International Journal of Engineering & Technology
CENTRE OF PROFESSIONAL RESEARCH PUBLICATIONS
ISSN 2049-3444

International Journal of Engineering and Technology Volume 2 No. 3, March, 2012

Robust Regression Models for Predicting PM_{10} Concentration in an Industrial Area

Ahmad Zia Ul-Saufie[1,2], Ahmad Shukri Yahaya[1], Nor Azam Ramli[1] , Hazrul Abdul Hamid[1,]
[1]School of Civil Engineering, Universiti Sains Malaysia, Pulau Pinang, Malaysia
[2]Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Malaysia

ABSTRACT

Particulate Matter (PM) is an air pollutant consisting of a mixture of solid and liquid particles suspended in the air with diameter less than or equal to 10 micrometers (PM_{10}). It can cause significant health effects, particularly among the elderly and infants, people with asthma and other respiratory diseases. The aim of this study is to determine the best robust regression models for future prediction of PM_{10} concentration in Pulau Pinang, Malaysia. Robust method is less sensitive than ordinary least squares (OLS) to large changes in small parts of the data. Robust regression works by assigning a weight to each data point. The weighting functions used in this study are Huber, Andrews, Bisquare, Cauchy, Fair, Logistic, Talwar, Welsch and OLS. Model comparison statistics using Prediction Accuracy (PA), Coefficient of Determination (R^2), Index of Agreement (IA) , Normalised Absolute Error (NAE) and Root Mean Square Error (RMSE) show that Fair is the best weighting function for next day (RMSE =11.077, NAE= 0.122, PA=0.927, IA = 0.961, R^2=0.858,) and next 2-day (RMSE = 14.153, NAE= 0.122, PA=0.927, IA = 0.961, R^2=0.773) prediction while Cauchy is the best for next 3-day (RMSE = 16.012, NAE= 0.122, PA=0.927, IA = 0.961, R^2=0.718). Performance indicators showed that the developed robust regression models can be used for long term prediction of PM_{10}.

Keywords: *Particulate Matter (PM_{10}), Robust Regression, Performance indicators, Ordinary Least Squares, Weighted Least Square.*

1. INTRODUCTION

The air quality status in Malaysia are based on the air quality monitoring in several cities for the hourly concentration of carbon monoxide (CO), sulphur dioxide (SO_2), nitrogen dioxide(NO_2), ozone (O_3) and particulate matter less than or equal 10 µm (PM_{10}). The aim of this study is to determine the best robust regression models for long term prediction of PM_{10} concentration based on meteorological parameters and gaseous concentration (SO_2, NO_2 and CO) in Pulau Pinang, Malaysia. PM_{10} are chosen because it is known to have significant impacts on the environment and human health [1].

According to Department of Environment, Malaysia [2], PM_{10} are emitted from motor vehicle exhausts, heat and power generation plants, industrial process and open burning activities. In Malaysia, high concentration level of particulate matter in the cities are closely linked with emission from vehicles [3]. The two major sources of PM_{10} in Malaysia are industries and heavy traffics. [4] indicated that traffic emission especially from lorries and diesel pick-up transporting agricultural goods from and into the town (Kuala Lumpur) have major influence on PM_{10} loadings. According to [5], particule matter in Malaysia have four major sources such as industrial and development, motor vehicle, open burning and power generation.

The presence of PM_{10} in atmosphere can cause severe health impacts to human such as allergies, asthma, respiratory related illnesses, nose and throat irritations, premature mortality and more severe impacts such as hospital admission [6] ;[7];[8];[9]. [4] found that PM_{10} may also bring negative impacts on the growth and productivity of small and short cycle plant species such as vegetables.

Statistical modelling could offer good insights in short term predicting of future air pollution levels (next day, next 2-day and next 3-day), hence allowing local environmental authorities to carry out daily air pollution forecasts. The public health advisors can use this information make decisions regarding air pollution abatement measures.

