Survey Methodology, December 1990
Vol. 16, No. 2, pp. 239–253
Statistics Canada

A Method for the Analysis of Seasonal ARIMA Models

DAVID A. BINDER and J. PETER DICK[1]

ABSTRACT

A commonly used model for the analysis of time series models is the seasonal ARIMA model. However, the survey errors of the input data are usually ignored in the analysis. We show, through the use of state-space models with partially improper initial conditions, how to estimate the unknown parameters of this model using maximum likelihood methods. As well, the survey estimates can be smoothed using an empirical Bayes framework and model validation can be performed. We apply these techniques to an unemployment series from the Labour Force Survey.

KEY WORDS: Kalman filter; Partial likelihood; Data smoothing.

1. INTRODUCTION

It is common practice to analyze data from surveys where similar data items are collected on repeated occasions, using time series analysis methods. Most standard methods for these analyses assume the data are either observed without error or have independent measurement errors. However, in the analysis of repeated survey data, when there are overlapping sampling units between occasions, the survey errors can be correlated over time.

A commonly used model in the analysis of time series is the seasonal integrated autoregressive-moving average (ARIMA) regression model, which we discuss in this paper. We show how to incorporate the (possibly correlated) survey errors into the analysis. In particular, we consider the case where the survey (design) error can be assumed to be an ARMA process up to a multiplicative constant.

When such a model for the behaviour of the population characteristics is assumed, the minimum mean squared error, or, equivalently, the Bayes linear estimator for the characteristic at a point in time can be derived. This estimator incorporates the model structure which the classical estimators, such as the minimum variance linear unbiased estimators, ignore. When the model parameters are estimated from the survey data, the estimators are empirical Bayes.

Blight and Scott (1973), Scott and Smith (1974), Scott, Smith and Jones (1977), Jones (1980), Rao, Srinath and Quenneville (1989) and others considered the implications of certain stochastic models for the population means over time. Hausman and Watson (1985) incorporate a measurement error model into the standard seasonal adjustment process. Miazaki (1985) assumed that the survey error could be modelled with a pure moving average process. In Binder and Dick (1989), these results were generalized using state space models and Kalman filters. In this paper, we extend the framework to include the model where differencing of the original series of the population means yields an ARMA model. We use the modified Kalman filter approach given by Kohn and Ansley (1986). To estimate the unknown parameters, we maximize the marginal likelihood function using the method of scoring. This approach can also handle missing data routinely. We also show how the survey estimates can be smoothed to incorporate the model features using empirical Bayes methods. Confidence intervals for these

[1] D.A. Binder, Business Survey Methods Division and J.P. Dick, Social Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6.

smoothed values are also given, using the method described by Ansley and Kohn (1986). Bell and Hillmer (1987) used a similar model but their initial conditions do not extend easily to include regression terms or missing values (while preserving the marginal likelihood approach).

An example of this model is described in Section 5 using unemployment data from the Canadian Labour Force Survey. This example shows the implications on the estimates of the model parameters when the survey errors are taken into account. We derive a smoothed estimate of the underlying process under the model assumptions. Recursive residuals are produced and validation techniques are used to evaluate the various models.

2. THE MODEL

Suppose we have a series of point estimates from a repeated survey of a population characteristic, given by $y_1, y_2, \ldots, y_T$. We assume that y_t can be decomposed into three components, so that

$$y_t = \boldsymbol{x}_t'\gamma + \theta_t + e_t, \tag{2.1}$$

where $\boldsymbol{x}_t'\gamma$ is a deterministic regression term, θ_t is a population parameter following a time series model, and e_t is the survey error, assumed to have zero expectation.

We first describe an integrated seasonal autoregressive-moving average model for $\{\theta_t\}$. We let B be the backshift operator; $\nabla = 1 - B$ and $\nabla_s = 1 - B^s$, where s is the seasonal period. We define the following polynomial functions:

$$\lambda(B) = 1 - \lambda_1 B - \lambda_2 B^2 - \ldots - \lambda_P B^P,$$

$$\alpha(B) = 1 - \alpha_1 B - \alpha_2 B^2 - \ldots - \alpha_p B^p,$$

$$\upsilon(B) = 1 - \upsilon_1 B - \upsilon_2 B^2 - \ldots - \upsilon_Q B^Q,$$

and

$$\beta(B) = 1 - \beta_1 B - \beta_2 B^2 - \ldots - \beta_q B^q.$$

The seasonal ARIMA $(p,d,q)\,(P,D,Q)_s$ model for $\{\theta_t\}$ is given by

$$\lambda(B^s)\alpha(B)\nabla^d \nabla_s^D \theta_t = \upsilon(B^s)\beta(B)\epsilon_t, \tag{2.2}$$

where the ϵ_t's are independent $N(0,\sigma^2)$. We define $a(B) = \lambda(B^s)\alpha(B)$, a $(p + sP)$-degree polynomial; $\Delta(B) = \nabla^d \nabla_s^D$, a $(d + sD)$-degree polynomial; $b(B) = \upsilon(B^s)\beta(B)$, a $(q + sQ)$-degree polynomial; $A(B) = a(B)\Delta(B)$, a $(p + d + sP + sD)$-degree polynomial; $u_t = \Delta(B)\theta_t$, an ARMA$(p + sP, q + sQ)$ process. Therefore, alternative representations of (2.2) are

$$a(B)\Delta(B)\theta_t = b(B)\epsilon_t, \tag{2.3}$$

$$A(B)\theta_t = b(B)\epsilon_t, \tag{2.4}$$

and

$$a(B)u_t = b(B)\epsilon_t. \tag{2.5}$$

We now consider the survey errors $\{e_t\}$ of expression (2.1). It will be assumed that the sample sizes of the repeated survey are sufficiently large that the errors for the survey estimates can be approximated by a multivariate normal distribution. In the simplest case, where the surveys are non-overlapping and the sampling fractions are small, the e_t's can be assumed to be independent. In a rotating panel survey, the survey errors are usually correlated. In this case, since the correlations between survey occasions are zero after panels have been rotated out, a pure moving average process can be used to describe the survey error process.

Alternatively, if a random sample of units are replaced on each survey occasion, a pure autoregressive process may best describe the process. More complicated models are also possible. For example, in a two-stage design, some of the first stage units may be replaced randomly on each occasion and the second stage units may have a rotating panel design. This might be approximated by an autoregressive-moving average process, as suggested by Scott, Smith and Jones (1977).

