Sains Malaysiana 40(2)(2011): 191–196

Improvement on the Innovational Outlier Detection Procedure in a Bilinear Model

(Pembaikan Prosedur Pengesanan Nilai Sisihan Inovasi dalam Model Bilinear)

I.B. MOHAMED*, M.I. ISMAIL, M.S. YAHYA, A.G. HUSSIN,
N. MOHAMED, A. ZAHARIM & M.S. ZAINOL

ABSTRACT

This paper considers the problem of outlier detection in bilinear time series data with special focus on BL(1,0,1,1) and BL(1,1,1,1) models. In the previous study, the formulations of effect of innovational outlier on the observations and residuals from the process had been developed and the corresponding least squares estimator of outlier effect had been derived. Consequently, an outlier detection procedure employing bootstrap-based procedure to estimate the variance of the estimator had been proposed. In this paper, we proposed to use the mean absolute deviance and trimmed mean formula to estimate the variance to improve the performances of the procedure. Via simulation, we showed that the procedure based on the trimmed mean formula has successfully improved the performance of the procedure.

Keywords: Bootstrap; bilinear; innovational outlier; least squares method

ABSTRAK

Kertas kerja ini mempertimbangkan masalah pengesanan nilai terpencil dalam data siri masa bilinear dengan fokus khas kepada model BL(1,0,1,1) dan BL(1,1,1,1). Dalam kajian terdahulu, formulasi kesan nilai terpencil inovasi ke atas cerapan dan ralat daripada proses di atas telah dibina dan penganggar kuasa dua terkecil kesan outlier telah diterbitkan. Justeru, prosedur pengesanan nilai terpencil menggunakan prosedur bootstrap untuk menganggar varians penganggar telah dicadangkan. Dalam kertas kerja ini, kami mencadangkan untuk menggunakan "min sisihan mutlak" dan formula "min terkemas" bagi menganggar varians untuk memperbaiki keupayaan prosedur. Melalui simulasi, kami menunjukkan bahawa prosedur berdasarkan formula "min terkemas" telah berjaya memperbaiki keupayaan prosedur.

Kata kunci: Bilinear; bootstrap; kaedah kuasa dua terkecil; nilai tersisih inovasi

INTRODUCTION

The existence of observations that deviate markedly from the rest of the observations occurs frequently in time series data. These observations are usually called outliers. In certain cases, visual inspection of data may be used to deal with outliers. However, it is preferable that a specific procedure could be developed based on, for example, a hypothesis testing approach. In the literature, extensive studies have been conducted on the occurrence of the additive outlier (AO) and the innovational outlier (IO) in linear time series models, for example, Fox (1972), Tsay (1986), Chang et al. (1988) and Chen et al. (1993).

On the other hand, few studies can be found on the detection of outliers in bilinear models. Chen (1997) used Gibbs sampling method to detect AO in a general bilinear model while Zaharim et al. *(*2006) used the least squares method to detect both AO and IO in the two simplest order of bilinear models. Zaharim et al. *(*2006) had used the bootstrapping method to estimate the variance of the measures via the standard variance formula. Later, Ismail et al. (2008) had proposed improved versions of the AO detection procedure by utilizing the Mean Absolute Deviance (MAD) formula and by trimming the bootstrap samples considered in the calculation of variance. In this paper, we follow the suggested improvised approach for the IO case. The performance of the three procedures are then compared.

This paper is organized as follows: First, we introduce the bilinear model in the second section followed by the description of the outlier detection procedure for IO case. Then we discuss the improvement made on the outlier detection procedure. A simulation study is then carried out to study the performance of the procedure as given in the fifth section. As an illustration, we apply the procedure on local rainfall in the last section.

BILINEAR MODEL

A real in-depth statistical study on bilinear models was started after Granger and Anderson (1978) published a manuscript on the model. The general bilinear model, denoted by BL(p,q,r,s) where p,q,r,s are positive integers or zero, is given by

$$Y_t = \mu + \sum_{i=1}^{p} a_i Y_{t-i} - \sum_{j=1}^{q} c_j e_{t-j} + \sum_{k=1}^{r} \sum_{l=1}^{s} b_{kl} Y_{t-k} e_{e-l} + e_t, \tag{1}$$

where a_i, c_j and b_{kl} are constants, and e_t's are assumed to follow a normal distribution with mean zero and precision τ, $\tau > 0$. The model is a simplified case of nonlinear Volterra series expansions and an extension of general linear autoregressive moving average model of orders p and q.

