VOL. 9, NO. 11, NOVEMBER 2014 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences


www.arpnjournals.com

PREDICTING OZONE CONCENTRATIONS LEVELS USING PROBABILITY DISTRIBUTIONS

Ghazali N.A[1], Yahaya A.S.[2], Nasir, M.Y[1] and Mokhtar M.I.Z.[1]
[1]School of Ocean Engineering, Universiti Malaysia Terengganu, Kuala Terengganu, Terengganu, Malaysia
[2]School of Civil Engineering, Engineering Campus, Universiti Sains Malaysia, Nibong Tebal, Pulau Pinang, Malaysia
E-Mail: nurul.adyani@umt.edu.my

ABSTRACT

Ozone (O_3) is one of the strongest atmospheric oxidants and is designated as a criteria pollutant in the atmospheric surface layer. Surface O_3 contributes to a number of environmental problem including adverse effects on health, vegetation and materials, as well as climate forcing. Thus it is necessary to gain a good understanding of the characteristics of O_3 pollution. In this research, four theoretical distributions namely Weibull, Beta, Lognormal and Inverse Gaussian distribution were used to find the best distribution that can fit the O_3 data at Cheras, Selangor. Statistical distribution models are based upon probability and capable of estimating the entire range of pollutant concentration. Probability density functions (pdf) and cumulative distribution functions (cdf) will be used to predict the time of the day with high ozone concentrations and hence can be used as a prediction tool. Parameter estimation for each type of distribution was estimated by using the method of maximum likelihood estimator (MLE). The best distribution was determined using the plots of cumulative distribution functions (cdf) and performance indicator including Root Mean Square Error (RMSE), Prediction Accuracy (PA) and Coefficient of Determination (R^2). The results revealed that the best distribution to represent O_3 concentration level in Cheras for 2010 is the Beta distribution. Based on the prediction using Beta distribution, it can be concluded that the O_3 concentration level in Cheras exceed the Malaysian Ambient Air Quality Guidelines of 0.01 parts per million (ppm).

Keywords: surface ozone, probability distributions, maximum likelihood estimator, performance indicators.

INTRODUCTION

The air pollution in Malaysia has not yet reached a critical level as in other metropolitan areas in Asia, like Jakarta or Manila [2]. However, even outside extreme haze periods, pollution levels increased despite tight regulations and this is exacerbated by the increase in the number of vehicles, distance travelled and growth in industrial production. The haze phenomenon in Malaysia especially in the Klang Valley region is an important and serious problem. The haze phenomenon in Malaysia which contribute to the air pollution event with most of the major air pollutant reading including O_3 concentration exceeds Malaysian standard [4] have already been observed in some urban and industrial regions of Malaysia [21] [26].

Analysis and forecasting of air quality parameters are important topics in current atmospheric and environmental research due to the health impact of air pollution [13]. Growth and environmental protection accompany visions of true sustainable development. Rapid economic growth has paid rich dividends to Malaysians over the years. Although several unhealthy air quality episodes have been recorded, Malaysia generally experienced good to moderate air quality status [9] [10].

Changes in the global climate due to increases in carbon dioxide and other greenhouse gases have created regional changes in temperature, humidity, and precipitation [22]. One of the major problems originating from air pollution in urban areas is pollution caused by photochemical oxidants [13] [5]. Among these pollutants, O_3 adversely affect human health and the environment. O_3 is formed by the photochemical reaction of incoming solar radiation and nitrogen oxides (NO_x), facilitated by a variety of volatile organic compounds (VOCs). Other parameters (temperature and wind speed) also influence the formation of O_3. The composition of O_3 concentrations may depend upon the time of day and the location (urban, industrial, coastal). O_3 plays a major role in oxidation processes and radiation transfer in the atmosphere. Along with carbon dioxide, methane, water vapour and nitrous oxide, it acts as a greenhouse gas and is the third largest contributor to global warming. High O_3 levels not only cause damage to plant, natural materials, but also lead to damage of lung tissues in humans [1]. O_3 can irritate lung airways and cause inflammation much like sunburn [18].

