Journal of Applied Statistics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/cjas20

Holt-Winters Forecasting: An Alternative Formulation Applied to UK Air Passenger Data

J.D. Bermúdez [a], J.V. Segura [b] & E. Vercher [a]

[a] Depto. Estadística e I.O., Fac. Matemáticas, Universitat de València, Burjassot, Spain

[b] Universidad Miguel Hernández de Elche, Centro de Investigación Operativa, Elche, Spain

Published online: 15 Nov 2007.

Journal of Applied Statistics
Vol. 34, No. 9, 1075–1090, November 2007

Holt–Winters Forecasting: An Alternative Formulation Applied to UK Air Passenger Data

J.D. BERMÚDEZ*, J.V. SEGURA** & E. VERCHER*

*Depto. Estadística e I.O., Fac. Matemáticas, Universitat de València, Burjassot, Spain;
**Universidad Miguel Hernández de Elche, Centro de Investigación Operativa, Elche, Spain

ABSTRACT *This paper provides a formulation for the additive Holt–Winters forecasting procedure that simplifies both obtaining maximum likelihood estimates of all unknowns, smoothing parameters and initial conditions, and the computation of point forecasts and reliable predictive intervals. The stochastic component of the model is introduced by means of additive, uncorrelated, homoscedastic and Normal errors, and then the joint distribution of the data vector, a multivariate Normal distribution, is obtained. In the case where a data transformation was used to improve the fit of the model, cumulative forecasts are obtained here using a Monte-Carlo approximation. This paper describes the method by applying it to the series of monthly total UK air passengers collected by the Civil Aviation Authority, a long time series from 1949 to the present day, and compares the resulting forecasts with those obtained in previous studies.*

KEY WORDS: Exponential smoothing, time series forecasting, prediction intervals, linear model, additive error, Monte-Carlo methods

Introduction

Exponential smoothing methods are among the most widely used forecasting techniques in industry and business. Forecasting competitions (Makridakis *et al.*, 1998; Makridakis & Hibon, 2000) have reported the surprising forecasting accuracy of Holt–Winters methods, obtained with minimal effort in computation and model identification. In addition, recent developments provide a class of state space models for which exponential smoothing methods are optimal (Ord *et al.*, 1997; Hyndman *et al.*, 2002), this class contains both ARIMA models and dynamic nonlinear statistical models.

While Holt–Winters methods produce accurate forecasts, they do not provide good prediction intervals. A lot of different formulae have been proposed for obtaining prediction intervals, but simulated and empirical studies (Yar & Chatfield, 1990; Ord *et al.*, 1997; Koehler *et al.*, 2001) have shown that the proposed intervals tend to be too narrow, in the sense that more observations than expected fall outside the prediction intervals, although unreasonably wide intervals have also been reported (Grubb & Mason, 2001).

Correspondence Address: E. Vercher, Depto. Estadística e I.O., Fac. Matemáticas, Universitat de València, Doctor Moliner 50, E-46100 Burjassot, Spain.

0266-4763 Print/1360-0532 Online/07/0901075–16
DOI: 10.1080/02664760701592125

The estimation of the unknowns in Holt–Winters methods is another controversial question. The smoothing parameters are usually estimated by minimising the one-step-ahead mean squared errors, once the starting values have been fixed. Some heuristic algorithms have been developed for estimating the starting values, although no empirical evidence seems to favour any particular method (Segura & Vercher, 2001). Recently, the joint estimation of smoothing parameters and starting values has been proposed using maximum likelihood estimation (Ord *et al.*, 1997; Hyndman *et al.*, 2002) and other optimisation methods (Segura & Vercher, 2002; Bermúdez *et al.*, 2006a, b).

In this paper we apply the additive Holt–Winters method and introduce the stochastic component of the model by means of additive, uncorrelated, homoscedastic and Normal distributed errors. The joint distribution of the data vector, giving the unknowns, is then a multivariate Normal distribution. We obtain their moments, mean vector and covariance matrix, as functions of the smoothing parameters and the initial conditions of the series. We then show that the Holt–Winters method can be formulated as a linear heteroscedastic model, whose coefficients are given by the initial conditions, and that its covariance matrix only depends on the smoothing parameters.

That alternative formulation allows us, first, to optimise the smoothing parameters and the initial values jointly and, second, facilitates obtaining reliable prediction intervals. In addition, our approach also allows us to work with series with multiplicative seasonality and multiplicative random error by means of suitable data transformations.

This paper is organised as follows. In the next section we describe the multivariate linear model for the additive Holt–Winters method, we also present the explicit formulae for the calculation of the maximum likelihood estimators and some computational issues. The third section is devoted to the development of the point forecast and prediction intervals. The fourth section presents our experiences in fitting and forecasting the UK airline passenger data series. A simulation study to measure the performance of our approach for calculating prediction intervals under normal errors is included in the fifth section. The paper ends with some concluding remarks.

A Multivariate Model for the Additive Holt–Winters Method

The additive Holt–Winters method is usually defined through the transition equations:

$$\begin{array}{lll} a_i & = \alpha(y_i - c_{i-p}) + (1-\alpha)(a_{i-1} + b_{i-1}) & \text{(level equation)} \\ b_i & = \beta(a_i - a_{i-1}) + (1-\beta)b_{i-1} & \text{(slope equation)} \\ c_i & = \gamma(y_i - a_{i-1} - b_{i-1}) + (1-\gamma)c_{i-p} & \text{(seasonal components equation)} \end{array}$$

where $\boldsymbol{Y} = (y_1, \ldots, y_n)'$ are the observed data, p is the length of the seasonal cycle and $\theta = (\alpha, \beta, \gamma)'$ is the vector of smoothing parameters. Using the vector of initial conditions, $\omega = (a_0, b_0, c_{1-p}, \ldots, c_0)'$, and recursively applying the transition equations, the h steps ahead prediction usually proposed is $\hat{y}_{n+h} = a_n + h\, b_n + c_{n+h-p}$. In practice, however, both the vector of initial conditions ω and the vector of smoothing parameters θ are unknown and have to be estimated from the data. The above equation for updating the seasonal index is similar to, but is not, that originally proposed by Winters (1960); instead, we used here the updated equation proposed by Ord *et al.* (1997), see also Hyndman *et al.* (2002).

We introduce the stochastic component of the model by means of additive, uncorrelated, homoscedastic and normally distributed errors. So, we assume that the ith observation y_i comes from the random variable

$$Y_i = a_{i-1} + b_{i-1} + c_{i-p} + \epsilon_i \tag{1}$$

where the error vector $\varepsilon = (\epsilon_1, \ldots, \epsilon_n)'$ follows a $\mathrm{N}_n(0, \sigma^2 I_n)$ distribution.