Regression analysis is an important statistical tool applied in most sciences studies. Ordinary least square (OLS) method has been generally adopted because of tradition and ease of computation [10]. But, real data sets frequently contain outliers, which may cause gross errors and exceptional observations. In such cases, OLS methods become unreliable. Hence robust regression can act as a complement to OLS method because robust regression provide answers similar to least square regression when the data are linear and have normally distributed errors [11]. The result will differ significantly when the error are not normally distributed or when the data contain significant outlier.

This paper discussed robust regression methods (weighted least square) for predicting PM_{10} concentration in an industrial area in Pulau Pinang because it is among the most densed populated states in Malaysia with 1490 persons per square kilometre [12].

2. Methodology

2.1 Site Description

Pulau Pinang is a state in Malaysia and located on the northwest coast of Peninsular Malaysia by the Strait of Malacca (Figure 1). Pulau Pinang State consists of two parts, Pulau Pinang Island and mainland Seberang Perai. The island has an area of 285 km^2 and is connected to Seberang Perai by ferry as well as a 13.5 km long Pulau Pinang bridge. The state has the highest population density in Malaysia. The whole of Pulau Pinang state has a density of 1,490 people per square kilometre and has a population of 1,596,900 as of 2010 [12]. Seberang Perai is the hinterland portion of Pulau Pinang state populated by 815,767 people in the 2010 census, and has a density 1,086 people per square kilometre [13]. Figure 2 shows the populations density in Pulau Pinang.

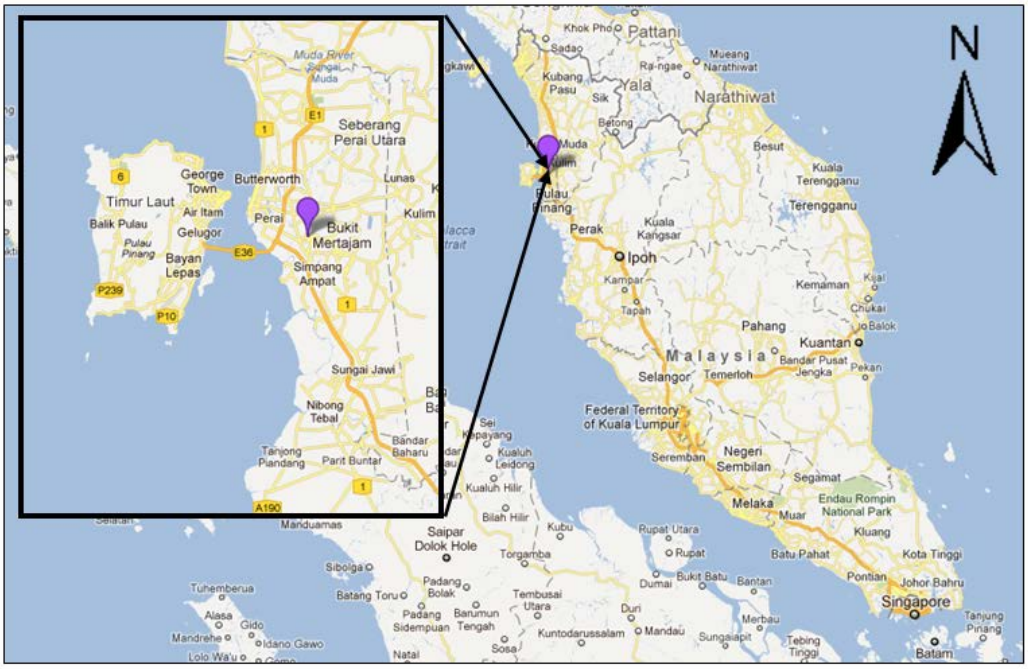


Figure 1. Site Location (Source: Adapted from Google Map, 2011)

