The Effect of Influential Outliers on Regression Analysis in Predicting Ground-Level Ozone Concentrations

Muqhlisah Muhamad, Ahmad Zia Ul-Saufie Mohamed Japeri, Sayang Mohd Deni, Ahmad Shukri Yahaya

Abstract— Outliers are data which deviate too far from other observations. The contamination of outliers in air pollution data are always associated with bias results which presents less accurate information of air pollution due to the non-reflecting conditions of the actual phenomenon. Most of data cannot be claimed to be free from influential outliers. However, the study of influential outliers received less attention and was not studied so frequently in order to check the adequacy of the regression model. Even a single observation of the influential outliers can have a severe distortion on the model of prediction: The aim of this study is to evaluate the influence of outliers using standardized residual and Cook’s distance on the prediction of ozone (O₃) concentrations level by excluding the point of outliers in the observation. The prediction models developed for Shah Alam, Pasir Gudang and Jerantut. The result will show the difference of regression coefficients between cases included (origin) and cases excluded for next day prediction (D+1) in Shah Alam, Pasir Gudang and Jerantut. The performance indicator of the models are measured by normalized absolute error (NAE) and prediction accuracy (PA). The average NAE and PA reduced by 0.26% and 0.23% respectively. The model developed could be implemented among local authorities and non-government organization (NGO) in order to prepare people with early action or precaution.

Index Terms— Influential Outliers, Ozone, Regression Coefficients, Standardized Residual, Cook’s Distance

1 INTRODUCTION

Ground-level ozone (O₃) has become a significant pollutant as a result of economy and population growth. O₃ is a secondary pollutant which means that it is not emitted into the air directly but it needs some combination of volatile organic compounds (VOC’s) and nitrogen oxides (NOₓ) under the influence of ultraviolet (UVB) [1]. According to Environmental Protection Agency [2], O₃ concentration was at high level from morning till afternoon and low level at night. The air pollution index reading was dominated by O₃ concentrations level around the afternoon and early evening [3].

Most of the data cannot be claimed to be free from the outliers including the data of O₃ concentrations. An outliers is a point that is far from another point and has a large residual. According to Berry & Feldman [4], an outliers is a data point where the dependent variable does not follow the trend of the rest of the data. If a regression analysis such as the predicted response, the estimated slope coefficients or the hypothesis test results is too much influenced by data point, so it is suspected as influential outliers as known as influential cases [5].

In regression analysis, the fit of a statistical model becomes lower significance with the presence of outliers [6]. The offense in data recording and sampling, the error in data acquisition or data management and the damage of monitoring instrument in data recording are the factors that contribute to the formation of outlier [7].

The influential outlier could be from x-space or y-space. Midi et al. [8], Sarkar, Midi, and Rana [9], pointed that, in x-space, the influential observation measured by leverage values in order to measure how far an independent variables deviate from its mean while standardized residual and Cook’s distance are used to measure the influential outlier from y-space by the change of the estimate of coefficients. The regression coefficients not too affected by the large leverage values [10].

According to Yahaya [5] and Field [10], the whole influential outlier from the observation could be measured using standardized residual or Cook’s distance because the influential for the overall of outliers depend on the dependent variable of y-space. This study attempted to present the result in regression analysis after considering the influential outlier from y-space based on the standardized residual and Cook’s distance in Shah Alam, Pasir Gudang and Jerantut for next day prediction (D+1) of O₃ concentrations level.

2 METHODOLOGY

2.1 Data Acquisition and Data Management

According to Ghazali et al. [11] and Banan et al. [12] the measurements of air pollutants and meteorological variables followed the procedure of monitoring record from the standard by United States Environmental Protection Agency (EPA) [2] [11].

The primary data was managed by Alam Sekitar Malaysia Sendirian Berhad (ASMA) [13], which is the private company under supervision of Department of Environmental Malaysia (DoE). Furthermore, the secondary data from 1st January 2002 until 31st December 2013 were obtained from Department of Environmental Malaysia (DoE).

