UNIVERSITI SAINS MALAYSIA

Laporan Akhir Projek Penyelidikan Jangka Pendek

Modelling of Carbon Monoxide Concentration in Major Towns in Malaysia : A Case Study in Penang, Kuching and Kuala Lumpur

by

Assoc. Prof. Ahmad Shukri Yahaya

Assoc. Prof. Dr. Nor Azam Ramli

2008

PUSAT PENGAJIAN KEJURUTERAAN AWAM
KAMPUS KEJURUTERAAN

FINAL REPORT

USM SHORT TERM GRANT

MODELLING OF CARBON MONOXIDE CONCENTRATION IN MAJOR TOWNS IN MALAYSIA: A CASE STUDY IN PENANG, KUCHING AND KUALA LUMPUR

PREPARED BY

AHMAD SHUKRI BIN YAHAYA

NOR AZAM RAMLI

FEBRUARY 2008

MODELLING OF CARBON MONOXIDE CONCENTRATION IN MAJOR TOWNS IN MALAYSIA: A CASE STUDY IN PENANG, KUCHING AND KUALA LUMPUR

Ahmad Shukri Yahaya
Nor Azam Ramli
Pusat Pengajian Kejuruteraan Awam
Kampus Kejuruteraan
Universiti Sains Malaysia
14200 Nibong Tebal
Seberang Perai Selatan
Pulau Pinang

TABLE OF CONTENTS

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 on CO as they are known to trigger adverse health impact to human as well as environment. Therefore, a well developed model need to be used in order to analyze the trends of the pollutants emission concentration. Eight distributions, Weibull, log-normal, gamma, Rayleigh, log-logistic, Pareto, Laplace and inverse Gaussian were used to find the best distribution that can fit the CO observations at three sites; Seberang Perai, Kuala Lumpur and Kuching. The characteristics of the observation were established and the probabilities of the exceedences concentration were calculated. Hourly data for 1999 and 2002 were used from this research. From this research, the best distributions that fit with the CO observations were obtained. Generalized lambda distributions were also used ti fit the CO data. The probabilities for air pollutants emissions exceeding the Malaysian Ambient Air Quality Guidelines (MAAQG) have been successfully predicted. For the 1998 data, Kuala Lumpur was predicted to exceed 9ppm for 2.5 days in 1999 with a return period of one occurrence per 146 days. However, Seberang Perai and Kuching do not exceed the MAAQG. Based on the 2002 data, it can be concluded that the CO concentration levels in Seberang Perai, Kuala Lumpur and Kuching does not exceed the Malaysian Ambient Air Quality Guidelines of 9 ppm. Similar results were also obtained using the generalized lambda distribution. The probability density functions and cumulative distribution functions for two extreme value distributions have been fitted. For 1998, the best distributions that fit the observations are the Frechet distribution using *ff2* for all three sites. By using the maximum daily data, the Gumbel distribution is the best distribution for the three sites. For 2002, the Gumbel distribution is the best distribution for all sites using both the *ff2* and maximum daily data. The probability density functions and cumulative distribution functions obtained in this research can be used to predict the return period for the coming year. From these extreme value distributions and its cumulative distribution functions, it can be concluded that the CO concentration levels in Seberang Perai and Kuching does not exceed the Malaysian Ambient Air Quality Guidelines of 9 ppm based on the 1998 and 2002 data. However, the CO concentration levels in Kuala Lumpur exceed the Malaysian Ambient Air Quality Guidelines of 9 ppm based on the 1998 data. The probabilities of exceedences are 0.14 and 0.18 respectively for *ff2* and maximum daily data. For 2002, the probability of exceedences is 0.016 and 0.061 respectively for *ff2* and maximum daily data. In this research, the probabilities for air pollutants emissions exceeding the Malaysian Ambient Air Quality Guidelines have been successfully predicted

CHAPTER 1

INTRODUCTION

Monitoring data and studies on ambient air quality show that some of the air pollutants in several large cities in Malaysia are increasing with time and are not always at acceptable levels according to the national ambient air quality standards. The industrialization policy has started to impose costs in terms of pollution and the degradation of urban environment. Depletion of air and water quality, and contamination by industrial wastes has become more serious in recent years. Among them, air pollution is one major issue that needs to be dealt with before it causes more harm to human health and the environment.

This research work develops a tool needed to analyze the statistical characteristics of carbon monoxide monitoring records in Kuala Lumpur, Kuching and Seberang Perai. Nine distributions were used namely Weibull distribution, log-normal distribution, beta distribution, gamma distribution, Rayleigh distribution, log-logistic distribution, Pareto distribution, Laplace distribution and inverse Gaussian distribution.

Three sites compared are Kuala Lumpur, Kuching and Seberang Perai. Kuala Lumpur and Seberang Perai are two major conurbations in West Malaysia while Kuching represents East Malaysia.

1.1 PROBLEM STATEMENT

Malaysia experienced good to moderate air quality status most of the time. However, several unhealthy air quality statuses were also recorded at several parts of the countries that include Kuala Lumpur, Kuching and Seberang Perai. These sites not only experienced a rapid growth of population but also industrialization which is accompanied by a growing number of vehicles that contribute to air pollution problem.

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.

In the Northern region of the West coast of Peninsular Malaysia, comprising the states of Perlis, Kedah, Pulau Pinang and Perak, the overall air quality ranged between good and moderate most of the time except at Perai. As Perai is a heavily industrialized area with several petrochemical complexes, the air quality remained at the moderate level more than 90 percent of the time.

The overall air quality in Sarawak deteriorated due to transboundary haze pollution between July to September 2002. Except for Limbang station, all other stations in Sarawak including Kuching recorded unhealthy levels between 3 to 22 days, due to high levels of particulate matter in the air.

The DoE is the body that is responsible for monitoring and acquiring air pollutants monitoring records in Malaysia. They provide continuous measurement and maintain records of pollutants in the ambient air. However, these data have not been analyzed statistically. Therefore, it is very important to conduct research on its characteristics of the air pollutants monitoring records and its statistical prediction. Based on those factors, a method is required to predict the number of days where air pollutant exceeds the Malaysian Ambient Air Quality Guidelines (MAAQG).

1.2 OBJECTIVES

The aim of this research was to obtain the best model to predict carbon monoxide (CO) concentration level in three major cities in Malaysia. Eight 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.

The objectives are as follows:

1. To fit all the distributions given in section (a).
2. To find the 'best' distribution to describe the data.
3. To find the exceedences and return period of the CO critical concentration.

For the first objective, eight probability distributions were used and the best distribution to describe the data was obtained. Then generalized lambda distributions were used and the best distribution to describe the data was obtained. Finally extreme value distributions were fitted to the CO data using the threshold method and maximum daily data.

CHAPTER 2

LITERATURE REVIEW

2.1 AIR POLLUTION

The natural problems, which are considered as "air pollution", can generally be accepted by the definition of pollution of ambient or outdoor air.

Canter (1996), defined air pollution as the presence in the outdoor atmosphere of one or more pollutants in such quantities and of such duration as may tend to be injurious to human, plant, or animal life or property, or which may unreasonably interfere with the comfortable enjoyment of life or property, or the conduct of business.

Among different environmental pollution problems, air pollution is reported to cause the greatest damage to health and loss of welfare from environmental causes in Asian countries (Hughes, 1997). Although data and study on air pollution in Malaysia are very limited, one review that had been done by Afroz *et al.* (2003), indicated that suspended particulate matter (SPM) and nitrogen dioxide (NO_2) are among the predominant pollutants.

In line with the need for regional harmonization and for easy comparison with countries in the region, the Department of Environment (Malaysia) revised its index system in 1996, and the Air Pollutant Index (API) was adopted. The API system of Malaysia closely follows the Pollutant Standard Index (PSI) system of the United States (Department of Environment, Malaysia, 1996).

The Malaysian Ambient Air Quality Guidelines (MAAQG), which forms the basis for calculating the API, is presented in Table 2.1 and the Air Pollutant Index (API) in Table 2.2.

Table 2.1: Malaysian Ambient Air Quality Guidelines

Pollutant	Averaging Time	Malaysia Guideline	
		ppm	$\mu g/m^3$
Particulate matter (PM_{10})	24 Hour 1 Year		150 50
Carbon monoxide	1 Hour 8 Hour	30 9	35 10
Nitrogen dioxide	1 Hour 24 Hour	0.17 0.04	320
Sulphur dioxide	1 Hour 24 Hour	0.13 0.04	350 105
Ozone (O_3)	1 Hour 8 Hour	0.10 0.06	200 120
Total Suspended Particulate (TSP)	24 Hour 1 Year		260 90
Lead	3 Month		1.5

(source: Department of Environment, Malaysia, 2002)

In MAAQG, the pollutants that are involved in this research are included as shown in Table 2.1. The guideline valuesfor carbon monoxide (CO) is at 30 ppm (for 1 hour) and 9 ppm (for 8 hours and above).

Table 2.2: Malaysia Air Pollutant Index (API)

API	DESCRIPTOR
0 – 50	Good
51 – 100	Moderate
101 – 200	Unhealthy
201 – 300	Very unhealthy
> 300	Hazardous

(Source: Department of Environment, Malaysia, 1996)

The air quality status for Malaysia is reported based on the Air Pollutant Index as shown in Table 2.2. The API is used to indicate the pollutants status based on their concentration. If the value of the API is 100 for CO this is equivalent to 9 ppm of concentration level which is the threshold value for CO for an 8 hour average. Based on Table 2.2, it shows that when the API is above 100, the air quality status is unhealthy.

2.2 AIR QUALITY IN MALAYSIA

In 1998, nine new Continuous Air Quality Monitoring (CAQM) stations were added in addition to the existing 29 stations. These 38 CAQM stations are located strategically throughout the country. The station serve to continuously monitor the presence of air pollutants emitted from sources such as motor vehicles, industries and open burning. On the average, the overall air quality in Malaysia was good throughout the year 1998, except in the vicinity of Miri, Sarawak. This is mainly due to forest and peat fires around Miri and aggravated by the dry weather conditions (Department of Environment, Malaysia, 1998).

In 2002, there are already 50 CAQM monitoring stations to monitor the status of air quality throughout the country. The overall air quality for Malaysia throughout 2002 was good to moderate most of the time. However, several unhealthy air quality status were also recorded at several parts of the countries especially in major cities for instance Kuala Lumpur that experienced 30 unhealthy days where the air quality hovered between API 101 – 200. The major contributor for this unhealthy event is peat swamp and forest around Selangor and Kuala Lumpur area (Department of Environment, Malaysia, 2002).

2.3 SOURCES OF AIR POLLUTION

The three major sources of air pollution in Malaysia are mobile sources, stationary sources, and open burning sources. Afroz *et al.,* (2003), review the air pollution in Malaysia based on the reports of the air quality monitoring in several large cities in Malaysia, which cover air pollutants such as carbon monoxide (CO), sulphur dioxide (SO_2), nitrogen dioxide (NO_2), ozone (O_3), and suspended particulate matter (SPM). However, PM_{10} concentration is more preferable than SPM for determining air pollution in Malaysia. The results indicated that particulate matter (PM_{10}) and nitrogen dioxide (NO_2) are the predominant pollutants. Other pollutants such as CO, NO_x, SO_2, and Pb are also observed in several big cities in Malaysia.

Based on Figure 2.1, from 1998 to 2002, emission from mobile sources were the most significant contributors to air pollution, followed by emission from stationary sources, such as power station, industrial fuel consumption, industrial process, municipal waste and domestic fuel consumption.

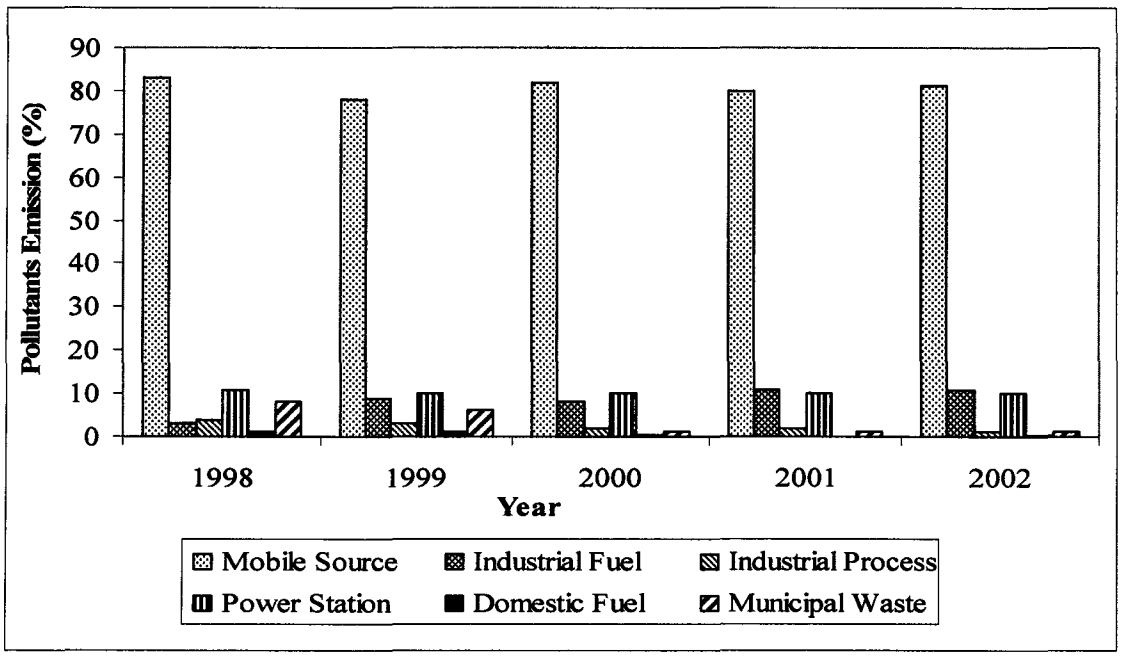


Figure 2.1: Estimated Air Pollution Load by Major Sources
(Source: Department of Environment, 2002)

2.4 HEALTH IMPACT OF AIR POLLUTION

The impact of air pollution is broad, especially for human being where it can cause several significant effects (including carcinogenic effects). Such health problems include cardiac arrhythmias, reducing lung function, asthma, chronic bronchitis and increasing respiratory symptoms, such as sinusitis, sore throat, dry and wet cough, and hay fever (WHO, 1998).

There are very limited number of studies that relate air pollution to its health impacts in Malaysia. The lack of data gathering for environmental epidemiological analysis makes it difficult to estimate the health impact of air pollution. During the Indonesian forest fires in 1997, outpatient visits in Kuching, Sarawak increased between two and three times through the peak periods of smoke haze and respiratory disease outpatient visits to Kuala Lumpur General Hospital increased from 250 to 800 per day (WHO, 1998).

2.5 CARBON MONOXIDE (CO)

Carbon monoxide is a colourless, odourless and tasteless gas. Carbon monoxide is produced in large quantities as a result of the incomplete combustion of fossil fuels and biomass (Godish, 1997). It is approximately 50% heavier than air, of which it is a normal constituent (WHO, 1987).

Carbon monoxide gas is chemically inert under normal conditions and has an estimated atmospheric mean life of about two and a half months (Peavy *et al.*, 1985). It is known that the main source of CO emission is from motor vehicle exhaust, while the other sources include industrial processes and open burning activities (Ibrahim, 2004).

The annual eight hourly average concentrations of carbon monoxide throughout the country measured from 1996 to 2002 were well below the Malaysian Ambient Air Quality Guideline as shown in Figure 2.2. The concentrations of CO were consistently higher in urban areas, principally due to motor vehicles.

