Exponentially Weighted Information Criteria for Selecting Among Forecasting Models

James W. Taylor

Saïd Business School

University of Oxford

International Journal of Forecasting, 2008, Vol. 24, pp. 513-524.

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

Author biographical sketch:

James W. Taylor is a Professor of Decision Sciences at the Saïd Business School, University of Oxford. His research interests include exponential smoothing, prediction intervals, quantile regression, combining forecasts, volatility forecasting, call centre forecasting, electricity demand forecasting and weather ensemble predictions.

Exponentially Weighted Information Criteria for Selecting Among Forecasting Models

Abstract

Information criteria (IC) are often used to select between forecasting models. Commonly used criteria are Akaike's IC and Schwarz's Bayesian IC. They involve the sum of two terms: the model's log likelihood and a penalty for the number of model parameters. The likelihood is calculated with equal weight given to all observations. We propose that greater weight should be put on more recent observations in order to reflect more recent accuracy. This seems particularly pertinent when selecting among exponential smoothing methods, as they are based on an exponential weighting principle. In this paper, we use exponential weighting within the calculation of the log likelihood for the IC. Our empirical analysis uses supermarket sales and call centre arrivals data. The results show that basing model selection on the new exponentially weighted IC can outperform individual models and selection based on the standard IC.

Keywords: Information criteria; Model selection; Exponential weighting; Exponential smoothing; SARMA.

1. Introduction

Model selection is an important issue for time series forecasting applications. The choice of a poor model can have substantial financial implications. Selection procedures can benefit from judgemental input, but often this is either not available or its use is impractical. This is the case, for example, when the aim is to select a model for each of a large number of series. In this situation a purely statistical approach must be used.

A common statistical approach to model selection is to use an information criterion (IC), such as Akaike's IC (AIC) or Schwarz's Bayesian IC (BIC). These measures involve the summation of two terms. The first is the model's log likelihood, which provides a natural assessment of the quality of the fit of the model. The second is a penalty term, which is a function of the number of parameters in the model and has the aim of avoiding over-fitting. In a recent study, Billah et al. (2005) present a novel new class of empirically based IC in which the data dictates the form of the penalty term. In this paper, we also present empirically based IC. However, by contrast with Billah et al. (2005), we use the standard penalty terms, and instead let the data dictate the form of the log likelihood measure of fit.

In the existing IC, the likelihood is calculated with equal weight given to all observations. We propose that greater weight should be put on more recent observations in order to reflect more recent accuracy. This principal underlies exponential smoothing methods because they involve exponentially decreasing weight being placed on older observations. With this in mind, in this paper, we propose the use of exponential weighting within the calculation of the log likelihood for the IC. More specifically, for our empirical work, we rewrite the likelihood in terms of the model's residual variance, and then use exponential weighting within the calculation of this variance. The exponential decay parameter is derived empirically using the in-sample data. Although the new class of IC would seem to have particular appeal for selecting among exponential smoothing methods, it

is not restricted to this application, and indeed could be used for selection between various different types of models.

In Section 2, we briefly review IC and their use in model selection. In Section 3, we introduce our new exponentially weighted IC. Section 4 provides an empirical evaluation of the use of the new criteria for a case study involving the forecasting of daily supermarket sales data. In Section 5, we provide further empirical evidence using a time series of half-hourly call centre arrivals. The final section provides a summary and concluding comments.

2. Model selection using IC

The selection of a statistical model is often based on IC. The theoretical basis is the Kullback-Leibler discrimination information (KL) (see Kullback, 1959), which is presented in expression (1).

$$KL = E\left(\ln\left(f\left(\mathbf{y}\right)\right) - \ln\left(f_1\left(\mathbf{y}\right)\right)\right) \tag{1}$$

where **y** is a vector of observations of the dependent variable; f is the true unknown joint density; f_1 is the joint density given by the model; and the expectation is taken with respect to f. The essential idea is to rank candidate models according to their KL value. As the first term on the right hand side of expression (1) is a constant for all models, it is equivalent to rank models according to $-E(\ln(f_1(\mathbf{y}))$. However, this expectation cannot be evaluated because f is unknown. This has prompted the development of IC, which are asymptotically unbiased estimators of this expression (Irizarry, 2001).

Formally, the use of an IC assumes that the model has been estimated by maximum likelihood. The IC is usually written as

$$IC = -\frac{2}{n}\sum_{i=1}^{n}\ln\left(L_i\right) + g\left(n,k\right) \tag{2}$$

where n is the number of observations used to fit the model; L_i is the contribution from the ith observation to the likelihood function for the fitted model; k is the number of parameters in

the model; and g is the penalty term aimed at avoiding over-fitting. The preferred model is the one for which the value of the IC is lowest. Making the assumption of an additive Gaussian model error term leads to a Gaussian likelihood and an IC expression in terms of the residual variance, which is commonly simplified to the following:

$$IC = \ln\left(\frac{1}{n}\sum_{i=1}^{n} e_i^2\right) + g(n,k) \qquad (3)$$

where e_i is the residual from the fitted model. A variety of IC have been proposed that differ only in their choice of penalty term (see Billah et al., 2006). Most popular amongst these are the AIC (Akaike, 1973, 1974) and the BIC (Schwarz, 1978).

For the AIC: $g(n,k) = 2k/n$

For the BIC: $g(n,k) = k\ln(n)/n$

From these functions, it is evident that the BIC carries a much stiffer penalty for the number of parameters than the AIC. Inoue and Kilian (2006) provide recent theoretical analysis for using IC to select from among forecasting models. Empirical evidence for their use in selecting among exponential smoothing methods is provided by Hyndman et al. (2002) and Billah et al. (2006). Interestingly, Gardner (1985, 2006) concludes both of his reviews of exponential smoothing by commenting on the need for greater practical guidance on how to select between exponential smoothing methods and other time series methods.