In this study, the hourly concentrations for each variables are transformed into daily 12 hours average concentrations, from 7am to 7pm because the level of O₃ concentrations level was suspected to be highly active during morning and
evening rush hour [14] According to Mohammed et. al. [15], most of the areas have a large emission of O₃ formation factor during morning to evening. Eighty percent of daily data records were randomly selected to develop the models and twenty percent were used for validation of the models [16].

### 2.2 Area

Monitoring station in Shah Alam is located at Taman Tun Dr. Ismail (TTDI) Jaya Primary School (N 3.077324°, E 101.510323°) and nearby residential area (Figure 1). At the same time, this station is located at the main transportation area such as major road, highways, and airport as well as surrounded by low industrial area [17]. Besides, Shah Alam city is located at the center of Petaling Jaya city (east) and Klang town (west) [18].

Pasir Gudang (N 1.470750°, E 103.895702°) is situated at Sekolah Menengah Pasir Gudang 2, Johor. It is the main location of industrial activities and main road for transportation. It is located in the middle of the Malaysian peninsular specifically at Meteorological Department at Batu Embun, Jerantut, Pahang.

According to Azmi et al. [17], Jerantut (N 3.949639°, E 102.364400°) is a background area surrounded by natural forest, low open burning and a low number of motor vehicles. It is located at the middle of the Malaysian peninsular specifically at Meteorological Department at Batu Embun, Jerantut, Pahang.

### 2.3 Influential Outliers

The influential point of outliers are the case where there is existing larger residual that differ from the other observation substantially [21]. Influential outliers are any point that has a large effect in analysis of regression. According to Sarkar et al. [9], the results of the analysis will lead to incorrect inferences by the unduly influence of outliers. The change of regression coefficients after removing some data observation shows that the data before was influenced by outliers. Table 1 shows the initial diagnostic to assess the influence of outlier.

### 2.4 Multiple Linear Regression (MLR)

Regression analysis that was used in this study is multiple linear regression (MLR) based on traditional approaches of ordinary least square estimate (OLS). MLR is an extension from a simple linear regression. In MLR, there are one dependent variable and several independent variables. Chatterjee & Hadi [25], define the general equation of MLR as follow,

\[
y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, 2, \ldots, n
\]

where,

- \(y_i\) = dependent variable
- \(x_{ip}\) = independent variable (predictor)
- \(\beta_p\) = represent values of the predictors for \(i\)th unit
- \(\beta_0\) = regression constant
- \(\varepsilon_i\) = regression coefficient (parameter estimate)

The common method in measurement of influential outliers is by using standardized residual. According to Blatna [22], there are two types to calculate the standardized residual for the \(i\)-observation, one uses the residual mean square error and the others use the residual mean square. Both of residuals are selected from the model fitted to the full dataset. Meanwhile, when the residual more than absolute 3, there must be a problematic case in the influence of outliers.

Researcher use Cook’s distance method to measure the influential outliers. Cook’s distance formula acts as a measurement to the overall influence of outliers on the model of prediction [10]. According to Cook & Weisberg [24], there must be a concerned matter if the values of Cook’s distance is greater than 1. Besides, Bohrnstedt & Knoke [21] noted that when the value of \(D_i\) is smaller than Cook’s distance there may be some underlying problems of outliers, where \(N\) is the number of observation.

These methods involve comparison of the regression analysis between the case included (origin) and the case excluded (after excluding the influential point of outliers). In this study, the data observation of the standardized residuals more than absolute 3 and the Cook’s distance more than 4/N will be deleted and the model will run again to obtain an appropriate regression coefficient after the exclusion of some cases [23].
(SO$_2$), nitrogen oxides (NO$_x$), ozone (O$_3$) and carbon monox-
ides (CO).

At the first step, the initial diagnostic of OLS will be con-
sidered to assure the data that satisfy the OLS as-
sumption [26]. The data will be run to get the initial model of prediction
without considering the influential outliers (origin).