In the past, researchers used empirical methods of classification and regression trees, regression models and neural networks to predict O_3, but these have limitations in interpretation and prediction. Therefore a new approach that can identify the factors influencing and involved in photochemical processes is crucial. Various statistical techniques have been proposed to predict O_3 concentrations. These include multiple linear regression (MLR) [13] [6] [15] [16]; neural network [25]; time series model [24]; classification and regression tree analysis [23]; and application of principal component analysis and clustering techniques [8]. In consideration of the aforementioned, the present study has developed a method for predicting O_3 concentration in Malaysia's urban area, which has a tropical climate and intensive vehicular traffic. Thus far, no prediction tools have been applied to predict surface O_3 in Malaysia except for a study conducted by [13] [14] [15] using MLR. Therefore, this study has assessed the applicability of distribution function

as a new method for prediction of O_3 concentration in Malaysia.

The aim of this research was to obtain the best model to predict surface O_3 concentration level in Cheras, Malaysia. Four theoretical distributions were used to fit the parent distribution of O_3. These distributions were later used to understand the characteristic of O_3 concentration for a one year cycle.

METHODOLOGY

In this research, the monitoring records that were obtained from Department of Environment (DoE) for sites in Cheras was chosen.

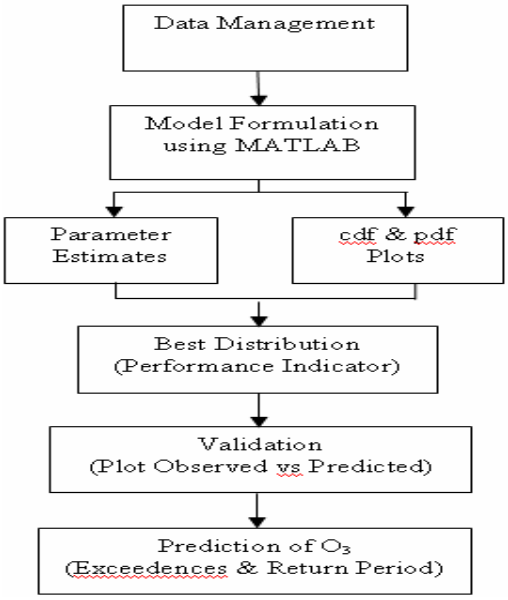

Figure-1. Flow chart of methods.

In order to do the data analysis, MATLAB version 7 R14 SP2 software has been utilized. MATLAB is an equation-solving software package that has proven to have a wide range of applicability to engineering problems. The analysis using MATLAB was done by writing programs as well as using MATLAB distribution functions.

Four parent distributions were used to fit the hourly O_3 concentration in this research. The estimation of two Weibull probability density function parameters (α = shape parameter and σ = scale parameter), Beta probability density function parameters (μ = location parameter and σ = scale parameter), Lognormal probability density function parameters (α = shape parameter and σ = scale parameter), and Inverse Gaussian probability density function parameters (μ = location parameter and σ = scale parameter) were done using the maximum likelihood estimators (MLE) method.

The best distribution was determined using the plots of cumulative distribution functions (cdf) and performance indicator. In this paper, three performance indicators have been used to determine the agreement between predicted and observed O_3 hourly concentration data. The tests are root mean square error (RMSE), prediction accuracy (PA), and coefficient of determination (R^2). The best distribution will represent the O_3 concentration data and will be utilized to predict the exceedences and return period. The overall flow of process in model development is shown in Figure-1.

PROBABILITY DISTRIBUTIONS

Four probability distributions were used for this research. The distributions are Weibull distributions [12], Beta distribution [11], lognormal distribution [19] and Inverse Gaussian distribution [7]. The probability density functions and the estimators of the parameters of the distributions are given in Table-1. The parameters were estimated using the method of maximum likelihood.

Table-1. Probability density functions and its parameter estimates.