In this paper, we assume that the survey error process is given by

$$e_t = k_t \omega_t, \tag{2.6}$$

where $\{\omega_t\}$ is an ARMA (m,n) process, given by

$$\phi(B)\omega_t = \psi(B)\eta_t \tag{2.7}$$

and

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_m B^m,$$

and

$$\psi(B) = 1 - \psi_1 B - \psi_2 B^2 - \ldots - \psi_n B^n.$$

The η_t's are independent $N(0,\tau^2)$. The factor k_t has been included in (2.6) to allow for non-homogeneous variances when the autocorrelation function is homogeneous in time.

In the model just described we assume that τ^2, the k_t's and the coefficients of $\phi(B)$ and of $\psi(B)$ can be estimated directly from the survey data, using design-based methods. However, in general, the other parameters are unknown. This includes γ, σ^2, and the coefficients of $\lambda(B)$, $\alpha(B)$, $\upsilon(B)$ and of $\beta(B)$. The x_t's in the regression term are assumed known.

3. STATE SPACE FORMULATION OF THE MODEL

3.1 General Formulation

The model described in Section 2 can be formulated as a state space model with partially improper priors. This has a number of advantages. It permits, through use of a modified Kalman filter, calculation of a marginal likelihood function, which can be maximized to estimate unknown parameters. It also accommodates smoothing of the original survey estimates, by removing the estimates of survey error from the data.

In the state space model, two processes occur simultaneously. The first process, the observation system, details how the observations depend on the current state of the process parameters. The second process, the transition system, details how the parameters evolve over time.

For the state space models we consider here, the observation equation is written as

$$y_t = \boldsymbol{h}_t' \boldsymbol{z}_t \tag{3.1a}$$

and the transition equation is

$$\boldsymbol{z}_t = \boldsymbol{F}\boldsymbol{z}_{t-1} + \boldsymbol{G}\boldsymbol{\xi}_t, \tag{3.1b}$$

where $\boldsymbol{z}_t$ is an $(r \times 1)$ state vector and $\boldsymbol{h}_t$ is a fixed $(r \times 1)$ vector. In the transition equation, $\boldsymbol{F}$ is a fixed $(r \times r)$ transition matrix, $\boldsymbol{G}$ is a fixed $(r \times m)$ matrix and the $\boldsymbol{\xi}_t$'s are independent normal vectors with mean zero and covariance $\boldsymbol{U}$.

The final requirement to complete the specification of the state space process is the initial conditions for $\boldsymbol{z}_0$. In this paper, we shall use the improper prior formulation given in Kohn and Ansley (1986). In general, we assume that $\boldsymbol{z}_0$ has a partially diffuse r-variate normal distribution with mean $\boldsymbol{m}(0 \mid 0) = 0$ and covariance matrix $\boldsymbol{V}(0 \mid 0)$, where

$$\boldsymbol{V}(0 \mid 0) = \kappa \boldsymbol{V}_1(0 \mid 0) + \boldsymbol{V}_0(0 \mid 0) \tag{3.2}$$

for large κ. The matrix $\boldsymbol{V}_1(0 \mid 0)$ specifies the diffuse part of the prior. We explain in Section 3.2 how to obtain $\boldsymbol{V}_1(0 \mid 0)$ and $\boldsymbol{V}_0(0 \mid 0)$ for our model.

We denote the conditional mean of $\boldsymbol{z}_t$ given the observations up to and including time t' by $\boldsymbol{m}(t \mid t')$, and the conditional variance by $\boldsymbol{V}(t \mid t')$, where

$$\boldsymbol{V}(t \mid t') = \kappa \boldsymbol{V}_1(t \mid t') + \boldsymbol{V}_0(t \mid t'). \tag{3.3}$$

Recursive formulae for the cases where $t = t'$ and $t = t' + 1$ are given in Kohn and Ansley (1986). They refer to this as the modified Kalman filter.

Since the model for $\{y_t\}$ given by (2.1) contains survey errors $\{e_t\}$ an estimate of the components without survey error, given by

$$y_t \text{ (smoothed)} = \boldsymbol{x}_t' \boldsymbol{\gamma} + \theta_t \tag{3.4}$$

is often of interest. When the right hand side of (3.4) can be expressed as $\boldsymbol{g}_t' \boldsymbol{z}_t$, for some $\boldsymbol{g}_t'$, then it is possible to obtain the conditional mean and variance of the linear combination $\boldsymbol{g}_t' \boldsymbol{z}_t$ given all the data, using the modified Kalman filter. To do this, the recursions are applied up to time t to obtain $\boldsymbol{m}(t \mid t)$ and $\boldsymbol{V}(t \mid t)$. Then the state vector $\boldsymbol{z}_t$ is augmented by the state $z_{t,r+1} = \boldsymbol{g}_t' \boldsymbol{z}_t$, and $\boldsymbol{m}(t \mid t)$ and $\boldsymbol{V}(t \mid t)$ are also appropriately augmented. The matrix $\boldsymbol{F}$ in (3.1b) is modified to add the equation $z_{t+1,\,r+1} = z_{t,r+1}$. After these modifications, the modified Kalman filter can be used as before, so that the last component of $\boldsymbol{m}(T \mid T)$ gives the conditional expectation of $\boldsymbol{g}_t' \boldsymbol{z}_t$, given all the data, $y_1, y_2, \ldots, y_T$. As well, the last diagonal component of $\boldsymbol{V}(t \mid t)$ gives the conditional variance. This procedure can be generalized to include any number of smoothed estimates and their conditional covariances. In applications, space limitations on the computer might preclude computing the smoothed values for a large number of time points.

3.2 Model for θ

Harvey and Phillips (1979) described a method to put the ARIMA model (2.4) into the state space form given by (3.1). The dimension of $\boldsymbol{z}_t$ is $r = \max(p + d + sP + sD, q + sQ)$. By augmenting $\boldsymbol{A} = (A_1, \ldots, A_{p+d+sP+sD})$ or $\boldsymbol{b} = (b_1, \ldots, b_{q+sQ})$ with zeroes

to have dimension r, the ARIMA model may be written in the form given by (3.1), where $\boldsymbol{h}_t' = (1, 0, \ldots, 0)$, $\boldsymbol{G}_t' = (1, -b_1, \ldots, -b_{r-1})$ and

$$\boldsymbol{F} = \left[\begin{array}{c|c} \begin{matrix} A_1 \\ \vdots \\ A_{r-1} \end{matrix} & I_{r-1} \\ \hline A_r & \mathbf{0}' \end{array}\right],$$

where I_{r-1} is the $(r-1)$ by $(r-1)$ identity matrix and $\mathbf{0}'$ is a row vector of zeroes.

