© Universiti Tun Hussein Onn Malaysia Publisher's Office

IJIE

Journal homepage: http://penerbit.uthm.edu.my/ojs/index.php/ijie
ISSN : 2229-838X e-ISSN : 2600-7916

The International Journal of Integrated Engineering

# Reduction of Response Variable Influential Outliers Using MEstimation in the Next Day Prediction of Ground-Level Ozone Concentration

**Muqhlisah Muhamad[1], Ahmad Zia Ul-Saufie[1*], Sayang Mohd Deni[2]**

[1]Faculty of Computer and Mathematical Sciences,
Universiti Teknologi MARA, 13500 Permatang Pauh, Pulang Pinang, MALAYSIA

[2]Faculty of Computer and Mathematical Sciences,
Universiti Teknologi MARA, 40450 Shah Alam, Selangor, MALAYSIA

*Corresponding Author

DOI: https://doi.org/10.30880/ijie.2020.12.01.027
Received 13 March 2019; Accepted 30 October 2019; Available online 16 February 2020

**Abstract:** Ground-level ozone concentration ($O_3$) is a second significant air pollutant in Malaysia after particulate matter concentration. It is a secondary pollutant that created by photochemical reaction of primary pollutant such as volatile organic compound (VOCs) and nitrogen oxides ($NO_x$) under the influence of solar radiation (UVB). $O_3$ photochemical reactions used solar radiation with certain wavelength as the catalyst. In statistical analysis of prediction, the concentration level of $O_3$ contains the influential outliers due to several factors such as offense in data recording and sampling, the error in data acquisition or data management and the damage of monitoring instrument in data recording that can lead to misleading result or information. The objective of this study is to predict the level of $O_3$ concentration for next day (D+1) by using predictors of wind speed (WS), temperature (T), relative humidity (RH), nitric oxide (NO), sulphur dioxide ($SO_2$), nitrogen dioxide ($NO_2$), ozone ($O_3$) and carbon monoxide (CO) for selected urban area of Shah Alam by the method of minimizing influential outliers from response variable using M-estimation. The influential outliers from response variable is minimized using tuning constant approached at 95% level of efficiency. The improvement has been proved when Fair method has minimized 5.34% influential outliers from response variable and the average accuracy of the model is 0.5134.

**Keywords:** Secondary pollutant, prediction, tuning constant, concentration

## 1. Introduction

Pollution is a dirty substance that pollutes water, air, and land [1]. Air pollution is poisonous gases and trapped particles that come from primary and secondary pollutants. Air pollutants that are emitted directly from a source are considered primary pollutants, as they can be released from natural ways or human action. Secondary pollutants are those that are not directly emitted from a source but form when primary pollutants react chemically in the atmosphere.
[2].

Malaysia has a department in monitoring air pollution and environment. The Department of Environment (DoE) is the main department under the Ministry of Natural Resources and Environment, responsible to ensure a healthy and safe environment for people in Malaysia [3]. Meanwhile, Alam Sekitar Malaysia Sdn. Bhd. (ASMA) is a private agency under DoE that provides environmental solutions to the government, industries, research institutions and individuals. The DoE

*Corresponding author: ahmadzia101@uitm.edu.my has 65 monitoring stations and all stations record on an hourly basis the measurement of ozone ($O_3$), particulate matter ($PM_{10}$), sulphur dioxide ($SO_2$), carbon monoxide (CO), nitrogen dioxide ($NO_2$) and particulate matter ($PM_{2.5}$).

The formation of secondary pollutants of $O_3$ needs a photochemical reaction between the primary pollutants of nitrogen oxide ($NO_x$) and volatile organic compounds (VOCs) combined with ultraviolet sunlight (UVB) [4]. $NO_x$ refers to the molecules of nitrogen dioxide ($NO_2$) and nitric oxide (NO). The primary sources of NOx are emitted from motor vehicles and combustion processes. Meanwhile, the presence of VOCs are mainly due to industries and high traffic volume.

According to the Malaysian Department of Environment [5], the perfect combination between the conducive atmospheric condition and the emission from motor vehicles and industrial activities will result in the formation of $O_3$. The burning of hydrocarbon fuels from transport, heating from homes, factories and business, manufacturing process and power plants are the main sources of $O_3$ [6]. $O_3$ was also detected to be highly active from 7 am to 7 pm due to the presence of sunlight [7].

The other contributors to the formation of $O_3$ are CO and methane ($CH_4$). Both air pollutants of CO and $CH_4$ are emitted from man-made activities such as motor vehicles, landfills, power plant and industries facilities. At times, natural sources such as, lightning, trees and soil can also contribute to the formation of $O_3$ concentration.

In statistical analysis and air pollution, regression analysis was widely used as a tool to predict the concentrations level of some air pollutants. This study used several independent variables to form an equation in terms to predict the level of $O_3$ concentrations and the equation was known as multiple linear regression. The relationship between dependent variable and independent variables could be concluded in multiple linear regression model with several analyses [8].

However, outlier is one of common issues in developing regression model. Even a single point of outlier could distort the regression analysis and lead to incorrect inferences [9]. Outlier is a huge different observation point from another point of observation. In other words, the point of observation which is different to the general trend of the observation [10]. The presence of outliers will affect the accuracy of the model prediction in the forming of regression coefficients [11]. Every data cannot be claimed to be free from outliers [12] and most of studies did not take the assessment of outliers into their prediction consideration. Sometimes 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 outliers [13].

According to Yahaya [11] and Field [14], the whole influential outlier from the data observation could be measured using standardized residual or Cook's distance because the influential for the overall of outliers depend on the response variable. The aim of this study to approach M-estimation method from robust regression to minimize the number of influential outliers from response variable of the next day prediction (D+1) of $O_3$ concentration level in urban area of Shah Alam.

## 1.1 Variable Selection

The variables used in this study were consist of ozone ($O_3$, ppb), wind speed (WS, km/h), ambient temperature (T, °C), relative humidity (RH, %), nitric oxide (NO: ppb), nitrogen dioxide ($NO_2$, ppb), carbon monoxide (CO, ppb) and sulphur dioxide ($SO_2$, ppb) that were chose as the predictors in order to predict the level of $O_3$ concentration level for next day (D+1). The variables are selected [15-24] by the factors that associated with high contribution to the formation of $O_3$ concentration such as the conducive meteorological factor of temperature, wind speed and relative humidity when they combine with the other air pollutants.

## 2. Methodology

The purpose of this study is to develop the next day prediction model (D+1) of $O_3$ concentration level for Shah Alam using robust method by M-estimation in order to minimize the influential outliers from response variable. The procedures of this study are illustrated in Fig. 1. Before the robust regression model developed, the assessment of the outliers will be conducted to identify the influential outliers from response variable using Cook's distance and standardized residual.

Thus, M-estimation method will be used to reduce the contamination of the influential outliers from response variable using tuning constant in developing the prediction model. Tuning constant controls how sharp M-estimation as an outliers detector that contaminated response variable data. Nine method from M-estimation has been introduced such as Huber, Andrew, Bisquare, Cauchy, Fair, Talwar, Logistic, Welsch and Hampel. This method will be compared with classical method of ordinary least square in order to determine a better model to be used in the prediction of $O_3$ concentration level.

## 2.1 Site Selection

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. Besides, Shah Alam city is located between Petaling Jaya city (east) and Klang town (west). Shah Alam station is selected due to the highest level of $O_3$ concentration in Malaysia according to the rising number of registered mobile vehicles in Shah Alam throughout 2003 [25] and commencing on 2003, the increased burning of industrial waste including from hotels, commercial centres, institutions and night markets tend to produce large emission of $NO_x$ and VOCs which are the main element of $O_3$ formation [26].

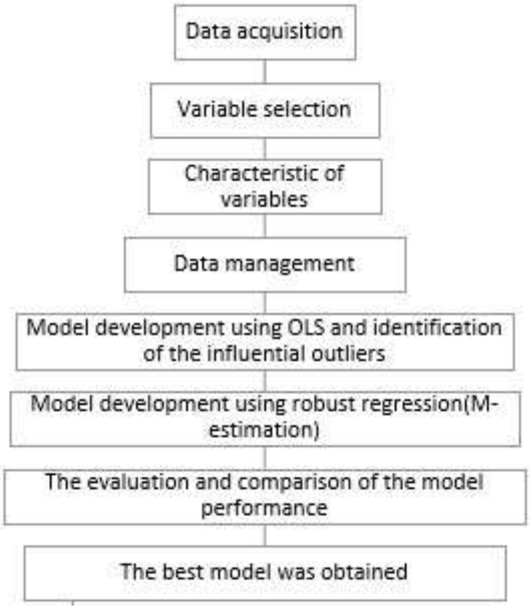

Data acquisition
Variable selection
Characteristic of variables
Data management
Model development using OLS and identification of the influential outliers
Model development using robust regression(M-estimation)
The evaluation and comparison of the model performance
The best model was obtained

**Fig. 1 - Research flow**

## 2.2 Data Acquisition

In air pollution monitoring procedures, United States Protection Agency (EPA) standardized some guidelines to measure air pollutants and meteorological variables [27]. The air pollutant and meteorological variables were monitored by Teledyne Ozone Analyzer Model 400A UV Absorption ($O_3$), Teledyne Model 200A (NO and $NO_2$), Teledyne Model 100A ($SO_2$), Teledyne Model 300 (CO), Met One 010C Sensor (WS), Met One 062 Sensor (T) and Met One 083D Sensor (RH) [28]. The primary data was managed by Alam Sekitar Malaysia Sendirian Berhad (ASMA), which is the private company under supervision of Department of Environmental Malaysia (DoE). The secondary data from 1st January 2002 until 31st December 2012 were obtained from Department of Environmental Malaysia (DoE).