Distribution	Probability density function	Parameter estimates
Weibull	$\left(\frac{\alpha}{\sigma}\right)\left(\frac{x}{\sigma}\right)^{\alpha-1} exp\left\{-\left(\frac{x}{\sigma}\right)^{\alpha}\right\}$	$\left(\frac{1}{\alpha}\right)-\left(\frac{\sum_{i=1}^{n} x_i ln(x_i)}{\sum_{i=1}^{n} x_i^{\alpha}}\right)+\left(\frac{1}{n}\right)\sum_{i=1}^{n} ln(x_i)=0 \; \sigma=\left(\frac{1}{n}\sum_{i=1}^{n} x_i^{\alpha}\right)$
Beta	$\frac{x^{\sigma-1}(1-x)^{\mu-1}}{B(\sigma,\mu)}$ where, $B(\sigma,\mu)=\int_0^1 u^{\sigma-1}(1-\mu)^{\mu-1}du \quad 0 \le x \le 1$	$\sigma=\left\{\bar{x}\left(\left\lvert\frac{\bar{x}(1-\bar{x})}{s^2}\right\rvert-1\right)\right\} \mu=\left\{(1-\bar{x})\left(\left\lvert\frac{\bar{x}(1-\bar{x})}{s^2}\right\rvert-1\right)\right\}$
Lognormal	$\left(\frac{1}{x\alpha\sqrt{2\pi}}\right)exp\left\{-\frac{1}{2}\left(\frac{ln(x)-\sigma}{\alpha}\right)^2\right\}$	$\alpha=\left(\frac{1}{n-1}\right)\sum_{i=1}^{n}(ln(x_i)-\sigma)^2 \quad \sigma=\frac{1}{n}\sum_{i=1}^{n} ln(x_i)$
Inverse Gaussian	$\left(\frac{\sigma}{2\pi x^3}\right)^{\frac{1}{2}} exp\left(\frac{-\sigma(x-\mu)}{2\mu^2 x}\right)$	$\mu=\bar{x}; \; \sigma=\frac{(n-1)}{\sum_{i=1}^{n}\left(\frac{1}{x_i}-\frac{1}{\bar{x}}\right)}$

Notation: σ = the scale parameter, α = the shape parameter, μ = the location parameter, n = number of observations, B = beta function, s= standard deviation

PERFORMANCE INDICATORS

Performance indicators were used to determine the distribution that can give the best fit to the data. The three performance indicators are root mean square error (RMSE), prediction accuracy (PA) and coefficient of determination (R^2). For PA and R^2, the value would be from 0 to 1 and the performance indicator with value that is closes to 1 gives the best fit. For RMSE, the value that is closest to zero gives the best fit. Table-2 gives the equations for the performance indicators which have been used by [19] and [17].

Table-2. Performance indicators.

Indicators	Equations
Root Mean Square Error	$\sqrt{\left(\frac{1}{N-1}\right)\sum_{i=1}^{N}(P_i - O_i)^2}$
Prediction Accuracy	$\frac{\sum_{i=1}^{N}(P_i - \bar{O})^2}{\sum_{i=1}^{N}(O_i - \bar{O})^2}$
Coefficient of Determination	$R^2 = \left(\frac{\sum_{i=1}^{N}(P_i - \bar{P})(O_i - \bar{O})}{N \cdot S_{pred} \cdot S_{obs}}\right)^2$

Notation: N = Number of observation, 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.

STUDY AREA

Selangor is the most developed city in the country and belongs to the region with the best infrastructure and telecommunication facilities. The state covers about 125,000 km² and has an annual average temperature of 26 °C. Selangor's geographical position near the center of Peninsular Malaysia has contributed to the state's rapid development (i.e., transportation and industrial hub), which in turn attracts migrants from other countries. The influx of immigrants has contributed further to Selangor's rapid population growth. Cheras is the federal territory of Kuala Lumpur which situated in the middle of Malaysia. Cheras was located at latitude 3° 6' 20.2428" north to the equator and longitude 101° 43' 31.1262" east of Meridian. It has 59.31 km^2 area. The population estimated was high because of job offering and it is a federal territory of Kuala Lumpur. It is a busy city and has a large number of traffic. Therefore, due to congestion of traffic, development and population, Cheras have become the ideal place to study on urban ozone concentration. Figure-2 shows the location of Cheras in Malaysia.

