THE SMOOTHNESS CRITERION AS A TREND DIAGNOSTIC

P. FOGARTY AND N. C. WEBER

The smoothness criterion is used in the design of symmetric moving average trend filters in time series and in graduation in actuarial studies. This measure of smoothness is used to motivate a diagnostic for determining the order of local polynomial trend.

1. Introduction

The standard approach to time series analysis is to decompose the series into three components, the trend, the seasonal component and the irregular noise. In this paper we are concerned with estimating the trend and so we will assume that all series have been seasonally adjusted. The structural form of the trend is often decided by physical considerations or by inspection of global time series plots. A global parametric model may be used to estimate the trend but a more common approach is to use a finite moving average filter to obtain non-parametric estimates of the local trend. Detailed accounts of recent developments in trend estimation can be found in Kenny and Durbin [5] and Cleveland et al. [1].

Finite symmetric moving average filters for identifying the underlying trend in non-seasonal time series are not new, dating from the seminal work of Henderson [3], in the actuarial literature, and Macaulay [7]. Determining the order of the local polynomial trend is often problematic, particularly if there is subtle local curvature in the presence of a dominant linear trend. The most common diagnostic tool is a sequence of time series plots for the original series and the residual series calculated after each attempt at estimation. In this paper a simple, alternative graphical diagnostic is provided to help select or confirm the order of any local polynomial term in the trend. These plots are motivated by the smoothness criterion used in designing optimal filter weights and appear to be sensitive to local curvature.

Consider the simple structural model for a non-seasonal or seasonally adjusted time series y_t with underlying trend process T_t,

$$y_t = T_t + \epsilon_t,\tag{1.1}$$

where ϵ_t is a zero mean noise process. Following Gray and Thomson [2] the trend may be described locally by a polynomial of order p plus some stochastic process,

$$T_t = \sum_{j=0}^{p} \beta_j t^j + \xi_t,\tag{1.2}$$

where the zero mean stochastic process ξ_t is assumed to be correlated, but uncorrelated with the noise process ϵ_t. Moreover, using Δ to denote the backwards difference operator, $\Delta y_t = y_t - y_{t-1}$, it is assumed that the differenced process $(\Delta^{p+1}\xi_t)$ is second order stationary.

The trend at some central time point t is estimated by a moving average of $n = 2r + 1$ consecutive observations. Let the estimator be

$$\widehat{T}_t = \sum_{s=-r}^{r} w_s y_{t+s}.\tag{1.3}$$

Denote the n vector of weights by $\mathbf{w} = (w_{-r},\dots,w_r)^T$. For $\widehat{T}_t$ to be an unbiased predictor of T_t, that is, $E(\widehat{T}_t - T_t) = 0$, we need $\sum_{s=-r}^{r} w_s = 1$ and $\sum_{s=-r}^{r} s^j w_s = 0$, $0 < j \le p$, so that the moving average filter passes polynomials of degree p. We write these constraints as $C^T\mathbf{w} = \mathbf{c}$, where $\mathbf{c} = (1,0,\dots,0)^T$ is a vector of length $(p+1)$.

Gray and Thomson [2] focus on the design of symmetric, moving average trend filters and derive expressions for the optimal weights using the fidelity and smoothness criteria. Fidelity is more commonly known as mean squared error,

$$F = E(\widehat{T}_t - T_t)^2,\tag{1.4}$$

and smoothness is measured by

$$S = E(\Delta^{p+1}\widehat{T}_t)^2.\tag{1.5}$$

These criteria have a long history, particularly in the actuarial literature (see, e.g., London [6]) going back to Henderson [4] and Whittaker [11]. Minimising F ensures the trend estimate is in some sense close to the "true" value, whereas minimising S ensures that the fitted trend polynomial is close to a smooth polynomial of degree p. For actuarial data where the underlying trend might reflect mortality rates, say, it is natural to insist on smoothness in the estimates.

