Journal of Modern Mathematics and Statistics 7 (1): 1-7, 2013
ISSN: 1994-5388

Using Normalized BIC to Improve Box-Jenkins Model Building

E.P. Clement
Department of Mathematics and Statistics, University of Uyo, Uyo, Nigeria

Abstract: The Box-Jenkins Model building approach is used to fit a statistical time series model to the chemical viscosity reading data. The data were extracted from Box-Jenkins in called series D. The Normalized BIC was explored to compare the fitted ARIMA (1, 1, 1) Model with both the AR (1) and IMA (1, 1) Models fitted originally to the same series by Box-Jenkins in 1976. Among this class of significantly adequate set of ARIMA (p, d, q) Models of the same data set, the ARIMA (1, 1, 1) Model was found as the most suitable model with least BIC value of -2.366, MAPE of 2.424, RMSE of 0.301 and R^2 of 0.749. Estimation by Ljung-Box test with Q (18) = 9.746, 16 d.f and p-value of 0.880 showed no autocorrelation between residuals at different lag times. Finally, a forecast for a lead time (l) of 12 was made.

Key words: ARIMA Model, normalized Bayesian Information Criterion (BIC), Box-Jenkins approach, Ljung-Box statistic, time series analysis

INTRODUCTION

The Box-Jenkins approach to modeling ARIMA (p, d, q) processes is adopted in this research. The original Box-Jenkins modeling procedure involved an iterative three-stage process of model selection, parameter estimation and model diagnostic checking. Recent explanations of the process (Makridakis *et al*., 1998) often include a preliminary stage of data preparation and a final stage of model application (or forecasting).

Although, originally designed for modeling time series with ARIMA (p, d, q) processes, the underlying strategy of Box-Jenkins is applicable to a wide variety of statistical modeling situations. It provides a convenient framework which allows an analyst to think about the data and to find an appropriate statistical model which can be used to help answer relevant questions about the data.

ARIMA Models describe the current behaviour of variables in terms of linear relationships with their past values. These models are also called Box-Jenkins Models on the basis of these researchers pioneering research regarding time series forecasting techniques. An ARIMA Model can be decomposed into two parts (Box *et al*., 1994). First, it has an Integrated (I) component (d) which represents the order of differencing to be performed on the series to attain stationarity. The second component of an ARIMA consists of an ARMA Model for the series rendered stationary through differentiation. The ARMA component is further decomposed into AR and MA components (Pankratz, 1983). The Auto Regressive (AR) components capture the correlation between the current values of the time series and some of its past values. For example, AR (1) means that the current observation is correlated with its immediate past values at time t = 1. The Moving Average (MA) component represents the duration of the influence of a random (unexplained) shocks. For example, MA (1) means that a shock on the value of the series at time t is correlated with the shock at time t = 1. The autocorrelation functions (acf) and partial autocorrelation functions (pacf) are used to estimate the values of p and q.

MATERIALS AND METHODS

The Box-Jenkins methodology, researchers adopted for this research is widely regarded to be most efficient forecasting technique and is used extensively. It involves the following steps: Model identification, model estimation, model diagnostic check and forecasting (Box and Jenkins, 1976).

Model identification: The foremost step in the process of modeling is to check for the stationarity of the time series data. This is done by observing the graph of the data or autocorrelation and the partial autocorrelation functions (Makridakis *et al*., 1998). Another way of checking for stationarity is to fit the first order AR Model to the raw data and test whether the coefficients ϕ is <1. The task is to identify an appropriate sub-class of model from the general ARIMA family:

$$\phi(B)\nabla^{d}X_{t} = \theta(B)_{e_{t}} \quad (1)$$

which may be used to represent a given time series. The approach will be:

- To difference X_1 as many times as is needed to produce stationarity

$$\phi(B)_{\omega_t} = \theta(B)_{e_t} \quad (2)$$

Where:

$$\omega_t = (1-B)^d X_t = \nabla^d X_t \quad (3)$$

- To identify the resulting ARIMA process

The principal tools for putting above points effect will be the autocorrelation and the partial autocorrelation functions. Non-stationary stochastic process is indicated by the failure of the estimated autocorrelation functions to die out rapidly to achieve stationarity, a certain degree of differencing (d) is required. The degree of differencing (d), necessary to achieve stationarity is attained when the autocorrelation functions of:

$$\omega_t = \nabla^d X_t \quad (4)$$

die out fairly quickly. The autocorrelation function of an AR (p) process tails off while its partial autocorrelation function has a cut off after lag p. Conversely, the acf of a MA (q) process has a cut off after lag q while its partial autocorrelation function tails off. However, if both the acf and pacf tail off, a mixed ARMA (p, q) process is suggested. The acf of a mixed ARMA (p, q) process is a mixture of exponentials and damped sine waves after the q-p lags. Conversely, the pacf of a mixed ARMA (p, q) process is dominated by a mixture of exponentials and damped sine waves after the first p-q lags.

