Malaysian Journal of Science **25** (1): 73 – 78 (2006)

Water Level Data Modeling with Bilinear Time Series Analysis

Mohd. Sahar Yahya[1], Ibrahim Mohamed[2,*], Azami Zaharim[3], Mohammad Said Zainol[4]

[1] Centre for Foundation Studies in Science. University of Malaya, 50603 Kuala Lumpur.
mohdsahar@um.edu.my
[2,*] Institute of Mathematical Sciences. University of Malaya, 50603 Kuala Lumpur.
imohamed@um.edu.my (corresponding author)
[3] Department of Architecture, Faculty of Engineering. National University of Malaysia. 43600 UKM Bangi, Selangor
azami@vlsi.eng.ukm.my
[4] Faculty of Information Technology and Quantitative Sciences. MARA University of Technology, 47000 Shah Alam, Selangor
mankidal@streamyx.com
Received 10th May 2005, accepted in revised form 17th April 2006.

ABSTRACT In the literature, many time series data, such as the economic and hydrological data, show various nonlinearity characteristics. The Keenan's test and F-test are employed in identifying a nonlinear data set. This article looks at the modeling of nonlinear time series data using bilinear time series model. The model is an extension of autoregressive model such that an extra term representing the bilinear characteristic is introduced. The estimation of bilinear models is obtained using nonlinear least squares method. As an illustration, analysis on water level of Sungai Kelantan using the above method is presented.

ABSTRAK Di dalam literatur, wujud data siri masa, seperti data ekonomi dan data hidrologi, yang menunjukkan pelbagai ciri-ciri tak linear. Ujian Keenan dan Ujian-F digunakan untuk mengenalpasti set data tak linear. Artikel ini melihat kepada pemodelan data siri masa tak linear menggunakan model siri masa Bilinear. Model ini merupakan model yang lebih am daripada model autoregresi di mana sebutan tambahan bagi mewakili ciri bilinear diperkenalkan. Penganggaran model bilinear diperolehi dengan menggunakan kaedah kuasa dua terkecil tak linear. Sebagai ilustrasi, analisis ke atas data aras sungai Kelantan dengan menggunakan kaedah di atas dibentangkan.

(Bilinear, nonlinear least squares method, hydrology)

INTRODUCTION

Water level has been used as an indicator to the occurrence of flooding. The univariate Box-Jenkins approach based on ARIMA modeling has been used in many applications. A good account of the approaches is available in, *inter alia*, Box and Jenkins [1], Fuller [2] and Chatfield [3]. However, there are time series data, including water level data, which are not suitable to be fitted by linear models. Such examples are the Castle River flow data in Alberta, Canada (Oyet [4]) and the average monthly flows of the Fraser River in British Columbia (Lewis and Ray [5]). These data should be fitted better by nonlinear models such as bilinear models. In this article, a comparative study between ARIMA and bilinear modeling is discussed.

DATA COLLECTION

Records of daily water level data at the Sungai Kelantan were obtained from the Department of Irrigation and Drainage of Selangor from 20 April 2001 till 22 September 2001. Figure 1 gives the plot of the data. The dashed line represents the mean of the data which equals to 9.003. The length of the data is 156. The east coast of Peninsular Malaysia is known to have a north-west monsoon during this period and has dry months in June and July. This is reflected in Figure 1 whereby the water levels are below the mean from middle of June till end of July.

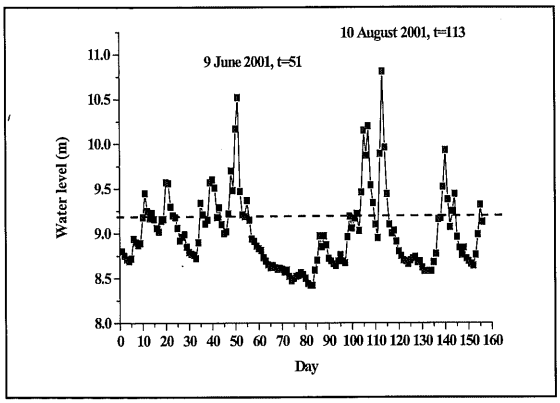

Figure 1. Plot of daily mean water level data

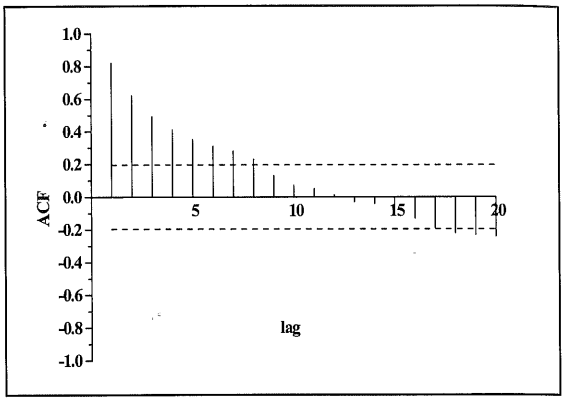

Figure 2. The ACF plot of the water level data

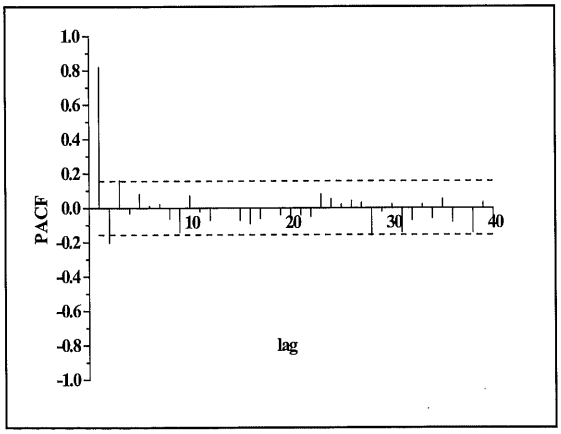

Figure 3. The PACF plot of the water level data

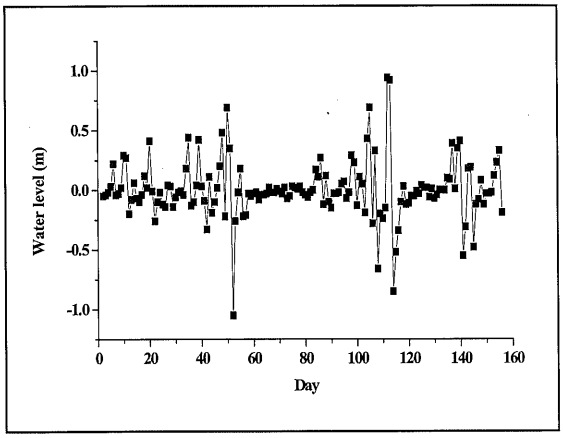

Figure 4. Plot of first differenced data

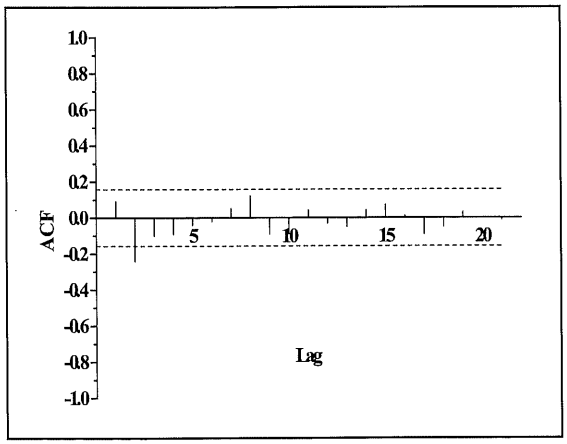

Figure 5. The ACF plot of first differenced

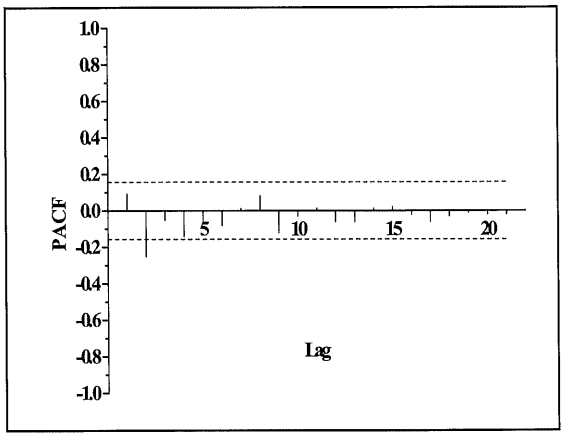

Figure 6. The PACF plot of first differenced data

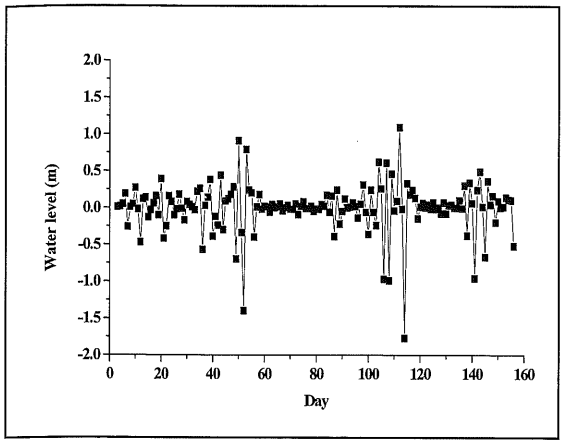

Figure 7. Plot of second differenced data

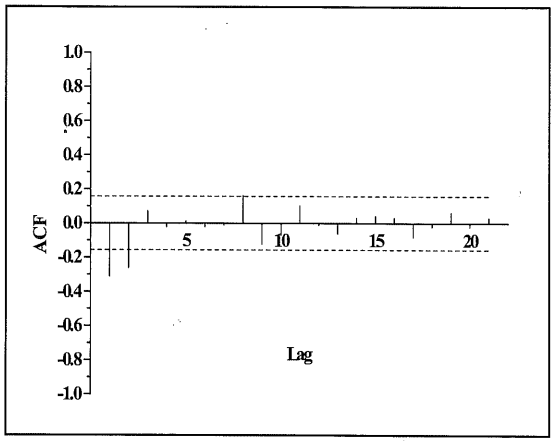

Figure 8. The ACF plot of second differenced data

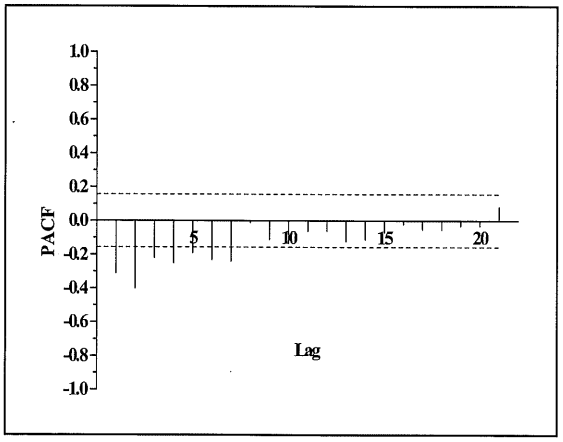

Figure 9. The PACF plot of second differenced data

ARIMA MODELING OF THE WATER LEVEL

The ACF and PACF plots of the original data are given by Figure 2 and Figure 3 respectively. It can be seen that the autocorrelations values die out very slowly suggesting that the original data is non-stationary. Hence, differencing of data is needed.

The plot of first differenced data and its ACF and PACF plots are given by Figures 4-6 respectively. However, it is difficult to identify a possible model for the data based on the ACF and PACF plots. Both plots have non-significant values of the first lag but significant values of the second lag. The plot of second differenced data is given in Figure 7. The ACF and PACF plots of the data are given in Figures 8 - 9. The magnitude of the PACF values is less than 4 and the PACF values die out after 7 lags. There are two large ACF values for the first two lags. The possible model for this data is ARIMA(0,2,2). The parameter estimates are 0.826 and 0.173 with the t-ratio values are 10.41 and 2.18 respectively. At 5% significant level, the critical value is 1.65 obtained from the *t*-distribution with 154 degree of freedom. Thus, both parameters should be included in the model. The ACF and PACF plots of the residuals are given in Figure 10 and Figure 11 respectively. Only the ACF and PACF values at lag 2 are significant. The residuals can be said to follow white noise process. The Ljung-Box statistics at lags 12 and 24 are 16.286 and 21.135 respectively whereas the 5% critical values based on the chi-square distribution with 10 and 22 degrees of freedom are 18.307 and 33.924 respectively. Hence, The Ljung-Box statistics do not suggest any inadequacy of the model.

To check on the normality of the residuals, the correlation test between the residuals and the normal scores is carried out. The correlation value is 0.928 which is below the corresponding 5% critical value of 0.987. This suggests that the normality is not totally satisfied. From the histogram of the standardized residuals given in Figure 12, there are quite a number of positive high residuals which are due to the high spikes in the original data.

