Distributional Fit of Carbon Monoxide Data

[1,4]A.M. RAZALI, [2]A.P. DESVINA, [3]M.S. SAPUAN, [4]A. ZAHARIM
[1]Faculty of Science and Technology,
Universiti Kebangsaan Malaysia
43600 UKM Bangi, Selangor D.E.
MALAYSIA

[2]Faculty of Science and Technology,
UIN Sultan Syarif Kasim Riau
INDONESIA

[3]PERMATApintar Gifted Centre,
Universiti Kebangsaan Malaysia
43600 UKM Bangi, Selangor D.E.
MALAYSIA

[4]Solar Energy Research Institute,
Universiti Kebangsaan Malaysia
43600 UKM Bangi, Selangor D.E.
MALAYSIA
mahir@ukm.my, syafiqsapuan87@ukm.my, azami@eng.ukm.my

Abstract: - Air pollution is a problem that concerns many of us all over the world and it is a negative side effect of industrial development. Air pollution from cars and factories, in conjunction with a very humid climate, produce a highly corrosive environment. Land transportation provide a significant contribution to half of the total emission of $PM_{2.5}$, CO, HC and NO_x, where air pollution levels have been exceeded or almost exceeded the ambient air quality standard. This study determine the distributional fit of carbon monoxide (CO) data obtained from Solar Energy Research Institute (SERI), Universiti Kebangsaan Malaysia, Bangi from 16 September 2008 to 16 January 2009. The distribution models used in this study were exponential, gamma, generalized extreme value, lognormal and Weibull distributions. Parameters for all distribution models were estimated by using maximum likelihood method. The goodness of fit of the models were determined by using Kolmogorov-Smirnov and Anderson Darling statistics. The lognormal distribution model was found to fit better than other distribution models.

Key-Words: - Statistical distribution models, air pollution, maximum likelihood method, goodness of fit tests.

1 Introduction

The air pollution problem primarily concerns industrialized countries. However it also has significant impact on countries in the developing world that are making steady growth with their industrial development. The main source of air pollution in urban environments is transportation activities especially from motor vehicles. Land transportation provide a significant contribution to half of the total emission of $PM_{2.5}$, CO, HC and NO_x, where air pollution levels have been or exceeded almost ambient air quality standard [1].

In this study, we fitted five statistical models to CO data. The distribution models used in this study were exponential, gamma, generalized extreme value, lognormal and Weibull distributions.

2 Carbon Monoxide (CO) Data

This CO data was obtained from SERI, Universiti Kebangsaan Malaysia, Bangi. Observations were taken every minute starting from 16 September 2008 to 16 January 2009. However, in this study the data used was the average data in each hour. Figure 1 below is the average CO index in each hour.

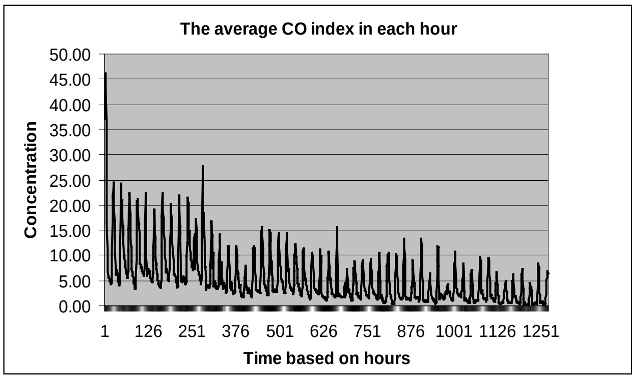

Fig. 1 The average CO index in each hour

3 Statistical Models

3.1 Exponential distribution

Exponential distribution is one of the simplest statistical distribution. It is characterized by λ the only parameter. The probability density function (pdf) for this exponential distribution is given in the equation below,

$$f\left(x\right)=\lambda e^{-\lambda x} \tag{1}$$

The cumulative distribution function (cdf) for this distribution function is shown below,

$$F\left(x\right)=1-e^{-\lambda x} \tag{2}$$

where $x \geq 0$ is the CO index data (in this study) and $\lambda > 0$ is the parameter [7,10]. The maximum likelihood estimator for the parameter λ is defined by equation (3),

$$\hat{\lambda}=\frac{n}{\sum_{i=1}^{n} x_i} \tag{3}$$

where *n* is the total number of observations.