Billah et al. (2005) introduce a new class of empirically based IC, which involves the use of in-sample data to derive the penalty term. They allow the penalty to be a simple function of the number of model parameters. For applications such as inventory control, where a collection of time series is available, they construct a common penalty function based on all the series. For cases involving just one series, the authors describe a form of bootstrapping. They use the standard likelihood or residual variance term to assess the fit within their criteria. Their study prompted our development of a new class of empirically based IC. By contrast with the IC of Billah et al. (2005), we use the standard penalty terms,

and instead let the data dictate the form of the measure of fit. We present the new criteria in the next section.

3. Model selection using exponentially weighted IC

In this section, we first consider the estimation of model parameters using maximum weighted likelihood. We then discuss a number of studies that have used weighted likelihood within IC. Finally, we introduce exponentially weighted IC, which is the focus of this paper.

3.1. Weighted likelihood estimation

There is a well established literature on the estimation of statistical model parameters using maximum weighted likelihood. In reviewing this area, Hu and Zidek (2002) cite one of the earliest contributions as being that of Tibshirani and Hastie (1987) who, in a regression context, use zero weights for data points outside a window around a given point in regressor space, and unit weights for all points within the window. This idea has been developed in the literature on nonparametric regression, where weighted likelihood estimation is presented as a form of kernel estimation (see Staniswalis, 1989).

A somewhat earlier discussion of weighted likelihood estimation is provided by Gilchrist (1967) in the context of smoothing and forecasting time series. Gilchrist considers an exponentially discounting weighting scheme, and links the idea to Brown's (1963) work on exponentially weighted least squares estimation and exponential smoothing (see the discussion of general exponential smoothing by Gardner, 1985). It is worth noting that, when viewed in the context of nonparametric estimation, exponentially weighted likelihood corresponds to the case where the kernel is defined as being one-sided with exponentially declining weight on past data (see Gijbels et al., 1999). There are recent examples of the use of exponentially weighted likelihood estimation in the area of value at risk in finance (see Mittnik and Paolella, 2000; Guermat and Harris, 2002; Fan and Gu, 2003).

3.2. Weighted IC

In addition to their use in estimation, weighted likelihood expressions have also been used within IC for model selection. Most of these uses have focused on the development of robust forms of model selection. Drawing on robust estimation methods, Ronchetti (1985, 1997) proposes that the log likelihood measure of fit in the IC be replaced by alternative measures involving robust functions of the residuals, including a weighted average of squared residuals that downweights outlying observations. Agostinelli (2002) presents a robust IC based on a weighted likelihood employing the weight function of Markatou et al. (1997).

A weighted likelihood is used within an IC by Hens et al. (2006), who propose a weighting scheme to help account for the deletion of incomplete cases from a multivariate sample of data. Irizarry (2001) addresses the problem of selecting between different locally approximating models to use within a local likelihood nonparametric regression procedure. He develops modified IC, which incorporate a local weighting scheme within the likelihood expression. In the next section, we propose the use of an exponentially weighted likelihood within an IC. As exponential weighting is a form of kernel weighting, in a sense our proposal can be viewed as an extension of Irizarry's use of a locally weighted likelihood within an IC.

3.3. Exponentially weighted IC

In the standard IC expressions (2) and (3), equal weight is given to all observations. In this paper, we suggest that when comparing two or more time series forecasting models greater weight should be put on more recent observations in order to reflect more recent accuracy. This seems particularly pertinent when selecting among exponential smoothing methods, as they are based on the principle of applying greater weight to more recent data. Although our proposal is not restricted to selecting among exponential smoothing methods, these methods do suggest a rather simple and intuitive weighting scheme. Our proposal is to use exponential weighting within the calculation of the IC. With regard to the theoretical

basis of the IC, discussed at the beginning of Section 2, this exponential weighting amounts to estimating the KL, or more specifically the expression -$E(\ln(f(\mathbf{y}))$, using a conditional estimate, rather than an unconditional estimate that is used in the standard IC.

Incorporating exponential weighting within the IC expression (2), we get the following exponentially weighted IC (EWIC) expression:

$$EWIC = -2\left(\frac{1-\lambda}{1-\lambda^n}\right)\sum_{i=1}^{n}\lambda^{n-i}\ln(L_i) + g(n,k) \tag{4}$$

The speed of the exponential decay is governed by the decay parameter $\lambda \in [0,1]$. As in expression (2), L_i is the contribution from the ith observation to the likelihood function for the fitted model. By contrast with the standard IC of expression (2), which involves the simple average of the L_i, expression (4) involves an exponentially weighted average of the L_i. It is worth noting that, in a sense, we apply the same exponential weighting principle to both the log likelihood calculation and the penalty term. This can see in the following expressions, where we rewrite expression (4):

$$\begin{aligned} EWIC &= -2\left(\frac{1-\lambda}{1-\lambda^n}\right)\sum_{i=1}^{n}\lambda^{n-i}\ln(L_i) + g(n,k) \\ &= \left(\frac{1-\lambda}{1-\lambda^n}\right)\sum_{i=1}^{n}\lambda^{n-i}\left(-2\ln(L_i) + g(n,k)\right) \end{aligned}$$

In expression (3), we presented the standard IC in terms of residual variance. Imposing exponential weighting on expression (3) delivers the following form for the EWIC:

$$EWIC = \ln\left(\left(\frac{1-\lambda}{1-\lambda^n}\right)\sum_{i=1}^{n}\lambda^{n-i}\,e_i^2\right) + g(n,k) \tag{5}$$

In the IC of expression (3), the residual variance is estimated as the mean squared residual. By contrast, in expression (5), it is estimated as an exponentially weighted average of the squared residuals. Using the same penalty functions from the AIC and BIC delivers new exponentially weighted criteria, EWAIC and EWBIC, respectively. In our empirical studies of Sections 4 and 5, we consider only models with a Gaussian additive error term, and

so we use the forms of the IC and EWIC in expressions (3) and (5). In those sections, we describe how we empirically derive the exponential decay parameter using in-sample data. For simplicity, in our work, for each time series, we have used a common value of the decay parameter for all lead times. However, derivation of a different parameter for each lead time would be straightforward to implement.

