2011 International Conference on Environment and Industrial Innovation
IPCBEE vol.12 (2011) © (2011) IACSIT Press, Singapore

Modelling of PM_{10} Concentration in Industrialized Area in Malaysia: A Case Study in Nilai

Norazian Mohamed Noor [1], Cheng Yau Tan [1], Mohd Mustafa Al Bakri Abdullah [2], Nor Azam Ramli [3] and Ahmad Shukri Yahaya [3]

[1] School of Environmental Engineering, Universiti Malaysia Perlis, Pejabat Pos Besar 01007 Kangar, Perlis
[2] School of Material Engineering, Universiti Malaysia Perlis, Pejabat Pos Besar 01007 Kangar, Perlis
[3]School of Civil Engineering, Universiti Sains Malaysia, Engineering Campus 14300 Nibong Tebal, Pulau Pinang

Abstract. Three distributions, namely Weibull, log-normal and gamma were chosen to model the PM_{10} observations at selected industrial area i.e. Nilai, Negeri Sembilan. One-year period hourly average data for 2006 and 2007 was used for this research. For parameters estimation, method of maximum likelihood estimation (*MLE*) was selected. Four performance indicators that are mean absolute error (*MAE*), root mean squared error (*RMSE*), coefficient of determination (R^2) and prediction accuracy (*PA*), were applied to determine the goodness-of-fit criteria of the distributions. The best distribution that fits with the PM_{10} observations in Nilai was found to be gamma distribution. The probabilities of the exceedences concentration were calculated and the return period for the coming year was predicted from the cumulative density function (*cdf*) obtained from the best-fit distributions. For the 2006 data, Nilai was predicted to exceed 150 $\mu g/m^3$ for 2.7 days in 2007 with a return period of one occurrence per 137 days.

Keywords: Particulate matter, statistical analysis, probability distributions, performance indicators, exceedences, return period.

1. Introduction

Exponential development of science and technology nowadays has lead to the rapid growing industrialization which is the major sources of various environmental pollutions, especially air pollution. Air pollutants, specifically particulate matter (PM) smaller than about 10 micrometers, referred as PM_{10}, have received extensive attention, due to its capability to settle in the bronchi and lungs and cause health problems.

Malaysian Ambient Air Quality Guidelines (MAAQG) were issued and target values for annual and daily mean mass concentrations for various air pollutant were established to control and reduce air pollutant levels in the atmosphere. Monitoring data and studies on ambient air quality show that some of the air pollutants in several large cities are increasing with time and are not always at acceptable levels according to the MAAQG. There are very limited data and case studies on air pollution in our country. While the application is almost non-existent in our country, it is an attractive analytical option as it can reasonably predict the return period and exceedences in the succeeding period to meet the evolving information needs of environmental quality management [1].

Many types of probability distributions have been used to fit air pollutant concentrations including Weibull distribution [2], lognormal distribution [3], gamma distribution [4] and Rayleigh distribution [5]. Lu [6] and Chen *et al.* [7] have studied the goodness-of-fit for selected probability distributions by using several performance indicators such as mean absolute error (*MAE*), root means error (*RMSE*), index of agreement (d_2), bias (*B*), normalized absolute error (*NAE*), prediction accuracy (*PA*) and coefficient of determination (R^2).

The goals of this research were to study the statistical characteristics of the observed data, as well as to select the best-fit distribution in order to predict the exceedences and return period of the PM_{10} critical concentration.

2. Materials and Methods

2.1. Data Sets

The datasets consisted of PM_{10} concentration on a time-scale of one per hour (hourly averaged) for 2006 and 2007. The PM_{10} data were taken from the air monitoring station in Nilai, Negeri Sembilan. Nilai is a rapidly growing town surrounded by many industrial areas leading to great air pollution.

2.2. Probability Distributions

Three theoretical distributions namely Weibull, gamma and log-normal distributions are used to fit the entire measured PM_{10} data [8,9]. For parameters estimation, method of maximum likelihood estimation (MLE) was selected.

2.3. Performance Indicators

Four performance indicators (PI) that are mean absolute error (*MAE*), root mean squared error (*RMSE*), coefficient of determination (R^2) and prediction accuracy (*PA*), were applied to determine the goodness-of-fit criteria as to judge which type of parent distribution is the most appropriate to represent the PM_{10} pollutant concentration [10].

2.4. Exceedences and Return Period

Once the best-fit distribution is determined, the cumulative distribution function (cdf) of the fitted distribution was used to calculate the exceedence, or the probability that the event is equaled or exceeded in computed period. The reciprocal of the exceedance probability was calculated so to obtain the return period (also known as the recurrence interval) of the event.