3.2 Gamma distribution

The probability density function (pdf) for this distribution is given by

$$f\left(x\right)=\frac{\lambda}{\Gamma\left(\gamma\right)}\left(\lambda x\right)^{\gamma-1}\exp\left(-\lambda x\right) \tag{4}$$

It is characterized by two parameters and they are γ and λ. The cumulative distribution function (cdf) for this distribution function is,

$$F\left(x\right)=\left({}^{x}\!/\!{}_{\lambda}\right)^{\gamma-1}\frac{\left[\exp\left({}^{-x}\!/\!{}_{\lambda}\right)\right]}{\left[\lambda\right]\Gamma\left(\gamma\right)} \tag{5}$$

where $x \geq 0$ is the CO index data (in this study), $\gamma > 0$ is a shape parameter, $\lambda > 0$ is a scale parameter and $\Gamma\left(\gamma\right)$ is the gamma function [7,10].

The maximum likelihood estimator for the shape and scale parameters are defined by the equations (6) and (7) shown below [4,7]:

$$\hat{\gamma}=\frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2} \tag{6}$$

and

$$\hat{\lambda}=\frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2} \tag{7}$$

3.3 Generalized extreme value distribution

The generalized extreme value distribution is a flexible three parameter model that combines the Gumbel, Frechet dan Weibull maximum extreme value distribution [8]. The probability density function (pdf) for this distribution is given by (8),

$$f\left(x\right)=\exp\left[-\left[1-k\left(\frac{\left(x-\lambda\right)}{\delta}\right)\right]^{1/k}\right]\left[1-k\left(\frac{\left(x-\lambda\right)}{\delta}\right)\right]^{\frac{1}{k-1}}\left(\frac{1}{\delta}\right) \tag{8}$$

for $k \neq 0$; and for $k = 0$:

$$f\left(x\right)=\exp\left[-\exp\left({}^{\left(x-\lambda\right)}\!/\!{}_{\delta}\right)\right]\exp\left({}^{\left(x-\lambda\right)}\!/\!{}_{\delta}\right)\left(\frac{1}{\delta}\right) \tag{9}$$

The parameter for this distribution are k, δ and λ. The cumulative distribution function (cdf) is:

$$F\left(x\right)=\exp\left[-\left[1-k\left({}^{\left(x-\lambda\right)}\!/\!{}_{\delta}\right)\right]^{1/k}\right],\ for\ k \neq 0 \tag{10}$$

$$F\left(x\right)=\exp\left[-\exp\left[-\left({}^{\left(x-\lambda\right)}\!/\!{}_{\delta}\right)\right]\right],\ for\ k = 0 \tag{11}$$

where $x \geq 0$ is the CO index data, $k > 0$ is a shape parameter, $\delta > 0$ is a scale parameter and $\lambda > 0$ is a location parameter.

According to [8], the probability weighted moments estimator for the shape, scale and location parameters are defined by the equations shown below,

$$q=\frac{2m_1-\bar{x}}{3m_2-\bar{x}}-\frac{\log(2)}{\log(3)} \qquad (12)$$

$$m_j=\tfrac{1}{n}\sum_{i=1}^{n}\tfrac{(i-1)(i-2)\cdots(i-j)}{(n-1)(n-2)\cdots(n-j)}x_{i:n},\ \ j=1,2,\cdots \qquad (13)$$

$$\hat{k}=q+q^2 \qquad (14)$$

$$\hat{\delta}=\frac{(2m_1-\bar{x})\hat{k}}{\Gamma\left(1+\hat{k}\right)\left(1-2^{\hat{k}}\right)} \qquad (15)$$

$$\hat{\lambda}=\bar{x}+\hat{\delta}\,\frac{\left\{\Gamma\left(1+\hat{k}\right)-1\right\}}{\hat{k}} \qquad (16)$$

where q is a quantile, m_j is a weighted moments.

