Jour. Ind. Soc. Ag. Statistics
54(1), 2001 : 43–51

State Space Modelling Versus ARIMA Time-Series Modelling

S. Ravichandran[1] and Prajneshu
Indian Agricultural Statistics Research Institute, New Delhi-110012
(Received : June, 2000)

SUMMARY

Box-Jenkins Autoregressive Integrated Moving Average (ARIMA) procedure is generally used for analyzing time-series data. In this article, another approach, which is quite promising, viz. State space modelling approach using Kalman filtering technique, is studied in detail. The advantage of this technique is that it can take into account the time dependency of the underlying parameters. As an illustration, the two approaches are compared for modelling and forecasting all-India marine products export data.

Key words : ARIMA time-series procedure, Kalman filtering technique, Marine products export data, State space modelling.

1. Introduction

In agriculture, data are generally collected over time and Box-Jenkins ARIMA time-series modelling procedure is used to analyze these data (see e.g. Box, *et al.* [2]). In this approach, the underlying parameters are assumed to be constants but in reality this assumption is rarely met. Evidently, the parameters are time-dependent, and to handle such a situation another approach, viz. State space modelling using Kalman filtering technique may be effectively used. This procedure has been mostly used in control engineering (Meinhold and Singpurwalla [5]) but, to the best of our knowledge, it has not so far been used to model data from Indian agriculture. Accordingly, the purpose of this paper is to highlight importance of this dynamical modelling technique so that researchers may start applying it in their respective fields of studies. As an illustration, the two procedures are compared to model data of all-India marine products export.

---

1 *Present address* : Scientist, Computer Centre, Indian Council of Agricultural Research, Krishi Bhawan, New Delhi-110001

2. *State Space Model*

As in Box-Jenkins approach, state space modelling is applicable only when the time-series data is made stationary. This latter approach represents a multivariate time-series through auxiliary variables (called state variables), some of which may or may not be directly observable. The state vector summarizes all the information from the present and past values of the time-series relevant to the prediction of future values of the series. The observed time-series are expressed as linear combinations of the state variables. The state space model is also called a "Markovian representation," or "Canonical representation" of a multivariate time-series process. It is defined with the help of two equations, viz. "State transition equation" and "Measurement equation". Let X_t be the $r \times 1$ vector of observed variables, after differencing and subtracting the sample mean and let $\underset{\sim}{Z}_t$ be the state vector of dimension s, $s \geq r$, where the first r components of $\underset{\sim}{Z}_t$ consist of X_t. Let $X_{t+k|t}$ denote the conditional expectation of X_{t+k} based on the information available at time t. Then the last $s - r$ elements of $\underset{\sim}{Z}_t$ consist of elements of $X_{t+k|t}$ where $k > 0$ is specified. Then the model is defined as

$$\underset{\sim}{Z}_{t+1} = F\underset{\sim}{Z}_t + G e_{t+1} \tag{1}$$

Eq. (1) is the state transition equation. F is the transition matrix of order $s \times s$. This matrix determines the dynamic properties of the model. The $s \times r$ coefficient matrix G is the input matrix which determines the variance structure of eq. (1). For model identification, the first r rows and r columns of G are set to an identity matrix of order $r \times r$. The input vector e_t is a sequence of independent, normally distributed random vectors of dimension r with mean vector 0 and covariance matrix Σ_{ee}. The random error e_t is sometimes called the "Innovation vector". In addition to the state transition equation, state space models usually include a measurement or observation equation that gives the observed values X_t as a function of the state vector $\underset{\sim}{Z}_t$. Since the observed values X_t are included in the state vector, the observation equation here merely represents the extraction of first r components of the state vector. Hence the measurement or observation equation is given by

$$X_t = [I_r \quad 0]\underset{\sim}{Z}_t \tag{2}$$

where I_r is an identity matrix of dimension $r \times r$.

