Abstract

Applications of exponential smoothing to forecasting time series usually rely on three basic methods: simple exponential smoothing, trend corrected exponential smoothing and a seasonal variation thereof. A common approach to selecting the method appropriate to a particular time series is based on prediction validation on a withheld part of the sample using criteria such as the mean absolute percentage error. A second approach is to rely on the most appropriate general case of the three methods. For annual series this is trend corrected exponential smoothing: for sub-annual series it is the seasonal adaptation of trend corrected exponential smoothing. The rationale for this approach is that a general method automatically collapses to its nested counterparts when the pertinent conditions pertain in the data. A third approach may be based on an information criterion when maximum likelihood methods are used in conjunction with exponential smoothing to estimate the smoothing parameters. In this paper, such approaches for selecting the appropriate forecasting method are compared in a simulation study. They are also compared on real time series from the M3 forecasting competition. The results indicate that the information criterion approaches provide the best basis for automated method selection, the Akaike information criteria having a slight edge over its information criteria counterparts.

Keywords: Model selection; Exponential smoothing; Information criteria; Prediction; Forecast validation

1. Introduction

The exponential smoothing methods are relatively simple but robust approaches to forecasting. They are widely used in business for forecasting demand for inventories (Gardner, 1985). They have also per-
structural changes in product markets. As these components may be combined by addition or multiplication operators, 24 variations of the exponential smoothing methods may be identified (Hyndman, Koehler, Snyder, & Grose, 2002). Given this proliferation of options, an automated approach to method selection becomes most desirable (Gardner, 1985; McKenzie, 1985).

Hyndman et al. (2002) provided a statistical framework for exponential smoothing based on the earlier work of Ord, Koehler, and Snyder (1997). The framework incorporated stochastic models underlying the various forms of exponential smoothing and enabled the calculation of maximum likelihood estimates of smoothing parameters. It also enabled the use of Akaike’s information criterion (Akaike, 1973) for method selection. One issue not addressed was the preference for Akaike’s information criterion over possible alternatives such as Schwarz (1978), Hannan and Quinn (1979), Mallows (1964), Golub, Heath, and Wahba (1979), and Akaike (1970). One aim, therefore, is to determine whether Akaike’s information criterion (AIC) has a superior performance compared to its alternatives. Given that it was developed to minimise the forecast mean squared error, it might be hypothesised that the AIC has a natural advantage over the alternatives in forecasting applications, except possibly for Akaike’s FPE which is asymptotically equivalent to the AIC.

The exponential smoothing methods were traditionally implemented without reference to a statistical framework so that other approaches were devised to resolve the method selection problem. Prediction validation (Makridakis, Wheelwright, & Hyndman, 1998) is one such approach. The sample is divided into two parts: the fitting sample and the validation sample. The fitting sample is used to find sensible values for the smoothing parameters, often with a sum of squared one-step ahead prediction error criterion. The validation sample is used to evaluate the forecasting capacity of a method with a criterion such as the mean absolute percentage error (MAPE). Another approach applies a general version of exponential smoothing on the assumption that it effectively reduces to an appropriate nested method when this is warranted by the data. Trend corrected exponential smoothing is applied to annual time series; Winter’s method is applied to sub-annual time series. A second aim is to gauge the effectiveness of these traditional approaches relative to the information criterion approach to method selection.

The plan of this paper is as follows. State space models for exponential smoothing and an approach to their estimation are introduced in Section 2. Criteria to be used in model selection and a measure for comparing resulting forecast errors are explained in Section 3. A simulation study is discussed in Section 4. An application of the model selection criteria to the M3 competition data (Makridakis & Hibon, 2000) is given in Section 5. The paper ends with some concluding remarks in Section 6.

2. State space models

The state space framework in Snyder (1985), and its extension in Ord et al. (1997), provides the basis of an efficient method of likelihood evaluation, a sound mechanism for generating prediction distributions, and the possibility of model selection with information criteria. Important special cases, known as structural models, that capture common features of time series such as trend and seasonal effects, provide the foundations for simple exponential smoothing, trend corrected exponential smoothing and Holt–Winters’ seasonal exponential smoothing. Of the 24 versions of exponential smoothing found in Hyndman et al. (2002), the scope of this study is limited to three linear cases.