Following Whittaker's approach Gray and Thomson [2] derive expressions for the weights obtained by minimising the compromise criterion

$$Q = \theta F + (1 - \theta)S, \quad \theta \in [0, 1].\tag{1.6}$$

For the classical case where ϵ_t is white noise Gray and Thomson [2] derive the optimal symmetric weights

$$\mathbf{w} = E_\theta^{-1}C(C^T E_\theta^{-1}C)^{-1}\mathbf{c}, \tag{1.7}$$

where

$$E_\theta = \theta(\sigma^2 I + \Omega) + (1-\theta)(B_{p+1} + \Gamma_{p+1}), \tag{1.8}$$

in which Ω is the covariance matrix with elements

$$\Omega_{jk} = \mathrm{cov}\,(\xi_{t+j} - \xi_t, \xi_{t+k} - \xi_t), \tag{1.9}$$

Γ_{p+1} is the covariance matrix of a sequence of n observations from $\Delta^{p+1}\xi_t$, and B_{p+1} is the covariance matrix of a sequence of n observations from the stationary moving average process $\Delta^{p+1}\epsilon_t$. Note that the classical Macaulay filters can be obtained by setting $\theta = 1$ and the Henderson filters follow by setting $\theta = 0$.

McLaren and Steel [9, 10] extended Gray and Thomson [2] to produce trend estimates using data from repeated surveys. The correlation structure induced by the survey rotation pattern is incorporated in deriving $\mathbf{w}$.

In all cases the optimal weights depend on the choice of n, p and the covariance structures of the processes ϵ_t and ξ_t.

The window width n is specified by the user according to the nature and volatility of the series. The covariance structures of the processes ϵ_t and ξ_t are used to construct the E_θ matrix. Because of the form of the optimal weights in (1.7) the covariance matrices need only be known up to a constant of proportionality. The process ϵ_t is generally taken to be white noise, although under conditions of rotation sampling there may be some autocorrelation structure which can be modelled according to the rotation pattern used. The process ξ_t is included if there is reason to assume the trend T_t will include some deviation from the local polynomial, and so should be determined by the nature of the series being examined.

The choice of p is an important factor in determining the optimal symmetric filter weights. In many instances the nature of the trend is clear but there are situations where the noise can mask some local curvature. In regression and other areas of time series analysis it is standard practice to use diagnostic checks to determine whether or not model assumptions are reasonable. In this paper we will investigate the local smoothness estimates

$$\hat{S}_t(p) = (\Delta^{p+1}\hat{T}_t)^2 \tag{1.10}$$

and propose using plots of $\hat{S}_t$ against t for various values of p as a diagnostic tool to identify p and/or support the choice of a particular p.

2. Smoothness plots

First we consider what happens if we misspecify p. Assume the series y_t is given locally by a polynomial of order q, so that

$$y_t = \sum_{j=0}^{q} \beta_j t^j + \xi_t + \epsilon_t, \tag{2.1}$$

but we believe the local polynomial is of order p. Let $\mathbf{w}_p$ denote the filter weights obtained from (1.7), where the local polynomial is assumed to be of order p. Then

$$\begin{aligned}
\hat{S}_t(p) &= \left(\Delta^{p+1} \sum_{s=-r}^{r} w_s y_{t+s} \right)^2 \\
&= \left(\Delta^{p+1} \sum_{s=-r}^{r} w_s \left(\sum_{j=0}^{q} \beta_j (t+s)^j + \xi_{t+s} + \epsilon_{t+s} \right) \right)^2 \\
&= \left(\sum_{s=-r}^{r} w_s \left(\Delta^{p+1} \sum_{j=0}^{q} \beta_j (t+s)^j + \Delta^{p+1} \xi_{t+s} + \Delta^{p+1} \epsilon_{t+s} \right) \right)^2 .
\end{aligned} \tag{2.2}$$

Now if $p \geq q$, that is, we use the correct order or over-fit the local polynomial, then

$$\hat{S}_t(p) = \left(\sum_{s=-r}^{r} w_s \left(0 + \Delta^{p+1} \xi_{t+s} + \Delta^{p+1} \epsilon_{t+s} \right) \right)^2 \tag{2.3}$$

which has expected value

$$E(\hat{S}_t(p)) = \mathbf{w}_p^T \Gamma_{p+1} \mathbf{w}_p + \mathbf{w}_p^T B_{p+1} \mathbf{w}_p. \tag{2.4}$$