3.4 Lognormal distribution

The lognormal distribution is specified by the two parameters, μ and σ^2. The probability density function (pdf) for this distribution is:

$$f(x)=\left[{}^{1}\!/_{x\sigma\sqrt{2\pi}}\right]\exp\left[\left({}^{-1}\!/_{2\sigma^2}\right)\left(\ln x-\mu\right)^2\right] \qquad (17)$$

where $x\ge 0$ is the CO index data, $\mu>0$ is the mean and $\sigma>0$ is the standard deviation of the lognormal distribution [7,10].

The maximum likelihood estimator for the shape and scale parameters are defined by the equation shown below [4,7]:

$$\hat{\mu}=\frac{1}{n}\sum_{i=1}^{n}\ln\left(x_i\right) \qquad (18)$$

$$\hat{\sigma}^2=\tfrac{1}{n}\left[\sum_{i=1}^{n}\left(\ln x_i\right)^2-\frac{\left(\sum_{i=1}^{n}\ln x_i\right)^2}{n}\right] \qquad (19)$$

3.5 Weibull distribution

The Weibull distribution is a generalization of the exponential distribution. The name of this distribution was taken from the name of the Swedish physicist, Wallodi Weibull. This distribution has been used in many studies, such as in the study of human disease mortality. Weibull distribution is specified by two parameters α and β. The probability density function (pdf) for this distribution is

$$f(x)=\frac{\beta}{\alpha}\left(\frac{x}{\alpha}\right)^{\beta-1}\exp\left[-\left(\frac{x}{\alpha}\right)^{\beta}\right] \qquad (20)$$

The cumulative distribution function (cdf) is:

$$F(x)=1-\exp\left[-\left(\frac{x}{\alpha}\right)^{\beta}\right] \qquad (21)$$

where $x\ge 0$ is the CO index data, $\beta>0$ is a shape parameter and $\alpha>0$ is a scale parameter [12].

The maximum likelihood estimator for the shape and scale parameters are defined by the equations (22) and (23) below [12]:

$$\hat{\alpha}=\left[\left(\frac{1}{n}\right)\sum_{i=1}^{n}x_i^{\hat{\beta}}\right]^{1/\hat{\beta}} \qquad (22)$$

$$\hat{\beta}=\frac{n}{\frac{1}{\hat{\alpha}}\sum_{i=1}^{n}x_i^{\hat{\beta}}\ln x_i-\sum_{i=1}^{n}\ln x_i} \qquad (23)$$

where x_i are the generated data sample and *n* the total number of sample in the data set.

4 4 Parameter Estimation

Parameters for all distribution models were estimated using statistical software such as S-Plus, Easyfit and SYSTAT. The methods of estimation used were maximum likelihood and probability weighted moments. The results were shown in Table 2 below.

5 Goodness of Fits Test

In general, the goodness of fit test described here rely on the relation of the empirical distribution function of the observations to the hypothesized distribution function. The distributional fit of CO data were determined by using Kolmogorov-Smirnov test statistics D_n and Anderson-Darling test statistics *AD*. Both were nonparametric test that calculated based on the cumulative distribution function (cdf) and the probability density function (pdf) of a continuous variable. The hypothesis test for goodness of fit will reject H_0 if the *p*-value for confidence interval at 90% falls below some critical value [14].

5.1 Kolmogorov-Smirnov statistics

This test statistic calculate the maximum vertical distance between the empirical and hypothetical cdf. It is applied on the assumption that a theoretical continuous cdf is completely specified with known parameters. It is defined as in equation (24) below

$$D_n=\sup_x\left|F_n(x)-F_0(x)\right| \qquad (24)$$

where $F_n(x)$ and $F_0(x)$ are the empirical and theoretical continuous cdf, respectively, D_n is the maximum vertical distance between $F_n(x)$ and $F_0(x)$. The random variable x is representing (CO data). The minimum the value of D_n statistic, the better is data fits to the distribution [14].