Then the following state space equations are obtained: $a_i = a_{i-1} + b_{i-1} + \alpha\epsilon_i$, $b_i = b_{i-1} + \alpha\beta\epsilon_i$ and $c_i = c_{i-p} + \gamma\epsilon_i$, substituting y_i in the transition equations by using equation (1). So, our formulation coincides with the one proposed by Ord *et al.* (1997) for the additive Holt–Winters forecasting procedure as a state-space model with a single source of error.

Using recursively equation (1) together with the transition equations, the data can be stated in terms of the initial conditions and smoothing parameters in the following way:

$$\begin{aligned}
Y_1 &= a_0 + b_0 + c_{1-p} + \epsilon_1 \\
Y_2 &= a_0 + 2b_0 + c_{2-p} + \alpha(1+\beta)\epsilon_1 + \epsilon_2 \\
Y_3 &= a_0 + 3b_0 + c_{3-p} + \alpha(1+2\beta)\epsilon_1 + \alpha(1+\beta)\epsilon_2 + \epsilon_3 \\
&\ \vdots \\
Y_{p+1} &= a_0 + (p+1)b_0 + c_{1-p} + \gamma\epsilon_1 + \alpha\Sigma_{r=1}^{p}(1+\beta(p+1-r))\epsilon_r + \epsilon_{p+1}
\end{aligned} \tag{2}$$

and so on. The matrix form of those equations is $\boldsymbol{Y} = A\omega + L\,\varepsilon$, where L is the $n \times n$ low triangular matrix given by:

$$L = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ l_2 & 1 & 0 & \cdots & 0 & 0 \\ l_3 & l_2 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ l_{n-1} & l_{n-2} & l_{n-3} & \cdots & 1 & 0 \\ l_n & l_{n-1} & l_{n-2} & \cdots & l_2 & 1 \end{pmatrix} \tag{3}$$

with $l_i = \alpha(1 + (i-1)\beta) + \gamma(i = 1 \bmod p)$ for $i = 2, \dots, n$; A is the $n \times (p+2)$ matrix whose first column is the vector $(1, 1, \cdots, 1)'$, its second column is the vector $(1, 2, \cdots, n)'$ and its last p columns consist of blocks of identity $p \times p$ matrices stacked one upon the other to cover the n rows.

Therefore the joint distribution of the data vector $\boldsymbol{Y}$ is multivariate Normal with mean $\mathrm{E}(\boldsymbol{Y}) = A\,\omega$ and variance–covariance matrix $\mathrm{V}(\boldsymbol{Y}) = \sigma^2 LL'$. This variance matrix depends on the smoothing vector θ, but it is always a positive definite matrix whatever the value of θ.

The design matrix A is constant and known, it does not depend either on the smoothing parameters or on the initial conditions, and it is not of complete rank because its first column is the sum of the last p columns. In fact $\mathrm{rank}(A) = p + 1$. So, the initial conditions are non-estimable functions and the model is not identifiable: a linear constraint is needed. Without loss of generality, and just for mathematical convenience, we will use here the restriction $a_0 + b_0 = 0$ instead of the commonly used constraint $c_{1-p} + \cdots + c_0 = 0$. Note that if the maximum likelihood estimator of the initial condition parameters vector subject to the constraint $a_0 + b_0 = 0$, $(\hat{a}_0, \hat{b}_0, \hat{c}_{1-p}, \dots, \hat{c}_0)'$, is obtained, then the maximum likelihood estimators subject to the usual constraint $c_{1-p} + \cdots + c_0 = 0$ can be obtained easily using the invariance property of the maximum likelihood estimation. In fact, they are given by: $\tilde{b}_0 = \hat{b}_0$, $\tilde{a}_0 = \hat{a}_0 + (\hat{c}_{1-p} + \cdots + \hat{c}_0)/p$ and $\tilde{c}_i = \hat{c}_i - (\hat{c}_{1-p} + \cdots + \hat{c}_0)/p$ for $i = 1 - p, \dots, 0$.

In our approach, using the constraint $a_0 + b_0 = 0$, the matrix representation of the $\boldsymbol{Y}$ data vector becomes:

$$\boldsymbol{Y} = M\psi + L\varepsilon \tag{4}$$

where $\psi = (b_0, c_{1-p}, \dots, c_0)'$ is the new vector of initial conditions which is now identifiable. Then M is a known $n \times (p+1)$ complete rank matrix, whose first column is given by the vector $(0, 1, \ldots, n-1)'$ and its last p columns are built with blocks of identity $p \times p$ matrices stacked one upon the other to cover the n rows.

Parameter Estimation

Let us assume that the model is given by equation (4), then the log-likelihood function of the data vector $\boldsymbol{Y}$ is:

$$-\frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}(\boldsymbol{Y} - M\psi)'(LL')^{-1}(\boldsymbol{Y} - M\psi) \tag{5}$$

Let us now introduce some useful notation. Let X be the matrix $L^{-1}M$, let P_X be the orthogonal projection matrix on the vector space generated by the columns of the X matrix, $P_X = X(X'X)^{-1}X'$, and let $\tilde{\psi}$ be the usual mean square estimator of ψ in linear heteroscedastic models, $\tilde{\psi} = (X'X)^{-1}X'L^{-1}\boldsymbol{Y}$. Hence, the quadratic form in equation (5) can be decomposed as:

$$\begin{aligned}
&(\boldsymbol{Y} - M\psi)'(LL')^{-1}(\boldsymbol{Y} - M\psi)\\
&\quad= (L^{-1}\boldsymbol{Y} - X\tilde{\psi} + X(\tilde{\psi} - \psi))'(L^{-1}\boldsymbol{Y} - X\tilde{\psi} + X(\tilde{\psi} - \psi))\\
&\quad= (\tilde{\psi} - \psi)'X'X(\tilde{\psi} - \psi) + (L^{-1}Y)'(I - P_X)L^{-1}\boldsymbol{Y}
\end{aligned}$$

so, the log-likelihood function is:

$$-\frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\,(\tilde{\psi} - \psi)'X'X(\tilde{\psi} - \psi) - \frac{1}{2\sigma^2}\,(L^{-1}\boldsymbol{Y})'(I - P_X)L^{-1}\boldsymbol{Y}$$

The first quadratic form in the above expression can always be annulled, whatever the value of θ, while the second quadratic form involves only the parameter θ, which appears in the matrix L as well as in the matrix $X = L^{-1}M$. So, $\hat{\theta}$, the maximum likelihood estimator of the smoothing vector $\theta = (\alpha, \beta, \gamma)'$ is obtained by minimising

$$\min_{\alpha,\beta,\gamma}(L^{-1}\boldsymbol{Y})'(I - P_X)L^{-1}\boldsymbol{Y} \tag{6}$$