3. Estimation Procedure

State space procedure employs canonical correlation analysis for the identification of state space model (see e.g. Aoki [1]). First fit a sequence of unrestricted vector autoregressive (VAR) models and compute Akaike's Information Criterion (AIC) for each model given by

$$AIC_p = n \ln (|\hat{\Sigma}_p|) + 2pr^2 \tag{3}$$

where $\hat{\Sigma}_p$ is an estimate of innovation variance matrix obtained by fitting an autoregressive model of order p. The smallest AIC value determines the number of autocovariance matrices to be included in the canonical correlation analysis. It may be mentioned that in this paper, the symbol "′" indicates transpose of a matrix. The vector autoregressive models are estimated using the sample autocovariances :

$$\Gamma_i = E(X_t \; X'_{t-i})$$

Thus the Yule-Walker equations are :

$$\begin{bmatrix} \Gamma_0 & \Gamma_1 & \cdots & \Gamma_{p-1} \\ \Gamma'_1 & \Gamma_0 & \cdots & \Gamma_{p-2} \\ \vdots & \vdots & \cdots & \vdots \\ \Gamma'_{p-1} & \Gamma'_{p-2} & \cdots & \Gamma_0 \end{bmatrix} \begin{bmatrix} \Phi_1^p \\ \Phi_2^p \\ \vdots \\ \Phi_p^p \end{bmatrix} = \begin{bmatrix} \Gamma_1 \\ \Gamma_2 \\ \vdots \\ \Gamma_p \end{bmatrix} \tag{4}$$

and

$$\begin{bmatrix} \Gamma_0 & \Gamma'_1 & \cdots & \Gamma'_{p-1} \\ \Gamma_1 & \Gamma_0 & \cdots & \Gamma'_{p-2} \\ \vdots & \vdots & \cdots & \vdots \\ \Gamma_{p-1} & \Gamma_{p-2} & \cdots & \Gamma_0 \end{bmatrix} \begin{bmatrix} \Psi_1^p \\ \Psi_2^p \\ \vdots \\ \Psi_p^p \end{bmatrix} = \begin{bmatrix} \Gamma'_1 \\ \Gamma'_2 \\ \vdots \\ \Gamma'_p \end{bmatrix} \tag{5}$$

Here Φ_i^p are the coefficient matrices for the past observation form of the vector autoregressive model, and Ψ_i^p are the coefficient matrices for the future observation form which are given respectively as

$$X_t = \sum_{i=1}^{p} \Phi_i^p \; X_{t-i} + e_t \tag{6}$$

where e_t is a vector white noise sequence with mean vector 0 and covariance matrix Σ_p. This is the forward autoregressive form based on the past

observations. The backward autoregressive form based on the future observations is written as

$$X_t = \sum_{i=1}^{p} \Psi_i^p \, X_{t+i} + \eta_t \tag{7}$$

where η_t is a vector of white noise sequence with mean vector 0 and covariance matrix Ω_p. Elements of state vector are determined with the help of a sequence of canonical correlation analysis of the sample autocovariance matrices through the selected order. This analysis computes the sample canonical correlations of the past with an increasing number of steps into the future. Variables that yield significant correlations are added to the state vector and the remaining are excluded from further consideration. Once the state vector is determined, the state space model is fit to the data. The parameters in F, G and Σ_{ee} are estimated using maximum likelihood procedure. Here

$$L = -n/2 \ln (|\Sigma_{ee}|) - 1/2 \text{ trace } (\Sigma_{ee}^{-1} EE') \tag{8}$$

where E is a $r \times n$ matrix of innovations given as

$$E = [e_1 \ldots e_n] \tag{9}$$

After the parameters are estimated, forecasts are obtained from the selected state space model using Kalman filtering technique (Harvey [3]). The m-step ahead forecast of Z_{t+m} is $Z_{t+m|t}$, where $Z_{t+m|t}$ denotes the conditional expectation of Z_{t+m} given the information available at time t, i.e.

$$X_{t+m|t} = H Z_{t+m|t} \tag{10}$$

where the matrix

$$H = [I_r \quad 0]$$

The m-step ahead forecast error is

$$Z_{t+m} - Z_{t+m|t} = \sum_{i=0}^{m-1} \Psi_i e_{t+m-i} \tag{11}$$

and its variance is

$$V_{z,m} = \sum_{i=0}^{m-1} \Psi_i \Sigma_{ee} \Psi'_i \tag{12}$$

Letting $V_{z,0} = 0$, $V_{z,m}$ can be computed recursively using Kalman filtering technique as

$$V_{z,m} = V_{z,m-1} + \Psi_{m-1}\ \Sigma_{ee}\ \Psi'_{m-1} \tag{13}$$

Thus the variance of m-step ahead forecast error of X_{t+m}, obtained from eq. (13), is

$$V_{x,m} = H\ V_{z,m}\ H' \tag{14}$$

Goodness of fit of the selected model is assessed using Akaike's Information Criterion (AIC), Schwartz- Bayesian Information Criterion (SBC), Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) (Shumway and Stoffer [6]). Lower the values of these statistics, better is the fitted model.