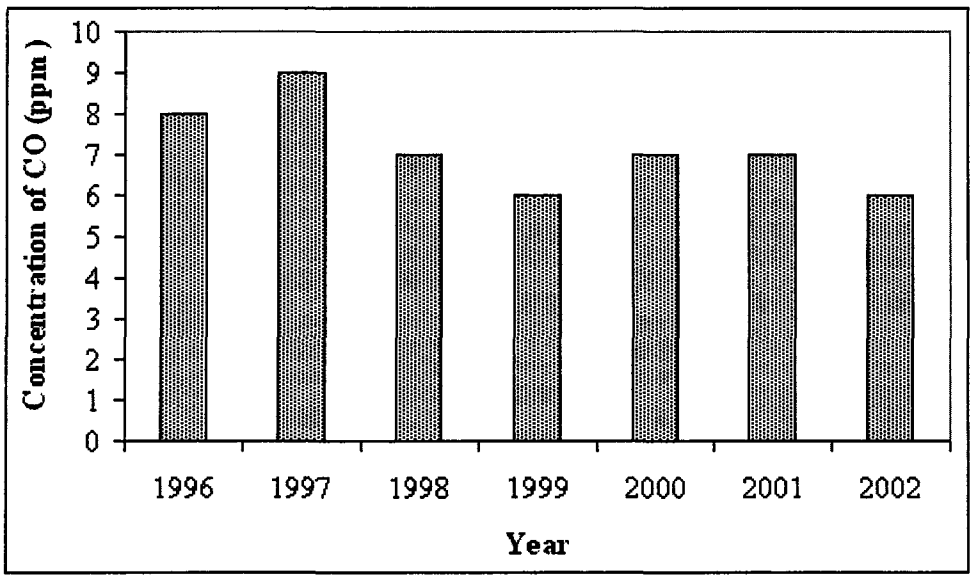


Figure 2.2: Annual Average of CO Concentration in Malaysia from 1996 to 2002 (Source: Department of Environment, 2002)

In Figure 2.3, the flow chart of the carbon monoxide peak episode is shown. Because of its atmosphere lifetime it is possible to neglect its chemical reactions and its physical removal term, so that carbon monoxide concentration at time *t*, [CO (*t*)], could be attributed to three main address: dispersion, emissions and concentrations at previous times (Maffeis, 1999).

Roughly, carbon monoxide emissions depend on ambient temperature (in the case of cold emissions), on traffic characteristics for examples the vehicle fleet composition and age, and traffic speed, on macroscale dispersion phenomena, that could be represented in a very simplified way by the synoptic weather type, on microscale turbulence which is strictly bound to wind speed and finally on local energy balance in the canopy layer, only partially explained by seasonality and synoptic conditions (Maffeis, 1999).

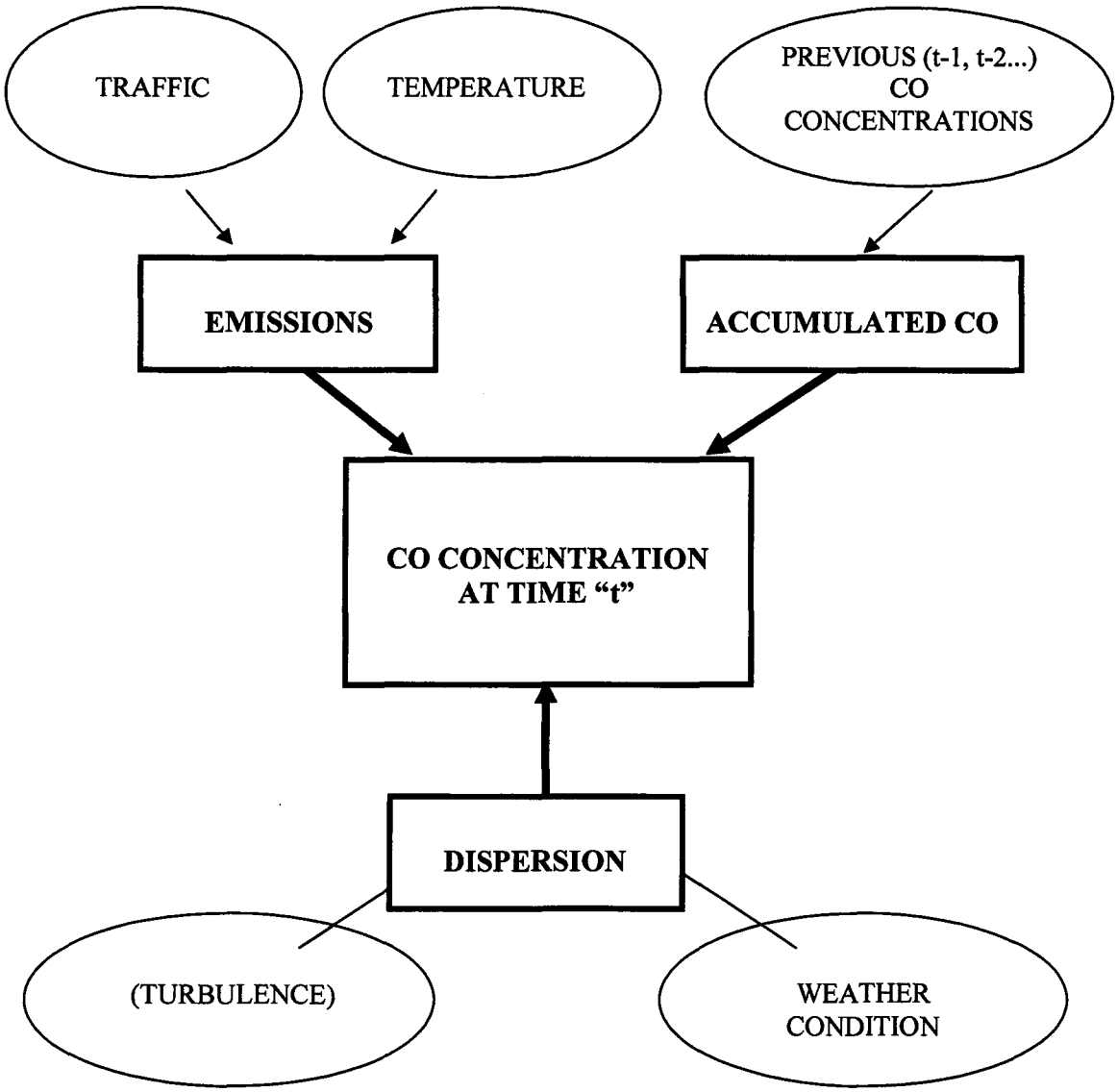


Figure 2.3: Flow chart of the carbon monoxide peak episode
(Source: Maffeis, 1999)

2.5.1 CO EFFECTS ON HUMAN AND ENVIRONMENT

Potentially harmful exposures to CO occur from a variety of sources and environments. At very high concentration which is more than 1000 ppm, CO exposures may be lethal with death resulting from asphyxiation. At lower concentration which is between 2 ppm to 80 ppm, CO may cause a variety of neurological symptoms including headache, fatigue, nausea and in some cases, vomiting. Asphyxiation and sub-lethal symptoms are usually caused by poorly vented combustion appliances, idling motor vehicles in closed

environments, excessive CO production and inadequate ventilation associated with a variety of industrial occupational activities and smoke inhalation from structural fires (Godish, 1997).

Fellenberg (2000) stated that CO endangers human specifically by its tendency to combine with hemoglobin as well as by the fact that it is an element in smog formation. The combination of CO and hemoglobin in blood will produce carboxyhemoglobin (COHb), thus reducing the capability of the blood to carry oxygen. The binding with other haeme proteins causes changes in the function of the affected organs such as the brain and the cardiovascular system, and also the developing fetus. It can impair human concentration, slow reflexes and make people confused and sleepy.

2.6 STATISTICS IN ENVIRONMENTAL ENGINEERING

Hoshmand (1998) suggested that statistics have played a significant role in the analysis and interpretation of data. With data interpretation using statistical analysis, outcomes from the analysis can be utilized as prediction tools that have become the major aim in environmental engineering.

There are large numbers of books available on general statistical methods, stochastic process and application of statistics to business problems, the social sciences, engineering and the health sciences. However, there are very few books in the environmental sciences (Ott, 1995). It is known that environmental data for air pollutants concentration comprise thousands of variables. Therefore, statistical analyses in air pollution are very important and need to be comprehensively developed.

There are many statistical methods developed to analyze data sets. However, environmental monitoring records are frequently asymmetrical and skewed to the right (that is with long tail towards high concentrations), so the validity of classical procedures may be questioned (Gilbert, 1987).

At present, the studies on fitting of distributions have not been done in Malaysia.

2.7 PROBABILITY DENSITY FUNCTIONS (PDF) APPLICATIONS

The emission levels and meteorological conditions influence the concentrations of air pollutants. When the parent probability distribution of air pollutants is correctly chosen, the specific distribution can be used to predict the mean concentration and probability of exceeding a critical concentration (Lu and Fang, 2003).

Selecting appropriate probability models for the data is an important step in environmental data analysis. These probability models may become the basis for estimating the parameters to meet the evolving information needs of environmental quality management (Singh *et al.*, 2001).

Probability density function (pdf) has been applied successfully in many physical phenomena such as wind speed, rainfall, river discharges and air quality. PDF was fitted to the data of vehicular emission in Chennai, India to predict the concentration of carbon monoxide in the ambient atmosphere (Harikrishna and Arun, 2003). In their research, the data was analyzed by using the Input Analyser Software developed by Software Inc. Ten standard probability density functions were fitted to the data based on the goodness of fit using Kolmogorov – Smirnov (KS) test and Anderson-Darling (AD) test.

Probability density function can be represented by its cumulative probability distribution function. It was obtained by adding (accumulating) the individual increments of the probability density function. The cumulative distribution function is defined as the probability that any outcome in X is less than or equal to a stated limiting value x (McBean and Rovers, 1998). In mathematical terms,

$$F(x) = \Pr[X \leq x] = \int_{-\infty}^{x} f(x)dx \tag{2.1}$$

There are many types of parent probability distribution used to fit the air pollutant concentration data. In this research, nine distributions were used to analyze the trending of PM_{10} and CO emissions in three sites.

2.8 DISTRIBUTIONS

In this research, probability distributions have been used to describe the air pollutant concentrations in Malaysia. The following continuous distributions have been used; log-normal distribution (Mage and Ott, 1984; Kao and Friedlander, 1995; Lu, 2002), gamma distribution (Berger *et al.*, 1982; Holland and Fitz-Simons, 1982), Weibull distribution (Georgepoulos and Seinfeld, 1982; Jakeman *et al.*, 1986), beta distribution (Morel *et al.*, 1999), Laplace distribution (Aryal and Rao, 2005), log-logistic distribution (Singh *et al.*, 2001), Pareto distribution (Singh, 2004), inverse Gaussian distribution (Chhikara and Folks, 1989; Mudholkar and Tian, 2002) and Rayleigh distribution (Celik, 2003).

2.8.1 THE WEIBULL DISTRIBUTION

The Weibull probability distribution function was selected because it is the most common and simple function when dealing with extreme events (Seinfeld and Pandis, 1998). The Weibull density function contains two parameters, sigma (σ) and lambda (λ). The σ value acts as a scale parameter and λ value acts as the shape parameter that determines the form and 'skew' of the distribution (Piegorsch and Bailer, 1997).

This distribution is versatile enough to be a good model for many distributions of data that are mound-shaped but skewed. Another advantage of the Weibull distribution is that the numbers of relatively straightforward techniques exist for actually fitting the model to

data (Scheaffer, 1995). Weibull distribution also has better flexibility and one of the asymptotic distributions of general extreme value theory (Kottegoda and Rosso, 1998).

Pang *et al.,* (2003) stated that estimators for Weibull distribution parameters has been considered by many researches such as Kao (1959), Dubey (1966), Wyckoff *et al.,* (1980), Zanakis (1997, 1979), Zanakis and Mann (1981) and Cohen and Whitten (1982). Kottegoda and Rosso (1998) cited that Weibull distribution has greater flexibility and closer fit to failure strengths and times of failure, and it is also one of the asymptotic distributions of general extreme value theory.

Wang and Mauzerall (2004) had conducted a study on distribution of surface ozone and its impact on grain production in China, Japan and South Korea from 1980 to the year of 2020. They used an integrated approach that incorporates atmospheric modeling, plant exposure yield response studies and economic assessment to estimate the value of the yield lost due to ozone exposure. Weibull function has been widely used in this study to express the relationship between ozone exposure and crop yield reduction. From the exposure-response function from individual case study plot, it was found that the function with the median Weibull parameter values best represent the characteristics of exposure-response function for a given crop species.

2.8.2 THE LOG – NORMAL DISTRIBUTION

The log-normal distribution is more widely used to represent the distribution of air pollutant concentration. The log-normal parent frequency distribution of air pollutants can provide a good result for evaluating the mean concentration. However, the tail of the theoretical parent distribution sometimes diverges in the high concentration region. It does not fit the high concentration very well and sometimes will cause large error in the high concentration region (Berger *et al.*, 1982).

Specifically, Ott (1990) demonstrated that under certain conditions the log-normal probability model theoretically is the correct model to choose for representing pollutant concentrations.

The log-normal distribution was found to be appropriate for representing particulate matter concentration distribution (Jakeman *et al.*, 1986; Kao and Friedlander, 1995; Lu, 2002).

Hadley and Toumi (2002) have applied log-normal distribution for assessing changes to the probability distribution of sulphur dioxide in UK. Their research has shown that the 2-parameter log-normal distribution can be a very good description of annual mean daily sulphur dioxide concentrations for a wide range of ambient levels, time periods and monitoring site types. By using log-normal distribution, they discover that the distribution gives consistently better fit to the data than the normal distribution. Seasonal and meteorological characterizations of daily data further confirm that the log-normal is good and, in most cases, significant fit.

2.8.3 THE GAMMA DISTRIBUTION

Lu (2003) found that the gamma distribution is the best distribution to represent the performance of high pollutants concentration in Taiwan. It is an important distribution for non - Gaussian statistical modeling. The distribution is positively skewed and it is often used as a model of life spans testing and many other related fields (Pang *et al.*, 2002).

Gamma distribution has been used for runoff modeling (Singh, 2004). Two parameter gamma distributions were used in his research for computing the direct runoff hydrograph resulting from a complex storm. To reduce the computational effort, unit kernel approach is suggested. The used of unit kernel renders the S-curve invariant with the sampling interval. The new procedure enhances the applicability of the two parameter gamma distribution or the Nash model for rainfall runoff modeling.

Romano *et al.* (2004) in his paper said that for the fuel oil plant, a gamma distribution has been chosen to represent particulate matter emissions which are measured from two electric power plants in Italy: a coal plant, consisting of two boilers and a fuel oil plant, of four boilers. The pollutants considered are SO_2, NO_x, CO and particulate matter.

2.8.4 THE RAYLEIGH DISTRIBUTION

The Rayleigh distribution is a special and simplified case of the Weibull distribution (Celik, 2003). The Weibull distribution is a special case of the generalized gamma distribution, while the Rayleigh distribution is a subset of the Weibull distribution. The Weibull distribution is a two parameter distribution, while the Rayleigh distribution has only one parameter. This makes the Weibull distribution somewhat more versatile and the Rayleigh distribution somewhat simpler to use (Johnson, 2001).

