Using Robust Outlier Detection To Identify Possible Flood Events

AZAMI ZAHARIM², RAFIZAH RAJALI¹, KAMARULZAMAN IBRAHIM¹
¹Department of Statistics, Universiti Kebangsaan Malaysia
²Fundamental Engineering Unit Faculty Engineering and Built Environment,
Universiti Kebangsaan Malaysia
43600 Bangi Selangor
MALAYSIA

Abstract:- The least squares method has been widely used in time series forecasting and outlier detection. However, the method is not very efficient in identifying outliers because it suffers the masking effect. The aim of this study is to overcome the masking effect by implementing the robust least median squares method in outlier detection. To illustrate, we identified the possible outliers from sixty-one readings of the daily rainfall recorded at Kajang JPS telemetric station. The outliers are then categorized into innovational outlier and additive outlier. Results based on both techniques were compared and it is found that the least median squares method effectively unmasked the effect of outliers as compared to the least squares method.

Key-Words: Outliers detection, robust least median squares, masking effect

1 Introduction
Robust method of least median squares (LMS) in outlier detection is more efficient than the classical method of least squares (LS) because the LMS method is immune to the masking effect suffered by the LS method (Rousseeuw & Leroy 2003). Masking effect is the failure of a procedure to detect one or more outliers in the existence of several suspected values (Barnet & Lewis 1994).

Daily amounts of rainfall in certain areas are recorded by telemetric stations to serve as a warning system against possible flood that occurs after the event of a heavy downpour. Outlier detection procedures can be used to detect an extreme rainfall on a particular day because an extreme rainfall is represented by an extreme value. As an illustration, outlier detection using the LS and LMS method were applied to sixty one daily readings of rainfall (mm) for the month of October and November 1999 recorded by the telemetric station of JPS Kajang.

The two types of outlier commonly found are additive outlier and innovational outlier. Additive outlier is deterministic in nature and it is usually associated with an isolated incident such as recording error. The innovational outlier on the other hand, contributes to an extraordinary shock in the
series at a particular time as well as in the subsequent observations (Zaharim, 1996). Thus, the ability to classify an outlier accurately will help identify a false alarm for the occurrence of flood.

2 Model Identification

The time series plot for the daily rainfall data (x_t) in Figure 1, suggest that x_t is non-stationary. First differencing is used to transform x_t into the stationary series y_t as illustrated in Figure 2. Plots of sample partial autocorrelation function (SPAC) and SAC for y_t are then used to tentatively identify plausible models for y_t. Since the SAC and the SPAC for y_t both cut off at lag 1, as shown in Figures 3 and 4 respectively, ARMA(1,0), ARMA(0,1) and ARMA(1,1) are possible models for y_t. Other ARMA(p,q) models that correspond to

\[ y_t = \phi_p y_{t-p} + a_{q-a_{t-q}} \]

with \( \phi_p \) and \( \theta_q \) being the estimated parameters for \( p = 0,1,2 \) and \( q = 0,1,2 \) are also fitted to y_t for assessment.

3 Diagnostic Checking

Diagnostic checks are performed on all possible models to validate the adequacy of each model. An adequate model is a model that has p-values of Ljung-Box statistics greater than 0.05 and significant parameter values for all the parameters in the model. Results for the diagnostic checks show that only ARMA (1, 0) and ARMA (0, 1) satisfy the model adequacy conditions as shown in Table 1.

Given more than one adequate model, the model that has the smallest standard deviation, Akaike information criterion (AIC) and Schwarz information criterion (SIC) will be chosen as the best model that describes the data. Results given in Table 1 suggest that ARMA (0, 1) is a
better model than ARMA (1, 0) as it has the smallest value of the three measures. Since the ARMA (0, 1) is more difficult to deal with numerically due to the moving average part as opposed to ARMA (1, 0), for further discussion on identification of outlier, it is sufficient to only consider the ARMA (1, 0) model (Rousseeuw & Leroy, 2003).

Diagnostic checking for the ARMA (1, 0) that uses the LMS method is performed to ensure adequacy before applying it to the outlier detection procedures. Results affirming the adequacy and goodness of fit can be found in Table 2.