Various methods of estimating the parameters of bilinear models are available. In this paper, the nonlinear least squares estimation method proposed by Priestley (1991) is used. The method is recursive in nature and the estimates are obtained when the convergence property is satisfied.

THE OUTLIER DETECTION PROCEDURE

The procedure of detecting AO and IO had been proposed by Zaharim et al. (2006) for BL(1,1,1,1) models. The BL(1,1,1,1) model is given by

$$Y_t = a_1 Y_{t-1} - c_1 e_{t-1} + b_{11} Y_{t-1} e_{t-1} + e_t. \tag{2}$$

The results hold for BL(1,0,1,1) model by taking $c_1 = 0$. For simplicity, we dropped the subscripts from the constants.

Let Y_t^* be the observed values from the BL(1,1,1,1) process with an IO occurring at time point $t = d$ with magnitudes ω and e_t^* being the resulting residuals when BL(1,1,1,1) is fitted on the contaminated data, $t = 1,2,\ldots,n$. Further, let Y_t and e_t be the observations and residuals that would have been obtained if there were no outliers in the data and they will be referred herewith as 'original observation' and 'original residual', respectively.

The procedure for detecting outliers is described here. The procedure is meant to detect IO in data generated from the BL(1,1,1,1) model. When IO occurs at time $t < d$, then $Y_t^* = Y_t$. On the other hand, for $t \geq d$, Zaharim et al. (2006) had shown that the formulation of IO effects on observations is given by

$$Y_{d+k}^* = Y_{d+k} + \omega \prod_{i=0}^{k-1} \left(a + b e_{d+i}\right), \tag{3}$$

for $k \geq d$ and $\prod_{i=0}^{-1}\left(a + b e_{d+i}\right)$ is defined to be unity. It is expected that IO will not only change the observation at $t = d$ but also several subsequent observations as illustrated in Figure 1.

Consequently, the residuals will also be affected and will differ from the outlier-free data set. When IO occurs at time $t < d$, then $e_{t,IO}^* = e_t$. On the other hand, for $t \geq d$ and $k \geq 0$, Zaharim et al. (2006) had shown that the formulation of IO effects on observations is as follows:

$$e_{d+k}^* = e_{d+k} + (-1)^k f_{d+k}(\omega) \tag{4}$$

where

$$f_{d+k} = \left\{ \left(bY_{d+(k-1)}^* + c\right)^{\omega} f_{d+(k-1)}(\omega) \right. \qquad \begin{array}{l} k = 0 \\ k = 1,2,\ldots \end{array}$$

It is expected that IO will not only change the residual at $t = d$, but it will also change several subsequent residuals as illustrated in Figure 2.

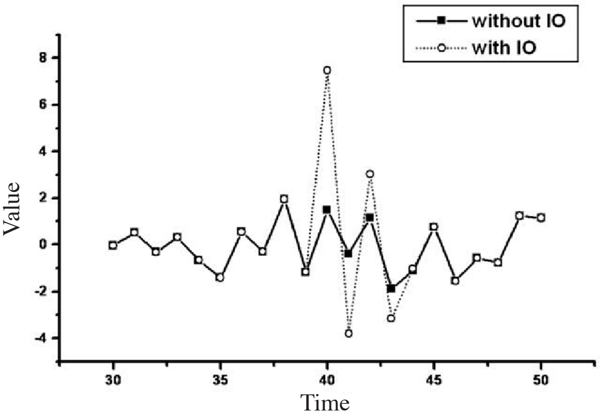


FIGURE 1. The effect of IO on observations

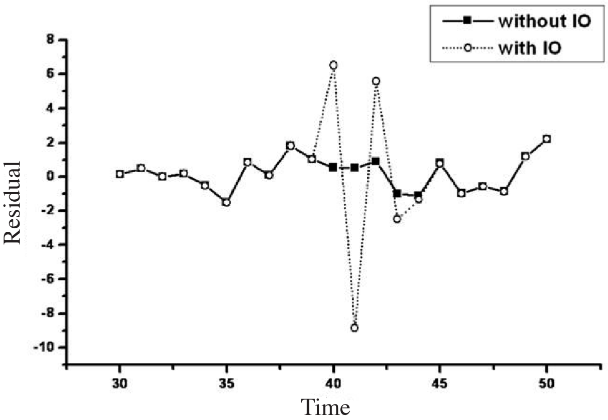


FIGURE 2. The effect of IO on residuals

The statistics to measure the magnitude of outlier effects for IO can be obtained using the least squares method. Consider the following equation:

$$S = \sum_{t=1}^{n} e_t^2 = \sum_{t=1}^{d-1} e_t^2 + \sum_{l=0}^{n-d} \left(e_{d+l}^* - \{-1\}^l f_{d+l}(\omega)\right). \tag{5}$$

Equation (5) is then minimized with respect to ω, yielding the following least squares measure of the IO effect:

$$\hat{\omega} = \frac{\sum_{k=0}^{n-d} \left[\{-1\}^k e_{d+k}^* A_k\right]}{\sum_{k=0}^{n-d} A_k^2} \tag{6}$$

where

$$A_k = \begin{cases} l & k = 0 \\ \left(bY_d^* + c\right) & k = 1 \\ \left(bY_{d+(k-1)}^* + c\right) A_{k-1} & k = 2,3,\ldots \end{cases}.$$

Zaharim et al. (2006) further used the bootstrap method to obtain the estimates of $\mathrm{Var}(\hat{\omega})$. It is carried out through the process of drawing random samples with replacement from the residuals as described below:

(a) Let (e_1, e_2, …, e_n) be the original residuals. Sampling with replacement is carried out from the original residuals giving a bootstrap sample of size n, say, $e^{*(1)} = (e_1^*, e_2^*, \ldots, e_n^*)$. This is repeated for a large number of times, say B times, giving B sets of bootstrap samples $e^{*(1)}, e^{*(2)}, \ldots, e^{*(B)}$.

(b) For each bootstrap sample $e^{*(M)}$, M = 1, 2, ..., B, we calculate $\tilde{\omega}_M$.

(c) The sample standard deviation of $\tilde{\omega}$ is given by

$$\tilde{\sigma}_{BS} = \left\{ \frac{\sum_{M=1}^{B} \left(\tilde{\omega}_M - \bar{\tilde{\omega}}_{BS}\right)^2}{(B-1)} \right\}^{1/2}, \tag{7}$$

where

$$\bar{\tilde{\omega}}_{BS} = B^{-1} \sum_{M=1}^{B} \tilde{\omega}_M .$$

Let H_0 denotes the hypothesis that $\omega = 0$ in the bilinear model considered and H_1 denotes the situations $\omega \neq 0$ in the bilinear model with IO at time t. The following test statistics can be used to test the hypothesis:

$$\hat{\tau}_t = \frac{\left(\hat{\omega}_t - \bar{\tilde{\omega}}_{BS,t}\right)}{\tilde{\sigma}_{BS,t}}, \qquad t = 1, \ldots, n. \tag{8}$$

The following procedure can now be used to detect the occurrence of IO at time t:

1) Compute the least squares estimates of the bilinear model based on the given data. Hence, obtain the residuals.
2) Compute $\hat{\tau}_t$ for t = 1, 2, ..., n using the residuals as obtained in part (a).
3) Let $\eta_t = \max_{t=1,2,\ldots,n} \left\{ |\hat{\tau}_t| \right\}$. Given a pre-determined critical value C, if $\eta_t > C$, then there is a possibility of an IO occurring at time t.

Through the suggested procedure, the occurrence of IO can be detected at any time t.

AN IMPROVED VERSION OF THE OUTLIER DETECTION PROCEDURE

In this paper, we attempted to improve the procedure of detecting IO as presented in the previous section.

1) The Mean Absolute Deviance (MAD) procedure

Instead of using equation (7) to calculate the standard deviation of $\hat{\omega}$, we utilize the procedure suggested by Hampel et al. (1986) in which the standard deviation is computed using the following relationship

$$\hat{\sigma}_{MAD} = 1.483 \times \text{median} \left\{ \left| \hat{\omega}_t - \tilde{\omega}_{MED} \right| \right\},$$

where $\tilde{\omega}_{MED}$ is the median of the bootstrap estimates, $\tilde{\omega}_M$, M = 1, 2, ..., B and B is the number of bootstrap resamples drawn.