Once $\hat{\theta}$ has been obtained, let $\hat{L}$ be the matrix L computed at $\hat{\theta}$ and $\hat{X} = \hat{L}^{-1}M$. The maximum likelihood estimator of ψ is then given by the vector $\tilde{\psi}$ computed at $\hat{\theta}$, *i.e.*

$$\begin{aligned}
\hat{\psi} &= (\hat{X}'\hat{X})^{-1}\hat{X}'\hat{L}^{-1}\boldsymbol{Y}\\
&= (M'\hat{L}'^{-1}\hat{L}^{-1}M)^{-1}M'\hat{L}'^{-1}\hat{L}^{-1}\boldsymbol{Y}
\end{aligned} \tag{7}$$

Finally, the maximum likelihood estimator of σ^2 is:

$$\begin{aligned}
\hat{\sigma^2} &= \frac{1}{n}\min_{\alpha,\beta,\gamma}(L^{-1}\boldsymbol{Y})'(I - P_X)L^{-1}\boldsymbol{Y} = \frac{1}{n}\,(\hat{L}^{-1}\boldsymbol{Y})'(I - P_{\hat{X}})\hat{L}^{-1}\boldsymbol{Y}\\
&= \frac{1}{n}\boldsymbol{Y}'\hat{L}'^{-1}(I - \hat{L}^{-1}M(M'\hat{L}'^{-1}\hat{L}^{-1}M)^{-1}M'\hat{L}'^{-1})\hat{L}^{-1}\boldsymbol{Y}
\end{aligned} \tag{8}$$

In fact we only need to solve an optimisation problem, the one given in equation (6) with respect to three decision variables, the three smoothing parameters; the estimators of the initial conditions and variance are obtained analytically with equations (7) and (8), respectively. On the contrary, a common practice in exponential smoothing procedures is to estimate ω, the initial conditions vector, using a heuristic procedure independently of the smoothing parameter vector θ and then to estimate θ, minimising the one-step-ahead mean square error. For state space models, Harvey (1989) also proposed treating the initial state vector as extra parameters that need to be estimated as part of the maximum likelihood procedure.

It must be pointed out – see the system of equations given by equation (2)– that the one-step-ahead errors, $\hat{\epsilon}_i = y_i - E(Y_i|y_1, \ldots, y_{i-1})$, can be computed as $L^{-1}(\boldsymbol{Y} - M\psi)$, therefore to minimise the one-step-ahead mean square error is equivalent to maximise the log-likelihood function given in equation (5).

Computational Remarks

The matrix calculus involved in the computation of equations (6), (7) and (8) is easy and fast to implement. Note that L is a triangular matrix that can be represented only through its first column – see equation (3) – so its inverse L^{-1} will have a similar form. Moreover, the first column of L^{-1} is given by the vector $(l^1, l^2, \ldots, l^n)'$, where $l^1 = 1$ and its other components can be obtained recursively from the equations $\Sigma_{i=1}^{k} l^i l_{k-i+1} = 0$, for $k = 2, \ldots, n$.

In addition, the computation of the orthogonal projection matrix P_X is numerically accurate enough using the singular value decomposition of matrix $X = L^{-1}M$. Let $X = UDV$ be the singular value decomposition of matrix X, where D is a $(p+1) \times (p+1)$ diagonal matrix, V is an orthogonal $(p+1) \times (p+1)$ matrix and U is a $n \times (p+1)$ matrix with orthogonal columns, then $P_X = X(X'X)^{-1}X' = UU'$. Moreover, the product matrix $(X'X)^{-1}X'$ needed to compute equation (7) is equal to $V'D^{-1}U'$.

On the other hand, the objective function of the optimisation problem given by equation (6) is clearly nonlinear with respect to the decision variables (*i.e.* the smoothing parameters), so for solving it a nonlinear optimisation procedure should be used. Since the smoothing parameters are constrained to the range $0 - 1$ it is not possible to use pure unconstrained optimisation methods but bound constrained nonlinear minimisation procedures.

Forecasting the Time Series

Now let $\boldsymbol{Y}_1$ be the $n \times 1$ vector of observed data and let $\boldsymbol{Y}_2$ be the $h \times 1$ vector of future data. Consider the joint $(n+h) \times 1$ vector $\boldsymbol{Y} = (\boldsymbol{Y}_1', \boldsymbol{Y}_2')'$ and suppose that it still follows the distribution given by equation (4), where the vector ε and the matrices M and L are partitioned in a similar way to the vector $\boldsymbol{Y}$:

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \end{pmatrix} = \begin{pmatrix} M_1 \\ M_2 \end{pmatrix} \psi + \begin{pmatrix} L_1 & 0 \\ L_{21} & L_2 \end{pmatrix} \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix}$$

Under those assumptions (see, for example, Seber, 1984) the conditional distribution of $\boldsymbol{Y}_2$ given $\boldsymbol{Y}_1$ is multivariate Normal with mean $\mu_{2.1}$ and variance matrix $\Sigma_{2.1}$ given by:

$$\begin{aligned}
\mu_{2.1} &= M_2\psi + L_{21}L_1^{-1}(Y_1 - M_1\psi) \\
\Sigma_{2.1} &= \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} \\
&= \sigma^2(L_{21}L_{21}' + L_2L_2') - \sigma^2 L_{21}L_1'(L_1L_1')^{-1}L_1L_{21}' \\
&= \sigma^2 L_2 L_2'
\end{aligned}$$

Point forecasts are then given by an estimator of the prediction mean, $\mu_{2.1}$. We propose to use the following one:

$$\hat{\mu}_{2.1} = M_2\hat{\psi} + \hat{L}_{21}\hat{L}_1^{-1}(\boldsymbol{Y}_1 - M_1\hat{\psi}) \tag{9}$$

Notice that $\hat{L}_1^{-1}(\boldsymbol{Y}_1 - M_1\hat{\psi})$ is the vector of one-step-ahead observed errors, so the forecasting vector given in the last expression agrees with the usual Holt–Winters predictor, but is computed using the estimates of ψ and θ obtained in the previous section.

In order to obtain prediction intervals, we are first going to suppose that the vector θ is known. In that case, $\hat{\psi}$ is the ordinary least square estimator, so it is unbiased, its variance is given by $\sigma^2(X'X)^{-1}$, its sampling distribution is Normal and independent of the random variable $n\hat{\sigma^2}/\sigma^2$, which follows a χ^2 distribution with degrees of freedom given by the rank of the matrix $I_n - P_X$, which is: $n - p - 1$.

