Malaysian Journal of Analytical Sciences, Vol 20 No 3 (2016): xxx – xxx
DOI: http://dx.doi.org/10.17576/mjas-xxxx-yyyy-zz

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

FITTING STATISTICAL DISTRIBUTIONS FUNCTIONS ON OZONE CONCENTRATION DATA AT COASTAL AREAS

(Penyesuaian Fungsi Taburan Statistik pada Data Kepekatan Ozon di Kawasan Pesisiran Pantai)

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

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

**Corresponding author: muhammad.yazidnasir@gmail.com*

Received: 24 February 2015; Accepted: 27 October 2015

Abstract

Ozone is known as one of the pollutant that contributes to the air pollution problem. Therefore, it is important to carry out the study on ozone. The objective of this study is to find the best statistical distribution for ozone concentration. There are three distributions namely Inverse Gaussian, Weibull and Lognormal were chosen to fit one year hourly average ozone concentration data in 2010 at Port Dickson and Port Klang. Maximum likelihood estimation (MLE) method was used to estimate the parameters to develop the probability density function (PDF) graph and cumulative density function (CDF) graph. Three performance indicators (PI) that are normalized absolute error (NAE), prediction accuracy (PA), and coefficient of determination (R^2) were used to determine the goodness-of-fit criteria of the distribution. Result shows that Weibull distribution is the best distribution with the smallest error measure value (NAE) at Port Klang and Port Dickson is 0.08 and 0.31, respectively. The best score for highest adequacy measure (PA: 0.99) with the value of R^2 is 0.98 (Port Klang) and 0.99 (Port Dickson). These results can give information to local authorities for prediction purpose.

Keywords: ozone concentration, coastal area, statistical distributions, goodness-of-fit, performance indicator

Abstrak

Ozon merupakan salah satu pencemar yang banyak menyumbang kepada masalah pencemaran udara. Maka kajian tentang ozon adalah penting untuk dijalankan. Objektif kajian ini adalah mencari taburan statistik yang terbaik untuk mewakili data kepekatan ozon. Tiga fungsi taburan yang digunakan dalam kajian ini adalah Gaussian songsang, Weibull dan Lognormal telah dipilih bagi menentukan taburan statistik yang terbaik untuk mewakili data taburan ozon per jam pada tahun 2010 di Port Dickson dan Port Klang. Kaedah penganggar kebolehjadian maksimum (MLE) telah digunakan untuk mengira parameter yang membentuk graf fungsi taburan kebarangkalian (PDF) dan fungsi taburan kumulatif (CDF). Tiga penunjuk prestasi (PI) iaitu ralat mutlak dinormalkan (NAE), kejituan ramalan (PA) dan pekali penentuan (R^2) telah digunakan untuk menguji prestasi kriteria taburan yang terbaik. Hasil kajian menunjukkan taburan Weibull adalah yang terbaik untuk mewakili data kepekatan ozon dengan nilai ukuran ralat terkecil (NAE) di Port Klang dan Port Dickson masing-masing ialah 0.08 dan 0.031. Skor terbaik juga terhasil untuk pengiraan kejituan tertinggi (PA: 0.99) dengan nilai R^2 di kedua-dua tempat ialah 0.98 (Port Klang) dan 0.99 (Port Dickson). Hasil kajian ini boleh digunapakai oleh penguatkuasa tempatan untuk tujuan ramalan kepekatan pada masa akan datang.

Kata kunci: Kepekatan ozon, pesisir pantai, taburan statistik, penyesuaian terbaik, penunjuk prestasi

Introduction

Ozone is one of air pollutant that exists in the atmosphere. Ozone is known as strong photochemical oxidants and one of the major problems originating from air pollution in urban areas [1]. Ozone formation at ground level was

originated by the ozone precursor which is Nitrogen oxides (NO_x) and Volatile Organic Compounds (VOC_s) has comes mainly from vehicles emission. The ozone concentration trend change when the emissions of its precursor change [2]. The ozone concentration that exists at ground level may be harmful to living organism especially human. Ozone was also known as the major contributing factor on chronic disease and mortality [3].