5.2 Anderson-Darling statistics

Anderson-Darling has suggested the following equation

$$A^2 = -\sum_{i=1}^{n} \frac{(2i-1)\left\{\ln F_0(x_i) + \ln\left[1 - F_0(x_{n-i+1})\right]\right\}}{n} - n \quad (25)$$

where $x_i, x_{i+1}, \cdots, x_{n-i+1}$ are the observations in increasing order and $F_0(x)$ are the empirical and theoretical continuous cdf. A smaller value of AD statistic gives the best distribution fit to the data [14].

6 Result and Discussion

6.1 Descriptive statistics

The histogram shown in Fig. 2 below is the frequency plots of the average CO index of each hour at the station in Universiti Kebangsaan Malaysia, Bangi.

Based on this Figure 2, it can be seen that the data is skewed to the right, so the five distribution functions above can be used to model CO data. The descriptive statistics for the data is shown in Table 1 below.

Table 1. Descriptive statistics for CO data index

Parameter	Statistics
Mean	5.3470
Standard deviation	5.1781
98th percentile	20.7880
Minimum	0.16
Maximum	46.15
Kurtosis	10.473
Skewness	2.484

Base on the above table, it can be seen that the mean CO index is 5.3470 *ppm*. The standard CO index is 10 *ppm*, this condition show that the average CO index is below the standard CO index harmful to human body. The standard deviation for CO index is 5.1781 and the 98th percentile is 20.7880. The minimum index is 0.16 and maximum is 46.15. The skewness and kurtosis for the data are 2.484 and 10.473 respectively. So, the distribution is skewed to the right.

6.2 Probability distribution functions

The results of the analysis can be seen in Table 2. The parameter for all distribution models were estimated by using either maximum likelihood or probability weighted moments. The estimated parameter values of each distribution are shown below.

Table 2
Parameters value for the distribution models

Distribution	Parameter
Exponential	$\lambda = 0.18702$
Gamma	$\gamma = 1.0663$
	$\lambda = 5.0146$
Gen. Extreme Value	$\lambda = 0.26195$
	$\delta = 2.6579$
	$k = 2.895$
Lognormal	$\mu = 1.2538$
	$\sigma = 0.9759$
Weibull	$\alpha = 1.286$
	$\beta = 5.4679$

The plot shown in Fig.2 is the probability density function (pdf) for exponential, gamma, generalized extreme value, lognormal and Weibull. A comparison between each plot, where all plot have skewness that leads to the right. This plot shows that all of the probability density function (pdf) is consistent with the results of the skewness calculated for CO index data. This plot show that the distribution models fit of CO index data approximately the lognormal distribution model.

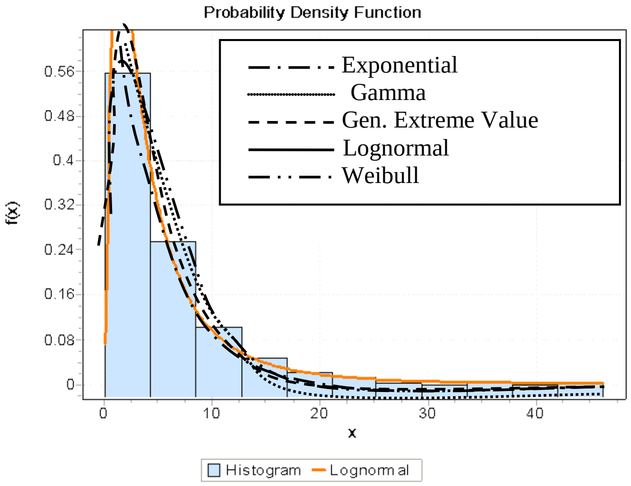

Fig. 2 The probability density function of CO index data for each distribution

Table 3 The goodness of fit test statistics for each distribution models

Distribution	Kolmogorov -Smirnov	Anderson-Darling
Exponential	0.06739	10.539
Gamma	0.05395	6.8528
Gen.Extreme Value	0.04607	4.756
Lognormal	**0.03674**	**2.8478**
Weibull	0.05713	7.8657

Table 3 shows the goodness of fit test statistics for all distribution models. Based on the table above the values for D_n and *AD* is minimum for lognormal distribution (D_n=0.03674 and *AD*=2.8478). This shows that the lognormal distribution is the best fit for CO index data compared with the other distributions.