If $p = q - 1$, that is we under-fit by one degree, then

$$\hat{S}_t(q-1) = \left(\sum_{s=-r}^{r} w_s \left(q! \beta_q + \Delta^q \xi_{t+s} + \Delta^q \epsilon_{t+s} \right) \right)^2 \tag{2.5}$$

which has expected value over the local window of length $(2r+1)$

$$E(\hat{S}_t(q-1)) = (q! \beta_q)^2 + \mathbf{w}_{q-1}^T \Gamma_q \mathbf{w}_{q-1} + \mathbf{w}_{q-1}^T B_q \mathbf{w}_{q-1}. \tag{2.6}$$

If $p < q - 1$, then

$$\hat{S}_t(p) = \left(\sum_{s=-r}^{r} w_s \left(\Delta^{p+1} \sum_{j=0}^{q} \beta_j (t+s)^j + \Delta^{p+1} \xi_{t+s} + \Delta^{p+1} \epsilon_{t+s} \right) \right)^2 \tag{2.7}$$

which has expected value

$$E(\widehat{S}_t(p)) = \left(\sum_{s=-r}^{r} w_s \left(\Delta^{p+1} \sum_{j=0}^{q} \beta_j (t+s)^j \right) \right)^2 + \mathbf{w}_p^T \Gamma_{p+1} \mathbf{w}_p + \mathbf{w}_p^T B_{p+1} \mathbf{w}_p. \tag{2.8}$$

Note that in this case $E(\widehat{S}_t(p))$ depends on t. From this we can see that when we over-fit or have a correct fit

$$E(\widehat{S}_t(p)) = \mathbf{w}_p^T \Gamma_{p+1} \mathbf{w}_p + \mathbf{w}_p^T B_{p+1} \mathbf{w}_p, \tag{2.9}$$

which will vary with p, depending on the structure of the processes ϵ_t and ξ_t, but will not necessarily suggest which of the $p \geq q$ is the correct choice. When we under-fit, $E(\widehat{S}_t(p))$ is increased by $(\sum_{s=-r}^{r} w_s (\Delta^{p+1} \sum_{j=0}^{q} \beta_j (t+s)^j))^2$ which does not involve ϵ_t and ξ_t but does depend on t if $p < q - 1$. This suggests the following diagnostic check for p. Obtain the filtered series for various choices of p and calculate

$$\widehat{S}_t(p) = (\Delta^{p+1} \widehat{T}_t)^2, \tag{2.10}$$

for each internal point t, that is, for $t = r + p + 1, \ldots, N - r$, where N is the length of the time series. Plot $\widehat{S}_t(p)$ against t and choose the value of p at which $\widehat{S}_t(p)$ is approximately time invariant, and does not differ drastically from $E(\widehat{S}_t(p)) = \mathbf{w}_p^T \Gamma_{p+1} \mathbf{w}_p + \mathbf{w}_p^T B_{p+1} \mathbf{w}_p$.

Let $\mathbf{d} = (\Delta^{p+1} \epsilon_{t+r}, \ldots, \Delta^{p+1} \epsilon_{t-r})^T$ and let $\mathbf{e}$ denote the $(n + p + 1)$ vector given by $\mathbf{e} = (\epsilon_{t+r}, \ldots, \epsilon_{t-r-p-1})^T$. As $\Delta^{p+1} y_t = \sum_{j=0}^{p+1} \binom{p+1}{j} (-1)^j y_{t-j}$, we can write

$$\mathbf{d} = D_{p+1} \mathbf{e}, \tag{2.11}$$

where D_{p+1} is the $n \times (n + p + 1)$ matrix with rows made up of zeroes and the coefficients of the operator Δ^{p+1},

$$D_{p+1} = \begin{pmatrix} 1 & -\binom{p+1}{1} & \cdots & (-1)^{p+1}\binom{p+1}{p+1} & & 0 \\ \mathbf{0} & \ddots & \ddots & \cdots & \ddots & \mathbf{0} \\ 0 & \cdots & 1 & -\binom{p+1}{1} & \cdots & (-1)^{p+1}\binom{p+1}{p+1} \end{pmatrix}. \tag{2.12}$$

If the errors are assumed to be normal and second order stationary, the following theorem applies when p is greater than or equal to the correct order of the polynomial term.