It may be pointed out that none of the standard statistical packages, viz. SPSS, GENSTAT, MSTAT are capable of handling state space analysis of time-series data. To the best of our knowledge, only Statistical Analysis System (SAS) software package contains programs for fitting state space models. SAS modules, viz. PROC STATESPACE and PROC ARIMA, available in SAS [8] are utilized for data analysis.

4. An Illustration

As an illustration, the univariate data pertaining to all-India marine products export during the period 1975-'98, obtained from various reports of MPEDA [4], is considered. The data is presented in Table 1 for ready reference. Following Venugopalan and Prajneshu [7], ARIMA time-series modelling technique is first applied to analyze the data. Autocorrelations (r_k) and Partial autocorrelations (Φ_{kk}) up to lag 10 are worked out. Since the computed r_k values do not tail off towards zero, the original series is found to be nonstationary and the stationarity is achieved after first differencing, *i.e.* d = 1. Patterns of r_k and Φ_{kk} in respect of differenced data show that the partial autocorrelations cut off after lag 1 and autocorrelations tail off towards zero after initial spikes. However, to save space, r_k and Φ_{kk} values are not presented. Hence, for describing the given data, ARIMA (1, 1, 0) is identified as the best model which is given by

$$X_t - X_{t-1} = \mu\,(1 - \Phi_1) + \Phi_1\,(X_{t-1} - X_{t-2}) + e_t \tag{15}$$

where μ is a constant and Φ_1 denotes the AR parameter. The above equation is fitted to the given data and the estimates of μ and the Φ_1 are obtained as

$$\hat{\mu} = 16.99,\ \hat{\Phi}_1 = 0.87$$

Table 1. All India marine products export data

Year	Quantity (in thousand tonnes)
1975-'76	54.5
1976-'77	66.8
1977-'78	66.0
1978-'79	86.9
1979-'80	86.4
1980-'81	75.6
1981-'82	70.1
1982-'83	78.2
1983-'84	92.7
1984-'85	86.2
1985-'86	83.7
1986-'87	85.8
1987-'88	97.2
1988-'89	99.8
1989-'90	110.2
1990-'91	139.4
1991-'92	171.8
1992-'93	208.6
1993-'94	244.0
1994-'95	307.3
1995-'96	296.3
1996-'97	378.2
1997-'98	385.8

Subsequently, state space modelling procedure is applied to the transformed data. AIC values are worked out to find out the number of autoregressive models fit to the series. Table 2 shows that the smallest AIC value is at lag 1, which determines the number of autocovariance matrices analyzed in the canonical correlation phase. The selection of a first-order autoregressive model by the AIC statistic looks reasonable for the given data because partial autocorrelations for lags greater than 1 are not significant. Next the Yule-Walker estimates for the selected autoregressive models are obtained. For the given data, number of lags comes out as 1. After the autoregressive order selection process has determined the number of lags, canonical correlation analysis phase selects the state vector. Once that is selected, the state space model is estimated by maximum likelihood procedure. Information from the canonical correlation analysis and from the preliminary autoregression analysis is used to form preliminary estimate of the parameter of state space model.

Table 2. AIC values for state space model

Lag	0	1	2	3	4	5
AIC	111.79	106.97	108.60	110.55	112.51	113.37
Lag	6	7	8	9	10	
AIC	114.35	115.90	117.78	119.76	121.36	

This preliminary estimate is used as starting value for the recursive estimation process.

For the given data, the final maximum likelihood estimate is obtained as F(1, 1) = 0.549. Hence the fitted state space model is

$$X_{t+1} = 0.549\, X_t + e_{t+1}, \quad \mathrm{Var}[e_{t+1}] = 250.76 \tag{16}$$

The identified ARIMA (1, 1, 0) and state space models given respectively by eqs. (15) and (16) are fitted to the data ignoring a few observations towards the end, say three, and forecast values on the basis of the identified models are compared with the actual observations for assessing accuracy of the fitted models. Forecast values for 1995-'96, 1996-'97, 1997- '98 using ARIMA and state space models are respectively 364.50, 416.41, 463.72 and 348.06, 376.44, 398.03 thousand tonnes against the actual respective values of 296.3, 378.2, 385.8 thousand tonnes, thereby indicating that the latter approach yields closer forecast values. Further, goodness of fit statistics, viz. AIC, SBC, RMSE and MAE presented in Table 3 also show that the state space model provides lower