At the beginning of this section, we suggested that exponentially decaying weights seems a particularly natural concept for IC when selecting between exponential smoothing methods. Conversely, it could be suggested that, because these methods already have exponential weighting within their state equations, exponentially weighting the IC is unnecessary and unappealing. It is certainly true that the exponential smoothing state equations deliver a fitted value for a particular period that is constructed from an exponential weighting of data prior to that period. However, in selecting among forecasting models, our proposal is that a new weighting scheme should be introduced that weights the fit for that period in proportion to the distance of the period from the forecast origin. By contrast, a standard IC applies no weighting, and assesses average fit across the estimation sample. In summary, we are proposing exponential weighting of the fit of each model, and it is unimportant how the fitted values were constructed.

4. Empirical analysis of daily supermarket sales

4.1. Description of the study

In this section, we compare the forecast accuracy resulting from model selection based on the standard IC with selection based on the new exponentially weighted IC. Our empirical study used the same dataset employed by Taylor (2007), which consists of 256 time series of daily observations of sales of different items from an outlet of a large UK supermarket chain. The series vary in length from 72 to 1436 observations with a median of 764 observations. Figure 1 is a plot of one of the series. Prior to the forecasting study, bank

holidays and the days of the Christmas periods were replaced by the average of the corresponding day of the week from the week before and the week after the day in question. This is consistent with the approach taken by the company, as they do not use their time series forecasting approach for these periods.

---------- Figure 1 ----------

The final 20% of each series was used for post-sample forecast evaluation. For each series, we rolled the forecast origin forward through the post-sample evaluation period to produce a collection of forecasts from each model for each horizon. We considered forecast horizons of one to 14 days in our study. For each new forecast origin, model parameters were re-estimated using a moving window of length equal to 80% of the total number of observations in each series.

4.2. Individual forecasting models

The supermarket sales application requires an automated forecasting process because predictions are required at frequent intervals for many different products. Exponential smoothing is widely used for such inventory control applications due to its simplicity, robustness and accuracy (see Gardner, 2006). We included in our study the company's own exponential smoothing approach and simple exponential smoothing applied to data deseasonalised using multiplicative seasonal decomposition. For each series, we calculated afresh the seasonal factors for each new forecast origin using only data prior to that origin. Of the methods considered by Taylor that involve exponential smoothing of the level of the series, the company's approach and simple exponential smoothing were the most accurate in that study. The company's approach involves smoothing the total weekly sales, W_t, and the split, L_t, of the weekly sales across the days of the week. With sales expressed as y_t, the formulation is given as

$$W_t = \alpha \sum_{i=0}^{6} y_{t-i} + (1-\alpha) W_{t-1} \tag{6}$$

$$L_t = \frac{\gamma\, y_t}{\sum_{i=1}^{7} y_{t-i}} + (1-\gamma) L_{t-7} \tag{7}$$

The forecasts are given by

$\hat{y}_t(m) = W_t L_{t+m-7}$ for m = 1 to 7

$\hat{y}_t(m) = W_t L_{t+m-14}$ for m = 8 to 14

Following the approach of Hyndman et al. (2002), both simple exponential smoothing and the company's approach can be expressed in state space form to deliver innovations models for which parameters can be estimated using maximum likelihood. The innovations form of the single source of error state space model for the company's method is given by the observation equation in expression (8) along with the state equations of expressions (6) and (7). For simplicity, in our analysis, we employed an additive Gaussian error term, ε_t, which means that the IC expressions (3) and (5) can be used.

$$y_t = W_{t-1} L_{t-7} + \varepsilon_t \tag{8}$$

Following common practice, we constrained the parameters to lie between zero and one, and we used simple averages of the first few observations to calculate initial values for the smoothed components (see, for example, Hyndman et al., 2002). Because we are not estimating the initial values in the same optimisation as the parameters, our optimisation amounts to maximising a conditional likelihood function.

4.3. Model selection based on IC

At each forecast origin, we used the AIC, BIC, EWAIC and EWBIC to select between the two individual forecasting models. For some series, the same model was selected at all forecast origins, whilst for others the IC dictated considerable switching between the two models. For each of the two EWIC, the parameter, λ, was optimised once for each series

using just the subsample consisting of the first 80% of the series. Of this subsample, the final 20% was used to evaluate the forecasting performance of the EWIC approach for different values of λ. We rolled the forecast origin forward through this final 20% to produce a collection of one day-ahead predictions. For each new origin, parameters were re-estimated for both individual models using a moving window of length equal to 80% of the number of observations in the subsample. The optimal value of λ was deemed to be the value that led to the lowest mean absolute one day-ahead prediction error for the final 20% of the subsample. Our choice of the mean absolute error was made arbitrarily, and alternative error summary measures could certainly be used. As in the work of Billah et al. (2005), we used a grid search to find the optimal value of λ because the objective function is somewhat complex. We considered values of λ from 0.8 to 1 with an increment of 0.005.

