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

Muqhlisah Muhamad, Ahmad Zia Ul-Saufie Mohamad 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 (O_3) 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 (O_3) has become a significant pollutant as a result of economy and population growth. O_3 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 (NO_x) under the influence of ultraviolet (UVB) [1]. According to Environmental Protection Agency [2], O_3 concentration was at high level from morning till afternoon and low level at night. The air pollution index reading was dominated by O_3 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 O_3 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 O_3 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 O_3 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 O_3 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 light 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 [19]. Since Pasir Gudang has become the largest industrial area in Malaysia, it has encouraged the investment from local and foreign investor [20].

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 in 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 several data observation shows that the data before was influnced by outliers. Table 1 shows the initial diagnostic to assess the influence of outlier.

TABLE 1
The Assessment of the Influential Outliers

Influential outliers (y-direction)	Description
Standardized residual: $ZRE = \frac{e_i}{\sqrt{MSE}}$, $^*e_i = y_i - \hat{y}_i$ [10]	Larger than $\lvert 3 \rvert$ deemed an outlier.
Cook's distance: $D_i^* = \frac{e_i^2}{pMSE}\left[\frac{h_{ii}}{(1-h_{ii})^2}\right] \sim F_{p,\,n-p}$ [21]	Greater than $\frac{4}{N}$ indicates that the data point strongly influences the estimated coefficients.

where,

y	=	dependent variable
x	=	independent variable (predictor)
p	=	represent values of the predictors for i^{th} unit
β_0	=	regression constant
β_P	=	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 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 $\frac{4}{N}$ is smaller than Cook's distance there may be some underlying problems of outliers, where N is the number of observation.

These methods involves 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].

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,

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_n x_{in} + \varepsilon, \quad \mathbf{i = 1, 2, \ldots n,} \qquad (1)$$

where,

y	=	dependent variable
x	=	independent variable (predictor)
p	=	represent values of the predictors for i^{th} unit
β_0	=	regression constant
β_P	=	regression coefficient (parameter estimate)

MLR was used as a way to predict O_3 concentrations level for the next day prediction (D+1), including several independent variables such as wind speed (WS), relative humidity (RH), temperature (T), nitric oxides (NO), sulphur oxides

(SO_2), nitrogen dioxides (NO_2), ozone (O_3) and carbon monoxides (CO).

$$3,D+1 \quad P_0 \quad P_1 \quad P_2 \quad P_3 \quad P_4 \qquad (2)$$

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

Then, the data will be ran 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 performance models for next day (D+1) prediction. The PI (Table 2) are consists of normalized absolute error (NAE) and prediction 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 2
PERFORMANCE INDICATORS [27]

Performances Indicator (PI)	Formulae	Notes
Normalized Absolute Error (NAE)	$\frac{\sum_{i=1}^{N}\lvert P_i - O_i\rvert}{\sum_{i=1}^{N} O_i}$	Close to 0, model is appropriate
Prediction Accuracy (PA)	$\frac{\sum_{i=1}^{N}\left(P_i - \overline{P}\right)^{\sim}}{\sum_{\cdot,\cdot}^{N}(O_i - \overline{O})^2}$	Close to 1, model is appropriate

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
$\overline{P}$	=	Mean of the predicted values of one set daily data
$\overline{O}$	=	Mean of the observed values of one set daily data

3 RESULTS AND DISCUSIONS

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]:

$$Outliers: UQ + 1.5IQR = M \qquad (3)$$

$$Extremevalue: UQ + 3.0IQR = M \qquad (4)$$

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 (equation 4) will be obtained. The percentage of the outliers and extreme value could be counted as follow,

Percentage of outliers;

$$\frac{totalofoutliers}{N} \times 100 \qquad (5)$$

Percentage of extreme value;

$$\frac{totalofextremevalue}{N} \times 100 \qquad (6)$$

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 major 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. This is due to open burning and smokes from vehicles [30]. According to Department of Environment, Malaysia [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 concentrations were detected at 3.91% and 0.25% respectively.