3. Results and Discussion

3.1. Data Description

Table 1 gives the summary of the descriptive statistics for PM_{10} hourly data of Nilai for 2006 and 2007. The mean values for in both years are higher than their respective median which indicates that the pollutants distributions are positively skewed (also called right-skewed). The standard deviations of 2006 data are higher than 2007 data showing that higher variability of the pollutant datasets were observed in 2006. The maximum value is 344.1 and decreased to below MAAQG limit of 150 $\mu g/m^3$ in 2007.

Figure 1 shows the time series plot for PM_{10} concentration in Nilai (2006). The high concentrations continued from the end of September (6520) to the mid of October (hour 7033) due to the land and forest fires in several provinces in Sumatra and Kalimantan, Indonesia coupled with the direct influence of south westerly winds, which had caused Malaysia to experience short periods of slight to moderate haze as a result of transboundary pollution [20]. There is no any concentration above 150 $\mu g/m^3$ for 2007 (Figure 2).

Table 1: Descriptive statistics for PM_{10} concentration

	Nilai	
	2006	**2007**
Valid Data	8760	8465
Missing Data	0	295
Mean	63.5	61.6
Median	57.9	60.4
Standard Deviation	29.3	18.1
Mode	61.8	41.9
Variance	856.6	328.9
Minimum Value	21.6	16.4
Maximum Value	344.1	122.1
Range	322.5	105.7

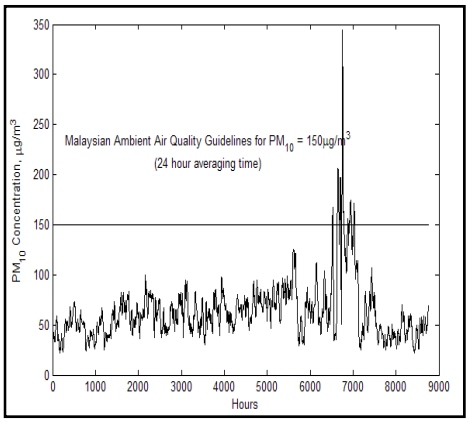

Fig. 1: Time series plot for PM_{10} concentration in Nilai (2006)

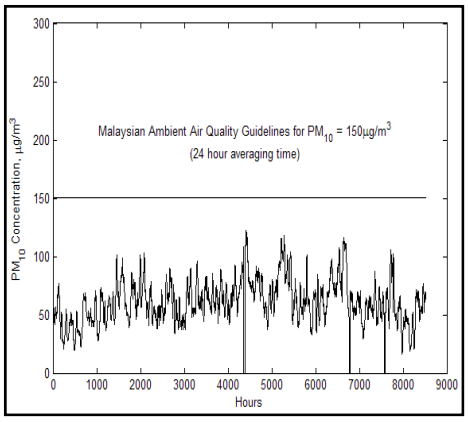

Fig. 2: Time series plot for PM_{10} concentration in Nilai (2007)

3.2. Probability Distributions

Table 4.2 shows the parameter estimates of the three distributions for 2006 and 2007. All the estimates have been obtained using maximum likelihood estimators (MLE).

Table 2: Parameter estimates

Distributions	**Nilai**			
	2006		**2007**	
	α	***β***	***α***	***β***
Weibull	2.22	71.599	3.622	68.261
Gamma	6.423	9.881	5.574	11.056
Log-Normal	0.385	4.071	0.312	4.075

3.2.1 Probability Distribution Functions (pdf)

All plots for both years in Nilai are positively skewed indicating most of the concentrations line within the range of lower value. They are narrowed towards centre indicating they have lesser low and high values than the data of 2006. Most of the mode concentration occurs at the value around 50 µg/m^3.