Figure 2 presents a histogram of the optimal EWAIC λ values for the 256 series. The figure shows that for about 30% of the series, the optimal value was found to be one, implying no exponential decay. By contrast, the low values of λ for many of the series implies very rapid exponential decay. We were concerned that the optimal values derived for the shorter length series were not as reliable as for the longer series. With this in mind, we implemented a second version of the EWIC approach; this time using for all 256 series a common value of λ, which we chose to be the mean of the optimised values displayed in Figure 2. Our use of a collection of time series to derive the value of λ is similar to the approach taken by Billah et al. (2005) to define the form of the empirically based penalty term within their new IC. It is also similar to the approach followed by Fildes et al. (1998) in their selection of exponential smoothing parameters. They find that using a commonly occurring value for all series can be preferable to using the value optimised for each series. For the EWAIC, the mean of the 256 optimised λ values was 0.9388. The mean value for the EWBIC was similar. Indeed, the results for the BIC and EWBIC were broadly similar to

those for the AIC and EWAIC, respectively. In view of this, in Section 4.4, for simplicity we present only the results for the following three IC: AIC, EWAIC with λ optimised for each series, and EWAIC with λ set as the mean of the 256 optimised parameter values.

---------- Figure 2 ----------

4.4. Results

We evaluated post-sample forecasting performance using the mean absolute error (MAE), median absolute error (MedAE) and the root mean squared error (RMSE). Percentage error measures were not appropriate as the sales values were often close to zero for many of the series. Since the order of magnitude of the forecast errors varied across the series, it was not appropriate to average the error measures across the 256 series. Therefore, we used the same measure employed by Taylor (2007) to summarise performance across all the series. For the MAE results, the calculation proceeded by computing, for each series and each forecast horizon, the ratio of the MAE for that model to the MAE for simple exponential smoothing. We then calculated the weighted geometric mean of this measure across the 256 series, where the weighting was proportional to the number of post-sample observations in each series. Expression (9) presents the measure:

$$\left(\prod_{i=1}^{256} \left(\frac{MAE_i}{MAE_i^{SES}} \right)^{\left(\frac{N_i}{\sum_{j=1}^{256} N_j} \right)} - 1 \right) \times 100 \qquad (9)$$

where MAE_i and MAE_i^{SES} are the MAE for the method being considered and for simple exponential smoothing, respectively; and N_i is the number of post-sample observations for series i. Lower values of the measure are better, with negative values indicating that the method outperforms simple exponential smoothing, and positive values indicating the opposite. Figure 3 plots the results for the measure of expression (9), and Figure 4 presents the results for the

measure based on the MedAE. We do not present the results for the measure based on the RMSE, because the ranking of methods for this measure was similar to those for the MAE results shown in Figure 3.

---------- Figures 3 and 4 ----------

Let us consider first the MAE results in Figure 3. The figure shows no clear dominance between the company's model and simple exponential smoothing, and that both are outperformed by model selection based on the AIC for all lead times except two to five days ahead and 10 days ahead. Interestingly, up to about 12 days ahead, the results for model selection based on the AIC are not as good as those for selection based on the EWAIC with λ optimised separately for each series. The figure shows further improvement resulting from the use of the EWAIC approach with λ set as the mean of the 256 optimised values.

The results in Figure 4 for the MedAE show the company's model outperforming simple exponential smoothing up to 10 days ahead. By contrast with the results for the MAE, Figure 4 shows the AIC only improving on the company's model at seven days ahead and beyond 10 days ahead. However, the plot shows that model selection is improved by using EWIC with λ optimised separately for each series, and that this approach is further improved by using EWIC with the common choice of λ. Indeed, overall, this version of the EWIC approach outperforms both of the individual models.

Closer inspection of the IC approaches revealed that the IC penalty terms generally did not have a substantial impact on model selection, and that selection tended to be governed by the residual variance term (which is replaced in the EWIC approaches by exponentially weighted estimates of residual variance). This is perhaps not surprising given that there is a small difference in the number of parameters in the two models; one parameter for simple exponential smoothing and two for the company's model. This is not the case in the next section, where there is a substantial difference in the number of parameters in the two models considered.

5. Empirical analysis of half-hourly call centre arrivals

5.1. Description of the study

In this section, we evaluate model selection based on the new IC using a time series of half-hourly arrivals at the call centre of a major retail bank in the UK. The data is used in the study of Taylor (2008). It covers the 36-week period from 3 January 2004 to 10 September 2004, inclusive. The series corresponds to calls received by the bank's four large call centres dealing with general customer enquiries. These centres are open on all seven days of the week, from 7am to 11pm, and together receive more than a quarter of a million calls each week.

Figure 5 shows the first four weeks of the series, which are very representative of the behaviour of the series throughout our full sample of data. The series exhibits no apparent trend and very clear seasonality. Indeed, an interesting feature of the series is the presence of both an intraday seasonal cycle and an intraweek seasonal cycle. Although the intraday cycle is less obvious, closer inspection reveals that, at least on weekdays, there is a peak around 11am followed by a second, lower peak around 2pm. Gans et al. (2003) explain that such an intraday pattern is reasonably typical.

---------- Figure 5 ----------

As with the supermarket sales data, prior to fitting and evaluating models, we smoothed out the 'special days', such as bank holidays, as their inclusion is likely to be unhelpful in our comparison of models. We replaced all special days by the average of the corresponding period in the two adjacent weeks. The variance in the arrivals appeared to be changing over time, and approximately proportional to the volume of arrivals, which is consistent with the common assumption that call centre arrivals obey a time-inhomogeneous Poisson process (see, for example, Brown et al. 2005). In order to reduce this heteroskedasticity, we applied a log transformation to this series before model fitting.

We used the final 12 weeks of data for post-sample forecast evaluation. As with the supermarket series, we rolled the forecast origin forward through the post-sample evaluation period to produce a collection of forecasts from each model for each horizon. We considered all forecast horizons from one half-hour ahead to two weeks ahead. Model parameters were re-estimated for each new forecast origin using a moving window of 24 weeks of data.