The mean average of O_3 concentration for Pasir Gudang is 0.020
erately
centrati
activitie
vironm
was rec
O_3 has
Accordi
data ob
tion is a

The
was rec
The ave
ground
the co
contam

TABLE 5
THE ASSESSMENT OF THE INFLUENTIAL OUTLIERS FOR NEXT DAY PREDICTION (JERANTUT)

Case number	ZRE	Case number	ZRE	Case number	ZRE
100		459		186	
(0.036)	4.4174	(0.0304)	3.53712	(0.0293)	3.11430
1816		1782		2351	
(0.0394)	4.11426	(0.0368)	3.49734	(0.0350)	3.06098
2354		971		2303	
(0.0378)	4.07292	(0.0328)	3.49330	(0.0329)	2.993147
2197		168		90	
(0.0442)	3.93880	(0.0384)	3.41256	(0.005)	-3.23109
2291		1316		2324	
(0.0443)	3.72980	(0.0372)	3.30114	(0.0163)	-3.34052
1256		2087		2090	
(0.0344)	3.69978	(0.0358)	3.29114	(0.0101)	-3.46733
1680		1810		760	
(0.0342)	3.62534	(0.0329)	3.25950	(0.0119)	-3.50936
892		1716		31	
(0.0298)	3.54436	(0.0309)	3.21717	(0.0094)	-6.04158

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 O_3 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 O_3 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)

Case number	ZRE	Case number	ZRE	Case number	ZRE
449		1738		802	
(0.0941)	6.50519	(0.0658)	3.68211	(0.0721)	3.31245
1470		2698		697	
(0.0941)	4.95935	(0.0745)	3.63997	(0.0748)	3.26471
2761		749		1904	
(0.0814)	4.65284	(0.0814)	3.63002	(0.0772)	3.05693
1033		1832		1499	
(0.0802)	4.54432	(0.0663)	3.54977	(0.0026)	-3.30420
1485		2262		43	
(0.0802)	4.34507	(0.0639)	3.52923	(0.001)	-4.07700
889		2062			
(0.0746)	4.09284	(0.0684)	3.36003		
2180		2678			
(0.0652)	3.87848	(0.0631)	3.32757		

TABLE 4
THE ASSESSMENT OF THE INFLUENTIAL OUTLIERS FOR NEXT DAY PREDICTION (PASIR GUDANG)

Case number	ZRE	Case number	ZRE	Case number	ZRE
1781		1742	3.4660	1106	
(0.0494)	4.50821	(0.045)	3	(0.0377)	3.15597
667		623	3.4140	1908	
(0.0427)	3.93482	(0.0418)	4	(0.038)	3.13228
1941		1895	3.2837	691	
(0.0472)	3.75056	(0.0378)	2	(0.0383)	3.11936
2204		2301	3.1730	1094	
(0.0441)	3.70298	(0.0376)	6	(0.005)	-3.24086
2293		2270	3.1714	839	
(0.0496)	3.66339	(0.0495)	7	(0.004)	-3.39054
2520		1659	3.1703	2112	
(0.0435)	3.61877	(0.0392)	5	(0.0056)	-3.44603

TABLE 6
SUMMARIZATION OF THE MODEL

Sites	N (Origin)	Model (Origin)	N (Case Excluded)	Model (Case Excluded)
Shah Alam	2843	$O_{3,\ D+1}$ = 0.065962 + 0.000201WS -0.001167T - 0.000248RH + 0.019016 NO - 0.012392 SO_2 + 0.071989 NO_2 + 0.446465 O_3 + 0.001877 CO	2684	$O_{3,\ D+1}$ = 0.065170+ 0.000064WS -0.001087T - 0.000257RH + 0.011587NO - 0.098755SO_2 + 0.046577NO_2 + 0.444678O_3 + 0.002415CO
Pasir Gudang	2638	$O_{3,\ D+1}$ = 0.016850+ 0.000223WS -0.000299T -0.000016RH - 0.077266NO -0.050056SO_2 + 0.207855NO_2 +0.461476O_3 + 0.001237CO	2470	$O_{3,\ D+1}$ = 0.019195+0.000320WS - 0.000398T -0.000032RH - 0.041060NO -0.077482SO_2 + 0.245734NO_2 +0.498056O_3 + 0.000106CO
Jerantut	2405	$O_{3,\ D+1}$ = 0.011844 - 0.000411WS - 0.000084T- 0.000048RH + 0.213771NO + 0.007003SO_2 + 0.539510NO_2 + 0.713415O_3 - 0.000568CO	2273	$O_{3,\ D+1}$ = 0.009988 - 0.000494WS -0.000045T- 0.000038RH + 0.156441NO - 0.029804SO_2 + 0.404445NO_2 + 0.760523O_3 -0.001158CO