Consider now the random vector $\boldsymbol{Y}_2 - \hat{\mu}_{2.1}$:

$$\begin{aligned}\boldsymbol{Y}_2 - \hat{\mu}_{2.1} &= M_2\psi + L_{21}\varepsilon_1 + L_2\varepsilon_2 - M_2\hat{\psi} - L_{21}L_1^{-1}(M_1\psi + L_1\varepsilon_1 - M_1\hat{\psi})\\ &= M_2(\psi - \hat{\psi}) + L_2\varepsilon_2 - L_{21}L_1^{-1}M_1(\psi - \hat{\psi})\\ &= (M_2 - L_{21}L_1^{-1}M_1)(\psi - \hat{\psi}) + L_2\,\varepsilon_2\end{aligned}$$

The estimator $\hat{\psi}$ is only a function of the random vector ε_1, independent of ε_2, therefore both summands in the last expression are independent, hence:

$$\begin{aligned}\mathrm{E}(\boldsymbol{Y}_2 - \hat{\mu}_{2.1}) &= (M_2 - L_{21}L_1^{-1}M_1)\mathrm{E}(\psi - \hat{\psi}) + L_2\mathrm{E}(\varepsilon_2) = 0\\ \mathrm{V}(\boldsymbol{Y}_2 - \hat{\mu}_{2.1}) &= (M_2 - L_{21}L_1^{-1}M_1)\mathrm{V}(\psi - \hat{\psi})(M_2 - L_{21}L_1^{-1}M_1)' + L_2\mathrm{V}(\varepsilon_2)L_2'\\ &= \sigma^2\left[(M_2 - L_{21}L_1^{-1}M_1)(M_1'L_1'L_1'^{-1}M_1)^{-1}(M_2 - L_{21}L_1^{-1}M_1)' + L_2L_2'\right]\end{aligned}$$

Moreover, the sampling distribution of $\boldsymbol{Y}_2 - \hat{\mu}_{2.1}$ is multivariate Normal, because it is a linear combination of the random vector ε, and it is independent of $\hat{\sigma^2}$, because it is only a function of the random vectors $\hat{\psi}$ and ε_2, both independent of $\hat{\sigma^2}$. Therefore, if $S = \sigma^{-2}\mathrm{V}(\boldsymbol{Y}_2 - \hat{\mu}_{2.1})$ and $v \neq 0$ is any known constant vector,

$$t_v = \sqrt{\frac{n-p-1}{n}}\frac{1}{\hat{\sigma}}(v'Sv)^{-1/2}\,v'(\boldsymbol{Y}_2 - \hat{\mu}_{2.1}) \tag{10}$$

follows a t-Student distribution with $n - p - 1$ degrees of freedom. This result allows us to build exact prediction intervals for different goals. For example, the one-step-ahead prediction interval is built using $v = (1, 0, \ldots, 0)'$, while the cumulative prediction interval for the first h steps is obtained by using $v = (1, \ldots, 1)'$. The interval for any other linear combination of the predictions is obtained in a similar way.

In the usual case where the vector θ is unknown, we propose to obtain an approximation to the prediction intervals using equation (10) with $\hat{\theta}$ instead of θ, approaching $\mathrm{V}(\boldsymbol{Y}_2 - \hat{\mu}_{2.1})$ with:

$$\hat{\sigma}^2(M_2 - \hat{L}_{21}\hat{L}_1^{-1}M_1)(M_1'\hat{L}_1'\hat{L}_1'^{-1}M_1)^{-1}(M_2 - \hat{L}_{21}\hat{L}_1^{-1}M_1)' + \hat{\sigma}^2\hat{L}_2\hat{L}_2' \tag{11}$$

The second summand in this variance-covariance matrix, $\hat{\sigma}^2\,\hat{L}_2\hat{L}_2'$, is the usual variance proposed to compute prediction intervals (Yar & Chatfield, 1990) that used to provide intervals that were too narrow. Our variances are greater because they incorporate the uncertainty about the initial values of the series, adding the first summand in equation (11), and hence the intervals should be wider.

Transformed Data

Model (4) could be unsatisfactory for data in the original scale, but may be appropriate after an adequate transformation. For example, model (4), with data in logarithmic scale, can be used as an alternative to the completely multiplicative Holt–Winters model. In the case of monotonic transformations, if the goal is the prediction of the series at a specific lag – i.e. v is a unitary vector – point and interval prediction is straightforward, a matter of just back-transforming the results obtained above.

If the purpose is a cumulative prediction, or in general if v is not an unitary vector, equation (10) is not helpful, but a Monte Carlo approximation is always feasible. Note that in the transformed scale $S^{-1/2}\ (\boldsymbol{Y}_2 - \hat{\mu}_{2.1})\ /\hat{\sigma}$ is a random vector with independent t-Student components, and therefore it is easy to obtain a random sample from the joint distribution of $\boldsymbol{Y}_2$. Accumulating that sample after back-transforming it, a random sample from the cumulative prediction distribution for the original scale is obtained. The desired forecast is obtained with the sample mean (or the sample median) of this last sample, and prediction intervals are obtained using appropriate sample quantiles.

UK Air Passenger Data

The time series used in this paper is the UK airline total passenger numbers from the Civil Aviation Authority, which is available from its website (http://www.caa.co.uk). This long time data set series includes the monthly total air passenger numbers at UK airports from January 1949 to December 2004. Following a common approach to dealing with this type of seasonality, Grubb & Mason (2001) consider a power transformation (Box & Cox, 1964) to make both the seasonality and the random component additive. The data have an upward trend together with seasonal variation whose size is roughly proportional to the local mean level; that is, they have multiplicative seasonality (see Figure 1). In our approach the power transformation with $\lambda = 0.25$ is used, since the variance of the transformed data seems to be stabilised for this parameter value. We will present the performance of our approach and evaluate the out-of-sample forecasts and the coverage of the prediction intervals for this data set.

This series corresponds to the aggregation of terminal passengers at all UK airports, for all passenger types, origins and destinations. The aggregation of these components reduces the shifts and variation in demand between airports or destinations, for example. The airport terminal capacities are also measured in numbers of passengers (pax) for capacity planning purposes.

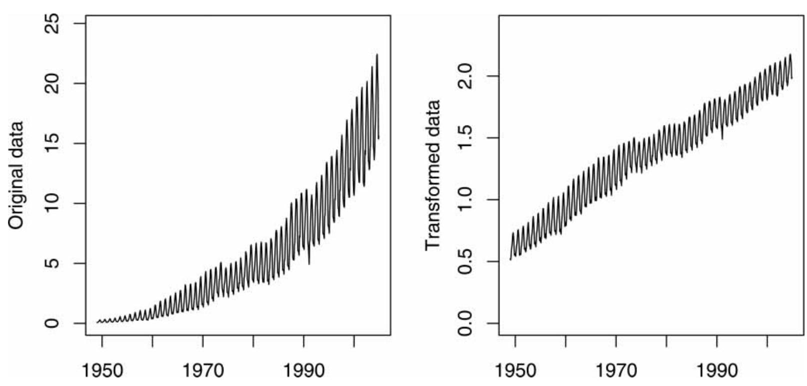

Figure 1. UK air passengers monthly series from January 1949 to December 2004. Left, data in the original scale (million passengers per month). Right, transformed data (the transformation used is the fourth root)

We choose to use the public domain software (R Development Core Team, 2005) to carry out our study. We use various R functions included in its base installation package in conjunction with some functions written by the authors. Specifically, to calculate the function of the fitting errors, we use the singular value decomposition given by the R function `svd` and its minimisation has been done using the R function `optim`. This last function includes an option for box-constrained optimisation. In fact, we have an unconstrained optimisation problem with three decision variables that have lower and upper bounds. For solving these kinds of problems the R package uses a limited-memory modification of the BFGS quasi-Newton method (Byrd *et al.*, 2005) which allows box-constraints.