Then, the data will be run again to obtain the model after
excluding the point of influential outliers by removing some
data observation. According to Field [10], when the value of
standardized residual is larger than absolute value of 3 and
when the value of Cook's distance is larger than 4/N, these
cases should be investigated. The analysis of regression such
as scatter plot and the regression coefficients (estimated
parameters) between these two models (origin and cases
excluded) will be compared.

2.5 Performance Indicator (PI)
Performance indicators (PI) are used to evaluate the per-
formance models for next day (D+1) prediction. The PI (Table 2)
are consists of normalized absolute error (NAE) and predic-
tion accuracy (PA). Both models of prediction before and after
considering the influence of outliers will be compared by the
PI to show the improvement.

<table>
<thead>
<tr>
<th>Performance Indicator (PI)</th>
<th>Formulae</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normalized Absolute Error (NAE)</td>
<td>$\sum_{i=1}^{N} \left</td>
<td>P_i - O_i \right</td>
</tr>
<tr>
<td>Prediction Accuracy (PA)</td>
<td>$\sum_{i=1}^{N} \left( \frac{P_i - P}{O_i - \bar{O}} \right)^2$</td>
<td>Close to 1, model is appropriate</td>
</tr>
</tbody>
</table>

where,

$N$ = Number of sample daily measurement of a selected sites

$P_i$ = Predicted value of one set daily data

$O_i$ = Observed values of one set daily data

$\bar{P}$ = Mean of the predicted values of one set daily data

$\bar{O}$ = Mean of the observed values of one set daily data

3 RESULTS AND DISCUSSIONS

3.1 Descriptive Statistics
Descriptive statistics was used to describe a situation [28]. In
order to display the graphical method of descriptive statistics,
boxplot was chosen as an alternative way to describe the value
of median, minimum, maximum, first quartile, third quartile

and the total observation of outliers [29]. The total number of
outliers and extreme value could be examined by using the
following formulas [28]:

$\text{Outliers : } UQ + 1.5IQR = M$  \hspace{1cm} (3)

$\text{Extreme value : } UQ + 3.0IQR = M$  \hspace{1cm} (4)

For this study, M represents the last value of O$_3$ that
consider as an outlier and extreme value after rearrange the
data. By the observation of O$_3$ till M, the total number of data
that contain the outliers (equation 3) and extreme value (equa-
tion 4) will be obtained. The percentage of the outliers and
extreme value could be counted as follow;

Percentage of outliers;
\[
\frac{\text{total of outliers}}{N} \times 100
\]

Percentage of extreme value;
\[
\frac{\text{total of extreme value}}{N} \times 100
\]

where, $N$ is the total number of O$_3$.

The monitoring station in Shah Alam is located at Taman Tun Dr. Ismail (TTDI) Jaya Primary School (N 3.077324°, E 101.510323°) nearby a residential area. At the same time, this
station is located at the main transportation area such as ma-
jor road, highways, and airport. The mean average of O$_3$
centration for Shah Alam is 0.032 ppm and the monitoring
record is assumed to be moderately skewed with the value of
0.717. The maximum amount of O$_3$ concentration recorded
was 0.097 ppm. The average value of O$_3$ has not been de-
tected at 3.91% and 0.25% respectively.

According to Department of Environment, Ma-
laysia [3], the unhealthy days from year 2001 to 2012 in Klang
Valley was mainly due to the high concentration level of O$_3$.
Shah Alam was recorded as having the highest number of
unhealthy days except for year 2005, 2010, 2011 and 2012. The
total number of outliers and extreme value for O$_3$ concen-
trations were detected at 3.91% and 0.25% respectively.