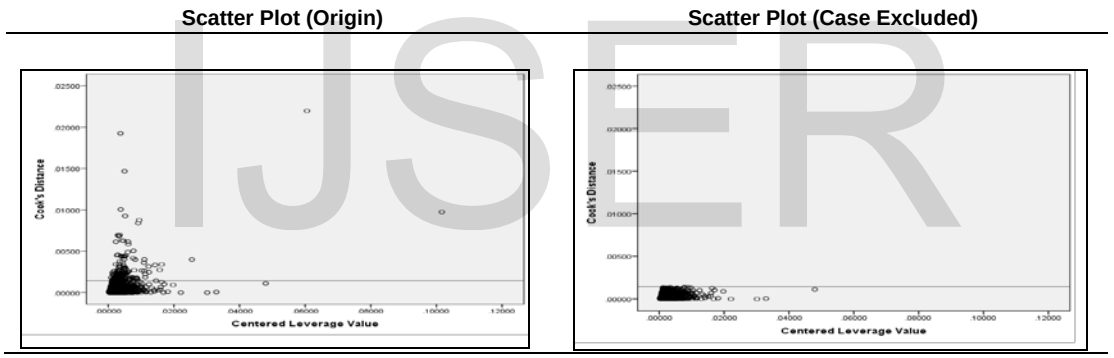

Fig. 1. Scatter Plot of Cook's Distance against Centered Leverage Value

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 0.00140696447, thus the model after cases excluded has been improved.

TABLE 7
PERFORMANCE INDICATOR

Sites	Model (Origin)		Model (Case Excluded)	
	NAE	PA	NAE	PA
Shah Alam	0.233824	0.452115	0.233445	0.452986
Pasir Gudang	0.253529	0.476902	0.252433	0.472564
Jerantut	0.176454	0.754636	0.176152	0.754887

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 O_3 to their health.

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

[1] E. Sanna, Air Pollution and Health, New York: Health and the Environment, 2009.

[2] EPA, "Ground Level Ozone Make it Harder to Breathe, United State Environmental Protection Agency. Online [Accessed 23 April 2015] Available from World Wide: http://www.epa.gov./.," 2012.

[3] DoE, "Department of Environment Malaysia. Malaysia Environmental Quality Report 2010.," Department of Environment, Ministry of natural Resources and Environment, Malaysia., Kuala Lumpur, 2011.

[4] W. D. Berry and S. Feldman, Multiple Regression in Practice. Quantitative Application in Social Sciences, Newbury Park, : Sage University Paper, 1985.

[5] A. S. Yahaya, "Applied Regression Models Using SPSS," Universiti Sains Malaysia, Nibong Tebal, Pulau Pinang, 2014.

[6] B. L. Bowerman and R. T. O'Connell, Linear statistical Models: An Applied Approach., PWS-Kent, Boston, 1990.

[7] J. W. Osborne and A. Overbay, "The Power of Outliers (and why researcher should always check for them)," *Practical Assessment, Research & Evaluation,* 2004.

[8] H. Midi, N. Mohamed Ramli and R. Imon, "The Performance of Diagnostic-Robust Generalized Potentials for the Identification of Multiple High Leverage Points in Linear Regression," *Applied Statistics,* pp. 507-520, 2009.

[9] S. K. Sarkar, H. Midi and R. Rana, "Detection of Outliers and Influential Observations in Binary Logistic Regression: An Empirical Study," *Applied Sciences,* pp. 11 (1): 26-35, 2011.

[10] A. Field, Discovering Statistics Using SPSS, Second ed., London: Sage, 2005.

[11] N. A. Ghazali, N. A. Ramli, A. s. Yahaya, N. F. F. MD Yusof, N. Sansuddin and W. A. Al Madhoun , "Transformation of Nitrogen Dioxide into Ozone and Prediction of Ozone Concentrations Using Multiple Linear Regression," *Environ Monit Assess,* pp. 165: 475-489, 2010.