(2) The 5% Trimmed Mean (TM) procedure

Here, the calculation of the standard deviation uses the trimmed sample such that the smallest and largest 5% of $\tilde{\omega}_M$ are removed from the calculation. Equation (7) is then used to give the standard deviation, $\hat{\sigma}_{TM}$.

The improved procedures are expected to be able to overcome the problem of overestimation in the computation of standard deviation.

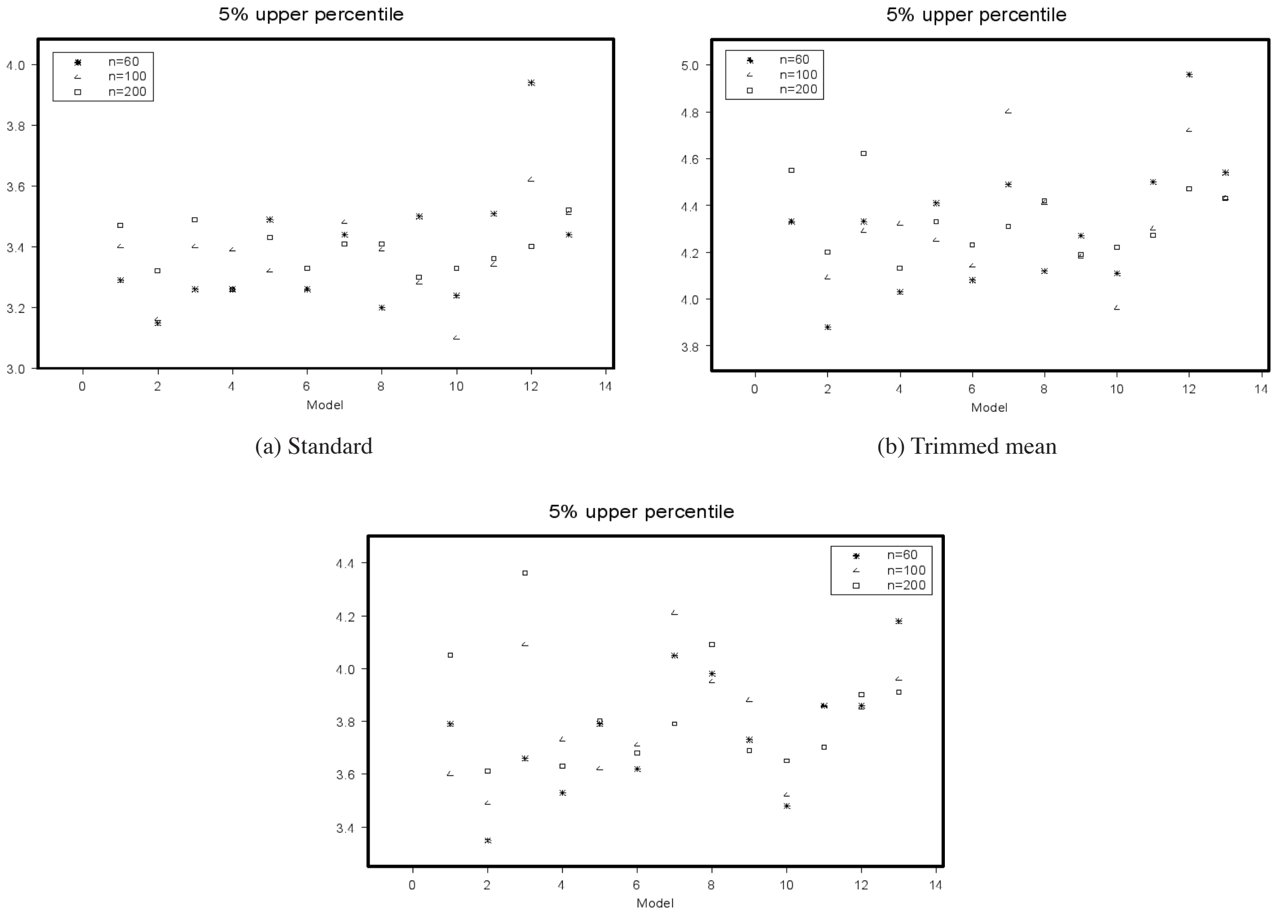
SIMULATION – CUTPOINTS

We consider 13 different models representing a broad choice of coefficients of BL(1,0,1,1) and BL(1,1,1,1) models satisfying the stationary condition of the bilinear model. Table 1 lists the full models for the BL(1,0,1,1) case. For each model, three cases of the sample are considered; n = 60, n = 100 and n = 200. The random errors, are assumed to follow the standard normal distribution. For each model and for each sample size, 100 series are generated. The test statistics for the IO given by equation (8) are calculated based on the standard, trimmed mean and MAD procedures. The focus is to examine the sampling behavior of $\eta_t = \max_{t=1,2,\ldots,n} \left\{ |\hat{\tau}_t| \right\}$. In particular, the percentiles of the test statistics at the 10%, 5% and 1% levels are estimated when no outlier is present in the series.

The plots for 5% percentiles values are given in Figure 3. From the figures, there is no clear pattern of increment or decrement of values in sample size of n, n = 60, 100, 200. For the standard and MAD methods, the values lie between 3-4, while for the trimmed mean method, the values lie in the range of 3.8-4.8. Based on the results, for the standard and MAD procedures, critical values of 2.5 to 4.0 seem to be the suitable choices for the series of size between 60-200, while we may use higher values between 3 to 4.5 for the trimmed mean procedure. In practice, more than one critical value is suggested for the analysis. Similar results are observed for the BL(1,1,1,1) model and they are not given here.