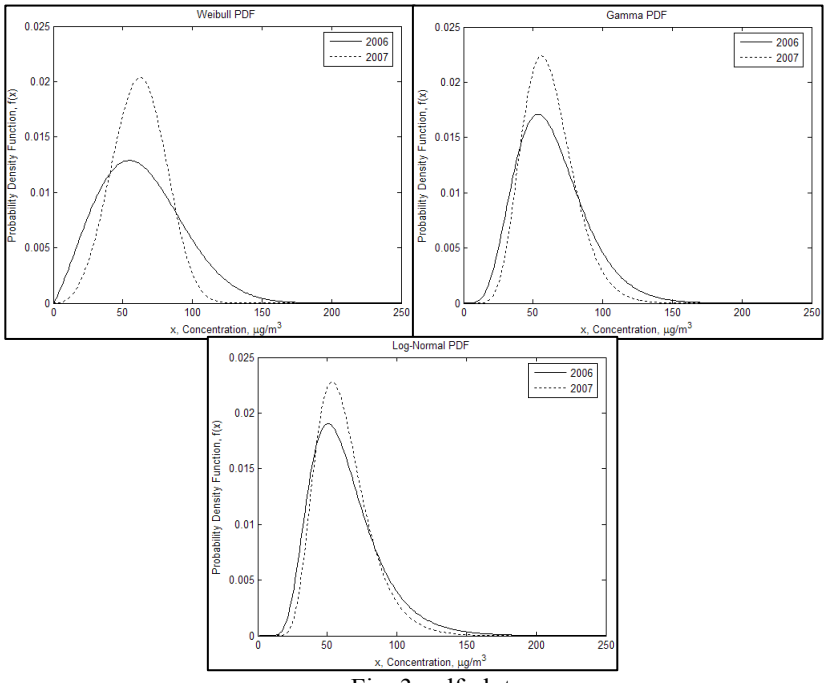

Fig. 3: pdf plots

3.2.2 Probability Distribution Functions (pdf)

cdf plots of 2006 for Nilai show that log-normal distributions fit the observed distribution very well compared to the others. The worse distributions are given by Weibull distribution. For 2007, gamma distribution best fit the observed monitoring data.

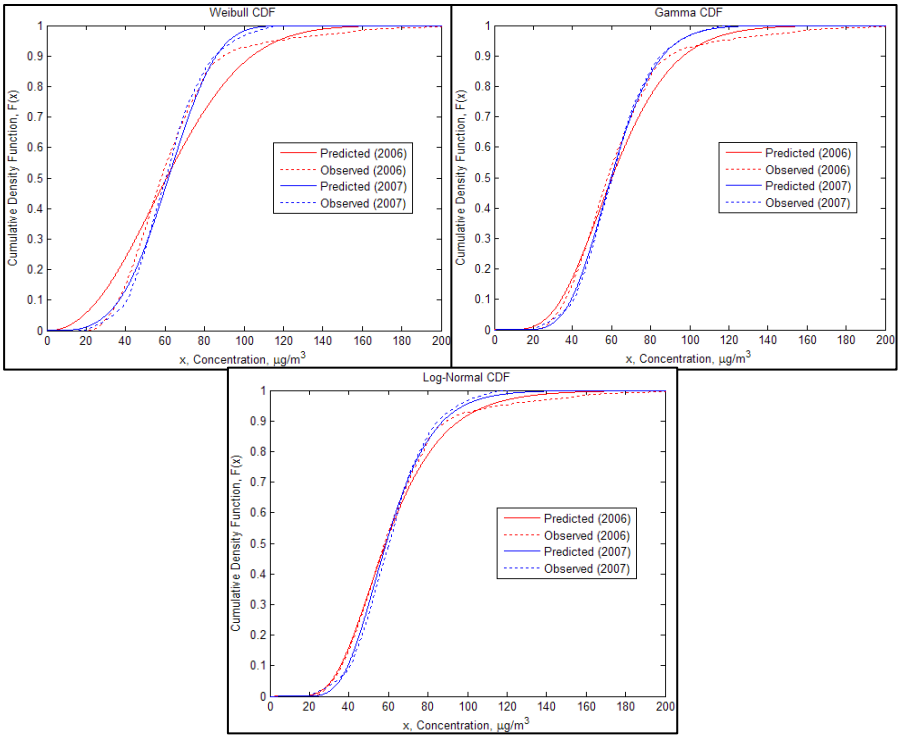

Fig. 4: cdf plots

3.3. Performance Indicators

From Table 3, log-normal distribution fits the 2006 data better than log-normal and Weibull distributions since it give the smallest error (MAE = 2.756; and RMSE= 7.835) and the highest value for R^2 (0.94) and *PA* (0.97). Performance indicators for 2007 show that the smallest values for *MAE* (0.8443) and *RMSE* (1.3896) are given by gamma distribution which also gives the highest values for R^2 (0.9944) and *PA* (0.9973). It can be concluded that the log-normal distribution is the best distribution for 2006 whereas 2007 data is best represented by gamma distribution.