[12] N. Banan, M. L. Latif, L. Juneng and F. Ahamad, "Characteristics of Surface Ozone Concentrations at Stations with Different Backgrounds in the Malaysian Peninsula," *Aerosol and Air Quality Research,* pp. 13: 1090-1106, 2013.

[13] ASMA, "Alam Sekitar Malaysia Sdn. Bhd.," *http://www.environmalaysia.com,* 2013.

[14] N. R. Awang, N. A. Ramli, N. I. Mohammed and A. S. Yahaya, "Times Series Evaluation of Ozone Concentrations in Malaysia Based on Location of Monitoring Stations," *International Journal of Engineering and Technology,* vol. 3, 2013.

[15] N. I. Mohammed, N. A. Ramli and A. S. Yahaya, "Ozone Phytotoxicity Evaluation and prediction of Crops Production in Tropical Regions," *Atmospheric Environment,* pp. 343-349, 2013.

[16] A. Z. Ul-Saufie, Future Daily Particulate Matter Concentrations Prediction Using Regression Artificial Neural Network and Hybrid Models in Malaysia, Pulau Pinang : Universiti Sains Malaysia, 2012.

[17] S. Z. Azmi, M. T. Latif, A. S. Ismail, L. Juneng and A. A. Jemain, "Trend and Status of Air Quality at Three Different Monitoring Stations in the Klang Valley, Malaysia," *Air Qual Atmos Health,* pp. 3(1): 53-64, 2010.

[18] O. L. H. Leh, S. N. A. Mohamed Musthafa and A. R. Abdul Rasam, "1. Oliver Ling Hoon Leh, Siti Nur AfiUrban Environmental Health: Respiratory Infection and Urban Factors in Urban Growth Corridor of Petaling Jaya, Shah Alam and Klang, Malaysia," *Sains Malaysiana,* pp. 43(9): 1405-1414, 2014.

[19] H. Ahmat, A. S. Yahaya and N. A. Ramli, "Particulate Matter Analysis for Three Industrialized Areas Using Extreme Value," *Sains Malaysiana,* pp. 44 (2): 175-185, 2015.

[20] R. Ahmad, Z. Majid, M. R. M. Yusoff, M. Z. Abdullah and A. Othman, "Air Quality of Pasir Gudang Industrial Estate," *FNEHR Conference,* 1994.

[21] G. Bohrnstedt and D. Knoke, "Norusis's SPSS 11 Chapter 22 on "Analyzing Residual:" Hamilton's Chapter on "Robust Regression"," in *Statistics for Social Data Analyis*, 1982.

[22] D. Blatna, "Outliers in Regression," University of Economic Prague, 2005.

[23] A. H. Klym, "Detecting Outliers and Influential Data Points in Receiver Operating Characteristic (ROC) Analysis," 2007.

[24] R. D. Cook and S. Weisberg, Residuals and Influence in Regression, Great Britain: Clrapnan atd HUN Ltd, 1982.

[25] S. Chatterjee and A. S. Hadi, Regression Analysis by Example, Fourth ed., New Jersey: John Wiley, 2006.

[26] D. C. Montgomery, E. A. Peck and G. G. Vining, Introduction to Linear Regression Analysis, Fifth ed., New Jersey: John Wiley, 2012.

[27] O. Gervasi, "Computational Science and Its Applications, Italy. Springer," 2008.

[28] A. G. Bluman, Elementary Statistics A Step by Step Approach, New York: McGraw Hill, 2009.

[29] Y. Sun and M. G. Genton, "Functional Bloxplot," *Journal of Computational and Graphical Statistics,* vol. 20, pp. 316-334, 2011.

[30] DoE, "Department of Environment, Malaysia. Malaysia Environmental Quality Report 2003," Department of Environment, Ministry of Natural Resources and Environment, Malaysia, Kuala Lumpur, 2004.

[31] DoE, "Department of Environment, Malaysia. Malaysia Air Quality Report 2012," Department of Environment, Ministry of Natural Resources and Environment, Malaysia, Kuala Lumpur, 2013.