Figure-2. Location of the study area.

RESULTS AND DISCUSSIONS

The hourly O_3 concentration data was obtained for 2010. Table-3 gives summaries of O_3 concentration for Cheras.

Table-3. Descriptive statistics for O_3 concentration.

	Value
Minimum value	0.001
Maximum value	0.169
Mean	0.021
Variance	0.001
Standard deviation	0.026
Median	0.008
Skewness	1.568
Kurtosis	2.121

From Table-3, it shows that minimum O_3 concentration is 0.001 ppm and its maximum value is 0.169 ppm. The mean value was higher than median value; indicate the ozone concentration was high. The coefficient of skewness and kurtosis are greater than zero showing that right skewed distributions are more appropriate to fit the data.

The parameter estimates and performance indicators for the four distributions are given in Table-4.

Table-4. Parameter estimates and performance indicators.

Distribution	Parameter estimates	RMSE	PA	R^2
Weibull	$\sigma = 0.079$ $\alpha = 9.119$	0.057	0.806	0.650
Beta	$\sigma = 0.453$ $\mu = 3.158$	0.146	**0.995**	**0.992**
Lognormal	$\sigma = 1.437$ $\alpha = -4.747$	0.042	0.752	0.565
Inverse Gaussian	$\sigma = 0.021$ $\mu = 0.005$	**0.027**	0.837	0.700

From Table-4, it can be seen that the Beta distribution is the best distribution that can fit the data since it gives the best results for RMSE, PA and R^2. Thus, the Beta distribution can be used for prediction purposes.

The pdf plot of the best distributions using the final scale and location parameters was presented in Figure-2. From Figure-3, the mode of the O_3 concentration data is 0.1 ppm. The graph shows the line is skewed positively to the right indicate that the higher O_3 concentration occur in 2010.

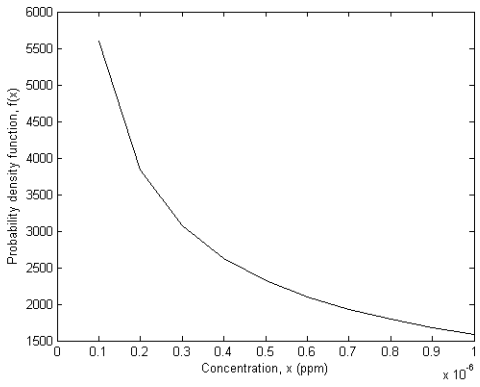

Figure-3. Pdf plots for Beta distribution on 2010.

Figure-4 shows the cdf plot for the best distribution. Based on Figure-4, the observation line was underestimating from start and start overestimate at concentration less than 0.1 ppm. However, the observation line was underestimating again at concentration between 0.1 ppm and 0.2 ppm before it fitted to the theoretical line at concentration 0.4 ppm.

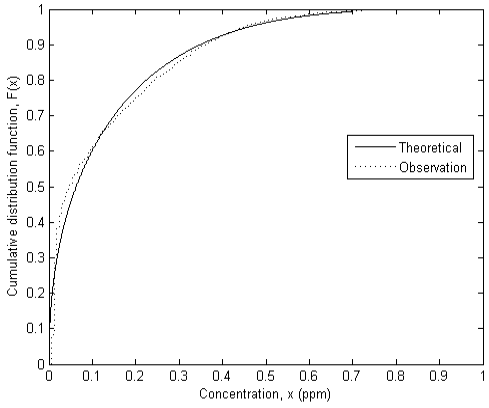

Figure-4. Cdf plots for Beta distribution on 2010.