Ozone was expected to be existing at coastal site by the influence of the geographical structure of the site and others parameters. These secondary pollutants were affected by the continental and maritime wind as well as the sea breeze [4,5]. The diurnal ozone coastal has been investigated by Nair et al. [6], state that the ozone mixing ratio increase during early morning and reach maximum at 1100 hour and start decreasing after 1600 hour. However, the pattern has secondary peak appearing about 1900 hour shows the ozone is closely associate with the circulation pattern from sea-breeze to land breeze. Other studies found that ship emission also give a significant impact to the ozone flocculation. The impact of ships emission to ozone concentration at coastal site up to 15 ppb and about 5 ppb at location 2 km from that coastal site [7]. Hence, this study of ozone concentration is important to predict ozone exceedances to assess and monitor the air quality.

There are many statistical approaches used to study the ozone concentration such as probability distribution, multiple linear regression and artificial intelligence. Among that, probability distribution has been widely used by researcher to predict the air pollutant concentration [8,9]. This study will focus on three distributions namely Lognormal, Weibull and inverse Gaussian distribution for prediction purposes. Those distributions used certain parameter to form their own shape and characteristic [10]. These parameters were commonly determined by using few techniques such as maximum likelihood estimator (MLE), method of moment (MOM) and least square error. However, MLE method is selected in parameter estimation cause its efficiency and good theoretical properties [11].

The two-parameter Weibull distribution (scale (σ) and shape (λ)) is widely used in data analysis because of its flexibility in modeling. There are numerous papers and books dealing with various aspects of Weibull modeling, inference, applications, as well as parameter estimation [12]. The others competitor of Weibull distribution is Lognormal distribution which also good in most cases significant fit in seasonal and meteorological characterizations of daily data [13]. They found that 2-parameter log-normal distribution (location (μ) and scale (σ)) give the best description of annual mean daily sulphur dioxide concentration for a wide range of ambient level. The others competitor of the Weibull and Log-normal distribution in modelling asymmetric data from various scientific fields is Inverse Gaussian distribution. In reliability and life testing, the inverse Gaussian distribution is particularly useful in situations where early failures dominate [14]. Inverse Gaussian consists of two identical parameters which is location (μ) parameter and scale (σ) parameter. It is a two-parameter family of continuous probability distribution [15].

The first aim of this study is to determine the goodness-of-best fit distribution data representing ozone concentration by using distribution function method. The second aim of this study is to determine the best distribution for ozone concentration level at Port Dickson and Port Klang by using three performance indicators namely normalized absolute error (NAE), prediction accuracy (PA), and coefficient of determination (R^2).

Materials and Methods

Study area

Two sites were selected for this study, Port Dickson and Port Klang as illustrated in Figure 1.

Site 1 (Port Dickson) is located 1 km from sea and it has large contributions to the local economy in oil and gas sector. It has two refineries that have been operating by Shell Refining and Petron Company which expected to contribute with air pollution problem. According to Department of Environment Malaysia (DoE) [16], Port Dickson was recorded 7 days unhealthy during 2012 which means that the air quality in Port Dickson needs to be investigated. Site 2 (Port Klang) has the main business in marine-based activity. A lot of ships docked here for export and import purposes. Highly shipping emission and traffic congestions here due to truck involvement was expected to trigger the ozone formation contribute to high ozone concentration. DoE [16] stated that Port Klang had 13 unhealthy days on 2012 which represents third highest in Klang Valley based from the annual Environment Quality Report.