7 Conclusion

This study was carried out to determine the distributional fit of carbon monoxide (CO) data obtained from Solar Energy Research Institute (SERI), Universiti Kebangsaan Malaysia, Bangi from 16 September 2008 to 16 January 2009. The five statistical distributions used were exponential, gamma, generalized extreme value, lognormal and Weibull distributions. The goodness of fit test of Kolmogorov-Smirnov and Anderson-darling statistics were used to determine the distribution models fit of CO data. The results shows that the lognormal distribution models is the best fit for CO index data than other distribution models.

Acknowledgement

This research was financed in part by the Fundamental Research Grant (Code: FRGS/1/2011/ST/UKM/02/8) which is duly acknowledge by authors.

References:

[1] Balakrishnan, N. and Nevzorov, V.B. *A primer on statistical distributions*. John Wiley & Sons, Inc., 2003.

[2] Cassidy, B.E., Mary A.A.A., Timothy A.A., Daniel B.H., P. Barry, R., Charlee, W.B. and Luke, P.N. Particulate matter and carbon monoxide multiple regression models using environmental characteristics in a high diesel-use area of Baguio City, Philippines. *Science of the Total Environmental* Vol. 381, 2007, pp. 47-58.

[3] Chen, C. (2006). Test of fit for the three-parameter lognormal distribution, *Computational Statistics & Data Analysis*, Vol. 50, 2006, pp. 1418-1440.

[4] Evans, M., Nicholas, H. and Brian, P. *Statistical distribution*. John Wiley & Sons, Inc. 2000.

[5] Gokhale, S. and Khare, M. Statistical behavior of carbon monoxide from vehicular exhausts in urban environments. *Environmental Modelling & Software* Vol. 22, 2007, pp. 526-535.

[6] Jakeman, A.J., Simpson, R.W. and Taylor, J.A. Modelling distributions of air pollutant concentrations-III: hybrid modelling deterministic statistical distributions. *Atmospheric Environment* Vol. 22, No.1, 1988, pp. 163-174.

[7] Krishnamoorthy, K. *Handbook of statistical distributions with applications*. Chapman & Hall/CRC. 2006.

[8] Kotz, S. and Saralees, N. *Extreme value distributions theory and applications*. Imperial College Press. 2000.

[9] Kuchenhoff, H. and Thamerus, M. Extreme value analysis of Munich air pollution data. *Environment and Ecological Statistics* Vol. 3, 1996, pp. 127-141.

[10] Lee, T.E. and John, W.W. *Statistical methods for survival data analysis*. John Wiley & Sons, Inc. 2003.

[11] Modic, J. Carbon monoxide and COHb concentration in blood in various circumstances. *Energy and Buildings* Vol. 35, 2003, pp. 903-907.

[12] Rinne, H. *The weibull distribution a handbook*. Chapman & Hall/CRC. 2009.

[13] Rumburg, B., Alldredge, R. and Claiborn, C. Statistical distributions of particulate matter and the error associated with sampling frequency. Hennes *Atmospheric Environment* Vol. 35, 2001, pp. 2907-2920.

[14] Thode, H.C. *Testing for normality*. Marcel Dekker, Inc. 2002.

[15] Viras L.G., Paliatsos A.G. and Fotopoulos A.G. Nine year trend of air pollution by CO in Athens, Greece. *Environment Monit Assess*. Vol. 40, 1996, pp. 203-214.