According to Celik (2003), the research on wind power at the southern region of Turkey shows that the Weibull model is better in fitting the measured monthly probability density distributions than the Rayleigh model. It is shown from the monthly correlation coefficient values of the fits, the Weibull model provided better power density estimations in all 12 months than the Rayleigh model but from the annually correlation coefficient values of the fits, the Rayleigh model is better than Weibull model.

2.8.5 THE LOG – LOGISTIC DISTRIBUTION

The two parameter log-logistic distribution model is the most appropriate distribution for the CO concentration data sets. Moreover, the log-logistic model predicts the 'most frequently occurring' values as well as 'rare' events for example the extreme percentiles with reasonable accuracy (Gokhale and Khare, 2005).

Schorp and Leyden (2001) have used log-normal, log-logistic and Pearson (Type V) distributions for airborne nicotine concentrations in hospitality facilities. They discover that the Pearson Type V distribution is somewhat better, but clearly, the experimental data do not fit a smooth function. When a large data set is available, a log-normal distribution function fits well. In this case, the log-logistic distribution is used with log-normal and Pearson (Type V) to define the best distribution that fits well with the data.

2.8.6 THE PARETO DISTRIBUTION

The Pareto distribution is used to model the distribution of personal income, the distribution of population sizes and stock price fluctuations, where the parameter a's are used for minimum income, minimum population size and minimum stock price, respectively (Singh, 2004). So far the Pareto distribution has not been used in air pollution modeling.

Mudholkar and Tian (2002) in their paper said that the time evolution of the average output shows almost no aggregate fluctuations for the log-normal economy, but large fluctuations in the Pareto economy even in the absence of aggregate shocks. In particular, the variance of the average aggregate output in the Pareto case is one order of magnitude greater than the variance of the log-normal case.

2.8.7 THE LAPLACE DISTRIBUTION

It also called the double exponential distribution. It is the distribution of differences between two independent variates with identical exponential distributions (Abramowitz and Stegun, 1972). So far, this distribution has not been used in air pollution modeling but it is used in financial and economic behavior modeling.

Stanley *et al.* (1996) and Bottazzi and Sechhi (2003) show that the growth rates of firms are generally well fitted by a Laplace distribution. Such a finding can be shown to derive from the firm's size being distributed as a power law. In fact, simulated data for the firm's growth rates are well approximated by a (asymmetric) Laplace distribution.

According to Gatti *et al.* (2005), simulations of the model replicate surprisingly well an impressive set of stylized facts, particularly two well-known universal laws: (i) the distribution of firm's size (measured by the capital stock) is skewed and described by a power law; (ii) the distributions of the rates of change of aggregate and firm's output follow a similar Laplace distribution.

2.8.8 THE INVERSE GAUSSIAN DISTRIBUTION

The inverse Gaussian distribution is a well-known competitor of the Weibull, gamma and log-normal distributions in modeling asymmetric data from various scientific fields. In reliability and life testing, the inverse Gaussian distribution is particularly useful in situations where early failures dominate (Chhikara and Folks, 1989).

Rao (1973) describes the Gaussian law as having the maximum entropy among all distributions on the real line and with given mean and variance. However, it is easy to see that in the characterization, it is enough to restrict the support to the real line and fix only the variance.

Mudholkar and Tian (2002) obtain an entropy characterization of the inverse Gaussian distribution, and used it to develop an inverse Gaussian entropy test of the composite hypothesis of normality. In their paper, they have offered an entropy characterization of the inverse Gaussian distribution and used it to develop a test for the corresponding composite goodness-of-fit hypothesis. The test is consistent against all absolutely continuous distributions with nonnegative support and is seen to have good power properties compared with its competitors. The inverse Gaussian distribution cannot be directly characterized as a maximum entropy distribution. Therefore, it does not fall in the categories of maximum entropy distributions.

2.9 GENERALIZED LAMBDA DISTRIBUTION (GLD)

The generalized lambda distribution which will be used in this research was developed by Karian and Dudewicz (1999). Research on fitting a distribution has been actively done a long time ago. Pearson (1895) (see Karian and Dudewicz (2000)) gave a four parameter system of probability density functions and fitted the parameters by what he called the method of moments. Ramberg and Schmeiser (1972, 1974) introduces the four parameter generalized lambda distribution for generating random variates using Monte Carlo simulation method. A system with tables was developed for fitting a wide variety of curve shapes by Ramberg *et al.* (1979). Dudewicz and Karian (1996, 1999) developed the generalized lambda distribution by the method of moments and percentiles respectively. King and Macgillivray (1999) developed the starship estimation method which was used for the generalized lambda distributions. This method can be used for the full parameter space and is flexible, allowing choice of both the form of the generalized lambda distribution and of the nature of fit required. The authors gave examples of its use in fitting data and approximating distributions were also given. However, care is needed when fitting and using such quantile-defined distributional families that are rich in shape but have complex properties.

Okur (1988) used the generalized lambda distribution of Ramberg and Schmeiser (1974) to fit the air pollutant concentrations in Turkey. The data used are the daily smoke and sulphur dioxide (SO_2) concentrations from the urban areas of Ankara, Turkey for the years of 1984 and 1985. The data were obtained from two representative stations. The

generalized lambda distribution was compared with the two-parameter lognormal distribution. Two goodness-of-fit criteria were adopted to judge the relative successes of individual fittings. These are the (1) Kolmogorov-Smirnov type statistic and (2) absolute deviation. The author found that the expected number of exceedences under the GLD is in close agreement with the actual exceedences. However, the performance of the GLD and lognormal models were found to be similar when the two goodness-of-fit criteria were used.

Ahmad Shukri Yahaya *et al.* (2006) used the generalized lambda distribution of Ramberg and Schmeiser (1974) to fit the rainfall data in Malaysia. . The data used was the monthly data (measured in millimeters, mm) from 1951 from 1980 obtained from one of the meteorological stations in Pulau Pinang The chi-square goodness-of-fit test was used to judge the fit of the distribution. The author found that the generalized lambda distribution fits well the rainfall data and can be used to predict the probability of exceedences.

2.10 EXTREME VALUE DISTRIBUTIONS (EVD)

The utilization of an extreme value distribution involves the fitting of a continuous distribution such as Pearson or log Pearson and/or an asymptotic distribution such as Gumbel's extreme value distribution. Whereas normal and lognormal distributions are referred to as central-fitting distributions, the tails of distributions are better described by distribution other than the normal or lognormal distributions (McBean and Rovers, 1998). Kottegoda and Rosso (1998) stated that extreme value theory which is used in storm, flood, wind, sea waves, and earthquake estimation, dates back to the pioneering works by Frechet (1927) and Fisher and Tipett (1928). This theory was extensively developed by Gumbel(1958) following the extremal type theorem originated by Ginedenko (1943) (Please refer to Kottegoda and Rosso, 1998) . The extreme value theory is concerned with probability calculations and the statistical inference associated with the extreme values of random processes (Leong *et al.*, 2001). Leong *et al.*, (2001) stated that Tipett laid the theoretical foundations in 1928 when he showed that there could be only three possible types of extreme value limit distributions that are Gumbel distribution, Frechet distribution and Weibull distribution.

The extreme value theory has also been used in air pollution study. Lu (2003) used extreme value theory to fit the monthly maximum data and high concentration data of air pollutants concentration over a specific percentile. By this, the cumulative probability extremes and return period can be computed. Lu and Fang (2003) used the log-normal, Weibull and Type V Pearson distributions to compare with Type I asymptotic distribution of extreme value theory and Type I two parameter exponential distributions. The data used are monthly maximum PM_{10} concentration and high PM_{10} concentration over a specific percentile. For high concentrations, Type I two parameter exponential distribution give the best fit. Lu (2004) used the daily average PM_{10} concentration at five monitoring stations in Central Taiwan to fit extreme value distributions. The data was for six years from January 1994 to December 1999. Extreme values greater then the 85^{th} percentile were considered. Lu (2004) fitted the two parameter exponential

distribution and asymptotic distribution of extreme value. He concluded that the two parameter exponential distribution gives the best fit.

2.11 THE PERFORMANCE INDICATORS

The distribution parameters could be estimated from the measured data by the maximum-likelihood estimate (MLE). The MLE always gives a minimum variance estimate of parameters (Lu, 2003).

Once the distribution parameters were determined, the goodness-of-fit criteria were used to judge which type of parent distribution is the most appropriate one to represent air pollutant concentration in high concentration region (Lu, 2003).

Lu (2003) used three statistical indicators which are Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and Index of Agreement (IA) in his research to evaluate the goodness-of-fit for the theoretical distributions and judge which distribution is appropriate to represent the PM_{10} distribution in high concentration region in Taiwan.

The results of the goodness-of-fit criteria in the high PM_{10} concentration region for different fitted distributions were shown and all the criteria indicate that the two-parameter exponential distribution is the best distribution and the gamma distribution is the next best distribution to represent the performance for high PM_{10} concentration. However, the log-normal distribution is better than the gamma distribution at His-Twen station. Therefore, the two-parameter exponential distribution is the most appropriate distribution to represent high PM_{10} concentrations.

CHAPTER 3

METHODS

In this research, the monitoring records that were obtained from Department of Environment (DoE) for sites in Kuala Lumpur, Seberang Perai and Kuching were chosen. The hourly CO concentrations for 2002 were obtained.

In order to do the data analysis, MATLAB version 7 R14 SP2 (Chapman, 2002) software was used. MATLAB is an equation-solving software package that has proven to have a wide range of applicability to engineering problems. It can perform a variety of data analysis and presentation functions, including statistical analyses and graphical presentation of data. The analysis using MATLAB was done by writing programs as well as using MATLAB distribution functions.

3.1 STUDY AREA

Three sites have been chosen for this research which is Kuala Lumpur, Seberang Perai and Kuching. Figure 3.1 shows the location of these three sites in Malaysia. Kuala Lumpur is the capital of Malaysia and Seberang Perai is situated in Penang. Kuching is situated in East Malaysia and is the capital of Sarawak, on the island of Borneo.

Figure 3.1: Map of Malaysia and the 3 chosen sites
(Source: Encarta, 2006)

3.1.1 Kuala Lumpur

Figure 3.2 shows the location of Kuala Lumpur in Peninsular 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.

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. Ozone was the main contributory pollutant giving rise to unhealthy air quality in Shah Alam, Kajang, Gombak and Kuala Lumpur, especially between 2 pm to 5 pm in the daytime. About 70 percent of the unhealthy air quality condition was due to the presence of high particulate matter (PM_{10}) and the remaining 30 percent due to high ozone levels (Department of Environment, 2002).

Figure 3.2: The location of Kuala Lumpur
(Source: Encarta,2006)

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.

Referring to the Ministry of Transport (2006) website, the vehicle numbers for Kuala Lumpur is increasing with time. Table 3.1 shows the total numbers of vehicles in 1998 and 2000. The vehicle numbers for 2002 is estimated from 1998 and 2000 as given in the

table. According to Table 3.1 below, the percentage increase of vehicle numbers per year is about 2.5%.

Table 3.1: Vehicle numbers for Kuala Lumpur

Year	1998	2000	2002 (Estimated)
Vehicle size	759 690	801 493	843 296

(source: Ministry of Transport, 2006)

3.1.2 Seberang Perai

Penang is actually the one and only island state situated in the North-West coast of Peninsular Malaysia. The state capital of Penang is Georgetown which is situated in the island itself and there are a total of five main districts in Penang which is South-West District, Southern Seberang Perai, Central Seberang Perai, Northern Seberang Perai and North-East District. Figure 3.3 shows the location of Seberang Perai in Penang, Malaysia.

According to the Ministry of Housing and Local Government (2006), the population size of Seberang Perai in 1998 is about 251 553 people and in 2002 is about 268 180 people. From the population size in 1998 and 2002, the estimated population growth for Seberang Perai is almost 1.6% per year.

Figure 3.3: The location of Seberang Perai
(Source: Encarta, 2006)

Referring to the Ministry of Transport (2006) website, the vehicle numbers for Seberang Perai is increasing with time. Table 3.2 shows the total numbers of vehicles in 1998 and 2000. The vehicle numbers for 2002 is estimated from 1998 and 2000 as given in the table. According to Table 3.2, the percentage of vehicle numbers increasing per year is about 1.4%. The percentage increase of vehicle numbers per year is lower than Kuala Lumpur.

Table 3.2: Vehicle numbers for Seberang Perai

Year	1998	2000	2002 (Estimated)
Vehicle size	350 921	361 643	372 365

(Source: Ministry of Transport, 2006)

As Perai is a heavily industrialized area with several petrochemical complexes, the air quality remained at the moderate level more than 90 percent of the time. The main pollutant of concern is sulphur dioxide (SO_2) caused by industrial fuel combustion (Department of Environment, 2002).

3.1.3 Kuching

Kuching Division is one of the eleven administrative divisions in Sarawak, East Malaysia, on the island of Borneo. Formerly called "First Division", it is the center and the starting point of modern Sarawak. Kuching Division contains three administrative districts: Kuching, Bau and Lundu. Figure 3.4 shows the location of Kuching in Sarawak.

According to the Ministry of Housing and Local Government (2006), the population size of Sarawak in 1998 is about 429 667 people and in 2002 is about 458 300 people. From the population size in 1998 and 2002, the estimated value for population growth in Sarawak is almost 1.6% per year.

Figure 3.4: The location of Kuching
(Source: Encarta, 2006)

Table 3.3 shows the total numbers of vehicles in 1998 and 2000. The vehicle numbers for 2002 is estimated from 1998 and 2000 as given in the table. According to Table 3.3, the percentage decrease of vehicle numbers per year is about – 4.09 %. The percentage of vehicle numbers per year for Kuching is the lowest among the three sites.

Table 3.3: Vehicle numbers for Kuching

Year	1998	2000	2002 (Estimated)
Vehicle size	168 631	156 785	144 939

(source: Ministry of Transport, 2006)

3.2 DESCRIPTIVE ANALYSIS

In this sub section, numerical summaries of the observation are included for application of statistics, probability and reliability in this research. This is a complementary method through which much of the information contained in a monitoring record can be represented economically and conveyed or transmitted with greater precision. This method utilizes a set of characteristic numbers to summarize the observation and highlight their main features.

The most important purpose of these descriptive numerical summaries is for statistical inference, a role that graphs cannot fulfill. Among these, the most important statistics are the mean, median, mode, variance, standard deviation, skewness and kurtosis.

3.2.1 MEAN

The arithmetic mean of a set of n measurements, $x_1, x_2, ..., x_n$, is the average of the measurements. Typically, the symbol $\bar{x}$ is used to represent the sample mean for example, the mean of a sample of n measurements. Equation (3.1) shows the mean formula that is given by Mendenhall and Sincich (1995).

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{3.1}$$

where,

x_i = number of measurement per year for particular sites
n = total number of annual measurement for particular sites

3.2.2 MEDIAN

The median of a set of n measurements, $x_1, x_2, ..., x_n$, is the middle number when the measurements are arranged in ascending (or descending) order, for example, the value of x located so that half the area under the relative frequency histogram lies to its left and half the area lies to its right. Equation (3.2) shows the median formula that is given by Mendenhall and Sincich (1995).

$$m = \begin{cases} x_{i[(n+1)/2]} & \text{if } n \text{ is odd} \\ \dfrac{x_{i(n/2)} + x_{i(n/2+1)}}{2} & \text{if } n \text{ is even} \end{cases} \tag{3.2}$$

where,

x_i = number of measurement per year for particular sites
n = total number of annual measurement for particular sites

3.2.3 MODE

The mode of a set of n measurements, $x_1, x_2, ..., x_n$, is the value of x that occurs with the greatest frequency (Mendenhall and Sincich, 1995).

