Aust. N. Z. J. Stat. **47**(1), 2005, 87–99

LOCAL LINEAR FORECASTS USING CUBIC SMOOTHING SPLINES

ROB J. HYNDMAN[1]*, MAXWELL L. KING[1], IVET PITRUN[1] AND BAKI BILLAH[1]

Monash University

Summary

This paper shows how cubic smoothing splines fitted to univariate time series data can be used to obtain local linear forecasts. The approach is based on a stochastic state–space model which allows the use of likelihoods for estimating the smoothing parameter, and which enables easy construction of prediction intervals. The paper shows that the model is a special case of an ARIMA(0, 2, 2) model; it provides a simple upper bound for the smoothing parameter to ensure an invertible model; and it demonstrates that the spline model is not a special case of Holt's local linear trend method. The paper compares the spline forecasts with Holt's forecasts and those obtained from the full ARIMA(0, 2, 2) model, showing that the restricted parameter space does not impair forecast performance. The advantage of this approach over a full ARIMA(0, 2, 2) model is that it gives a smooth trend estimate as well as a linear forecast function.

Key words: ARIMA models; exponential smoothing; Holt's local linear forecasts; maximum likelihood estimation; non-parametric regression; smoothing splines; state–space model; stochastic trends.

1. Introduction

Suppose we observe a univariate time series $\{y_t\}$, $t = 1, \dots, n$, with nonlinear trend. We are interested in forecasting the series by extrapolating the trend using a linear function estimated from the observed time series.

Linear trend extrapolation is very widely used and performs relatively well in practice. For example, Makridakis & Hibon (2000), Assimakopoulos & Nikolopoulos (2000) and Hyndman & Billah (2003) discuss the excellent performance of linear trend methods in the M3-competition. In this paper, we discuss a method for local linear extrapolation of a stochastic trend based on cubic smoothing splines.

For equally spaced time series, a cubic smoothing spline can be defined as the function $\hat{f}(t)$ which minimizes

$$\sum_{t=1}^{n} \big(y_t - f(t)\big)^2 + \lambda \int_{\mathcal{S}} \big(f''(u)\big)^2 \, du \tag{1}$$

over all twice-differentiable functions f on $\mathcal{S}$ where $[1, n] \subseteq \mathcal{S} \subseteq \mathbb{R}$. The smoothing parameter λ is controlling the 'rate of exchange' between the residual error described by the sum of squared residuals and local variation represented by the square integral of the second derivative

Received August 2002; revised January 2004; accepted April 2004.
* Author to whom correspondence should be addressed.
[1] Dept of Econometrics and Business Statistics, Monash University, Clayton VIC 3800, Australia.
e-mail: Rob.Hyndman@buseco.monash.edu.au
Acknowledgments. The first author thanks the Australian National University for their hospitality while much of this work was being done. The authors also thank Andrew Harvey, two referees and two associate editors for helpful comments on earlier versions of this paper.

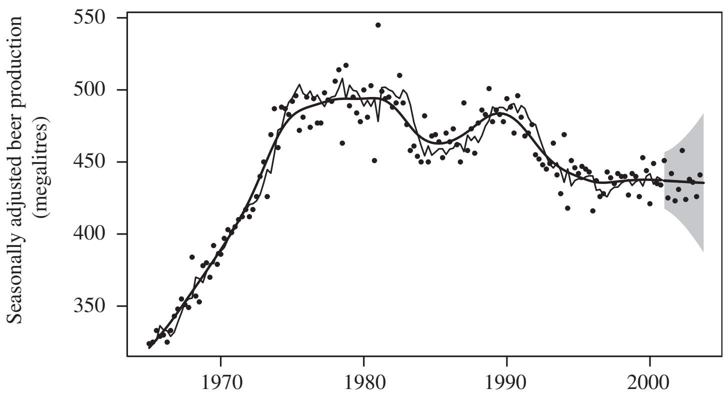

Figure 1. Cubic spline forecasts of Australian quarterly beer production (seasonally adjusted) for March 2001 – December 2003, with 80% prediction intervals. Actual values for March 1965 – September 2003 are also shown. The thick line through the historical data shows the fitted cubic spline $\hat{f}(t)$; the forecasts are obtained by a linear extrapolation of $\hat{f}(t)$; the prediction intervals are obtained from the state–space model described in Section 2. Here $\lambda = 210.5$. The thin line shows the 'trend' obtained from the one-step forecasts using an ARIMA(0, 2, 2) model.

of f. For a given λ, fast algorithms for computing $\hat{f}(t)$ are described by Green & Silverman (1994 pp. 19–21). Large values of λ give $\hat{f}(t)$ close to linear while small values of λ give a very wiggly function $\hat{f}(t)$. In practice, λ is not generally known.

The solution to (1) consists of piecewise cubic polynomials joined at the times of observation, $t = 1, 2, \ldots, n$. Furthermore, the solution has zero second derivative at $t = n$. Therefore, an extrapolation of $\hat{f}(t)$ for $t > n$ is linear. The linear extrapolation of $\hat{f}(t)$ provides our point forecasts.

We derive a new method for computing prediction intervals for these forecasts, making use of a stochastic model formulation due to Wahba (1978) and Wecker & Ansley (1983). We also provide a new method for estimating the smoothing parameter λ.

Figure 1 gives an example of our forecast procedure applied to seasonally adjusted Australian quarterly beer production (March 1965 – December 2000). The data were obtained from the Australian Bureau of Statistics (CAT8301.0), and seasonally adjusted using the X11-based seasonal adjustment method (Ladiray & Quenneville, 2001). The fitted spline curve is shown, along with the associated linear forecast function and 80% prediction intervals. Our method provides a smooth historical trend, a linear forecast function and prediction intervals.

Forecasts are usually made using models that give most weight to recent observations, and negligible weight to the distant past. This means that the smoothing parameter λ should not be too big for forecasting purposes. We make this explicit by finding the bounds on λ required for our model to be invertible. (Specifically, we find that $\lambda < 1.640519n^3$.)

Some linear forecast methods assume there is an underlying linear trend (e.g. a random walk with constant drift). We do not make this assumption. Our forecast function is linear, but the underlying trend $f(t)$ is allowed to be nonlinear. Further, the possible future changes in trend direction are explicitly accommodated in the stochastic model for trend.

For example, with the beer production data of Figure 1, if the forecasts were made in 1990 (just before the downturn in beer production) then the point forecasts would show an upward trend, whereas the actual values would trend downwards. However, the prediction intervals would be wide enough to accommodate the downturn that occurred.

An alternative approach to local linear forecasting is to allow a deterministic nonlinear trend. This is the approach followed by Nottingham & Cook (2001), for example. We prefer the stochastic trend approach as it allows the uncertainty in the trend to be explicitly included in the measures of forecast uncertainty. A hybrid approach, combining both deterministic and stochastic trends, is provided through SEMIFAR models (see Beran & Ocker, 1999; Beran & Feng, 2002).

Other local linear forecast models with stochastic trends include an $\text{ARIMA}(0,2,2)$ model, Harvey's (1989 p.45) local linear growth model and the AN model of Hyndman *et al.* (2002) which underlies Holt's (1957, 2004) linear trend method. In fact, these are all connected — Harvey's model is asymptotically equivalent to the AN model, and the AN model is a reparameterization of an $\text{ARIMA}(0,2,2)$ model.

Our model is also equivalent to an $\text{ARIMA}(0,2,2)$ model, but with a restricted parameter space. The restricted space seems to have a minor effect on forecast performance. The gain in using the spline approach, compared to the full $\text{ARIMA}(0,2,2)$ approach, is that it provides a smooth trend function that is useful for aiding interpretation of the historical data. Figure 1 also shows the 'trend' constructed from the one-step forecasts obtained using an $\text{ARIMA}(0,2,2)$ model fitted to the beer data with exact likelihood estimation (Gardner, Harvey & Phillips, 1980). The spline method provides a much superior estimate of the smooth trend through the historical data.

Our paper is structured as follows. Section 2 describes the stochastic model formulation for the cubic smoothing spline forecasts and Section 3 shows how to estimate the smoothing parameter. Simple expressions for obtaining point forecasts and prediction intervals are given in Section 4. In Section 5 we discuss the relationship between our model, an $\text{ARIMA}(0,2,2)$ model and a state–space model underlying Holt's linear trend forecasts. These relationships enable us to obtain the maximum bound for the smoothing parameter λ to ensure invertibility. Finally, in Section 6 we compare the forecasting performance of our model with other local linear forecasting models.

2. State–space model

The definition of cubic smoothing splines given in Section 1 provides suitable point forecasts, but does not allow estimation of forecast uncertainty. To do that, we use the stochastic process formulation proposed by Wahba (1978) and developed in subsequent work of Wecker & Ansley (1983). We present Wecker & Ansley's state–space model in the special case of cubic smoothing splines applied to equally spaced data.

First, we transform the observation time space to $[0,1]$ by defining the transformed observation times as $\{t_1,\dots,t_n\}$ where $t_i = i/n$. This transformation means that λ is rescaled also. Our transformed value of λ is $\lambda_* = n^{-3}\lambda$.

Then, for $i = 1,2,\dots,$ we define

$$g(t_i) = \tau \int_0^{t_i} (t_i - u)\, dW(u)\,,$$

where $\tau > 0$ and $W(u)$ is a standard Wiener process. Also

$$\boldsymbol{u}_i = \tau \begin{bmatrix} \int_{t_{i-1}}^{t_i} (t_i - u)\, dW(u) \\ W(t_i) - W(t_{i-1}) \end{bmatrix} \quad \text{and} \quad \boldsymbol{\alpha}_i = \begin{bmatrix} g(t_i) - g(t_1) \\ \tau(W(t_i) - W(t_1)) \end{bmatrix}.$$

Then we assume Y_i satisfies the state–space model

$$Y_i = \boldsymbol{s}_i^\top \boldsymbol{\beta} + [1\ 0]\ \boldsymbol{\alpha}_i + e_i\,, \tag{2}$$

$$\boldsymbol{\alpha}_i = \boldsymbol{T}_i \boldsymbol{\alpha}_{i-1} + \boldsymbol{u}_i \qquad (i = 1, \ldots, n)\,, \tag{3}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1)$ is normally distributed with zero mean and covariance matrix $c\boldsymbol{I}$,

$$\boldsymbol{T}_i = \begin{bmatrix} 1 & i/n \\ 0 & 1 \end{bmatrix},$$

e_i are iid $\mathrm{N}(0, \sigma^2)$ and $\boldsymbol{s}_i = (1, t_i)$. The starting condition is $\boldsymbol{\alpha}_0 = (0, 0)$. The state $\boldsymbol{\alpha}_{i-1}$ is assumed independent of $\boldsymbol{u}_i$.

Wahba (1978) showed that

$$\lim_{c\to\infty} \mathrm{E}\big(\boldsymbol{s}_i^\top \boldsymbol{\beta} + [1\ 0]\ \boldsymbol{\alpha}_i \mid Y_1, \ldots, Y_n\big)$$

is the cubic smoothing spline $\hat{f}(t)$ with $\lambda_* = \sigma^2/\tau^2$. Thus $\hat{f}(t)$ is the mean of Y_t, and we can obtain point forecasts which extrapolate $\hat{f}(t)$ by applying the Kalman recursions to the state–space model (2) and (3). Furthermore, we can also obtain forecast variances in this way.

However, a more direct approach is possible using a matrix formulation of the model. Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$, $\boldsymbol{e} = (e_1, \ldots, e_n)$ and $\boldsymbol{g} = (g(t_1), \ldots, g(t_n))$. Then

$$\boldsymbol{Y} = \boldsymbol{S}\boldsymbol{\beta} + \boldsymbol{g} + \boldsymbol{e}\,, \tag{4}$$

where the ith row of $\boldsymbol{S}$ is $\boldsymbol{s}_i^\top$.

Proposition 1. *Let $\boldsymbol{Y}$ be given by (4). Then $\boldsymbol{Y}$ is normally distributed with mean $\boldsymbol{0}$ and covariance matrix*

$$\boldsymbol{\Omega} = \sigma^2(c\boldsymbol{S}\boldsymbol{S}^\top + \lambda_*^{-1}\boldsymbol{\Sigma} + \boldsymbol{I}_n)\,,$$

where c has been rescaled and where $\boldsymbol{\Sigma}$ is symmetric with the (j,k)th element on or above the diagonal given by $\Sigma_{jk} = \frac{1}{6}(\sigma^2 n^{-3} j^2(3k - j))$ $(k \geq j)$. That is

$$\boldsymbol{\Sigma} = \tfrac{1}{6}\sigma^2 n^{-3} \begin{bmatrix} 2 & 5 & \cdots & 3n-1 \\ 5 & 16 & \cdots & 4(3n-2) \\ \vdots & \vdots & \ddots & \vdots \\ 3n-1 & 4(3n-2) & \cdots & 2n^3 \end{bmatrix}.$$

We provide a proof for this result in the Appendix.

We use the stochastic formulation given by (4) and Proposition 1 to obtain point forecasts and prediction intervals.

3. Estimation

Estimates of the smoothing parameter λ_* can be obtained by maximizing the likelihood function of the model which is given by

$$\ell(\lambda_* \mid \boldsymbol{Y}) = |\boldsymbol{\Omega}|^{-1/2}(\boldsymbol{Y}^\top \boldsymbol{\Omega}^{-1}\boldsymbol{Y})^{-n/2}\,. \tag{5}$$

Let $\boldsymbol{P}$ denote the upper-triangular matrix from the Choleski decomposition of $\sigma^2\boldsymbol{\Omega}^{-1}$. ($\boldsymbol{P}$ depends only on λ_*.) Then, we can write

$$|\boldsymbol{\Omega}|^{-1/2} = \sigma^{-1}|\boldsymbol{P}| \tag{6}$$

and
$$(\boldsymbol{Y}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{Y})^{-n/2} = \sigma^n(\boldsymbol{Y}^{*\top}\boldsymbol{Y}^*)^{-n/2} = \sigma^n\left(\sum_{i=1}^n w_i^2\right)^{-n/2}, \tag{7}$$

where $\boldsymbol{Y}^* = \boldsymbol{PY}$ and w_i is the ith element of $\boldsymbol{Y}^*$. Using (5)–(7), the log-likelihood is given by

$$\log\ell(\lambda_* \mid \boldsymbol{Y}) = (n-1)\log\sigma + \log|\boldsymbol{P}| - \tfrac{1}{2}n\log\left(\sum_{i=1}^n w_i^2\right).$$

Thus we can estimate λ_* by maximizing

$$\log|\boldsymbol{P}| - \tfrac{1}{2}n\log\left(\sum_{i=1}^n w_i^2\right).$$

This is a new method for selecting a bandwidth for smoothing splines, although it is similar in spirit to the likelihood-based method of Wecker & Ansley (1983). (Our method is much faster as we do not need to iteratively apply GLS estimation or the Kalman filter.)

4. Prediction

We now wish to use the fitted model to predict the next n_0 observations. Because the model is formulated in state–space form, we could simply use the iterative Kalman filter equations. However, in this special case, we can derive a direct solution which is faster.

We express the next n_0 observations as the n_0-vector $\boldsymbol{Y}_0 = \boldsymbol{S}_0\boldsymbol{\beta} + \boldsymbol{g}_0 + \boldsymbol{e}_0$ where $\boldsymbol{Y}_0 = (Y_{n+1}, \dots, Y_{n+n_0})$ and $\boldsymbol{g}_0$, $\boldsymbol{e}_0$ are defined analogously, and where $\boldsymbol{S}_0$ has ith row $[1, t_{n+i}]$, $i = 1, \dots, n_0$. We also define $\boldsymbol{\Sigma}_0$ as the symmetric $n_0 \times n_0$ matrix with (j,k)th element $\frac{1}{6}\sigma^2 n^{-3}(n+j)^2(2n+3k-j)$ for $k \geq j$. It is assumed that $\boldsymbol{Y}_0$ has the same properties as the observed vector $\boldsymbol{Y}$. Then $\boldsymbol{\Omega}_0 = \sigma^2(c\boldsymbol{S}_0\boldsymbol{S}_0^\top + \lambda_*^{-1}\boldsymbol{\Sigma}_0 + \boldsymbol{I}_{n_0})$ is the variance–covariance of $\boldsymbol{Y}_0$.