5.2. Individual forecasting models

We included two individual models in the study: a multiplicative double seasonal ARMA model and an innovations model based on Taylor's (2003) double seasonal exponential smoothing. We followed the Box-Jenkins methodology to identify the most suitable SARMA model based on the first estimation sample of 24 weeks. No differencing was performed because the series did not appear to have a unit root in the level or seasonality. We considered lag polynomials up to order three, and this resulted in the following model.

$$\left(1-\phi_1 L-\phi_2 L^2\right)\left(1-\phi_{32}L^{32}-\phi_{64}L^{64}-\phi_{96}L^{96}\right)\left(1-\phi_{224}L^{224}-\phi_{448}L^{448}-\phi_{672}L^{672}\right)\left(y_t-c\right) = \left(1-\theta_2 L^2-\theta_3 L^3\right)\left(1-\theta_{96}L^{96}\right)\left(1-\theta_{448}L^{448}-\theta_{672}L^{672}\right)\varepsilon_t$$

where y_t is the log of the arrivals, ε_t is an error term, L is the lag operator, and ϕ_i, θ_i and c are parameters, which are re-estimated at each forecast horizon using maximum likelihood based on the standard Gaussian assumption. The SARMA terms are represented by powers of the lag operator that are multiples of 32 and 224, which are the lengths of the intraday and intraweek seasonal cycles, respectively.

Taylor's (2003) double seasonal exponential smoothing formulation for intraday data is presented in the following expressions:

$$l_t = \alpha\left(y_t - d_{t-32} - w_{t-224}\right) + (1-\alpha)\, l_{t-1} \tag{10}$$

$$d_t = \delta\left(y_t - l_t - w_{t-224}\right) + (1-\delta)\, d_{t-32} \tag{11}$$

$$w_t = \omega\left(y_t - l_t - d_{t-32}\right) + (1-\omega)\, w_{t-224} \tag{12}$$

$$\hat{y}_t(k) = l_t + d_{t-32+k} + w_{t-224+k} + \phi^k\left(y_t - \left(l_{t-1} + d_{t-32} + w_{t-224}\right)\right) \qquad \text{for } k \le 32 \tag{13}$$

where l_t is the smoothed level; d_t and w_t are the seasonal indices for the intraday and intraweek seasonal cycles, respectively; α, δ and ω are the smoothing parameters (which were constrained to lie between zero and one); and $\hat{y}_t(k)$ is the k step-ahead forecast made from forecast origin t, where $k \leq 32$. It is straightforward to rewrite the forecast function for longer lead times. The term involving the parameter ϕ, in expression (13), is a simple adjustment for first-order autocorrelation. We used simple averages of the first few observations to calculate initial values for the smoothed components. As we commented in Section 4.2, this implies that our optimisation can be viewed as maximising a conditional likelihood function. The method can be expressed in state-space form to deliver a model for which parameters can be estimated using maximum likelihood. The innovations form of the single source of error state space model is given by the formulation of expressions (10) to (12) with the forecast function of expression (13) replaced by expressions (14) and (15).

$$y_t = l_{t-1} + d_{t-32} + w_{t-224} + \phi e_{t-1} + \varepsilon_t \quad (14)$$

$$e_t = y_t - \left(l_{t-1} + d_{t-32} + w_{t-224}\right) \quad (15)$$

5.3. Model selection based on IC

As in the supermarket sales study, at each forecast origin, we used the AIC, BIC, EWAIC and EWBIC to select between the two individual forecasting models. In both studies, our implementation of the IC for selection at each forecast origin can also be termed model-switching. The SARMA model and exponential smoothing method contained 14 and four parameters, respectively, implying a substantial difference in their respective IC penalty terms.

The weighting parameter λ for the EWIC was optimised using a very similar procedure to that described for the supermarket application. The optimisation used just the subsample consisting of the first 24 weeks of the series. Of this subsample of observations, the final six weeks was used to evaluate the forecasting performance of the EWIC approach

for different values of λ. For each new origin, parameters were re-estimated for both individual models using a moving window of length equal to 18 weeks. The optimal value of λ was deemed to be the value that led to the lowest mean absolute percentage one day-ahead forecast error for the final six weeks of the subsample. We considered values of λ from 0.8 to 1 with an increment of 0.0005. In comparison with the supermarket case, we used a finer increment for our call centre study because it only involved one time series. The result was optimal λ values of 0.9785 and 0.9865 for the EWAIC and EWBIC, respectively. These relatively high values seem reasonable when one considers the high frequency nature of the data, which has a seasonal cycle consisting of a relatively large number of periods. It is worth noting that these values correspond to half-lives of 32 and 51 half-hour periods, respectively. As in our reporting of the supermarket study, in the next section, for simplicity, we present only the results for model selection based on the AIC and EWAIC.

5.4. Results

In Figures 6 and 7, we present the post-sample mean absolute percentage error (MAPE) and median absolute percentage error (MedAPE), respectively. Comparing the two individual models, the figures show that for prediction up to about a week ahead, exponential smoothing performed the better, but that the superiority was reversed for the longer lead times.

---------- Figures 6 and 7 ----------

Turning to the IC, we found that the AIC selected the SARMA model at all forecast origins, which explains why the AIC model selection results coincide with those of the SARMA model in Figures 6 and 7. In Figure 6, the MAPE results for the EWAIC approach are very encouraging, as they match or outperform the better of the individual models for lead times up to a week ahead. For longer lead times, the EWAIC approach compares well with the SARMA model, which was the better of the individual models for these longer lead times. The relative

performance of the EWAIC approach is similar for the MedAPE to that for the MAPE, with particularly good results for prediction up to five days ahead.