3.2.4 VARIANCE

The variance of a sample of n measurements, $x_1, x_2, ..., x_n$, is defined as shown in equation (3.3). Equation (3.3) shows the variance formula that is given by Mendenhall and Sincich (1995).

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1} \tag{3.3}$$

where,

x_i = number of measurement per year for particular sites
n = total number of annual measurement for particular sites
$\bar{x}$ = mean of one set of annual monitoring record

3.2.5 STANDARD DEVIATION

The standard deviation of a sample of n measurements is equal to the square root of the variance. Equation (3.4) shows the standard deviation formula that is given by Mendenhall and Sincich (1995).

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \tag{3.4}$$

where,

x_i = number of measurement per year for particular sites
n = total number of annual measurement for particular sites
$\bar{x}$ = mean of one set of annual monitoring record
s^2 = variance of one set of annual monitoring record

3.2.6 SKEWNESS

The skewness measures the asymmetry of a set of data about its mean. Equation (3.5) shows the skewness formula that is given by Kottegoda and Rosso (1998).

$$g_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{ns^3} \tag{3.5}$$

where,

x_i = number of measurement per year for particular sites

n = total number of annual measurement for particular sites
$\bar{x}$ = mean of one set of annual monitoring record
s = standard deviation of one set of annual monitoring record

Division by the cube of the sample standard deviation gives a dimensionless measure. A plot is said to have positive skewness if it has a longer tail on the right, which is toward increasing values, than on the left. In this case, the number of values less than the mean is greater than the number that exceeds the mean. A symmetrical plot suggests zero skewness (Kottegoda and Rosso, 1998). For standard normal distribution, the value for skewness is 0.

3.2.7 KURTOSIS

The extent of the relative steepness of ascent in the vicinity and on either side of the mode in a plot is said to be measure of its peakedness or tail weight. Equation (3.6) shows the kurtosis formula that is given by Kottegoda and Rosso (1998).

$$g_2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{ns^4} \tag{3.6}$$

where,

x_i = number of measurement per year for particular sites
n = total number of annual measurement for particular sites
$\bar{x}$ = mean of one set of annual monitoring record
s = standard deviation of one set of annual monitoring record

For standard normal distribution, the value for kurtosis is 3.

3.3 THE DISTRIBUTIONS

Eight distributions were used. They are Weibull, gamma, log-normal, Rayleigh, Pareto, log-logistic, Laplace and inverse Gaussian distributions which were used to fit the CO concentrations using MATLAB. MATLAB was used to obtain the parameters of the distributions.

Before statistical analyses for the data set were done, the statistical characteristics for every data set were obtained. Summaries of each data such as the mean, median, variance, standard deviation, minimum and maximum concentration, kurtosis and skewness were examined.

3.3.1 THE WEIBULL DISTRIBUTION

Weibull distribution has been used to model hourly CO emission in three sites which are Kuala Lumpur, Kuching and Seberang Perai. The Weibull probability distribution function was selected because it is the most common and simple function when dealing with extreme events (Seinfeld and Pandis, 1998).

This distribution is versatile enough to be a good model for many distributions of data that are mound-shaped but skewed. Another advantage of the Weibull distribution is that the numbers of relatively straightforward techniques exist for actually fitting the model to data (Scheaffer, 1995).

3.3.1.1 THE DENSITY AND DISTRIBUTION FUNCTIONS

Equation (3.7) and (3.8) are given in Evans *et al.*, (2000). The probability density function (pdf) is defined as:

$$f(x)=\left(\frac{\lambda}{\sigma}\right)\left(\frac{x}{\sigma}\right)^{\lambda-1}\exp\left\{-\left(\frac{x}{\sigma}\right)^{\lambda}\right\} \tag{3.7}$$

and the cumulative distribution function (cdf) is defined as:

$$F(x)=1-\exp\left\{-\left(\frac{x}{\sigma}\right)^{\lambda}\right\} \tag{3.8}$$

where $x \geq 0$, σ represents a scale parameter and λ represents a shape parameter for annual measurement at particular sites.

3.3.1.2 PARAMETER ESTIMATIONS

The MLE for the Weibull distribution are given as follows:

$$\left(\frac{1}{\lambda}\right)-\left(\frac{\sum_{i=1}^{n} x_i \ln(x_i)}{\sum_{i=1}^{n} x_i^{\lambda}}\right)+\left(\frac{1}{n}\right)\sum_{i=1}^{n}\ln(x_i)=0 \tag{3.9}$$

$$\sigma=\left(\frac{1}{n}\sum_{i-1}^{n} x_i^{\lambda}\right)^{\frac{1}{\lambda}} \tag{3.10}$$

Equations (3.9) and (3.10) show the MLE formulae. The values of λ can be obtained from equation (3.9) by numerical method.

3.3.2 THE LOG-NORMAL DISTRIBUTION

The log-normal distribution is more widely used to represent the distribution of air pollutant concentration (Berger *et al.*, 1982).

3.3.2.1 THE DENSITY FUNCTION

Equation (3.11) shows the pdf (Evans *et al.*, 2000).

$$f(x)=\left(\frac{1}{x\lambda\sqrt{2\pi}}\right)\exp\left\{-\frac{1}{2}\left(\frac{\ln(x)-\sigma}{\lambda}\right)^2\right\} \tag{3.11}$$

where $x \geq 0$, λ represents a shape parameter and σ represents a scale parameter for annual measurement at particular sites.

3.3.2.2 PARAMETER ESTIMATIONS

For determining the maximum likelihood estimators of the log-normal distribution with parameters λ and σ, the same procedure can be used as for the normal distribution. Equations (3.12) and (3.13) show the MLE formula (Evans *et al.*, 2000).

$$\sigma=\left(\frac{1}{n}\right)\sum_{i=1}^{n}\ln(x_i) \tag{3.12}$$

$$\lambda=\left(\frac{1}{n-1}\right)\sum_{i=1}^{n}\left(\ln(x_i)-\sigma\right)^2 \tag{3.13}$$

3.3.3 THE GAMMA DISTRIBUTION

The gamma distribution is the best distribution to represent the performance of high pollutants concentration in Taiwan (Lu, 2003). It is an important distribution for non - Gaussian statistical modeling. The distribution is positively skewed and it is often used as a model of life spans testing and many other related fields (Pang *et al.*, 2002).

3.3.3.1 THE DENSITY FUNCTION

The probability density function of the gamma distribution can be expressed in terms of the gamma function as shows in equation (3.17). Equation (3.17) is given by Evans *et al.* (2000). The cumulative distribution function (CDF) is obtained from integrating the PDF.

$$f(x)=\left(\frac{1}{\sigma\Gamma(\lambda)}\right)\left(\frac{x}{\sigma}\right)^{\lambda-1}\exp\left(-\frac{x}{\sigma}\right) \tag{3.14}$$

where

$\Gamma(\lambda)$ = gamma function with argument λ

$$= \int_0^{\infty} e^{-t}t^{\lambda-1}dt$$

And $x \geq 0$, λ represents a shape parameter and σ represents a scale parameter for annual measurement of particular sites. For integer n, gamma $(n+1) = n!$

3.3.3.2 PARAMETER ESTIMATIONS

The MLE for the gamma distribution is given in equation (3.18) and (3.19). Equation (3.18) and (3.19) are given by Evans *et al.* (2000).

$$\ln(\lambda)-\psi(\lambda)=\ln\left(\frac{\bar{x}}{g}\right) \tag{3.15}$$

$$\sigma\lambda=\bar{x} \tag{3.16}$$

where,

g = geometric sample mean

$$= \prod_{i=1}^{n} x_i^{1/n}$$

and,

$\psi(\lambda)$ is the digamma function.

3.3.4 THE RAYLEIGH DISTRIBUTION

The Weibull distribution is a special case of the generalized gamma distribution, while the Rayleigh distribution is a subset of the Weibull distribution. The Weibull distribution is a two parameter distribution, while the Rayleigh distribution has only one parameter (Johnson, 2001).

3.3.4.1 THE DENSITY AND DISTRIBUTION FUNCTIONS

The equation for the probability density function and the cumulative distribution function for Rayleigh distribution are given by Evans *et al.* (2000) as shown in equation (3.17) and (3.18).

$$f(x)=\left(\frac{x}{\sigma^2}\right)\exp\left[-\left(\frac{x^2}{2\sigma^2}\right)\right] \tag{3.17}$$

$$F(x)=1-\exp\left[-\left(\frac{x^2}{2\sigma^2}\right)\right] \tag{3.18}$$

where $x>0$, $\sigma>0$ and σ represents a scale parameter for annual measurement at particular sites.

3.3.4.2 PARAMETER ESTIMATIONS

The only equation for maximum likelihood estimation of the Rayleigh distribution is given in equation (3.19). This equation is given by Evans *et al.* (2000).

$$\sigma=\left(\left(\frac{1}{2n}\right)\sum_{i=1}^{n}x_i^2\right)^{1/2} \tag{3.19}$$

3.3.5 THE LOG-LOGISTIC DISTRIBUTION

Referring to Gajjar and Khatri (1969), the log-logistic distribution seemed to be the most appropriate distribution model in the development of a hybrid model. The log-logistic distribution have two parameters which are μ, the location parameter and σ, the scale parameter.

3.3.5.1 THE DENSITY AND DISTRIBUTION FUNCTIONS

The probability density function and the cumulative distribution function for log-logistic distribution are given by Evans *et al.,* (2000) and is shown in equation (3.20) and (3.21) respectively.

$$f(x)=\frac{e^{-(\ln x-\mu/\sigma)}}{\sigma\{1+\exp[-(\ln x-\mu)/\sigma]\}^2} \tag{3.20}$$

$$F(x) = \{1 + \exp[-(\ln x - \mu)/\sigma]\}^{-1} \tag{3.21}$$

where $-\infty < x < \infty$, $\sigma > 0$ where σ represents a scale parameter and μ represents a location parameter for annual measurement at particular sites.

3.3.5.2 PARAMETER ESTIMATIONS

The maximum likelihood (MLE) estimation of the parameters is calculated as follows:

$$\sum_{i=1}^{n}\left[1 + \exp\left(\frac{\ln x_i - \mu}{\sigma}\right)\right]^{-1} = \frac{n}{2} \tag{3.22}$$

$$\sum_{i=1}^{n}\left(\frac{\ln x_i - \mu}{\hat{\sigma}}\right)\frac{1 - \exp[\ln x_i - \mu/\sigma]}{1 + \exp[(\ln x_i - \mu/\sigma)]} = 2 \tag{3.23}$$

Equations (3.22) and (3.23) are given by Evans *et al.* (2000).

3.3.6 THE PARETO DISTRIBUTION

The Pareto distribution is used to model the distribution of personal income, the distribution of population sizes and stock price fluctuations, where the parameter a's are used for minimum income, minimum population size and minimum stock price, respectively (Singh, 2004).

3.3.6.1 THE DENSITY AND DISTRIBUTION FUNCTIONS

The equation for the probability density function and the cumulative distribution function are given by Evans *et al.* (2000) as shown in equation (3.24) and (3.25). The equation for the pdf is:

$$f(x) = \frac{\lambda \mu^{\lambda}}{x^{\lambda+1}} \tag{3.24}$$

and the equation for the cumulative distribution function is:

$$F(x) = 1 - \left(\frac{\mu}{x}\right)^{\lambda} \tag{3.25}$$

where, $x \geq \mu$, $\mu > 0$, $\lambda > 0$, μ represents a location parameter and λ represents a shape parameter for annual measurement at particular sites.

3.3.6.2 PARAMETER ESTIMATIONS

The maximum likelihood for the parameter estimations of Pareto distribution are:

$$\frac{1}{\lambda} = \left(\frac{1}{n}\right)\sum_{i=1}^{n} \ln\left(\frac{x_i}{\mu}\right) \tag{3.26}$$

$$\mu = \text{minimum}(x_i) \tag{3.27}$$

Equations (3.26) and (3.37) are given by Evans *et al.* (2000).

3.3.7 THE LAPLACE DISTRIBUTION

It also called the double exponential distribution. It is the distribution of differences between two independent variates with identical exponential distributions (Abramowitz and Stegun, 1972).

3.3.7.1 THE DENSITY AND DISTRIBUTION FUNCTIONS

The equation for the probability density function and the cumulative distribution function for Laplace distribution are given by Evans *et al.* (2000) and are shown in equation (3.28) and (3.29) respectively.

$$f(x) = \frac{1}{2\sigma}\exp\left[\left(\frac{|x-\mu|}{\sigma}\right)\right] \tag{3.28}$$

$$F(x) = \begin{cases} \frac{1}{2}\exp\left[-\left(\frac{\mu - x}{\sigma}\right)\right] & \text{if}(x < \mu) \\ 1 - \frac{1}{2}\exp\left[-\left(\frac{x-\mu}{\sigma}\right)\right] & \text{if}(x \geq \mu) \end{cases} \tag{3.29}$$

where, $-\infty < x < \infty$, $-\infty < \mu < \infty$, $\sigma > 0$, μ represents a location parameter and σ represents a scale parameter for annual measurement at particular sites.

3.3.7.2 PARAMETER ESTIMATIONS

The MLE for the parameters of the Laplace distribution are given by Evans *et al.* (2000) and are given in equation (3.30) and (3.31).

$$\mu = \text{median} \tag{3.30}$$

$$\sigma = \left(\frac{1}{n}\right)\sum_{i=1}^{n}|x_i - \mu| \tag{3.31}$$

3.3.8 THE INVERSE GAUSSIAN DISTRIBUTION

In reliability and life testing, the inverse Gaussian distribution is particularly useful in situations where early failures dominate. This is due to the non-monotonic behaviour of its hazard function (Chhikara and Folks, 1989).

3.3.8.1 THE DENSITY FUNCTION

The equation for the probability density function is given by Evans *et al.* (2000) as shown in equation (3.32).

$$f(x) = \left(\frac{\sigma}{2\pi x^3}\right)^{1/2} \exp\left(\frac{-\sigma(x-\mu)^2}{2\mu^2 x}\right) \tag{3.32}$$

where $x > 0$, $\mu > 0$, $\sigma > 0$, μ represents a location parameter and σ represents a scale parameter for annual measurement at particular sites.

3.3.8.1 PARAMETER ESTIMATIONS

The parameter estimations for the inverse Gaussian distribution are given in equation (3.36) and (3.37). The equation (3.36) and (3.37) are given by Evans *et al.* (2002).

$$\mu = \bar{x} \tag{3.33}$$

$$\sigma = \frac{(n-1)}{\sum_{i=1}^{n}\left(\frac{1}{x_i} - \frac{1}{\bar{x}}\right)} \tag{3.34}$$

3.4 GENERALIZED LAMBDA DISTRIBUTION

The generalized lambda distribution (GLD) with parameters $\lambda_1, \lambda_2, \lambda_3, \lambda_4$, GLD($\lambda_1, \lambda_2, \lambda_3, \lambda_4$) is defined by

$$Q(y) = Q(y, \lambda_1, \lambda_2, \lambda_3, \lambda_4) = \lambda_1 + \frac{y^{\lambda_3} - (1-y)^{\lambda_4}}{\lambda_2} \tag{3.35}$$

with $0 \leq y \leq 1$ and $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ are the location, scale, skewness and kurtosis parameters respectively.