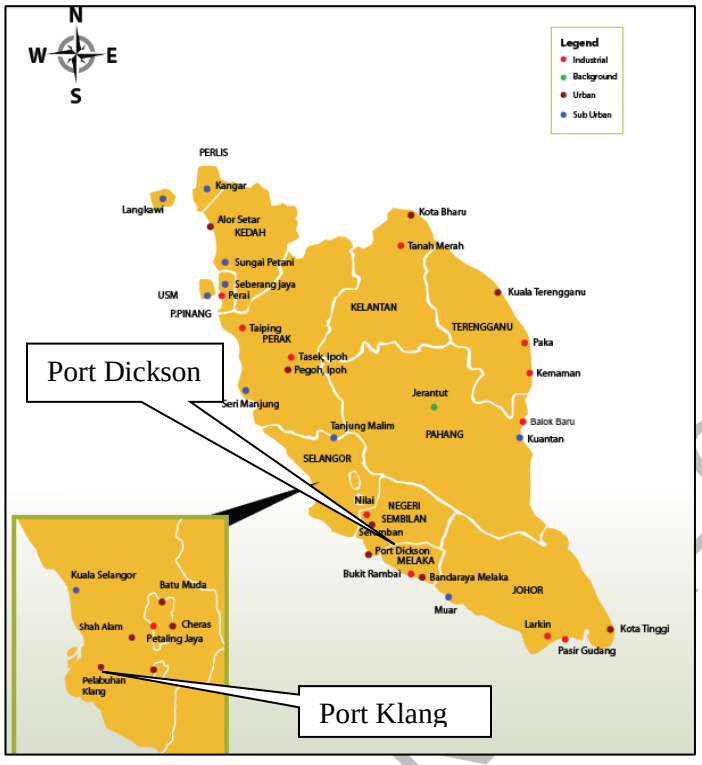


Figure 1. Description of the selected study area (Source: EQR 2012).

Probability distribution

Probability distribution is mathematical form to construct theoretical and idealization of real data set [17]. The probability density function (PDF) and cumulative density function (CDF) were used to estimate the probability of the exceedences of ozone based on Malaysian Ambient Air Quality Guideline (MAAQG). The PDF plot was used to identify the skewness of the distribution and it plot by using the value of parameter estimation [18]. The derivation of PDF was used for prediction of exceedences [19] and the CDF was used to determine the probability of air pollutant concentration [20]. Table 1 shows the PDF and the parameter estimation used [21].

Weibull distribution

The formula used for parameter estimation that was given by Lu [9]:

$$f(x) = \left(\frac{\lambda}{\sigma}\right)\left(\frac{x}{\sigma}\right)^{\lambda-1} \exp\left[-\left(\frac{x}{\sigma}\right)^{\lambda}\right]$$

The cumulative distribution function takes the form as

$$F(x) = 1 - \exp\left[-\left(\frac{x}{\beta}\right)^{\alpha}\right]$$

where $x \geq 0$, α = shape parameter, β = scale parameter. β was obtain by

$$\frac{\left(\sum_{i=1}^{n}(x_i^{\sigma} \ln x_i)\right)}{\left(\sum_{i=1}^{n} x_i^{\sigma}\right)} - \left(\frac{1}{\sigma}\right) - \left(\frac{1}{n}\right)\left(\sum_{i=1}^{n} \ln x_i\right) = 0$$

and α was then calculated using following equation

$$\lambda = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^{\hat{x}}\right)^{\frac{1}{\sigma}}$$

Table 1. Probability density function and its parameter estimates

Distribution	Probability Density Function	Parameter Estimates
Weibull	$f(x) = \left(\frac{\lambda}{\sigma}\right)\left(\frac{x}{\sigma}\right)^{\lambda-1} \exp\left[-\left(\frac{x}{\sigma}\right)^{\lambda}\right]$	$\frac{\left(\sum_{i=1}^{n}(x_i^{\sigma}\ln x_i)\right)}{\left(\sum_{i=1}^{n} x_i^{\sigma}\right)} - \left(\frac{1}{\sigma}\right) - \left(\frac{1}{n}\right)\left(\sum_{i=1}^{n}\ln x_i\right) = 0$; $\lambda = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^{\hat{x}}\right)^{\frac{1}{\sigma}}$
gnormal	$f(x) = \frac{1}{x\lambda\sqrt{2\Pi}} \exp\left[-\frac{1}{2}\left(\frac{\ln(x)-\sigma}{\lambda}\right)^2\right]$	$\sigma = \frac{1}{n}\sum_{i=1}^{n}\ln(x_i)$; $\mu = \sqrt{\frac{1}{n}\sum_{i=1}^{n}[\ln(x_i)-\sigma]}$
Inverse gaussian	$\left(\frac{\sigma}{2\pi x^3}\right)^{\frac{1}{2}} \exp\left(-\frac{\sigma(x-\mu)^2}{2\mu^2 x}\right)$	$\mu = \bar{x}$; $\sigma = \frac{n-1}{\sum_{i=1}^{n}\left(\frac{1}{x_i} - \frac{1}{\bar{x}}\right)}$