Monthly Forecasts

The results reported here for UK air passengers were computed by fitting the additive Holt–Winters method in five different settings. We use the first 39, 44, 47, 51 and 56 years of transformed monthly data to make forecasts and measure their accuracy using the remaining data: the last 204, 144, 108 and 60 monthly observations respectively. The initial conditions and smoothing parameters were chosen by minimising the mean square one-step-ahead error (MSE) over the fitting period data, that is maximising the log-likelihood function (equation (5)). The procedure performs a multi-search to locate the 'best-practice' optimal solution. In fact, it makes consecutive restarts of the fitting algorithm, with 27 different initial guesses for the unknowns, and typically found several different local minima. Once the estimators have been calculated for a given set of monthly observations in the fitting phase, they are used for computing both the point forecasts and interval predictions, without any re-estimation. As a measure of fitting and forecast accuracy the mean absolute percentage error (MAPE) is used.

The long lead-time is usually recommended for aviation forecasting and it would be useful to present the prediction intervals for different horizons, since the length of the data series allows us to do it. The results for the medium and short lead-time are included in order to measure the performance of the model with respect to the effect of the decrease in the number of air passengers due to the two Gulf Wars and September 11. In addition, in each setting we use the available data to compute maximum likelihood estimates and to calculate forecast values for the different forecasting horizons.

Table 1 shows that there is a little variation both in the smoothing parameter values of the linear model and the initial components of level and trend, which have been evaluated for the transformed data. In addition, the measures of fitting error, which have been calculated for the raw data for all the observed periods, are similar, being very different from the post-sample accuracy shown in Table 2. Note that $\beta = 0$ throughout which implies that there are no variations in the trend whose estimate is $b_0 = 0.002$.

Table 1. Smoothing parameters, initial values, RMSE and MAPE for the UK passenger data across different fitting periods

Fitting phase	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\gamma}$	a_0	b_0	RMSE	MAPE
1949 to 1987	0.276	0.000	0.657	0.595	0.002	0.133	3.977
1949 to 1992	0.310	0.000	0.628	0.596	0.002	0.167	3.922
1949 to 1995	0.300	0.000	0.610	0.596	0.002	0.172	3.800
1949 to 1999	0.305	0.000	0.600	0.596	0.002	0.181	3.636
1949 to 2004	0.314	0.000	0.577	0.596	0.002	0.226	3.531

Figure 2 shows the monthly forecasts for the following six years obtained in each one of the four settings considered; that is, the observations up to December 1987, 1992, 1995 and 1999, respectively. In each case the observed data, the out-of-sample forecasts and the 90% prediction intervals for the following six years are shown. The observed data are plotted using a thicker line, while the dashed lines show the prediction intervals and the solid line represents the out-of-sample forecasts.

Notice that in Figure 2 we are using some overlapping forecast horizons (the same as also used in Tables 1 and 2) in order to visualise the effect of the decrease in passenger numbers after the first Gulf War. So, when the UK airline passenger series was fitted with the observations up to 1987, the forecast is accurate enough up to the summer of 1990, when the first Gulf War began, but after that date the forecast becomes too optimistic, although the observed data are outside but very close to the prediction intervals. If the fitting period includes data up to December 1992, only a few months after the first Gulf War, the actual values are bigger than the forecast ones, showing that the forecast is not able to predict the recovery of the series once the effect of the first Gulf War had passed. Instead, when

Table 2. Post-sample MAPE for the UK passenger data across different forecasting horizons

Fitting phase	4 years	7 years	9 years	12 years	17 years
1949 to 1987	6.669	8.515	8.735	8.163	8.177
1949 to 1992	5.538	7.169	7.686	7.645	
1949 to 1995	2.794	3.206	3.125		
1949 to 1999	3.679				

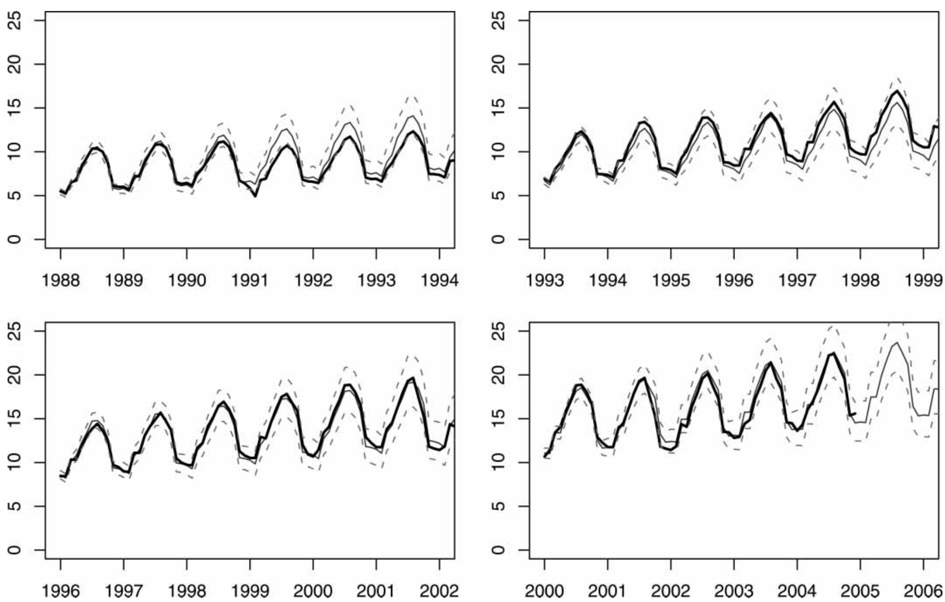

Figure 2. Observed data (thicker line), out-of-sample forecasts (solid line) and 90% prediction intervals (dashed lines) for UK air passenger data, in million of passengers per month. The fitting phase contains data from January 1949 to December 1987, December 1992, December 1995 and December 1999 respectively. The following 72 monthly forecasts are made from January 1988, 1993, 1996 and 2000 respectively

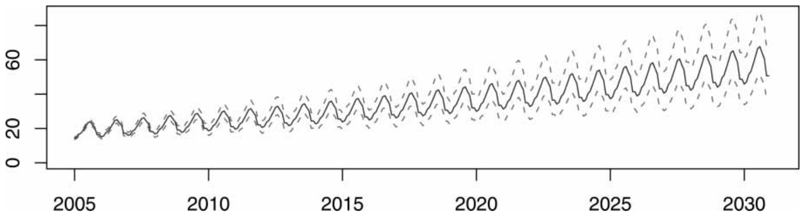

Figure 3. Forecasts and 90% prediction intervals for UK air passenger data, monthly totals in million of passengers

the data series was fitted up to December 1995, the observed data are well-covered by the prediction intervals, even in the year 2001. Finally, in the fourth picture, which includes the observations up to December 1999 for the fitting phase, it is possible to appreciate the decrease in passenger numbers due to the effect of September 11 and the second Gulf War, but in this case the actual values are inside the prediction intervals.