In this formulation, the state vector $z_t = (z_{1t}, \ldots, z_{rt})'$ is defined as

$$\begin{aligned} z_{it} = {} & A_i\theta_{t-1} + A_{i+1}\theta_{t-2} + \ldots + A_r\theta_{t-(r-i+1)} \\ & - b_{i-1}\epsilon_t - b_i\epsilon_{t-1} - \ldots - b_{r-1}\epsilon_{t-(r-i)}, \end{aligned} \tag{3.5}$$

for $i = 2, 3, \ldots, r$ and $z_{1t} = \theta_t$.

To complete the specification for $\{\theta_t\}$, initial conditions for z_0 are required. These are given in Ansley and Kohn (1985), a summary of which is provided here.

From expression (2.5), $\{u_t\}$ is an ARMA process. We define

$$\theta_- = (\theta_0, \theta_{-1}, \ldots, \theta_{-S})',$$

where $S = \max(0, p + sP + d + sD - 1)$. We let

$$u_- = (u_0, u_{-1}, \ldots, u_{-R})',$$

where $R = \max(0, p + sP - 1)$. Finally, we let

$$w_- = (\theta_{-R-1}, \theta_{-R-2}, \ldots, \theta_{-S})',$$

when $S > R$.

Now, u_- is assumed to be a stationary ARMA process, so that its covariance matrix can be derived from expression (2.5). It is assumed that w_- is $N(\mathbf{0}, \kappa I)$ and is independent of u_-. Since $(u_-', w_-')'$ is a non-singular linear combination of θ_-, the covariance matrix for θ can be derived. Using the form of expression (3.5) for z_0, the initial covariance matrix can be computed. Note that when both d and D are zero, so that no differencing takes place in the model, then w_- is the null vector and we have $u_- = \theta_-$.

3.3 Model for the Observed Data

In Section 2 we assumed that $e_t = k_t\,\omega_t$, where ω_t is an ARMA(m,n) model. Therefore, from the discussion in Section 3.2, it is clear that e_t can be represented in state space form, with $\boldsymbol{h}_t = (k_t, 0, \ldots, 0)'$, and $e_t = \boldsymbol{h}_t' z_t$.

The regression component can be similarly represented by adding γ to the state vector and initially, assuming that γ has mean zero and covariance κI. Note that in the transition equation γ remains constant.

Since we can represent each of the components of y_t in expression (2.1) by a state space model, it is straightforward to combine the individual models into an overall model, by extending the state vector to include the state vectors from the individual components. The observation equation is then the sum of the three individual components.

4. ESTIMATION OF THE STATE SPACE MODEL

4.1 Estimation of the Parameters

The unknown parameters of this model are σ^2, and the coefficients of $\lambda(B)$, $\alpha(B)$, $\upsilon(B)$ and $\beta(B)$. We transformed σ^2 to $\log(\sigma^2)$, in the numerical maximization procedure described below to avoid problems with negative parameter values. The model for the vector of observations $y = (y_1, y_2, \ldots, y_T)'$ given in Section 3 is equivalent to

$$y = M\eta + \zeta, \tag{4.1}$$

where η is j-variate $N(\mathbf{0}, \kappa I)$, ζ is T-variate $N(\mathbf{0}, W)$, and M is some fixed $T \times j$ matrix. We note that η contains unknown constants including the regression coefficients; W is a function of the ARMA parameters; M is a function of the differencing structure.

Kohn and Ansley (1986) recommended maximizing the limit of $\kappa^{j/2}$ times the likelihood function for the data, as κ tends to infinity. It can be shown that this limit of the likelihood function is equivalent to the marginal likelihood function of $y - M\hat{\eta}$, where $\hat{\eta}$ is the maximum likelihood estimate of η when M and W are known. Tunnicliffe-Wilson (1989) has shown that the Jacobian of the transformation from the data y to $(\hat{\eta}, y - M\hat{\eta})$ does not depend on the model parameters of W whenever M is known. Ansley and Kohn (1985) have shown that M does not depend on the unknown parameters. By using the modified Kalman filter, the computations for the marginal likelihood function are more straightforward than the approach given by Tunnicliffe-Wilson.

The procedure we employed computes both the marginal likelihood function and its first derivatives with respect to the unknown parameters. This involves taking first derivatives of the initial conditions and of $m(t \mid t')$ and the components of $V(t \mid t')$ for $t = t'$ and $t = t' + 1$. All the computations were done using PROC IML in SAS.

The likelihood function was maximized using a modification of the method of scoring. This modification allowed for varying step sizes. On each iteration, the likelihood function was computed at the previous step size, as well as at this step size multiplied and divided by a predetermined constant. (We used 1.1 as the factor.) The next step size was to choose the point which maximized the likelihood function among the three points. Each time a check was made to determine whether the parameters were in range. This was done by checking for positive semi-definiteness of the initial covariance matrix of the state vector. If it was out of range, the step size was divided again by the constant and the procedure repeated.

To estimate the variance matrix for the estimated parameters, the inverse of the Fisher information matrix was used. This is readily computed since the first derivatives of the likelihood function are available.

4.2 Estimation of the Smoothed Values

Smoothed values as defined in (3.4) for the estimates can be obtained by zeroing out that component of the state vector which corresponds to the survey error. However, this still leaves open the question of how to estimate its variance. To derive the standard error of the smoothed

estimate it is necessary to account for the fact that the unknown parameters have been estimated from the data, particularly when the data series is short; see Jones (1979).