The parameters $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ can be estimated using the method of moments and the method of percentiles. For this research, both these methods will be used.

The first four moments for the GLD are given by

$$\alpha_1 = \mu = E(X) = \lambda_1 + \frac{A}{\lambda_2} \tag{3.36}$$

$$\alpha_2 = \sigma^2 = E\left[(X - E(X))^2\right] = \frac{B - A^2}{\lambda_2^2} \tag{3.37}$$

$$\alpha_3 = E\left[(X - E(X))^3\right] \Big/ \sigma^3 = \frac{C - 3AB + 2A^3}{\lambda_2^3 \sigma^3} \tag{3.38}$$

$$\alpha_4 = E\left[(X - E(X))^4\right] \Big/ \sigma^4 = \frac{D - 4AC + 6A^2B - 3A^4}{\lambda_2^4 \sigma^4} \tag{3.39}$$

where

$$A = \frac{1}{1+\lambda_3} - \frac{1}{1+\lambda_4} \tag{3.40}$$

$$B = \frac{1}{1+2\lambda_3} + \frac{1}{1+2\lambda_4} - 2\beta(1+\lambda_3, 1+\lambda_4) \tag{3.41}$$

$$C = \frac{1}{1+3\lambda_3} - \frac{1}{1+3\lambda_4} - 3\beta(1+\lambda_3, 1+\lambda_4) + 3\beta(1+\lambda_3, 1+2\lambda_4) \tag{3.42}$$

$$D = \frac{1}{1+4\lambda_3} + \frac{1}{1+4\lambda_4} - 4\beta(1+3\lambda_3, 1+\lambda_4) + 6\beta(1+2\lambda_3, 1+2\lambda_4) - 4\beta(1+\lambda_3, 1+3\lambda_4) \tag{3.43}$$

$\alpha_1, \alpha_2, \alpha_3, \alpha_4$ can be estimated using equations (3.36) until (3.39). $\beta(a,b)$ is the beta function.

If $X_1, X_2, \ldots, X_n$ are the observed data, then the sample moments corresponding to $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ and denoted by $\hat{\alpha}_1, \hat{\alpha}_2, \hat{\alpha}_3, \hat{\alpha}_4$ are given by

$$\hat{\alpha}_1 = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{3.44}$$

$$\hat{\alpha}_2 = \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X})^2 \tag{3.45}$$

$$\hat{\alpha}_3 = \frac{1}{n\hat{\sigma}^3}\sum_{i=1}^{n} (X_i - \bar{X})^3 \tag{3.46}$$

$$\hat{\alpha}_4 = \frac{1}{n\hat{\sigma}^4}\sum_{i=1}^{n} (X_i - \bar{X})^4 \tag{3.47}$$

The percentiles for the GLD are given by

$$\rho_1 = Q\left(\frac{1}{2}\right) = \lambda_1 + \frac{\left(\frac{1}{2}\right)^{\lambda_3} - \left(\frac{1}{2}\right)^{\lambda_4}}{\lambda_2} \tag{3.48}$$

$$\rho_2 = Q(1-u) - Q(u) = \frac{(1-u)^{\lambda_3} - u^{\lambda_4} + (1-u)^{\lambda_4} - u^{\lambda_3}}{\lambda_2} \tag{3.49}$$

$$\rho_3 = \frac{Q\left(\frac{1}{2}\right) - Q(u)}{Q(1-u) - Q\left(\frac{1}{2}\right)} = \frac{(1-u)^{\lambda_4} - u^{\lambda_3} + \left(\frac{1}{2}\right)^{\lambda_3} - \left(\frac{1}{2}\right)^{\lambda_4}}{(1-u)^{\lambda_3} - u^{\lambda_4} + \left(\frac{1}{2}\right)^{\lambda_4} - \left(\frac{1}{2}\right)^{\lambda_3}} \tag{3.50}$$

$$\rho_4 = \frac{Q\left(\frac{3}{4}\right) - Q\left(\frac{1}{4}\right)}{\rho_2} = \frac{\left(\frac{3}{4}\right)^{\lambda_3} - \left(\frac{1}{4}\right)^{\lambda_4} + \left(\frac{3}{4}\right)^{\lambda_4} - \left(\frac{1}{4}\right)^{\lambda_3}}{(1-u)^{\lambda_3} - u^{\lambda_4} + (1-u)^{\lambda_4} - u^{\lambda_3}} \tag{3.51}$$

If $\hat{\pi}_p$ is the $(100)^{th}$ data percentile, then the sample statistics that will be used to estimate $\rho_1, \rho_2, \rho_3, \rho_4$ and denoted by $\hat{\rho}_1, \hat{\rho}_2, \hat{\rho}_3, \hat{\rho}_4$ are given by

$$\hat{\rho}_1 = \hat{\pi}_{0.5} \tag{3.52}$$

$$\hat{\rho}_2 = \hat{\pi}_{1-u} - \hat{\pi}_u \tag{3.53}$$

$$\hat{\rho}_3 = \frac{\hat{\pi}_{0.5} - \hat{\pi}_u}{\hat{\pi}_{1-u} - \hat{\pi}_{0.5}} \tag{3.54}$$

$$\hat{\rho}_4 = \frac{\hat{\pi}_{0.75} - \hat{\pi}_{0.25}}{\hat{\rho}_2} \tag{3.55}$$

u is a value between 0 and $1/4$.

The probability density function for the GLD($\lambda_1, \lambda_2, \lambda_3, \lambda_4$) is given by

$$f(x) = \frac{\lambda_2}{\lambda_3 y^{\lambda_3 - 1} + \lambda_4 (1-y)^{\lambda_4 - 1}} \quad \text{at } x = Q(y) \tag{3.56}$$

3.5 THE EXTREME VALUE DISTRIBUTIONS

According to the theory of extreme values, the largest or smallest value from a set of independent and identically distributed random variables tends to an asymptotic distribution that only depends on the tail of the distribution of the basic variables (Kottegoda and Rosso, 1998). There are two main extreme value distributions which will be considered here that are as follows:

i If the tail of the $f(y, \theta)$ is unbounded and decreases at least as rapidly as an exponential form, the asymptotic extreme value distribution is termed type I of maxima. This is the Gumbel distribution.

ii If the initial distribution of $f(y, \theta)$ features an unbounded upper tail, but not all its moment is finite, then the asymptotic extreme value distribution is term type II of maxima. This is Frechet distribution.

3.5.1 Threshold Values

In order to determine the threshold values, the following equation was applied with different frequency factors (Madsen and Rosbjerg, 1997).

$$ffc = E(Q) + cS(Q) \tag{3.57}$$

where;

$E(Q)$ is the mean,

c is the frequency factor (2, 3, 4 and 5), and

$S(Q)$ is thestandard deviation

By using SPSS, concentration more than *ffc* from each data set were selected as new set of extreme data.

3.5.2 The Gumbel Distribution

The Gumbel distribution was extensively developed and applied to flood flows by Gumbel in 1954 and 1958. This distribution results from any underlying distribution of the X_i's of the exponential type.

The probability density function for the Gumbel distribution is as follows (Kottegoda and Rosso, 1998);

$$f(x;\sigma,\mu)=\frac{1}{\sigma}\exp\left[-\frac{x-\mu}{\sigma}-\exp\left(-\frac{x-\mu}{\sigma}\right)\right],0<x<\infty,-\infty<\mu<\infty,\sigma>0 \tag{3.58}$$

The cumulative distribution function for the Gumbel distribution is as follows;

$$F(x,\mu,\sigma)=\exp\left[-\exp\left(-\frac{x-\mu}{\sigma}\right)\right],0<x<\infty,-\infty<\mu<\infty,\sigma>0 \tag{3.59}$$

Parameters σ and μ (scale and location) in this distribution can be estimated by using method of moments and maximum likelihood estimators.

3.5.2.1 The Method of Maximum Likelihood Estimators

The likelihood function of θ, where θ represents the set of unknown parameters is defined as;

$$L(\theta)=\prod_{i=1}^{n} fx(x_i\ |\theta) \tag{3.60}$$

Substituting Equation 3.58 for *f(x)* in Equation 3.60 the log-likelihood function becomes

$$\ln L = -\sum_{i=1}^{n} \frac{x_i - \mu}{\sigma} - \sum_{i=1}^{n} \exp\left(-\frac{x_i - \mu}{\sigma}\right) - n \ln \sigma \tag{3.61}$$

The partial derivatives of $\ln L$ are;

$$\frac{\partial \ln L}{\partial \sigma} = \sum \frac{x_i - \mu}{\sigma^2} + \sum \frac{x_i - \mu}{\sigma^2} \exp\left(-\frac{x_i - \mu}{\sigma}\right) - \frac{n}{\sigma}, \tag{3.62}$$

and;

$$\frac{\partial \ln L}{\partial \mu} = \frac{n}{\sigma} - \sum \frac{1}{\sigma} \exp\left(-\frac{x_i - \mu}{\sigma}\right) \tag{3.63}$$

By setting $\frac{\partial \ln L}{\partial \sigma} = 0$ and $\frac{\partial \ln L}{\partial \mu} = 0$, parameters σ and μ can be obtained. From Equation (3.63),

$$\exp\left(\frac{\mu}{\sigma}\right) = \frac{n}{\sum \exp(-x_i / \sigma)} \tag{3.64}$$

This is used in Equation (3.62) to obtain σ and μ, after simplifying;

$$\sigma = \bar{x} - \frac{\sum x_i \exp(-x_i / \sigma)}{\sum \exp(-x_i / \sigma)} \tag{3.65}$$

and;

$$\mu = -\sigma \ln\left(\frac{1}{n} \sum_{i=1}^{n} \exp\left(-\frac{x_i}{\sigma}\right)\right) \tag{3.66}$$

3.5.3 The Frechet Distribution

The Frechet distribution was first developed and applied to flood flows by Frechet (1927). The probability density function of the Frechet distribution is as follows (Kottegoda and Rosso, 1998);

$$f(x,\sigma,\lambda)=\frac{\lambda}{\sigma}\left(\frac{\sigma}{x}\right)^{\lambda+1}\exp\left[-\left(\frac{\sigma}{x}\right)^{\lambda}\right], x\geq 0,\sigma,\lambda\geq 0 \tag{3.67}$$

The CDF for the Frechet distribution is as shown below;

$$F(x,\sigma,\lambda)=\exp\left[-\left(\frac{\sigma}{x}\right)^{\lambda}\right], x\geq 0,\sigma,\lambda\geq 0 \tag{3.68}$$

Just like the Gumbel distribution, the scale and shape parameter (σ and λ) in the Frechet distribution can also be estimated by using the method of moments and maximum likelihood estimators.

3.5.3.1 The Method of Maximum Likelihood Estimators

In this method, λ is the solution of the following equation (Kottegoda and Rosso, 1998);

$$\frac{1}{\lambda}+\frac{\sum_{i=1}^{n} x_i^{-\lambda}\ln(x_i)}{\sum_{i=1}^{n} x_i^{-\lambda}}=\frac{1}{n}\sum_{i=1}^{n}\ln(x_i) \tag{3.69}$$

Thus, σ can be estimated by the following equations;

$$\sigma=\left(\frac{1}{n}\sum_{i=1}^{n} x_i^{-\lambda}\right)^{-\frac{1}{\lambda}} \tag{3.70}$$

3.6 PERFORMANCE INDICATORS

Four performance indicators are used in this research which is root mean square error (RMSE), index of agreement (IA), prediction accuracy (PA) and coefficient of determination (R^2). The equations (3.71), (3.72), (3.73) and (3.74) are given by Lu (2003).

The root mean square error (RMSE) is given by:

$$RMSE = \sqrt{\left(\frac{1}{N-1}\right)\sum_{i=1}^{N}(P_i - O_i)^2} \tag{3.71}$$

For a good model, the RMSE should approach zero. Therefore, a smaller RMSE means the model is more appropriate.

The index of agreement (IA) is given by:

$$IA = 1 - \left[\frac{\sum_{i=1}^{N}(P - O_i)^2}{\sum_{i=1}^{N}\left(\left|P_i - \overline{O}\right| + \left|O_i - \overline{O}\right|\right)^2}\right] \tag{3.72}$$

where $0 \le IA \le 1$. When IA is closer to 1, then the model is appropriate to simulate the experimental data.

The prediction accuracy (PA) is given by:

$$PA = \frac{\sum_{i=1}^{N}(P_i - \overline{O})^2}{\sum_{i=1}^{N}(O_i - \overline{O})^2} \tag{3.73}$$

where $0 \le PA \le 1$. When PA is closer to 1, then the model is appropriate to simulate the experimental data.

The coefficient of determination (R^2) is given by:

$$R^2 = \left(\frac{\sum_{i=1}^{N}(P_i - \overline{P})(O_i - \overline{O})}{N.S_{pred}.S_{obs}}\right)^2 \tag{3.74}$$

where;

N = Total number of annual measurements of a particular site
P_i = Predicted values of one set annual monitoring record
O_i = Observed values of one set annual monitoring record
$\overline{P}$ = Mean of the predicted values of one set annual monitoring record
$\overline{O}$ = Mean of the observed values of one set annual monitoring record
S_{pred} = Standard deviation of the predicted values of one set annual monitoring record

S_{obs} = Standard deviation of the observed values of one set annual monitoring record

When R^2 is closer to 1, then the model is appropriate to simulate the experimental data.

CHAPTER 4

RESULTS AND DISCUSSIONS

Chapter 4 discussed the characteristic of the observations as well as obtaining the parameter estimates of the nine distributions. Performance indicators of the nine distributions were also obtained to determine the best distribution.

4.1 DATA DESCRIPTION

Tables 4.1 give summaries of CO concentration in 1998 and 2002 for Kuala Lumpur, Kuching and Seberang Perai.

Table 4.1: Descriptive statistics for CO (1998 and 2002)

	Sites					
	Kuala Lumpur		Kuching		Seberang Perai	
	1998	2002	1998	2002	1998	2002
Total, N	8760	8200	8185	6244	7949	8231
Min value	0.00	0.01	0.00	0.01	0.00	0.01
Max value	14.03	3.13	3.14	2.86	3.74	3.13
Mean	2.27	0.61	0.60	0.60	0.73	0.61
Variance	3.45	0.15	0.13	0.14	0.19	0.15
Standard deviation	1.86	0.38	0.36	0.37	0.44	0.38
Median	1.87	0.52	0.51	0.51	0.64	0.52
Skewness	1.49	1.73	1.69	1.57	1.47	1.73
Kurtosis	3.16	7.64	4.72	6.56	3.62	7.62

Table 4.1 shows that the minimum value for CO concentrations in Kuala Lumpur, Kuching and Seberang Perai are the same which is 0.00 ppm in 1998 but increased to 0.01ppm in 2002.. The highest value for the maximum concentrations is represented at two sites which are Kuala Lumpur and Seberang Perai with the value of 3.13 ppm in 1998 but the maximum value is 14.03ppm in 2002 which occurs at Kuala Lumpur. All sites have positive values for skewness showing that the distribution of CO concentrations are skewed to the right. Kuala Lumpur has the highest mean of CO concentration for both years followed by Seberang Perai and then Kuching.

Table 4.1 and 4.2 give the characteristics of the extreme values using frequency factors 2 and 3 respectively. The mean values of CO concentrations for both years are the highest for Kuala Lumpur.