Model estimation: Preliminary estimates of the parameters are obtained from the values of appropriate autocorrelation of the differenced series. These can be used as starting values in the search for appropriate least square estimates. In practice not all parameter in the models are significant. The ratios:

$$\left|\frac{\text{Parameter}}{1.96 \times \text{SE}}\right| > 1 \quad (5)$$

may suggest trying a model in which some of the parameters are set to zero (Enders, 2003). Then, researchers need to re-estimate the model after each parameter is set to zero.

Diagnostic check: The diagnostic check is a procedure that is used to check residuals. The residual should fulfill the models assumption of being independent and normally distributed. If these assumptions are not fulfilled then another model is chosen for the series. Researchers will use the Ljung-Box test statistic for testing the independency of the residuals. Also, statistical inferences of the parameters and the goodness of fit of estimated statistical models will be made.

Ljung-Box statistics: Ljung and Box (1978) statistic tests whether a group of autocorrelations of a time series are <0, the test statistic is given as:

$$Q = T(T+2)\sum_{k=1}^{s}\frac{r_k^2}{T-K} \quad (6)$$

Where:

T = Number of observations
s = Length of coefficients to test autocorrelation
r_k = Autocorrelation coefficient (for lag k)

The hypothesis of Ljung-Box test are:

H_0: Residual is white noise
H_1: Residual is not white noise

If the sample value of Q exceeds the critical value of a χ^2 distribution with s degrees of freedom then at least one value of r is statistically different from zero at the specified significance level.

Normalized Bayesian Information Criterion (BIC): In statistics, the Bayesian Information Criterion (BIC) or Schwarz criterion (also SBC, SBIC) is a criterion for model selection among a finite set of models. It is based in part on the likelihood function and it is closely related to Akaike Information Criterion (AIC).

When fitting models, it is possible to increase the likehood by adding parameters but doing so may result in over fitting. The BIC resolves this problem by introducing a penalty term for the number of parameters in the model. The penalty term is large in BIC than in AIC.

The BIC was developed by Schwarz (1978) who gave a Bayesian argument for adopting it. It is closely related to the Akaike Information Criterion (AIC). In fact, Akaike was so impressed with Schwarz's Bayesian formalism that he developed his own Bayesian formalism, now often referred to as the ABIC for a Bayesian information criterion or more casually Akaike's Bayesian information criterion (Akaike, 1977).

The BIC is an asymptotic result derived under the assumptions that the data distribution is in the exponential family. Let:

x: The observed data
n: The number of data points in x, the numbers of observations or equivalently the sample size

k: The numbers of free parameters to be estimated. If the estimated model is a linear regression k is the number of regressors including the intercept

p (x/k): The probability of the observed data given the number of parameters or the likelihood of the parameters given the dataset

L: The maximized value of the likelihood functions for the estimated model. The formula for the BIC is:

$$-2.\ln p(x \mid k) \approx BIC = -2.\ln L + k\ln(n) \quad (7)$$

with the assumption that the model errors or disturbances are independent and identically distributed according to normal distribution and that the boundary condition that the derivative of the log likelihood with respect to the true variance is zero, this becomes (up to an additive constant which depends only on n and not on the model; Priestley, 1981):

$$BIC = n.\ln\left(\hat{\sigma}_e^2\right) + k.\ln(n) \quad (8)$$

where, $\hat{\sigma}_e^2$ is the error variance. The error variance in this case is defined as:

$$\hat{\sigma}_e^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \quad (9)$$

One may point out from probability theory that $\hat{\sigma}_e^2$ is a biased estimator for the true variance, σ^2. Let $\bar{\sigma}_e^2$ denote unbiased form of approximating the error variance. It is defined as:

$$\bar{\sigma}_e^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x - \bar{x})^2 \quad (10)$$

Additionally, under the assumption of normality the following version may be more tractable:

$$BIC = \chi^2 + k.\ln(n) \quad (11)$$

Note that there is a constant added that follows from transition from log-likelihood to χ^2, however in using the BIC to determine the best model, the constant becomes trivial.

Given any two estimated models, the model with the lower value of BIC is the one to be preferred. The BIC is an increasing function of σ_e^2 and an increasing function of k. That is unexplained variations in the dependent variable and the number of explanatory variables increase the value of BIC. Hence, lower BIC implies either fewer explanatory variables, better fit or both. The BIC generally penalizes free parameters more strongly than does the Akaike information criterion, though it depends on the size of n and relative magnitude of n and k.

It is important to keep in mind that the BIC can be used to compare estimated models only when the numerical values of the dependent variable are identical for all estimates being compared. The models being compared need not be nested, unlike the case when models are being compared using an F or likelihood ratio test.

Characteristic of the Bayesian information criterion:

- It is independent of the prior or the prior is vague (a constant)
- It can measure the efficiency of the parameterized model in terms of predicting the data
- It penalizes the complexity of the model where complexity refers to the number of parameters in models
- It is approximately equal to the minimum description length criterion but with negative sign
- It can be used to choose the number of clusters according to the intrinsic complexity present in a particular dataset
- It is closely related to other penalized likelihood criteria such as RIC and the Akaike Information Criterion (AIC)

Implications of the Bayesian information criterion: BIC has been widely used for model identification in time series and linear regression. It can, however be applied quite widely to any set of maximum likelihood-based models. However in many applications (for example, selecting a black body or power law spectrum for an astronomical source), BIC simply reduces to maximum likelihood selection because the number of parameters is equal for the models of interest.

RESULTS

Having discussed some basic concepts and theoretical foundation of time series that will enable us analyze the data. Researchers now present a step by step analysis of the dataset of series D.

Model identification: The graphical plot of the original series of the chemical process viscosity reading: Every hour is given in Fig. 1. It is observed that the series exhibits non-stationary behaviour indicated by its growth.

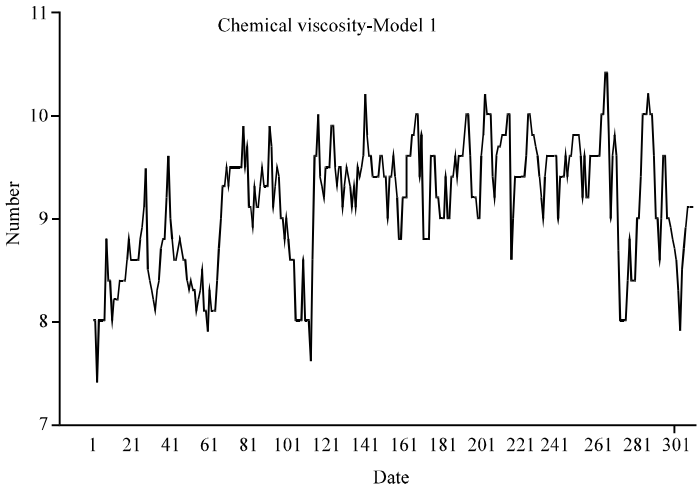

Fig. 1: Graph of original series X_t (observed)

The sample autocorrelations of the original series in Fig. 2 failed to die out quickly at high lags, confirming the non-stationarity behaviour of the series which equally suggests that transformation is required to attain stationarity. Consequently, the difference method of transformation was adopted and the first difference (d = 1) of the series was made. The plot of the stationary equivalent is given in Fig. 3 while the plots of the autocorrelation and partial autocorrelation functions of the differenced series are given in Fig. 4 and 5, respectively.

The autocorrelation and partial autocorrelation functions of the differenced series indicated no need for further differencing as they tend to be tailing off rapidly. They also indicated no sign of seasonality since they do not repeat themselves at lags that are multiples of the number of periods per season.