To obtain the variance of $g'z_t$, it is sufficient to derive the variance $z_T - \hat{m}(T \mid T)$, where $\hat{m}(T \mid T)$ is the estimate of $m(T \mid T)$ at the estimated parameter values. This is because the state vector has been augmented to include $g'z_t$. Now,

$$z_T - \hat{m}(T \mid T) = [z_T - m(T \mid T)] + [m(T \mid T) - \hat{m}(T \mid T)]. \tag{4.2}$$

The first component of the right hand side of (4.2) has conditional variance $V(T \mid T) = V_0(T \mid T)$, assuming that $V_1(T \mid T) = \mathbf{0}$. The second component of (4.2) represents a bias term and is independent of the first term, since it depends only on the data y. By taking a Taylor series expansion of the second term around the true parameter values and ignoring higher terms, we have the second component of (4.2) is

$$m(T \mid T) - \hat{m}(T \mid T) = \left[\frac{-\partial \hat{m}(T \mid T)}{\partial \phi}\right]' (\hat{\phi} - \phi), \tag{4.3}$$

where ϕ is the vector of unknown parameters and $\hat{\phi}$ is its estimate. Therefore, the asymptotic variance of (4.2) is approximately

$$\mathrm{Var}[z_T - \hat{m}(T \mid T)] = V_0(T \mid T) + \left[\frac{\partial \hat{m}(T \mid T)}{\partial \phi}\right]' V_\phi \left[\frac{\partial \hat{m}(T \mid T)}{\partial \phi}\right], \tag{4.4}$$

where V_ϕ is the covariance matrix for the unknown parameters. Expression (4.4) is estimated by using the estimated parameter values. This is the same approach as that given by Ansley and Kohn (1986).

4.3 Generalized Recursive Residuals

As Harvey and Durbin (1986) pointed out, useful quantities for performing model diagnostics are the generalized recursive residuals. In terms of our state space model, this is the difference between the observation and the one-step ahead prediction from the Kalman filter. These can be used for all time points t where $V_1(t + 1 \mid t) = \mathbf{0}$. Under the model, these residuals are approximately independent normal. They can be standardized to have an estimated variance of unity under the model. Diagnostics similar to those used in classical regression models can then be performed.

5. ANALYSIS OF LABOUR FORCE DATA

5.1 Parameter Estimation

To demonstrate this procedure, we take data from the Canadian Labour Force Survey (LFS). The LFS is a monthly rotating panel survey with each panel containing one-sixth of the selected households. A panel will remain in the sample for six consecutive months while the primary sampling units will rotate out after approximately two years. The sample selection follows a stratified multi-stage design.

The data were the monthly number of unemployed as published from January 1977 to December 1986 for the province of Nova Scotia and for the subprovincial region within Nova Scotia corresponding to Cape Breton Island. This province was selected because the sampling errors are moderate compared to the larger provinces. Cape Breton Island was selected because its smaller sample size provides estimates with a larger relative variance. Graph 1a displays the logarithm of the Nova Scotia series and Graph 1b shows the similarly transformed Cape Breton Island series. We used the logarithms as our inputs.

Lee (1990) estimated the autocorrelations for the Nova Scotia survey error up to a lag of eleven. We derived the coefficients of the ARMA (m,n) survey error process given in (2.7) by matching these autocorrelations. A good fit was found using an ARMA (3,6) model. The resulting coefficients were:

$$\begin{aligned} \phi_1 &= 0.2575 & \psi_1 &= -0.1847 \\ \phi_2 &= -0.3580 & \psi_2 &= -0.5873 \\ \phi_3 &= -0.6041 & \psi_3 &= 0.3496 \\ & & \psi_4 &= 0.0647 \\ \tau^2 &= 0.7246 & \psi_5 &= 0.0982 \\ & & \psi_6 &= 0.0347. \end{aligned}$$

The k_t's of (2.6) were the estimated standard errors of the estimates, derived by taking a Taylor series approximation for the logarithms.

A series of models were fitted to the Nova Scotia data with an assumption of no sampling error. The same models were then refitted, incorporating the model for the survey error process. In this case we could also compute smoothed values for the survey estimates and compare their standard errors with the standard errors of the original series.

The preliminary model selected for the Nova Scotia data, ignoring the sampling error, was a seasonal ARIMA $(1,1,0)(0,1,1)_{12}$. However the moving average term for the seasonal component was estimated to be one, so a deterministic regression term was used to account for the seasonality. The 12 regression variables included a linear term and a dummy variable for each of the first 11 months. The dummy variable for a reference month took the value 1 for the reference month, -1 for December and 0 for the other months. Note that an intercept term is not appropriate for this model because the first differences of the data are fitted.

Further analysis of this reduced model showed that the moving average seasonal component was not required in the model. The final model selected for the Nova Scotia data was an ARIMA (1,1,0) with a deterministic regression component. This same model was then used for the Nova Scotia data with the survey error process incorporated. The same structural model was used for the Cape Breton Island series.

Table 1 displays the parameter estimates. The estimates that do not incorporate the survey error component are in the **Without Sampling Errors** columns. First, examining the models for Cape Breton Island shows that the regression estimates are similar, as would be expected. Note that the autoregressive estimates (AR) are also similar and that the **With Sample Error** model has reduced the estimated model variance substantially. The column headed ***T*-value** displays the estimated parameter divided by its standard error. Note that the t-values for the autoregressive parameter are substantially different (-0.68 vs -2.85). This would lead to

Table 1

Parameter Estimates - Unemployment Series 1977-1986

Parameter	Nova Scotia				Cape Breton Island			
	Without Sampling Error		With Sampling Error		Without Sampling Error		With Sampling Error	
	Estimate	*T*-value	Estimate	*T*-value	Estimate	*T*-value	Estimate	*T*-value
Alpha	−0.296	−3.23	0.862	2.08	−0.260	−2.85	−0.231	−0.68
Sigma	0.0597	-	0.0032	-	0.1049	-	0.0520	-
Trend	0.00427	1.01	0.00420	1.89	0.00607	0.79	0.00598	1.50
January	0.064	3.60	0.048	1.93	−0.007	−0.23	−0.003	−0.10
February	0.083	4.80	0.078	3.30	0.027	0.89	0.028	0.97
March	0.166	10.20	0.165	6.40	0.171	5.76	0.164	5.76
April	0.106	6.60	0.104	4.10	0.099	3.33	0.089	3.19
May	0.009	0.60	0.016	0.70	−0.008	−0.28	−0.007	−0.24
June	−0.101	−6.00	−0.088	−3.30	−0.029	−0.96	−0.033	−1.17
July	−0.016	−1.20	−0.014	−0.63	0.082	2.77	0.081	3.13
August	−0.058	−3.60	−0.062	−2.37	−0.011	−0.37	−0.009	−0.30
September	−0.106	−6.60	−0.105	−3.96	−0.104	−3.51	−0.098	−3.18
October	−0.081	−4.80	−0.071	−3.08	−0.084	−2.83	−0.069	−2.44
November	−0.026	−1.80	−0.029	−1.08	−0.063	−2.10	−0.074	−2.46