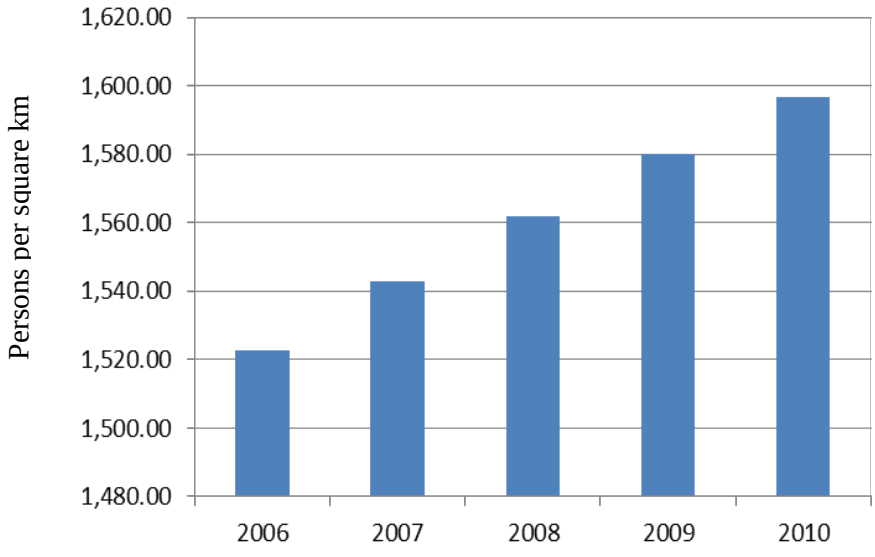


Figure 2. Population density in Pulau Pinang (Department of Statistics, Malaysia, 2010)

2.2 Robust Regression

Robust regression works by assigning a weight to each data point. The weighting functions used in this study are 'Huber', 'Andrews', 'Bisquare', 'Cauchy', 'Fair', 'Logistic', 'Talwar', 'Welsch' and 'OLS' using the command robustfit in MATLAB 2010a. The algorithm uses iteratively reweighted least squares with a bisquare weighting function as default. Tune is a tuning constant that is divided into the residual vector before computing weights. The weighting function can be any one of the following strings (Table 1):

Table 1: Weighting Function Equations for Robust Regression (MATLAB, 2009a)

Weight Function	Equation	Default Tuning Constant
'andrews'	w = (abs(r)<pi)* sin(r)/ r	1.339
'bisquare' (default)	w = (abs(r)<1)* (1 - r^2)^2	4.685
'cauchy'	w = 1/ (1 + r^2)	2.385
'fair'	w = 1/ (1 + abs(r))	1.400
'huber'	w = 1/ max(1, abs(r))	1.345
'logistic'	w = tanh(r)/ r	1.205
'ols'	no weighting function	None
'talwar'	w = 1 * (abs(r)<1)	2.795
'welsch'	w = exp(-(r^2))	2.985

In this study, tune is unspecified and the default value in the Table 1 is used. Default tuning constants give coefficient estimates that are approximately 95% as statistically efficient as the ordinary least-squares estimates, provided the response has a normal distribution with no outliers. The value *r* in the weight functions is

$$r = resid/(tune*s*sqrt(1-h)) \quad (1)$$

where *resid* is the vector of residuals from the previous iteration, *h* is the vector of leverage values from a least-squares fit, and *s* is an estimate of the standard deviation of the error given by

$$s = MAD/0.6745 \quad (2)$$

Here *MAD* is the median absolute deviation of the residuals from their median. The constant 0.6745 makes the estimate unbiased for the normal distribution [14].

2.3 Performance Indicators

Performance indicators were used to evaluate the goodness of fit for the robust regression to determine which weighting function is appropriate to represent the PM_{10} concentration in Seberang Prai, Pulau Pinang. Performance indicators that are used to determine the best method for predicting PM_{10} concentration are normalized absolute error (NAE), root mean square error (RMSE), index of agreement (IA), prediction accuracy (PA), and coefficient of determination (R^2). The equations used are previously reported by [15].

Table 2. Performance Indicators [15]

Performance index	Equation	Description
Mean absolute error (MAE)	$MAE = \frac{\sum_{i=1}^{n} \lvert P_i - O_i \rvert}{n}$	MAE value closer to zero indicates better method
Normalized absolute error (NAE)	$NAE = \frac{\sum_{i=1}^{n} Abs(P_i - O_i)}{\sum_{i=1}^{n} O_i}$	NAE value closer to zero indicates better method.
Index of agreement	$IA = 1 - \left[\frac{\sum_{i=1}^{n} (P - O_i)^2}{\sum_{i=1}^{n} \left(\lvert P_i - \overline{O} \rvert + \lvert O_i - \overline{O} \rvert \right)^2} \right]$	IA value closer to 1 indicates better method.