In order to validate the performance of chosen distribution can fits the data, a plot of observed O_3 concentration versus predicted O_3 concentration using the best distribution which is Beta distribution was done. This plot is given in Figure-5. From the plot, it shows a very good agreement with the value of the coefficient of determination of 0.9920. However, the extreme observations on the right side of the graph cannot be predicted that well using Beta distribution.

Prediction of the probability that the O_3 concentration exceeds the Malaysian Ambient Air Quality Guidelines (0.01 ppm) or not was then tested by using the Beta distribution. As a result, it was found that the value equals to 0.000866 which mean that the predicted return period is equal to 0.3 day. Thus, this shows that the O_3 concentration exceeds the Malaysian standard.

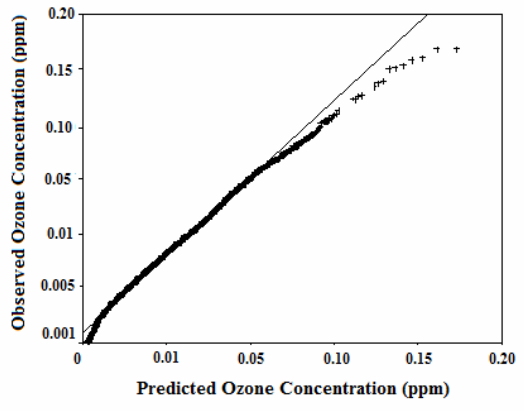

Figure-5. Plot of observed value versus predicted values

CONCLUSIONS

The study had revealed that the air quality status in Cheras were not good at all time. Four distributions were compared and the Beta distribution gives the best fit since two performance indicators gives the best results for this distribution. The scatter plot of observed O_3 concentrations versus predicted values obtained from the Beta distribution shows a very good fit with the coefficient of determination value of 0.9920. However, this prediction is not very good at the extreme right tail of the concentration. The probability that the O_3 concentration exceeds the Malaysian Ambient Air Quality Guidelines was also investigated. The value of the probability is 0.000866 showing that there is exceedences value.

ACKNOWLEDGMENT

This study was funded by Ministry of Higher Education through the Fundamental Research Grant Scheme (FRGS), project number 59314. We express our appreciation to the Universiti Malaysia Terengganu for providing the financial support to carry out this study. We also thank the Department of Environment of Malaysia for their support.

REFERENCES

[1] Abdul-Wahab S. A., Bakheit C. S., Al-Alawi S. M. 2005. Principle component and multiple regression analysis in modelling of ground-level ozone and factors affecting its concentrations. Environmental Modelling and Software. 20, pp. 1263-1271.

[2] Amir A. 2007. Air pollution trends in Petaling Jaya, Selangor, Malaysia. Unpublished Master Thesis. Universiti Putra Malaysia, Malaysia.

[3] Andrew D. Steer, Thomas E. Walton. 2003. Indonesia environment monitor 2003 special focus: Reducing Pollution. Indonesia Air Quality Report, World Bank Indonesia Office http://www.worldbank.or.id.

[4] Awang M., Jaafar A.B., Abdullah A.M., Ismail M., Hassan M.N., Abdullah R., Johan S., Noor H. 2000. Air quality in Malaysia: Impacts, management issues and future challenges. Respirology. 5, pp. 183-196.

[5] Azmi S.Z., Latiff M.T., Ismail A.S., Liew J., Jemain A.A. 2010. Trends and Status of Air Quality at Three Different Monitoring Stations in the Klang Valley, Malaysia. Air Quality, Atmosphere and Health 3, pp. 53-64.

[6] Cardelino C., Chang M., John J.S. 2001. Ozone Prediction in Atlanta, Georgia: Analysis of the 1999 Ozone Season. Air and Waste Management Association 51, pp. 1227-1236.

[7] Chhikara R.S. and Folks J.L. 1989. The Inverse Gaussian distribution: theory, methodology, and applications. CRC Press, USA. ISBN 0824779975.