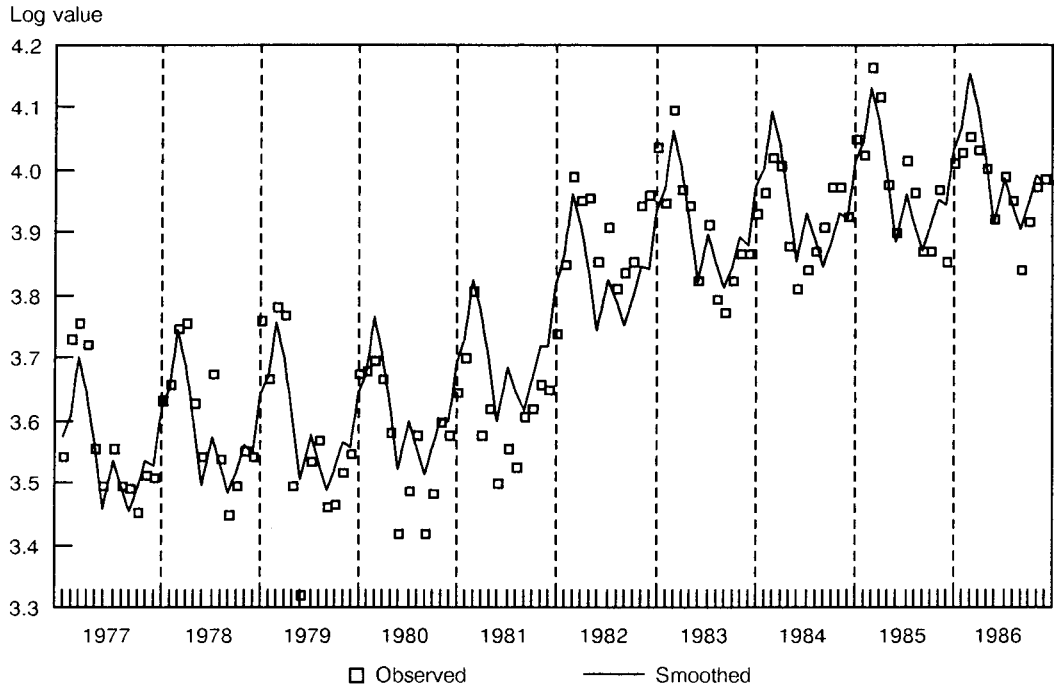
accepting a model for the Cape Breton Island data with only a deterministic regression term when the survey error process is incorporated into the model. However, if the survey error is ignored in the analysis, too much significance would be attached to the autoregressive parameter.

The results for the Nova Scotia models are also displayed on Table 1. Note that the reduction in the estimate of the model variance by incorporating the sampling error structure is much greater for the Nova Scotia series than was achieved for the Cape Breton data. An important result in the Nova Scotia models is the difference in the estimates for the autoregressive component. Both models show that the AR component is highly significant in each model. The **Without Sample Error** model gives an estimate of $\alpha = -0.296$; whereas the **With Sample Error** model gives an estimate of $\alpha = 0.862$. Clearly, the interpretations that would be associated with these two estimates are entirely different.

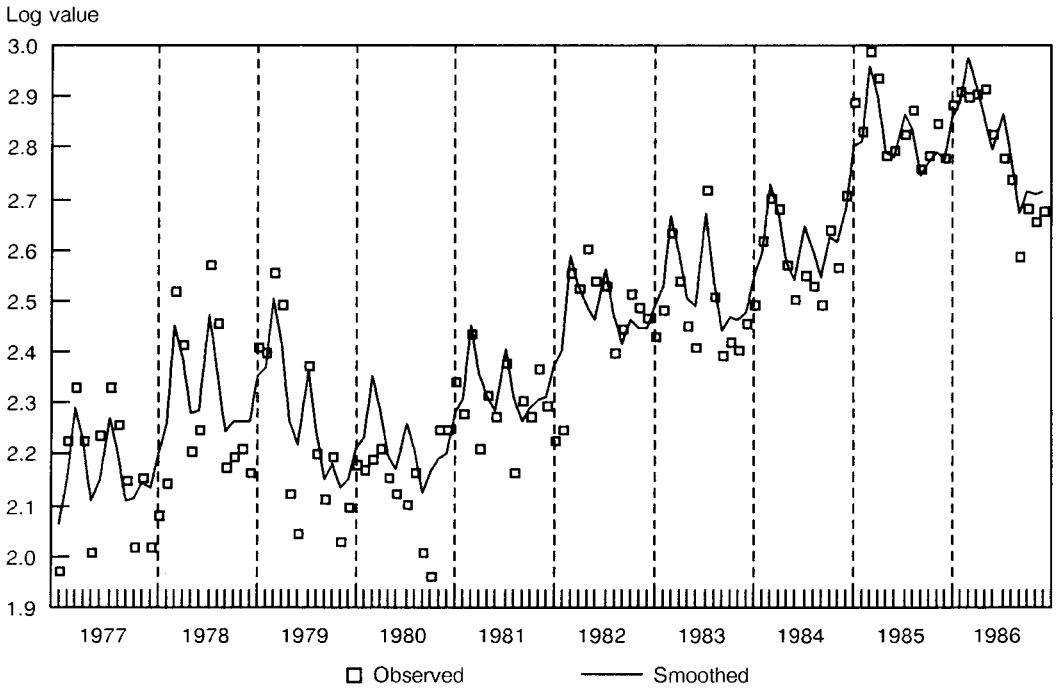
The smoothed estimates for the model incorporating sampling error are shown superimposed on the original data series in Graph 1a. Graph 1b shows the smoothed estimates for Cape Breton Island superimposed on the original series. The most notable item in these plots is the impact of the recession of 1981 on the smoothed estimates. Prior to the recession, the model tends to overestimate unemployment and after 1981 the model tends to underestimate the number of unemployed.