## 2.3 Data Management

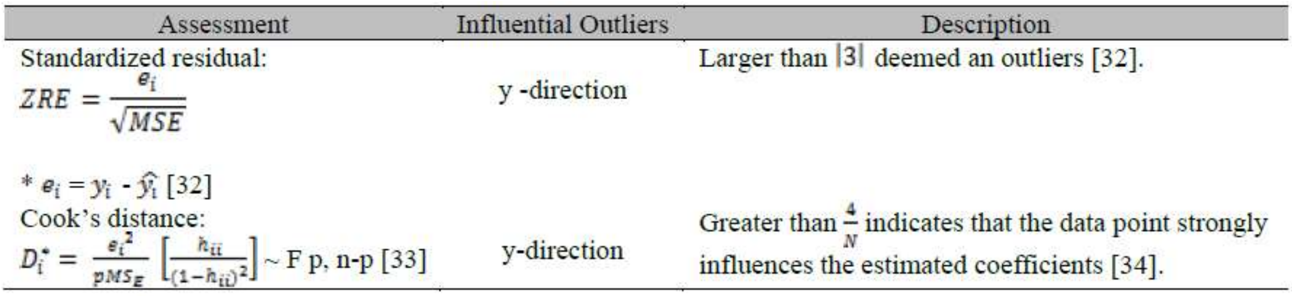
In this study, the hourly concentrations for each variable selected were transformed into daily 12 hours average concentration, from 7am to 7pm because the level of $O_3$ concentration level was suspected to be highly active during morning and evening [29]. According to Mohammed, Ramli, and Yahaya [7], most of the areas have a large emission of $O_3$ formation factor from morning till evening. This study only used 12 hours average ozone concentrations from 7 am until 7 pm to predict the next day (from 7 am until 7 pm) ozone concentrations.

Missing values is one of the problems in the process of data acquisition that may lead to the interrupted analysis. The offence in data recording and sampling, the failure of machine and human error are the several reasons that contributed to the missing values observation [30]. One of the suitable methods in the imputation of missing values for air pollution data is by using the mean imputation technique as suggested by [31]. The missing values will be replaced by the mean obtained between the above value and the below value also known as mean above below method (MAB) as followed by Noor et al. [30]. The data from 2002 until 2012 were randomized into 80% and 20% as suggested by [12] where 80% of the data were used for training while the other 20% were used for validation.

## 2.4 Influential Outliers Identification

The influential point of outliers is the case where there is existing larger residual that differs substantially from the other observations [32]. Influential outliers are any point that has a large effect on the analysis of regression. According to Sarkar, Midi, and Rana [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 showed that the data before was influenced by outliers. This study only considers the influential outliers from response variable by computing standardized residual and Cook's distance assessment as shown in Table 1.

**Table 1 - The assessment of the influential outliers from response variable**



| Assessment | Influential Outliers | Description |
|---|---|---|
| Standardized residual:<br>$ZRE = \frac{e_i}{\sqrt{MSE}}$<br>* $e_i = y_i - \hat{y}_i$ [32] | y -direction | Larger than \|3\| deemed an outliers [32]. |
| Cook's distance:<br>$D_i^* = \frac{e_i^2}{pMS_E}\left[\frac{h_{ii}}{(1-h_{ii})^2}\right]$ ~ F p, n-p [33] | y-direction | Greater than $\frac{4}{N}$ indicates that the data point strongly influences the estimated coefficients [34]. |

## 2.5 M-estimation

Huber [35] introduced M-estimation as the simplest method in the detection of outliers from response variable. Mestimation is an extension from the maximum likelihood estimation where the main principle in M-estimation is to minimize the residual function of weighting function and the steps of M-estimation are shown as follows [36] and [37],

- Step 1: Test assumptions of ordinary least square.
- Step 2: Detect the presence of outliers in the data.
- Step 3: Calculate regression coefficient ($\beta_i$) with ordinary least square.
- Step 4: Calculate initial residual value:

$$e_i = y_i - \hat{y}_i \tag{1}$$

- Step 5: Calculate initial value of standard deviation:

$$\hat{\sigma}_i = 1.4826\ MAD \tag{2}$$

- Step 6: Calculate value:

$$u_i = \frac{e_i}{\hat{\sigma}_i} \tag{3}$$

- Step 7: Calculate the weighted value of Bisquares (Tukey):

$$\omega_i = \begin{cases} \left[1 - \left(\frac{u_i}{4.685}\right)^2\right]^2, & |u_i| \le 4.685; \\ 0 & ,|u_i| > 4.685. \end{cases} \tag{4}$$

- Step 8: Calculate $\beta_i$ using weighted least square method with weighted $\omega_i$.

At the first step, each of the regression model for $O_3$ (D+1, D+2 and D+3) was checked to ensure the test assumptions in ordinary least square (OLS) were satisfied. The presence of the outliers were detected using standardized residual and Cook's distance at step two. Then, the regression coefficient $(\beta_i)$ was calculated to obtain the predicted value ($\hat{y}_i$) for $O_3$ using the method of OLS estimate. In order to obtain regression coefficients $(\beta_i)$ from Mestimation, the estimator for standard deviation that we denoted as sigma $(\hat{\sigma}_i)$ is rescaled to median absolute deviation (MAD) by the factor 1.4826 at step 5, where 1.4826 is the value when the residual is normally distributed and the sample is large.

From step 6, the proportion of the residual from the estimated scale of standard deviation ($u_i$) was obtained to be used in the Bisquare weighting function (step 7) where 4.685 represents the value of tuning constant of Bisquare at 95% level of efficiency. Step 4 until 7 were repeated until the value of regression coefficient $(\beta_i)$ in step 8 converged. The same steps were applied in other weighting function (Huber, Andrew, Cauchy, Fair, Talwar, Logistic, Welsch and Hampel) to obtain the value of regression. Table 2 shows all weighting function used in M-estimation.

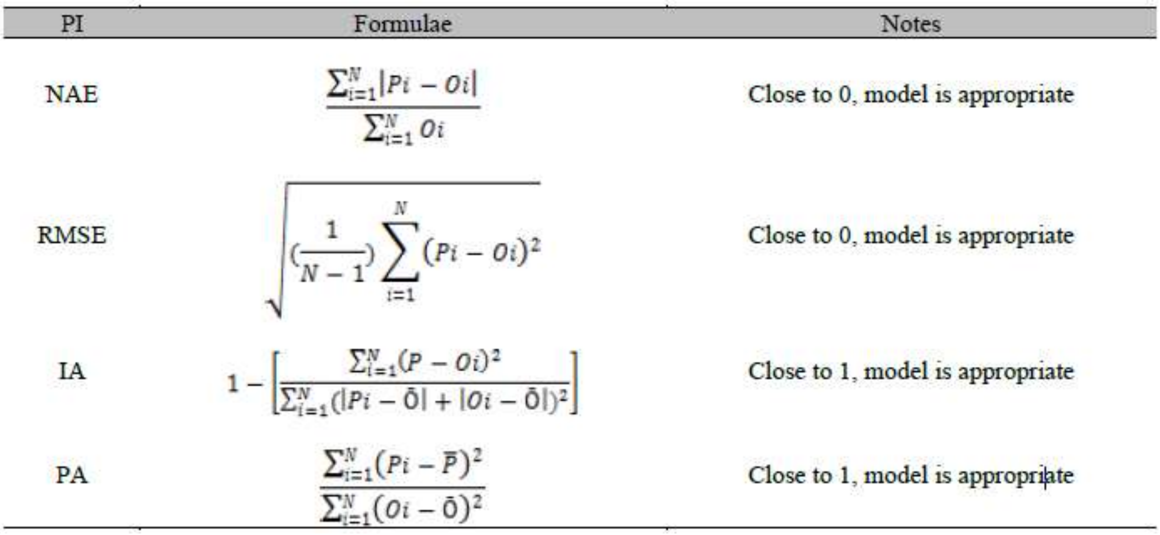
## 2.6 Performance Indicator

Performance indicators are used to evaluate the performance and the adequacy of the models. The performance indicators (Table 3) are consists of normalized absolute error (NAE), Root Mean Square Error (RMSE), Index of Agreement (IA) and prediction accuracy (PA). The best prediction model will be obtained by comparing the performance indicators between the models.

**Table 2 - Weighting function of M-estimation [38]**

| Weighting Function | ρ ( $r$ ) | ψ ( $r$ ) |
|---|---|---|
| Huber | $\frac{r^2}{2}, \|r\| \le c$<br>$c\|r\| - \frac{r^2}{2}, \|r\| > c$ | $r, \|r\| \le c$<br>$sign\,(r)c$ |
| Andrew | $c^2\left[1 - cos\left(\frac{r}{c}\right)\right], \|r\| \le c$<br>$c^2, \|r\| > c$ | $c \sin\left(\frac{r}{c}\right), \|r\| \le c$<br>$0, \|r\| > c$ |
| Bisquare | $\frac{c^2}{6}\{1 - \left[1 - \left(\frac{r}{c}\right)^2\right]^3, \|r\| \le c$<br>$\frac{c^2}{6}, \|r\| < c$ | $3r\left[1 - \left(\frac{r}{c}\right)^2\right]^2, \|r\| \le c$<br>$0, \|r\| > c$ |
| Cauchy | $\left(\frac{c^2}{2}\right) log\left[1 + \left(\frac{r}{c}\right)^2\right]$ | $r\left[1 + \left(\frac{r}{c}\right)^2\right]^{-1}$ |
| Fair | $c\left[\|r\|c - log\left(1 + \frac{\|r\|}{c}\right)\right]$ | $r\left(1 + \frac{\|r\|}{c}\right)^{-1}$ |
| Logistic | $c^2 log\left[cosh\left(\frac{r}{c}\right)\right]$ | $c\ tan\ h\left(\frac{r}{c}\right)$ |
| Talwar | $\frac{r^2}{2}, \|r\| \le c$<br>$\frac{c^2}{2}, \|r\| > c$ | $r, \|r\| \le c$<br>$0, \|r\| > c$ |
| Welsch | $\frac{c^2}{2}\left[1 - exp\left[-\left(\frac{r}{c}\right)^2\right]\right]$ | $\left[r\ exp\left[-\left(\frac{r}{c}\right)^2\right]\right]$ |
| Hampel | $\frac{1}{2}r^2\ if\ \|r\| < a$<br>$a\|r\| - \frac{1}{2}a^2\ if\ a \le \|r\| < b$<br>$a\frac{c\|r\| - \frac{1}{2}r^2}{c - b} - \frac{7a^2}{6}\ if\ b \le \|r\| \le c$<br>$a(b + c - a), otherwise$ | $r\ if\ \|r\| < a$<br>$a\ sign\ r\ if\ a \le \|r\| < b$<br>$a\frac{c\ sign\ r - r}{c - b}\ if\ b \le \|r\| \le c$<br>$0, otherwise$ |