6. Summary and concluding comments

In this paper, we have introduced a new class of IC for selecting among forecasting models. It involves the use of exponential weighting within the measure of fit in the standard IC, such as the AIC and BIC. The implication is greater weight being placed on more recent observations in order to reflect more recent accuracy. The criteria can be viewed as empirically based as the exponential weighting parameter is derived using the in-sample data. We demonstrated the use of model selection based on the new IC using two case studies, one involving supermarket sales data and the other using a series of call centre arrivals. The results showed that overall selection based on the new IC was able to outperform use of the individual models, as well as selection based on the standard IC.

In terms of future research, it would be interesting to have further empirical results for different datasets. With regard to developing the method, one possibility is to investigate the potential for a synthesis of our EWIC with the empirically based IC of Billah et al. (2005). The resultant hybrid IC would involve an exponential weighting within the measure of fit, and an empirically driven penalty term.

Acknowledgements

We are grateful to the members of the forecasting groups at the collaborating supermarket and bank for providing the data and background information for the study. We also acknowledge the helpful comments of Everette Gardner, Rob Hyndman and Patrick McSharry on an earlier version of the paper. We are also grateful for the insightful comments of an associate editor and two referees.

References

Agostinelli, C. (2002). Robust model selection in regression via weighted likelihood methodology, *Statistics and Probability Letters*, 56 289-300.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, in 2nd International Symposium on Information Theory, Petrov, B.N. & Csáki, F. (eds), Akadémia Kiadó: Budapest, 267–281.

Akaike, H. (1974). A new look at the statistical model identification, *IEEE Transactions on Automatic Control*, AC-19, 716-723.

Billah, M. B., Hyndman, R.J. & Koehler, A.B. (2005). Empirical information criteria for time series forecasting model selection, *Journal of Statistical Computation and Simulation*, 75, 831-840.

Billah, M. B., King, M.L., Snyder, R.D. & Koehler, A.B. (2006). Exponential smoothing model selection for forecasting, *International Journal of Forecasting*, 22, 239-249.

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

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

Fan, J. & Gu, J. (2003). Semiparametric estimation of value at risk, *Econometrics Journal*, 6, 261-290.

Fildes, R., Hibon, M., Makridakis, S. & Meade, N. (1998). Generalising about univariate forecasting methods: Further empirical evidence, *International Journal of Forecasting*, 14, 339-358.

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.

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

Gijbels, I., A. Pope & Wand, M.P. (1999). Understanding exponential smoothing via kernel regression, *Journal of the Royal Statistical Society Series B*, 61, 39-50.

Gilchrist, W.G. (1967). Methods of estimation involving discounting, *Journal of the Royal Statistical Society Series B*, 29, 355-369.

Guermat, C. & Harris, R.D.F. (2002). Forecasting value at risk allowing for time variation in the variance and kurtosis of portfolio returns, *International Journal of Forecasting*, 18, 409-419.

Hens, N., Aerts, M., & Molenberghs, G. (2006). Model selection for incomplete and design-based samples, *Statistics in Medicine*, 25, 2502-2520.

Hu, F. & Zidek, J.V. (2002). The weighted likelihood, *The Candaian Journal of Statistics*, 30, 347-371.

Hyndman, R.J., Koehler, A.B., Snyder, R.D. & Grose, S. (2002). A state space framework for automatic forecasting using exponential smoothing methods, *International Journal of Forecasting*, 18, 439-454.

Inoue, A. & Kilian, L. (2006). On the selection of forecasting models, *Journal of Econometrics*, 130, 273-306.

Irizarry, R.A. (2001). Information and posterior probability criteria for model selection in local likelihood estimation, *Journal of the American Statistical Association*, 96, 303-315.

Kullback, S. (1959), *Information Theory and Statistics*, New York: Wiley.

Markatou, M., Basu, A., & Lindsay, B.G. (1997). Weighted likelihood estimating equations: the discrete case with applications to logistic regression, *Journal of Statistical Planning and Inference*, 57, 215–232.

Mittnik, S. & Paolella, M.S. (2000). Conditional density and value-at-risk prediction of Asian currency exchange rates, *Journal of Forecasting*, 19, 313-333.

Ronchetti, E. (1985). Robust model selection in regression, *Statistics and Probability Letters*, 3, 21-23.

Ronchetti, E. (1997). Robustness aspects of model choice, *Statistica Sinica*, 7 327-338.

Schwarz, G. (1978). Estimating the dimension of a model, *Annals of Statistics*, 6, 461-464.

Staniswalis, J.G. (1989). The kernel estimate of a regression function in likelihood-based models, *Journal of the American Statistical Association*, 84, 276-283.

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

Taylor, J.W. (2007). Forecasting daily 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.

Tibshirani, R.J. & Hastie, T. (1987). Local likelihood estimation, *Journal of the American Statistical Association*, 82,559-567.

Figure 1 Daily supermarket sales of a single item.