5.2 Model Validation

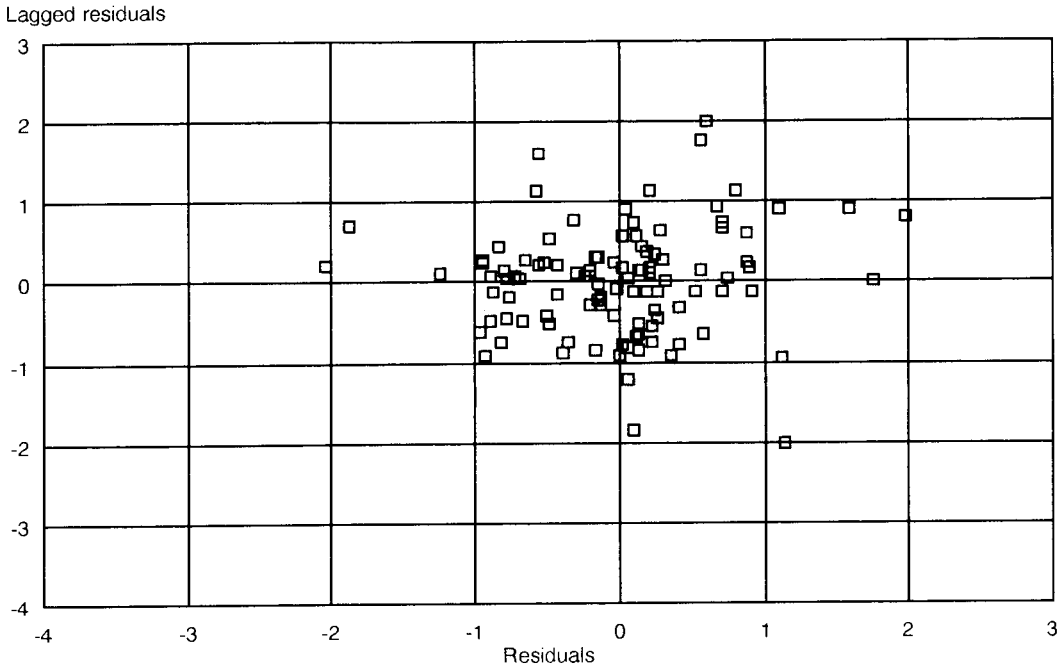
The plots of the generalized recursive residuals (described in Section 4.3) against the lagged generalized recursive residuals were produced for all the models. Graphs 2a and 2b show these plots for the two models for Nova Scotia. Note that Graph 2a shows less dispersion around the origin than Graph 2b, indicating a better fit when survey error is incorporated in the model.



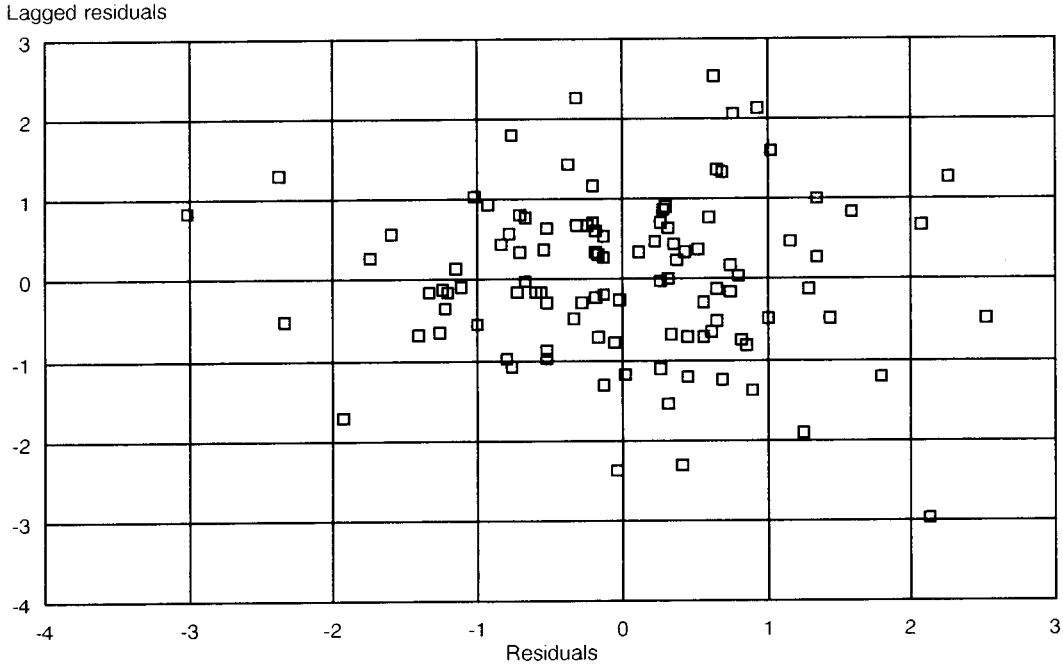
Graph 1a Nova Scotia Observed and Smoothed Values (Log Transform)



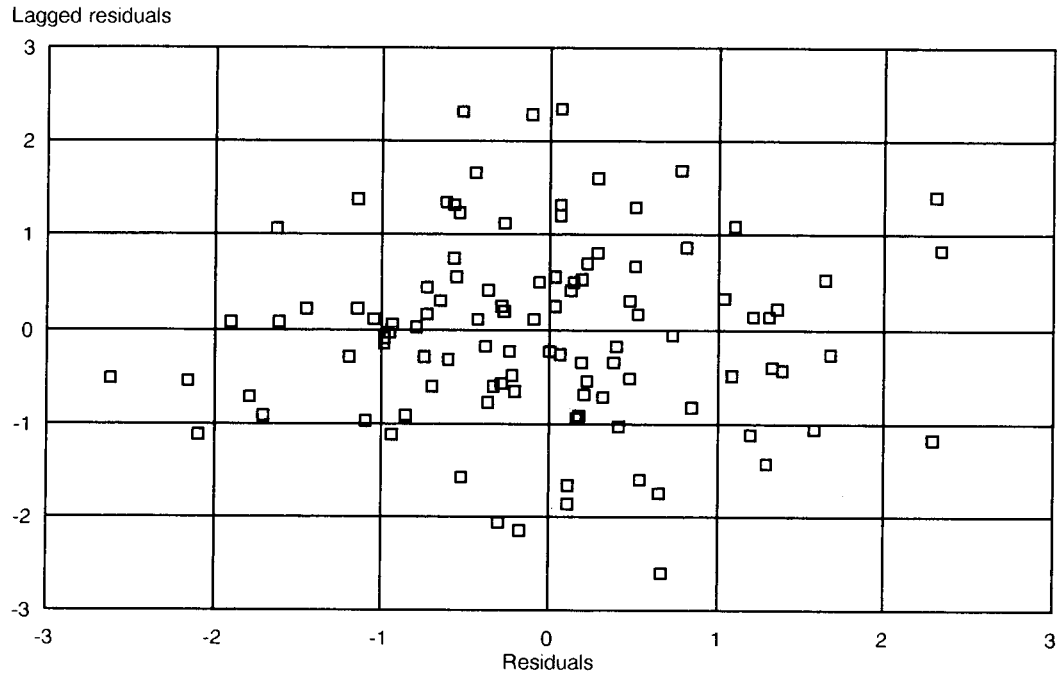
Graph 1b Cape Breton Island Observed and Smoothed Values (Log Transform)