Table 4.2: Descriptive statistics for CO(1998) for *ff2* and maximum daily data

	Sites					
	Kuala Lumpur		Kuching		Seberang Perai	
	ff2	*max*	*ff2*	*max*	*ff2*	*max*
Total, N	381	365	381	363	303	358
Minimum	6.09	0.82	1.60	0.32	1.34	0.44
Maximum	14.03	14.03	3.14	3.14	3.74	3.74
Mean	7.72	6.54	1.99	1.24	1.69	1.64
Variance	2.38	5.99	0.14	0.28	0.09	0.31
Standard deviation	1.54	2.45	0.37	0.53	0.30	0.56
Median	7.21	6.28	1.87	1.13	1.61	1.55
Skewness	1.41	0.38	1.67	1.06	1.22	0.78
Kurtosis	1.95	0.04	3.32	1.32	1.18	0.61

Table 4.3: Descriptive statistics for CO(2002) for *ff2* and maximum daily data

	Sites					
	Kuala Lumpur		Kuching		Seberang Perai	
	ff2	*max*	*ff2*	*max*	*ff2*	*max*
Total, N	385	363	303	292	382	364
Minimum	5.45	2.30	1.34	0.30	1.39	0.24
Maximum	11.22	2.86	2.86	2.86	3.13	3.13
Mean	6.56	5.86	1.69	1.19	1.78	1.50
Variance	1.10	2.61	0.09	0.24	0.12	0.28
Standard deviation	1.05	1.62	0.30	0.49	0.35	0.53
Median	6.25	5.68	1.61	1.09	1.68	1.39
Skewness	1.45	0.44	1.22	0.92	1.50	0.73
Kurtosis	2.11	-0.008	1.18	0.58	2.40	0.34

4.2 PARAMETER ESTIMATES USING PROBABILITY DISTRIBUTIONS

Tables 4.4 and 4.5 show the parameter estimates of the eight distributions for the three sites. All the estimates have been obtained using maximum likelihood estimators or method of moments as discussed in Chapter 3.

Table 4.4: Parameter estimates for the three sites (1998)

Distributions	Seberang Perai	Kuala Lumpur	Kuching
Weibull	$\sigma = 0.82$ $\lambda = 1.76$	$\sigma = 2.51$ $\lambda = 1.26$	$\sigma = 0.68$ $\lambda = 1.80$
Log-normal	$\mu = -0.51$ $\sigma = 0.69$	$\mu = 0.46$ $\sigma = 1.02$	$\mu = -0.67$ $\sigma = 0.62$
Gamma	$\sigma = 2.76$ $\lambda = 0.26$	$\sigma = 1.44$ $\lambda = 1.63$	$\sigma = 3.12$ $\lambda = 0.19$
Rayleigh	$\sigma = 0.60$	$\sigma = 2.12$	$\sigma = 0.50$
Log-log logistic	$\mu = -0.46$ $\sigma = 0.36$	$\mu = 0.56$ $\sigma = 0.55$	$\mu = -0.65$ $\sigma = 0.33$
Pareto	$\mu = -0.23$ $\sigma = 0.86$	$\mu = -0.17$ $\sigma = 2.72$	$\mu = -0.22$ $\sigma = 0.71$
Laplace	$\mu = 0.64$ $\sigma = 0.31$	$\mu = 1.92$ $\sigma = 1.39$	$\mu = 0.51$ $\sigma = 0.26$
Inverse Gaussian	$\mu = 0.73$ $\sigma = 0.97$	$\mu = 2.34$ $\sigma = 0.99$	$\mu = 0.60$ $\sigma = 1.01$

Table 4.5: Parameter estimates for the three sites (2002)

Distributions	Seberang Perai	Kuala Lumpur	Kuching
Weibull	$\sigma = 0.69$ $\lambda = 1.72$	$\sigma = 0.69$ $\lambda = 1.72$	$\sigma = 0.67$ $\lambda = 1.74$
Log-normal	$\mu = -0.67$ $\sigma = 0.64$	$\mu = -0.67$ $\sigma = 0.64$	$\mu = -0.70$ $\sigma = 0.65$
Gamma	$\sigma = 2.85$ $\lambda = 0.22$	$\sigma = 2.85$ $\lambda = 0.22$	$\sigma = 2.84$ $\lambda = 0.21$
Rayleigh	$\sigma = 0.51$	$\sigma = 0.51$	$\sigma = 0.50$
Log-log logistic	$\mu = -0.65$ $\sigma = 0.35$	$\mu = -0.65$ $\sigma = 0.35$	$\mu = -0.67$ $\sigma = 0.35$
Pareto	$\mu = -0.23$ $\sigma = 0.73$	$\mu = -0.23$ $\sigma = 0.73$	$\mu = -0.25$ $\sigma = 0.72$
Laplace	$\mu = 0.52$ $\sigma = 0.27$	$\mu = 0.52$ $\sigma = 0.27$	$\mu = 0.51$ $\sigma = 0.26$
Inverse Gaussian	$\mu = 0.61$ $\sigma = 1.06$	$\mu = 0.61$ $\sigma = 1.06$	$\mu = 0.60$ $\sigma = 0.96$

From Table 4.5, it can be seen that the parameters for Seberang Perai are the same as the parameter values for Kuala Lumpur for the 2002 data.

4.3 RESULTS FOR SEBERANG PERAI DATA

The best distribution that fits the observed distribution can be chosen by looking at the performance indicators. Table 4.6 shows the values of the performance indicators for the CO concentration in Seberang Perai for 1998 and 2002.

Table 4.6: The performance indicators value for CO concentration in Seberang Perai

DISTRIBUTION	PERFORMANCE INDICATOR							
	RMSE		IA		PA		R^2	
	1998	2002	1998	2002	1998	2002	1998	2002
Weibull	0.066	0.072	0.994	0.991	0.988	0.982	0.977	0.965
Gamma	0.034	0.048	0.998	0.996	0.997	0.993	0.994	0.986
Log-normal	0.177	0.078	0.970	0.991	0.985	0.994	0.970	0.989
Laplace	0.167	0.170	0.963	0.950	0.949	0.935	0.900	0.874
Pareto	0.166	0.133	0.973	0.977	0.988	0.986	0.976	0.971
Rayleigh	0.092	0.101	0.988	0.980	0.982	0.973	0.965	0.946
Log-logistic	0.310	0.211	0.917	0.945	0.925	0.940	0.856	0.883
Inverse Gaussian	0.217	0.092	0.958	0.988	0.982	0.995	0.964	0.990

Table 4.6 shows that the smallest value for RMSE is given by the gamma distribution for both years. The highest value for IA is also given by the gamma distribution. The inverse Gaussian distribution gives the highest values for PA and R^2 which are 0.995 and 0.990 respectively for 2002. Based on the results, it can be concluded that the gamma distribution is the best distribution that can fit the CO concentration in Seberang Perai for 1998 and the inverse Gaussian distribution for 2002.

4.4 RESULTS FOR KUALA LUMPUR DATA

Table 4.7 shows the value of the four performance indicators for the CO concentration in Kuala Lumpur. From Table 4.7, the smallest value for RMSE is given by the Weibull distribution for 1998 and the gamma distribution for 2002. The gamma and Weibull distributions give the highest value for Index of Agreement or IA in 1998 but the gamma is the best for 2002 in terms of IA. The highest value for PA is given by two distributions which are the gamma and Weibull distributions for 1998. The highest value for R^2 is given by the inverse Gaussian distribution which is 0.990. Based on the results, it can be concluded that the best distribution that fits the observed distribution in Kuala Lumpur is the Weibull distribution for 1998 and the inverse Gaussian distribution for 2002.

Table 4.7: The performance indicators value for CO concentration in Kuala Lumpur

DISTRIBUTION	PERFORMANCE INDICATOR							
	RMSE		IA		PA		R^2	
	1998	2002	1998	2002	1998	2002	1998	2002
Weibull	0.066	0.072	0.999	0.991	0.999	0.982	0.999	0.965
Gamma	0.102	0.049	0.999	0.996	0.999	0.993	0.998	0.986
Log-normal	1.970	0.078	0.862	0.991	0.917	0.994	0.841	0.989
Laplace	0.850	2.05	0.949	0.415	0.928	0.935	0.861	0.874
Pareto	0.201	0.134	0.997	0.977	0.997	0.986	0.993	0.971
Rayleigh	0.665	0.101	0.959	0.980	0.981	0.973	0.961	0.946
Log-logistic	5.040	0.209	0.562	0.945	0.707	0.940	0.500	0.884
Inverse Gaussian	1.988	0.092	0.861	0.988	0.915	0.995	0.838	0.990

4.5 RESULTS FOR KUCHING DATA

Figure 4.1 shows the cumulative distribution plots for CO concentration in Kuching where seven distributions were plotted and compared with the observed distribution.

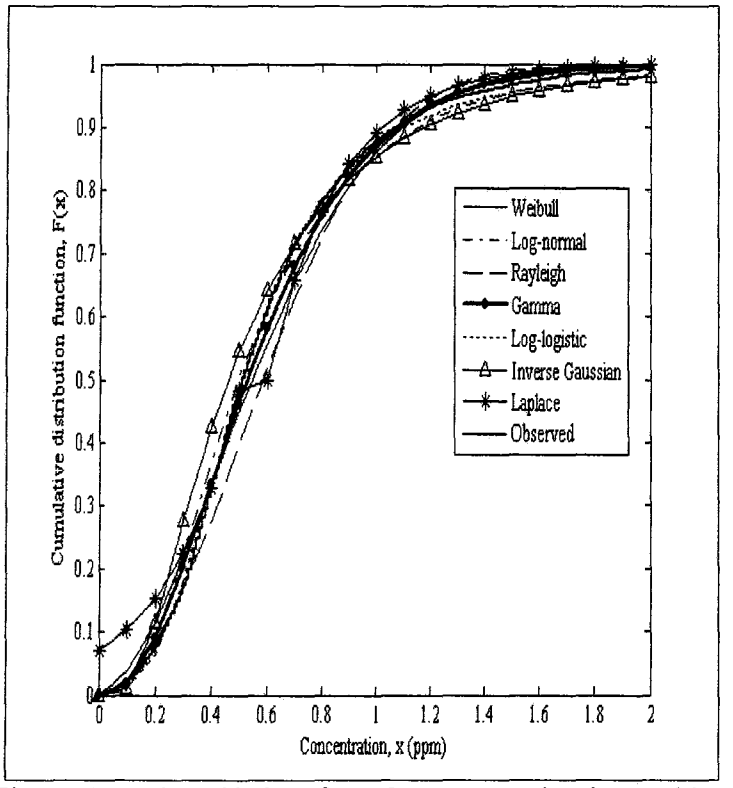


Figure 4.1: The cdf plots for CO concentration in Kuching

Referring to Figure 4.1, it shows that the Gamma distribution fits the observed distribution very well compared to the Weibull, Rayleigh, log-normal, log-logistic, Laplace and inverse Gaussian distributions. From Figure 4.1, the worst distribution is given by the Rayleigh distribution.

Table 4.5 shows the values of the performance indicators which compare the seven distributions for fitting the CO concentration in Kuching.

Table 4.5: The performance indicators value for CO concentration in Kuching

DISTRIBUTION	PERFORMANCE INDICATOR							
	RMSE		IA		PA		R^2	
	1998	2002	1998	2002	1998	2002	1998	2002
Weibull	0.069	0.063	0.990	0.992	0.981	0.986	0.962	0.971
Gamma	0.045	0.038	0.996	0.997	0.993	0.995	0.986	0.990
Log-normal	0.071	0.099	0.992	0.985	0.996	0.990	0.992	0.980
Laplace	0.161	0.154	0.950	0.954	0.937	0.941	0.878	0.885
Pareto	0.144	0.126	0.970	0.977	0.987	0.987	0.973	0.973
Rayleigh	0.086	0.088	0.984	0.984	0.975	0.978	0.950	0.955
Log-logistic	0.169	0.219	0.957	0.936	0.953	0.932	0.907	0.868
Inverse Gaussian	0.095	0.117	0.986	0.980	0.995	0.991	0.991	0.981

From Table 4.5, the smallest value for RMSE is given by the gamma distribution for 1998 and 2002. The highest value for IA is also given by the gamma distribution for both years. Based on the analysis of these results, the best distribution that can fit the CO concentration in Kuching is the Gamma distribution for 1998 and 2002.

4.6 THE EXCEEDENCES VALUE FOR CO OBSERVATIONS

The exceedences for CO observations are based on the distributions that have been chosen as the best distribution for CO concentrations at the three sites. The values for the exceedences are defined from the cdf plots for the distributions.

4.6.1 SEBERANG PERAI

The distribution that fits the CO concentration in Seberang Perai is the gamma distribution for 1998 and the inverse Gaussian distribution for 2002. For both distributions, the probability that the concentration level is less than or equal to 9 ppm is equal to 1 [that is, $F(X \leq 9) = 1$] and the probability that the CO concentration is more

than 9 ppm is equal to 0 $[\text{that is}, F(X>9)=0]$. This shows that there is no incidence where the CO concentration exceeds 9ppm for 1999 and 2003 in Seberang Perai.

4.6.2 KUALA LUMPUR

The distribution that fits the CO concentration in Kuala Lumpur for 1998 is the Weibull distribution and for 2002 it is best represented by the inverse Gaussian distribution. It can be shown that for 1998, the probability that the CO concentration is more than 9 ppm is 0.0068 $[\text{that is}, F(X>9)=0.0068]$. This shows that there will be 2.5 days where the CO concentration in 1999 which will exceed 9ppm. Thus the return period for 1999 is once per 146 days.

For 2002, there is no incidence where the CO concentration exceeds 9ppm for 2003 in Kuala Lumpur.

4.6.3 KUCHING

The distribution that fits the CO concentration in Kuching is the gamma distribution for both years. The probability that the concentration level is less than or equal to 9 ppm is equal to 1 $[\text{that is}, F(X\le 9)=1]$ and the probability that the CO concentration is more than 9 ppm is equal to 0 $[\text{that is}, F(X>9)=0]$. This shows that there is no incidence where the CO concentration exceeds 9ppm for 1999 and 2003 in Kuching.

4.7 PARAMETER ESTIMATES USING THE GLD

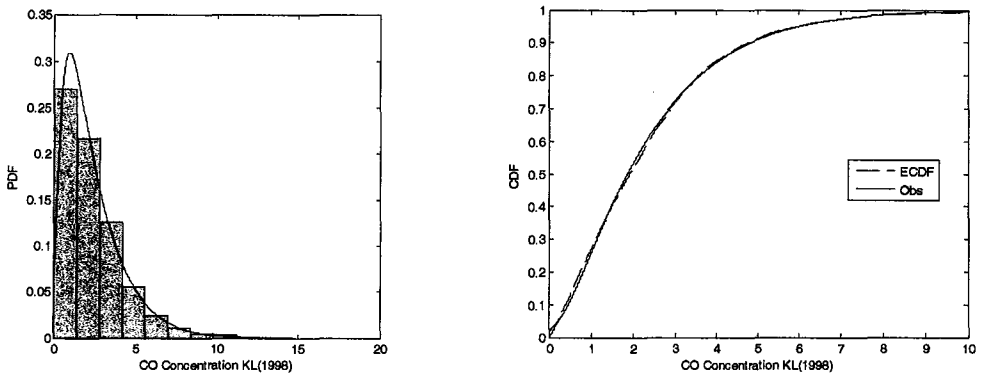
Table 4.6 shows the parameter estimates of the GLD for the three sites for 1998 and 2002. All the estimates have been obtained using the method of moments as discussed in Chapter 3.