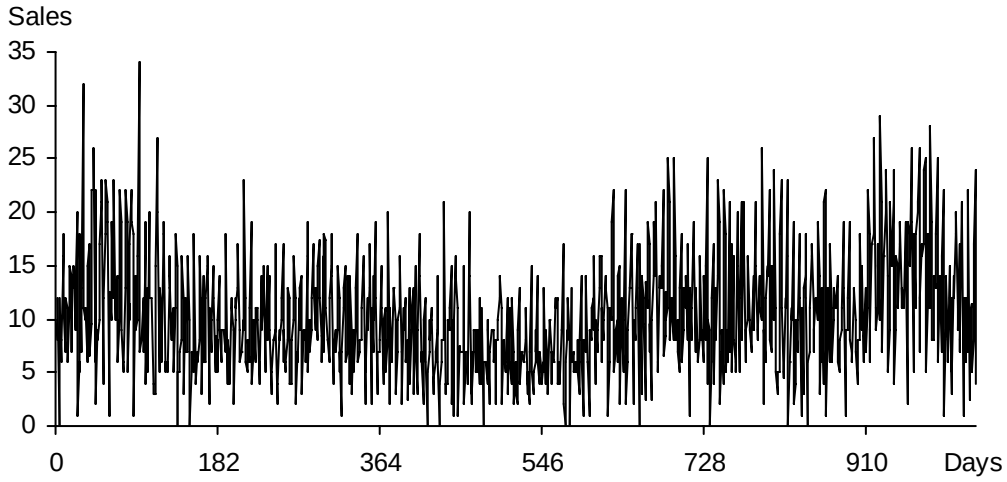

Figure 2 Histogram for optimised values of exponential weight parameter λ for the 256 series.

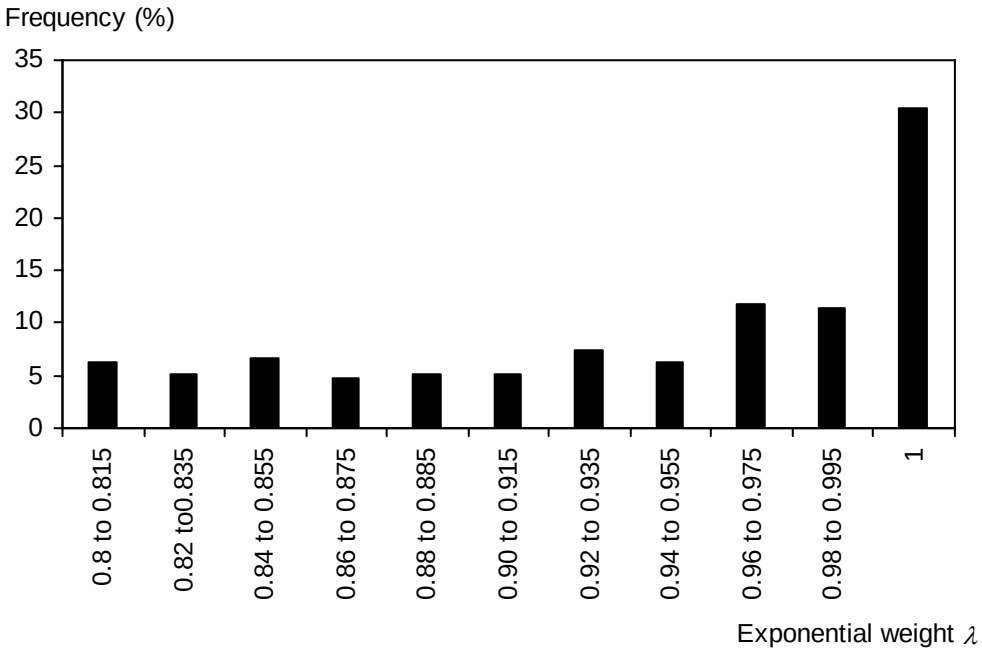

Figure 3 Relative MAE measure of expression (9), calculated for all 256 series. Accuracy measured relative to simple exponential smoothing. Lower values are better.

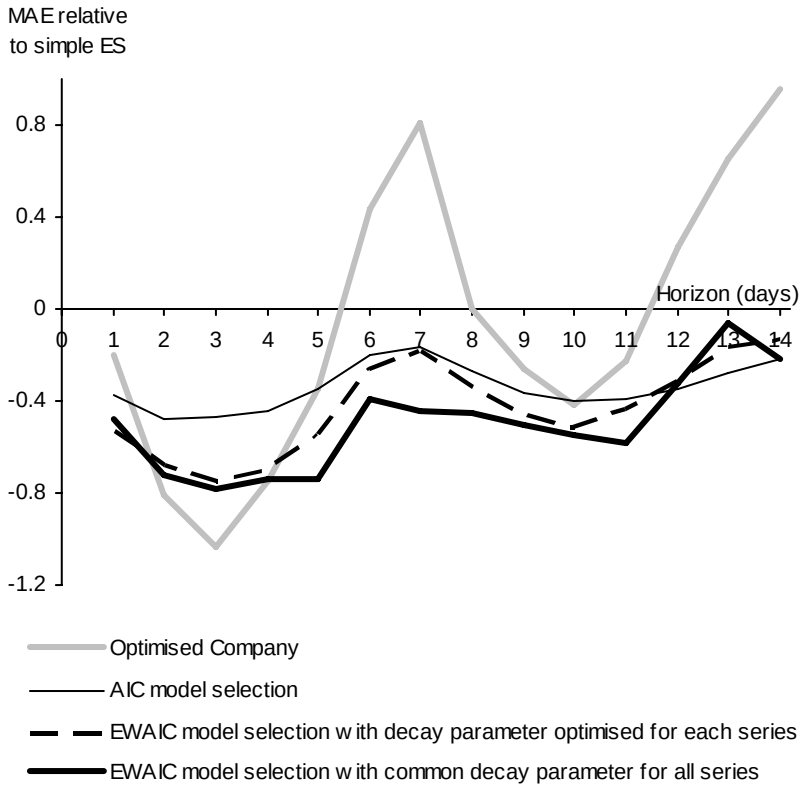

Figure 4 Relative MedAE measure calculated for all 256 series. Accuracy measured relative to simple exponential smoothing. Lower values are better.