Prediction accuracy	$PA = \frac{\sum_{i=1}^{n}(P_i - \overline{O})^2}{\sum_{i=1}^{n}(O_i - \overline{O})^2}$	PA value closer to 1 indicates better method
Coefficient of determination (R^2)	$R^2 = \left(\frac{\sum_{i=1}^{n}(P_i - \overline{P})(O_i - \overline{O})}{n.S_{pred}.S_{obs}} \right)^2$	R^2 value closer to 1 indicates better method

Where n = total number of annual measurements of a particular site, P_i = predicted values of one set of annual monitoring record, O_i= observed values of one set of annual monitoring record, $\bar{P}$= mean of the predicted values of one set of annual monitoring record, $\bar{O}$= mean of the observed values of one set of annual monitoring record, S_{pred}= standard deviation of the predicted values of one set of annual monitoring record, and S_{obs}= standard deviation of the observed values of one set of annual monitoring record between input and output vectors.

3. RESULTS AND DISCUSSIONS

Box plots for PM_{10} at Seberang Perai are shown in Figure 3. The box plots in Figure 3 show that there were several occurrences of extreme event every year. The median value or 50th percentile of the monitoring records are always lower than the mean. These indicate that the concentrations of PM_{10} recorded are probably high or extremely high. Positive skewness at this site signify the occurrence of extreme events. The annual mean are 2004 (91.76 µg/m^3); 2005 (78.99 µg/m^3); 2006 (49.81 µg/m^3) and 2007 (45.45 µg/m^3). Percentage of missing values for each year are 2004 (4.9%), 2005 (0.55%), 2006 (6.28%) and 2007 (4.76%). Annual mean values for 2004 and 2005 are similar, but the values dropped 45% in 2006 and 2007 because Department of Environment shifted monitoring station from Institut Latihan Perindustrian to Sek. Keb.Cenderawasih for industrial area in Pulau Pinang due to unsuitability.

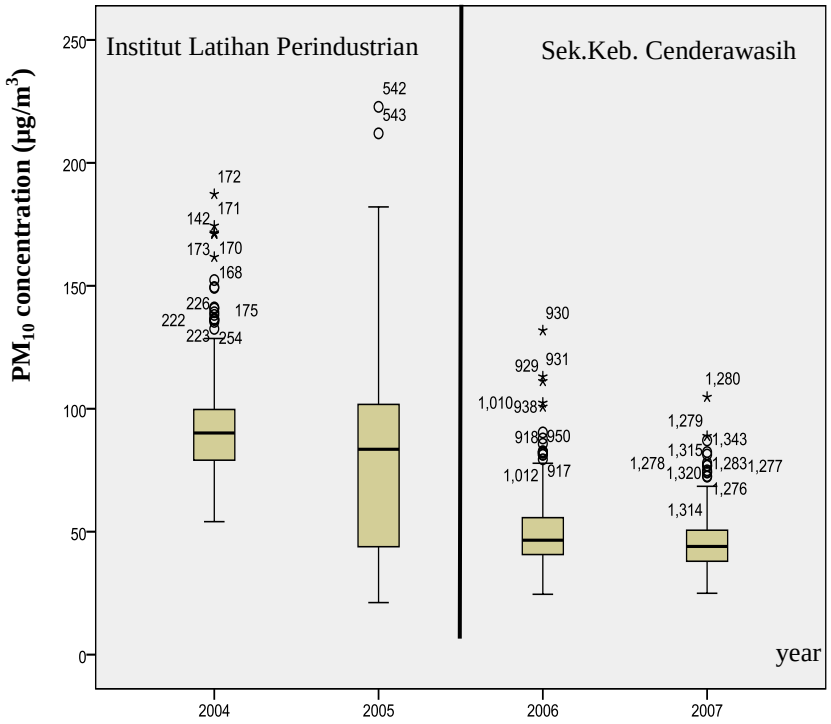
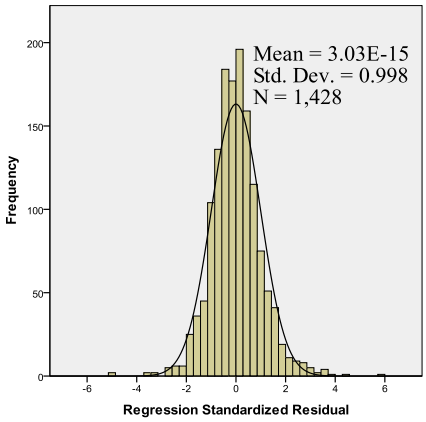