To derive the best linear unbiased predictor for $\boldsymbol{Y}_0$ and the variance–covariance matrix of the associated prediction error, we first combine past and future values of $\{Y_t\}$ to obtain $\boldsymbol{Z} = (\boldsymbol{Y}, \boldsymbol{Y}_0)$ with covariance matrix

$$\mathrm{E}(\boldsymbol{ZZ}^\top) = \begin{bmatrix} \boldsymbol{\Omega} & \boldsymbol{U} \\ \boldsymbol{U}^\top & \boldsymbol{\Omega}_0 \end{bmatrix} = \sigma^2(c\boldsymbol{S}_1\boldsymbol{S}_1^\top + \lambda_*^{-1}\boldsymbol{\Sigma}_1 + I_{n+n_0}),$$

where $\boldsymbol{S}_1$ and $\boldsymbol{\Sigma}_1$ are constructed analogously to $\boldsymbol{S}$, $\boldsymbol{S}_0$, $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}_0$. Then, using standard results for conditional expectations of multivariate normal random variables (e.g. Rao, 1973 Section 8a), we obtain

$$\mathrm{E}(\boldsymbol{Y}_0 \mid \boldsymbol{Y}) = \boldsymbol{U}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{Y} \tag{8}$$

and
$$\mathrm{var}(\boldsymbol{Y}_0 \mid \boldsymbol{Y}) = \boldsymbol{\Omega}_0 - \boldsymbol{U}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{U}. \tag{9}$$

Equations (8) and (9) allow point forecasts and associated prediction intervals to be computed. In particular, the h-step ahead point forecast $\hat{Y}_{n+h}$ is the hth element of $\boldsymbol{U}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{Y}$, and its variance v_h is the hth diagonal element of the matrix $\boldsymbol{\Omega}_0 - \boldsymbol{U}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{U}$. Since Y_t is assumed normal, prediction intervals can be constructed from these first two moments in the usual way. A 95% prediction interval is given by $\hat{Y}_{n+h} \pm 1.96\sqrt{v_h}$.

These results assume that c, λ_* and σ^2 are known. In reality, c is any sufficiently large number (in the empirical calculations described in this paper we use $c = 100$), and the parameter λ_* can be estimated using the procedure described in Section 3.

To estimate σ^2, we first calculate one-step forecasts $\hat{Y}_t$ and associated 'variances' v_t from (8) and (9), with $\sigma^2 = 1$. This has no effect on the forecast means, but the forecast variances are incorrect by a factor of σ^2. So σ^2 can be estimated as

$$\hat{\sigma}^2 = \frac{1}{v_t}\sum_{t=1}^{n}(Y_t - \hat{Y}_t)^2 .$$

It is also possible to obtain the forecast means and variances using the Kalman filter algorithm. However, the advantage of the approach outlined here is that it gives explicit one-line expressions for the forecast means and variances ((8) and (9)) and so is more amenable to further analysis. For example, it is easy using our results to find the mean and variance of the sum of the next h observations (something that is useful in inventory planning), but this is difficult to obtain using a Kalman filter.

We provide an R function on the website http://www.statsoc.org.au/Publications/anzjsfiles/data.dwt that implements the spline forecast method described here.

5. Comparisons with other approaches

The spline model described above gives local linear forecasts based on a stochastic trend. We now explore connections between this model and other models which also have stochastic trends and produce local linear forecast functions.

In particular, we look at the range of values for λ which can lead to an invertible model. Invertibility is a desirable property of a forecasting model because we want to avoid models where the distant past has a non-negligible effect on the present.

5.1. ARIMA(0,2,2) models

It is known (see Wecker & Ansley, 1983) that the cubic spline state–space model described in Section 2 is equivalent to an ARIMA$(0, 2, 2)$ model with some restrictions on parameters. However, no-one seems to have explicitly worked out the connection, or the implications it has for forecasting with the cubic spline state–space model.

We define the ARIMA$(0, 2, 2)$ model as

$$Y_t - 2Y_{t-1} + Y_{t-2} = \varepsilon_t - \theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2} ,$$

where ε_t are independent $\mathrm{N}(0, \sigma_\varepsilon^2)$. For invertibility, we also require $|\theta_2| < 1$, $\theta_2 - \theta_1 < 1$ and $\theta_2 + \theta_1 < 1$ (Box, Jenkins & Reinsel, 1994 p.73). Then the ARIMA$(0, 2, 2)$ forecast

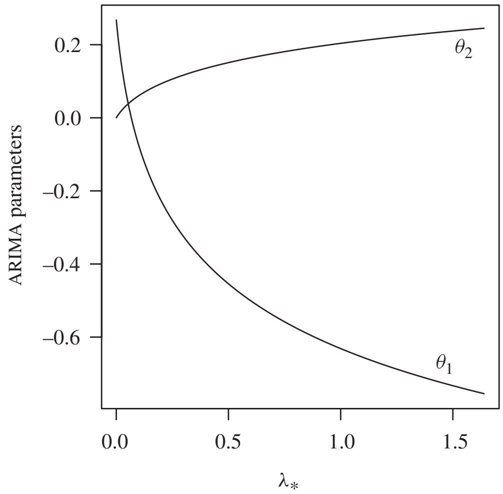

Figure 2. The relationship between the ARIMA parameters θ_1 and θ_2 and the cubic spline parameter λ_*

function is $\hat{Y}_{n+h} = \ell_n + b_n h$ where $\ell_n = Y_n - \theta_2 \hat{e}_n$ and $b_n = Y_n - Y_{n-1} + \theta_1 \hat{e}_n + \theta_2(\hat{e}_n + \hat{e}_{n-1})$. (Here, $\hat{e}_j$ denotes the jth residual.)