TABLE 1. List of models used for the determination of the critical values for BL(1,0,1,1)

Model	Full Model
1	$Y_t = 0.1Y_{t-1} + 0.1Y_{t-1}e_{t-1} + e_t$
2	$Y_t = 0.1Y_{t-1} + 0.3Y_{t-1}e_{t-1} + e_t$
3	$Y_t = 0.1Y_{t-1} + 0.5Y_{t-1}e_{t-1} + e_t$
4	$Y_t = 0.2Y_{t-1} + 0.2Y_{t-1}e_{t-1} + e_t$
5	$Y_t = 0.3Y_{t-1} + 0.3Y_{t-1}e_{t-1} + e_t$
6	$Y_t = 0.4Y_{t-1} + 0.2Y_{t-1}e_{t-1} + e_t$
7	$Y_t = 0.5Y_{t-1} + 0.1Y_{t-1}e_{t-1} + e_t$
8	$Y_t = 0.1Y_{t-1} + 0.1Y_{t-1}e_{t-1} + e_t$
9	$Y_t = 0.1Y_{t-1} + 0.3Y_{t-1}e_{t-1} + e_t$
10	$Y_t = 0.2Y_{t-1} + 0.2Y_{t-1}e_{t-1} + e_t$
11	$Y_t = 0.4Y_{t-1} + 0.2Y_{t-1}e_{t-1} + e_t$
12	$Y_t = 0.3Y_{t-1} + 0.1Y_{t-1}e_{t-1} + e_t$
13	$Y_t = 0.5Y_{t-1} + 0.1Y_{t-1}e_{t-1} + e_t$



(a) Standard

(b) Trimmed mean

(c) MAD

FIGURE 3. Plot of critical values of IO procedure for BL(1,0,1,1)

SIMULATION - PERFORMANCE

The outlier detection procedure is now applied to cases characterized by a combination of the following factors:

1) Two underlying models; BL(1,0,1,1) and BL(1,1,1,1), with different combinations of coefficients
2) A single IO at $t = 40$ in samples of size 100.
3) Two different values of outlier effect; $\omega = 3, 5$.
4) Three different levels of critical values; 2.5, 3.0, 3.5.

Series are generated to contain a single IO. The standard deviation of the noise process for each model is set to be unity. For the given model, 500 series of length 100 are generated using the *rnorm* procedure in S-Plus. The summary of the performance of the procedure is given in Tables 2 and 3 for BL(1,0,1,1) and BL(1,1,1,1) models respectively. In each table, the values in columns 4-6 represent relative frequency or proportion of correctly detecting IO with correct location at t=40 for critical values equal 2.5, 3.0 and 3.5 respectively, using different procedures and different magnitudes of the outlier.