*μ is the location parameter, σ is the scale parameter, λ is the shape parameter, n is the total number of data

Lognormal Distribution

Lognormal was used to fit the ozone concentration data. Lognormal distribution with probability density function is given by Lu [9]:

$$f(x) = \frac{1}{x\lambda\sqrt{2\Pi}} \exp\left[-\frac{1}{2}\left(\frac{\ln(x)-\sigma}{\lambda}\right)^2\right]$$

The cumulative density function form for normal distribution is

$$f(x) = \frac{1}{2\Pi}\int_{-\infty}^{\frac{\ln(x)-\sigma}{\alpha}} e^{-\frac{x^2}{2}}\, dt$$

σ is obtain by solution below

$$\sigma = \frac{1}{n}\sum_{i=1}^{n}\ln(x_i)$$

and α by using solution below

$$\lambda = \sqrt{\frac{1}{n}\sum_{i=1}^{n}[\ln(x_i)-\sigma]}$$

Inverse Gaussian distribution
The probability density function (pdf) for Inverse Gaussian distribution given by Tweedie [15] is:

$$\left(\frac{\sigma}{2\pi x^3}\right)^{\frac{1}{2}} \exp\left(-\frac{\sigma(x-\mu)^2}{2\mu^2 x}\right)$$

where λ is scale parameter and μ is location parameter. The parameter estimation used is:

$$\sigma = \frac{n-1}{\sum_{i=1}^{n}\left(\frac{1}{x_i}-\frac{1}{\bar{x}}\right)}$$

for shape parameter, and

$$\mu = \bar{x}$$

for location parameter.

Performance indicator
The performance indicators used for this study were NAE, PA, and R^2 to identify the best distribution. The best distribution were selected by range of value that closest to 1 for adequacy measure (PA and R^2) and the value closest to zero for error measure (NAE). Table 2 shows three equations used for the performance indicator [22].

Table 2. Performance indicator equation

Indicators	Equation
Normalize absolute error	$\frac{\sum_{i=1}^{n} Abs(P_i - O_i)}{\sum_{i=1}^{n} O_i}$
Prediction of accuracy	$\frac{\sum_{i=1}^{N}(P_i - \bar{O})^2}{\sum_{i=1}^{N}(O_i - \bar{O})^2}$
Coefficient of determination	$\left(\frac{\sum_{i=1}^{N}(P_i - \bar{P})(O_i - \bar{O})}{NS_{pred}S_{obs}}\right)^2$

* N = Number of observations, P_i = Predicted values, O_i = Observed values, $\bar{P}$ = Mean of the predicted values, $\bar{O}$ = Mean of the observed values, S_{pred} = Standard deviation of the predicted values, S_{obs} = Standard deviation of the observed values

Results and Discussion

Ozone concentration data
The ozone concentration data collected over a year period 2010 from January to December by DOE Malaysia and managed by private company Alam Sekitar Malaysia Sdn Bhd (ASMA). The ozone concentration was measured by ASMA using Teledyne Ozone Analyzer Model 400 A UV Absorption [23]. The data was collected every hour and the unit measurement is part per million (ppm). All descriptive statistic ozone concentration for Port Klang, and Port Dickson is run by using IBM SPSS statistic 20 analytical and MATLAB r2014a computational software. Table 3 is the descriptive statistic of ozone concentration result from the software used in this study.

Table 3. Descriptive statistics of ozone concentrations

Descriptive	Port Klang (ppm)	Port Dickson (ppm)
Maximum	0.123	0.140
Mean	0.020	0.026
Median	0.014	0.022
Standard Deviation	0.018	0.019
Skewness	1.321	1.013

From Table 3, the maximum value of ozone concentration recorded in Port Dickson was higher than Port Klang. However, both sites were exceeding the MAAQG (0.1ppm). The mean of both sites were greater than median indicated there was high concentration recorded during period of study. The result also shows positively skewed to the right for Port Klang (sk = 1.321) and Port Dickson (sk = 1.013) means the right tails of data is longer and higher concentration has been occurred.