THEOREM 2.1. *Consider a series given by the local dynamic model (2.1), with $\xi_t = 0$ and optimal symmetric central filter weights $\mathbf{w}_p$ given by (1.7) for $p \geq q$. If the noise process $\mathbf{e}$ is normally distributed and second order stationary with known covariance structure then*

$$\widehat{S}_t(p) \sim \sigma^2 \left(\mathbf{w}_p^T D_{p+1} V_\epsilon D_{p+1}^T \mathbf{w}_p \right) \chi_1^2, \tag{2.13}$$

where $\sigma^2 V_\epsilon$ is the covariance matrix of a vector of error terms of length $(n + p + 1)$.

Proof. Under the assumption that $\xi_t = 0$, and for $p \geq q$, we have

$$\widehat{S}_t(p) = \left(\sum_{s=-r}^{r} w_s \Delta^{p+1} \epsilon_{t+s} \right)^2. \tag{2.14}$$

Note that we can write

$$\sum_{s=-r}^{r} w_s \Delta^{p+1} \epsilon_{t+s} = \mathbf{w}_p^T \mathbf{d}, \tag{2.15}$$

where $\mathbf{w}_p = (w_r, \ldots, w_{-r})^T$. Hence $\sum_{s=-r}^{r} w_s \Delta^{p+1} \epsilon_{t+s} = \mathbf{w}_p^T D_{p+1} \mathbf{e}$, and

$$\widehat{S}_t(p) = \left(\mathbf{w}_p^T D_{p+1} \mathbf{e} \right)^2 = \mathbf{w}_p^T D_{p+1} \mathbf{e} \mathbf{e}^T D_{p+1}^T \mathbf{w}_p = \mathbf{a}^T M \mathbf{a}, \tag{2.16}$$

where $\mathbf{a} = D_{p+1}^T \mathbf{w}_p$ is an $(n+p+1)$ vector of known constants, and $M = \mathbf{e}\mathbf{e}^T$.

Now use the assumption that the error process ϵ_t is normally distributed and second order stationary, then the vector $\mathbf{e}$ is multivariate normal with mean $\mathbf{0}$ and covariance matrix $\sigma^2 V_\epsilon$, where $\sigma^2 V_\epsilon$ is known. Under this assumption M has a Wishart distribution,

$$M = \mathbf{e}\mathbf{e}^T \sim W_{n+p+1}(\sigma^2 V_\epsilon, 1). \tag{2.17}$$

Using results for the Wishart distribution

$$\frac{\mathbf{a}^T M \mathbf{a}}{\mathbf{a}^T \sigma^2 V_\epsilon \mathbf{a}} \sim \chi_1^2. \tag{2.18}$$

See, for example, Mardia et al.[8, Theorem 3.4.2, page 67]. Now from (2.16) and (2.18) it is evident that

$$\widehat{S}_t(p) \sim \sigma^2 \left(\mathbf{w}_p^T D_{p+1} V_\epsilon D_{p+1}^T \mathbf{w}_p \right) \chi_1^2, \tag{2.19}$$

as required. □

Thus under the assumptions that $\xi_t = 0$ and $\mathbf{e}$ is multivariate normal with zero mean and known covariance structure we can find the distribution of $\widehat{S}_t(p)$ for $p \geq q$. Note that if the errors are independent then $V_\epsilon = I$. The above result can easily be extended to include the case where ξ_t is normally distributed with known covariance structure by treating $(\xi_t + \epsilon_t)$ as the noise term.

To apply this in practice, calculate $\widehat{S}_t(p)$ for every point in the body of the series so as to avoid issues with end effects. Then using the above result, compare the values of $\widehat{S}_t(p)$ to some chosen reference level such as

$$r = \sigma^2 \left(\mathbf{w}_p^T D_{p+1} V_\epsilon D_{p+1}^T \mathbf{w}_p \right) \chi_1^2(0.95), \tag{2.20}$$

based on the 95th percentile of the χ_1^2 distribution. Recall that $\hat{S}_t(p)$ will tend to be larger than expected when $p < q$, so if there are a large number of values greater than the reference level, this would suggest q is greater than the current value of p.