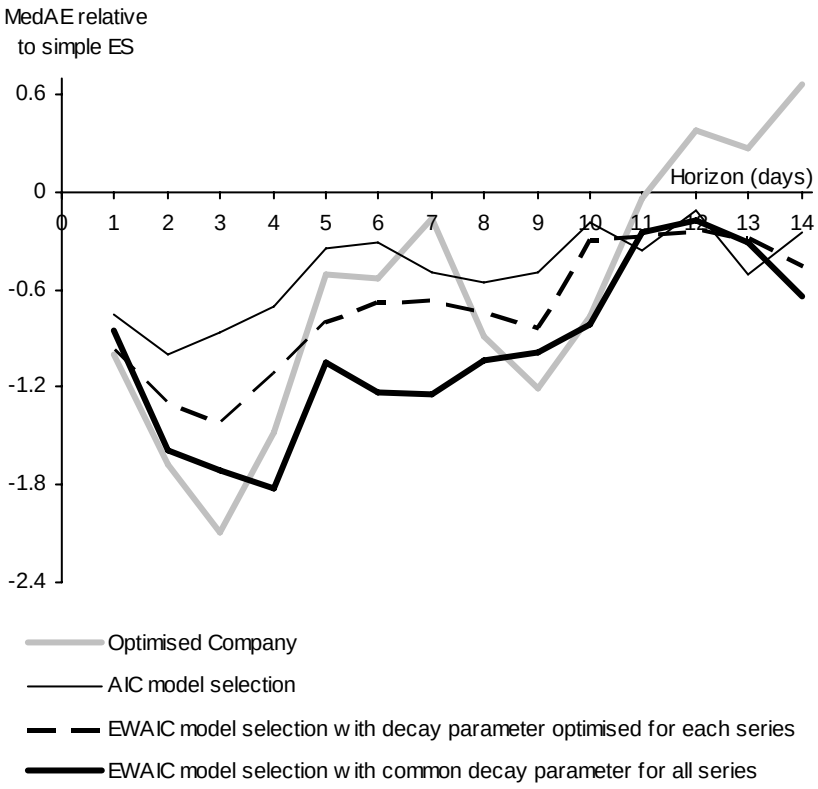

Figure 5 Half-Hourly call centre arrivals from Saturday 3 January to Friday 30 January 2004. The centres are open for 224 half-hours each week.

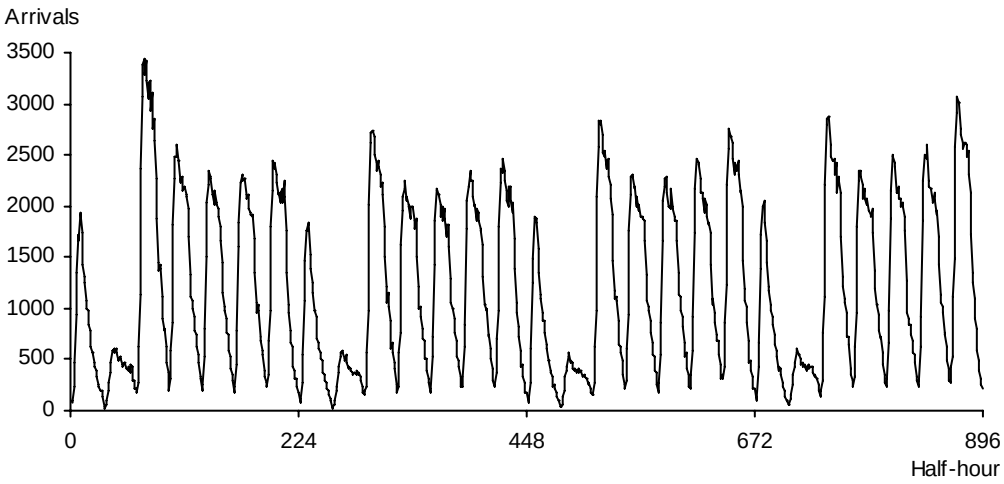

Figure 6 MAPE for the call centre arrivals series for lead times from one half-hour to 14 days.

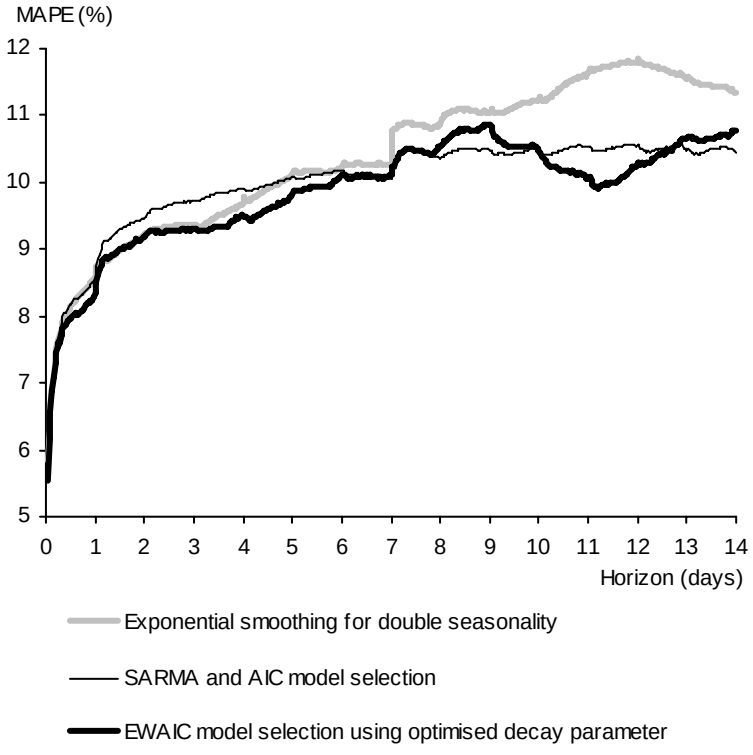

Figure 7 MedAPE for the call centre arrivals series for lead times from one half-hour to 14 days.