Note: ψ ( ) = ρ’ , = residual and a, b, c = tuning constant.



| PI | Formulae | Notes |
|---|---|---|
| NAE | $\frac{\sum_{i=1}^{N}\lvert Pi - Oi\rvert}{\sum_{i=1}^{N} Oi}$ | Close to 0, model is appropriate |
| RMSE | $\sqrt{(\frac{1}{N-1})\sum_{i=1}^{N}(Pi - Oi)^2}$ | Close to 0, model is appropriate |
| IA | $1 - \left[\frac{\sum_{i=1}^{N}(P - Oi)^2}{\sum_{i=1}^{N}(\lvert Pi - \bar{O}\rvert + \lvert Oi - \bar{O}\rvert)^2}\right]$ | Close to 1, model is appropriate |
| PA | $\frac{\sum_{i=1}^{N}(Pi - \bar{P})^2}{\sum_{i=1}^{N}(Oi - \bar{O})^2}$ | Close to 1, model is appropriate |

**Table 3 - Performance indicator (PI) [39]**

Note: N = Number of sample daily measurement of a selected sites, $\hat{y}_i$ = Predicted value of one set daily data,
$y_i$ = Observed values of one set daily data, $\bar{\hat{y}}_i$ = Mean of the predicted values of one set daily data
$\bar{y}_i$ = Mean of the observed values of one set daily data

# 3. Results and Discussion

## 3.1 Air Pollutants Characteristics

Shah Alam was located at a busy main transportation area and thus exposed to traffic congestion. Almost daily sees Shah Alam polluted by 32.421 ppb of $O_3$ concentration. The reading of $O_3$ concentration was very high on 15th March 2003 at 97.00 ppb. This is due to the rising number of registered mobile vehicles in Shah Alam throughout 2003 [25], and that same year saw the increased burning of industrial waste including from hotels, commercial centres, institutions and night markets [26]. Table 4 summarizes the characteristics of $O_3$ concentrations level for Shah Alam.

**Table 4 - Summarizes the characteristics of $O_3$ concentrations level for Shah Alam**

| Shah Alam | $O_3$ |
|---|---|
| Mean | 32.421 |
| Median | 31.250 |
| Mode | 32.421 |
| Standard Deviation | 11.441 |
| Variance | 130.907 |
| Skewness | 0.657 |
| Kurtosis | 1.030 |
| Maximum | 97.000 |
| Percentiles 95% | 52.750 |
| Percentiles 99% | 65.709 |

Following up to these scenes, the formation of $O_3$ become active due to the high presence of $NO_x$ (108.330 ppb), CO (6214.167 ppb) and VOCs from the vehicles and industrial burning respectively. The measurement of dispersion showed that the level of $O_3$ concentrations for Shah Alam is moderately skewed when the skewness value is 0.657) and this tended to have a standard deviation with value 11.441. Besides, the distribution for $O_3$ concentrations has a positive kurtosis with value 1.030.

The percentage of outliers for $O_3$ in Shah Alam was described using box and whisker plot. The points that were out of the box and whisker plot indicated that the observation of the outliers. The percentage of outliers suspected for Shah Alam and is 1.69 as shown in Fig. 2.

