Exponentially Weighted Methods for Forecasting Intraday Time Series with Multiple Seasonal Cycles

James W. Taylor

Saïd Business School

University of Oxford

International Journal of Forecasting, 2010, Vol. 26, pp. 627-646.

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

Exponentially Weighted Methods for Forecasting Intraday Time Series with Multiple Seasonal Cycles

Abstract

This paper introduces five new univariate exponentially weighted methods for forecasting intraday time series that consist of both intraweek and intraday seasonal cycles. Applications of relevance include forecasting volumes of call centre arrivals, transportation, e-mail traffic and electricity load. The first method that we develop extends an exponential smoothing formulation that has been used for daily sales data, and which involves smoothing the total weekly volume and its split across the periods of the week. Two new methods are proposed that use discount weighted regression (DWR). The first uses DWR to estimate the time-varying parameters of a model with trigonometric terms. The second introduces DWR splines. We also consider a time-varying spline that uses exponential smoothing. The final new method presented involves the use of singular value decomposition followed by exponential smoothing. Empirical results are provided for a series of intraday call centre arrivals.

Keywords: Seasonality; Intraday data; Call centre arrivals; Exponential smoothing; Exponential weighting; Discount weighted regression; Regression splines; Singular value decomposition.

1. Introduction

The increasing availability of time series recorded at an intraday frequency has led to a growing interest in forecasting methods for such data. An application where this is the case is call centre management. The dramatic rise in the number and size of call centres in recent years has emphasised the need for accurate call centre arrivals forecasts, which are a key input to staffing decisions (Gans et al., 2003). With this forecasting application, there are very often no explanatory variables available, and so univariate modelling is needed for all lead times of interest. Fig. 1 presents a two-week sample of half-hourly total call arrivals at the call centres of NHS Direct, which is the 24-hour telephone helpline provided by the National Health Service (NHS) in England and Wales. The figure shows evidence of a repeating intraweek pattern of length 336 half-hourly periods, and also, at least for the weekdays, there is similarity between the intraday patterns of length 48 periods.

---------- Fig. 1 ----------

Other examples of data, exhibiting both intraday and intraweek cycles and requiring univariate forecasting methods, are electricity load, transportation counts, and hospital admissions (see, for example, Taylor, 2003; Lam et al., 2006; Gould et al, 2008). Although weather-based models are widely employed for load forecasting, univariate methods can be useful for lead times less than one day, and for regions where weather predictions are not available.

These applications require univariate methods for intraday series that are suitable for use in automated forecasting systems. Exponential smoothing is an obvious candidate, given its frequent use in automated applications in business and industry (Hyndman et al., 2008). Several exponential smoothing methods have previously been proposed for forecasting intraday series. These include Taylor's (2003) adaptation of the Holt-Winters method, and the recent developments of Gould et al. (2008) and Taylor and Snyder (2009) that aim to produce models with lower dimensionality than the Holt-Winters adaptation, in order to achieve greater parsimony and efficiency. The issue of dimension reduction is a theme running through this paper. Exponential smoothing has performed well in empirical studies with intraday data, particularly for lead times of less than approximately two days ahead (see, for example, Taylor and McSharry, 2007, and Taylor, 2008). This motivates us to consider in this paper new methods based on exponential weighting.

We present a new exponential smoothing method that involves smoothing the total weekly volume and its split across the periods of the week. Christiaanse (1971) uses exponentially weighted regression to estimate the time-varying parameters of a linear model with trigonometric terms for intraday load forecasting. We develop this idea by using Harrison and Johnston's (1984) discount weighted regression (DWR) to enable more than one discount factor to be used. Poirier (1973) introduces regression splines as a means of fitting a spline function to a nonlinear relationship between two variables. Our new proposal is to use DWR to fit a time-varying regression spline to the intraweek seasonal cycle. Harvey and Koopman (1993) model the intraweek cycle as a time-varying spline using a multiple source of error state space model. We introduce an alternative version of this that replaces this model with exponential smoothing.

Previous research has reduced the complexity of the modelling of intraday data by applying singular value decomposition (SVD), prior to time series forecasting (e.g. Shen and Huang, 2005; Taylor et al., 2006). The application of SVD proceeds by arranging the intraday data as a ($d \times m_1$) 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. We present a new approach that follows the use of SVD with exponential smoothing.

We should comment that the methods introduced in this paper will not be useful for all intraday series. For example, they would have to be adapted for use with intraday financial returns and electricity price data because such series exhibit complex structure in the volatility.

In Section 2, we describe the structure of the empirical analysis that we use to compare the forecast accuracy of the various methods. Section 3 reviews the existing exponential smoothing formulations for seasonal intraday data. Section 4 introduces a method that we term 'double seasonal total and split exponential smoothing'. Section 5 considers DWR with trigonometric terms. In Section 6, we introduce DWR splines, and a method that uses exponential smoothing to model the intraweek cycle in terms of a time-varying spline. Section 7 presents a procedure that involves the use of SVD followed by exponential smoothing. In Section 8, we summarise and provide concluding comments.

2. Structure of empirical analysis

Our empirical analysis, comparing forecast accuracy, uses a 35-week time series of half-hourly call arrivals for NHS Direct. This data is also studied by Taylor (2010). Fig. 1 shows a two-week sample of this series, and Fig. 2 presents the full series. Although this is count data, the volumes are high, and so we treat the data as values from a continuous variable. To stabilize the variance, we applied a logarithmic transformation. The series contained five public holidays. On these days, call arrivals differed greatly from the regular seasonal pattern. Forecasts for such days would typically be prepared offline. As our study is concerned with automated online prediction, we smoothed out the bank holidays using simple averaging procedures. We used a similar procedure for 162 half-hourly periods for which the observations were missing. These periods and the bank holidays were not included in our post-sample forecast evaluation. For all methods implemented in this paper, parameters were estimated once using the first 25 weeks of data. (We describe parameter optimisation in detail in Section 3.1.) The forecast origin was then rolled forward through the remaining 10-week period to produce a collection of post-sample forecasts for half-hourly lead times up to two weeks ahead. The focus of our empirical analysis is on point forecast accuracy. In the final section of the paper, we briefly consider the generation of prediction intervals from the various methods.

We evaluate point forecast accuracy using the mean absolute error (MAE), which is calculated for lead time k using the following expression:

$$\frac{1}{(3{,}360-k+1)}\sum_{t=8{,}400}^{11{,}760-k}\left|y_{t+k}-\hat{y}_t(k)\right|$$

where y_{t+k} is the actual number of call arrivals in half-hour t+k, and $\hat{y}_t(k)$ is the k step-ahead forecast from forecast origin t. Note that 25×336=8,400 is the length of the 25-week estimation sample, 35×336=11,760 is the length of the total 35-week sample, and 3,360 is the length of the post-sample period. We also evaluated the root mean squared error and mean absolute percentage error, but we do not report these results because the rankings of the methods for these measures were broadly similar to those for the MAE. Overall, our empirical results were better for the methods that involve exponential smoothing than for those that use DWR. Of the exponential smoothing methods, Taylor's adaptation of the Holt-Winters method performed particularly well up to a day ahead, but beyond this, it was matched or outperformed by several of the more recent exponential smoothing approaches, including some of the new formulations introduced in this paper.

---------- Fig. 2 ----------

3. Review of exponential smoothing methods for intraday data

3.1. HWT exponential smoothing

Holt-Winters exponential smoothing has been extended by Taylor (2003) to accommodate the intraday and intraweek cycles in intraday data. We refer to the method as 'HWT exponential smoothing', and present it in expressions (1)-(5) in error correction form.

$$\hat{y}_t(k) = l_t + d_{t-m_1+k_1} + w_{t-m_2+k_2} + \phi^k e_t \quad (1)$$

$$e_t = y_t - \left(l_{t-1} + d_{t-m_1} + w_{t-m_2}\right) \quad (2)$$

$$l_t = l_{t-1} + \alpha\, e_t \quad (3)$$

$$d_t = d_{t-m_1} + \delta\, e_t \quad (4)$$

$$w_t = w_{t-m_2} + \omega\, e_t \quad (5)$$

m_1 and m_2 are the number of periods in the day and week, respectively; l_t is the smoothed level; d_t is the seasonal index for the intraday cycle; w_t is the seasonal index for the intraweek cycle that remains after the intraday cycle is removed; α, δ and ω are smoothing parameters; and k_1=[(k-1) mod m_1]+1 and k_2=[(k-1) mod m_2]+1. The term involving the parameter ϕ, in the forecast function of expression (1), is an autoregressive (AR) adjustment for first-order residual autocorrelation. Inclusion of this term substantially improves forecast accuracy, and so is integral to the method (Taylor, 2003). A trend is not included in any of the exponential smoothing formulations in this paper because, in our empirical work, its inclusion resulted in no change in forecast accuracy. We present all exponential smoothing formulations with additive seasonality. Multiplicative seasonality formulations led to similar results.

A convenient way to generate prediction intervals for an exponential smoothing method is to express it as a statistical model. Hyndman et al. (2008) describe how innovations state space models are an attractive class of models for representing exponential smoothing methods. These models contain a single source of error. Gould et al. (2008) present advantages of these models over multiple source of error models. Parameter estimation for the innovations state space models avoids the use of the Kalman filter, which makes the models accessible to a broader audience, although it must be acknowledged that the Kalman filter is a standard tool for many practitioners. For the innovations state space models, the updating equations are identical in form to the model equations, and this enables easier interpretation and manipulation. Monte Carlo simulation provides a simple way to generate prediction intervals from innovations state space models. Taylor (2010) introduces the following innovations state space model, which captures the essence of HWT exponential smoothing:

$$y_t = l_{t-1} + d_{t-m_1} + w_{t-m_2} + \phi e_{t-1} + \varepsilon_t \tag{6}$$

$$e_t = y_t - \left(l_{t-1} + d_{t-m_1} + w_{t-m_2}\right) \tag{7}$$

$$l_t = l_{t-1} + \alpha e_t \tag{8}$$

$$d_t = d_{t-m_1} + \delta e_t \tag{9}$$

$$w_t = w_{t-m_2} + \omega e_t \tag{10}$$

where $\varepsilon_t \sim \mathrm{NID}(0,\sigma^2)$. We use this notation to indicate that the ε_t are independent and normally distributed with zero mean and constant variance σ^2. Expressions (6) and (7) can be rewritten as:

$$y_t = l_{t-1} + d_{t-m_1} + w_{t-m_2} + e_t$$

$$e_t = \phi e_{t-1} + \varepsilon_t$$

These expressions indicate that the pragmatic correction for autocorrelation leads us to a model with residual modelled as an AR process. We are not aware of previously developed exponential smoothing models with this structure.