Table 3. Goodness of fit statistics

Statistic	Model	
	ARIMA (1, 1, 0)	State Space Model
AIC	155.73	106.97
SBC	157.62	108.27
RMSE	13.87	9.45
MAE	12.59	6.87

values as compared to ARIMA model. Thus, for the data set under consideration, state space modelling technique performs better than the ARIMA approach. Finally, forecasts of all-India marine products for the next five years starting from 1998-'99, using state space model, are respectively 415.87, 431.68, 446.36, 460.15 and 472.28 thousand tonnes. In order to get visual insight, the graph of fitted state space model along with data points is exhibited in Fig.1.

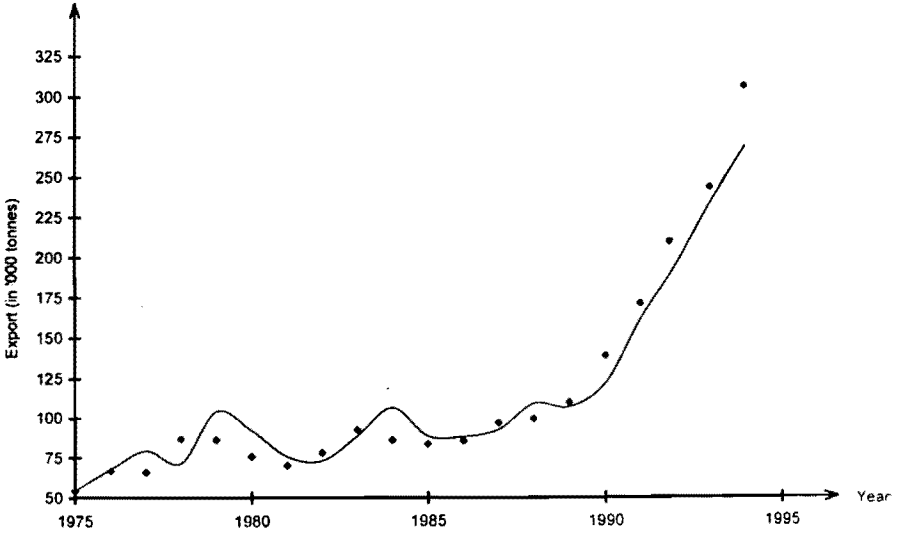

Fig. 1. Fitting of all-India marine products export data using state space model

5. *Concluding Remarks*

The importance of state space modelling using Kalman filtering technique for the dynamic linear models is highlighted. It is suggested that researchers start using this technique for modelling and forecasting time-series data. It will also be worthwhile to study Generalized Kalman Filter (GKF) and Extended Kalman Filter (EKF) involving nonlinear state space models. However, this is a very difficult task as the software for fitting such models is not readily available. Work in the direction is in progress and shall be reported in due course of time.

ACKNOWLEDGEMENT

The authors are grateful to the referee for his valuable suggestions.

REFERENCES

[1] Aoki, M., (1987). *State Space Modelling of Time Series.* Springer Verlag, New York.

[2] Box, G.E.P., Jenkins, G.M. and Reinsel, G.C., (1994). *Time Series Analysis : Forecasting and Control.* 3rd edition. Prentice Hall, New Jersey.

[3] Harvey, A.C., (1984). A unified view of statistical forecasting procedures. *J. Forecasting,* **3,** 245-275.

[4] Marine Products Export Development Authority, (1996). *Statistics of Marine Products Exports.* India.

[5] Meinhold, R.J. and Singpurwalla, N.D., (1983). Understanding Kalman filter. *Amer. Statistician, 37, 123-127.*

[6] Shumway, R.H. and Stoffer. D.S., (2000). *Time Series Analysis and its Applications.* Springer Verlag, New York.

[7] Venugopalan, R. and Prajneshu, (1996). Trend analysis in all-India marine products export using statistical modelling techniques. *Ind. J. Fish.,* **43,** 107-113.

[8] SAS, (1990). *SAS User's Guide*, Version 6.12. SAS Institute Incorporation, U.S.A.