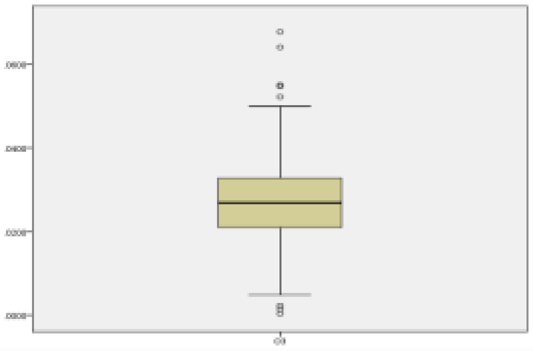


**Fig. 2 - Box and Whisker plot for $O_3$ concentration level**

## 3.2 Ordinary Least Square

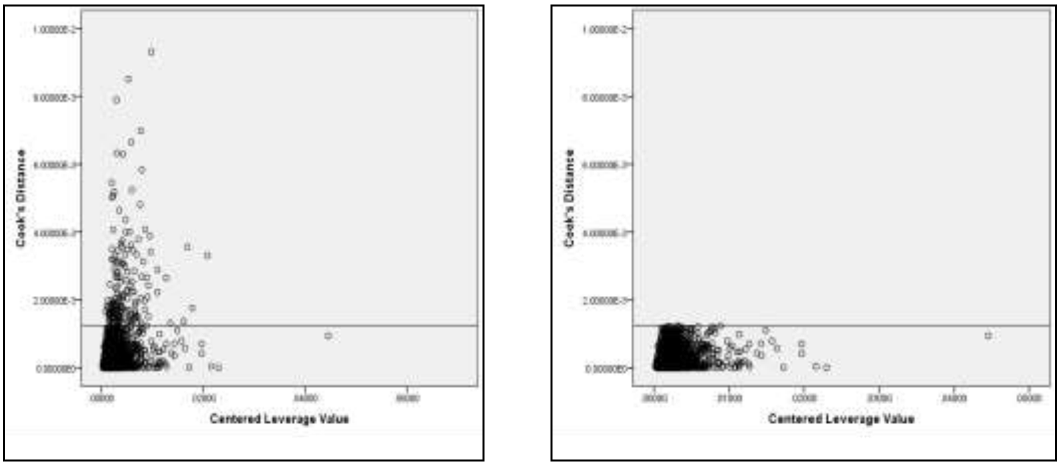
Since all of the assumption of ordinary least square estimate have been fulfilled, thus the multiple linear regression model for next day (D+1) prediction has been developed as show in Table 4. The multiple linear regression model in Table 5 are evaluated by performance indicators. The average error is 5.488 and the average accuracy is 0.5127. Meanwhile, the model show that 20.56% of the total variation in $O_{3,\ D+1}$ is explained by the regression line using the predictors.

**Table 5 - Multiple linear regression model and performance indicator**

| Model | Description of the Model |
|---|---|
| $O_{3,\ D+1} = 62.009500 + 0.258307WS - 1.100540T - 0.214758RH + 0.012961NO - 0.068383SO_2 + 0.069568NO_2 + 0.437523O_3 + 0.001931CO$ | NAE = 0.244080<br>RMSE = 10.731951<br>IA = 0.571453<br>PA = 0.454015<br>$R^2$ (100%) = 20.56 |

## 3.3 Influential Outliers Identification

The influential outliers from response variable was identifed at 172 observation after compute the assessment of standardized residual and Cook's distance. Furthermore, the influence of the outliers from response variable could be illustrated using scatter plot of Cook's distance against centered leverage value. The point of observation that was above the line indicates the observation of influential outliers from y-direction after conducting the assessment of standardized residual and Cook's distance. After removing the outliers from response variable, the uncontaminated data from response variable is remained below the horizontal line of Cook's as in Fig. 3.



(a) The influential outliers from response variable

(b) After remove the influential outliers from response variable

**Fig. 3 - Box and Whisker plot for $O_3$ concentration level**

## 3.4 M-Estimation Models

The models developed by M-estimation use tuning constant to detect the outliers from response variable. Tuning constant is located in the weighting function of Huber, Andrew, Bisquare, Cauchy, Fair, Logistic, Talwar, Welsch and Hampel. The tuning constant has an important role in determining the regression coefficient in order to develop the model in the prediction of $O_3$ concentrations level for next day (D+1). Table 6 shows the models developed to predict the level of $O_3$ concentrations in Shah Alam using the nine M-estimation methods.

Then, each next day (D+1) prediction model of $O_3$ concentrations level from M-estimation was evaluated using performance indicators. The lower error (NAE and RMSE) and the higher accuracy (IA and PA) indicated that the model is appropriate. In order to determine a good model among these nine methods, the ranking method of performance indicators [40] was conducted where the error of NAE value and RMSE value are ranked in increasing order (1 = the smallest value of error to 9 = the largest value of error) and the accuracy of IA value and PA value are ranked in decreasing order (1 = the largest value of accuracy to 9 = the smallest value of accuracy). Hence, a good Mestimation method was found by the smallest summation from ranking the performance indicators of each method as shown in Table 7.

The ranking with the smallest summation (Table 8) which is the Fair method is seen as a good M-estimation model to predict the level of $O_3$ concentrations in Shah Alam for next day (D+1). The comparison of the result between Mestimation (Fair) and ordinary least square method was shown in Table 9.