Table 1: Diagnostic checking for ARMA(1,0) and ARMA(0,1) that use LS methods

<table>
<thead>
<tr>
<th>Model</th>
<th>Parameter</th>
<th>Ljung-Box Statistic</th>
<th>$\hat{\sigma}_e$</th>
<th>AIC</th>
<th>SIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARMA (1,0)</td>
<td>$\hat{\phi}_1 = -0.4716; p$-value = 0.00</td>
<td>Lag 12 24 36 48 Chi-Square 10.6 14.0 28.4 48.7 DF 11 23 35 47 p-Value 0.48 0.93 0.78 0.40</td>
<td>22.16</td>
<td>534.12</td>
<td>368.76</td>
</tr>
<tr>
<td>ARMA (0,1)</td>
<td>$\hat{\theta}_1 = 0.9488; p$-value = 0.00</td>
<td>Lag 12 24 36 48 Chi-Square 5.9 10.9 22.2 43.7 DF 11 23 35 47 p-Value 0.88 0.98 0.96 0.61</td>
<td>19.52</td>
<td>528.08</td>
<td>359.90</td>
</tr>
</tbody>
</table>

Table 2: Diagnostic checking for ARMA(1,0) that uses LMS method

<table>
<thead>
<tr>
<th>Model</th>
<th>Parameter</th>
<th>Ljung-Box Statistic</th>
<th>$\hat{\sigma}_e$</th>
<th>AIC</th>
<th>SIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARMA (1,0)</td>
<td>$\hat{\phi}_1 = -0.6579; p$-value = 0.02</td>
<td>Lag 12 24 36 48 Chi-Square 14.45 18.93 30.99 56.23 DF 11 23 35 47 p-Value 0.21 0.71 0.66 0.17</td>
<td>22.65</td>
<td>536.73</td>
<td>371.37</td>
</tr>
</tbody>
</table>
4 Outlier Detection

The iterative procedures that are used for detecting the innovational outlier (IO) and additive outlier (AO) are based on works by Chen and Liu(1993) and also Azami Zaharim(1996). Results for the outlier detection of the LS and LMS methods are shown in Tables 3 and 4 respectively.

Using the critical value (C) equal to 2.5, both the LS and LMS have identified the 6th and 16th observations as outliers. However, the LMS method has detected an additional outlier at the 61st observation. As the outlier detection is based on the residuals, line plots of the residuals for both the method can be seen in Figures 5 and 6.

Table 3 : Summary for the outlier detection using ARMA(1,0) that is obtained by means of LS method with $C= 2.5$

<table>
<thead>
<tr>
<th>Iteration</th>
<th>$\hat{\phi}$</th>
<th>$\sigma_\varepsilon$</th>
<th>Type</th>
<th>Time in Day</th>
<th>Standardized Outlier Effect, $\tau$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-0.4716</td>
<td>22.16</td>
<td>IO</td>
<td>16</td>
<td>3.64</td>
</tr>
<tr>
<td>1</td>
<td>-0.4248</td>
<td>19.61</td>
<td>IO</td>
<td>6</td>
<td>2.84</td>
</tr>
</tbody>
</table>

No additional outlier detected

Table 4 : Summary for the outlier detection using ARMA(1,0) that is obtained by means of LMS method with $C= 2.5$

<table>
<thead>
<tr>
<th>Iteration</th>
<th>$\hat{\phi}$</th>
<th>$\sigma_\varepsilon$</th>
<th>Type</th>
<th>Time in Day</th>
<th>Standardized Outlier Effect, $\tau$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-0.6579</td>
<td>22.65</td>
<td>IO</td>
<td>16</td>
<td>3.39</td>
</tr>
<tr>
<td>1</td>
<td>-0.6579</td>
<td>20.48</td>
<td>IO</td>
<td>6</td>
<td>2.69</td>
</tr>
<tr>
<td>2</td>
<td>-0.6579</td>
<td>19.27</td>
<td>IO</td>
<td>61</td>
<td>2.56</td>
</tr>
</tbody>
</table>

No additional outlier detected
5 Conclusion
LS and LMS outlier detection were successful in identifying the heavy downpour on the 16th of October (16th observation) which leads to the worst flood that hit Kajang town in 28 years on the 17th of October 1999. (Malay Mail, 19/10/1999). However, the LS method failed to detect the IO at the 61st observation as it suffers the masking effect. This problem has been overcome by the usage of LMS technique in outlier detection.

References: