Malaysian Journal of Analytical Sciences, Vol 20 No 4 (2016): xxx - xxx
DOI: http://dx.doi.org/10.17576/mjas-2016-2004-zz

MALAYSIAN JOURNAL OF ANALYTICAL SCIENCES
Published by The Malaysian Analytical Sciences Society
ISSN 1394 - 2506

MODELLING DISTRIBUTION FUNCTION OF SURFACE OZONE CONCENTRATION FOR SELECTED SUBURBAN AREAS IN MALAYSIA

(Permodelan Fungsi Taburan Kepekatan Permukaan Ozon di Kawasan Sub-Bandar yang Terpilih di Malaysia)

Muhammad Izwan Zariq Mokhtar, Nurul Adyani Ghazali*, Muhammad Yazid Nasir, Norhazlina Suhaimi

School of Ocean Engineering,
Universiti Malaysia Terengganu, 21030 Kuala Terengganu, Terengganu, Malaysia

**Corresponding author: nurul.adyani@umt.edu.my*

Received: 24 February 2015; Accepted: 27 October 2015

Abstract

Ozone is known as an important secondary pollutant in the atmosphere. The aim of this study is to find the best fit distribution for calculating exceedance and return period of ozone based on suburban areas; Perak (AMS1) and Pulau Pinang (AMS2). Three distributions namely Gamma, Rayleigh and Laplace were used to fit 2 years ozone data (2010 and 2011). The parameters were estimated by using Maximum Likelihood Estimation (MLE) in order to plot probability distribution function (PDF) and cumulative distribution function (CDF). Four performance indicators were used to find the best distribution namely, normalized absolute error (NAE), prediction accuracy (PA), coefficient of determination (R^2) and root mean square error (RMSE). The best distribution to represent ozone concentration at both sites in 2010 and 2011 is Gamma distribution with the smallest error measure (NAE and RMSE) and the highest adequacy measure (PA and R^2). For the 2010 data, AMS1 was predicted to exceed 0.1 ppm for 2 days in 2011 with a return period of one occurrence per 179 days. However, AMS2 do not exceed MAAQG limit (0.1 ppm) based on both 2010 and 2011. From the results, the exceedance and return period can be used as a guidance to overcome future air pollution problem.

Keywords: ozone, performance indicators, probability distribution function, return period, cumulative distribution function

Abstrak

Ozon dikenali sebagai pencemar sekunder yang paling penting di atmosfera. Matlamat kajian ini adalah untuk mengenalpasti taburan yang terbaik dalam mengira jumlah kepekatan yang melepasi had yang ditetapkan dan tempoh kembali untuk ozon berdasarkan kawasan sub-bandar; Perak (AMS1) dan Pulau Pinang (AMS2). Pembolehubah – pemboleubah telah dianggarkan melalui penggunaan penganggar kebolehjadian maksimum (MLE) untuk plot fungsi taburan kebarangkalian (PDF) dan fungsi taburan kumulatif (CDF). 4 penunjuk prestasi telah digunakan untuk mencari taburan terbaik antaranya, ralat mutlak dinormalkan (NAE), kejituan ramalan (PA)', pekali penentuan (R^2)' dan punca min ralat kuasa dua (RMSE). Taburan terbaik yang mewakili kepekatan ozon di kedua – dua tempat adalah taburan Gamma dengan nilai ukuran ralat yang terkecil (NAE dan RMSE) dan nilai ukur kadaran yang terbesar (PA dan R^2). Data bagi tahun 2010, AMS1 dijangkakan akan melepasi had 0.1 ppm dalam masa 2 hari pada tahun 2011 dengan tempoh kembali satu kejadian dalam masa 179 hari. Walaubagaimanapun, AMS2 tidak melebihi had MAAQG (0.01 ppm) untuk kedua – dua tahun 2010 dan 2011. Berdasarkan hasil kajian, didapati jumlah kepekatan yang melepasi had yang ditetapkan dan tempoh kembali boleh digunakan sebagai panduan untuk mengatasi masalah pencemaran udara pada masa akan datang.