For lead times beyond one step-ahead, the optimal point forecasts from the HWT model of expressions (6)-(10) differ slightly from those produced from the HWT method using expression (1). For example, for $k \leq m_1$, point forecasts from the HWT model are given by the following expression:

$$\hat{y}_t(k) = l_t + \alpha\phi\left(1-\phi^{k-1}\right)/\left(1-\phi\right)e_t + d_{t-m_1+k} + w_{t-m_2+k} + \phi^k e_t \tag{11}$$

This is the expectation of expression (6) in period $t+k$, conditional on information up to period t. The sum of the first two terms on the right hand side of expression (11) is the expectation of the lagged level term. The two terms are required because the state equation for the level, in expression (8), contains both the lagged level and the autocorrelated residual. The latter is ignored in the HWT method's forecast function of expression (1). Having said this, for the NHS Direct data, we found that the HWT model produced extremely similar point forecasts to those from the HWT method. The reason for this was that the optimised value for α was close to 0 and ϕ was well below 1, which led to the second term on the right hand side of expression (11) being small in value.

For all the exponential smoothing methods in this paper, we used the first three weeks of data to initialise the state variables, which for the HWT model are l_t, d_t and w_t. We constrained the parameters to lie between 0 and 1, and estimated them by minimising the sum of squared in-sample forecast errors (SSE). (An alternative is to constrain the parameters to their admissibility region. See Hyndman et al., 2008, Chapter 10.) We followed an optimisation procedure similar to that used by Engle and Manganelli (2004) for a different type of model. Our procedure involved first generating 10^5 vectors of parameters from a uniform random number generator between 0 and 1. For each of the vectors, we then evaluated the SSE. The 10 vectors that produced the lowest SSE values were used, in turn, as the initial vector in a quasi-Newton algorithm. Of the 10 resulting vectors, the one producing the lowest SSE value was chosen as the final parameter vector. (For most of the exponential smoothing methods, replacing 10^5 in this procedure with a substantially smaller number led to the same optimised values.) We obtained the following optimised values for the HWT method: α=0.026, δ=0.054, ω=0.164 and ϕ=0.372.

An advantage of HWT exponential smoothing is its relative simplicity. Due to the widespread familiarity of the standard Holt-Winters method, the HWT formulation seems reasonably natural. Furthermore, the method is straightforward to implement. For criticisms of the method, we turn to Gould et al. (2008) who point out that it is unappealing to use the same intraday cycle state variable, d_t, for each of the seven days of the week. In defence of the method, it could be argued that d_t models the commonality in the intraday cycle for each day of the week, and the differences between days is accommodated by the intraweek cycle state variable w_t. Nevertheless, this may well not be the best way to model commonality among the seven intraday cycles of the week. Furthermore, the HWT method can be viewed as being of high dimension, as it involves updating the level, plus the m_1 periods of the intraday cycle, as well as the m_2 periods of the intraweek cycle. Indeed, the method requires the updating and initialisation of $(1+m_1+m_2)$=385 terms. Having said this, we note that, for the NHS Direct series, initialising all seasonal states with a value of zero led to no loss in post-sample forecast accuracy. Regardless of this, the high dimensionality of the method seems inefficient, and so, with this in mind, in Sections 3.2 and 3.3, we review two formulations that are more parsimonious. In Section 4.2, we report empirical forecasting results for the three methods of Section 3.

3.2. IC exponential smoothing

Gould et al. (2008) present an exponential smoothing method that allows the intraday cycle for the different days of the week to be represented by different seasonal state variables. They propose a common intraday cycle for days of the week that exhibit similar patterns of demand. Due to its focus on intraday cycles, the method has been termed 'intraday cycle (IC) exponential smoothing'.

---------- Fig. 3 ----------

To decide which days of the week can be treated as having a common intraday cycle, we follow the approach of Gould et al., which relies on a plot of the type shown in Fig. 3. This figure shows, for each day of the week, the average demand for each period of the day, calculated using only the in-sample observations. From the figure, we can deduce that Saturday and Sunday should each be treated as having their own distinct cycles. The patterns for Monday morning and afternoon suggest that there should be a distinct Monday cycle. It seems reasonable to use a common intraday cycle for Tuesday, Wednesday and Thursday. The particularly low level of arrivals on Friday afternoon, and slightly higher level on Friday evening, suggests that this day should perhaps be allocated its own distinct cycle. We, therefore, had a choice between using four separate cycles, or five cycles with a distinct one being used for Friday. The lack of clarity as to how to cluster the days of the week is a potential weakness of the method. In our empirical analysis, we implemented the method based on either four or five cycles, and found that the post-sample forecasting results for the model with five distinct cycles was a little better than the version with just four. For simplicity, in the remainder of the paper, we present the formulation and results for the model with five distinct cycles. IC exponential smoothing is presented in the form of an innovations state space model in the following expressions:

$$y_t = l_{t-1} + \sum_{i=1}^{5} I_{it}\, d_{i,t-m_1} + \phi e_{t-1} + \varepsilon_t$$

$$e_t = y_t - \left(l_{t-1} + \sum_{i=1}^{5} I_{it}\, d_{i,t-m_1} \right)$$

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

$$d_{it} = d_{i,t-m_1} + \left(\sum_{j=1}^{5} \gamma_{ij} I_{jt} \right) e_t \qquad (i = 1 \text{ to } 5)$$

$$I_{it} = \begin{cases} 1 & \text{if time period } t \text{ occurs on a day of type } i \\ 0 & \text{otherwise} \end{cases}$$

where $\varepsilon_t \sim \text{NID}(0,\sigma^2)$; l_t is the smoothed level; d_{it} is the value of the intraday cycle of type i in period t (for i=1 to 5); I_{it} indicates the day type on which period t falls; and α and the γ_{ij} are smoothing parameters. The γ_{ij} can be viewed as a 5×5 matrix of parameters that enables the five types of intraday cycle to be updated at different rates. It also enables intraday cycle of type i to be updated when the current period is not on a day of type i. We included in our empirical study two forms of the method; one involved estimation of the matrix of γ_{ij} parameters with just the constraint that the parameters lie between 0 and 1, and the other involved the additional restrictions of common diagonal elements and common off-diagonal elements. The results for these two forms of the method were similar, and so when discussing forecast accuracy in Section 4.2, we present just the results for the latter 'restricted' form. By contrast with the Gould et al. model, we have included a residual AR term, as in the HWT method. For the NHS Direct data, inclusion of this term in the IC method greatly benefited forecast accuracy.

Gould et al. note that the restricted form is identical to the HWT method, provided seven distinct intraday cycles are used and the seasonal parameters for the two methods satisfy $\gamma_{ii}=\delta+\omega$ and $\gamma_{ij}=\delta$. For our NHS Direct data, the optimized parameters for the restricted form of IC exponential smoothing, with five distinct intraday cycles, were α=0.028, γ_{11}=0.185, γ_{12}=0.053, and ϕ=0.378. These α and ϕ values are close to those reported for the HWT method in Section 3.1, and $\gamma_{11}\approx\delta+\omega$ and $\gamma_{12}\approx\delta$. This suggests that the restricted IC model with five distinct cycles is similar to the HWT method, and so we should expect to achieve rather similar results from the two methods.

As we noted at the end of Section 3.1, the HWT method involves initialising and updating $(1+m_1+m_2)$=385 terms. By contrast, the IC method, with five distinct intraday cycles, involves initialising and updating the level and $5m_1$=240 seasonal terms. Therefore, the IC method can be viewed as being of lower dimension than the HWT method.

3.3. Parsimonious seasonal exponential smoothing

The parsimonious seasonal exponential smoothing method of Taylor and Snyder (2009) was motivated by the desire to find a more parsimonious and flexible method than IC exponential smoothing. They point out that an unattractive feature of the IC method is that it allows only whole days to be clustered together, such as Tuesdays, Wednesdays and Thursdays being treated as having a common intraday cycle. Their view is that it makes more sense to allow parts of days to be clustered together. For example, Fig. 3 might motivate the pattern of call arrivals for night hours to be treated as the same for all weekdays, with a different clustering being used for the other periods of the day.

Parsimonious seasonal exponential smoothing allows the clustering of any chosen periods within the intraweek cycle of length m_2 periods. In this method, the term 'season' refers to a set of periods in the intraweek cycle for which demand is assumed to be identical. The method proceeds by clustering the periods so that each belongs to one of M distinct seasons, where $M \leq m_2$. For each of the M seasons, a seasonal state s_{it} is defined and updated using exponential smoothing, which is presented in expressions (12)-(15) in the form of an innovations state space model. α and γ are smoothing parameters, and ϕ is the coefficient of a residual AR term. The level is captured within the M seasonal states, and updated each period using the parameter α. Indeed, all M seasonal states are updated each period, with the degree of updating depending on whether or not the current period t is in season i.

$$y_t = \sum_{i=1}^{M} I_{it} s_{i,t-1} + \phi e_{t-1} + \varepsilon_t \tag{12}$$

$$e_t = y_t - \sum_{i=1}^{M} I_{it} s_{i,t-1} \tag{13}$$

$$s_{it} = s_{i,t-1} + \left(\alpha + \gamma I_{it}\right) e_t \qquad \left(i = 1 \text{ to } M\right) \tag{14}$$

$$I_{it} = \begin{cases} 1 & \text{if period } t \text{ occurs in season } i \\ 0 & \text{otherwise} \end{cases} \tag{15}$$

where $\varepsilon_t \sim \text{NID}(0,\sigma^2)$.

In Section 4.2, we show the results for the method with the following two alternative judgementally chosen clusters of the m_2 periods of the week:

(i) In view of our comments in Section 3.2 regarding Fig. 3, we defined there to be five distinct days within the week, with Tuesday, Wednesday and Thursday treated as identical. As Fig. 3 shows a similar pattern for the night hours on all five weekdays, we defined the pattern for the first 12 half-hourly periods of the day to be the same on all the weekdays. This implies M=2m_1+3(m_1-12)+12=216 seasons. We obtained the following optimised parameter values: α=0.022, γ=0.173 and ϕ=0.425.

(ii) We defined the night hours to have the same pattern for all seven days of the week. This leads to M=5(m_1-12)+12=192 seasons. For this formulation, we obtained: α=0.087, γ=0.168 and ϕ=0.423.

Since *M* is the number of states in the model, both *M*=216 and *M*=192 imply a lower number of terms, that need initialising and updating, than the HWT and IC methods. This dimension reduction explains why the authors gave the method the name ‘parsimonious’. Other judgemental selections of clusters could be considered, but we have opted for simplicity. Taylor and Snyder investigated statistical approaches to clustering, including the use of hierarchical cluster analysis and k-mean cluster analysis based on the Euclidean distance between observations for different periods of the same historical week in the estimation sample. However, these procedures did not lead to improved accuracy over the better of their judgemental selections of clusters.

Taylor and Snyder show that simple extensions of their parsimonious seasonal exponential smoothing model include HWT and IC exponential smoothing as special cases. To obtain the HWT method, we define M=m_2, so that each period of the week is treated as a distinct season, and include an additional term within the parenthesis on the right hand side of expression (14). This term is the product of a new smoothing parameter and an indicator variable signalling whether or not *t* occurs at the same period of the day as season *i*.