This can be done graphically by plotting all the $\hat{S}_t(p)$ values together with a horizontal line corresponding to the relevant reference value. When examining these diagrams it is important to recall that we expect some points to lie above the line even when p is correct, and that neighbouring points are not independent. Adjacent values of $\hat{S}_t(p)$ will have some y_t values in common so one aberrant point may affect the behaviour of a block of consecutive values of the smoothness measure. This dependence may result in the points appearing to clump together.

3. Examples

The method is best illustrated via the following simple examples.

Example 3.1. A linear series with normal noise is generated, and a second series is generated by adding a cosine term to the first so that there is one turning point within any filter window. The first series is given by

$$y_{1,t} = a + bt + \epsilon_t, \tag{3.1}$$

and the second by

$$y_{2,t} = a + bt + d\cos\left(\frac{2\pi t}{k}\right) + \epsilon_t, \tag{3.2}$$

where $\epsilon_t \sim N(0,1)$ and a and b are common. The second series approximates a locally quadratic series since $\cos(x)$ is locally approximated by $\cos(x) \approx 1 - x^2$. The time series plots for two series of length $N = 100$ are given in Figure 3.1.

By visual inspection alone one could easily be fooled into thinking that the second series can be modelled by a local linear trend. We then filter each series with weights given by (1.7) for $r = 6$, $p = 0,1,2,3,4$ and with $E_\theta = B_{p+1}$. The weights with $\theta = 0$ were selected because they are the ones that minimise the smoothness criterion for the filtered series, given that the local trend is a polynomial of degree p. Thus they will be the most sensitive to the case where the trend cannot be described by a smooth order p polynomial. For both filtered series we calculate $\hat{S}_t(p) = (\Delta^{p+1}\hat{y}_t)^2$ for $t = p + 7,\ldots,94$, giving a series of smoothness values. Next, calculate the reference level r for each p where

$$r = \sigma^2 \left(\mathbf{w}_p^T D_{p+1} D_{p+1}^T \mathbf{w}_p\right) \chi_1^2(0.95), \tag{3.3}$$

and plot the series of smoothness values with a horizontal line at r.

The diagrams in Figure 3.2 demonstrate typical patterns for these plots. Note that the order of the smoothness values changes markedly with p, as does the cut off value r. The series $y_{1,t}$ can clearly be identified as linear, as it has few points above the line for the plots corresponding to $p = 1,2,3,4$. The series $y_{2,t}$ has many points above the line in the plot for $p = 1$, suggesting the series is given by a local polynomial of degree greater than 1. Even in this case where the curvature is not strong the diagnostic plots send a clear
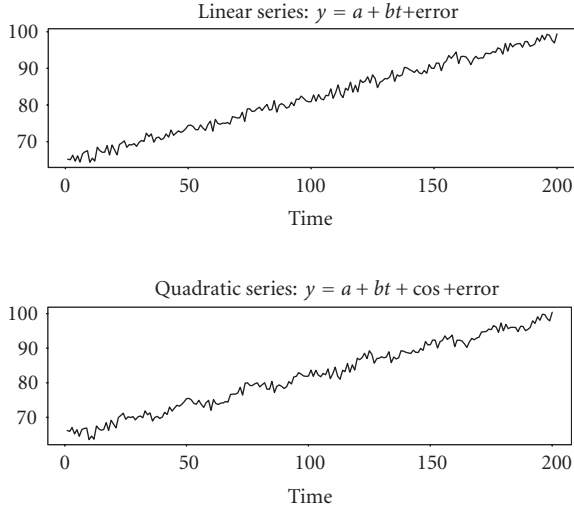
Figure 3.1. Time series plots.

message that there is some quadratic effect in the data. Note that the plots for $p = 4$ for both series are very similar as almost all curvature has been differenced out of both series.