The mean average of O$_3$ concentration for Pasir Gudang

TABLE 2

<table>
<thead>
<tr>
<th>Performance Indicators [27]</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Performance Indicator (PI)</strong></td>
</tr>
<tr>
<td>Normalized Absolute Error (NAE)</td>
</tr>
<tr>
<td>Prediction Accuracy (PA)</td>
</tr>
</tbody>
</table>

TABLE 5

<table>
<thead>
<tr>
<th>Case number</th>
<th>ZRE</th>
<th>Case number</th>
<th>ZRE</th>
<th>Case number</th>
<th>ZRE</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>4174</td>
<td>(0.036)</td>
<td>3.53712</td>
<td>(0.0293)</td>
<td>3.11430</td>
</tr>
<tr>
<td>1816</td>
<td>1782</td>
<td>(0.0394)</td>
<td>3.49734</td>
<td>(0.0350)</td>
<td>3.06908</td>
</tr>
<tr>
<td>2354</td>
<td>971</td>
<td>(0.0378)</td>
<td>3.49330</td>
<td>(0.0329)</td>
<td>2.993147</td>
</tr>
<tr>
<td>2197</td>
<td>168</td>
<td>(0.0442)</td>
<td>3.41256</td>
<td>(0.0005)</td>
<td>-3.23109</td>
</tr>
<tr>
<td>2291</td>
<td>1316</td>
<td>(0.0443)</td>
<td>3.30114</td>
<td>(0.0163)</td>
<td>-3.34052</td>
</tr>
<tr>
<td>1256</td>
<td>2087</td>
<td>(0.0344)</td>
<td>3.29984</td>
<td>(0.0729)</td>
<td>-3.46733</td>
</tr>
<tr>
<td>1680</td>
<td>1810</td>
<td>(0.0345)</td>
<td>3.25950</td>
<td>(0.0119)</td>
<td>-3.50936</td>
</tr>
<tr>
<td>892</td>
<td>1716</td>
<td>(0.0298)</td>
<td>3.54436</td>
<td>(0.0094)</td>
<td>-6.04158</td>
</tr>
</tbody>
</table>
3.2 Influential Outliers and Analysis of Regression

In Shah Alam, the origin model was built by 2843 observation. After checking the standardized residuals (ZRE), 19 points observation have been detected as influential points since the value of ZRE are greater than \[3\]. All cases number (data observation) as shown in Table 3, will be truncated, for example, the case number for 449 was truncated at 0.0941 ppm level of ozone concentration. Besides, Cook’s distance also was assessed and 140 data observation were detected as an influential points when the values of Cook’s distance are greater than 0.00140696 (4/2843). Finally, the data will be ran at a second time to obtain the model after excluding several data of influential points \[10\].

These assessment of influential point of outliers will be applied and repeated in the prediction of O3 concentrations level in Pasir Gudang (Table 4) and Jerantut (Table 5). For Pasir Gudang, 168 data observation have been removed by 18 data observation of ZRE and 150 data observation of Cook’s assessment. Meanwhile, 132 data observation for Jerantut have been truncated by excluding 24 cases number of ZRE and 108 cases number of Cook’s distance.

After the cases number for each sites were excluded, the new model for each site would be obtained. All of regression coefficients have been changed when the influential outliers are considered. The result of models are summarized in Table 6. Both models of included (origin) and excluded show the differences in regression coefficients (parameter estimates). The differences between regression coefficient proved that the influential outliers give big impact to the prediction model of O3 concentrations level. Besides, the example of scatter plot for Shah Alam (Fig. 1) also shows pretty much well after the drop of problematic cases of the influential outliers. Most of outliers above the line have been removed after considering the standardized residual (ZRE) and Cook’s distance.