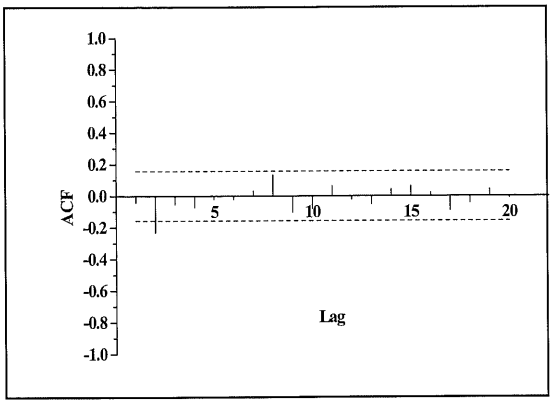

Figure 10. The ACF plot of residuals

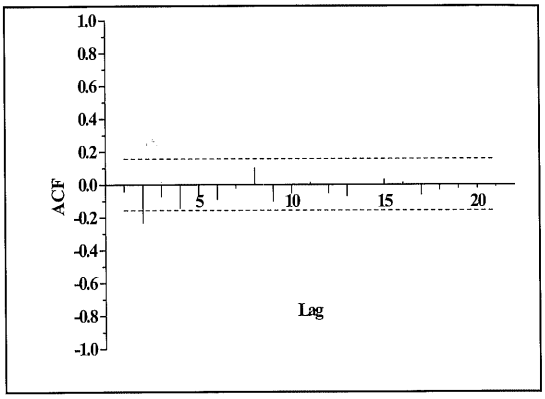

Figure 11. The PACF plot of residuals

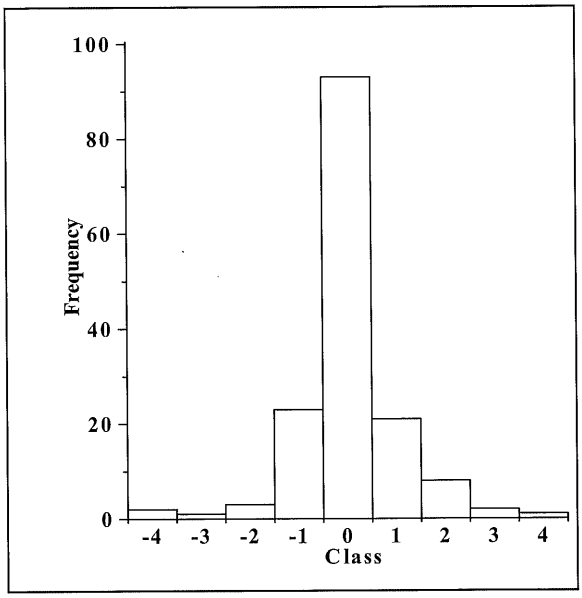

Figure 12. The histogram of the standardized residuals

Other linear models are also fitted in order, if possible, to improve the modeling. They will be compared based on three types of order selection criteria. They are the Akaike's information criteria denoted by AIC (see Akaike [6]), Akaike's Bayesian information criteria denoted by BIC (see Akaike [7]) and Schwarz's criteria denoted by SBIC (see Schwarz [8]). The possible models that improve the modeling are listed in Table 1. It can be seen that ARIMA(0,2,3) reduces the AIC, BIC and SBIC values the most. The coefficients values are 0.9354, 0.3549 and -0.2911 which are all significant at 5% significance level. The Ljung-Box statistics do not suggest any inadequacy of the model. The correlation between the residuals and normal scores is 0.943. Although this correlation value is still not significant, it is better than that for ARIMA(0,2,2)

Table 1. Summary of results of selected linear models

MODEL	AIC	BIC	SBIC	VAR
ARIMA(0,2,2)	13.51	-421.05	-423.05	0.06225
ARIMA(0,2,3)	5.75	-424.81	-427.81	0.05846
ARIMA(1,2,2)	12.61	-417.95	-420.95	0.06108

NONLINEARITY TEST

The Keenan's test (see Keenan [9]) and the F-test (see Tsay [10]) are used to investigate whether the water level data belongs to a nonlinear model. Both tests suggest that the data is nonlinear with p-values 0.00001 and 0.00003 respectively. This should be true as the data contains several spikes which will not be explained fully by any linear model.