Using Fig. 4 and 5, the differenced series will be denoted by $\{\omega_t\}$ for t = 1, 2, ..., 309 where, $\omega_t = \nabla X_t$. It is observed that both the autocorrelation and partial autocorrelation functions of ω_t are characterized by correlations that alternate in sign and which tend to damp out with increasing lag. Consequently, a mix autoregressive moving average of order (1,1,1) was proposed since both the autocorrelation and partial autocorrelation functions of the w_t seem to be tailing off. Thus using Eq. 1, the proposed model is an ARIMA (1, 1, 1):

$$\phi(B)\nabla X_t = \theta(B)e_t \tag{12}$$

$$\left(1-\phi_1^B\right)\omega_t = \left(1-\theta_1^B\right)e_t \tag{13}$$

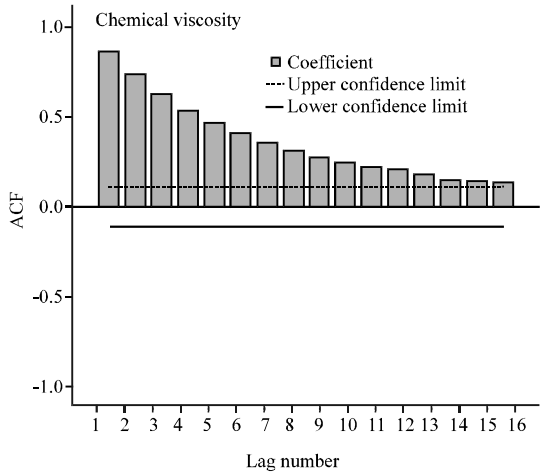

Fig. 2: Plot of autocorrelation functions of the original series

$$\left(1-\phi_1^B\right)\left(X_t - X_{t-1}\right) = \left(1-\theta_1^B\right)e_t \tag{14}$$

The plot of the autocorrelation and partial autocorrelation functions of the residuals from the tentatively identified ARIMA (1, 1, 1) Model are given in Fig. 6.

Estimation of parameters: Having tentatively identified what appears to be a suitable model, the next step is to obtain the least squares estimates of the parameters of the model. The SPSS 17 Expert Modeler was used to fit the model to the data. The coefficient of both the AR and the MA were not significantly different from zero with values

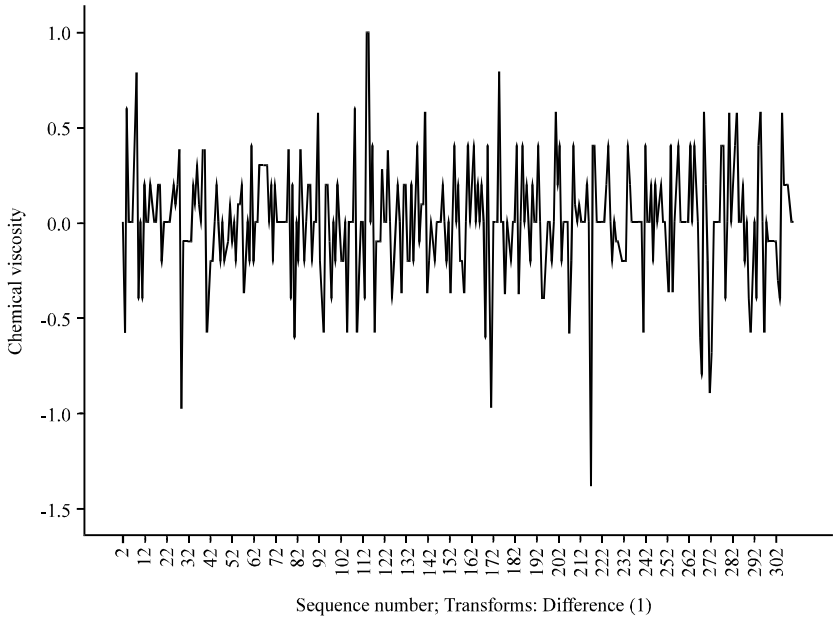

Fig. 3: Graph of the differenced series D (-1)