Kata kunci: ozon, penunjuk prestasi, fungsi taburan kebarangkalian, fungsi taburan kumulatif, tempoh kembali

MJAS Vol 20 No 4 (2016)

Introduction

Owing to the burgeoning development of infrastructure in all areas in Malaysia including suburban areas, surface ozone occurrence is very likely. The higher entropic emissions such as heating emission from building and the unfavourable dispersion conditions can initiate the temperate climate of high air pollution [1]. The precursors of surface ozone are contributed by anthropogenic and natural sources. Anthropogenic sources such as deforestation, logging, fuel combustion, biomass burning, industrial causes and marine emission [2]. The examples of natural sources are nitrogen cycle, sea salt aerosol, cud-chewing animal waste and atmospheric lightning [2]. The ozone (O_3) can irritate lung airways and cause inflammation much like sunburn [3]. In Malaysia, many studies applied statistical distribution on particulate matter (PM_{10}) but less in ozone (O_3). Most of the studies on ozone use different method which are multiple linear regressions (MLR), variation and correlation between parameters such as reported in previous studies [4, 5], but there are lack of studies applying statistical distribution on ozone.

The aim of this study is to find the best fit distribution for determining exceedance and return period of ozone in AMS1 and AMS2 via statistical distribution. Three distributions namely Gamma, Rayleigh and Laplace were used to fit 2 years ozone data (2010 and 2011).

Materials and Methods

Study area

Both sites selected in this study were shown in Figure 1. The coordinate of AMS1 is 3°41’17.00’’ N and 101°31’11.79’’ E. AMS1 situated at Tanjung Malim district. Surrounded by many residential and construction areas, contained partly of Lebuhraya Utara-Selatan (main highway in Malaysia) can catalyse the ozone formation. The coordinate of AMS2 is 5°21’38.12’’ N and 100°17’55.74’’ E. AMS2 located at an education based centre in Pulau Pinang and congested with resident areas and traffics including Pulau Pinang bridge which is connected with the mainland.

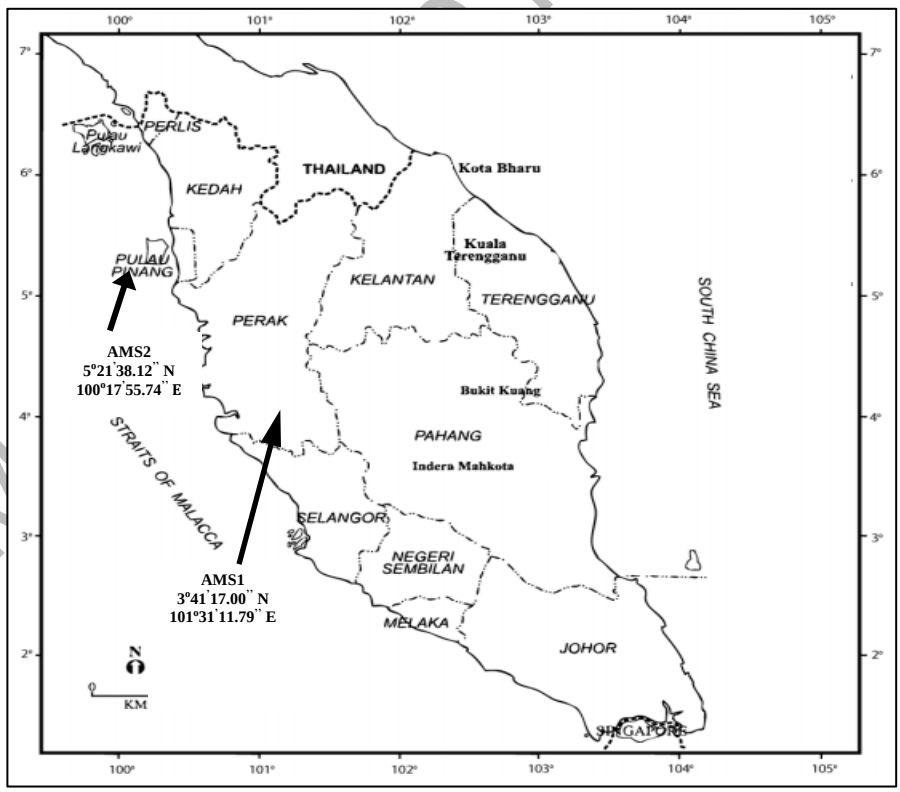

Figure 1. Map of Malaysia Peninsula (site location and its description [6])

Instrumentation and data collection

The hourly O_3 data is procured from both sites in year 2010 and 2011 which were owned from Department of Environment Malaysia (DoE) and managed by Alam Sekitar Sdn. Bhd (ASMA). The data are subjected to standard quality control processes and quality assurance procedures [7]. ASMA applied Beer-Lambert law to measure low range surface ozone concentration by using Teledyne Ozone Analyzer Model 400E UV Absorption [7]. Those best fit distribution, best distribution, exceedance and return period were obtained by analysing selected distributions via MATLAB R2014a. Statistical Package for Social Science version 19 (SPSS) was implemented to identify descriptive statistics for both stations.