It is worth noting that parsimonious seasonal exponential smoothing is very suited to modelling intraday series where the number of periods in each day is not the same. Taylor and Snyder present the case of a call centre that is open for a shorter duration at weekends than on weekdays.

4. Double seasonal total and split exponential smoothing

4.1. Extending total and split exponential smoothing

Total and split exponential smoothing was developed by a supermarket company for daily sales forecasting (see Taylor, 2007). It involves smoothing both the total weekly sales and the split of the total sales among the days of the week. The method is suitable for data with a single repeating seasonal cycle. Applied to an intraweek cycle of length m_2 periods, the method is expressed as:

$$\hat{y}_t(k) = T_t \, S_{t-m_2+k_2} \tag{16}$$

$$T_t = \alpha \sum_{i=0}^{m_2-1} y_{t-i} + (1-\alpha) T_{t-1} \tag{17}$$

$$S_t = \gamma \, y_t \Big/ \sum_{i=0}^{m_2-1} y_{t-i} + (1-\gamma) S_{t-m_2} \tag{18}$$

where T_t is the smoothed total across the week; S_t is the smoothed value of y_t as a proportion of the most recently observed weekly total; α and γ are smoothing parameters; and k_2=[(k-1) mod m_2]+1. The method can be viewed as a hybrid of the ratio-to-moving average seasonal decomposition procedure and Holt-Winters exponential smoothing with multiplicative seasonality and no trend. In the total and split method, the Holt-Winters smoothing of the level is replaced by smoothing of the total over the seasonal cycle. With regard to the seasonal decomposition approach, the total and split method can be viewed as replacing simple averages by exponentially weighted moving averages.

With the aim of capturing the intraday and intraweek cycles in intraday data, in expressions (19)-(23), we introduce a new 'double seasonal total and split exponential smoothing' formulation, presented as a state space model.

$$y_t = T_{t-1}\, D_{t-m_1} W_{t-m_2} + \phi\, e_{t-1} + \varepsilon_t \tag{19}$$

$$e_t = y_t - T_{t-1}\, D_{t-m_1} W_{t-m_2} \tag{20}$$

$$T_t = \alpha \sum_{i=0}^{m_2-1} y_{t-i} + (1-\alpha) T_{t-1} \tag{21}$$

$$D_t = \delta\, y_t \Big/ \sum_{i=0}^{m_1-1} y_{t-i} + (1-\delta)\, D_{t-m_1} \tag{22}$$

$$W_t = \omega\, y_t \Big/ \sum_{i=0}^{(m_2/m_1)-1} y_{t-i\times m_1} + (1-\omega) W_{t-m_2} \tag{23}$$

T_t is the smoothed weekly total; D_t is the smoothed value of y_t as a proportion of the most recently observed daily total; W_t is the smoothed value of y_t as a proportion of the sum of the observed values occurring on the same period of the day during the most recent week; and α, δ and ω are smoothing parameters. The optimised values of the parameters were: α=0.840, δ=0.060, ω=0.144 and ϕ=0.463.

4.2. Empirical evaluation of exponential smoothing methods considered so far

In Fig. 4, we show post-sample forecasting accuracy for the exponential smoothing methods considered so far, when applied to the NHS Direct data. The figure shows the MAE, calculated as described in Section 2, for lead times from one half-hour to two weeks ahead. We also implemented two simple benchmark methods. The first takes as a forecast the most recently observed value for the same half-hour of the week as the period to be predicted. The second was a 'seasonal moving average', calculated as the mean of the observations for the same half-hour of the week in each of the most recent four weeks. With our choice of scaling of the axes, only the second of these two simple methods appears in Fig. 4.

---------- Figs. 4 and 5 ----------

The figure shows the seasonal moving average benchmark method performing relatively poorly up to about four days ahead, but reasonably well beyond that. The HWT method performed the best up to about one and a half days ahead. Beyond this, it is reasonably competitive with the better of the other methods. Overall, the results for the IC method were slightly poorer than those for the HWT method. The figure shows the results for the two versions of parsimonious seasonal exponential smoothing. Assuming the pattern of night hours to be identical for just the five weekdays led to results that, overall, outperformed the HWT method beyond about two days ahead. However, assuming the pattern of night hours to be identical for all seven days of the week led to very poor results. The reason for this is difficult to understand from the plot of average intraday cycles in Fig. 3. Instead, it is useful to consider Fig. 5, which shows the average intraday cycles calculated using the logarithmically transformed data. As stated in Section 2, this transformation was applied prior to the use of each method. In Fig. 5, the difference between the patterns of night hours on weekdays and those on weekends is really quite pronounced, and this helps explain why the second version of the parsimonious seasonal exponential smoothing method performed so poorly. In summary, it seems that this method can perform well, but that its accuracy relies on a suitable clustering of the seasons. Fig. 4 shows the double seasonal total and split exponential smoothing method competing well beyond about two days ahead. Although not shown in Fig. 4, for all lead times, this method outperformed the single seasonal version of the method presented in expressions (16)-(18).

We also considered a periodic implementation of the standard Holt-Winters exponential smoothing model. This involved treating the data as 48 series of daily observations, where each series corresponded to observations for the same half-hour period of the day. We produced forecasts for each of the 48 series by applying standard Holt-Winters separately to each series. Beyond about three days ahead, this approach delivered MAE values competitive with the best of the other exponential smoothing methods. For simplicity, we do not show these results in Fig. 4.

5. DWR with trigonometric terms

5.1. DWR

In this section, we introduce discount weighted regression (DWR) before describing our use of the procedure for intraday data in Section 5.2. If the parameters of a model are believed to be changing over time, it is appealing to use an estimation approach that puts more weight on more recent information. One such approach is exponentially weighted least squares. This is considered in its recursive form by Ameen and Harrison (1984), who refer to it as 'exponentially weighted regression' (EWR). Let us consider the following model:

$$y_t = \boldsymbol{x}_t'\boldsymbol{\beta}_t + \varepsilon_t$$

where $\boldsymbol{x}_t$ is a vector of regressors; $\boldsymbol{\beta}_t$ is a vector of time-varying parameters; $\varepsilon_t \sim$ NID(0,V_t); and V_t is a potentially time-varying variance term. EWR involves the following minimisation:

$$\sum_{i=1}^{t} \lambda^{t-i} \left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right)^2$$

where λ is a discount factor; and the optimised value of the vector $\boldsymbol{\beta}$ is the estimator for $\boldsymbol{\beta}_t$ at time *t*. It is well known that the simple exponential smoothing estimator is equivalent to EWR for a model with constant term but no regressors. If a time series is trending, a linear trend term could be included in the EWR, which leads to Brown's (1963) double exponential smoothing. If the data is seasonal, trigonometric terms or dummy variables can be included. Using EWR to fit models that are functions of time is termed 'general exponential smoothing' (GES). Gardner (1985) makes the interesting point that, in contrast to Holt-Winters, GES has the feature that all seasonal terms are revised with each observation, which should make the forecasts more responsive to changing seasonal patterns. Christiaanse (1971) uses GES with trigonometric terms for intraday electricity load forecasting. He investigates four different subjectively chosen values of λ and four different choices of trigonometric terms. In this paper, we develop this method in the light of methodological and computational advances that have taken place since Christiaanse's study.

Ameen and Harrison (1984) write that the use of just one parameter is a notable disadvantage of EWR. They suggest that it may well be appropriate for information on different components of a time series to be discounted at different rates. The use of a single discount factor in EWR contrasts with the exponential smoothing methods of Sections 3 and 4, which use two or three smoothing parameters. Ameen and Harrison and Harrison and Johnston (1984) introduce ‘discount weighted regression’ (DWR) to enable a different discount factor to be used for each parameter. DWR is an adaptation of the updating equations for recursive least squares. The DWR updating equations are presented in expressions (25)-(27) for the model of expression (24). These expressions are essentially those of the Kalman filter applied to a regression model with the incorporation of exponential weighting. (The link between EWR and Kalman filtering is considered by Li, 2008.)

$$y_t = \boldsymbol{x}_t'\boldsymbol{\beta}_t + \varepsilon_t \tag{24}$$

$$\hat{\boldsymbol{\beta}}_t = \hat{\boldsymbol{\beta}}_{t-1} + \boldsymbol{Q}_t^{-1}\boldsymbol{x}_t u_t \tag{25}$$

$$u_t = y_t - \boldsymbol{x}_t'\hat{\boldsymbol{\beta}}_t \tag{26}$$

$$\boldsymbol{Q}_t = \boldsymbol{\lambda}^{\frac{1}{2}}\boldsymbol{Q}_{t-1}\boldsymbol{\lambda}^{\frac{1}{2}} + \boldsymbol{x}_t\boldsymbol{x}_t' \tag{27}$$

ε_t ~ NID(0,V_t), where V_t is a potentially time-varying variance term; and $\boldsymbol{\lambda}^{\frac{1}{2}} = \mathrm{diag}\left(\lambda_1^{\frac{1}{2}}, \lambda_2^{\frac{1}{2}}, \ldots, \lambda_M^{\frac{1}{2}}\right)$, where λ_i can be interpreted as the discount factor for information concerning the *i*th parameter ($0 \leq \lambda_i \leq 1$), and M is the dimension of the parameter vector $\boldsymbol{\beta}_t$. We used OLS regression applied to the first three weeks of data to estimate $\boldsymbol{\beta}_0$, and to express strong confidence in this initialisation, we set $\boldsymbol{Q}_0 = \boldsymbol{I}/10^6$, where $\boldsymbol{I}$ is an identity matrix. In our empirical work, we used the formulae of Harrison and Johnston (1984), which avoids the use of matrix inversion.

5.2. DWR with trigonometric terms for intraday data

For our NHS Direct data, we fitted the following model using DWR:

$$y_t = b_0 + \sum_{i\in S_1}\left(b_{1i}\sin\left(2i\pi\tfrac{d(t)}{m_1}\right) + b_{2i}\cos\left(2i\pi\tfrac{d(t)}{m_1}\right)\right) + \sum_{i\in S_2}\left(b_{3i}\sin\left(2i\pi\tfrac{w(t)}{m_2}\right) + b_{4i}\cos\left(2i\pi\tfrac{w(t)}{m_2}\right)\right) + e_t \tag{28}$$

where b_0 and the b_{ji} are model parameters; $d(t)$ and $w(t)$ are repeating step functions defined so that $d(t)$ counts from 1 to m_1 within each day, and $w(t)$ counts from 1 to m_2 within each week; and S_1 and S_2 are chosen sets of positive integers, which must be less than $m_1/2$ and $m_2/2$, respectively. When selecting these sets, the same trigonometric term must not appear in both summations.

We felt it was impractical to allow a different DWR discount factor for each parameter in expression (28), so we considered two different specifications. We refer to the first as ‘EWR’ because it involved the use of a single, common discount factor for all parameters in expression (28). We refer to the second specification as 'DWR'. It allows three different discount factors: λ_1 for the constant term, a common discount factor λ_2 for all coefficients in the first summation (corresponding to harmonics of the intraday cycle), and a common discount factor λ_3 for all coefficients in the second summation (corresponding to the remaining harmonics of the intraweek cycle).