**Table 6 - Robust regression model by m-estimation for Shah Alam next day prediction (D+1)**

| Method | Model for Shah Alam |
|---|---|
| Huber<br>(1.345) | $O_{3,\ D+1} = 58.34081 + 0.218359WS - 0.999551T - 0.201683RH - 0.003597NO - 0.080723SO_2 + 0.071342NO_2 + 0.427896O_3 + 0.002304CO$ |
| Andrew<br>(1.3390 | $O_{3,\ D+1} = 59.29342 + 0.211954WS - 1.016579T - 0.206885RH - 0.008400NO - 0.080975SO_2 + 0.068869NO_2 + 0.424208O_3 + 0.002495CO$ |
| Bisquare<br>(4.685) | $O_{3,\ D+1} = 59.28020 + 0.211684WS - 1.01620T - 0.206799RH -0.008292NO - 0.080967SO_2 + 0.068917NO_2 + 0.424289O_3 + 0.002487CO$ |
| Cauchy<br>(2.385) | $O_{3,\ D+1} = 58.50722 + 0.215841WS - 1.01235T - 0.200501RH - 0.004714NO - 0.078353SO_2 + 0.070814NO_2 + 0.431530O_3 + 0.002348CO$ |
| Fair<br>(1.400) | $O_{3,\ D+1} = 58.33889 + 0.220274WS - 1.02206T - 0.197090RH -0.002412NO - 0.075385SO_2 + 0.071074NO_2 + 0.438312O_3 + 0.002279CO$ |
| Logistic<br>(1.205) | $O_{3,\ D+1} = 58.35987 + 0.217172WS - 1.01374T - 0.198907RH - 0.003527NO - 0.077309SO_2 + 0.071205NO_2 + 0.434015O_3 + 0.002310CO$ |
| Talwar<br>(2.975) | $O_{3,\ D+1} = 61.93210 + 0.192428WS - 1.04524T - 0.225391RH -.003655NO - 0.063148SO_2 + 0.058234NO_2 + 0.414953O_3 + 0.002605CO$ |
| Welsch<br>(2.985) | $O_{3,\ D+1} = 58.929651 + 0.213073WS - 1.013085T - 0.204147RH - 0.007236NO - 0.080173SO_2 + 0.069647NO_2 + 0.426784O_3 + 0.002440CO$ |
| Hampel<br>(2,4 and 5) | $O_{3,\ D+1} = 61.60702 + 0.235683WS -1.07250T - 0.216456RH + 0.001681NO - 0.075576SO_2 + 0.068695NO_2 + 0.427674O_3 + 0.002221CO$ |

**Table 7 - Performance Indicators of M-estimation for Shah Alam Next Day Prediction Model (D+1)**

| Method | NAE | RMSE | IA | PA |
|---|---|---|---|---|
| Huber | 0.244177 | 10.755146 | 0.569628 | 0.453377 |
| Andrew | 0.244382 | 10.762230 | 0.568863 | 0.452901 |
| Bisquare | 0.244377 | 10.762060 | 0.568862 | 0.452915 |
| Cauchy | 0.244172 | 10.756848 | 0.571193 | 0.453215 |
| Fair | 0.244047 | 10.753994 | 0.573525 | 0.453260 |
| Logistic | 0.244112 | 10.755345 | 0.572023 | 0.453286 |
| Talwar | 0.244590 | 10.753717 | 0.566154 | 0.453188 |
| Welsch | 0.244302 | 10.760139 | 0.569656 | 0.453013 |
| Hampel | 0.244248 | 10.743702 | 0.568980 | 0.453509 |

**Table 8 - The ranking of performance indicators of M-estimation for Shah Alam next prediction (D+1)**

| Method | NAE | RMSE | IA | PA | Sum |
|---|---|---|---|---|---|
| Huber | 4 | 4 | 5 | 2 | 15 |
| Andrew | 8 | 9 | 7 | 9 | 33 |
| Bisquare | 7 | 8 | 8 | 5 | 28 |
| Cauchy | 3 | 6 | 3 | 8 | 20 |
| **Fair** | **1** | **3** | **1** | **4** | **9** |
| Logistic | 2 | 5 | 2 | 3 | 12 |
| Talwar | 9 | 2 | 9 | 6 | 26 |
| Welsch | 6 | 7 | 4 | 7 | 24 |
| Hampel | 5 | 1 | 6 | 1 | 13 |

**Table 9 - The comparison between ordinary least square method and M-estimation**

| Method | NAE | RMSE | IA | PA |
|---|---|---|---|---|
| OLS | 0.244080 | 10.731951 | 0.571453 | 0.454015 |
| Fair | 0.244047 | 10.753994 | 0.573525 | 0.453260 |

## 4. Conclusion

This study has proved that robust method is better than ordinary least square method since the influential outliers in air pollution data have been reduced using the weightage approached. The average accuracy of M-estimation by Fair method is 0.5134 which is better than ordinary least square method where the average accuracy is 0.5129. Besides, the improvement has been proved when Fair method has minimized 5.34% influential outliers from response variable. Therefore, these models could be implemented among heath public health, government, citizen and the other authorities to prepare and can take an early action to avoid the negative impact of $O_3$ concentration.

## Acknowledgement

This study was funded by 600-RMI/FRGS 5/3 (40/2014). We would also like to extend our appreciation to the Department of Environmental Malaysia (DoE) for providing the air quality data for this research.

## References

[1] H. Collins, Collin COBUILD Advanced Learner's English Dictionary, Great Britain, 2006.

[2] S. Azmi, Isu Alam Sekitar di Malaysia (Ancaman Alam dan Atmosfera), Kuala Lumpur, 2007.

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

[4] DoE, "Department of Environment, Malaysia. Malaysia Environmental Quality Report 2013," Department of Environment, Ministry of Natural Resources and Environment, Malaysia, Kuala Lumpur, 2014.

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

[6] R. Mozer, "Ground Level Ozone," 2008.

[7] 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.

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

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

[11] A. S. Yahaya, "Applied Regression Models Using SPSS," 2014.

[12] 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.

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

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

[15] E. Agirre-Basurko, G. I. Berastegi and I. Madariaga, "Regression and Multilayer Perceptron-Based Models to Forecast Hourly Ozone and Nitrogen Dioxides levels in the Bilbao area," Environmental Modelling & Software, 2006.

[16] M. Musa, A. A. Jemain and W. Z. Wan Zin, "Scaling and Persistence of Ozone Concentrations in Klang Valley," Journal of Quality Measurement and Analysis, pp. 9(1): 9-20, 2013.

[17] K. Jaioun, K. Saithanu and J. Mekparyup, "Multiple Linear Regression Model to Estimate Ozone Concentration in Chonburi, Thailand," International Journal of Applied Environmental Sciences, pp. 4: 1305-1308, 2014.

[18] W. Wang, W. Lu, X. Wang and A. Y. Leung, "Prediction of Maximum Daily Ozone Level Using Combined Neural Network and Statistical Characteristics," Environmental International, pp. 29: 555-562, 2003.

[19] 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.

[20] J.-S. Heo, K.-H. Kim and D.-S. Kim, "Pattern Recognition of high Ozone Episodes in Forecasting Daily Maximum Ozone Levels," TAO, vol. 15, pp. 199-220, 2004.

[21] A. W. Delcloo and H. d. Backer, "Modelling Planetary Boundary Layer Ozone, Using Meteorological Parameters at Uccle and Payerne," Atmospheric Environment, pp. 39: 5067-5077, 2005.

[22] N. A. Ramli, N. A. Ghazali and A. S. Yahaya, "Diurnal Fluctuations of Ozone Concentrations and its Precursors and Prediction of Ozone Using Multiple Linear Regression," Malaysian Journal of Environmental Management, pp. 11(2): 57-69, 2010.

[23] N. Banan, M. T. Latif, L. Juneng and M. F. Khan, "An Application of Artificial Neural Network for the Prediction of Surface Ozone Concentrationns in Malaysia," Springer Science, 2014.

[24] U. Schlink, M. Richter, S. Dorling, G. Nunnari, G. Cawley and E. Pelikan, "Statistical Models to Assess the Health Effect and to Forecast Ground-Level Ozone," Environmental Modelling & Software, pp. 21: 547-558, 2006.

[25] S. H. M. Shafie and M. Mahmud, "Analisis Pola Taburan Reruang PM10 dan O3 di Lembah Klang dengan Mengaplikasi Teknik Geographic Information System," Malaysian Journal of Society and Space, vol. 11, no. 3, pp. 61-73, 2015.

[26] DoE, Department of Environment, Malaysia. Malaysia Air Quality Report 2011, Kuala Lumpur: Department of Environment, Ministry of Natural Resources and Environment, Malaysia, 2012.

[27] F. Ahamad, M. T. Latif, R. Tang, L. D. Juneng and H. Juahir, "Variation of Surface Ozone Exceedane Around Klang Valley, Malaysia," Atmospheric Research, 2014.

[28] 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.

[29] N. M. Noor, A. S. Yahaya and N. A. Ramli, "Estimation of Missing Values for Air Pollution Data Using Mean Imputation Techniques," 2008.

[30] A. S. Yahaya, N. A. Ramli and N. Fitri, "Effects of Estimating Missing Values on Fitting Distribution: International Conference on Quantitative Sciences and Its Applications," 2005.

[31] 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.

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

[33] D. L. Stevens, Sampling Design and Statistical Analysis Methods for the Integrated Biological and Physical Monitoring of Oregon Streams, Corvalis, Oregon, 2002.

[34] P. J. Huber, Robust Regression: Asymptotics, Conjectures and Monte Carlo, Ann. Stat., 1973, pp. 1, 799-821

[35] Y. Susanti, H. Pratiwi, S. Sulistijowati H and L. Twenty, "M Estimation, S Estimation, and MM Estimation in Robust Regression," International Journal of Pure and Applied Mathematics, 2014.

[36] C. Stuart, Robust Regression, 2011.

[37] C. Chen and G. Yin, "Computing the Efficiency and Tuning Constant for M-Estimation," Joint Statistical Meetings-Statistical Computing Section, 2002.

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

[39] N. I. Mohammed, "Developement and Assessment of New AOTX Models for Ozone Phytotoxicity Effect on Paddy Yield Reductions Malaysia," 2012.