Probability distribution and performance indicator

The parameters for the distributions were estimated by maximum likelihood estimation (MLE). The Gamma distribution is comprised of Chi-Squared, Erlang and Exponential distribution. It has scale and shape parameter in which growing rate of shape parameter signifies the large dataset [8]. Laplace distribution is known as double exponential distribution because of its CDF pattern [8]. It has location and scale parameter in which increasing number of scale parameter epitomizes heavier tail and lower kurtosis in PDF figure. Rayleigh distribution has scale parameter only which is simplified from Weibull distribution [8]. The formulas are as shown in Table 1. The formulas for the performance indicators are tabulated in Table 2.

Table 1. Each distribution with respective probability distribution function (PDF), cumulative distribution function (CDF) and maximum likelihood estimation (MLE)

Distribution	**PDF and CDF**	**MLE parameter**
Gamma	PDF: $f\left(x; a, B\right) = \frac{1}{\Gamma(a)B^a} x^{a-1}\exp\left(-x/B\right)$ (1) CDF: $\frac{1}{\Gamma(a)}\gamma(a, \frac{x}{B})$ $a > 0, B > 0$ (3)	Scale Parameter: $\hat{a} = \bar{x}/\hat{B}$ (2) Shape Parameter : $log\hat{a} - \varphi(\hat{a}) = \left[\bar{x}/(\prod_{i=1}^{n} x_i)\,\hat{}\,\frac{1}{n}\right]$ (4)
Rayleigh	PDF: $f(x; a) = \frac{x}{a^2} e^{\frac{-x^2}{2a^2}}$ (5) CDF: $1 - e^{\frac{-x^2}{2a^2}}$ $a > 0$ (7)	Scale Parameter : $a = \left(\left(\frac{1}{2n}\sum_{i=1}^{n} x_i^2\right)\right)\,\hat{}\,\frac{1}{2}$ (6)
Laplace	PDF: $f(x; \mu, \lambda) = \frac{1}{2\lambda} e\left(-\frac{\lvert x-\mu\rvert}{\lambda}\right)$ (8) CDF: $\begin{cases} \frac{1}{2} e\left(\frac{x-\mu}{\lambda}\right) \; if \; x < \mu \\ 1 - \frac{1}{2} e\left(\frac{-x-\mu}{\lambda}\right) \; if \; x \geq \mu \end{cases}$ (10)	Scale Parameter: $\frac{1}{n}\sum_{i=1}^{n}\lvert x_i - \mu\rvert$ (9) Location Parameter: Median (11)

Notation: μ is the scale parameter for Laplace distribution, a is the scale parameter for Gamma and Rayleigh distribution, λ is the location parameter and B is the shape parameter [8].

Table 2. The formula and best value for each performance indicator

Type	**Performance Indicator**	**Formula**	**Best Value**
Error Measure	Normalized Absolute Error (NAE)	$NAE = \frac{1}{\sum_{i=1}^{N}(O_i)}\sum_{i=1}^{N}\lvert P_i - O_i\rvert$ (12)	Near 0
	Root Mean Square Error (RMSE)	$RMSE = \sqrt{\left(\frac{1}{N-1}\right)\sum_{i=1}^{N}(P_i - O_i)\hat{}2}$ (13)	Near 0
Adequacy Measure	Prediction Accuracy (PA)	$PA = \sum_{i=1}^{N}\left[\frac{(P_i-\bar{P})(O_i-\bar{O})}{(N-1)\sigma_P\sigma_O}\right]$ (14)	Near 1
	Coefficient of Determination (R^2)	$R^2 = \left[\frac{1}{N}\frac{\sum_{i=1}^{N}(P_i-\bar{P})(O_i-\bar{O})}{\sigma_P\sigma_O}\right]$ (15)	Near 1

MJAS Vol 20 No 4 (2016)

Exceedance and return period

Actual exceedance can be obtained by calculating the cases number exceeding the threshold of Malaysia Ambient Air Quality Guideline (MAAQG) limit and dividing this number by the number of available data [12]. Predicted exceedance can be determined by using the data collected from both actual and previous years [12]. Thus, the formula of exceedance is written in equation 16 below [9].

Exceedance: $P\{X > 0.1\} = 1 - P\{X \leq 0.1\} = 1 - F(0.1)$ (16)

Return period is a time period to calculate how many days the ozone data exceed the standard. The formula of return period is stated as below [9].