Parameter estimation

Table 4 shows the parameter estimation values for Port Dickson and Port Klang, respectively. From the table, the shape parameter (λ) is higher than scale parameter (σ) for Weibull distributions indicate higher ozone concentration has occurred at both sites. The Lognormal parameter shows higher ozone concentration occur by the scale parameter (σ) is higher than location parameter (μ). The inverse Gaussian parameter also show the location or mean parameter (μ) is higher than shape parameter (λ) which indicates high ozone concentration was occurred during this period.

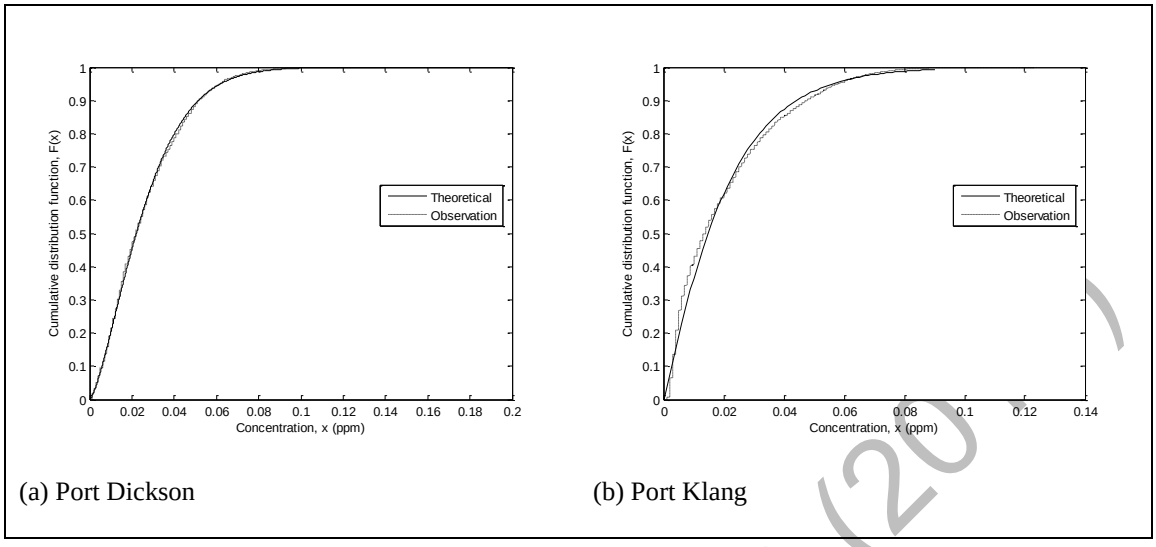
Table 4. Parameter estimation for Port Dickson and Port Klang

Distribution	Site	Parameter	
Weibull	Port Dickson	σ= 0.02863	λ= 1.417844
	Port Klang	σ= 0.02066	λ= 1.098527
Lognormal	Port Dickson	μ= -3.96320	σ= 0.900715
	Port Klang	μ= -4.38629	σ= 1.035732
Inverse Gaussian	Port Dickson	μ= 0.02604	λ= 0.019379
	Port Klang	μ= 0.01991	λ= 0.011716

The location parameter (μ), the scale parameter (σ) is the shape parameter, λ.