The series of monthly forecasts built with all the available data is shown in Figure 3, and the 90% prediction intervals are plotted there too. It must be noticed that these intervals are growing in length with the prediction lag, but their lengths remain at a reasonable level even for long lead-times (312 months).

The length of the prediction intervals is an increasing function of both the prediction lag and the level of the series. In order to compare their lengths we use a relative measure: for each month we evaluate the ratio between the semi-interval length and the point forecast and, for each year, we calculate the mean percentage of these 12 ratios. Table 3 shows some of these annual mean percentage semi-interval lengths for lags of between one and 26 years, and for different fitting periods: using all available data until 2004 and using only data until 1984, when the level of the series was approximately half the present level. Notice that this measure, the annual mean percentage semi-interval length, is hardly influenced by the level of the series.

Prediction Intervals for Annual Forecasts

Once monthly forecasts are built, annual forecasts can be obtained easily by just accumulating the 12 predictions for each year. However, as a data transformation has been used, obtaining prediction intervals is not so easy.

Here, we will use a Monte Carlo approximation to the annual forecasts and prediction intervals. For each year, we simulate a random sample from the joint predictive distribution of their 12 months, with data in the transformed scale. Then we add up the 12 predictions to obtain the annual forecast of the back-transformed simulated sample, which is a random sample from the predictive distribution for that year. The empirical distribution of this last

Table 3. Annual mean percentage semi-interval length, AMPSIL, of the prediction intervals for two different fitting periods: 1949–2004 and 1949–1984

1949–2004	2005	2006	2007	2008	2009	2010	2015	2020	2025	2030
AMPSIL	5.6	8.3	10.3	12.0	13.4	14.6	19.5	23.1	25.9	28.3
1949–1984	1985	1986	1987	1988	1989	1990	1995	2000	2005	2010
AMPSIL	4.6	7.9	10.2	11.9	13.8	14.9	20.4	24.4	27.5	30.2

Table 4. Forecast terminal passenger numbers at UK airports 2005 to 2030, in millions per annum. The fitting phase contains data from 1949 to 2004

Year	10th percentile	Median	90th percentile
2005	220.4	227.2	234.2
2010	261.7	289.8	320.5
2015	317.2	364.6	418.1
2020	383.5	453.1	533.1
2025	461.0	556.8	669.5
2030	549.8	677.4	828.9

sample is a Monte Carlo approach to the predictive distribution, and from it we can obtain the desired point forecast and prediction interval.

Table 4 shows the median and some percentiles for long lead-time forecasts obtained using all available data. Note that the total terminal passengers in all UK airports during 2004 was 215.6 millions. From our point forecast, which is the median of the predictive distribution, we can say that the actual number of passengers would be duplicated in only 15 years, and that moreover, the 500 million passengers a year would be reached before 2025.

Figure 4 shows the observed yearly data from 1988 on, with small circles, point forecasts are shown with a solid line and the 90% predictive intervals with dashed lines. The point forecasts and predictive intervals are computed at four different times, shown by vertical dashed lines, as was done in the previous subsection. The first forecasts are computed with data up to 1987, and they are plotted from 1988 to 1992, then the model parameters are recomputed using data up to 1992 and plotted from 1993 to 1995, and so on. Notice that although the length of the intervals grows with the forecast lag, it remains at a reasonably wide level in contrast to the intervals obtained by Grubb & Mason (2001).

Econometric models are usually used for long-term planning. In particular, the UK Department of the Environment, Transport and the Regions (DETR) made forecasts of some independent variables, using alternative scenarios, then predicted future passenger demand for different market segments and obtained the aggregate passenger demand by means of the aggregation of the forecasts of the segments.

Table 5 shows a comparison of our forecasts with those proposed in the DETR 2000 study. In order to be comparable, our forecasts have been built using data until December 1998, the same data set used in the DETR 2000 study. That year, 1998, the passenger terminal numbers at UK airports was 159.0 million. From past DETR studies it is known that the mid scenario predictions consistently underestimate the passenger terminal numbers (DETR, 2000), so it is interesting to note that our 2005 point prediction is similar to the

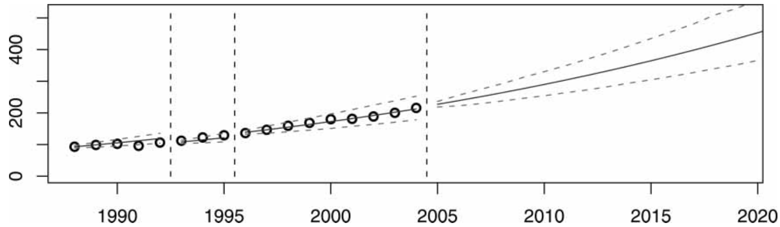

Figure 4. Annual forecast for UK air passenger data, in millions per annum. The dashed lines show the 90% predictive interval while the solid line is the point prediction

Table 5. Comparison of our forecasts for terminal passenger numbers at UK airports, in millions per annum, with DETR forecast. The fitting phase contains data till 1998

	DETR forecast			Our forecasts		
Year	Low	Mid	High	10th percentile	Median	90th percentile
2005	220.5	228.8	237.4	204.5	229.9	257.7
2010	256.8	276.1	296.8	250.9	293.4	341.7
2015	299.5	333.2	370.6	308.7	369.3	438.4
2020	348.5	400.7	460.8	376.5	459.0	558.0

mid scenario one, but for the other horizons our point prediction is comparable to the high scenario.

On the Uncertainty Remaining in the Prediction Intervals

Concerning the influence of our formulation on the width of the prediction intervals note that the vector of initial conditions is always associated with the matrix $M_2 - L_{21}L_1^{-1}M_1$, both in the prediction (equation (9)), and in the variance of the forecasting error (equation (11)). If the vector of smoothing parameters is known and this matrix is null, the uncertainty of initial conditions will not affect the variance and therefore the prediction intervals will coincide with those provided by Yar & Chatfield (1990). The matrix $M_2 - L_{21}L_1^{-1}M_1$ is a continuous function of θ, which can be computed for different values of θ, obtaining that for $\beta > 0$ this matrix converges to zero, as $n \longrightarrow \infty$.

We undertook a simulation study to measure the performance of our approach for calculating prediction intervals under normal errors. The main fact involved in this simulation study is that $\beta = 0$, in such a way that we only compare the effect of adding to the usual variance $\sigma^2\ L_2L_2'$, the new term, $\sigma^2(M_2 - L_{21}L_1^{-1}M_1)(M_1'L_1'L_1'^{-1}M_1)^{-1}(M_2 - L_{21}L_1^{-1}M_1)'$, which incorporates the uncertainty caused by the initial conditions.

