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

International Journal of Engineering and Technology Volume 3 No. 9, September, 2013

Predicting CO Concentrations Levels Using Probability Distributions

[1.]A. S. Yahaya, [1]N.A. Ramli, [2]A.Z Ul-Saufie, [1]H. A. Hamid, [1]H. Ahmat, [1]Z.A Mohtar

[1]School of Civil Engineering, Engineering Campus, Universiti Sains Malaysia,
14300 Nibong Tebal, Seberang Perai Selatan, Pulau Pinang Malaysia.
[2]Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Malaysia.

ABSTRACT

In Malaysia, air pollutant emissions were monitored all over the country to detect any significant change which may cause harm to human health and the environment. This research is focused on carbon monoxide (CO) concentration as it is known to cause severe health impact to human as well as environment. Therefore a well developed system need to be used in order to analyze the trending of all of the pollutants emission inventories. In this research, seven theoretical distributions that are Weibull, gamma, lognormal, Laplace, Rayleigh, log-logistic and inverse Gaussian distributions were developed. It is used to verify and simulate the trend of the monitoring data for CO emission in Kuala Lumpur, which is the capital of Malaysia, in the form of probability density function and hence can be used as a prediction tool. The method of maximum likelihood estimates (MLE) was used for estimating the parameters of the distributions. The best distribution was determined using the plots for the cumulative distribution functions (cdf) and performance indicators. Five performance indicators used are the root mean square error (RMSE), index of agreement (IA), prediction accuracy (PA) and coefficient of determination (R^2). From the performance indicators, it was found that the best distribution to represent the CO concentration level in Kuala Lumpur for 2002 is the inverse Gaussian distribution. Based on the prediction using the inverse Gaussian distribution, it can be concluded that the CO concentration level in Kuala Lumpur does not exceed the Malaysian Ambient Air Quality Guidelines of 9 parts per million (ppm).

Keywords: *Carbon monoxide, probability distributions, performance indicators*

1. INTRODUCTION

Air pollution is the presence of contaminants in the air in concentrations that prevent the normal dispersive ability of the air and interfere with biological processes and human economics. In Malaysia, the Department of Environment, Malaysia (DoE) is one of the bodies that is responsible in monitoring the status of air quality throughout the country to perceive any significant change which may cause harm to human health and the environment. There are 52 monitoring locations throughout the country that belong to the Department of Environmentα- The DoE Malaysia describe that air pollution occurs when air impurities in the form of gaseous or particles are emitted into the atmosphere. Peavy (1985) defined air pollution as the presence in the outdoor atmosphere of one or more air contaminants (i.e., dust, fumes, gas, mist, odor, smoke, or vapor) in sufficient quantities, of such characteristics, and of such duration as to be or to threaten to be injurious to human, plant, or animal life or to property, or which reasonably interferes with the comfortable enjoyment of life or property.

Based on Department of Environment (DoE) report on air quality in 1998, motor vehicles remained the major source of air pollution in the country (Department of Environment, 1998). From 8.9 million motor vehicles registered in 1998, approximately 2 million tonnes of carbon monoxide, 237 000 tonnes of oxides of nitrogen, 111 000 of hydrocarbons, 38 000 tonnes of sulphur dioxide and 17 000 tonnes of particulate matters were emitted into the atmosphere. Generally, in 2002 the air quality was between good to moderate most of the time, except for a number of unhealthy days at various locations in the states of Selangor, Negeri Sembilan and Sarawak.

From the geographical and development point of view, the Klang Valley is the most prone to serious air pollution compared to other parts of the country in 2002. During February to March 2002, the Klang Valley experienced hot and dry weather with reduced rainfall, conditions ideal for peat swamp and forest fires in many areas of Selangor and Kuala Lumpur. This has caused the air quality to deteriorate from moderate to unhealthy level. Based on DoE data air quality status for the Klang Valley in 2002, the number of days with unhealthy air quality conditions ranged from 17 to 67 days.

The aim of this research was to obtain the best model to predict carbon monoxide (CO) concentration level in Kuala Lumpur, Malaysia. Seven theoretical distributions were used to fit the parent distribution of CO. These distributions were later used to understand the characteristic of CO concentration for a one year cycle.

2. PROBABILITY DISTRIBUTIONS

Seven probability distributions were used for this research. The distributions are (1) log-normal distribution (Kao and Friedlander, 1995) (2) gamma distribution (Berger et al.,1982) (3) Weibull distribution (Georgepoulus and Seinfeld ,1982),(4) Laplace distribution (Aryal and Rao, 2005), (5) log-logistic distribution (Singh et al., 2001), (6) inverse Gaussian distribution (Chhikara and Folks, 1989), and (7) Rayleigh distribution (Celik, 2003). 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
Log-normal	$\left(\frac{1}{x\lambda\sqrt{2\pi}}\right)\exp\left\{-\frac{1}{2}\left(\frac{\ln(x)-\sigma}{\lambda}\right)^2\right\}$	$\sigma=\left(\frac{1}{n}\right)\sum_{i=1}^{n}\ln(x_i)$; $\lambda=\left(\frac{1}{n-1}\right)\sum_{i=1}^{n}\left(\ln(x_i)-\sigma\right)^2$
Gamma	$\left(\frac{1}{\sigma\Gamma(\lambda)}\right)\left(\frac{x}{\sigma}\right)^{\lambda-1}\exp\left(-\frac{x}{\sigma}\right)$	$\ln(\lambda)-\psi(\lambda)=\ln\left(\frac{\bar{x}}{g}\right)$; $\sigma\lambda=\bar{x}$
Weibull	$\left(\frac{\lambda}{\sigma}\right)\left(\frac{x}{\sigma}\right)^{\lambda-1}\exp\left\{-\left(\frac{x}{\sigma}\right)^{\lambda}\right\}$	$\left(\frac{1}{\lambda}\right)-\left(\frac{\sum_{i=1}^{n}x_i\ln(x_i)}{\sum_{i=1}^{n}x_i^{\lambda}}\right)+\left(\frac{1}{n}\right)\sum_{i=1}^{n}\ln(x_i)=0$ $\sigma=\left(\frac{1}{n}\sum_{i-1}^{n}x_i^{\lambda}\right)^{\frac{1}{\lambda}}$
Laplace	$\frac{1}{2\sigma}\exp\left[\left(\frac{\lvert x-\mu\rvert}{\sigma}\right)\right]$	$\mu=\text{median}$; $\sigma=\left(\frac{1}{n}\right)\sum_{i=1}^{n}\lvert x_i-\mu\rvert$
Log-logistic	$\frac{e^{-(\ln x-\mu/\sigma)}}{\sigma\{1+\exp[-(\ln x-\mu)/\sigma]\}^2}$	$\sum_{i=1}^{n}\left[1+\exp\left(\frac{\ln x_i-\mu}{\sigma}\right)\right]^{-1}=\frac{n}{2}$ $\sum_{i=1}^{n}\left(\frac{\ln x_i-\mu}{\hat{\sigma}}\right)\frac{1-\exp[\ln x_i-\mu/\sigma]}{1+\exp[(\ln x_i-\mu/\sigma)]}=2$
Inverse Gaussian	$\left(\frac{\sigma}{2\pi x^3}\right)^{1/2}\exp\left(\frac{-\sigma(x-\mu)^2}{2\mu^2 x}\right)$	$\mu=\bar{x}$; $\sigma=\frac{(n-1)}{\sum_{i=1}^{n}\left(\frac{1}{x_i}-\frac{1}{\bar{x}}\right)}$
Rayleigh	$\left(\frac{x}{\sigma^2}\right)\exp\left[-\left(\frac{x^2}{2\sigma^2}\right)\right]$	$\sigma=\left(\left(\frac{1}{2n}\right)\sum_{i=1}^{n}x_i^2\right)^{1/2}$

Notation: μ is the location parameter, σ is the scale parameter, λ is the shape parameter

3. PERFORMANCE INDICATORS

Performance indicators were used to determine the distribution that can give the best fit to the data. The four performance indicators are root mean square error (RMSE), index of agreement (IA), prediction accuracy (PA) and coefficient of determination $\left(R^2\right)$ Table 2 gives the equations for the performance indicators which have been used by Lu (2003) and Junninen et al. (2002).

Table 2: Performance Indicators

Indicators	Equations
Root Mean Square Error	$\sqrt{\left(\frac{1}{N-1}\right)\sum_{i=1}^{N}\left(P_i - O_i\right)^2}$
Index of Agreement	$1-\left[\frac{\sum_{i=1}^{N}\left(P-O_i\right)^2}{\sum_{i=1}^{N}\left(\left\|P_i-\overline{O}\right\|+\left\|O_i-\overline{O}\right\|\right)^2}\right]$
Prediction Accuracy	$\frac{\sum_{i=1}^{N}\left(P_i-\overline{O}\right)^2}{\sum_{i=1}^{N}\left(O_i-\overline{O}\right)^2}$
Coefficient of Determination	$R^2=\left(\frac{\sum_{i=1}^{N}\left(P_i-\overline{P}\right)\left(O_i-\overline{O}\right)}{N.S_{pred}.S_{obs}}\right)^2$

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

4. STUDY AREA

Kuala Lumpur has been chosen as the site for this research as Kuala Lumpur is the capital city of Malaysia. Kuala Lumpur is a federal territory situated in the middle of Malaysia. It is a developing city and the most important city in Malaysia. Figure 1 shows the location of Kuala Lumpur in Malaysia.

From the geographical and development point of view, the Klang Valley is most prone to serious air pollution compared to other parts of the country. In Malaysia, CO emission is mainly due to mobile sources. Large cities like Kuala Lumpur has higher CO concentration principally due to the number of motor vehicles (Department of Environment, 2002).

According to the Ministry of Housing and Local Government (2006), the population size of Kuala Lumpur in 1998 is about 1, 355 558 people and in 2002 is about 1, 445 158 people. From the population size in 1998 and 2002, the estimated value for population growth in Kuala Lumpur is almost 1.6% per year. According to the data from the Ministry of Transport (2006) the vehicle numbers for Kuala Lumpur is increasing with time. The percentage increase of vehicle numbers per year is estimated to be about 2.5%.

Figure 1. Location of the study area

According to the Ministry of Housing and Local Government (2006), the population size of Kuala Lumpur in 1998 is about 1, 355 558 people and in 2002 is about 1, 445 158 people. From the population size in 1998 and 2002, the estimated value for population growth in Kuala Lumpur is almost 1.6% per year. According to the data from the Ministry of Transport (2006) the vehicle numbers for Kuala Lumpur is increasing with time. The percentage increase of vehicle numbers per year is estimated to be about 2.5%.

5. DATA

The CO concentration data was obtained for 2002 and it was collected every hour. Table 3 give summaries of CO concentration for Kuala Lumpur. The unit of measurement is parts per million (ppm).

Table 3: Descriptive statistics for CO concentration

	Value
Total, N	8200
Minimum value	0.01
Maximum value	3.13
Mean	0.61
Variance	0.15
Standard deviation	0.38
Median	0.52
Skewness	1.73
Kurtosis	7.64

Table 3 shows that the minimum CO concentration is 0.01 ppm and its maximum value is 3.13 ppm. There are some missing values (about 6% of the total data). For this research, the missing values are not included in the analysis. The coefficient of skewness and kurtosis are greater then zero showing that right skewed distributions are more appropriate to fit the data.

5. RESULTS

The parameter estimates and performance indicators for the seven distributions are given in Table 4. From Table 4, it can be seen that the inverse Gaussian distribution is the best distribution that can fit the data since it gives the best results for IA, PA and R^2. Thus the inverse Gaussian distribution can be used for prediction purposes. To look at how well the chosen distribution fits the data, a plot of the observed CO concentration versus the predicted values using the inverse Gaussian distribution was done. This plot is given in Figure 2. The plot shows a very good agreement with the value of the coefficient of determination of 0.9904. However the extreme right observations cannot be predicted that well.

Table 4: Parameter Estimates and Performance Indicators

Distributions	Parameter estimates	RMSE	IA	PA	R^2
Weibull	$\sigma = 0.69$ $\lambda = 1.72$	0.072	0.757	0.983	0.965
Gamma	$\sigma = 2.85$ $\lambda = 0.22$	0.049	0.763	0.993	0.986
Log-normal	$\mu = -0.67$ $\sigma = 0.64$	0.078	0.710	0.995	0.989
Laplace	$\mu = 0.52$ $\sigma = 0.27$	0.164	0.699	0.936	0.876
Rayleigh	$\sigma = 0.51$	0.101	0.779	0.973	0.946
Log-logistic	$\mu = -0.65$ $\sigma = 0.35$	0.209	0.658	0.940	0.884
Inverse Gaussian	$\mu = 0.61$ $\sigma = 1.06$	0.092	0.799	0.995	0.990

The inverse Gaussian distribution was then used to predict the probability that the CO concentration exceeds the Malaysian Ambient Air Quality Guidelines (Department of Environment, 2002) which is 9ppm. It was found that this value equals zero. Thus this shows that the CO concentration does not exceed the Malaysian standards.

