varphi\mu_t^2)$.

	α	δ	δ'	ω	ω'	ϕ	ϕ'	ρ	φ
HWT	0.083	0.106		0.318		1.037		-0.273	0.00447
HWT with Stationary Level		0.222		0.353		1.478		-0.432	0.00508
HWT with Stationary Level and Generalized		0.364	0.139	0.253	0.297	1.851	0.829	-0.121	0.00434

Figure 1: Credit Card Company Half-Hourly Arrivals from Saturday 11 November 2006 to Friday 14 March 2008.

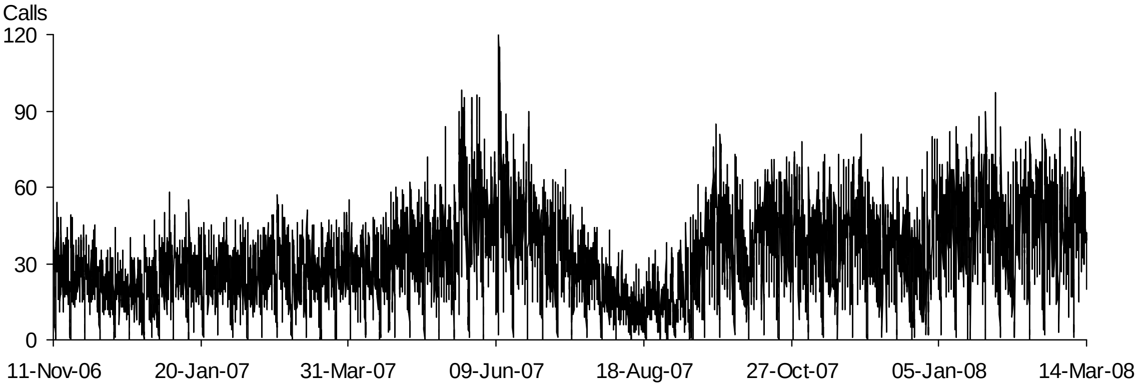


Figure 2: Credit Card Company Half-Hourly Arrivals from Saturday 23 June to Friday 20 July 2007.

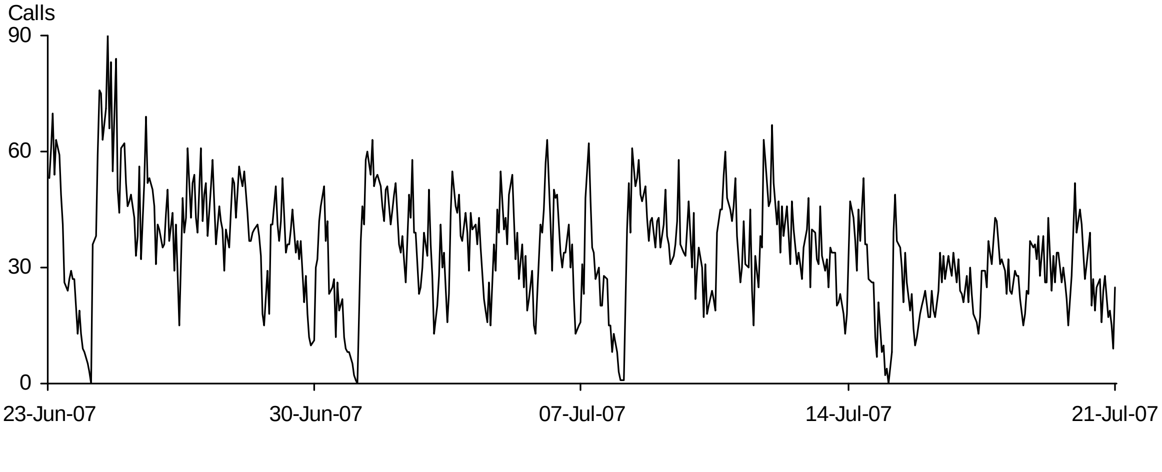


Figure 3: Average Intraday Cycle for Each Day of the Working Week for Credit Card Company Half-Hourly Arrivals.

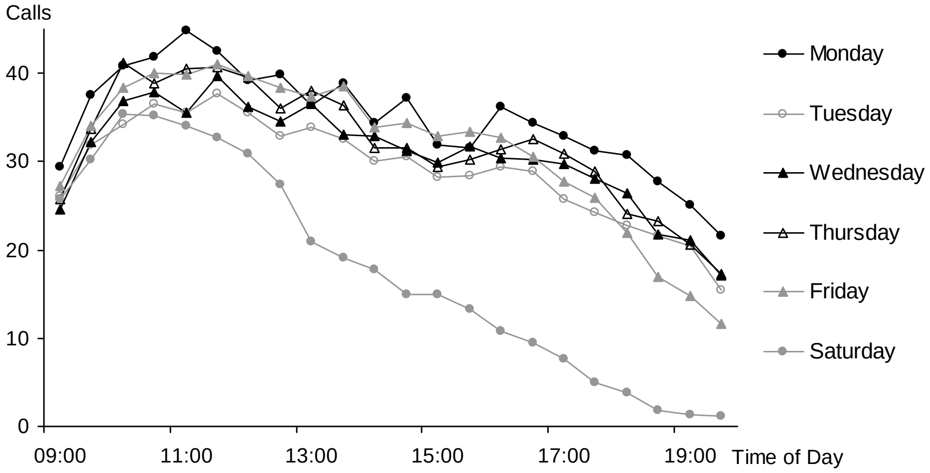


Figure 4: Mean RPS for Density Forecasting of the NHS Direct Series. Lower Values are Better.

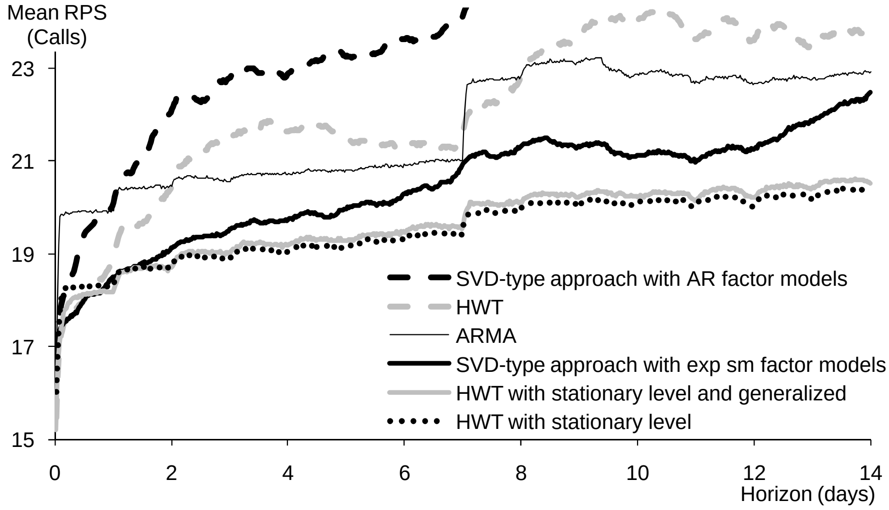


Figure 5: Hit % for Forecasting of the 95% Quantile of the NHS Direct Series. Horizontal Lines Indicate Ideal Value and 95% Acceptance Region.

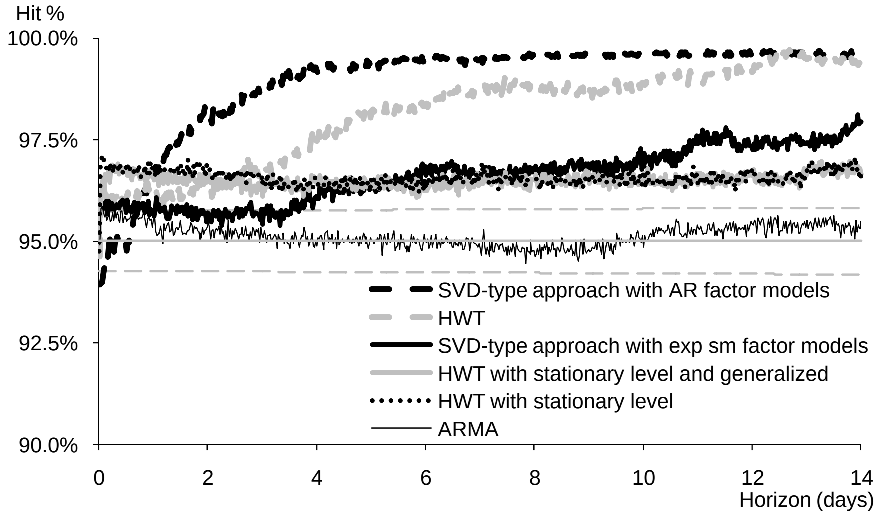


Figure 6: Comparison of Models with Different Distributions. Hit % for Forecasting of the 95% Quantile of the NHS Direct Series. Horizontal Lines Indicate Ideal Value and 95% Acceptance Region.

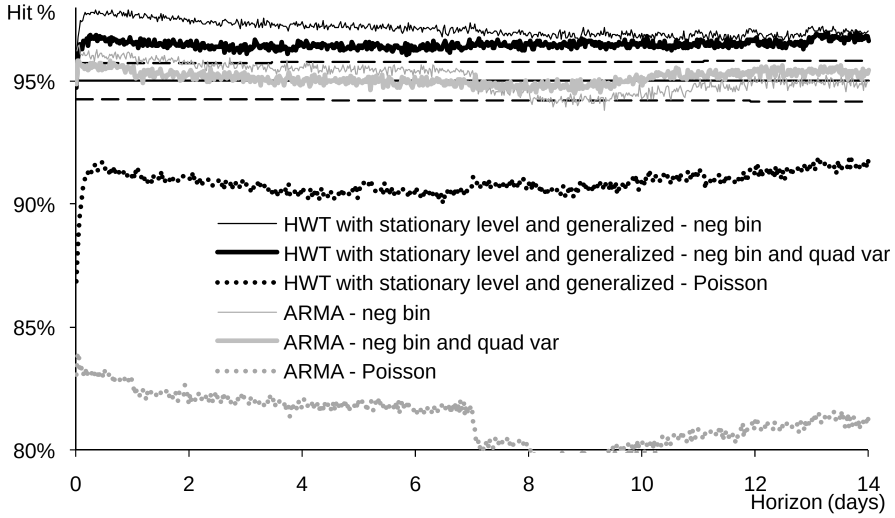


Figure 7: 90% Prediction Intervals for Call Volumes and Arrival Rates of NHS Direct Series. Forecast Origin is Final In-Sample Period. Model is HWT with Stationary Level and Generalized.

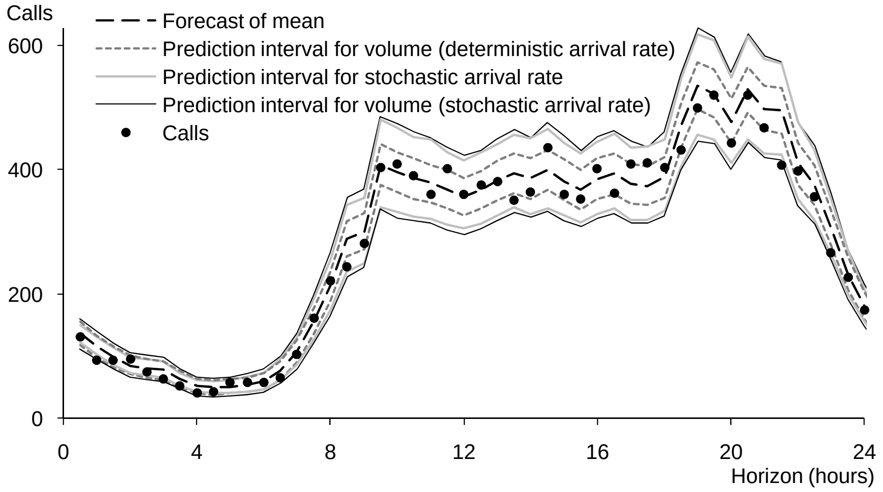


Figure 8: Mean RPS for Density Forecasting of Credit Card Company Series. Lower Values are Better.

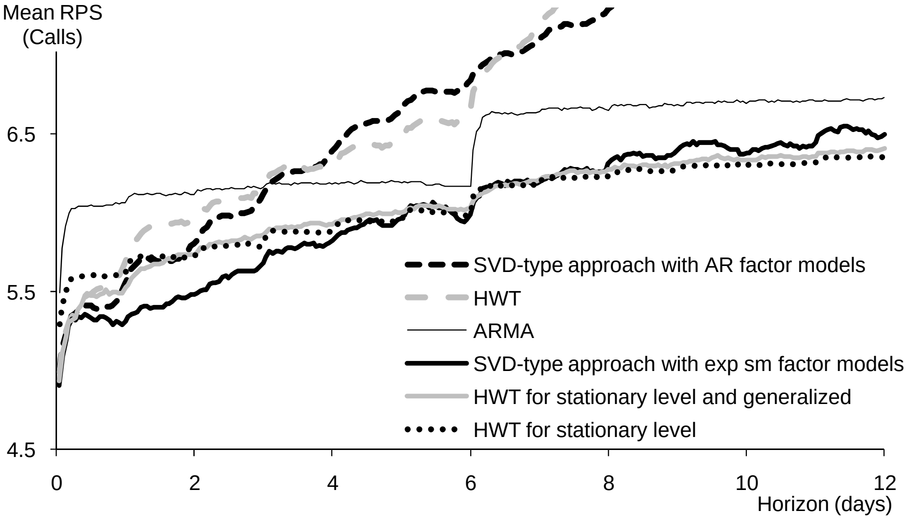


Figure 9: US Bank Half-Hourly Arrivals from Monday 3 March 2003 to Friday 24 October 2003.