The focus is on a time series that is governed by the innovations model (Snyder, 1985):

\[ y_t = hx_{t-1} + \varepsilon_t \]  

\[ x_t = Fx_{t-1} + \alpha \varepsilon_t. \]

Eq. (2.1), called the measurement equation, relates an observable time series value \( y_t \) in typical period \( t \) to a random \( k \)-vector \( x_{t-1} \) of unobservable components from the previous period. \( h \) is a fixed \( k \)-vector, while the \( \varepsilon_t \), the so-called innovations, are independent and normally distributed random variables with mean zero and a common variance \( \sigma^2 \). The inter-temporal dependencies in the time series are defined in terms of the unobservable components with the so-called transition equation (2.2). \( F \) is a fixed \( k \times k \) ‘transition’ matrix and \( \alpha \) is a \( k \)-vector of smoothing parameters.
The following special cases, termed structural models, provide the statistical underpinning for common forms of exponential smoothing. The models most naturally relate to the error correction versions of exponential smoothing (Gardner, 1985), but also underpin the more traditional and equivalent ‘weighted average’ versions of the method.

- **Local Level Model (LLM)**: \( y_t = \ell_{t-1} + \epsilon_t \), where \( \ell_t \) is a local level governed by the recurrence relationship \( \ell_t = \ell_{t-1} + \alpha x_t \), where \( 0 \leq \alpha \leq 1 \). It underpins the simple exponential smoothing method (Brown, 1959).
- **Local Trend Model (LTM)**: \( y_t = \ell_{t-1} + b_{t-1} + \epsilon_t \), where \( b_t \) is a local growth rate. The local level and local growth rates are governed by the equations \( \ell_t = \ell_{t-1} + b_{t-1} + \alpha x_t \) and \( b_t = b_{t-1} + \beta \epsilon_t \), respectively, where \( 0 \leq \alpha \leq 1 \) and \( 0 \leq \beta \leq \alpha \). Note that \( \alpha' = [\alpha \beta] \). This model underpins trend corrected exponential smoothing (Holt, 1957).
- **Additive Seasonal Model (ASM)**: \( y_t = \ell_{t-1} + b_{t-1} + s_{t-m} + \epsilon_t \), where \( s_t \) is the local seasonal component. The local level, growth and seasonal components are governed by \( \ell_t = \ell_{t-1} + b_{t-1} + \alpha x_t \), \( b_t = b_{t-1} + \beta \epsilon_t \), and \( s_t = s_{t-m} + \gamma \epsilon_t \), respectively, where \( m \) is the number of seasons in a year, \( 0 \leq \alpha \leq 1 \), \( 0 \leq \beta \leq \alpha \), and \( 0 \leq \gamma \leq 1 - \alpha \). In this case \( x_t' = [\ell_t \ b_t \ c_t \ldots c_{t-m+1}] \) and \( \alpha' = [\alpha \beta \gamma \ 0 \ldots 0] \). This model is the basis of Holt–Winters’ additive method (Winters, 1960).

Traditionally, the smoothing parameters \( \alpha \), \( \beta \) and \( \gamma \) were set to fixed values determined subjectively by users on the basis of personal experience. The studies of Chatfield (1978) and Bartolomei and Sweet (1989) show that this can be problematic and that parameters are best estimated from data. Ord et al. (1997) recommend that estimates of the parameters be obtained by maximizing the conditional log-likelihood. For the class of linear state space models (2.1) and (2.2), the conditional likelihood function based on a sample of size \( n \) is given by

\[
\log L = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{t=1}^{n} \epsilon_t^2
\]  

(2.3)

where the errors are calculated recursively with a general linear form of exponential smoothing defined by the relationships \( \epsilon_t = y_t - \mathbf{h}'x_{t-1} \) and \( x_t = \mathbf{Fx}_{t-1} + \mathbf{a} \) \( \epsilon_t \) for \( t = 1, 2, \ldots, n \).

This conditional likelihood is not only a function of \( \alpha \) but also of the unobserved random vector \( x_0 \). Following a tradition from exponential smoothing, the seed state vector \( x_0 \) is approximated with plausible heuristics such as the following:

<table>
<thead>
<tr>
<th>Model</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Local level model</strong></td>
<td>Seed level estimate equals average of first three observations in sample.</td>
</tr>
<tr>
<td><strong>Local trend model</strong></td>
<td>Fit a linear trend line to the first five observations. The seed level is set equal to its intercept. The seed rate is set equal to its slope.</td>
</tr>
<tr>
<td><strong>Local seasonal model</strong></td>
<td>Fit a linear trend line with seasonal dummy variables to the first two years of data. The seed level and rate are chosen as before. The seasonal factors are set equal to coefficients of the seasonal dummies after normalisation to sum to zero.</td>
</tr>
</tbody>
</table>

The resulting heuristic estimates \( \hat{x}_0 \) of \( x_0 \) are entered into the conditional likelihood function, so that the latter is then only maximized with respect to the parameter vector \( \alpha \).