[8] Davis J.M., Speckman P. 1999. A Model for Predicting Maximum and 8 h Average Ozone in Houston. Atmospheric Environment. 33, pp. 2487-2500.

[9] Department of Environment (DoE), Malaysia. 2010. Malaysia Environmental Quality Report 20011. Kuala Lumpur: Department of Environment, Ministry of Sciences, Technology and the Environment, Malaysia.

[10] Department of Environment (DoE), Malaysia. 2011. Malaysia Environmental Quality Report 2012. Kuala Lumpur: Department of Environment, Ministry of Sciences, Technology and the Environment, Malaysia.

[11] Evans M., Hastings N. and B. Peacock. 2000. Statistical Distribution. 3rd Edition. Wiley-Inter Science, New York. ISBN: 10: 0471371246, pp. 34-42.

[12] Georgepoulos P. and Seinfeld J. 1982. Statistical distributions of air pollutant concentrations. Environmental Science and Technology. 16(54): 401A-415A.

[13] Ghazali N. A., Ramli N. A., Yahaya A. S., Yusof N. F., Sansuddin N., Madhoun W. A. (2010). Transformation of nitrogen dioxide into ozone and prediction of ozone concentrations using multiple linear regression techniques. Environmental Monitoring and Assessment. 165(1): 475-489.

[14] Ghazali N.A., Ramli N.A., Yahaya A.S. 2009. A Study to investigate and Model the Transformation of Nitrogen Dioxide into Ozone Using Time Series Plot. European Journal of Scientific Research. 37(2): 192-205.

[15] Ghazali N. A., Ramli N. A., Yahaya A. S., Yusof N. F., Sansuddin N., Madhoun W. A. 2008. Regression modelling and temporal analysis of hourly ozone concentrations in urban environment of Malaysia. In J. M. Jahi, K. Aiyub, M. R. Razman, K. Ariffin and A. Awang (Eds.), Proceeding International Conference of Human Habitat and Environmental Change 2008. Bangi: UKM.

[16] Hubbard M.C., Cobourn W.G. 1998. Development of A Regression Model to Forecast Ground-level Ozone Concentration in Louisville, KY. Journal of Atmospheric Environment. 32(14/15): 2637-2647.

[17] Junninen H., Niska H., Tuppurrainen K., Ruuskanen J., Kolehmainen M. 2002. Methods for imputation of missing values in air quality data sets. Journal of Atmospheric Environment. 38, pp. 2895-2907.

[18] Kampa M., Castanas E. 2007. Human Health Effects of Air Pollution. Environmental Pollution. 151, pp. 362-367.

[19] Lu H. C. 2002. The statistical character of PM10 concentration in Taiwan Area. Atmospheric Environment. 36(9): 491-502.

[20] Nghiem L.H., Oanh N.T.H. 2009. Comparative Analysis of Maximum Daily Ozone Levels in Urban Areas Predicted by Different Statistical Models. Science Asia. 35, pp. 276-283.

[21] Nichol J. 1998. Smoke Haze in South East Asia: A Predictable Recurrence. Atmospheric Environment, 32, p. 2715-2716.

[22] Ramli N.A., Ghazali N.A., Yahaya A.S. 2009. Modelling of Ozone in Urban Environment in Malaysia. Malaysia: Pusat Pengajian Kejuruteraan Awam, Universiti Sains Malaysia.

[23] Ryan W.F. 1995. Forecasting Severe Ozone Episodes in the Baltimore Metropolitan Area. Atmospheric Environment. 29, pp. 2387-2398.

[24] Slini T., Karatzas K., Moussiopoulos N. 2002. Statistical Analysis of Environmental Data as the Basis of Forecasting: An Air Quality Application. The Science of the Total Environment. 288, pp. 227-237.

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

[26] Yusoff N. F., Ramli N. A., Yahaya A. S., Sansuddin N., Ghazali N.A., Madhoun W.A. 2010. Monsoonal differences and probability distribution of PM_{10} concentration. Environmental Monitoring and Assessment. 163(1-4): 655-667.