### Table 3
The Assessment of the Influential Outliers for Next Day Prediction (Shah Alam)

<table>
<thead>
<tr>
<th>Case number</th>
<th>ZRE</th>
<th>Case number</th>
<th>ZRE</th>
<th>Case number</th>
<th>ZRE</th>
</tr>
</thead>
<tbody>
<tr>
<td>449</td>
<td>0.0941</td>
<td>1738</td>
<td>0.0658</td>
<td>802</td>
<td>0.0721</td>
</tr>
<tr>
<td>1470</td>
<td>0.0941</td>
<td>2698</td>
<td>0.0745</td>
<td>697</td>
<td>0.0748</td>
</tr>
<tr>
<td>2761</td>
<td>0.0814</td>
<td>749</td>
<td>0.0721</td>
<td>1904</td>
<td>0.0026</td>
</tr>
<tr>
<td>1033</td>
<td>0.0802</td>
<td>1832</td>
<td>0.0772</td>
<td>1499</td>
<td>-3.30420</td>
</tr>
<tr>
<td>1485</td>
<td>0.0802</td>
<td>2262</td>
<td>0.0663</td>
<td>43</td>
<td>-4.07700</td>
</tr>
<tr>
<td>889</td>
<td>0.0746</td>
<td>2062</td>
<td>0.0639</td>
<td>697</td>
<td>0.049759</td>
</tr>
<tr>
<td>2180</td>
<td>0.0652</td>
<td>2678</td>
<td>0.0684</td>
<td>3.52923</td>
<td>-0.001</td>
</tr>
</tbody>
</table>

### Table 4
The Assessment of the Influential Outliers for Next Day Prediction (Pasir Gudang)

<table>
<thead>
<tr>
<th>Case number</th>
<th>ZRE</th>
<th>Case number</th>
<th>ZRE</th>
<th>Case number</th>
<th>ZRE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1781</td>
<td>0.0494</td>
<td>1742</td>
<td>0.045</td>
<td>3</td>
<td>0.0377</td>
</tr>
<tr>
<td>667</td>
<td>0.0427</td>
<td>623</td>
<td>0.0418</td>
<td>4</td>
<td>0.0383</td>
</tr>
<tr>
<td>1941</td>
<td>0.0472</td>
<td>1895</td>
<td>0.0378</td>
<td>2</td>
<td>0.0383</td>
</tr>
<tr>
<td>2204</td>
<td>0.0441</td>
<td>2301</td>
<td>0.0376</td>
<td>6</td>
<td>0.005</td>
</tr>
<tr>
<td>2293</td>
<td>0.0496</td>
<td>2270</td>
<td>0.0495</td>
<td>7</td>
<td>0.004</td>
</tr>
<tr>
<td>2520</td>
<td>0.0435</td>
<td>1659</td>
<td>0.0392</td>
<td>5</td>
<td>0.0056</td>
</tr>
</tbody>
</table>

After the cases number for each sites were excluded,
3.3 Performance Indicators

The performance indicators for next day prediction (D+1) in Shah Alam, Pasir Gudang and Jerantut are shown in Table 7. The average of NAE and PA for Shah Alam show the improvement by the decreasing 0.16% of the average NAE and the increasing 0.19% of the average PA. Since 159 observation have been truncated, including several large residual values and 140 number of data Cook’s distance which are larger than 0.0014069447, thus the model after cases excluded has been improved.

For Pasir Gudang, the model (case excluded) has been declined by 0.91% of the average PA even 168 of data observation have been removed due to the small value of residual (close to 3 and -3). The small value of residual cannot give the high impact to the excluded model. However, this model still shows an improvement by the decreasing 0.43% of the average NAE.

The performance indicators of the excluded model for Jerantut has been improved when the truncation process of the 132 cases number of the influential points gave an improvement when the average NAE has been decreased by 0.17% and average PA has been increased by 0.03%.

These three models from Shah Alam, Pasir Gudang and Jerantut could be used as a precaution among the citizens and local authorities. People can take an early action as a way to avoid the impact of O₃ to their health.