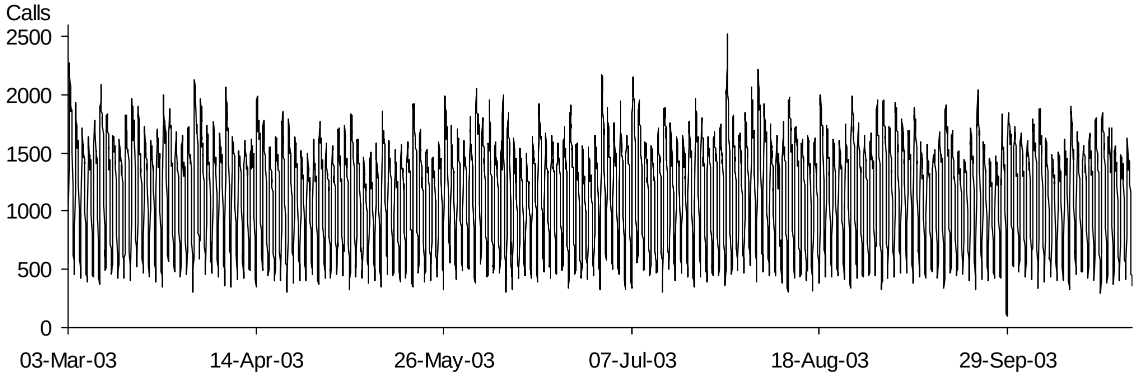


Figure 10: Mean RPS for Density Forecasting of US Bank Call Center Series. Lower Values are Better.

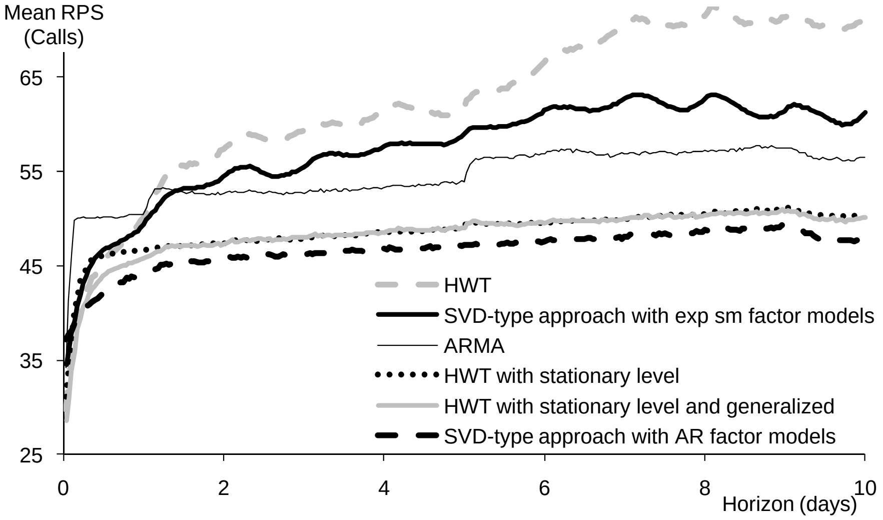
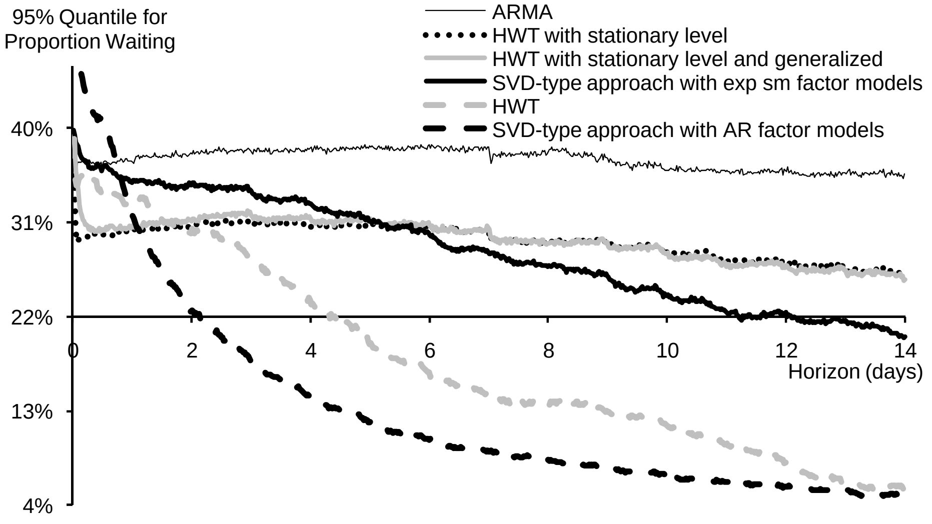


Figure 11: 95% Quantile of the Proportion of Calls Delayed in the Simulation Model for NHS Direct. (Quantile Averaged Over 100 Simulation Runs.) Ideal is 22%.


95% Quantile for
Proportion Waiting
ARMA
HWT with stationary level
HWT with stationary level and generalized
SVD-type approach with exp sm factor models
HWT
SVD-type approach with AR factor models
Horizon (days)