Return Period: $\frac{1}{P\{X>0.1\}} \times 365\ days$ (17)

Results and Discussion

The description statistic for AMS1 and AMS2 were shown in Table 3. The skewness showed positive numbers that define the occurrence of extreme events and high ozone emissions. These results showed that the ozone distributions are skewed to the right represent most data is concentrated on the left of the PDF plots with few high values. The mean, median, skewness and kurtosis were increasing for both sites from 2010 until 2011, indicating growing air pollution problem.

Table 3. The descriptive statistic for both sites

Data parameters	AMS1		AMS2	
	2010	2011	2010	2011
Maximum	0.127	0.165	0.086	0.089
Mean	0.018	0.019	0.019	0.020
Standard Deviation	0.018	0.021	0.014	0.015
Median	0.011	0.012	0.015	0.017
Skewness	1.13	1.6	0.9	1.005
Kurtosis	3.76	6.03	3.18	3.55

Table 4 shows that the parameter estimates are different when dissimilar statistical distributions models were applied via Maximum Likelihood Estimation (MLE). All shape parameter (a) in Gamma and Rayleigh distribution and scale parameter (λ) in Laplace distribution were less than 0.5 for both sites in 2 years. These showed that the PDF plot for all distributions had high and sharp central peak (represent high ozone concentration) and the tails are long and fat (represent few high values). The scale parameters (B) were greater than shape parameters (a) in Gamma distribution. The scale parameters (λ) were larger than location parameter (μ) in Laplace distribution. From the results, these parameters showed higher concentration for both sites in 2011 compared to 2010.

Table 4. MLE parameter estimates for both sites

Data parameters for distributions	AMS1		AMS2	
	2010	2011	2010	2011
Gamma	a=0.019 B=0.95	a=0.017 B=1.12	a=0.019 B=0.994	a=0.013 B=1.656
Rayleigh	a=0.018	a=0.020	a=0.017	a=0.018
Laplace	μ=0.011 λ=0.014	μ=0.011 λ=0.015	μ=0.015 λ=0.011	μ=0.017 λ=0.012

MJAS Vol 20 No 4 (2016)

From the parameter estimates, CDF and PDF for Gamma, Rayleigh and Laplace distribution were plotted. Among these distributions, Gamma was the best fitted distribution as shown in Figure 2 for both sites from 2010 until 2011. PDF plots for both sites in 2010 and 2011 are positively skewed and the densities were lower meaning that ozone concentration had risen. Long tail recorded in both PDF plots indicated the presence of extreme value. The CDF plot was used to indicate the best distribution that fit the ozone concentration data. Observational line in CDF plot for AMS1 in 2010 fit at 0.044 ppm with theoretical line. Meanwhile, for AMS1 in 2011, the observational line fit at 0.02 ppm. The observational line in CDF plot for AMS2 in 2010 fitted to theoretical line at 0.04 ppm and for AMS2 in 2011, most of ozone concentration data fit to the theoretical line.