In practice the residual variance σ^2 is typically unknown. If we have a very long time series, and we can assume constant variance, it would be possible to estimate σ^2 using early values of the time series and then use this estimate to examine smoothness values at an independent segment of the series. If we can find an estimate of the variance $\hat{\sigma}^2$ that is independent of $\hat{S}_t(p)$ and such that $\hat{\sigma}^2 \sim \sigma^2 \chi_l^2 / l$, for some degrees of freedom l, then

$$\frac{\hat{S}_t(p)}{\hat{\sigma}^2} \sim \left(\mathbf{w}_p^T D_{p+1} V_\epsilon D_{p+1}^T \mathbf{w}_p \right) F_{1,l}. \tag{3.4}$$

This result can be used in the same manner as above to provide a diagnostic check for the order of the local polynomial when σ^2 is unknown.

If the form of V_ϵ is known, as is the case for data from rotation group sampling (see, e.g., McLaren and Steel [9]), then estimates for σ^2 and V_ϵ can be obtained from early values of the time series. If the series is sufficiently long then there should be a clear time sequence gap between the values used to estimate the covariance structure and those used to calculate $\hat{S}_t(p)$ so that the dependence is minimised. If this is possible then the estimates for σ^2 and V_ϵ can be used to calculate the weights, and for diagnostic purposes they can be used to approximate the value for r via (2.20).

The following example demonstrates the use of this technique to determine the local order of the time series for the number of adult males employed in Australia (Australian Labour Force Survey, cat no 6202).
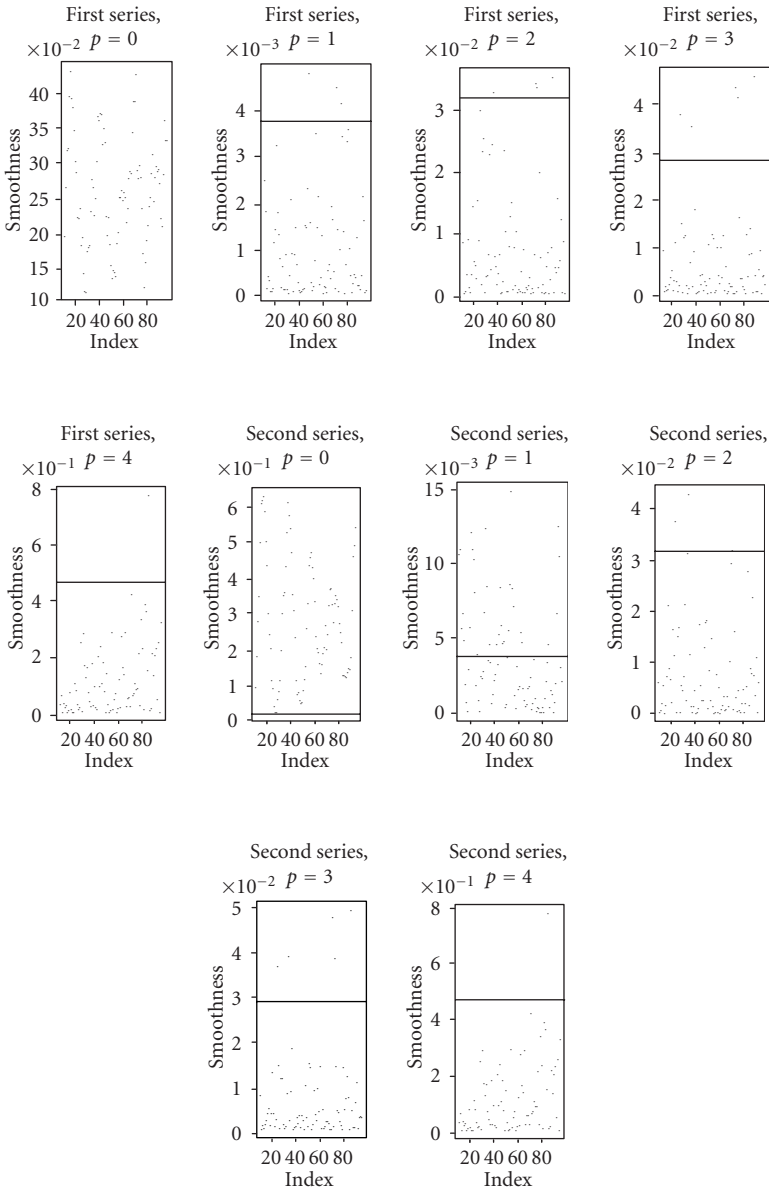
Figure 3.2. Empirical smoothness plots for Example 3.1.