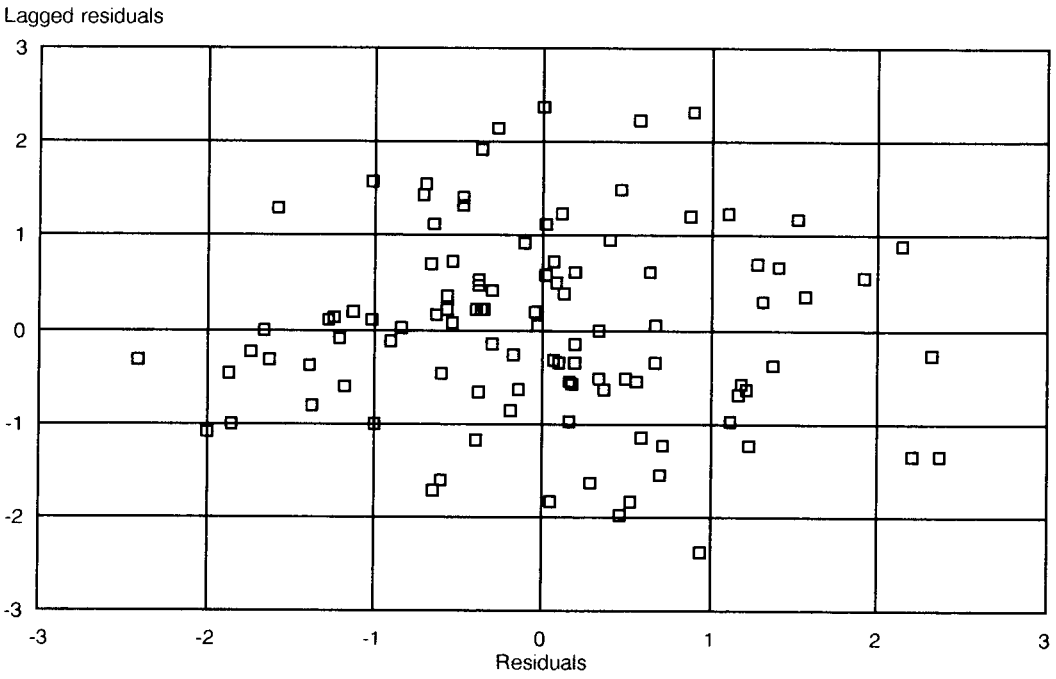
Graph 2a Nova Scotia One Step Ahead Prediction Errors – Survey Error Included



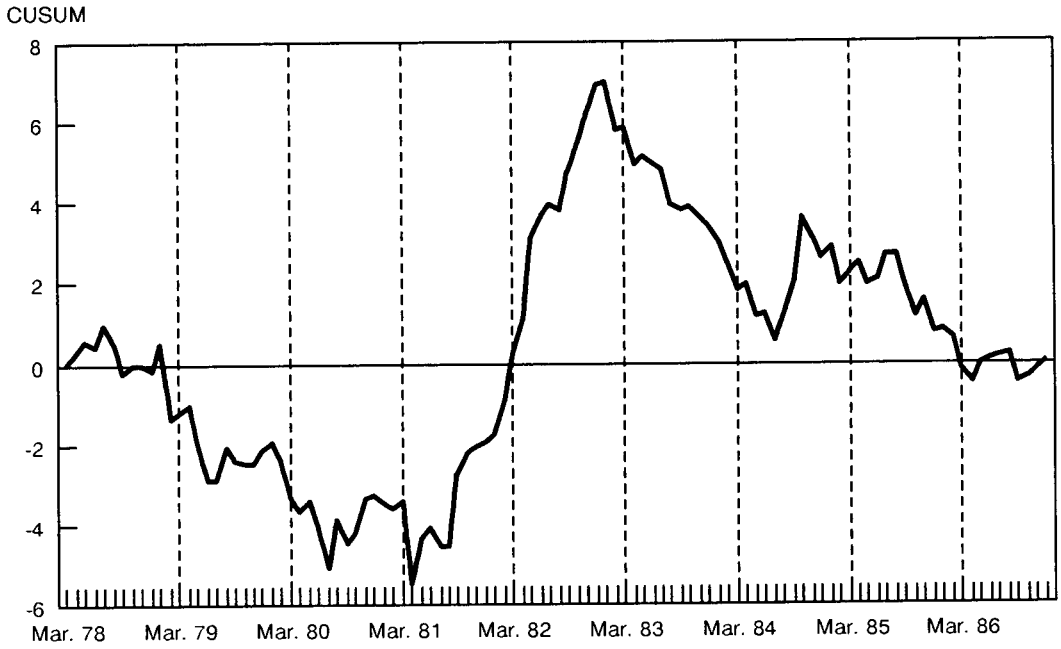
Graph 2b Nova Scotia One Step Ahead Prediction Errors – Survey Error Ignored