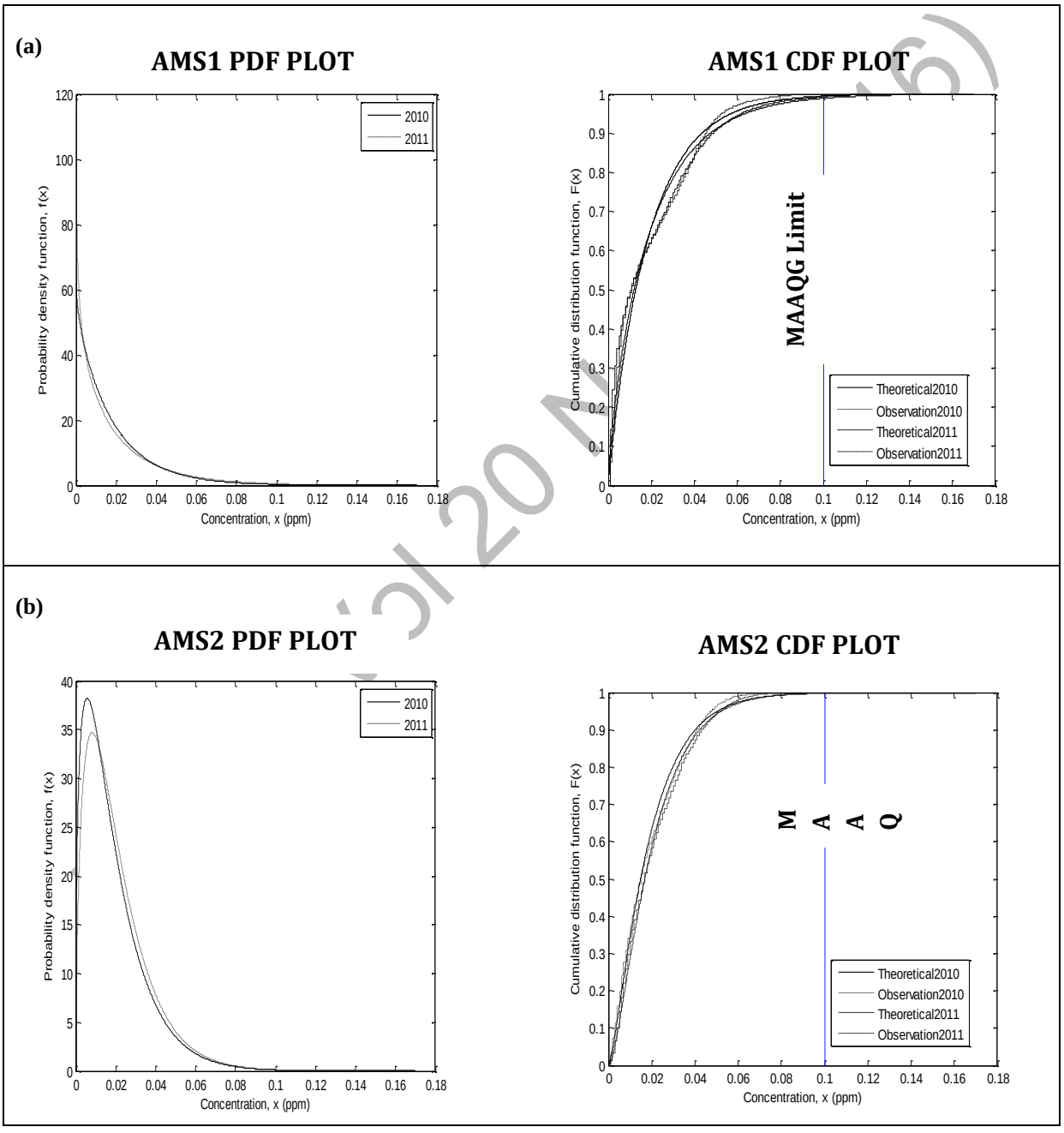

Figure 2. Gamma distribution PDF and CDF plots for both sites in 2010 and 2011 with MAAQG limit

The role of performance indicator is to determine the best fit distribution. NAE and RMSE were in error measure category which decides the smallest value gives the best fit. The R^2 and PA were categorized in adequacy measure which indicates the closer value to 1 gives the best fit. The result shows that the Gamma distribution is the best for both sites from 2010 to 2011 as shown in Table 5.

Table 5. The performance indicators for ozone concentration in both sites

Performance Indicator	Sites (Year)	Gamma Distribution	Rayleigh Distribution	Laplace Distribution	Best Distribution
NAE (near to 0)	AMS1 (2010)	0.151	0.405	0.361	Gamma
	AMS1 (2011)	0.181	0.496	0.410	
	AMS2 (2010)	0.150	0.223	0.206	
	AMS2 (2011)	0.054	0.198	0.197	
RMSE (near to 0)	AMS1 (2010)	0.005	0.009	0.008	Gamma
	AMS1 (2011)	0.006	0.011	0.008	
	AMS2 (2010)	0.004	0.005	0.012	
	AMS2 (2011)	0.002	0.005	0.013	
R^2 (near to 1)	AMS1 (2010)	0.950	0.939	0.790	Gamma
	AMS1 (2011)	0.989	0.919	0.784	
	AMS2 (2010)	0.943	0.980	0.851	
	AMS2 (2011)	0.982	0.980	0.858	
PA (near to 1)	AMS1 (2010)	0.975	0.969	0.889	Gamma
	AMS1 (2011)	0.994	0.959	0.886	
	AMS2 (2010)	0.971	0.990	0.923	
	AMS2 (2011)	0.991	0.990	0.926	

Table 6 shows that the predicted exceedances were underestimated for AMS1 from 2010 to 2011. However, the differences between predicted and actual days were small for both years. The gap of more than 1-day difference for AMS1 which is surrounded by residential areas showed that there was a few of consistent high hourly ozone detection in 2010. For 2011, the gap was below 1-day difference showed that there was detection of greater amount of consistent high hourly ozone concentration. From the result, the decreasing of difference between predicted and actual exceedances showed that the ozone concentration had become worsen from 2010 to 2011.