The exact likelihood function could potentially be obtained by integrating \( x_0 \) out of (2.3). The conditional likelihood is used instead because the state variables in the models are generated by non-stationary processes so that the seed state vector has an improper unconditional distribution. The situation is similar to one considered in Bartlett’s paradox (Bartlett, 1957) from Bayesian statistics. Exact likelihood values for models with different state dimensions are non-comparable and information criteria based on them are therefore also non-comparable. This is not an issue for the conditional likelihood because the use of an improper unconditional distribution of the seed state vector is avoided. It is for this reason that the conditional likelihood is used instead of the exact likelihood for estimating the parameter vector \( \alpha \).

On obtaining the estimates \( \hat{x}_0 \) and \( \hat{\alpha} \), the exponential smoothing algorithm is used to obtain the corresponding estimate \( \hat{x}_n \) of the state vector at the end of the sample. Point forecasts are then generated recursively with the equations \( \hat{y}_{n+j} = \mathbf{h}' \hat{x}_{n+j-1} \) and \( \hat{x}_{n+j} = \mathbf{Fx}_{n+j-1} \) for \( j = 1, 2, \ldots, r \) where \( r \) is the prediction horizon.

### 3. Model selection approaches and a measure for comparing them

An information criterion has the general form \( \log L(\hat{\alpha}) - p(n,q) \), where \( p(n,q) \) is the so-called...
penalty function, \( q \) being the number of free parameters. Various forms of the penalty have been suggested, as may be seen from Table 1. Note that \( q^* \) is the number of free parameters in the smallest model that nests all models under consideration and \( c = n - q^* \).

No clear theory exists for deciding which of these information criteria is best suited for choosing the appropriate method of exponential smoothing. Thus, a simulation study was undertaken to compare them. The simulation also included a comparison with two other approaches for model selection. The prediction validation approach (Val) selects the model with the smallest mean absolute percentage error (MAPE) for forecasting withheld data, and the encompassing model approach always selects LTM for annual data and ASM for quarterly and monthly data.

The performance of each approach was gauged in terms of the median absolute prediction error as a percentage of the standard deviation, given by

\[
\text{MdAPES} = \text{median} \left( \frac{\left| y_{n+j} - \hat{y}_n(j) \right|}{\sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2}} \times 100 \right).
\]

This measure was chosen for several reasons. It is a unit free measure that permits comparisons between real time series measured in different units. It avoids the problems encountered with the MAPE (and its variations) for series values close to zero. Most importantly, it gives a fair comparison when applied to time series with different standard deviations. The forecast error may display a tendency to be larger for time series with a large standard deviation. APES is a measure that does not produce larger values just because there is more variability in the time series. Thus, such time series will not necessarily cause an increase in the MdAPES just because of the larger variability. More variable time series can still have APES values near the median value and play a central role in the evaluation process. In the comparisons, both simulated data and real data with different amounts of variability are included. The median was used instead of the average to eliminate the distorting effect of outlier APES.

### 4. Simulation study

A simulation study was conducted to determine whether any of the approaches to model selection displayed a superior performance as measured in terms of forecast accuracy instead of the more usual criteria of the proportion of correctly selected models. It consisted of many experiments carried out under a wide variety of conditions. Depending on the type of data, the time series were generated by the three models: the local level model (LLM), local trend model (LTM) and additive seasonal model (ASM). For the ASM, the seed seasonal components were generated from the equation \( s_{j-m} = A \sin(2j\pi/m) \), \( j=1,2, \ldots, m \), where \( A \) is the seasonal amplitude and \( m \) is the number of seasons in a year.