Brown & de Jong (2001) show that the cubic spline state–space model can be written as an ARIMA(0,2,1) process observed with error:

$$Y_t = X_t + \eta_t\,, \qquad (1-B)^2 X_t = \xi_t + \psi \xi_{t-1}\,,$$

where $\psi = 2-\sqrt{3}$, $(X_1, X_2 - X_1)$ is fully diffuse, and η_t and ξ_t are uncorrelated white noise series with means zero and variances σ^2 and τ^2 respectively. We can show this is equivalent to an ARIMA(0, 2, 2) model with θ_2 obtained by solving the following quartic equation:

$$\theta_2^4 - c_1\theta_2^3 + c_2\theta_2^2 - c_1\theta_2 + 1 = 0\,, \quad \text{where} \quad \theta_1 = \frac{\theta_2}{1+\theta_2}\Big(\frac{\psi}{\lambda_*} - 4\Big)\,, \quad \sigma_\varepsilon^2 = \frac{\sigma^2\lambda_*}{\theta_2}\,,$$
$$c_1 = 4 + \frac{1+\psi^2}{\lambda_*} \quad \text{and} \quad c_2 = 6 - \frac{2(1+4\psi+\psi^2)}{\lambda_*} + \frac{\psi^2}{\lambda_*^2}\,.$$

Numerical calculations show that the above quartic equation has at most one root which gives an invertible solution, and that an invertible solution is obtained if and only if $0 < \lambda_* < 1.640519$. Figure 2 shows the values of θ_1 and θ_2 as functions of λ_*. In the original time space (where observation times are $1, 2, \ldots, n$), the upper bound on λ is $1.640519n^3$. This upper bound on λ should be imposed whenever the spline model is used for forecasting purposes. If the model is simply used to describe the historical trend, invertibility is not relevant and so the bound need not be imposed.

The range of ARIMA(0, 2, 2) models that can be fitted in this way is greatly restricted. A wider range of models with linear forecast functions can be obtained by fitting a general ARIMA(0, 2, 2) model. In fact, Box *et al.* (1994 p.146) show that all ARIMA$(p, 2, q)$ have forecast functions which are asymptotically linear (the 'eventual forecast function'), and that the forecast function is exactly linear if and only if $p = 0$ and $q \leq 2$.

5.2. Holt's local linear forecasts

Holt's local trend method has been used in forecasting for many decades and it has proved remarkably versatile and useful. Point forecasts (e.g. Makridakis, Wheelwright & Hyndman, 1998 p. 158) are given by $\hat{Y}_{n+h} = \ell_n + b_n h$ where ℓ_n and b_n are computed recursively as follows:

$$\ell_t = \alpha Y_t + (1-\alpha)(\ell_{t-1} + b_{t-1})\,, \qquad b_t = \beta(\ell_t - \ell_{t-1}) + (1-\beta)b_{t-1}\,,$$

for $t = 2, \ldots, n$. Starting values for these recursions are often set to $\ell_1 = Y_1$ and $b_1 = Y_2 - Y_1$, although we choose the starting values optimally (see below).

The unobserved components ℓ_t and b_t represent the level and slope of the series at time t, and α and β are constants. We normally restrict the parameters such that $0 \le \alpha \le 1$ and $0 \le \beta \le 1$.

Recently, Hyndman *et al.* (2002) (hereafter HKSG) provided a general modelling framework for exponential smoothing methods, including Holt's method. This enables the forecasts to be obtained from a state–space model, thus providing facilities for maximum likelihood estimation, calculation of prediction intervals, and so on. HKSG actually provide two state–space models for Holt's method, which give identical point forecasts but have different properties for higher forecast moments. In this paper, we consider only the additive error version.

The model can be written as follows:

$$Y_t = \ell_{t-1} + b_{t-1} + \varepsilon_t\,, \qquad \ell_t = \ell_{t-1} + b_{t-1} + \alpha\varepsilon_t\,, \qquad b_t = b_{t-1} + \alpha\beta\varepsilon_t\,,$$

where ℓ_t denotes the level at time t, b_t denotes the slope of the trend at time t, and ε_t is a Gaussian white noise process with zero mean and variance σ^2. We estimate the parameters α and β and the initial state vector (ℓ_0, b_0) by maximizing the conditional likelihood as described in HKSG.

Hyndman *et al.* (2005) show that the forecast mean of this model is identical to Holt's local trend forecast, and the forecast variance of the model is

$$v_h = \sigma^2\big(1 + \alpha^2(h-1)\big(1 + \beta h + \tfrac{1}{6}\beta^2 h(2h-1)\big)\big)\,.$$

Using this expression, prediction intervals can be constructed in the usual way.

The above state–space model underlying Holt's method is equivalent to an ARIMA(0, 2, 2) model where $\alpha = \theta_2 + 1$ and $\beta = (1 - \theta_1 - \theta_2)/(1 + \theta_2)$. In theory, the parameter space for (α, β) could be taken as the whole invertible region for the ARIMA model (in which case we would have $0 < \alpha < 2$ and $0 < \beta < (4/\alpha) - 2$). However, it is usual to restrict the space further and require $0 < \alpha < 1$ and $0 < \beta < 1$ which leads to more interpretable models.