Two main results are observed. Firstly, all three procedures perform quite well. As expected, the performance of the procedures improves when larger values of ω are used. Also, as larger critical values are used, the proportions of detection decrease. However, the performance is reduced when larger coefficient values are used. It is known that when larger coefficient values are used, there tends to be more spikes appearing in the data generated from the bilinear process. Consequently, it is expected to be harder to detect the outlier especially for small values of ω. Secondly, in general, the procedure based on the trimmed mean has improved the detection of IO compared to the standard procedure. However, the performance of the procedure based on MAD does not differ much from the standard procedure.

APPLICATION: KAMPUNG ARING MONTHLY RAINFALL DATA

The analysis of the rainfall data is carried out. The data were collected from Kampung Aring weather station, Kelantan, Malaysia for the period of August 1995 to July 2002. The plot of the monthly average in millimeters is given in Figure 4. It can be observed that the data are generally stationary in mean and variance except at time point 41 and 77, where rainfalls were heavy.

The possibility of fitting a non-linear model on the rainfall data is investigated. The non-linearity test has been

TABLE 2. The performance of three procedures for BL (1,0,1,1) models

BL(1,0,1,1)					
Co-Efficients	Magnitude of Outlier	Procedures	Proportion of Correct Detection		
			2.5	3.0	3.5
a=0.1 b=0.3	3	Standard	0.75	0.41	0.26
		TM	0.67	0.65	0.52
		MAD	0.70	0.50	0.35
	5	Standard	0.94	0.94	0.87
		TM	0.96	0.96	0.96
		MAD	0.94	0.94	0.92
a=0.5 b=0.1	3	Standard	0.50	0.33	0.22
		TM	0.50	0.50	0.39
		MAD	0.44	0.33	0.28
	5	Standard	1.00	0.83	0.72
		TM	1.00	1.00	0.95
		MAD	0.83	0.83	0.78
a=-0.1 b=-0.1	3	Standard	0.42	0.28	0.15
		TM	0.40	0.34	0.30
		MAD	0.37	0.33	0.14
	5	Standard	0.95	0.90	0.80
		TM	1.00	1.00	1.00
		MAD	1.00	0.90	0.85
a=-0.5 b=-0.1	3	Standard	0.47	0.47	0.13
		TM	0.60	0.53	0.53
		MAD	0.50	0.43	0.21
	5	Standard	1.00	1.00	0.88
		TM	1.00	1.00	1.00
		MAD	0.94	0.94	0.94

TABLE 3. The performance of the three procedures for BL(1,1,1,1) models

BL(1,0,1,1)					
Co-Efficients	Magnitude of Outlier	Procedures	Proportion of Correct Detection		
			2.5	3.0	3.5
a=0.1 c=0.1 b=0.3	3	Standard	0.49	0.32	0.22
		TM	0.51	0.51	0.46
		MAD	0.49	0.32	0.27
	5	Standard	0.71	0.69	0.60
		TM	0.76	0.76	0.74
		MAD	0.74	0.67	0.60
a=0.3 c=0.1 b=0.1	3	Standard	0.60	0.52	0.38
		TM	0.65	0.65	0.56
		MAD	0.65	0.56	0.42
	5	Standard	0.78	0.75	0.70
		TM	0.70	0.70	0.68
		MAD	0.58	0.58	0.55
a=-0.4 c=0.2 b=-0.2	3	Standard	0.63	0.48	0.32
		TM	0.62	0.59	0.55
		MAD	0.55	0.45	0.34
	5	Standard	0.69	0.69	0.69
		TM	0.72	0.72	0.72
		MAD	0.66	0.66	0.66
a=-0.5 c=-0.1 b=-0.1	3	Standard	0.57	0.45	0.29
		TM	0.57	0.57	0.53
		MAD	0.53	0.50	0.33
	5	Standard	0.84	0.81	0.72
		TM	0.82	0.82	0.82
		MAD	0.84	0.83	0.74

widely used to determine whether a given data set is linear or non-linear. Two such tests are Keenan's test by Keenan (1985) and F-test by Tsay (1986). When non-linearity tests are applied, the p-values of the Keenan' test and the F-test are 0.0293 and 0.2777, respectively. The Keenan's test strongly suggests that the data is non-linear. Ismail et al. (2008) had shown that the data is best fitted using the BL(1,0,1 ,1) model with the parameter estimates $\hat{a} = 0.364$ (standard error = 0.151) and $\hat{b}$ = -0.001 (standard error = 0.0004).