Candidate values for the various factors in these models were:

\[
\begin{align*}
\sigma &= 10, 20 \quad \text{(for all models)} \\
n &= 24, 40 \quad \text{(for LLM and LTM)} \\
n &= 24, 60 \quad \text{(for quarterly/m = 4 ASM)} \\
n &= 48, 96 \quad \text{(for monthly/m = 12 ASM)} \\
\alpha &= 0.1, 0.5, 0.9; \quad \epsilon_0 = 100 \quad \text{(for LLM)} \\
\beta &= \begin{pmatrix} 0.1 \\ 0.05 \\ 0.1 \\ 0.7 \end{pmatrix}; \\
\epsilon_0 &= \begin{pmatrix} 100 \\ 100 \\ 1 \\ 5 \end{pmatrix} \quad \text{(for LTM)}
\end{align*}
\]

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Penalty function</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIC</td>
<td>( q )</td>
<td>Akaike (1973)</td>
</tr>
<tr>
<td>BIC</td>
<td>( \log(n)/2 )</td>
<td>Schwarz (1978)</td>
</tr>
<tr>
<td>HQ</td>
<td>( \log(\log(n)) )</td>
<td>Hannan and Quinn (1979)</td>
</tr>
<tr>
<td>MCp</td>
<td>( n\log(1+2q/c)/2 )</td>
<td>Mallows (1964)</td>
</tr>
<tr>
<td>GCV</td>
<td>( -n\log(1-q/n) )</td>
<td>Golub et al. (1979)</td>
</tr>
<tr>
<td>FPE</td>
<td>( (n\log(n+q)-n\log(n-q))/2 )</td>
<td>Akaike (1970)</td>
</tr>
</tbody>
</table>

\( q \) is the number of free parameters. \( q^* \) is the number of free parameters in the smallest model nesting all other models. \( c = n - q^* \).
\[
\begin{pmatrix}
  z \\
  \beta \\
  \gamma \\
  \sigma \\
  b_0
\end{pmatrix}
= 
\begin{pmatrix}
  0.1 \\
  0.05 \\
  0.1 \\
  0.2 \\
  100
\end{pmatrix}
, 
\begin{pmatrix}
  0.4 \\
  0.1 \\
  0.2 \\
  100 \\
  5
\end{pmatrix}
; 
A = 0, 25, 50 (for ASM).
\]

The various combinations of these factors lead to 180 scenarios for the simulation study. The 180 scenarios were repeated 10 times (i.e., 10 trials) so that the study consisted of 1800 simulation experiments.

Each experiment consisted of the following four steps:

1. Generate a time series, from a specified model, consisting of a) a tuning sample of a specified size \(n\), and b) an evaluation sample for \(r\) succeeding periods: for annual data \(r=6\), for quarterly data \(r=8\), and for monthly data \(r=18\).

2. Fit a collection of models to the tuning sample using the conditional likelihood function: the LLM and LTM for annual data, and additionally the ASM model for quarterly and monthly data.

3. Select the best model by one of the model selection approaches:
   a) For the six information criteria approaches, choose the model that is best according to each information criterion.
   b) For the prediction validation approach (Val):
      i) Withhold the last \(r\) periods of the tuning sample and fit the local level model, local trend model, and additive seasonal model to the first \(n-r\) values by maximizing the conditional likelihood function.
      ii) Choose the model with the smallest MAPE for the forecasts of the \(r\) periods of withheld data.
   c) For the encompassing approach (Enc):
      i) Choose the local trend model for annual data.
      ii) Choose the additive seasonal model for quarterly and monthly data.

4. Using the estimates from Step 2 for the model chosen in Step 3:
   i) Generate predictions for each of the \(r\) periods in the evaluation sample.
   ii) Calculate the absolute prediction error as a percentage of the standard deviation of the tuning sample (APES) for each of the time periods in the evaluation sample.

Overall, 20,880 APES were calculated for each model selection method. As their relative frequency distribution displayed a strong positive skew, their central tendency was summarised by their median. The effect of sampling error on their median was gauged with a bootstrap study. The 20,880 APES were treated as a surrogate population. Five hundred samples of size 20,880 were randomly selected with replacement. The distribution of the medians of these 500 samples is summarised by its median and 90% confidence interval shown for each model selection method in Fig. 1a. It may be observed that:

a) The confidence intervals are quite tight so that the effect of sampling error has largely been eliminated by the use of the large surrogate population, thereby reducing room for ambiguity in the comparison of the various model selection methods.

b) There are no marked differences between the various information criteria.

c) Information criteria are better for model selection than the prediction validation approach.

d) The encompassing approach, to the surprise of the authors, is just as effective as the information criteria.

Observation c) is important because the prediction validation approach is often recommended and used in practice (Makridakis et al., 1998). The reason for its inferior performance is most likely to be the fact that it does not rely on the entire sample for the fitting phase and so the estimates of parameters lack the level of statistical efficiency that the other approaches have.