Table 4.6: Parameter estimates using method of moments for the three sites

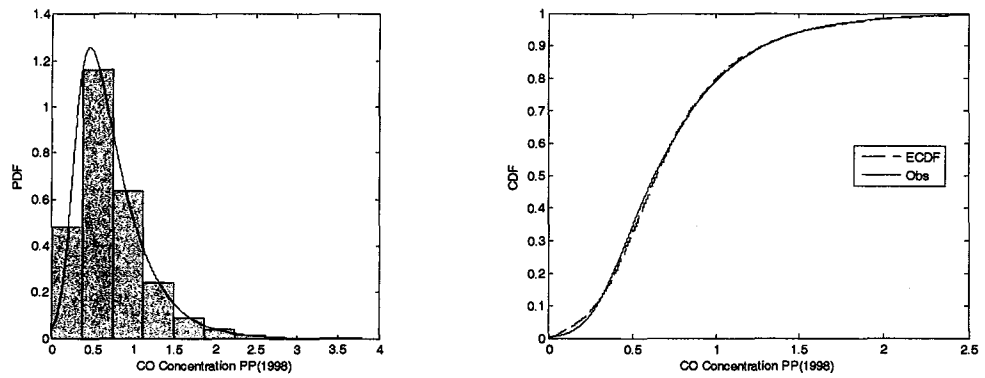
Sites	Year	λ_1	λ_2	λ_3	λ_4
Seberang Perai	1998	0.3919	0.0397	0.0025	0.0161
	2002	0.2649	0.0380	0.0009	0.0144
Kuala Lumpur	1998	0.6756	0.0251	0.0044	0.0483
	2002	0.2650	0.0380	0.0009	0.0144
Kuching	1998	0.3149	-0.0153	-0.0006	-0.0050
	2002	0.2872	0.0819	0.0032	0.0296

Figure 4.1 show the probability density functions and cumulative distributions for the three sites for 1998. Figure 1(a), (b) and (c) show a very good fit to the observed data.

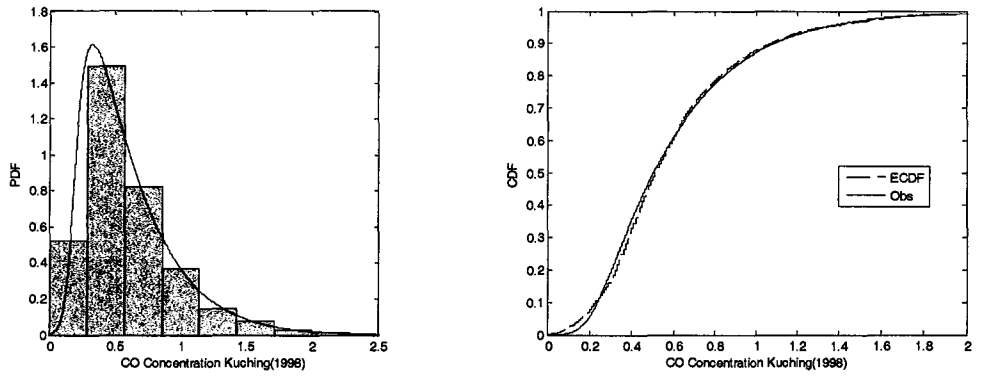
Table 4.7 shows the parameter estimates of the GLD for the three sites using the method of percentiles for 1998 and 2002.



(a) Kuala Lumpur CO Concentration (1998)



(b) Seberang Perai CO Concentration (1998)



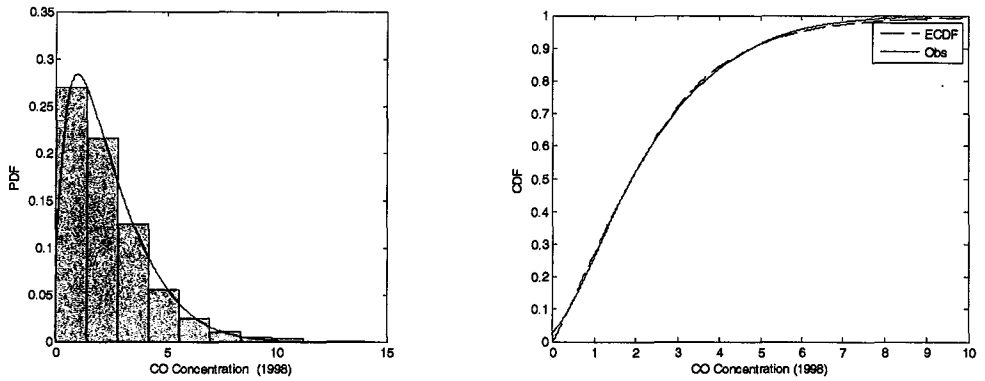
(c) Kuching CO Concentration (1998)

Figure 4.1: PDF and CDF for (a) Kuala Lumpur (b) Seberang Perai and (c) Kuching CO Concentrations Using Method Of Moments for 1998

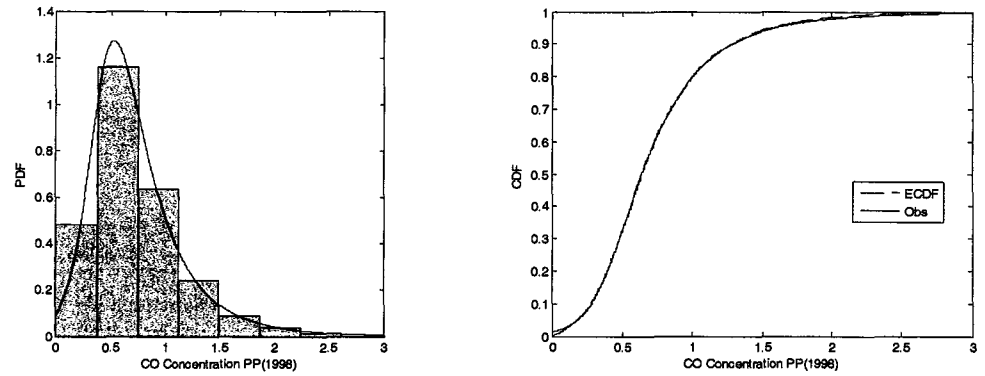
Table 4.7: Parameter estimates using method of percentiles for the three sites

Sites	Year	λ_1	λ_2	λ_3	λ_4
Seberang Perai	1998	0.5008	-0.5351	-0.0548	-0.1548
	2002	0.3687	-0.5242	-0.0344	-0.1421
Kuala Lumpur	1998	0.6897	0.0646	0.0142	0.1349
	2002	0.3687	-0.5242	-0.0344	-0.1421
Kuching	1998	0.2980	0.1105	0.0036	0.0379
	2002	0.3908	-0.9497	-0.0704	-0.2181

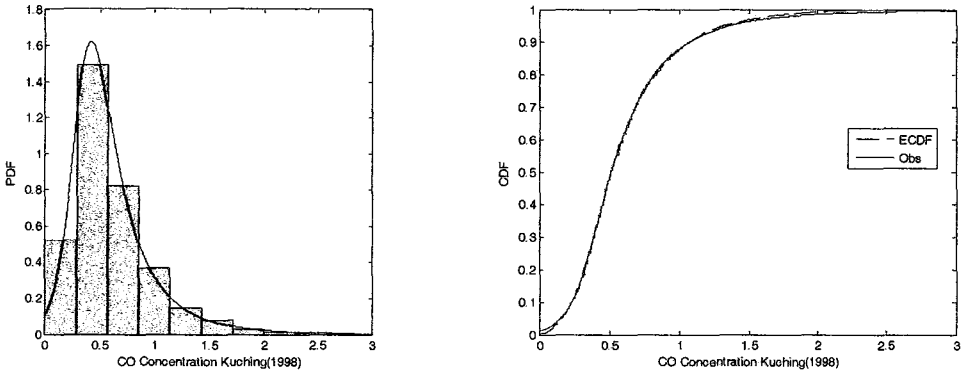
Figure 4.2 show the probability density functions and cumulative distributions for the three sites using the method of percentiles. Figure 4.2 also show very good fit to the observed data.



(a) Kuala Lumpur CO Concentration (1998)



(b) Seberang Perai CO Concentration (1998)



(c) Seberang Perai CO Concentration (1998)

Figure 4.2: PDF and CDF for (a) Kuala Lumpur (b) Seberang Perai and (c) Kuching CO Concentrations using method of percentiles for 1998

4.8 PERFORMANCE INDICATORS USING GLD

To compare the performance between the method of moment estimator with the method of percentile estimator for fitting the GLD, performance indicators were calculated. Five performance indicator measurements for 1998 were calculated and the results are given in Table 4.8 while Table 4.9 provides performance indicator measurements for 2002.

Table 4.8: Performance indicators for the three sites (1998)

Site	PERFORMANCE INDICATOR									
	NAE		*PA*		R^2		*RMSE*		*IA*	
	Mom	Per	Mom	Per	Mom	Per	Mom	Per	Mom	Per
Seberang Perai	0.0364	**0.0211**	0.9916	**0.9967**	0.9830	**0.9931**	0.0713	**0.0472**	0.9925	**0.9968**
Kuala Lumpur	**0.0415**	0.0437	**0.9938**	0.98819	**0.9874**	0.9763	**0.2711**	0.3537	**0.9942**	0.9899
Kuching	**0.0306**	0.0429	**0.9900**	0.9879	**0.9799**	0.9759	**0.0626**	0.0741	**0.9914**	0.9875

* Mom: Method of moments; Per: Method of percentiles

From Table 4.8, the CO concentration is best fitted with the method of percentiles for Seberang Perai. The method of moments give better fit for the CO concentration in Kuala Lumpur and Kuching.

Table 4.9: Performance indicators for the three sites (2002)

Site	PERFORMANCE INDICATOR									
	NAE		*PA*		R^2		*RMSE*		*IA*	
	Mom	Per	Mom	Per	Mom	Per	Mom	Per	Mom	Per
Seberang Perai	0.0399	**0.0280**	0.9898	**0.9956**	0.9794	**0.9909**	0.0672	**0.0512**	0.9914	**0.9951**
Kuala Lumpur	0.0399	**0.0279**	0.9897	**0.9955**	0.9792	**0.9908**	0.0674	**0.0514**	0.9914	**0.9951**
Kuching	0.0374	**0.0227**	0.9918	**0.9991**	0.9834	**0.9979**	0.0585	**0.0273**	0.9929	**0.9985**

* Mom: Method of moments; Per: Method of percentiles

From Table 4.9, the CO concentration for 2002 is best fitted with the method of percentiles for all three sites since the error measurements are smallest and accuracy measurements are close to 1.

4.9 THE EXCEEDENCES VALUE FOR CO CONCENTRATION USING GLD

The exceedences for CO observations are based on the distributions that have been chosen as the best distribution for CO concentrations at the three sites. The values for the exceedences are obtained from the cumulative distribution functions.

4.9.1 SEBERANG PERAI

For both distributions, the probability that the concentration level is less than or equal to 9 ppm is equal to 1 $[\text{that is, } F(X \leq 9) = 1]$ and the probability that the CO concentration is more than 9 ppm is equal to 0 $[\text{that is, } F(X > 9) = 0]$. This shows that there is no incidence where the CO concentration exceeds 9ppm for 1999 and 2003 in Seberang Perai.

4.9.2 KUALA LUMPUR

It can be shown that for 1998, the probability that the CO concentration is more than 9 ppm is 0.0069 $[\text{that is, } F(X > 9) = 0.0068]$. This shows that there will be 2.5 days where the CO concentration in 1999 which will exceed 9ppm. Thus the return period for 1999 is once per 146 days.

For 2002, there is no incidence where the CO concentration exceeds 9ppm for 2003 in Kuala Lumpur.

4.9.3 KUCHING

For both distributions, the probability that the concentration level is less than or equal to 9 ppm is equal to 1 $[\text{that is, } F(X \leq 9) = 1]$ and the probability that the CO concentration is more than 9 ppm is equal to 0 $[\text{that is, } F(X > 9) = 0]$. This shows that there is no incidence where the CO concentration exceeds 9ppm for 1999 and 2003 in Kuching.

4.10 EXTREME VALUE DISTRIBUTIONS

Two types of extreme value distributions (EVD) are used that is Gumbel and Frechet distributions to fit the CO concentration for Seberang Perai, Kuala Lumpur and Kuching. Frequency factor of $c = 2$ and maximum daily CO concentration were used.

4.10.1 FITTING GUMBEL DISTRIBUTIONS

The parameters of the Gumbel distribution for the three sites using $c = 2$ and maximum daily concentration are given in Table 4.10 below. The parameters were estimated using the maximum likelihood estimator.

Table 4.10: Parameter estimates for Gumbel distribution

Site	Year	$c = 2$		max	
		σ	μ	σ	μ
Seberang Perai	1998	0.238	1.832	0.454	1.377
	2002	0.235	1.629	0.439	1.249
Kuala Lumpur	1998	1.020	7.067	2.211	5.362
	2002	0.693	6.116	1.417	5.083
Kuching	1998	0.205	1.556	0.404	1.007
	2002	0.205	1.556	0.377	0.972

From Table 4.6, it can be seen that the location and scale parameter have the largest value for the distribution of the CO concentration in Kuala Lumpur. This means that CO monoxide concentration is highest in Kuala Lumpur. This is followed by the CO concentration in Seberang Perai and then in Kuching.

4.10.2 FITTING FRECHET DISTRIBUTIONS

The parameters of the Frechet distribution for the three sites using *ff2* and and maximum daily concentration are given in Table 4.11 below.

Table 4.11: Parameter estimates for Frechet distribution

Site	Year	$c = 2$		max	
		σ	λ	σ	λ
Seberang Perai	1998	1.816	8.373	1.298	2.705
	2002	1.613	7.519	1.165	2.274
Kuala Lumpur	1998	6.996	7.587	4.777	1.793
	2002	6.077	9.503	4.873	3.347
Kuching	1998	1.543	8.138	0.923	2.318
	2002	1.543	8.138	0.898	2.432

From Table 4.11, it can be seen that the scale parameter have the largest value for the distribution of the CO concentration in Kuala Lumpur. This means that CO monoxide concentration has more variability in Kuala Lumpur. This is followed by the CO concentration in Seberang Perai and then in Kuching.

4.11 PERFORMANCE INDICATORS USING EVD

The best distribution that fits the observed distribution can be chosen by looking at the performance indicators. Table 4.12 and Table 4.13 show the values of the performance indicators for the CO concentration in Seberang Perai, Kuala Lumpur and Kuching for the year 1998.

Table 4.12 shows that for *ff2*, the Frechet distribution provides the best fit for the CO concentration for all three sites in 1998. This is because the *NAE* and *RMSE* have the smallest values and the *PA*, R^2 and *IA* are near 1. From Table 4.13, for the maximum daily data the Gumbel distribution provides the best fit for the CO concentration for all three sites.

Table 4.14 and Table 4.15 give the values of the performance indicators for the three sites for the year 2002. From Table 4.14 and 4.15, the Gumbel distribution gives the best fit for the three sites using *ff2* and maximum daily data.