When the detection procedure based on the BL(1,0,1,1) model is applied to the data, an innovational outlier (IO) is detected at time point 41 for all the methods; with values of test statistics for standard method (5.430), trimmed mean method (6.98) and MAD method (5.14). Time point 41 corresponds to December 1998.

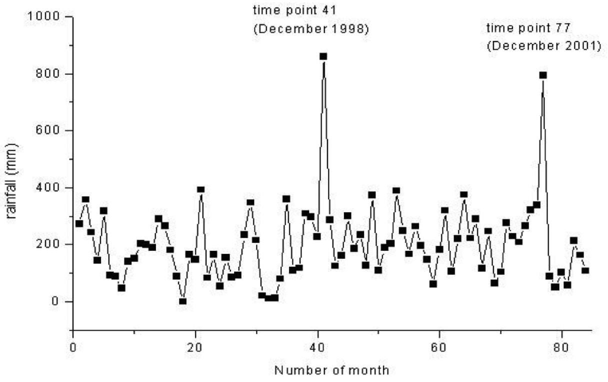


FIGURE 4. Plot of the rainfall data in Kampung Aring

CONCLUSION

An improved version of the outlier detection procedure for BL(1,0,1,1) and BL(1,1,1,1) models to detect IO is proposed in this paper. This simulation study shows that, in general, the three procedures work well in detecting IO with the procedure based on a trimmed mean, shows better results compared to the others. The proportion of correct detection is higher when the magnitude of the outlier effect is large. The detection procedure is applied on a local rainfall data set and is able to detect an IO in the data set.

REFERENCES

Chang, I., Tiao, G.C. & Chen, C. 1988. Estimation of time series parameters in the presence of outliers. *Technometrics* 30: 193-204.

Chen, C. & Liu, L.-M. 1993. Joint estimation of model parameters and outlier effects in time series. *Journal of American Statistical Society* 88: 284-297.

Chen, C.W.S. 1997. Detection of additive outliers in bilinear time series. *Computational Statistics and Data Analysis* 24: 283-294.

Fox, A.J. 1972. Outliers in time series. *Journal of the Royal Statistical Society* B 34: 350-363.

Granger, C.W.J. & Andersen, A.P. 1978. *Introduction to Bilinear Time Series Models*. Gottinge: Vandenhoeck and Ruprecht.

Hampel, F.R., Ronchetti, E.O., Rousseeuw, P.J. & Stahel, W.A. 1986. *Robust Statistics: The Approach Based on Influence Functions*. Toronto: John Wiley Inc.

Ismail, M.I, Mohamed, I.B. & Yahya, M.S. 2008. Improvement on additive outlier detection procedure in bilinear model. *Malaysian Journal of Science* 27(2): 107-114.

Keenan, D.M. 1985. A Tukey non-additivity type test for time series nonlinearity. *Biometrika* 72: 39-44.

Priestley, M.B. 1991. *Non-linear and Non-stationary Time Series Analysis*. San Diego: Academic Press.

Tsay, R. S. 1986a. Nonlinearity test for time series. *Biometrika* 73: 461-466.

Tsay, R.S. 1986b. Time series model specification in the presence of outliers. *Journal of the American Statistical Association* 81: 132-141.

Zaharim, A., Mohamed, I. B., Ahmad, I., Abdullah, S. & Omar, M.Z. 2006. Performances Test statistics for single outlier detection in bilinear (1,1,1,1) models, *WSEAS Transactions on Mathematics* 5(12): 1359-1364.

Ibrahim Mohamed* & Mohd Isfahani Ismail
Institute of Mathematical Sciences
University of Malaya
50603 Kuala Lumpur
Malaysia

Mohd Sahar Yahya & Abdul Ghapor Hussin
& Noraini Mohamed
Centre for Foundation of Studies in Sciences
University of Malaya
50603 Kuala Lumpur
Malaysia

Azami Zaharim
Faculty of Engineering and Built Environment
Universiti Kebangsaan Malaysia
43600 UKM Bangi, Selangor D.E.
Malaysia

Mohammad Said Zainol
Fakulti Teknologi Maklumat dan Sains Kuantitatif
Universiti Teknologi MARA
47000 Shah Alam, Selangor D.E.
Malaysia

*Corresponding author; email: imohamed@um.edu.my

Received: 29 December 2009
Accepted: 4 August 2010