The design of our simulation study was intended to explore the neighbourhood, in the parametric space, of the estimates obtained for the UK air passengers data (see Table 1). So, we always used for $c_{-11}, \ldots, c_0$ the values of their estimates obtained for the UK air passenger data, while the specified values used in the simulation for the smoothing parameters and initial conditions are the following: for α were 0.2, 0.3 and 0.4; for γ were 0.5, 0.6 and 0.7; for b_0 were 0.003 and -0.003, and for σ were 0.01 and 0.04. In addition, we simulated series with three different lengths: $n = 120$, 240 and 480 months. All the factor combinations give a total of 108 different simulated scenarios.

For each scenario we randomly generate 1000 realisations of length $3/2n$ of the series. The first n values of each realisation are the fitting data, so we used them to estimate all the parameters and obtain prediction intervals: first with our proposal, computed with the variance estimator giving by equation (11), and second with the usual variance estimator $\hat{\sigma}^2\ \hat{L}_2\hat{L}_2'$. Note that we call the usual prediction intervals those computed with that second variance estimator, although we always estimate the unknowns using equations (6), (7) and (8), which is not the usual point estimation procedure for the Holt–Winters method. The last $1/2n$ values of each realisation are used only to evaluate the coverage of both prediction intervals. For each period, the simulated coverage of the associated prediction interval is the percentage of the sample values of that period that fall within the limits of the intervals.

Figure 5 shows a box-plot summary of the results obtained for $n = 120$ and for a nominal coverage of the prediction intervals of 90%. For each scenario and each year ahead the mean of the simulated coverage of its 12 months have been computed; each box is the box and

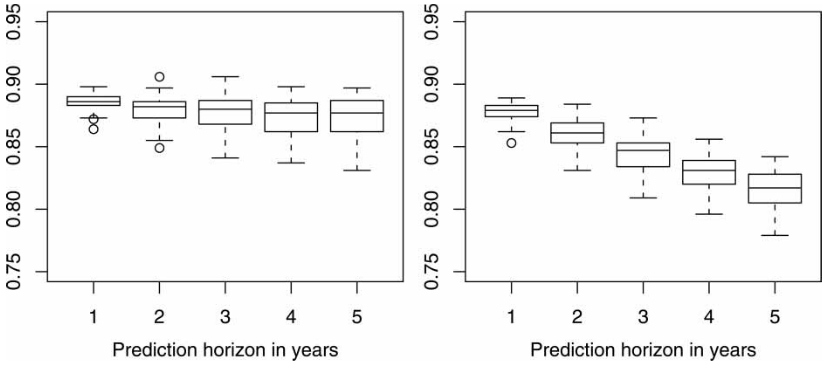

Figure 5. Simulated coverage of the prediction intervals with a nominal coverage of 90%, for $n = 120$. Left: our prediction intervals. Right: the usual ones

Table 6. Simulated coverage of the prediction intervals with a nominal coverage of 90%, for $n = 120$ and different pairs of parameter values. In this simulation study $\beta = 0$. In each cell: our proposal (the usual one)

α	γ	1st year	2nd year	3rd year	4th year	5th year
0.2	0.5	87.1 (86.1)	85.5 (83.7)	85.0 (81.7)	84.2 (80.1)	83.6 (78.3)
0.2	0.6	88.5 (87.8)	87.7 (86.0)	86.7 (83.9)	86.2 (82.3)	86.2 (81.0)
0.2	0.7	88.6 (88.0)	88.0 (86.4)	87.5 (84.9)	87.2 (83.4)	87.0 (82.2)
0.3	0.5	88.1 (87.1)	87.1 (84.9)	86.6 (82.9)	86.2 (81.1)	86.1 (79.7)
0.3	0.6	88.5 (87.7)	87.9 (85.8)	88.2 (84.8)	87.7 (83.1)	87.6 (81.6)
0.3	0.7	89.1 (88.4)	88.7 (86.6)	88.3 (85.2)	88.3 (83.8)	88.3 (82.4)
0.4	0.5	88.7 (87.9)	88.4 (86.1)	88.3 (84.6)	88.4 (83.0)	88.4 (81.8)
0.4	0.6	89.4 (88.5)	89.8 (87.5)	89.8 (86.4)	89.4 (84.7)	89.3 (83.3)
0.4	0.7	89.1 (88.3)	89.1 (87.0)	89.0 (85.7)	88.9 (84.1)	89.0 (83.2)

whiskers representation of those means for the 36 scenarios. The results for our proposed prediction intervals are shown in the figure at the left, the simulated coverage is about two points under the nominal coverage and this difference remains stable as the prediction lag increases. The results for the usual prediction intervals are shown at the right; they are very similar to our results for the first year but they get worse as the prediction lag increases: the simulated coverage of the fifth year ahead is only around 82%.

Table 6 shows a numerical summary of the means used to draw Figure 5. Each row shows the mean of the simulated coverage obtained in four scenarios, those with the same values for α and γ. In each cell, the mean simulated coverage for our proposal appears and, between brackets, for the usual prediction intervals. Our simulated coverage values are still smaller than the 90% nominal value, but they improve substantially those of the usual prediction intervals.

The simulation results obtained with $n = 240$ and $n = 480$ have been very similar, so we only present here the case of $n = 480$ months, which is almost the length of the UK air passenger data used in this paper, see Table 7 and Figure 6. In this case, our simulated coverage values are very close to the 90% nominal value, remaining almost invariable even for a prediction lag of twenty years. On the contrary, the usual prediction intervals show simulated coverage values decreasing quickly as the prediction lag increases. So, when $\beta = 0$, the uncertainty on the smoothing parameters seems not to be influenced by the

Table 7. Simulated coverage of the prediction intervals with a nominal coverage of 90%, for $n = 480$ and different pairs of parameter values. In this simulation study $\beta = 0$. In each cell: our proposal (the usual one)

α	γ	1st year	5th year	10th year	15th year	20th year
0.2	0.5	89.6 (89.5)	89.5 (88.2)	89.2 (86.4)	89.1 (84.8)	89.0 (83.1)
0.2	0.6	89.9 (89.8)	89.5 (88.2)	89.1 (86.5)	89.1 (84.9)	89.0 (83.4)
0.2	0.7	90.3 (90.2)	90.0 (88.9)	89.6 (87.1)	89.7 (85.6)	90.0 (84.9)
0.3	0.5	89.6 (89.5)	89.3 (87.7)	89.3 (85.8)	88.8 (83.7)	89.2 (82.5)
0.3	0.6	89.8 (89.6)	89.8 (88.2)	89.6 (86.4)	89.6 (84.7)	89.5 (83.1)
0.3	0.7	89.5 (89.2)	89.8 (88.4)	89.6 (86.5)	89.3 (84.7)	89.3 (83.0)
0.4	0.5	90.1 (89.9)	90.2 (88.5)	90.1 (86.5)	90.1 (84.9)	90.0 (83.1)
0.4	0.6	89.8 (89.6)	90.0 (88.3)	90.2 (86.8)	90.1 (85.1)	90.0 (83.5)
0.4	0.7	89.7 (89.5)	89.9 (88.3)	89.7 (86.3)	90.2 (85.1)	90.1 (83.2)