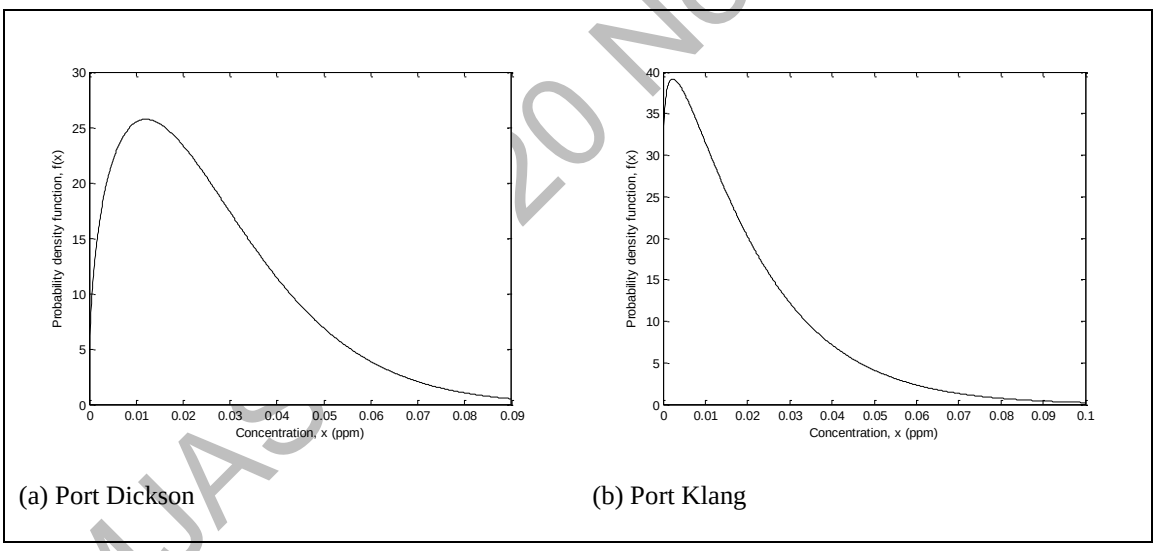
Probability distribution graph

From the parameter estimation, the CDF graph can be plotted to estimate goodness-of-fit of data. CDF graph for all distribution used were plotted to determine the best goodness-of-fit and the result were found to be Weibull distribution. Figure 2, shows the Weibull CDF plot of Port Dickson and Port Klang, respectively. The observation line was underestimates at approximately 0.038 ppm before it fitted again with theoretical line between 0.04 ppm and 0.06 ppm for Port Dickson site. Meanwhile, observation line was overestimates at 0.08 until 0.02 ppm then underestimate between 0.02 ppm and 0.06 ppm before it fitted again with theoretical line at 0.07 ppm for Port Klang site. Figure 3 shows the PDF plot of Weibull distribution depicted as it skewed to the right. The result also showed that the mode for Port Dickson site is approximately about 0.012 ppm and Port Klang site is approximately 0.008 ppm.



(a) Port Dickson (b) Port Klang

Figure 2. Weibull distribution CDF plot for Port Dickson and Port Klang



(a) Port Dickson (b) Port Klang

Figure 3. Weibull distribution PDF plot for Port Dickson and Port Klang

Performance indicators

Table 5 shows the best performance for both sites is Weibull distribution that computes by using computational MATLAB software.

The best result for adequacy measure PA and R^2 at Port Dickson is 0.99 and 0.98, respectively. Meanwhile, Weibull distribution scores the best for adequacy measure (PA = 0.99 and R^2 = 0.99) and error measure (NAE = 0.08) at Port Klang site. Thus the Weibull distribution can be used for prediction purposes at both sites.

Table 5. Descriptive of performance indicator

Distribution	Site	NAE	PA	R^2
Weibull	Port Dickson	0.308948	0.992197*	0.984230*
	Port Klang	0.077347*	0.993167*	0.986156*
Lognormal	Port Dickson	0.205216*	0.910724	0.829229
	Port Klang	0.203610	0.886573	0.785832
Inverse Gaussian	Port Dickson	0.263952	0.92141	0.848803
	Port Klang	0.212635	0.923070	0.851863

*best distribution

Conclusion

The descriptive of ozone concentrations in Port Dickson and Port Klang were statistically investigated. This study concluded the best distribution to represent ozone concentration at Port Dickson and Port Klang were found to be Weibull distribution. This distribution performs the best goodness-of-fit data from CDF graph plotted. This distribution also shows best performance indicator with highest score in NAE with smallest error measure (value closest to zero) for Port Klang. It is also scored well in PA and R^2 with highest adequacy measure (value closest to 1) for both sites.

Acknowledgement

Authors would like to thank the Ministry of Higher Education that was funded this study under Fundamental Research Grant Scheme 2013. Last but not least, DoE Malaysia for their permission to utilize the air quality data for this study and Universiti Malaysia Terengganu for their financial support.