For the spline model, we found that $\theta_2 > 0$. Therefore $\alpha > 1$, which means that the spline model falls outside the usual range of parameters considered for Holt's method. (We also found that $\beta > 1$ when $\lambda_* > 0.14514$.) This means that Holt's method and the cubic spline state–space model are both special but non-overlapping cases of the ARIMA(0, 2, 2) model. The various parameter spaces are shown in Figure 3.

6. Empirical comparison of models

Given that the cubic spline state–space model is a special case of an ARIMA(0, 2, 2) model, it is interesting to see if the restricted parameter space results in poorer forecasting

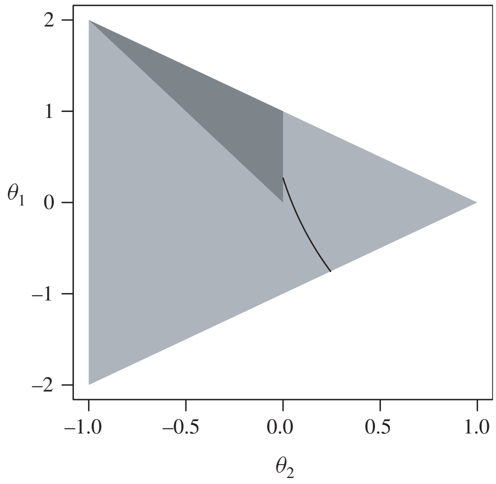

Figure 3. Parameter spaces for each model in terms of the ARIMA(0, 2, 2) parameters. The large triangular shaded region shows the parameter space for the ARIMA(0, 2, 2) model. The dark shaded region shows the usual restricted space for the Holt's model. The solid line shows the parameter space for the cubic spline state–space model.

performance. We also compare the forecasts from Holt's method (using the traditional Holt's parameter space).

We compare the three models by applying them to the 645 annual series which were part of the M3 forecasting competition (Makridakis & Hibon, 2000). For each series, six observations were withheld at the end of the series for comparisons. The remaining observations were used for estimating parameters.

For each series, we estimate the parameters using likelihood methods. We use the methods described in Sections 3 and 5.2 for the spline model and the state–space model underlying Holt's method, and for the full ARIMA model we use the exact likelihood method of Gardner *et al.* (1980) as implemented in the stats library distributed with R 1.9.0. We experimented with alternative optimization criteria, including optimizing for multi-step forecasting, but it made little difference to the results.

Each model is used to forecast the remaining six observations in the series. The forecasts are compared by computing the absolute percentage errors for each series and each forecast horizon. These are then averaged across series to obtain the mean absolute percentage error for each forecast horizon.

The results are given in Table 1; Figure 4 shows the quartiles of the absolute prediction errors for each model over the six forecast horizons. Together, these highlight some interesting similarities and differences between models. The restricted parameter space for Holt's method has not resulted in noticeable deterioration in forecast performance. The restricted parameter space for the cubic spline state–space method seems to have led to a minor deterioration in forecast performance for the largest forecast horizons.

We have investigated the effect of the length of series and the type of data (e.g. economic, demographic, finance) on these results and find the conclusions above are consistent across sample size and series classification.

TABLE 1

Mean absolute percentage error for each model, computed by averaging the absolute percentage error across all 645 annual series

	Forecast horizon					
Method	$h = 1$	$h = 2$	$h = 3$	$h = 4$	$h = 5$	$h = 6$
Spline	9.8	23.0	26.8	32.0	37.6	41.9
ARIMA(0, 2, 2)	8.6	21.6	26.9	30.3	35.9	37.8
Holt	8.6	20.8	25.0	29.1	33.6	36.2

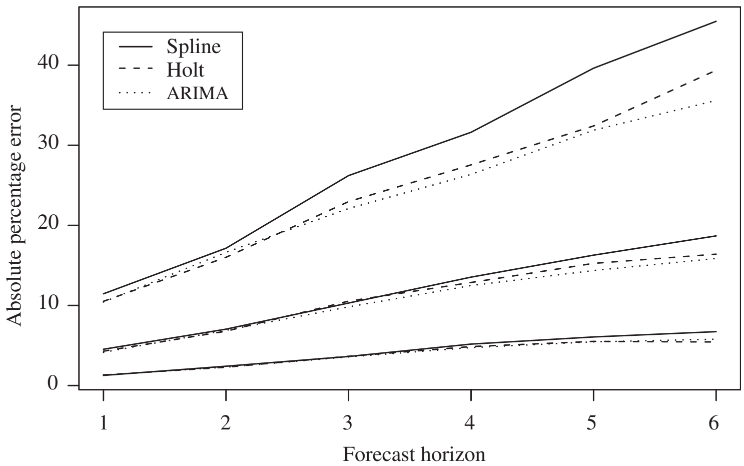

Figure 4. Quartiles of the absolute prediction errors for each model over the six forecast horizons, calculated across all 645 annual series

6.1. Conclusions

We have shown how cubic smoothing splines can be used to obtain local linear forecasts for a univariate time series. New results include a bound on the smoothing parameter to achieve invertibility; explicit and closed-form expressions for the point forecasts and prediction intervals; a new method for obtaining the smoothing parameter; and an empirical comparison with other local linear forecast methods.

It is possible to generalize our model to allow for seasonality and the effect of covariates. A non-stochastic covariate z_i can be included in the model by extending $\boldsymbol{\beta}$ to have an additional row (the coefficient of z_i), and extending $\boldsymbol{S}$ to include an additional column consisting of $z_1, \ldots, z_n$. Fixed seasonal factors can be included in an analogous way (with several such covariates corresponding to dummy variables for each season).

Stochastic seasonal factors could be included by adding a term z_i in (2) and letting $\sum_{i=1}^{m} z_i = \omega_i$ where m is the period of seasonality. (This is similar to the way seasonality is handled in the Basic Structural Model of Harvey (1989).) However, in this case, the Kalman filter would need to be used for calculation of the likelihood and for forecasting, because (4) and Proposition 1 no longer hold true.

Spline forecasts provide an alternative approach to ARIMA(0, 2, 2) models for local linear forecasting. The main advantage of the spline approach over the ARIMA approach is that it is directly associated with a smooth estimate of historical trend that is consistent with the forecasts. This can aid interpretation of the historical data as well as provide information about the trend used in forecasting. For example, the smooth trend through the beer production data in Figure 1 clearly shows the trend away from beer in Australia since about the late 1970s (partly explained by an increase in wine consumption). It also shows a brief resurgence in beer production in the late 1980s (when Australian beer exports led to increased production), before the production settled down to the current level.

A common criticism of non-parametric methods in general, and cubic splines in particular, is that they can be considered as special cases of more general time series models (e.g. Harvey & Koopman, 2000; Brown & de Jong, 2001). The (usually unstated) implication is that the more general model is better. We have shown that this restriction does not much reduce forecast performance, and so for forecasting purposes the criticism is not valid.

Appendix: Proof of proposition

This result follows directly from the state–space formulation except for the form of $\operatorname{var}(\boldsymbol{g})$ which we write as $\lambda_*^{-1}\boldsymbol{\Sigma}$.

Let $\boldsymbol{\Gamma}_i(j) = \mathrm{E}(\boldsymbol{\alpha}_i\boldsymbol{\alpha}_{i-j}^{\mathsf{T}})$, $j = 0, 1, \ldots$; $\boldsymbol{\alpha}_i$ follows a vector autoregressive model of order one in (3). Thus we obtain the Yule–Walker equations (Reinsel, 1997 Section 2.2.3),

$$
\begin{aligned}
&\boldsymbol{\Gamma}_0(0) = \mathbf{0} \\
&\boldsymbol{\Gamma}_i(0) = \boldsymbol{V}_i + \boldsymbol{T}_i\boldsymbol{\Gamma}_{i-1}(0)\boldsymbol{T}_i^{\mathsf{T}} \qquad (i = 1, \ldots, n) \\
&\boldsymbol{\Gamma}_i(j) = \boldsymbol{T}_i\boldsymbol{\Gamma}_{i-1}(j-1) \qquad (j = 1, 2, \ldots),
\end{aligned}
\tag{A1}
$$

where $\boldsymbol{V}_i$ is the covariance matrix of $\boldsymbol{u}_i$ and $\boldsymbol{\Gamma}_i(j) = \mathbf{0}$ if $j \geq i$. We can use these equations to iteratively calculate the values of $\boldsymbol{\Gamma}_i(j)$ for $i = 1, \ldots, n$ and $j = 1, 2, \ldots$. Then the (i, j)th element of $\lambda_*^{-1}\boldsymbol{\Sigma}$ is the top left element of $\boldsymbol{\Gamma}_j(j-i)$ if $i \leq j$ and the top left element of $\boldsymbol{\Gamma}_i(i-j)$ if $i \geq j$.

De Jong & Mazzi (2001) show that for any t_i where $0 < t_i < t_{i+1} < 1$ for $i = 1, 2, \ldots, n-1$, $\boldsymbol{V}_i$ has (j, k)th entry

$$
[\boldsymbol{V}_i]_{jk} = \tau^2 \int_{t_{i-1}}^{t_i} \frac{(t_i - u)^{2-j}(t_i - u)^{2-k}}{(2-j)!(2-k)!}\, du = \tau^2 \frac{h_i^{5-j-k}}{(4-j-k+1)(2-j)!(2-k)!},
$$

where $h_i = t_{i+1} - t_i$. Thus, in this special case where $h_i = h = n^{-1}$, we have

$$
\boldsymbol{V}_i = \tau^2 \begin{bmatrix} \frac{1}{3}h^3 & \frac{1}{2}h^2 \\ \frac{1}{2}h^2 & h \end{bmatrix}. \tag{A2}
$$

By substituting (A2) into (A1), we can construct $\boldsymbol{\Sigma}$.

First we show by induction that

$$
\boldsymbol{\Gamma}_i(0) = \tau^2 \begin{bmatrix} \frac{1}{3}i^3h^3 & \frac{1}{2}i^2h^2 \\ \frac{1}{2}i^2h^2 & ih \end{bmatrix}. \tag{A3}
$$

For $i = 0$, $\boldsymbol{\Gamma}_0(0) = \mathbf{0}$, so (A3) is true. Now assume (A3) is true for $i = k$. Then from (A1) we obtain

$$
\begin{aligned}
\boldsymbol{\Gamma}_{k+1}(0) &= \tau^2 \begin{bmatrix} \frac{1}{3}h^3 & \frac{1}{2}h^2 \\ \frac{1}{2}h^2 & h \end{bmatrix} + \tau^2 \begin{bmatrix} 1 & h \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{3}k^3h^3 & \frac{1}{2}k^2h^2 \\ \frac{1}{2}k^2h^2 & kh \end{bmatrix} \begin{bmatrix} 1 & 0 \\ h & 1 \end{bmatrix} \\
&= \tau^2 \begin{bmatrix} \frac{1}{3}(k+1)^3h^3 & \frac{1}{2}(k+1)^2h^2 \\ \frac{1}{2}(k+1)^2h^2 & (k+1)h \end{bmatrix}.
\end{aligned}
$$

So (A3) is true for $i = k + 1$ and for $i = 1, 2, 3, \ldots$ by induction. Now from (A1) we have

$$\boldsymbol{\Gamma}_i(j) = \boldsymbol{T}\boldsymbol{\Gamma}_{i-1}(j-1) = \boldsymbol{T}^2\boldsymbol{\Gamma}_{i-2}(j-2) = \boldsymbol{T}^j\boldsymbol{\Gamma}_{i-j}(0) \qquad (i \geq j)\,,$$

and so $\boldsymbol{\Gamma}_i(i-j) = \boldsymbol{T}^{i-j}\boldsymbol{\Gamma}_j(0)$. Thus

$$\boldsymbol{\Gamma}_i(i-j) = \tau^2 \begin{bmatrix} 1 & (i-j)h \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{3}j^3h^3 & \frac{1}{2}j^2h^2 \\ \frac{1}{2}j^2h^2 & jh \end{bmatrix} = \begin{bmatrix} \frac{1}{6}h^3j^2(3i-j) & \frac{1}{2}jh^2(2i-j) \\ \frac{1}{2}j^2h^2 & jh \end{bmatrix}.$$

Thus $\boldsymbol{\Sigma}$ is symmetric with the (j,k)th element on or above the diagonal given by $\Sigma_{jk} = \Sigma_{kj} = \frac{1}{6}\sigma^2h^3j^2(3k-j)$ $(k \geq j)$

and so $$\Sigma_{jk} = \tfrac{1}{6}\sigma^2n^{-3}k^2(3j-k) \qquad (j \geq k)\,.$$

References

ASSIMAKOPOULOS, V. & NIKOLOPOULOS, K. (2000). The theta model: a decomposition approach to forecasting. *Internat. J. Forecasting* **16**, 521–530.

BERAN, J. & FENG, Y. (2002). SEMIFAR models — a semiparametric approach to modelling trends, long-range dependence and nonstationarity. *Comput. Statist. Data Anal.* **40**, 393–419.

BERAN, J. & OCKER, D. (1999). SEMIFAR forecasts, with applications to foreign exchange rates. *J. Statist. Plann. Inference* **80**, 137–153.

BOX, G.E.P., JENKINS, G.M. & REINSEL, G.C. (1994). *Time Series Analysis: Forecasting and Control*, 3rd edn. Englewood Cliffs, NJ: Prentice-Hall.

BROWN, P.E. & DE JONG, P. (2001). Nonparametric smoothing using state space techniques. *Canad. J. Statist.* **29**, 37–50.

DE JONG, P. & MAZZI, S. (2001). Modeling and smoothing unequally spaced sequence data. *Stat. Inference Stoch. Process.* **4**, 53–71.

GARDNER, G., HARVEY, A.C. & PHILLIPS, G.D.A. (1980). Algorithm AS154. An algorithm for exact maximum likelihood estimation of autoregressive-moving average models by means of Kalman filtering. *Appl. Statist.* **29**, 311–322.

GREEN, P.J. & SILVERMAN, B.W. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach.* London: Chapman and Hall / CRC.

HARVEY, A.C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter.* Cambridge: Cambridge University Press.

HARVEY, A.C. & KOOPMAN, S.J. (2000). Signal extraction and the formulation of unobserved components models. *Econometrics J.* **3**, 84–107.

HOLT, C.C. (1957). Forecasting seasonal and trends by exponentially weighted moving averages. Office of Naval Research, Research Memorandum No. 52.

HOLT, C.C. (2004). Forecasting seasonals and trends by exponentially weighted moving averages. *Internat. J. Forecasting* **20**, 5–10. Reprint of Holt (1957).

HYNDMAN, R.J. & BILLAH, B. (2003). Unmasking the Theta method. *Internat. J. Forecasting* **19**, 287–290.

HYNDMAN, R.J., KOEHLER, A.B., ORD, J.K. & SNYDER, R.D. (2005). Prediction intervals for exponential smoothing state space models. *J. Forecasting* **24**, 17–37.

HYNDMAN, R.J., KOEHLER, A.B., SNYDER, R.D. & GROSE, S. (2002). A state space framework for automatic forecasting using exponential smoothing methods. *Internat. J. Forecasting* **18**, 439–454.

LADIRAY, D. & QUENNEVILLE, B. (2001). *Seasonal Adjustment With the X-11 Method.* Lecture Notes in Statistics, **158**. New York: Springer-Verlag.

MAKRIDAKIS, S.G. & HIBON, M. (2000). The M3-competition: results, conclusions and implications. *Internat. J. Forecasting* **16**, 451–476.

MAKRIDAKIS, S., WHEELWRIGHT, S.C. & HYNDMAN, R.J. (1998). *Forecasting: Methods and Applications*, 3rd edn. New York: John Wiley & Sons.

NOTTINGHAM, Q.J. & COOK, D.F. (2001). Local linear regression for estimating time series data. *Comput. Statist. Data Anal.* **37**, 209–217.

RAO, C.R. (1973). *Linear Statistical Inference and its Applications*, 2nd edn. New York: John Wiley & Sons.

REINSEL, G.C. (1997). *Elements of Multivariate Time Series Analysis*, 2nd edn. New York: Springer-Verlag.

WAHBA, G. (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regression. *J. Roy. Statist. Soc. Ser. B* **40**, 364–372.

WECKER, W.E. & ANSLEY, C.F. (1983). The signal extraction approach to nonlinear regression and spline smoothing. *J. Amer. Statist. Assoc.* **78**, 81–89.