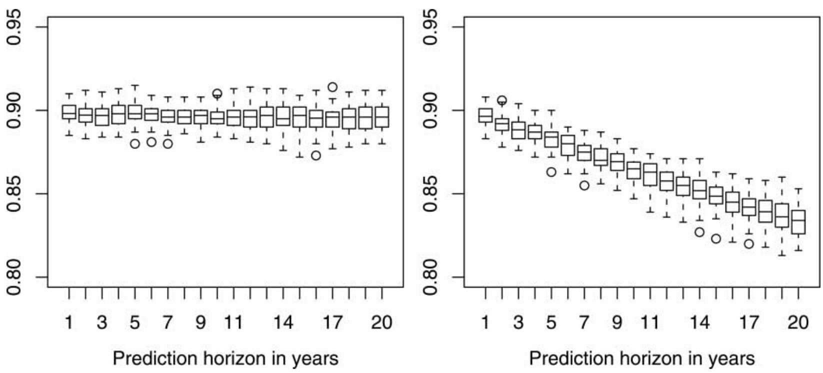

Figure 6. Simulated coverage of the prediction intervals with a nominal coverage of 90%, for $n = 480$. Left: our prediction intervals. Right: the usual ones

coverage of the prediction intervals once the uncertainty on the initial values have been taken into account.

Other nominal coverage of the prediction intervals (from 50% to 99%) have been considered, obtaining the same patterns shown here for the 90% nominal coverage.

Concluding Remarks

Holt–Winters methods are much-used forecasting techniques for univariate time series to support decision planning. The method used to specify initial values for the level, trend and seasonal components has been very influential on this forecasting procedure. Our linear modelling approach allows us to evaluate the values after obtaining the maximum likelihood estimation of the smoothing parameters. Assuming that the stochastic component of the model is introduced by means of additive homocedastic uncorrelated normal errors, it is shown that the conditional distribution of the future data is multivariate normal and approximate prediction intervals can be obtained using the maximum likelihood estimators of the smoothing vector.

The multivariate linear formulation proposed in this paper has two main advantages. First, the joint maximisation of the likelihood function with respect to both smoothing parameters and initial values is greatly simplified: only a numerical optimisation problem in three parameters has to be solved, the optimisation of the other parameters being analytical. Second, this formulation provides an expression for the variance–covariance matrix of the forecast vector that is fast to compute; hence, prediction intervals for individual lag forecasting as well as for linear combinations are easily obtained. For example, the annual prediction intervals proposed here for the UK air passengers series are obtained by computing the variance of the sum of the monthly forecasts.

The actual coverage of the prediction intervals built here is very similar to their nominal coverage, see Figure 2, but no general conclusions could be obtained due to the autocorrelation of the series. The previously proposed prediction intervals are known to be too narrow. However, when $\hat{\beta} = 0$ with our formulation we obtain a variance that is the usual one plus a positive term, so our intervals are wider and should behave better.

It is well-known that the additive Holt–Winters model is inappropriate if the seasonal components or the error variance depend on the level of the series, but it could also be useful after an adequate data transformation, as the UK air passenger series shows. The computational advantages obtained with the multivariate linear formulation of the additive Holt–Winters model made the search for such a transformation worthwhile.

Acknowledgements

The research leading to this paper was partially supported by the Ministerio de Educación y Ciencia of Spain, Grant No. TIN2005-08404-C04-04.

References

Bermúdez, J.D., Segura, J.V. & Vercher, E. (2006a) Improving demand forecasting accuracy using non-linear programming software, *Journal of the Operational Research Society*, 57, pp. 94–100.

Bermúdez, J.D., Segura, J.V. & Vercher, E. (2006b) A decision support system methodology for forecasting of time series based on Soft Computing, *Computational Statistics & Data Analysis*, 51, pp. 177–191.

Box, G.E.P. & Cox, D.R. (1964) An analysis of transformations (with discussion), *Journal of the Royal Statistical Society, B*, 26, pp. 211–252.

Byrd, R.H., Lu, P., Nocedal, J. & Zhu, C. (1995) A limited memory algorithm for bound constrained optimization, *SIAM Journal Scientific Computing*, 16, pp. 1190–1208.

DETR (2000) *Air Traffic Forecast for the United Kingdom* (Department of the Environment, Transport and the Regions).

Grubb, H. & Mason, A. (2001) Long lead-time forecasting of UK air passengers by Holt–Winters methods with damped trend, *International Journal of Forecasting*, 17, pp. 71–82.

Harvey, A. (1989) *Forecasting, Structural Time Series Model and the Kalman Filter* (Cambridge: Cambridge University Press).

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

Koehler, A.B., Snyder, R.D. & Ord, J. K. (2001) Forecasting Models and prediction intervals for the multiplicative Holt–Winters method, *International Journal of Forecasting*, 17, pp. 269–286.

Makridakis, S. & Hibon, M. (2000) The M3-competition: results, conclusions and implications, *Journal of Forecasting*, 16, pp. 451–476.

Makridakis, S., Wheelwright, S.C. & Hyndman, R.J. (1998) *Forecasting. Methods and Applications*, 3rd edn (New York: Wiley).

Ord, J. K., Koehler, A.B. & Snyder, R.D. (1997) Estimation and prediction for a class of dynamic nonlinear statistical models, *Journal of the American Statistical Association*, 92, pp. 1621–1629.

R Development Core Team (2005) *R: A Language and Environment for Statistical Computing* (Vienna, Austria: R Foundation for Statistical Computing). (http://www.R-project.org.)

Seber, G.A.F. (1984) *Multivariate Observations* (New York: Wiley).

Segura, J.V. & Vercher, E. (2001) A spreadsheet modeling approach to the Holt-Winters optimal forecasting, *European Journal of the Operational Research*, 131, pp. 375–388.

Segura, J.V. & Vercher, E. (2002) *A Mathematical Programming Model for the Holt-Winters multiplicative method* (Technical Report of the Operations Research Center, Universidad Miguel Hernandez de Elche, Spain).

Yar, M. & Chatfield, C. (1990) Prediction Intervals for the Holt-Winters Forecasting Procedure, *International Journal of Forecasting*, 6, pp. 127–137.

Winters P.R. (1960) Forecasting sales by exponentially weighted moving averages, *Management Science*, 6, pp. 324–342.