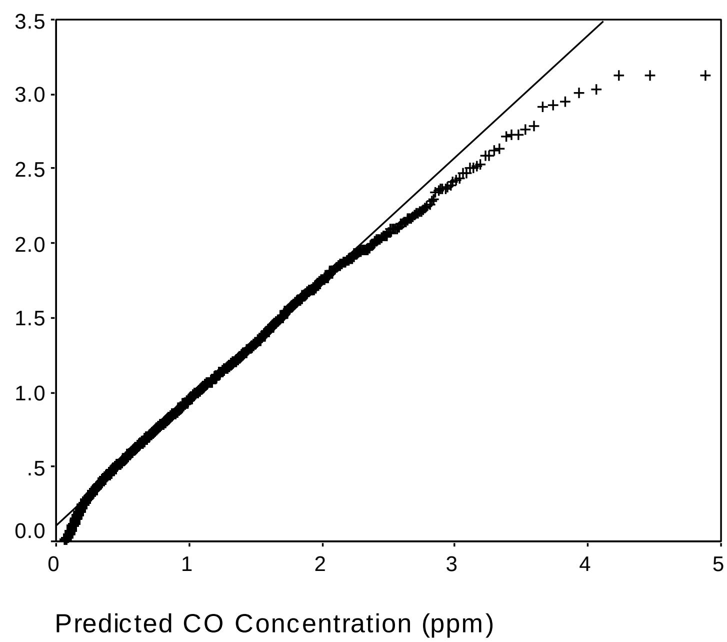

Figure 2: Plot of observed values versus predicted values

5. CONCLUSION

The characteristics of CO concentrations in Kuala Lumpur were investigated. The results show that the mean CO concentration in Kuala Lumpur for 2002 is 0.61ppm with a standard deviation of 0.38ppm which is well below the Malaysian Ambient Air Quality Guidelines of 9ppm. The maximum value is 3.13ppm

Seven distributions were compared and the inverse Gaussian distribution gives the best fit since three performance indicators gives the best results for this distribution. The scatter plot of observed CO concentrations versus the predicted values obtained from the inverse Gaussian distribution shows a very good fit with the coefficient of determination value of 0.9904. However this prediction is not very good at the extreme right tail of the concentration.

The probability that the CO concentration exceeds the Malaysian Ambient Air Quality Guidelines was also investigated. The value of the probability is zero showing that there is exceedences value.

Acknowledgment

Financial supports from the Ministry of Science, Technology and Innovation through the Science Fund project number 04-01-05-SF0115 and Universiti Sains Malaysia are highly appreciated.

REFERENCES

[1]. Aryal, G. and Rao, A. N. V. (2005) Reliability model using truncated skew – Laplace distribution. Nonlinear Analysis. 63 (5 – 7), 2575 – 2585

[2]. Berger, A., Melice, J. L. and Demuth, C. L. (1982) Statistical distributions of daily and high atmospheric SO_2 – concentrations. Atmospheric Environment. 16 (5), 2863 – 2877

[3]. 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 (7), 593 – 604

[4]. Chhikara, R. S. and Folks, J. L. (1989) The inverse Gaussian distribution as a lifetime model. Thechnometrics. 19 (4), 461 – 468

[5]. Department of Environment, Malaysia (1998) Malaysia environment quality report 1998. Malaysia: Department of Environment

[6]. Department of Environment, Malaysia (2002) Malaysia environment quality report 2002. Malaysia: Department of Environment

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

[8]. Kao, A. S. and Friedlander, S. K. (1995) Frequency distributions of PM_{10} chemical components and their sources. Environmental Science and Technology. 29(5), 19 – 28

[9]. Lu, H. C. (2002) The statistical character of PM_{10} concentration in Taiwan Area. Atmospheric Environment. 36 (9), 491 – 502

[10]. Ministry of Housing and Local Government (2007) [Online], [Accessed 25th January 2007]. Available from World Wide Web: http:// www.kpkt.gov.my

[11]. Ministry of Transport (2007) [Online], [Accessed 25th January 2007]. Available from World Wide Web: http:// www.mot.gov.my

[12]. Peavy, H. S., Rowe, D. R. and Tchobanoglous, G. (1985) Environmental engineering. Singapore: McGraw-Hill Co Singh, P. (2001) Simultaneous confidence intervals for the successive ratios of scale parameters. Journal of Statistical Planning and Inference. 36 (3), 1007 – 1019

[13]. 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, 2895-2907.

ABOUT AUTHORS

Ahmad Shukri Yahaya is presently an Associate Professor at the School of Civil Engineering, Universiti Sains Malysia. His main research area is on statistical data analyses, air pollution modeling and simulation.

Nor Azam Ramli received his Ph.D. in environmental engineering from the University of Wales. He is presently a Senior Lecturer at the School of Civil Engineering, Universiti Sains Malysia His main research area is on air pollution and environmental impact assessment.