The residuals from the various specifications tended to exhibit autocorrelation. In view of this, after estimating the parameters of expression (28) using DWR or EWR, we fitted an AR(1) model to the residuals. When forecasting, we simply added the AR model forecasts to those from the model of expression (28). A reasonable alternative way to model the residual autocorrelation would be to include an AR(1) term alongside the trigonometric terms in the regression model. (A discount factor of one would be appropriate for this term in the DWR estimation.) This approach would lead to the AR parameter being estimated recursively. By avoiding this recursive estimation, our modelling of the residual autocorrelation was more analogous to that in the exponential smoothing methods.

We optimised the discount factors λ_i and the residual AR model parameter ϕ in the same procedure by minimising the sum of squared in-sample one step-ahead forecast errors. The optimisation was performed by evaluating this objective function for many candidate values of the λ_i and ϕ sampled randomly over a suitable range of values established by initial experimentation. We found that candidate values greater than 0.9 were suitable for the DWR discount factors. This optimisation of the parameters is similar to the first stage of the procedure described in Section 3.1 to optimise the parameters of the exponential smoothing methods.

An important aspect of the method is the choice of trigonometric terms. We experimented with S_1=S_2=$\{1,2,\ldots,6\}$, but this led to poor post-sample forecast accuracy relative to much larger sets. For example, setting S_1 to be an empty set and S_2=$\{1,2,\ldots,167\}$ led to far better results. To select trigonometric terms based on only the estimation sample, we used the signal coherence procedure that was employed by Brooks and Hinich (2006) in their analysis of intraday exchange rate data. They describe the method as honing in on the frequency components of the Fourier transforms of the signal that are the most stable over time. We used the same signal coherence "floor value" of 0.45 that was used by Foster et al. (2008), who apply the signal coherence measure to investigate the cyclical structure within half-hourly electricity load and price data. For our data, the procedure selected eight integers (ranging between 1 and 11) for set S_1, and 57 integers (ranging between 1 and 161) for set S_2. Using these in the model of expression (28), we implemented EWR and DWR.

With DWR, we obtained the following values for the discount factors and the AR parameter: λ_1=0.922, λ_2=0.993, λ_3=0.998 and ϕ=0.190. These high discount factors seem reasonable when one considers the high frequency nature of the data. A discount factor of 0.993 corresponds to a half-life of approximately 99 half-hour periods, which is approximately two days, and a discount factor of 0.998 corresponds to a half-life of approximately 346 half-hour periods, which is approximately one week. The discount factor for the constant term is notably lower than the two discount factors for the coefficients of the trigonometric terms. This indicates faster evolution in the level of the data than in the seasonality. For EWR, we obtained: λ=0.998 and ϕ=0.519. This discount factor is similar to the values optimised for the two DWR discount factors for the coefficients of the trigonometric terms, but the value of ϕ is notably larger than in the DWR case, which enables a faster changing level, and thus compensates for the fact that the constant term does not have its own discount factor.

In a recent study, De Livera and Hyndman (2009) present an exponential smoothing approach that also uses a trigonometric formulation for the seasonality in intraday data. They allow the coefficients of the trigonometric terms to be time-varying by specifying the coefficients as exponential smoothing state variables that are updated with each new observation.

5.3. Empirical evaluation of the DWR methods

In Fig. 6, we present post-sample results for the EWR and DWR methods with trigonometric terms selected using the signal coherence method. As a benchmark, we show the results of HWT exponential smoothing. We also implemented a version of DWR with no discounting (λ_i=1 for all i) and the AR residual adjustment included, but its results were so poor that the MAE values do not appear on the graph. The figure shows EWR competing well with DWR beyond about two days ahead. DWR is much better for the early lead times because a different rate is allowed for the constant, which amounts to separate smoothing of the level. Overall, HWT exponential smoothing outperformed both the EWR and DWR methods.

---------- Fig. 6 ----------

6. Time-varying splines

6.1. Regression splines

In this section, we present an overview of regression splines before introducing two new forecasting methods in Sections 6.2 and 6.3. In essence, these methods involve the fitting of a spline to the intraweek seasonal cycle in intraday data.

A cubic interpolating spline is a function in (*x*,*y*) space that consists of cubic polynomials smoothly joined at coordinates (x_i^*, s_i^*). The points x_i^* are referred to as 'knots'. The smoothness of the spline function is ensured by constraining the function and its first and second derivatives to be continuous at the knots. If we make assumptions regarding the spline at its end points, then the value of the spline at a point *x* between these end points can be analytically derived. Indeed, the value is given by a linear function *f* of the value of the spline at the knots (see, for example, Poirier, 1973):

$$f(x) = \boldsymbol{w}'\boldsymbol{s}^* \tag{29}$$

where $\boldsymbol{s}^*$ is a vector containing the s_i^*; and ***w*** is a vector that can be analytically calculated from the location of the knots, the distance between knots, and the value of *x*.

Poirier (1973) addresses the situation where it is assumed that the relationship between two variables, y_t and x_t, can be described by a spline function plus error, e_t. In view of the structure of expression (29), Poirier proposes the following model, which he terms a 'regression spline':

$$y_t = \boldsymbol{w}_t'\boldsymbol{s}^* + e_t \tag{30}$$

Before the model can be estimated, the locations of the knots must be selected. In this regression model, the data vector is $\boldsymbol{w}_t$, as it is a known function of x_t, and the parameter vector is $\boldsymbol{s}^*$. This parameter vector contains the values of the spline function at the selected knots. As the model of expression (30) is linear, formulae from OLS regression can be used to estimate $\boldsymbol{s}^*$.

We fitted an OLS regression spline to the repeating intraweek cycle in the 25-week estimation sample of the NHS Direct data. With this application, the variable x_t is the period of the week, which ranges from 1 to 336, and the intraweek cycle is assumed to be the same for each week in the series.

We relax this assumption in Sections 6.2 and 6.3, where we consider time-varying regression splines. As noted by Harvey and Koopman (1993) for their method, selecting the knot locations is the main practical problem for the implementation of a regression spline. They explain that sections of the seasonal cycle displaying sharp peaks require relatively more knots than sections with less variability. Following Harvey and Koopman, we selected the knots subjectively, and we set the number of knots to be a little less than half the number of periods within the cycle. We chose the following 18 knots: 3am, 6am, 7am, 8am, 9am, 10am, 11am, midday, 1pm, 2pm, 3pm, 4pm, 5pm, 6pm, 7pm, 8pm, 10pm, 11.30pm. With 18 knots for each of the seven days, and an additional one to mark the start of the week, this spline consisted of (7×18)+1=127 knots. We constrained the spline function at the first and last knots of the week to be equal in value and in first and second derivatives.

In Sections 3.2 and 3.3, we presented exponential smoothing formulations that had been designed with the aim of achieving dimension reduction. For example, the parsimonious seasonal exponential smoothing method enabled the night hours to be the same for weekdays. Similar constraints can be imposed on the regression spline to enable dimension reduction. Harvey and Koopman (1993, Section 3.3) describe how to impose constraints to force the value of the spline function to be identical at selected knots. The constraints are specified using a selection matrix $\boldsymbol{C}$ consisting of 0's and 1's. Using this, we write vector $\boldsymbol{s}^*$ in terms of $\bar{\boldsymbol{s}}^*$, a lower dimension vector: $\boldsymbol{s}^* = \boldsymbol{C}\,\bar{\boldsymbol{s}}^*$. If we also define vector $\bar{\boldsymbol{w}}_t' = \boldsymbol{w}_t'\boldsymbol{C}$, we can rewrite expression (30) as:

$$y_t = \bar{\boldsymbol{w}}_t'\bar{\boldsymbol{s}}^* + e_t \qquad (31)$$

In view of our comments in Sections 3.2 and 3.3 regarding Fig. 3, we chose to specify the selection matrix $\boldsymbol{C}$ so that the spline function at the knots on Tuesday, Wednesday and Thursday are identical, and the spline function at the knots at 3am and 6am are identical for all five weekdays. This led to the new parameter vector $\bar{\boldsymbol{s}}^*$ being of dimension (2×18)+(3×16)+2+1=87. In other words, the spline consisted of 87 distinct knots. In essence, the dimensionality of the problem has been reduced because instead of modelling the intraweek cycle at all 336 periods of the week, we are now focussing on just 87 periods of the week.

6.2. DWR spline

As the intraweek cycle in the NHS Direct series is not constant, it came as no surprise that using the OLS regression spline as a forecast of future periods led to poor forecast accuracy. In Section 5, we presented DWR as a simple approach to allow time-variation in the parameters of a linear model. In this section, we propose the use of DWR to estimate a time-varying regression spline. This amounts to using DWR to estimate the parameter vector $\bar{\boldsymbol{s}}^*$ of the model in expression (31). In this context, we have a time-varying parameter vector $\bar{\boldsymbol{s}}_t^*$, and so we rewrite the model as in expression (32). DWR enables more weight to be put on more recent observations, and re-estimation of the time-varying regression spline with each new observation.

$$y_t = \bar{\boldsymbol{w}}_t' \bar{\boldsymbol{s}}_t^* + e_t \tag{32}$$

In the application of DWR with trigonometric terms in Section 5.2, there was a substantial difference between the value of the discount factor for the constant term and the values of the two discount factors for the coefficients of the seasonal terms. The inclusion of the constant term enabled the level of the series to be modelled with a different discount rate to that of the seasonality. The regression spline models considered so far in Section 6 do not include constant terms. To enable separate modelling of the level, we defined the first knot of each intraweek spline function to be a 'base knot', with the spline function at all other knots defined as the spline function at this base value plus a positive or negative increment. (This is equivalent to setting the value of the spline function at the first knot to be zero and including a constant term in the regression model.) When using DWR, time-variation in the regression spline at this base knot will capture time-variation in the level of the series, and the value of the regression spline at all other knots will capture time-variation in the seasonality. We implemented the base knot by defining a new matrix $\boldsymbol{G}$ as being equal to an identity matrix with first column altered so that it consists entirely of ones. With this matrix, we can write $\bar{\boldsymbol{s}}_t^*$ in terms of $\tilde{\boldsymbol{s}}_t^*$, a new vector: $\bar{\boldsymbol{s}}_t^* = \boldsymbol{G}\tilde{\boldsymbol{s}}_t^*$. The first entry of the new parameter vector, $\tilde{\boldsymbol{s}}_t^*$, is the value of the spline function at the base knot. The values of the spline function at the other distinct knots are given by the other entries of $\tilde{\boldsymbol{s}}_t^*$ plus the value of the spline function at the base knot. If we define $\tilde{\boldsymbol{w}}_t' = \bar{\boldsymbol{w}}_t'\boldsymbol{G}$, we can rewrite expression (32) as:

$$y_t = \tilde{\boldsymbol{w}}_t' \tilde{\boldsymbol{s}}_t^* + e_t$$

As in the DWR approach with trigonometric terms of Section 5.2, we included the residual AR term and considered two different discount factor specifications. The first relies on a common discount factor for the regression spline at all knots, and so we refer to this as an 'EWR regression spline'. For this specification, we obtained λ=0.995 and ϕ=0.341. The second allows three different discount factors: λ_1 for the regression spline at the base knot, λ_2 for the knots occurring during the night (at 3am and 6am), and λ_3 for the remaining knots. Using the same optimisation procedure described in Section 5.2, we obtained λ_1=0.943, λ_2=0.996, λ_3=0.995 and ϕ=0.195. As in Section 5.2, the value of λ_1 is noticeably lower than the values of λ_2 and λ_3. This is consistent with our interpretation that the use of the base knot in the regression spline provides a similar service to the constant term in the DWR trigonometric model. These terms capture the level in the data, allowing the other terms to focus on the modelling of the seasonality.