Table 6. The exceedance probability and return period of AMS1 for 2010 and 2011

	Elements	AMS1	
		2010	2011
	Probability ≤ 0.1 ppm	0.9944	0.9938
	Exceedance Probability > 0.1 ppm	5.6×10^{-3}	6.2×10^{-3}
Return Period	Predicted Exceedances (day)	2	2.3
	Actual Exceedances (day)	0.3	2.7
	Difference of predicted and actual exceedances (day)	1.7	0.4

MJAS Vol 20 No 4 (2016)

Conclusion

This study shows Gamma distribution is the best fit distribution for both suburban-based sites in 2010 and 2011. Both sites show the mean are greater than median indicating the observation data are positively skewed in all three distributions. For the 2010 data, AMS1 was expected to surpass 0.1 ppm for 2 days in 2011 with a return period of one occurrence per 179 days. For the 2011 data, AMS1 was expected to surpass 0.1 ppm for 2.3 days in 2012 with a return period of one occurrence per 159 days. The results showed the ozone concentration at AMS1 became worsen from 2010 to 2011. AMS2 do not have data that exceed 0.1 ppm for both years, therefore no exceedance and return period need to be considered.

Acknowledgement

The authors would like to thank to Ministry of Higher Education of Malaysia for funding FRGS grant 59314, University Malaysia Terengganu for supporting this project and Department of Environment Malaysia (DoE) for providing the O_3 data on 2010 and 2011 to complete this study.

References

1. Yusof, N. F. F. M., Ramli, N. A., Yahaya, A. S., Sansuddin, N., Ghazali, N. A. and Madhoun, W. A. (2010). Monsoonal differences and probability distribution of PM_{10} concentration. *Environmental Monitoring and Assessment*, 163: 655 – 667.
2. Wallace, J. M. and Hobbs, P.V. (2006). Atmospheric Science. *Academic Press Publication*, 2: 153 – 198.
3. Ramli, N. A., Ghazali, N. A. and Yahaya, A. S. (2010). Diurnal fluctuations of ozone concentrations and its precursors and prediction of ozone using multiple linear regressions. *Malaysia Journal of Environmental Management*, 11(2): 57 – 69.
4. Banan, N., Latif, M.T. and Juneng, L. (2013). An assessment of ozone levels in typical urban areas in the Malaysian Peninsular. *International Journal of Environmental, Earth Science and Technology*, 7(2): 61 – 64.
5. Ahamad, F., Latif, M.T., Tang, R., Juneng, L., Dominick, D. and Juahir, H. (2014). Variation of surface ozone exceedance around Klang Valley, Malaysia. *Journal of Atmospheric Research*, 139: 116 – 127.
6. Ismail, A. S., Latif, M. T., Azmi, S. Z., Juneng, L. and Jemain, A. A. (2010). Variation of surface ozone recorded at the eastern coastal region of the Malaysian Peninsula. *American Journal of Environmental Sciences*, 6(6): 560 – 569.
7. Department of Environment, Ministry of Natural Resources and Environment (2012). Malaysia environmental quality report. ISSN 0127 – 6433.
8. Forbes, C., Evans, M., Hastings, N. and Peacock, B. (2010). Statistical Distributions. Wiley, 4: 109 – 176.
9. Noor, N. M., Tan, C., Ramli, N. A., Yahaya, A. S. and Yusof, N. F. F. M. (2011). Assessment of various probability distributions to model PM_{10} concentration for industrialized area in Peninsula Malaysia: A case study in Shah Alam and Nilai. *Australian Journal of Basic and Applied Sciences*, 5(12): 2796 – 2811.
10. Lu, H.C. (2003). Estimating the emission source reduction of PM_{10} in Central Taiwan. *Journal of Chemosphere*, 54: 805 – 814.
11. Ul-Saufie, A. Z., Yahaya, A. S., Ramli, N. A. and Hamid, H.A. (2012). Robust regression models for predicting PM_{10} concentration in an industrial area. *International Journal of Engineering and Technology*, 2(3): 364 – 370.
12. Sansuddin, N., Ramli, N. A., Yahaya, A. S., Yusof, N. F. F. M., Ghazali, N. A. and Madhoun, W. A. A. (2011). Statistical analysis of PM_{10} concentrations at different locations in Malaysia. *Environmental Monitoring Assessment*, 180: 573 – 588.

MJAS Vol 20 No 4 (2016)