### Table 6
**Summarization of the Model**

<table>
<thead>
<tr>
<th>Sites</th>
<th>N (Origin)</th>
<th>Model (Origin)</th>
<th>N (Case Excluded)</th>
<th>Model (Case Excluded)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shah</td>
<td>2843</td>
<td>O₃, D+1 = 0.065962 + 0.000201WS -0.001167T -0.000248RH +0.019316 NO -0.012392 SO₂ +0.071989 NO₃ +0.987555SO₃ +0.046577NO₂ +0.001877 CO</td>
<td>2684</td>
<td>O₃, D+1 = 0.065170 + 0.000064WS -0.001087T -0.000257RH +0.011587NO -0.098755SO₂ +0.046577NO₃ +0.002415CO</td>
</tr>
<tr>
<td>Alam</td>
<td>2638</td>
<td>O₃, D+1 = 0.016850 + 0.000223WS -0.000299T -0.000016RH -0.077266NO₅-0.050056SO₂ +0.207855NO₂ +0.461467CO +0.001237CO</td>
<td>2470</td>
<td>O₃, D+1 = 0.011844 - 0.000411WS -0.000084T-0.000048RH +0.213771NO +0.007035SO₂ +0.539510NO₂ +0.713415O₃ -0.000568CO</td>
</tr>
<tr>
<td>Pasir</td>
<td>2405</td>
<td>O₃, D+1 = 0.011844 - 0.000411WS -0.000084T-0.000048RH +0.213771NO +0.007035SO₂ +0.539510NO₂ +0.713415O₃ -0.000568CO</td>
<td>2273</td>
<td>O₃, D+1 = 0.009988 - 0.000494WS -0.000045T-0.000038RH +0.156441NO -0.029804SO₂ +0.404445NO₂ +0.760523O₃ -0.001158CO</td>
</tr>
<tr>
<td>Gudang</td>
<td>2273</td>
<td>O₃, D+1 = 0.009988 - 0.000494WS -0.000045T-0.000038RH +0.156441NO -0.029804SO₂ +0.404445NO₂ +0.760523O₃ -0.001158CO</td>
<td>2273</td>
<td>O₃, D+1 = 0.009988 - 0.000494WS -0.000045T-0.000038RH +0.156441NO -0.029804SO₂ +0.404445NO₂ +0.760523O₃ -0.001158CO</td>
</tr>
<tr>
<td>Jerantut</td>
<td>2405</td>
<td>O₃, D+1 = 0.011844 - 0.000411WS -0.000084T-0.000048RH +0.213771NO +0.007035SO₂ +0.539510NO₂ +0.713415O₃ -0.000568CO</td>
<td>2273</td>
<td>O₃, D+1 = 0.009988 - 0.000494WS -0.000045T-0.000038RH +0.156441NO -0.029804SO₂ +0.404445NO₂ +0.760523O₃ -0.001158CO</td>
</tr>
</tbody>
</table>

### Table 7
**Performance Indicator**

<table>
<thead>
<tr>
<th>Sites</th>
<th>Model (Origin)</th>
<th>Model (Case Excluded)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shah</td>
<td>0.233824</td>
<td>0.233445</td>
</tr>
<tr>
<td>Alam</td>
<td>0.253529</td>
<td>0.252433</td>
</tr>
<tr>
<td>Pasir</td>
<td>0.176454</td>
<td>0.176152</td>
</tr>
<tr>
<td>Gudang</td>
<td>0.452115</td>
<td>0.452986</td>
</tr>
<tr>
<td>Jerantut</td>
<td>0.754636</td>
<td>0.754887</td>
</tr>
</tbody>
</table>
4 CONCLUSION

The prediction of \( O_3 \) concentration level has become an important matter in developing the model since it would be the \( O_3 \) concentration descriptor for future action or precaution. However, the presence of the influential outliers has slightly ruin the accuracy of prediction information. By the assessment of the influential outliers using standardized residual and Cook’s distance, the effect of the influential outliers on regression analysis could be seen by the performance indicators of NAE and PA. Thus, the better model obtained after exclude several influential outliers could be used in order to predict the level of \( O_3 \) concentration for next day.

ACKNOWLEDGMENT

A special appreciation to Department of Environmental Malaysia (DoE) for providing the air quality data for this research and special thanks to Institute of Research Management and Innovation (IRMI), Universiti Teknologi MARA, Malaysia for funding this study under 600-600-RMI/IRAGS 5/3 (36/2015) Grant and 600-RMI/FRGS 5/3 (40/2014).

References