6.3. Spline-based exponential smoothing

Harvey and Koopman (1993) model the intraweek seasonal cycle in electricity load data as a time-varying spline using a state space model with a separate state defined for the spline function at each knot. A multiple source of error state space model is used with values of the spline function at the knots specified as following a multivariate random walk. In this section, we adopt the approach of Harvey and Koopman, but instead of their model, we use the exponential smoothing model of expressions (33)-(37), which is an innovations state space model. As we discussed briefly in Section 3.1, a number of advantages have been put forward in favour of models of this type.

$$y_t = \overline{\boldsymbol{w}}_t' \overline{\boldsymbol{s}}_{t\text{-}1}^* + \phi e_{t-1} + \varepsilon_t \tag{33}$$

$$e_t = y_t - \overline{\boldsymbol{w}}_t' \overline{\boldsymbol{s}}_{t\text{-}1}^* \tag{34}$$

$$\overline{s}_{it}^* = \overline{s}_{i,t-1}^* + \left(\alpha + \kappa I_{it}^{knot} + \eta I_{it}^{nearby}\right) e_t \qquad \left(i = 1 \text{ to } M\right) \tag{35}$$

$$I_{it}^{knot} = \begin{cases} 1 & \text{if period } t \text{ is the location of knot } i \\ 0 & \text{otherwise} \end{cases} \tag{36}$$

$$I_{it}^{nearby} = \begin{cases} 1 & \text{if period } t \text{ is between knots } (i-1) \text{ and } (i+1) \\ 0 & \text{otherwise} \end{cases} \tag{37}$$

where $\varepsilon_t \sim \mathrm{NID}(0,\sigma^2)$; $\bar{\boldsymbol{w}}'_t$ has the same definition as in Section 6.1; $\bar{\boldsymbol{s}}^*_t$ is a state vector with ith element $\bar{s}^*_{it}$, which represents the value of the spline function at the ith knot in period t; α, κ and η are smoothing parameters; ϕ is the parameter of an AR residual term; and M is the number of distinct knots defined for the intraweek cycle. In our empirical analysis, we used the set of M=87 distinct knots described in Section 6.1. We did not specify a base knot, as in Section 6.2, as this is useful only in the DWR context. Expression (35) is similar in structure to the state equations of the parsimonious seasonal exponential smoothing model of Section 3.3. The two indicator variables enable the state variable for the spline function at the ith knot to be updated only if the current period t falls at or nearby the ith knot. We obtained the following optimised parameter values: α=0.039, κ=0.092, η=0.049, ϕ=0.320. The values for κ and η indicate that the state variable for the spline function at the ith knot is adjusted noticeably more if period t falls at the ith knot than if it falls nearby.

6.4. Empirical evaluation of spline methods

Fig. 7 shows post-sample forecasting performance for the EWR and DWR spline methods, and for the spline-based exponential smoothing method. As a benchmark, we also include in the figure the results of HWT exponential smoothing. The relative performances of the EWR and DWR methods have similarities with the relative performances of EWR and DWR in Section 5.3, where trigonometric terms were used. The DWR method is superior only for the early lead times. Interestingly, the EWR and DWR spline methods are outperformed, at almost all lead times, by the spline-based exponential smoothing method. However, this exponential smoothing method was, in turn, outperformed by HWT exponential smoothing, although for many lead times the superiority of the HWT method is not substantial. We note that these results are only illustrative because the spline methods could be improved by using a different choice for the number and locations of the knots, selected either subjectively or perhaps using some statistical guidance, such as cross-validation.

---------- Fig. 7 ----------

7. SVD with exponential smoothing

7.1. Applying SVD to a time series of intraday observations

In this section, we describe how singular value decomposition (SVD) can be used as the basis for forecasting intraday time series. In Section 7.2, we present a new SVD-based forecasting method.

In Section 3.3, we described how the use of parsimonious seasonal exponential smoothing requires the clustering of periods of the intraweek cycle that can be treated as identical. However, it could be argued that this is unappealing because, although some periods may be similar, few are identical. As an alternative to clustering, in this section, we use another dimension reducing technique from multivariate data analysis. We investigate the use of singular value decomposition (SVD), which is an approach similar to principal component analysis (PCA). SVD enables the reduction of the dimension of a multivariate dataset, where variables are highly correlated, to a smaller set of variables that are linear combinations of the original variables. The new variables are uncorrelated and explain most of the variation in the data. In our context, each of the original correlated variables corresponds to a particular period of the intraday cycle. Instead of having to model each intraday period, the forecasting task is simplified to one that requires the modelling of just the reduced set of uncorrelated variables. This is the approach adopted by Shen and Huang (2005, 2008a, 2008b) for intraday call centre data, and by Taylor et al. (2006) for intraday electricity load data.

In these studies, the application of SVD to a series of intraday observations proceeds by arranging the data as a ($d \times m_1$) matrix $\boldsymbol{Y}$, where d is the number of days in the estimation sample. Each column of $\boldsymbol{Y}$ has observations for a particular period of the day. The application of SVD to this matrix requires that $d \geq m_1$. The SVD of $\boldsymbol{Y}$ yields $\boldsymbol{Y}'\boldsymbol{Y} = \boldsymbol{VSV}'$, where $\boldsymbol{V}$ is an ($m_1 \times m_1$) orthogonal matrix, and $\boldsymbol{S}$ is a positive definite diagonal matrix. The positive square roots of the entries in $\boldsymbol{S}$ are known as the singular values of $\boldsymbol{Y}'\boldsymbol{Y}$. These singular values are typically arranged in decreasing order. The columns of $\boldsymbol{V}$ are orthogonal basis functions, which Shen and Huang (2008a) refer to as ‘intraday feature vectors’. The daily profiles (rows of $\boldsymbol{Y}$) can be projected onto the basis functions to give what Shen and Huang call ‘interday feature series’. Collecting these series as columns of a ($d \times m_1$) matrix $\boldsymbol{P}$, we have $\boldsymbol{P} = \boldsymbol{YV}$. Note that each interday feature series (column of $\boldsymbol{P}$) corresponds to a single intraday feature vector (column of $\boldsymbol{V}$). Dimension reduction occurs by keeping only the first k pairs of feature vectors and series corresponding to the largest k singular values. Shen and Huang (2008a) write that the singular values provide a natural ordering of the importance of the feature vectors and series in describing the original dataset. Time series forecasting methods can then be applied to each of the interday feature series, under the assumption that the intraday feature vectors remain constant. The resulting predictions for the interday feature series are then projected back onto the $\boldsymbol{Y}$ space to deliver forecasts for the original variable. Shen and Huang (2005, 2008a, 2008b) use a form of AR model for the interday feature series, while Taylor et al. (2006) use regression modelling with dummy variables to model the seasonality.

Let us briefly consider the link between PCA and SVD. PCA involves the application of SVD to a column-centred matrix. In the context of our data matrix $\boldsymbol{Y}$, this leads to the matrix decomposition being applied to the covariance matrix of $\boldsymbol{Y}$, rather than the matrix $\boldsymbol{Y}'\boldsymbol{Y}$. In the PCA literature, the interday feature series are known as the principal components, and the squared singular values are proportional to the variances of these principal components. By selecting the first k principal components, we are capturing the k most important contributors to the variance in the original dataset.

Following Shen and Huang (2008a), in Figs. 8-11, we plot intraday feature vectors and interday feature series for the NHS Direct data, with SVD performed using the estimation sample of 175 days. The first intraday feature vector seems to show the intraday cycle averaged across the seven days of the week. The corresponding interday feature series shows a seven-day seasonal cycle with higher (log) calls at the weekends. Figs. 3 and 5 show that the average intraday patterns for Saturday and Sunday are much larger around late morning and early afternoon. This is, at least partly, captured in the second intraday feature vector. The weekend peaks in the second interday feature series confirms that this second intraday feature vector is particularly relevant for explaining the weekend patterns.

---------- Figs. 8-11 ----------

The choice of k could be made based on a scree plot, which is often used in PCA, or perhaps cross-validation. In analysing intraday data for only weekdays, Shen and Huang (2008a) evaluate k=1 to 4, and write that k=5 may be needed as they identify different intraday patterns on each of their five weekdays. This suggests $k\leq 7$ for our series, as it includes all seven days of the week. In fact, in

Section 3.2, we concluded that there are just five distinct intraday cycles among the seven days of the week, which suggests $k \leq 5$. Such low values of k represent a considerable dimension reduction.

7.2. SVD-based exponential smoothing

In this section, we present a new exponential smoothing model formulated in terms of the k interday feature series, which are the first k columns of the matrix $\boldsymbol{P}$. The approach involves modelling these interday feature series, and then projecting them onto $\boldsymbol{Y}$ space to produce forecasts. The modelling produces updated estimates of the k interday feature series, as each new observation is received. This amounts to producing updated estimates of the k columns of matrix $\boldsymbol{P}$. In view of this, the state variables, in our exponential smoothing model, are defined as the values in period t of the first k interday feature series. In order to capture the intraweek cycle, we treat the NHS Direct data as consisting of the five distinct intraday cycle types identified in Section 3.2. The model is presented in the following expressions:

$$y_t = \sum_{i=1}^{5} I_{it}\, \boldsymbol{p}_{t-1}^{(i)}\, \tilde{\boldsymbol{v}}_{[t \bmod m_1]}{}' + \phi e_{t-1} + \varepsilon_t \tag{38}$$

$$e_t = y_t - \sum_{i=1}^{5} I_{it}\, \boldsymbol{p}_{t-1}^{(i)}\, \tilde{\boldsymbol{v}}_{[t \bmod m_1]}{}' \tag{39}$$

$$\boldsymbol{p}_t^{(i)} = \boldsymbol{p}_{t-1}^{(i)} + \left(\alpha \mathbf{1}_{m_1} \tilde{\boldsymbol{V}} + (\delta + \omega I_{it}) \tilde{\boldsymbol{v}}_{[t \bmod m_1]}\right) e_t \qquad (i = 1 \text{ to } 5) \tag{40}$$

$$I_{it} = \begin{cases} 1 & \text{if period } t \text{ occurs on a day of type } i \\ 0 & \text{otherwise} \end{cases} \tag{41}$$

where $\varepsilon_t \sim \text{NID}(0,\sigma^2)$; $\boldsymbol{p}_t^{(i)}$ is a $(1 \times k)$ state vector representing the values in period t of the first k interday feature series for a day of type i (for i=1 to 5); $\tilde{\boldsymbol{V}}$ is a $(m_1 \times k)$ matrix containing the first k intraday feature vectors (columns of $\boldsymbol{V}$); $\tilde{\boldsymbol{v}}_{[t \bmod m_1]}$ is the $[t \bmod m_1]$ row of $\tilde{\boldsymbol{V}}$, which corresponds to the period of the day on which t falls; $\mathbf{1}_{m_1}$ is a $(1 \times m_1)$ vector of 1's; α, δ and ω are smoothing parameters; and ϕ is the parameter of an AR residual term. Note that the matrix $\tilde{\boldsymbol{V}}$ and the vectors $\tilde{\boldsymbol{v}}_{[t \bmod m_1]}$ are derived by the SVD, and are, therefore, known prior to the exponential smoothing modelling. By contrast, the vectors $\boldsymbol{p}_t^{(i)}$ are estimated by the model.