The effect of sample size, collation period and prediction horizon are shown in Figs. 1b, c and d. To interpret these results, it should be understood that:

a. Small sample is \(n=24\) for annual and quarterly series, and \(n=48\) for monthly series;
b. Large sample is \(n=40\) for annual series, \(n=60\) for quarterly series and \(n=96\) for monthly data;
c. The short run consists of future periods 1–3 for annual series, 1–4 for quarterly series and 1–9 for monthly series;
d. The long run comprises future periods 4–6 for annual data, 5–8 for quarterly data and 10–18 for monthly data.

Partitioning the original sample of 20,880 APES according to these factors results in many smaller surrogate populations, their sizes being those shown in the second column of Table 2. The bootstrap was applied as described above to these smaller samples to again obtain the median and 90% confidence intervals of the median APES. The results obtained for sample size and prediction horizon were as expected. Larger sample sizes lead to smaller APES. There is more volatility in long run predictions. A key point, however, is that the basic conclusions above remain unchanged for these cases. The poorer performance of the prediction evaluation method does not change with sample size or prediction horizon. The results for the collation period tell a slightly different story. They

![Fig. 1. Simulation results (median APEs and 90% confidence intervals for the criteria given in Table 1).](image-url)
should be highly correlated with the sample size results because the monthly time series are longer than the quarterly time series, which in turn are longer than the annual series. Interestingly, in the case of the annual series, the median for prediction validation is lower than the medians of the other approaches, but the wider confidence intervals suggest that this difference is not statistically significant.

5. Application to the M3 competition data

The use of simulated data does raise the criticism that in real life the true model is unknown. Furthermore, real series are not as well behaved as the simulated series. This happens even when random errors and outliers are included in the simulated series. In this section, we investigate how forecasting performance is affected by the eight approaches to model selection on real data.

The eight model selection approaches were applied to the M3 competition data (Makridakis & Hibon, 2000) to see whether the results of simulated data carry through for real data. It was necessary to remove time series that were too short. Each time series had a tuning sample of a specified size \( n \) and an additional evaluation sample of size \( r \) where again \( r=6 \) for annual data, \( r=8 \) for quarterly data, and \( r=18 \) for monthly data. For the prediction validation approach, it was necessary to fit models to \( \frac{n}{C_0}r \) observations where \( n \) must exceed \( r \). Moreover, to obtain plausible results with a satisfactory level of statistical reliability,

![Fig. 2. M3 data results (median APEs and 90% confidence intervals for the criteria given in Table 1).](image-url)
the focus of the study was restricted to those series where \( n \geq 20 \) for annual data, \( n \geq 28 \) for quarterly data, and \( n \geq 72 \) for monthly data. After culling series that are too small according to this definition, 1452 of the 2829 time series in the M3 data remained for the comparative study.

The procedures that were described in Steps 2 and 3 of Section 3 for the simulation study were applied to the 1452 time series from the M3 competition data. The results involved 35,622 APES for each method. The bootstrapped sampling distribution of medians is summarised in Fig. 2a. The results are similar to those from the simulation study except that the encompassing approach is no longer competitive with the information criteria.

When examined by sample size, collation period and prediction horizon, as shown in Figs. 2b, c and d, the results are again similar to those from the simulation study. Now, however, the quarterly time series has the lowest MdAPES. Moreover, although the validation method appears to be better than its competitors for this particular case, the difference is barely significant at the 90% level. The distinction between the small and large samples used in this study depends on the collation period. A small sample was one with a length that did not exceed the median length of all the samples with a corresponding collation period.

6. Conclusions

In this paper various approaches to model selection were compared using time series simulated from statistical models underlying exponential smoothing. They were also evaluated using a subset of the time series from the M3 competition database. Their relative performances were judged in terms of the prediction capacities of their selected models.

Results from the simulated and real time series data proved to be remarkably consistent. They indicated that there is little to distinguish the various information criteria. The most important finding was that the information criteria approaches appear to be superior to the commonly used prediction validation approach. Moreover, there was some evidence, using the real data, that suggested that the information criteria approaches are better than the encompassing approach.

Overall, these studies indicate that the best practice would be to adopt an information criteria approach for choosing between the various common exponential smoothing methods. They indicate that the AIC probably has a slight edge over its counterparts. As expected, however, the FPE had a similar performance to the AIC, reflecting the asymptotic equivalence of these criteria.

References


Meeting of the Institute of Mathematical Statistics, Manhattan, Kansas.