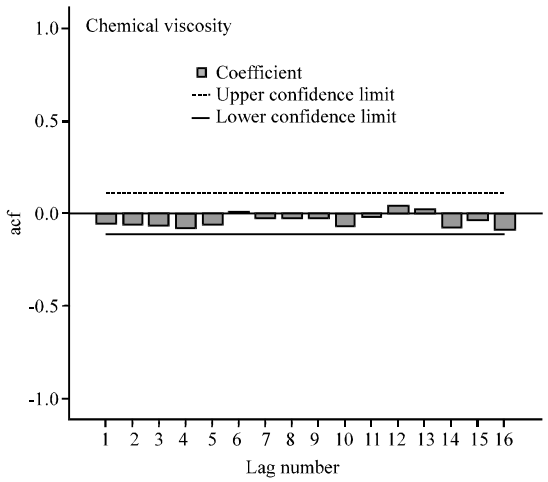

Fig. 4: Plot of the autocorrelation functions of the differenced series

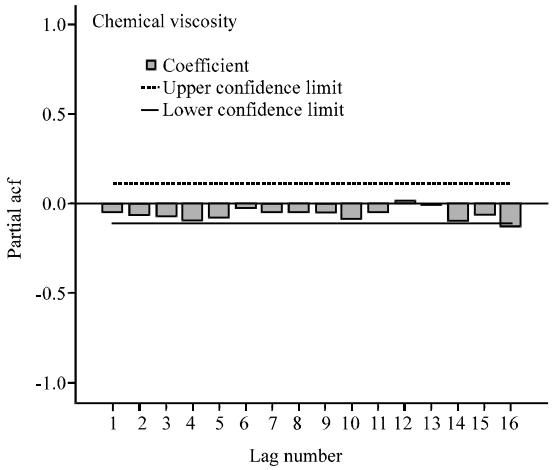

Fig. 5: Plot of the partial autocorrelation functions of the differenced series

of 0.814 and 0.972, respectively. This model enables us to write the model equation as:

$$X_t = 0.814X_{t-1} + 0.972e_{t-1} + e_t \quad (15)$$

That is the AR coefficient ϕ_1 was estimated to be 0.814 with standard error of 0.045 and a t-ratio of 18.024 while the MA coefficient θ_1 was estimated to be 0.972 with standard error of 0.020 and a t-ratio of 49.007.

For this model Q = 9.746. The 10 and 5% points of χ^2 with 16 degree of freedom are 23.50 and 26.30, respectively. Therefore, since Q is not unduly large and the evidence does not contradict the hypothesis of white noise behaviour in the residuals, the model is very adequate and significantly appropriate.

Model diagnostic check: It is concerned with testing the goodness of fit of the model. From plots of the residual acf and pacf, it can be seen that all points are randomly distributed and it can be concluded that there is an irregular pattern which means that the model is adequate. Also, the individual residual autocorrelations are very small and are generally within significance bounds.

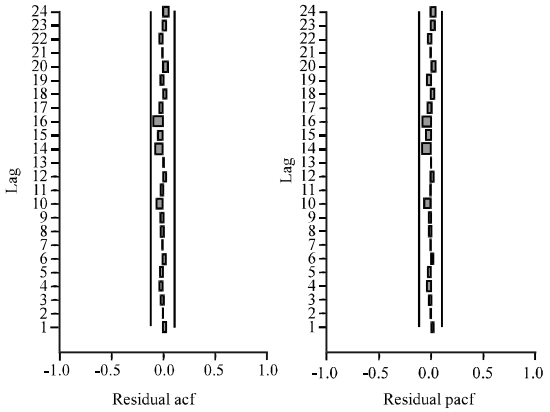

Fig. 6: Autocorrelation and partial autocorrelation functions of the residuals

Also, the statistical $\pm 2\sqrt{n}$ significance of the model was checked. Five criteria: The normalized Bayesian Information Criterion (BIC), the R^2, Root Mean Square Error (RMSE), the Mean Absolute Percentage Error (MAPE) and the Ljung-Box Q statistic were used to test for the adequacy and statistical appropriateness of the model.

First, the Ljung-Box (Q) Statistic test was performed using SPSS 17 Expert Modeler (Table 1 and 2), the Ljung-Box statistic of the model is not significantly different from zero with a value of 9.746 for 16 d.f. and associated p-value of 0.880, thus failing to reject the null hypothesis of white noise. This indicates that the model has adequately captured the correlation in the time series. Moreover, the low value of RMSE indicates a good fit for the model.

Also, the high value of the R^2 and MAPE indicate a perfect prediction over the mean.

Again, the model is adequate in the sense that the plots of the residual acf and pacf in Fig. 6 show a random variation thus from the origin zero (0), the points below and above are all uneven hence the model fitted is adequate.

The adequacy and significant appropriateness of the model was confirmed by exploring the normalized Bayesian Information Criterion (BIC). In a class of statistically significant ARIMA (p,d,q) Models fitted to the series, the ARIMA (1, 1, 1) Model had the least BIC value of -2.366.

Forecasting with the model: Forecasting based on the fitted model was computed up to lead time of 12 and the one-step forecasting and the 95% confidence limits are displayed in Table 3.