The observation equation (38) shows the state vector projected onto $\boldsymbol{Y}$ space using the intraday feature vectors. The state vector $\boldsymbol{p}_t^{(i)}$ is updated in expression (40), which is of a similar form to the state equations of the parsimonious seasonal exponential smoothing model of Section 3.3. To understand expression (40), it is helpful to recall from the previous section that $\boldsymbol{P}=\boldsymbol{YV}$, which reminds us that a value in $\boldsymbol{Y}$ space can be projected onto $\boldsymbol{P}$ space using $\boldsymbol{V}$. Since the error e_t is a value in $\boldsymbol{Y}$ space, we must use $\widetilde{\boldsymbol{V}}$ to transform e_t before using it to update $\boldsymbol{p}_t^{(i)}$. The smoothing parameters α, δ and ω play a similar role to the smoothing parameters with the same names in the HWT exponential smoothing model of Section 3.1. The term involving α updates the level of $\boldsymbol{p}_t^{(i)}$; the term involving δ updates the elements of $\boldsymbol{p}_t^{(i)}$ to differing degrees depending on their relationship to the period of the day on which t falls; and the term involving ω has a similar effect to that of the term involving δ, except that ω only has an effect on $\boldsymbol{p}_t^{(i)}$ if period t falls on a day of type i. We obtained the following optimised parameter values: α=0.0304 δ=0.296 ω=0.241 ϕ=0.275.

To gain a little more insight into the state equations in expression (40), consider the case where $\boldsymbol{V}$ is the identity matrix, and we impose no dimension reduction by setting $k=m_1$. Expression (40) then takes the form of a vector representation of the state equations of a version of the parsimonious seasonal exponential smoothing model with a distinct season defined for each period of the day, on each of five different day types, so that $M=5m_1$ in the model of expressions (12)-(15).

The methods of Shen and Huang (2005) and Taylor et al. (2006) first perform SVD, and then apply simple time series models to the first k interday feature series. These are daily series, and they can only be updated after the observations for a complete day have been observed. As Shen and Huang (2008a, Section 3.3) note, this implies that a forecast for day $n+h$ can only be produced at the end of day n. They explain that, as calls arrive during day $n+1$, a manager may want to update the forecast for the remainder of the day using the observations from the earlier part of that day. This intraday forecast updating is very useful because it allows dynamic intraday updating of the deployment of call centre agents (Hur et al., 2004). Shen and Huang (2008a) extend their approach to allow intraday forecast updating. This extension involves forecasting the interday feature series for day $n+1$ from a combination of the day-ahead forecast from a model applied to the daily interday feature series up to day *n*, and a prediction based on a second model that uses just data observed so far within day *n*+1. We would suggest that our SVD-based exponential smoothing approach is appealing because it involves just the one model, which can be used directly to produce forecasts for all lead times and from all forecast origins. However, the choice between methods will ultimately be decided by forecast accuracy, and by how easily the method can be understood and implemented. Shen and Huang (2008b) further develop their approach to enable the SVD to be updated with each new observation, and to accommodate a Poisson assumption to allow density forecasting of both the arrival volume and arrival rate. This density forecasting is also the focus of the Poisson HWT models of Taylor (2010).

7.3. Empirical evaluation of SVD with exponential smoothing

Fig. 12 presents post-sample results for our SVD approach, and HWT exponential smoothing. The results correspond to different values for *k*, the number of pairs of interday feature series and intraday feature vectors included in the model. The figure shows that using *k*=3 was inadequate, but that larger values of *k* led to reasonable forecast accuracy. The best results for the SVD method correspond to *k*=7, which confirms Shen and Huang's (2008a) suggestion that *k* should not need to be more than the number of days in a week of the time series. The results for the SVD method with *k*=7 are competitive with the HWT method, particularly beyond two days ahead.

---------- Fig. 12 ----------

8. Summary and concluding comments

In previous empirical studies with intraday data, exponential smoothing methods have performed well. This motivated us to develop the following five new exponentially weighted methods, which we have introduced in this paper:

(i) A double seasonal version of the total and split exponential smoothing method.

(ii) DWR used to update recursively the parameters of a linear model with trigonometric terms.

(iii) DWR used to update recursively a regression spline.

(iv) Spline-based exponential smoothing. In the approach of Harvey and Koopman (1993), we simply replace the multiple source of error state space model with an exponential smoothing model. The model is formulated in terms of time-variation in the values of a spline function at its knots.

(v) SVD-based exponential smoothing. First, SVD is applied to the intraday cycle of the time series in order to extract the main underlying features. An exponential smoothing model is then used with states defined as these main features.

A recurring theme in this paper has been the reduction of the dimensionality of a forecasting model in order to achieve parsimony and greater efficiency. Dimension reduction was the motivation behind the development of the IC and parsimonious seasonal exponential smoothing methods. Modelling intraday data using trigonometric terms is an attempt to concisely summarise the seasonal cycles. The fundamental appeal of spline methods is that they reduce the problem from modelling the intraweek cycle at all the periods of the week to modelling the cycle for a substantially smaller number of knots. Dimension reduction is clearly the motivation for using an SVD-based approach.

In Table 1, we provide an informal and subjective summary of the practical usefulness of the methods implemented in this paper. The bottom five rows of the table correspond to the five new methods that we have presented. The final column of the table summarises the ease with which prediction intervals can be generated for the forecasts from each method. In essence, when exponential smoothing is formulated as a statistical model, prediction intervals can be produced using analytical formulae or Monte Carlo simulation. However, with DRLS, the situation is less clear, and an empirical or Bayesian approach may be needed for estimating prediction intervals.

Table 1 shows that the HWT method is a challenging benchmark. It is relatively straightforward to understand and implement, and it performed well in terms of forecast accuracy in our study. However, beyond about one day ahead, it was matched or outperformed by several of the other methods that involved exponential smoothing. Indeed, overall, our forecasting results suggest that the methods involving exponential smoothing have greater potential than those using DWR. However, more empirical evidence is needed regarding relative accuracy. Further empirical work could also help address practical issues, such as the role of judgement. For example, the use of parsimonious seasonal exponential smoothing relies on judgemental clustering of intraweek periods, and the spline methods involve judgemental selection of the number and locations of the knots. For other methods, the role of statistical procedures is unclear. For example, perhaps cross-validation should replace the use of signal coherence for selecting trigonometric terms prior to estimation using DWR. Perhaps cross-validation should also be used to select the number of feature series and vectors to use in our SVD-based approach. We would also welcome additional empirical evidence using other data, as well as empirical comparison with methods from outside the exponentially weighted domain, such as ARIMA modelling, the dynamic harmonic regression of Tych et al. (2002), or the random effects model of Weinberg et al. (2007). The sizeable literature on forecast combining has numerous examples where combining has led to improved accuracy. It would be interesting to see the results of combining several methods considered in this paper, either with each other, or with methods not based on exponential weighting.

---------- Table 1 ----------

Acknowledgements

We are grateful to NHS Direct for providing the data, and to three referees for providing comments that helped greatly to improve the paper.

References

Ameen, J.R.M. & Harrison, P.J. (1984). Discounted weighted estimation, *Journal of Forecasting*, 3, 285-296.

Brooks, C. & Hinich, M.J. (2006). Detecting intraday periodicities with application to high frequency exchange rates, *Applied Statistics*, 55, 241-259

Brown, R.G. (1963). *Smoothing, Forecasting and Prediction of Discrete Time Series*, NJ: Prentice-Hall, Englewood Cliffs.

Christiaanse, W.R. (1971). Short-term load forecasting using general exponential smoothing, *IEEE Transactions on Power Apparatus and Systems*, PAS-90, 900-910.

De Livera, A. M. & Hyndman, R.J. (2009). Forecasting time series with complex seasonal patterns using exponential smoothing, Department of Econometrics and Business Statistics Working Paper 15/09, Monash University.

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

Foster, J., Hinich, M.J. & Wild, P. (2008). Randomly modulated periodic signals in Australia's national electricity market, *Energy Journal*, 29, 105-129.

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

Gardner, E.S. Jr. (1985). Exponential smoothing: The state of the art, *Journal of Forecasting*, 4, 1-28.

Gould, P.G., Koehler, A.B., Ord, J.K., Snyder, R.D., Hyndman, R.J. & Vahid-Araghi, F. (2008). Forecasting time-series with multiple seasonal patterns, *European Journal of Operational Research*, 191, 207-222.

Harrison, P. J. & Johnston, F. R. (1984). Discount weighted regression, *Journal of the Operational Research Society*, 35, 923-932.

Harvey, A. & Koopman, S.J. (1993). Forecasting hourly electricity demand using time-varying splines, *Journal of the American Statistical Association*, 88, 1228-1236.

Hur, D., Mabert, V.A. & Bretthauer, K.M. (2004). Real-time work schedule adjustment decisions: An investigation and evaluation, *Productions and Operations Management*, 13, 322-339.

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

Lam, W.H.K., Tang, Y.F., Chan, K.S. & Tam, M.-L. (2006). Short-term hourly traffic forecasts using Hong Kong annual traffic census, *Transportation*, 33, 291-310.

Li, T.H. (2008). On exponentially weighted recursive least squares for estimating time-varying parameters, *Journal of Statistical Theory and Practice*, 2, 339-354.

Poirier, D.J. (1973). Piecewise regression using cubic splines, *Journal of the American Statistical Association*, 68, 515-524.

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

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

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

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. (2007). Forecasting supermarket sales using exponentially weighted quantile regression, *European Journal of Operational Research*, 178, 154-167.

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). Density forecasting of intraday call center arrivals using models based on exponential smoothing, Working paper, University of Oxford.

Taylor, J.W., M. de Menezes, L.M. & McSharry, P.E. (2006). A comparison of univariate methods for forecasting electricity demand up to a day ahead, *International Journal of Forecasting*, 22, 1-16.

Taylor, J.W. & McSharry, P.E. (2007). Short-term load forecasting methods: An evaluation based on European data, *IEEE Transactions on Power Systems*, 22, 2213-2219.

Taylor, J.W. & Snyder, R.D. (2009). Forecasting intraday time series with multiple seasonal cycles using parsimonious seasonal exponential smoothing, Department of Econometrics and Business Statistics Working Paper 9/09, Monash University.

Tych, W., Pedregal, D.J., Young, P.C. & Davies J. (2002). An unobserved component model for multi-rate forecasting of telephone call demand: The design of a forecasting support system, *International Journal of Forecasting*, 18, 673-695.

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

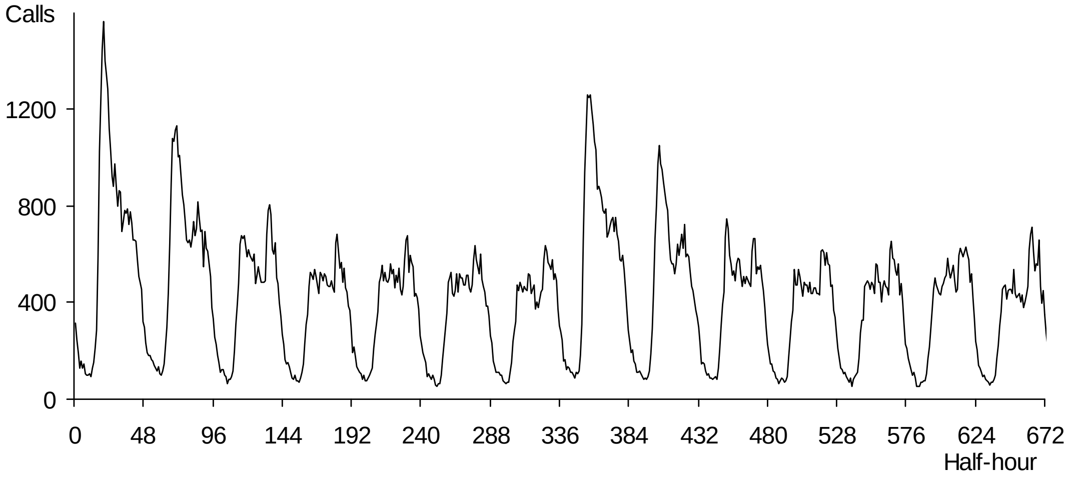

Fig. 1. Half-hourly arrivals at NHS Direct for the fortnight from Saturday 6 January 2007 to Friday 19 January 2007.

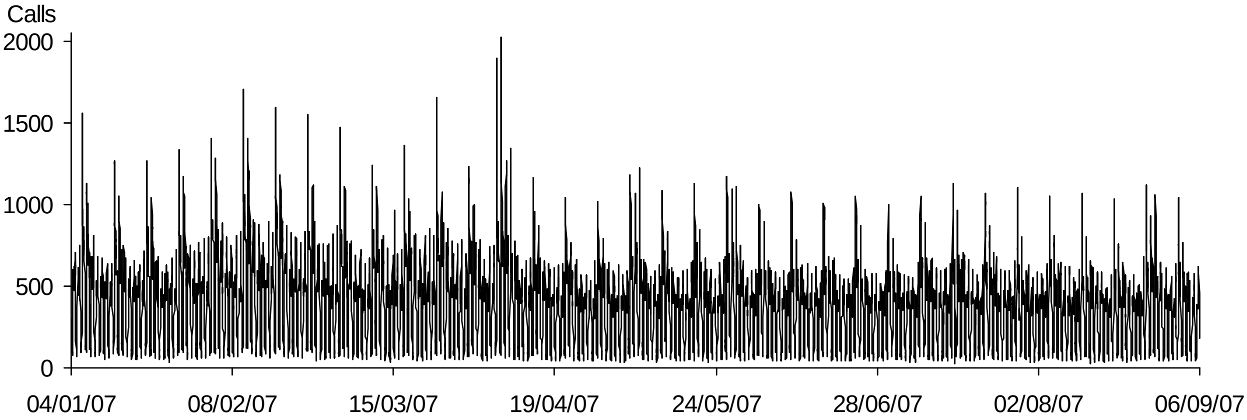

Fig. 2. Half-Hourly Arrivals at NHS Direct for a 35-week period in 2007.

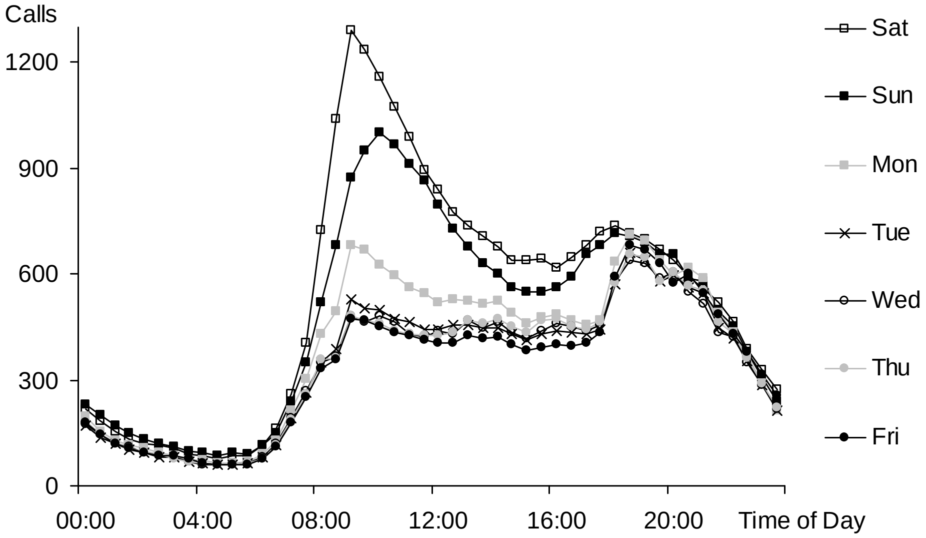

Fig. 3. Average intraday cycle for each day of the week for the NHS Direct data.

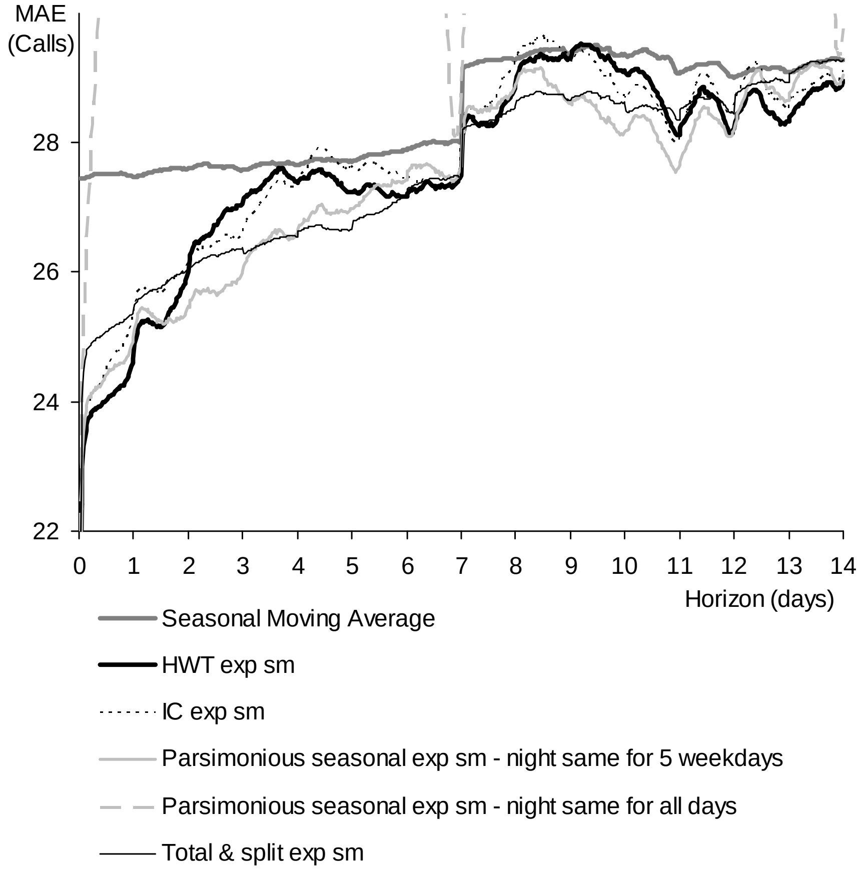

Fig. 4. Post-sample comparison of exponential smoothing methods for lead times from one half-hour ahead to two weeks ahead.

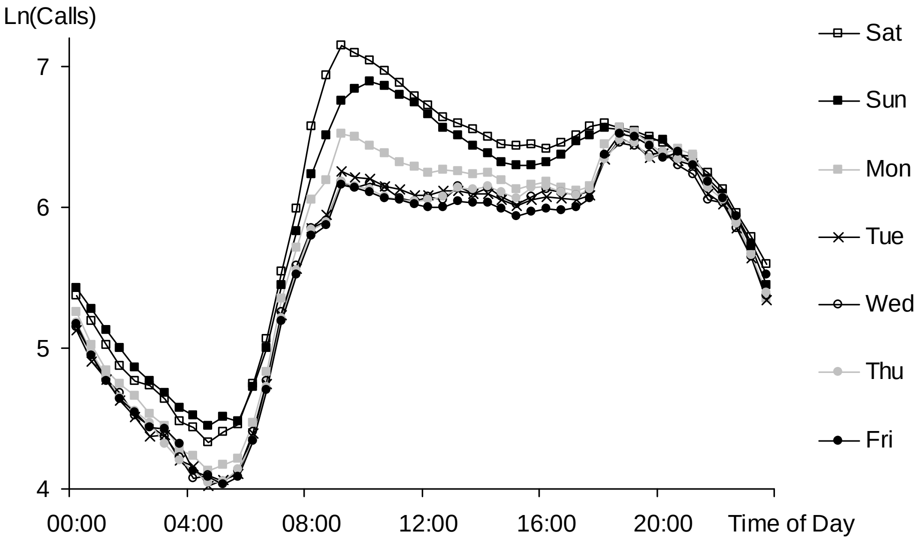

Fig. 5. Average intraday cycle for each day of the week for the logarithmic transformation of the NHS Direct data.

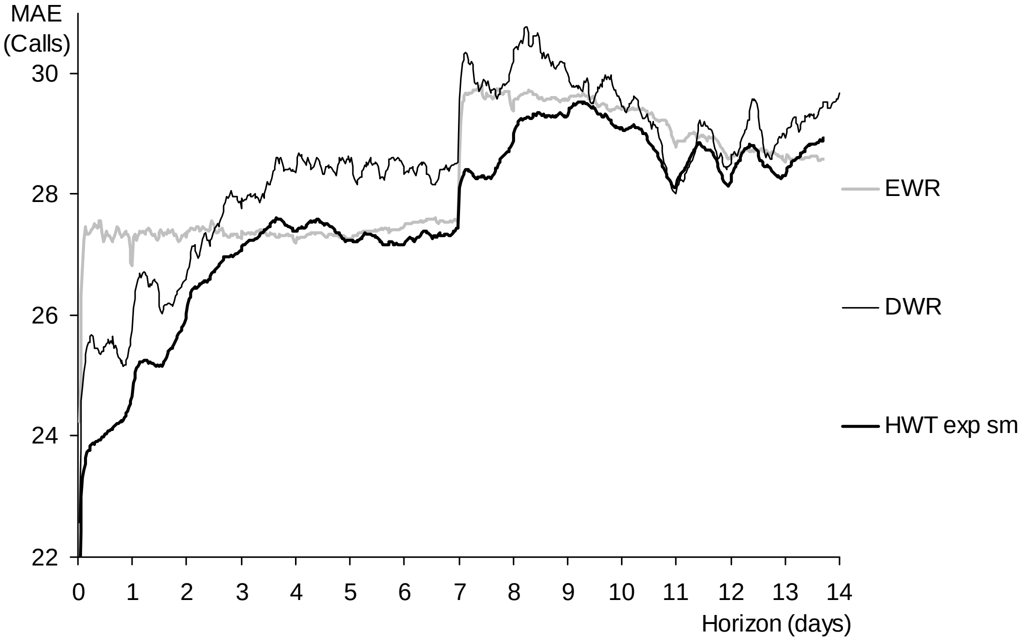

Fig. 6. Post-sample comparison of EWR and DWR with trigonometric terms chosen by signal coherence. Results also shown for HWT exponential smoothing. Results for lead times from one half-hour ahead to two weeks ahead.

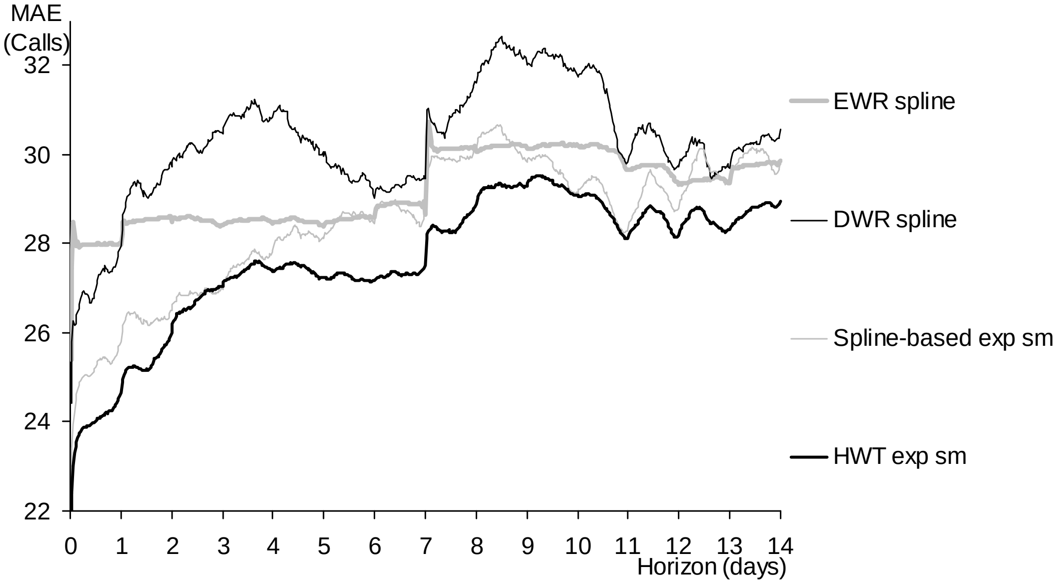

Fig. 7. Post-sample comparison of spline methods and HWT exponential smoothing for lead times from one half-hour ahead to two weeks ahead.

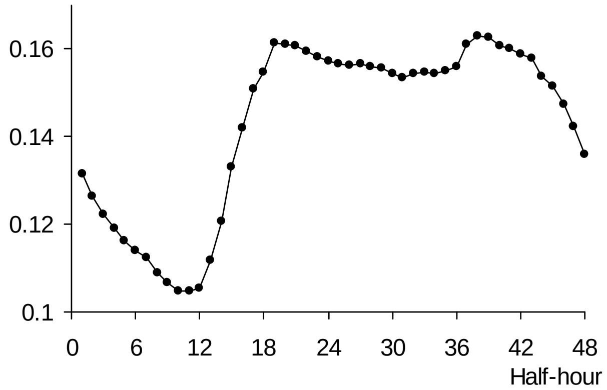

Fig. 8. First intraday feature vector (column of ***V***).

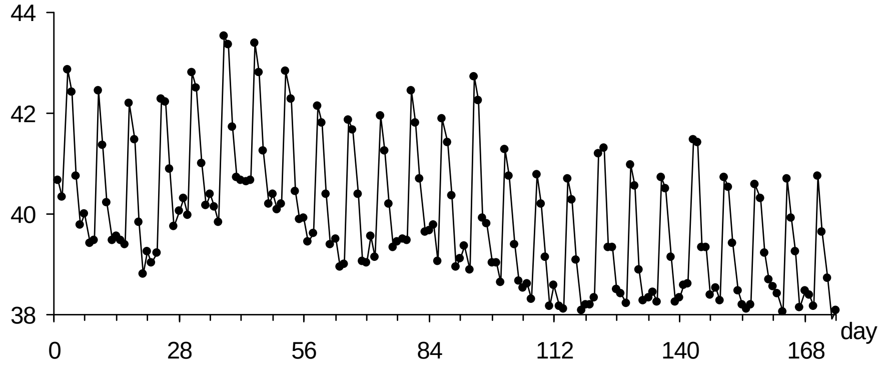

Fig. 9. First interday feature series (column of ***P***).

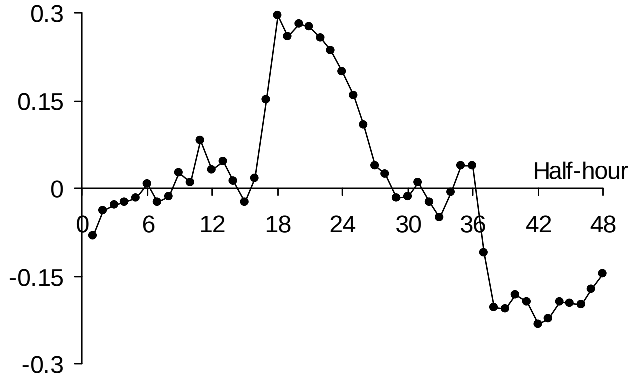

Fig. 10. Second intraday feature vector (column of ***V***).

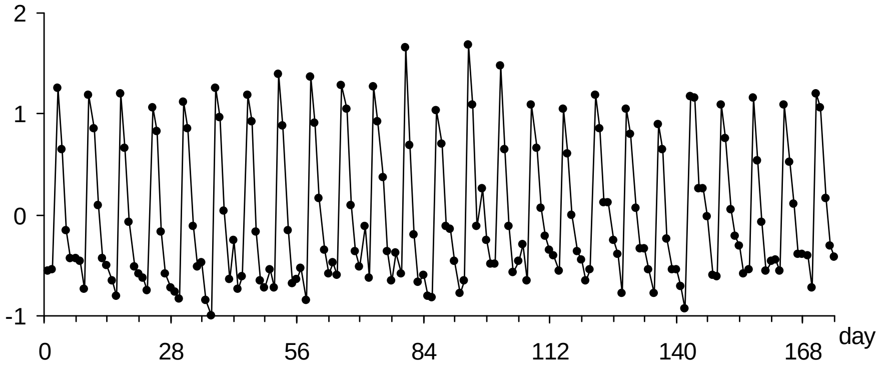

Fig. 11. Second interday feature series (column of ***P***).

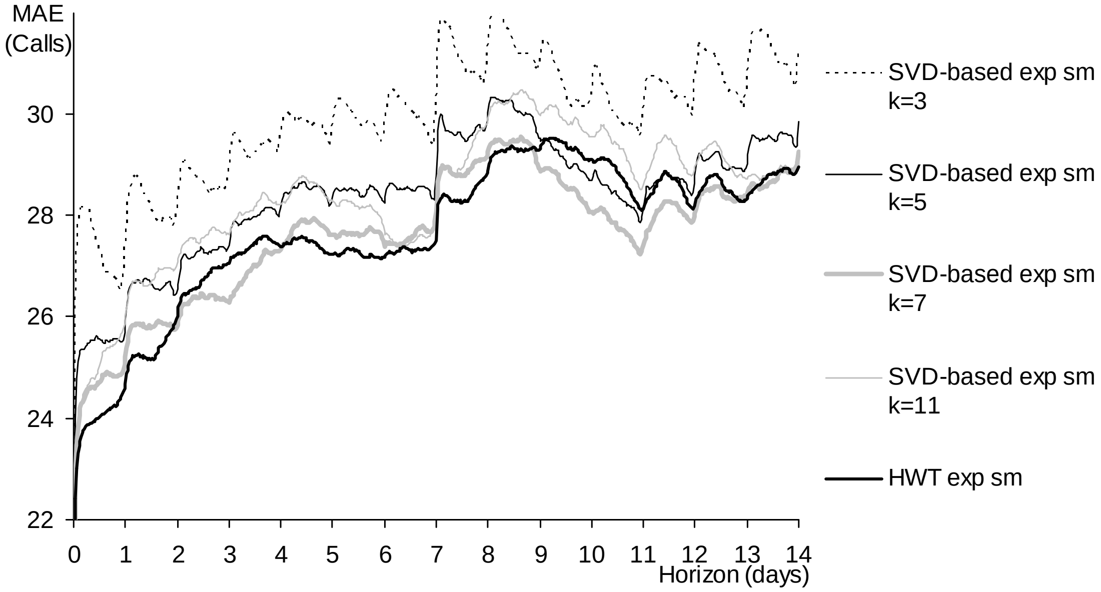

Fig. 12. Post-sample comparison of SVD-based exponential smoothing and HWT exponential smoothing for lead times from one half-hour ahead to two weeks ahead.

Table 1
Informal assessment of the practicality of the methods implemented in this paper. The five new methods are presented in the bottom five rows. Each method is graded on a scale of A to C, with an A grade being preferable.

	Conceptual simplicity	Ease of implementation	Judgement required	Overall point forecast accuracy	Ease of estimating prediction intervals
HWT exp sm	A	A	A	A	A
Intraday cycle exp sm	B	B	C	A	A
Parsimonious seasonal exp sm	B	B	C	A	A
Double seasonal total & split exp sm	A	A	A	A	A
DWR with trigonometric terms	B	C	A	B	C
DWR splines	B	C	C	C	C
Regression splines with exp sm	C	C	C	B	A
SVD with exp sm	C	B	B	A	A