Figure 4. Box plots for PM_{10} at Seberang Perai

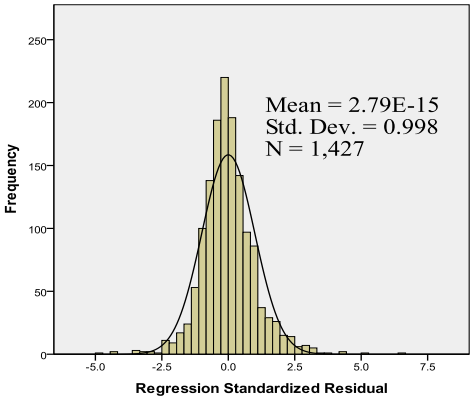
Ordinary least square method produced the best linear unbiased estimators under the normal error distribution. Real data usually do not meet the normality assumption, the OLS fit may not be optimal. Figure 4 shows error distribution for next day, next 2-day and next 3-day for predicting of PM_{10} concentration. In this case, for "next

day" and "next 2-day", the error look normally distributed and for the "next 3-day", the error look skewed to the right. From the first inference, it can be concluded that for the "next 3-day", robust regression method is more suitable because it can minimize influential outliers data automatically.

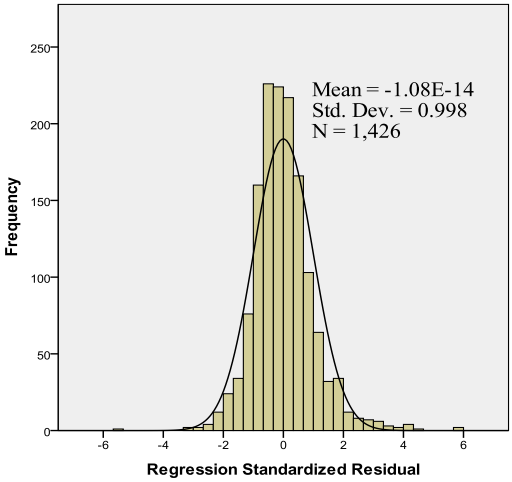
In this research, eight weighting function were used for robust regression i.e 'Andrews', 'bisquare', 'Cauchy', 'Fair', 'Huber', 'logistic', 'Talwar' and 'Welsch'. Table 3 showed performance comparison between weighting function for next day. Accuracies measured were prediction accuracy, coefficient of determination, and index of agreement, while the error measures were normalized absolute error and root mean square error. From the result, 'Fair' is the best weighting function for the robust regression for predicting PM_{10} concentrations. On the other hand, 'Cauchy' gives the lowest normalize absolute error.



(a)



(b)



(c)

Figure 4. Standardized residual analysis of PM_{10} for (a) next day (b) next 2-day (3) next 3-day.

Table 3. Performance Indicators for Next Day PM_{10} Concentration Prediction

Weighting Function	NAE	RMSE	PA	R^2	IA
Andrews	0.122352	11.085348	0.927057	0.858232	0.960631
Bisquare	0.122356	11.085003	0.927063	0.858243	0.960628
Cauchy	**0.122380**	11.078316	0.927114	0.858337	0.960760
Fair	0.122388	**11.074547**	**0.927142**	**0.858389**	**0.960841**
Huber	0.122422	11.076809	0.927128	0.858362	0.960770
Logistic	0.122386	11.076542	0.927128	0.858362	0.960799
Talwar	0.122394	11.086073	0.927053	0.858223	0.960533
Welsch	0.122367	11.082224	0.927085	0.858283	0.960683

Repeating the procedure revealed the best weighting function for next 2-day and next 3-day. Table 4 showed the best weighting function for next day, next 2-day and next 3-day for predicting PM_{10} concentration. The performance indicators reflected greater accuracy in next day PM_{10} concentration prediction compared to the next 2-day and next 3-day predictions. However, the result showed that robust regression could predict future PM_{10} concentration accurately until the next 3-day.