Table 1: Model parameters

Coefficients	Estimates	SE	t-radio	Sig.
AR Lag1	0.814	0.045	18.024	0.000
Difference	1	-	-	-
MA Lag1	0.972	0.020	49.007	0.000

Table 2: Model statistics

Model fit statistics				Ljung-box Q (18)			
R^2	RMSE	MAPE	BIC	Statistics	DF	Sig.	No. of outliers
0.749	0.301	2.424	-2.366	9.746	16	0.880	0

Table 3: One-step forecast of the ARIMA (1, 1, 1) Model

Lead time	Forecast	95% lower limit	95% upper limit
1	9.12	9.00	10.02
2	8.65	8.02	9.63
3	10.14	9.86	10.84
4	12.02	10.63	12.46
5	9.38	8.86	10.41
6	9.02	8.92	9.65
7	8.86	8.02	9.85
8	7.80	7.04	8.06
9	10.81	9.84	11.00
10	9.16	9.03	10.06
11	8.28	7.89	9.40
12	10.02	9.88	10.64

DISCUSSION

The sample acf and pacf of the original series (series D) were computed using the SPSS 17 Expert Modeler and their graphs were plotted. These were used in identifying the appropriate model. The series exhibited non-stationary behaviour following the inability of the sample acf of the series to die the rapidly even at high lags. The series was transformed by differencing once and stationarity was attained. The plot of the differenced series indicated that the series is evenly distributed around the mean.

Following the distribution of the acf and pacf of the differenced series an ARIMA (1, 1, 1) Model given by $X_t = 0.814X_{t-1} + 0.972e_{t-1} + e_t$ was identified. The parameters of the fitted model were estimated. The model was then subjected to statistical diagnostic check using the Ljung-Box test statistic and the normalized Bayesian Information Criterion (BIC). Analysis proved that the model is statistically significant, appropriate and adequate.

The fitted model was used to forecast values of the chemical viscosity readings for a lead time (l) of 12. The forecast is a good representation of the original data which neither decreases nor increases.

The fitted Model (ARIMA (1, 1, 1)) was compared with the two original models fitted to the same series by Box and Jenkins (1976). That is AR (1) Model given by:

$$z_t = 0.87z_{t-1} + a_t \quad (16)$$

and IMA (1, 1) Model given by:

$$\nabla z_t = -0.06\, a_{t-1} + a_t \quad (17)$$

were fitted to the series D data. The normalized Bayesian information criterion was used in comparing these three models. That is AR (1):

$$z_t = 0.87 z_{t-1} + a_t \quad (18)$$

IMA (1, 1):

$$\nabla z_t = -0.06 a_{t-1} + a_t \quad (19)$$

and ARIMA (1, 1, 1):

$$X_t = 0.814 X_{t-1} + 0.972 e_{t-1} + e_t \quad (20)$$

Analysis showed that the ARIMA (1, 1, 1) Model is superior to the two other models having the least BIC value.

CONCLUSION

The ARIMA (1, 1, 1) Model fitted to the chemical viscosity data is a better model than both the AR (1) and IMA (1, 1) Models fitted originally to the same series by Box-Jenkins in 1976. This showed that the Box-Jenkins Model Building Approach needs modification which this research has presented.

REFERENCES

Akaike, H., 1977. On Entropy Maximization Principle. In: Application of Statistics, Krishnaiah, P.R. (Ed.). North-Holland, Amsterdam, The Netherlands, pp: 27-41.

Box, G.E.P. and G.M. Jenkins, 1976. Time Series Analysis: Forecasting and Control. Holden-Day, San Francisco, USA.

Box, G.E.P., G.M. Jenkins and G.C. Reinsel, 1994. Time Series Analysis: Forecasting and Control. 3rd Edn., Prentice Hall, Delhi, India, ISBN-13: 9780130607744, Pages: 598.

Enders, W., 2003. Applied Econometric Time Series. 2nd Edn., John Wiley and Sons, New York, USA., ISBN-13: 9780471230656, Pages: 480.

Ljung, G.M. and G.E.P. Box, 1978. On a measure of lack of fit in time series models. Biometrika, 65: 297-303.

Makridakis, S., S.C. Wheelwright and R.J. Hyndman, 1998. Forecasting Methods and Applications. 3rd Edn., John Wiley and Sons Inc., New York.

Pankratz, A., 1983. Forecasting with Univariate Box-Jenkins Models-Concepts and Cases. John Wiley, New York.

Priestley, M.B., 1981. Spectral Analysis and Time Series Analysis. Vols. 1 and 2 Academic Press, London.

Schwarz, G., 1978. Estimating the dimension of a model. Ann. Stat., 6: 461-464.