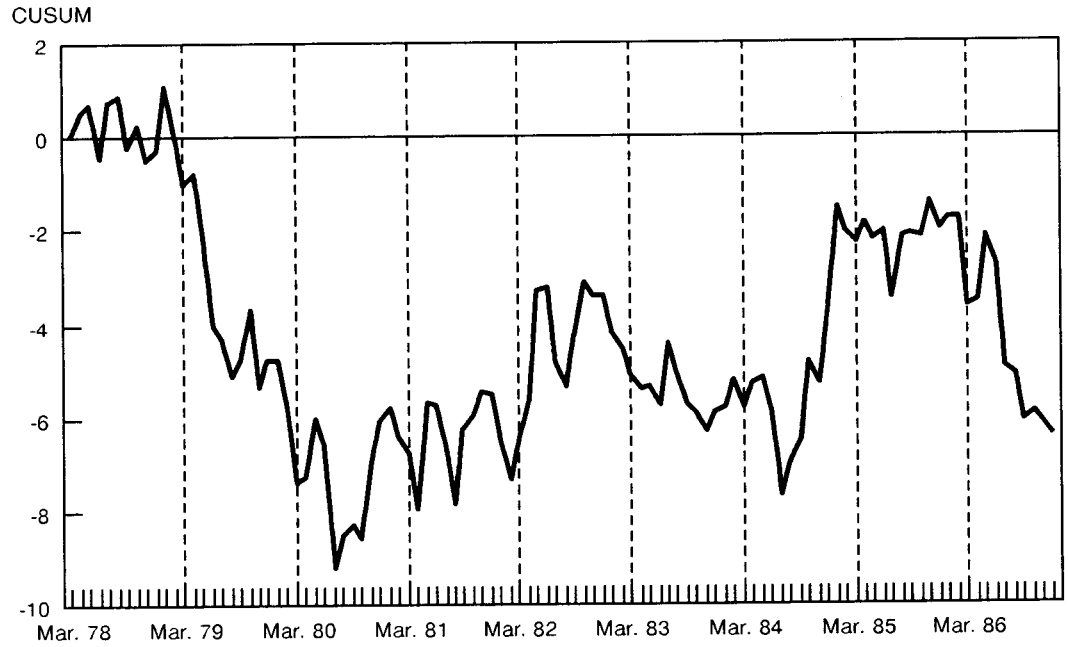
Graph 3a Cape Breton Island One Step Ahead Prediction Errors – Survey Error Included



Graph 3b Cape Breton Island One Step Ahead Prediction Errors – Survey Error Ignored



Graph 4a Nova Scotia CUSUM of One Step Ahead Prediction Errors



Graph 4b Cape Breton Island CUSUM of One Step Ahead Prediction Errors

The same plots for Cape Breton Island are shown in Graph 3a and 3b. There is a striking similarity in the resulting residual plots for the two models from Cape Breton. However, none of the four plots give any compelling reason to doubt the underlying normal assumption of any of the models.

To test that the models did not undergo a structural change, the recursive residuals can be cumulatively summed to create a CUSUM chart. Whereas using the tests described in Brown, Durbin and Evans (1975) produced no significant results, the chart does suggest some structural change may be occurring. The CUSUM for Nova Scotia, as displayed in Graph 4a, shows quite clearly that prior to the recession the residuals are generally negative, implying that the model predictors are too large. During the 1981 recession the model produces mainly positive residuals. This implies that the model predictors are too small. The CUSUM for the Cape Breton Island models is shown in Graph 4b. Here we can see that the model that includes the survey error undergoes an earlier structural change.

We see, therefore, that model improvements can be made. By incorporating an extra regression variable corresponding to the structural changes noted in the CUSUM chart, further analysis can be performed within the same general framework. The form of such a variable is currently being investigated.

5.3 Summary

These examples demonstrate the importance of accounting for survey errors in certain time series analyses. Using the modified Kalman filter, we have developed a flexible method for parameter estimation, data smoothing and model validation for a wide variety of commonly used models.

ACKNOWLEDGEMENTS

The authors are grateful to Bill Steele of Social Survey Methods Division for his work in programming the algorithm described in Section 4. We are also grateful to the Labour Force Survey for supplying the data series. The referee and the associate editor both supplied many useful comments that improved the paper.

REFERENCES

ANSLEY, C.F., and KOHN, R. (1985). A structured state space approach to computing the likelihood of an ARIMA process and its derivatives. *Journal of Statistical Computation and Simulation*, 21, 135-169.

ANSLEY, C.F., and KOHN, R. (1986). Prediction mean squared error for state space models with estimated parameters. *Biometrika*, 73, 467-473.

BELL, W.R., and HILLMER, S.C. (1987). Time Series methods for survey estimation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 83-92.

BINDER, D.A., and DICK, J.P. (1989). Modelling and estimation for repeated surveys. *Survey Methodology*, 15, 29-45.

BLIGHT, B.J.N., and SCOTT, A.J. (1973). A stochastic model for repeated surveys. *Journal of the Royal Statistical Society*, Series B, 35, 61-68.

BROWN, R.L., DURBIN, J., and EVANS, J.M. (1975). Techniques for testing the consistency of regression relationships over time. *Journal of the Royal Statistical Society*, Series B, 37, 149-163.

HARVEY, A.C., and DURBIN, J. (1986). The effects of seat belt legislation on British road casualties: A case study in structural time series modelling. *Journal of the Royal Statistical Society*, Series A, 149, 187-222.

HARVEY, A.C., and PHILLIPS, G.D.A. (1979). Maximum likelihood estimation of regression models with autoregressive-moving average disturbances. *Biometrika*, 66, 49-58.

HAUSMAN, J.A., and WATSON, M.W. (1985). Errors in variables and seasonal adjustment procedures. *Journal of the American Statistical Association*, 80, 531-540.

JONES, R.G. (1979). The efficiency of time series estimators for repeated surveys. *Australian Journal of Statistics*, 21, 45-56.

JONES, R.G. (1980). Best linear unbiased estimators for repeated surveys. *Journal of the Royal Statistical Society*, Series B, 42, 221-226.

KOHN, R., and ANSLEY, C.F. (1986). Estimation, prediction and interpolation for ARIMA models with missing data, *Journal of the American Statistical Association*, 81, 751-761.

LEE, H. (1990). Estimation of panel correlations for the Canadian Labour Force Survey. *Survey Methodology*, 16, 283-292.

MIAZAKI, E.S. (1985). Estimation for time series subject to the error of rotation sampling. *Ph. D. Thesis*, Iowa State University, Ames, Iowa.

RAO, J.N.K., SRINATH, K.P., and QUENNEVILLE, B. (1989). Optimal estimation of level and change using current preliminary data. In *Panel Surveys* (Eds. D. Kasprzyk, G. Duncan, G. Kaltin and M.P. Singh), New York: Wiley, 457-479.

SCOTT, A.J., and SMITH, T.M.F. (1974). Analysis of repeated surveys using time series methods. *Journal of the American Statistical Association*, 69, 674-678.

SCOTT, A.J., SMITH, T.M.F., and JONES, R.G. (1977). The application of time series methods to the analysis of repeated surveys, *International Statistics Review*, 45, 13-28.

TUNNICLIFFE-WILSON, G. (1989). On the use of marginal likelihood in time series model estimation. *Journal of the Royal Statistical Society*, Series B, 51, 15-27.