Table 4. The Best Weighting Function for Next Day, Next 2-day and Next 3-day.

Method	NAE	RMSE	PA	R^2	IA
Next Day (Fair)	0.122388	11.074547	0.927142	0.858389	0.960841
Next 2-day (Fair)	0.150352	14.074034	0.880680	0.774511	0.929396
Next 3-day (Cauchy)	0.165920	15.693665	0.848589	0.719094	0.910322

Multiple Linear Regression analysis based on the ordinary least square (OLS) method have been done for comparing performance between robust regression (RR) and multiple linear regression (MLR). Table 5 showed the model for predicting PM_{10} concentration using OLS and robust regression. The model showed that the concentration of PM_{10} in an industrial was influenced by ambient temperature, wind speed, relative humidity, carbon monoxide, sulphur dioxide, nitrogen dioxide and previous PM_{10} concentrations. Differences in constant value indicated the influence of outlier data for all the models.

Negative correlation between temperature and PM_{10} attributed to insignificant temperature fluctuation in Malaysia (24^0C-32^0C). PM_{10} level plummeted during strong wind events due to the dispersion by strong wind. SO_2 had positive correlation with PM_{10} in the area because most SO_2 came from diesel fueled vehicle motor emissions. Besides that, SO_2 was also contributed by industrial activities. For NO_2 and CO, the main sources were petrol fueled vehicle emission.

Table 5. Model for Predicting PM_{10} Concentration using OLS and Robust Regression.

Next day	MLR (OLS)	$PM_{10,t+1} = 12.50+0.95PM_{10} - 0.14ws - 0.1T+ 0.05RH -589.91NO_2 - 9.30CO + 172.06SO_2$
	Robust (Fair)	$PM_{10,t+1} = 14.30+0.96PM_{10} - 0.07ws - 0.1T+ 0.04RH -589.24NO_2 - 8.74CO + 108.73SO_2$
Next 2-day	MLR (OLS)	$PM_{10,t+2} = 39.9+ 0.9PM_{10} - 0.2ws - 0.8T+ 0.1RH -1046.9NO_2 – 18.2CO + 305.2SO_2$
	Robust (Fair)	$PM_{10,t+2} = 46.3+ 0.9PM_{10} - 0.3ws - 0.9T+ 0.1RH -1065NO_2 – 15.6CO + 219.8SO_2$
Next 3-day	MLR (OLS)	$PM_{10,t+3} = 53.3+ 0.9PM_{10} -0.76ws – 1.1T+ 0.2RH -1298.9NO_2 – 23.9CO + 433.7SO_2$
	Robust (Cauchy)	$PM_{10,t+3} = 65.8+ 0.9PM_{10} -1.0ws – 1.3T+ 0.1RH -1335.7NO_2 – 18.3CO + 358.9SO_2$

Table 6 also showed the performance indicators between OLS method and robust regression. All the result showed that robust regression is better than MLR (OLS) for future PM_{10} concentration prediction in an industrial area, Pulau Pinang. The result showed MLR (OLS) is a less suitable method compared to robust regression method in predicting PM_{10} concentration for data with outliers. On the other hand, this research found robust regression procedure provides a good mechanism for outliers identification as well as evaluating their relative influence on the final regression coefficient estimations. However, MLR is still applicable for PM_{10} concentration predictions.

Table 6. Performance indicators between OLS method and robust regression.