BILINEAR MODEL

The general bilinear model, denoted by BL(p,q,r,s), is given by

$$Y_t = \sum_{i=1}^{p} a_i Y_{t-1} + \sum_{j=1}^{q} c_j e_{t-j} + \sum_{k=1}^{r}\sum_{\lambda=1}^{s} b_{k\lambda} Y_{t-k} e_{t-\lambda} + e_t \qquad (1)$$

where a_i, c_j and $b_{k\lambda}$ are constant, and e_t's are assumed to follow normal distribution with mean zero and precision τ, $\tau > 0$. The first two components on the right-hand side of (1) are basically the ARMA model with parameters p and q. The second last component is nonlinear which helps to explain the nonlinearity characteristic of the data being modeled. Thus, ARMA (p,q) is a special case of the BL(p,q,r,s) when $r = s = 0$. In this article, the parameters of bilinear models are estimated using the nonlinear least squares method as suggested by Priestly [11].

Several bilinear models are fitted on the data. The diagnostic results based on the AIC, BIC and SBIC together with their respective residual variances are given in Table 2. It is clear that either BL(2,0,1,1) or BL(1,1,1,1) have lower values of the order selection criteria compared to the other two models. The parameter estimates of the fitted BL(1,1,1,1) model are $a_1 = 0.7692$, $c_1 = 0.4882$ and $b_{11} = -0.3942$. The correlation between residuals and normal scores is 0.9124. The correlation value is lower than that of ARIMA(0,2,2) or ARIMA(0,2,3) models. Again, the existence of few outliers might affect the results of normality test. The histogram of the standardized residuals is given in Figure 7. As for BL(2,0,1,1) models, the parameter estimates are $a_1 = 1.2803$, $a_2 = -0.3928$ and $b_{11} = -0.4802$. The correlation between residuals and normal scores is 0.909 which is lower than that for BL(1,1,1,1) model.

Table 2. Summary of results of selected bilinear models

MODEL	AIC	BIC	SBIC	VARIANCE
BL(1,0,1,1)	-3.279	-437.888	-439.888	0.0559
BL(2,0,1,1)	-17.112	-447.671	-450.671	0.0505
BL(3,0,1,1)	-16.963	-443.473	-447.473	0.0499
BL(1,1,1,1)	-17.938	-448.498	-451.498	0.0502

Table 3. Summary of diagnostic results

MODEL	AIC	BIC	SBIC	VARIANCE
BL(1,1,1,1)	-17.94	-448.50	-451.50	0.05022
ARIMA(0,2,3)	5.75	-424.81	-427.81	0.05846

MODEL COMPARISON

Table 3 gives the summary of diagnostic results based on BL(1,1,1,1) and ARIMA(0,2,3) models. It can be seen that, in general, bilinear models improves the modeling if compared to the fitted linear models. For variance of the residuals, $\hat{\sigma}_e^2$, the reduction by 14.1% is observed. The values of AIC, BIC and SBIC are also reduced. Hence, we can conclude that bilinear modeling improves the modeling compared to the best linear model.

CONCLUSION

The application of bilinear modeling has been illustrated by the Sungai Kelantan water level data. Results from the nonlinearity test confirm that the data is nonlinear. The results further show that bilinear model fits the data better if compared to best fitted linear models.

Acknowledgement The original data is obtained from the Department of Irrigation and Drainage of Selangor.

REFERENCES

1. Box, G.E.P. and Jenkins, G.M. (1976). *Time Series Analysis Forecasting and Control.* Holden-Day, San Francisco.
2. Fuller, W.A. (1976). *Introduction to Statistical Time Series*. Wiley, New York.
3. Chatfield, C. (1996). *The Analysis of Time Series: An Introduction*. Chapman and Hall, London.
4. Oyet, A.J. (2001). Nonlinear time series modeling : Order identification and wavelet filtering. *Interstat*, April 2001.
5. Lewis, P.A.W. and Ray, B.K. (2002). Nonlinear modeling of periodic threshold autoregressions using TSMARS. *Journal of Time Series Analysis* **23** (4): 272 - 285.
6. Akaike, H. (1969). Fitting autoregressive model for prediction. *Annals Institute of Statistical Mathematics* **21**: 203 - 217.
7. Akaike, H. (1979). A Bayesian extension of the minimum AIC procedure of autoregressive modeling. *Biometrika* **66**: 237 - 242.
8. Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* **6**: 461 - 464.
9. Keenan, D. M. (1985). A Tukey non-additivity type test for time series nonlinearity. *Biometrika* **72** (1): 39 - 44.
10. Tsay, R.S. (1986). Nonlinearity test for time series. *Biometrika* **73** (2): 461 - 466.
11. Priestly, M.B. (1991). *Non-linear and Non-stationary Time Series Analysis*. Academic Press, San Diego.