Example 3.2 (Australian Labour Force Survey data). Since the filters are designed to iden-
tify the trend in the presence of noise, the ABS seasonally adjusted figures are used, since
these have had seasonal variations removed. In this case we are examining the most recent
ten years of the seasonally adjusted time series, shown in **Figure 3.3**.
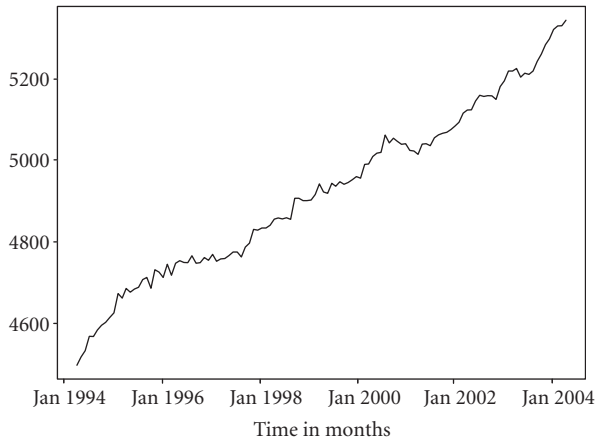
Figure 3.3. Seasonally adjusted Australian adult male employment data.

Smoothness estimates for this series are obtained assuming a locally constant, linear, quadratic, cubic or quartic model, according to (1.10), that is, allowing p to take the values 0, 1, 2, 3 or 4.

Estimates of the variance of the noise process are obtained from the data in the period 1984–1992, so that the data do not overlap with the span being examined. The variance is estimated by calculating the sample variance of the residuals that are obtained when the filter being tested has been applied to these earlier data. This variance estimate is treated as being approximately independent of the smoothness estimates and so is used to define the reference level for the smoothness values.

Plots of the smoothness estimates and the reference level for each value of p are shown in Figure 4.1.

These plots show that the series is best described by a locally quadratic trend, and that filters based on $p = 2$ are most appropriate. One could argue that the series is essentially locally linear, with a few periods of change such as late 1997 and 2000 to 2002, but these changes are more fully explained by a quadratic model. The locally cubic and quartic filters show no great advantage over the quadratic one, and so these more complex filters are not recommended.