	Method	NAE	RMSE	PA	R^2	IA
Next Day	**Robust (Fair)**	**0.122**	**11.075**	**0.927**	**0.858**	**0.961**
	MLR (OLS)	0.126	11.374	0.923	0.851	0.959
Next 2-day	**Robust (Fair)**	**0.150**	**14.074**	**0.881**	**0.775**	**0.929**
	MLR (OLS)	0.161	14.815	0.865	0.748	0. 923
Next 3-day	**Robust (Cauchy)**	**0.166**	**15.694**	**0.849**	**0.719**	**0.910**
	MLR (OLS)	0.181	16.799	0.823	0.676	0.895

4. CONCLUSION

The robust regression method (weighted least square) should be considered as an alternative approach for predicting PM_{10} concentration in Malaysian industrial area. Assessment of model performance indicated that robust regression with weighted least square has more advantages than OLS. Robust regression can minimize influence of outliers in data sets. Subsequent analysis of residual is unnecessary because this method can detect outliers automatically. If the data sets do not have any outlier, the result of OLS and robust regression are similar. Due to the prevalence of outliers (5% per year) in air pollution data such as extreme events, robust regression should be considered as a substitute to OLS method.

ACKNOWLEDGEMENTS

This study was funded by Universiti Sains Malaysia under Grant 304/PAWAM/60311017. Thank you to Universiti Sains Malaysia and Universiti Teknologi MARA for providing financial support to carry out this study and also thanks to the Department of Environment Malaysia for their support.

REFERENCES

[1] Lee S. W., (2010), *Fine particulate matter measurement and international Standardization for air quality and emissions from stationary sources.* Fuel, 89 (4), p.874 – 882

[2] Department of Environment (DoE), (2008), Malaysia, *Malaysia Environmental Quality Report 2008.* Kuala Lumpur: Department of Environment, Ministry of Sciences, Technology and the Environment, Malaysia.

[3] Ramli N. A., and Ibrahim W. H. W., (2003), PM10 concentration measurements at four selected sites in Semenanjung Malaysia: A comparison between sites with different background. In Proceeding ENVIRONMENT 2003, Pulau Pinang, p. 236 – 238 [

[4] Sedek J. N. M., Ramli N. A., and Yahaya A. S., (2006), *Air quality predictions using lognormal distribution functions of particulate matter in Kuala Lumpur, Malaysia.* Journal of Environmental Management, 7, p. 33 – 41

[5] Sansuddin N., (2010), Modeling Locational differences and prediction of temporal concentration of PM10 using time series analysis, Phd Dissertaion, Universiti Sains Malaysia.

[6] Fellenberg G., (2000), *The Chemistry of Pollution. England* , John Wiley and Sons Ltd.

[7] Godish T., (2004), *Air Quality (Fourth Edition).* Florida, Washington: Lewis Publishers.

[8] Tam B. N., and Neumann C. M, (2004), *A human health assessment of hazardous air pollutants in Portland, OR.* Journal of Environmental Management, 73(2004), p. 131 – 145

[9] Baccini M., Biggeri A, Grillo P., Consonni D.and Bertazzi P.A, (2011), Health Impact Assessment of Fine Particle Pollution at the Regional Level, American Journal of Epidemiology, 174(12), pp 1396-1405.

[10] Rousseeuw P.J., and Leroy A.M., (2008), *Robust Regression and Outlier Detection*, Wiley Series In Probability and Statistics.

[11] Mahajan V., Sharma S., and Wind Y., (1984) *Parameter estimation in marketing models in the presence of influential response Data: Robust Regression and applications*, Journal of Marketing Research, vol. XXI, pp 288-277

[12] Department of Statistics, Malaysia, (2010) *Population and Housing Census of Malaysia*.

[13] Socio-economic and Environmental Reseach Institute (SERI), (2011), *Statistics: Population*. Accessed 11/11/2011http://www.seri.com.my/v3/index.php?option=com_content&view=category&layout=blog&id=61&Itemid =80].

[14] MATLAB version R2009a, (2009), *computer software*, The MathWorks Inc.

[15] Lu H. C., (2003), *Estimating the emission source reduction of PM_{10} in central Taiwan*. Chemosphere, 54 (7), pp. 805 – 814.