Table 3: Performance indicators value for PM_{10} concentration in Nilai

Distributions	Performance Indicators							
	Mean Absolute Error (*MAE*)		Root Mean Squared Error (*RMSE*)		Coefficient of Determination (R^2)		Prediction Accuracy (*PA*)	
	2006	**2007**	**2006**	**2007**	**2006**	**2007**	**2006**	**2007**
Weibull	8.039	1.973	11.673	2.386	0.852	0.985	0.923	0.993
Gamma	4.324	0.844	9.822	1.39	0.896	0.994	0.947	0.997
Log-Normal	2.756	1.738	7.835	3.042	0.94	0.981	0.97	0.991

3.4. Exceedences and Return Period

The distribution that fits the PM_{10} concentration for 2006 is log-normal distribution and gamma distribution for 2007. From the log-normal cdf plot in Figure 5, the probability that the PM_{10} concentration for 2006 equal or less than 150 µg/m^3 is 0.9927 [that is, $P\{X \leq 150\} = 0.9927$] and the probability that the concentration greater than 150 µg/m^3 is 0.0073 [that is, $P\{X > 150\} = 1 - P\{X \leq 150\} = 0.0073$]. This shows that there will be 2.7 days where the PM_{10} concentration in 2007 which will exceed 150 µg/m^3. Thus the

return period for 2007 is once per 137 days. For 2007, Figure 6 shows the probability that the concentration greater than 150 μg/m^3 is 0 [that is, $P\{X > 150\} = 0$]. There is no incidence where the PM_{10} concentration exceeds 150 μg/m^3 for 2007. Hence, no return period is estimated for the next year.

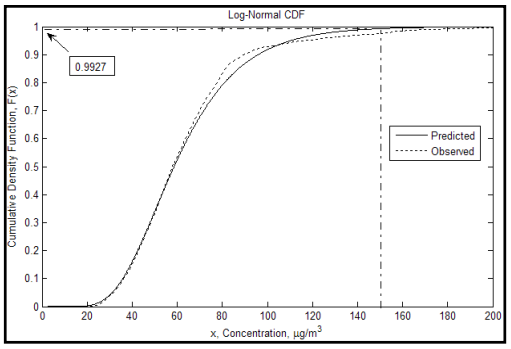

Fig. 5: Estimation of exceedences above MAAQG (150 μg/m^3) in Nilai for 2006 using log-normal cdf plot

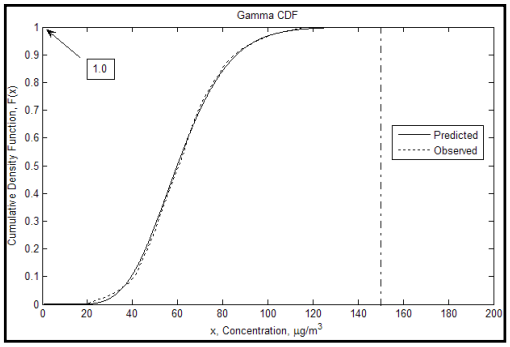

Fig.6: Estimation of exceedences above MAAQG (150 μg/m^3) in Nilai for 2007 using gamma cdf plot

4. Acknowledgement

We wish to thank Department of Environment (DoE) Malaysia for the data.

5. References

[1] Singh, K.P., Warsono, Bartolucci, A.A., and Bae, S. (2001). Mathematical Modeling of Environmental Data. Mathematical and Computer Modelling, 33, pp. 793-800.

[2] Wang, X. and Mauzerall, D. L. (2004). Characterizing Distributions of Surface Ozone and its Impact on Grain Production in China, Japan and South Korea: 1990 and 2020. Journal of Atmospheric Environment, 38 (74), pp. 4383-4402.

[3] Hadley, A. and Toumi, R. (2002). Assessing Changes to the Probability Distribution of Sulphur Dioxide in the UK Using Lognormal Model. Journal of Atmospheric Environment, 37 (24), pp. 455-467.

[4] Singh, P. (2004) Simultaneous Confidence Intervals for the Successive Ratios of Scale Parameters. Journal of Statistical Planning and Inference, 36 (3), pp. 1007-1019.

[5] Celik, A. N. (2003) A Statistical Analysis of Wind Power Density Based on The Weibull and Rayleigh Models at The Southern Region of Turkey. Journal of Renewable Energy, 29, pp. 593-604.

[6] Lu, H. C. (2003) Estimating the Emission Source Reduction of PM_{10} in Central Taiwan. Journal of Chemosphere, 54, pp. 805-814

[7] Chen, J.L., Islam, S. and Biswas, P. (1998). Nonlinear Dynamics of Hourly Ozone Concentrations: Nonparametric Short Term Prediction. Atmospheric Environment, 32, pp. 1839-1848.

[8] Evans, M. Hastings, N. and Peacock, B. (2000) *Statistical Distributions.* 3rd Edition, Wiley, New York.

[9] Kottegoda, N. T. and Rosso, R. (1998) Statistic, Probability and Reliability for Civil and Environmental Engineers. McGraw-Hill, Singapore.

[10] Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J., and Kolehmainen, M. (2004). Methods for Imputation of Missing Values in Air Quality Data Sets. Atmospheric Environment, 38, pp. 2895-2907.