4. Conclusion

The above method provides a diagnostic check for p, the degree of the local polynomial specified by the user when determining the optimal filter weights for a series. Interpreting the plots is similar to investigating residual plots or sample autocorrelation function (acf) plots. It must be remembered that if there are a large number of smoothness values, we expect some of the points to be "large" even when p is chosen correctly, and that each point cannot be considered independent of all others.
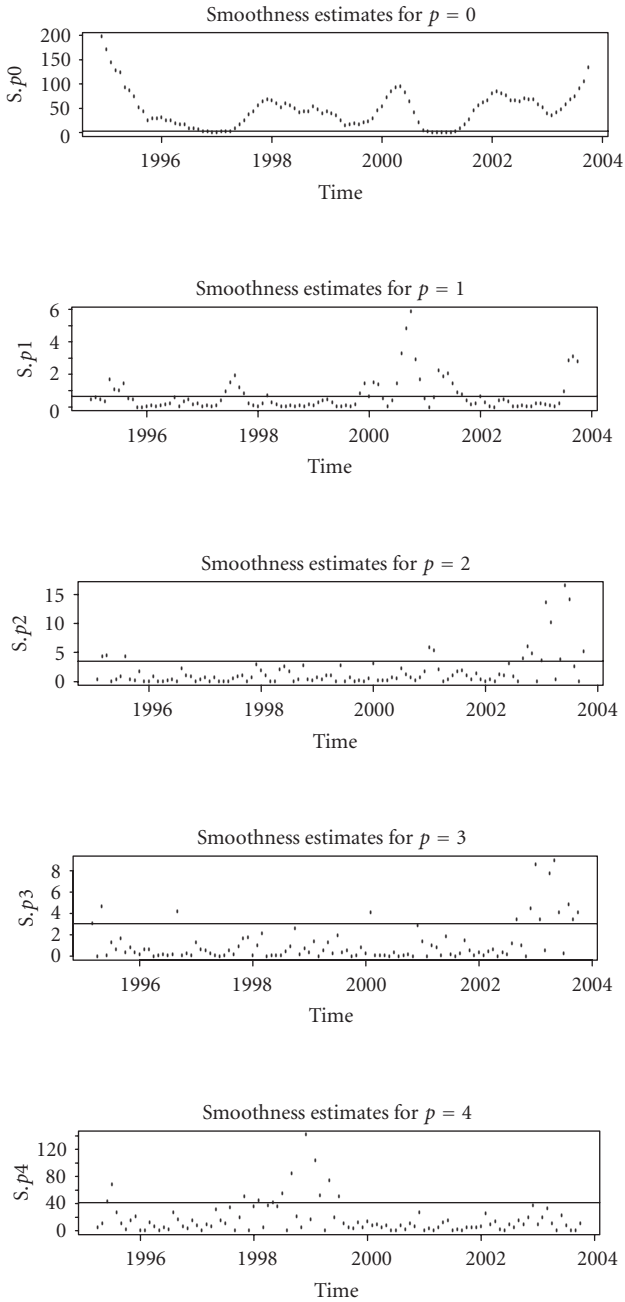
Smoothness estimates for $p = 0$

Smoothness estimates for $p = 1$

Smoothness estimates for $p = 2$

Smoothness estimates for $p = 3$

Smoothness estimates for $p = 4$

Figure 4.1. Empirical smoothness plots for the employment data.

Acknowledgment

References

[1] R. B. Cleveland, W. S. Cleveland, J. E. McRae, and I. J. Terpenning, *STL: A seasonal-trend decomposition procedure based on Loess*, Journal of Official Statistics **6** (1990), no. 1, 3–73.

[2] A. Gray and P. Thomson, *Design of moving-average trend filters using fidelity and smoothness criteria*, Athens Conference on Applied Probability and Time Series Analysis in Memory of E. J. Hannan, Vol. II (1995) (P. Robinson and M. Rosenblatt, eds.), Lecture Notes in Statist., vol. 115, Springer, New York, 1996, pp. 205–219.

[3] R. Henderson, *Note on graduation by adjusted average*, Transactions of the Actuarial Society of America **17** (1916), 43–48.

[4] ———, *A new method of graduation*, Transactions of the Actuarial Society of America **25** (1924), 29–40.

[5] P. B. Kenny and J. Durbin, *Local trend estimation and seasonal adjustment of economic and social time series*, Journal of the Royal Statistical Society. Series A **145** (1982), 1–41.

[6] D. London, *Graduation: The Revision of Estimates*, ACTEX, Connecticut, 1985.

[7] F. Macaulay, *The Smoothing of Time Series*, National Bureau of Economic Research, New York, 1931.

[8] K. V. Mardia, J. T. Kent, and J. M. Bibby, *Multivariate Analysis*, Probability and Mathematical Statistics, A Series of Monographs and Textbooks, Academic Press, London, 1979.

[9] C. H. McLaren and D. G. Steel, *Rotation patterns and trend estimation for repeated surveys using rotation group estimates*, Statistica Neerlandica **55** (2001), no. 2, 221–238.

[10] ———, *Designing trend filters and rotation patterns for repeated surveys*, School of Mathematics and Applied Statistics, University of Wollongong. preprint no. 19/98, 1998.

[11] E. T. Whittaker, *On a new method of graduation*, Proceedings of the Edinburgh Mathematical Society **41** (1923), 63–75.

P. Fogarty: School of Mathematics and Statistics, University of Sydney, NSW 2006, Australia

N. C. Weber: School of Mathematics and Statistics, University of Sydney, NSW 2006, Australia
E-mail address: neville@maths.usyd.edu.au