References

1. Ghazali N. A., Ramli N. A., Yahaya A. S., Md Yusof N. F. F., Sansuddin N. and Al Madhoun W. A. (2010). Tr ansformation of nitrogen dioxide into ozone and prediction of ozone concentrations using multiple linear regres sion techniques. *Environmental Monitoring and Assessment*, 165: 475 – 489.
2. Munir S, Chen H, and Ropkins K (2012). Modelling the impact of road traffic on ground level ozone concentrat ion using a quantile regression approach. *Atmospheric Environment*, 60: 283 – 291.
3. Yang, W. and Omaye, S., (2009). Air pollutants, oxidative stress and human health. *Research/Genetic tox icology and Environmental Mutagenesis*, 674: 45-54.
4. Shan W., Zhang J., Huang Z. and You L. (2010). Characterizations of ozone and related compounds under the influence of maritime and continental winds at a coastal site in the Yangtze Delta, nearby Shanghai. *Atmospheric Research*, 97: 26 – 34.
5. Donev E., Zeller K. and Avramov A., (2002). Preliminary background ozone concentration in the mountain and coastal areas of Bulgaria. *Environment Pollution*, 117: 281 – 286.
6. Nair P. R., Chand D., Lal S., Modh K. S., Naja M., Parameswaran K., Ravindran S. and Venkataramani S. (2002). Temporal variations in surface ozone at Thumba (8.6^{o}N, 77^{o}E)- a tropical coastal site in India. *Atmospheric Environment*, 36:603 – 610.
7. Song S. K., Shon Z. H., Kim Y. K., Kang Y. H., Oh I. B. and Jung C.H. (2010). Influence of ship emissions on ozone concentrations around coastal areas during summer season. *Atmospheric Environment*, 44: 713 – 723.
8. Lu, H. C., (2000). The statistical characters of PM_{10} concentration in Taiwan area. *Atmospheric Environment*, 36: 491 – 502.
9. Lu, H. C., (2003). Estimating the Emission Source Reduction of PM_{10} in Central Taiwan. *Chemosphere*, 54: 805-814.
10. Forbes C., Evan M., Hastings N. and Peacock B., (2011). Statistical Distributions. 4th Edition. John Wiley & sons. New Jersey.

11. Zerda I. (2012). An experimental comparison of popular estimation method for the Weibull, Gamma, and Gompertz distributions. *Schedae Informaticae*, 20: 67 – 82.
12. Zhibin T., (2009). A new approach to MLE of Weibull distribution with interval data. *Reliability Engineering and System Safety*, 94: 394 – 403.
13. Hadley A. and Toumi R., (2003). Assessing changes to the probability distribution of sulphur dioxide in the UK using a lognormal model. *Atmospheric Environment*, 37: 1461 – 1474.
14. Chikara R. S. and Folks J. L., (1989). The inverse Gaussian distribution: Theory methodology and applications. Marcel Dekker. New York.
15. Tweedie M. C. K. (1957). Statistical properties of inverse Gaussian distribution. I. *The Annal of Mathematical Statistics*, 362 – 367.
16. Department of Environment, Ministry of Natural Resources and Environment (2012). Malaysia Environmental Quality Report 2012; ISSN 0127-6433.
17. Wilks D. S. (2011). Statistical methods in the atmospheric sciences. 3rd Edition. Elsevier. USA.
18. Alshangiti, A.M., Kayid, M. and Alarfaj, B. (2014). A new family of Marshall–Olkin extended distribution. *Journal of Computational and Applied*, 271: 369 – 379.
19. Hamid, H. A., Yahaya, A. S., Ramli N. A., and Ul-Saufie, A. Z. (2013). Finding the best statistical distribution model in PM_{10} concentration modeling by using lognormal distribution. *Journal of Applied Science*, 13 (2): 294 – 300.
20. 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. *Environment Monitoring and Asse ssment*, 180: 573 – 588.
21. Sharma, P., Sharma, P., Jain, S. and Kumar, P., (2013). An intergrated statistical approach for evaluating the exceedance of criteria pollutants in the ambient air of megacity Delhi. *Atmospheric Environment*, 70: 7 – 17.
22. Yahaya, A. S. Ramli, N. A. Ul-Saufie, A. Z., Hamid, H. A., Ahmat, H. and Mohtar, Z. A. (2013). Predicting CO concentrations levels using probability distributions. *International Journal of Engineering and Technology*, 3: 900 – 905.
23. Banan N, Latif M. T., Juneng L. and Ahamad F., (2013). Characteristics of surface ozone concentrations at stations with different backgrounds in the Malaysia Peninsula. *Aerosol and Air Quality Research*, 13: 1090 – 1106.