Table 4.12: Performance Indicator using *ff2* for (1)Gumbel (G) and (2)Frechet (F) Distributions (1998)

Site	PERFORMANCE INDICATOR									
	NAE		*PA*		R^2		*RMSE*		*IA*	
	G	F	G	F	G	F	G	F	G	F
Seberang Perai	0.030	**0.023**	0.983	**0.994**	0.961	**0.982**	0.097	**0.066**	0.979	**0.991**
Kuala Lumpur	0.035	**0.026**	0.985	**0.989**	0.965	**0.973**	0.367	**0.277**	0.983	**0.991**
Kuching	0.025	**0.018**	**0.991**	**0.991**	**0.975**	**0.975**	0.056	**0.045**	0.990	**0.994**

Table 4.13: Performance Indicator using maximum daily data for (1)Gumbel (G) and (2)Frechet (F) Distributions (1998)

Site	PERFORMANCE INDICATOR									
	NAE		*PA*		R^2		*RMSE*		*IA*	
	G	F	G	F	G	F	G	F	G	F
Seberang Perai	0.015	0.621	0.997	0.878	0.989	0.766	0.043	1.137	0.998	0.427
Kuala Lumpur	0.039	0.817	0.989	0.908	0.972	0.819	0.517	5.872	0.990	0.378
Kuching	0.015	0.765	0.998	0.859	0.990	0.733	0.039	1.073	0.998	0.413

Table 4.14: Performance Indicator using *ff2* for (1)Gumbel (G) and (2)Frechet (F) Distributions (2002)

Site	PERFORMANCE INDICATOR									
	NAE		*PA*		R^2		*RMSE*		*IA*	
	G	F	G	F	G	F	G	F	G	F
Seberang Perai	0.027	0.302	0.987	0.771	0.969	0.592	0.079	0.561	0.985	0.540
Kuala Lumpur	0.026	0.131	0.986	0.748	0.968	0.557	0.245	1.013	0.984	0.447
Kuching	0.025	0.422	0.991	0.801	0.975	0.637	0.056	0.733	0.989	0.441

Table 4.15: Performance Indicator using maximum daily data for (1)Gumbel (G) and (2)Frechet (F) Distributions (2002)

Site	PERFORMANCE INDICATOR									
	NAE		*PA*		R^2		*RMSE*		*IA*	
	G	F	G	F	G	F	G	F	G	F
Seberang Perai	0.017	0.706	0.996	0.884	0.986	0.778	0.053	1.168	0.998	0.411
Kuala Lumpur	0.025	0.615	0.991	0.900	0.977	0.806	0.281	3.928	0.993	0.380
Kuching	0.023	0.748	0.997	0.869	0.986	0.751	0.044	1.003	0.998	0.418

4.12 THE EXCEEDENCES VALUE FOR CO CONCENTRATION USING EVD

The exceedences for CO observations are based on the distributions that have been chosen as the best distribution for CO concentrations at the three sites. The values for the exceedences are obtained from the respective cumulative distribution functions.

4.12.1 SEBERANG PERAI

The distribution that fits the CO concentration in Seberang Perai for 1998 is the Frechet distribution when *ff2* was used and it is represented by the Gumbel distribution when maximum daily concentration was used. For both these distributions, the probability that the concentration level is less than or equal to 9 ppm is equal to 1 $[\text{that is, } F(X \le 9) = 1]$ and the probability that the CO concentration is more than 9 ppm is equal to 0 $[\text{that is, } F(X > 9) = 0]$. This shows that there is no incidence where the CO concentration exceeds 9ppm for 1999 in Seberang Perai.

For 2002, the Gumbel distribution is the best distribution when *ff2* and maximum daily concentration was used. For both methods, the probability that the CO concentrations exceed 9ppm is 0. This shows that there is no incidence where the CO concentration exceeds 9ppm for 2003 in Seberang Perai.

4.12.2 KUALA LUMPUR

The distribution that fits the CO concentration in Kuala Lumpur for 1998 is the Frechet distribution when threshold values was used and it is represented by the Gumbel distribution when maximum daily concentration was used. Using *ff2*, the probability that the concentration level is less than or equal to 9 ppm is equal to 0.863 $[\text{that is, } F(X \le 9) = 0.863]$ and the probability that the CO concentration is more than 9 ppm is equal to 0.137. Using the maximum daily concentration, the probability that the CO concentration exceeds 9 ppm is 0.175.

For 2002, the Gumbel distribution is the best distribution when *ff2* and maximum daily concentration was used. Using the threshold value, the probability that the concentration

level is more than 9 ppm is 0.016. Using the maximum daily concentration, the probability that the CO concentration exceeds 9 ppm is 0.061.

4.12.3 KUCHING

For 1998, the distribution that fits the CO concentration in Kuching is the Frechet distribution using *ff2* and the Gumbel distribution for the maximum daily concentration. For both distributions, the probability that the concentration level is less than or equal to 9 ppm is equal to 1 $[\text{that is, } F(X \leq 9) = 1]$ and the probability that the CO concentration is more than 9 ppm is equal to 0 $[\text{that is, } F(X > 9) = 0]$. This shows that there is no incidence where the CO concentration exceeds 9ppm for 1999 in Kuching.

For 2002, the Gumbel distribution is the best distribution when threshold values and maximum daily concentration was used. For both methods, the probability that the CO concentrations exceed 9ppm is 0. This shows that there is no incidence where the CO concentration exceeds 9ppm for 2003 in Kuching.

CHAPTER 5

CONCLUSIONS

5.1 CONCLUSIONS

From this research, the following conclusions were obtained.

1. The characteristics of CO concentrations in Kuala Lumpur, Kuching and Seberang Perai were investigated. The pdf were established and discussed in Chapter 4. From the results of the statistical analysis, it indicates that the mean of the CO concentrations for the three monitoring records are greater than the median values. It shows that all of the observations are skewed to the right.

2. From the cdf plots and performance indicator values, the best distributions that fit the observations are given in Table 5.1.

Table 5.1: Best distributions to represent CO concentrations

Sites	Best distribution for 1998	Best distribution for 2002
Seberang Perai	Gamma	Inverse Gaussian
Kuala Lumpur	Weibull	Inverse Gaussian
Kuching	Gamma	Gamma

3. The pdf and cdf plots obtained in this research can be used to predict the return period for the coming year. In this research, the probabilities for air pollutants emissions exceeding the Malaysian Ambient Air Quality Guidelines (MAAQG) have been successfully predicted. For the 1998 data, Kuala Lumpur was predicted to exceed 9ppm for 2.5 days in 1999 with a return period of one occurrence per 146 days. However, Seberang Perai and Kuching does not exceed the MAAQG. Based on the 2002 data, it can be concluded that the CO concentration levels in Seberang Perai, Kuala Lumpur and Kuching does not exceed the Malaysian Ambient Air Quality Guidelines of 9 ppm.

4. The parameters of the GLD was also obtained. The parameters of the best GLD is given in Table 5.2.

Table 5.2: GLD parameters to represent CO concentrations

Sites	Year	λ_1	λ_2	λ_3	λ_4
Seberang Perai	1998	0.5008	-0.5351	-0.0548	-0.1548
	2002	0.3687	-0.5242	-0.0344	-0.1421
Kuala Lumpur	1998	0.6756	0.0251	0.0044	0.0483
	2002	0.3687	-0.5242	-0.0344	-0.1421
Kuching	1998	0.3149	-0.0153	-0.0006	-0.0050
	2002	0.3908	-0.9497	-0.0704	-0.2181

5. The probability of exceedences and return period using GLD was also obtained. The results are similar to that obtained using the probability distributions. Only

Kuala Lumpur in 1998 exceeds the MAAQG. It is predicted that in 1999, the exceedence of CO concentration exceeding 9ppm is 2.5 days with a return period of 146 days for Kuala Lumpur. Seberang Perai and Kuching does not exceed the MAAQG for 1998 and 2002.

6. The probability density functions and cumulative distribution functions for two extreme value distributions have been obtained. The best distributions between these two extreme value distributions were obtained by comparing its performance indicators. For 1998, the best distributions that fit the observations are the Frechet distribution using *ff2* for all three sites. By using the maximum daily data, the Gumbel distribution is the best distribution for the three sites. For 2002, the Gumbel distribution is the best distribution for all sites using both the *ff2* and maximum daily data.

7. From these distributions and its cumulative distribution functions, it can be concluded that the CO concentration levels in Seberang Perai and Kuching does not exceed the Malaysian Ambient Air Quality Guidelines of 9 ppm based on the 1998 and 2002 data. However, the CO concentration levels in Kuala Lumpur exceed the Malaysian Ambient Air Quality Guidelines of 9 ppm based on the 1998 data. The probabilities of exceedences are 0.14 and 0.18 respectively for *ff2* and maximum daily data. For 2002, the probability of exceedences is 0.016 and 0.061 respectively for *ff2* and maximum daily data.

8. The probability density functions and cumulative distribution functions obtained in this research can be used to predict the return period for the coming year. In this research, the probabilities for air pollutants emissions exceeding the Malaysian Ambient Air Quality Guidelines have been successfully predicted.

REFERENCES

Abramowitz, M. and Stegun, I. A. (1972) *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables.* New York: Dover. Pp. 930

Afroz, R., Hassan, M. N. and Ibrahim, N. A. (2003) Review of air pollution and health impacts in Malaysia. *Environment Research.* 92 (2), p.71 – 77

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

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

Beyer, W. H. (1987) *CRC Standard Mathematical Tables, 28th Edition.* Boca Raton FL: CRC Press. Pp. 534 – 535

Bottazzi, G. and Secchi, A. (2003) A General Explanation for the Laplace Distribution of Firm's Growth Rates. *Nature.* 58 (241), p. 754 – 758

Canter, L. W. (1996) *Environmental Impact Assessment, Second Edition.* New York: McGraw Hill

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), p. 593 – 604

Chapman, S. J. (2002) *MATLAB Programming for Engineers, Second Edition.* Australia: The Math Works, Inc.

Chhikara, R. S. and Folks, J. L. (1989) The inverse Gaussian Distribution as a Lifetime Model. *Thechnometrics.* 19 (4), p. 461 – 468

Department of Environment, Malaysia (1996) *Malaysia Environment Quality Report 1996.* Malaysia: Department of Environment

Department of Environment, Malaysia (1998) *Malaysia Environment Quality Report 1998.* Malaysia: Department of Environment

Department of Environment, Malaysia (2002) *Malaysia Environment Quality Report 2002.* Malaysia: Department of Environment

Encarta (2006) [Online], [Accessed 15th March 2006]. Available from World Wide Web: http://www.encarta.msn.com

Evans, M. Hastings, N. and Peacock, B. (2000) *Statistical Distributions, 3rd Edition.* New York: Wiley. Pp. 34 – 42

Fellenberg, G. (2000) *The Chemistry of pollution.* England: John Wiley & Sons Ltd.

Gajjar, A. V. and Khatri, C. G. (1969) Progressively Censored Samples From Log-normal and Log-logistic Distributions. *Thechnometrics.* 11 (94), p. 458 – 467

Gatti, D. D., Guilmi, C. D., Gaffeo, E., Giulioni, G., Gallegati, M. and Palestrini, A. (2005) A New Approach to Business Fluctuations: Heterogeneous Interacting Agents, Scaling Laws and Financial Fragility. *Journal of Economic Behavior and Organization.* 56 (3), p. 489 – 512

Georgepoulos, P. and Seinfeld, J. (1982) Statistical Distributions of Air Pollutant Concentrations. *Environmental Science and Technology.* 16 (54), p. 401A – 415A

Gilbert, O. R. (1987) *Statistical Method for Environmental Pollution Monitoring.* New York: Van Nostrand Reinhold Company Inc.

Godish, T. (1997) *Air Quality (Third Edition).* New York: Lewis Publishers.

Gokhale, S. and Khare, M. (2005) A Hybridmodel for Predicting Carbon Monoxide from Vehicular Exhausts in Urban Environments. *Atmospheric Environment.* 39 (22), p. 4025 – 4040

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), p. 455 – 467

Harikrishna, M. and Arun, C. (2003) Stochastic Analysis for Vehicular emissions on Urban Roads – A Case Study of Chennai. In: Martin J. Bunch, V. Madha Suresh and T Vasantha Kumaran (eds) Proceeding of the third International Conference on Environmental and Health, 15 – 17 December 2003, Chennai. India

Holland, D. M. and Fitz – Simons, T. (1982) Fitting Statistical Distribution to Air Quality Data by the Maximum Likelihood Method. *Atmospheric Environment.* 16 (74), p. 1071 – 1076

Hoshmand, A. R. (1998) *Statistical Method for Environmental and Agricultural Sciences.* Florida: CRC Press LLC

Hughes, G (1997) *Can the environment wait? Priority issues for East Asia.* Washington: World Bank

Ibrahim, C. A. (2004) Overview of Air Quality Management in Malaysia, University Science of Malaysia, 3 September 2004

Jakeman, A. J., Taylor, J. A. and Simpson, R. W. (1986) Modeling Distribution of Air Pollutant Concentrations – II. Estimation of One and Two Parameters Statistical Distribution. *Atmospheric Environment.* 20 (6), p. 2435 – 2447

Johnson, G.L. (2001) Wind Energy Systems. *Journal of Environmental Management.* 75 (43), p. 648 – 653

Kao, A. S. and Friedlander, S. K. (1995) Frequency Distributions of PM_{10} Chemical Components and Their Sources. *Environmental Science and Technology.* 29(5), p.19 – 28

Kottegoda, N. T. and Rosso, R. (1998) *Statistic, Probability and Reliability for Civil and Environmental Engineers.* Singapore: McGraw – Hill.

Lu, H. C. (2002) The Statistical Character of PM_{10} Concentration in Taiwan Area. *Atmospheric Environment.* 36 (9), p. 491 – 502

Lu, H. C. (2003) Estimating the Emission Source Reduction of PM_{10} in Central Taiwan. *Journal of Chemosphere.* 54 (7), p.805 – 814

Lu, H. C. and Fang, G. C. (2003) Predicting the Exceedences of a critical PM_{10} Concentration – A Case Study in Taiwan. *Atmospheric Environment.* 37 (4), p. 3491 – 3499

Maffeis, G. (1999) Prediction of Carbon Monoxide Acute Air Pollution Episodes. Model Formulation and First Application in Lombardy. *Atmospheric Environment.* 33 (12), p. 3859 – 3872

Mage, D. T. and Ott, W. R. (1984) An Evaluation of the Method of Fractiles, Moments and Maximum Likelihood for Estimating Parameters When Sampling Air Quality from a Stationary Log-normal Distribution. *Atmospheric Environment.* 18 (5), p. 163 – 171

Ministry of Housing and Local Government (2006) [Online], [Accessed 25th January 2006]. Available from World Wide Web: http:// www.kpkt.gov.my

Ministry of Transport (2006) [Online], [Accessed 25th January 2006]. Available from World Wide Web: http:// www.mot.gov.my

McBean E. A, Rovers, F. A. (1998) *Statistical Procedures for Analysis of Environmental Monitoring Data and Risk Assessment.* New Jersey: Prentice Hall PTR. Pp 71 – 72.

Mendenhall, W. and Sincich, T. (1995) *Statistics for Engineering and the Sciences: Fourth Edition.* Upper Saddle River, New Jersey: Prentice Hall. Pp. 40 – 44.

Morel, B., Yen, S. and Cifuentes, L. (1999) Statistical Distribution for Air Pollutant Applied to the Study of the Particulate Problem in Santiago. *Atmospheric Environment.* 33 (7), p. 2575 – 2585

Mudholkar, G.S. and Tian, L. (2002) An Entropy Characterization of the inverse Gaussian Distribution and Related Goodness-of-fit Test. *Journal of Statistical Planning and Inference.* 102 (2), p. 211 – 221

Ott, W. R. (1990) A Physical Explanation of the Log-normality of Pollutant Concentrations. *Journal of the Air and Waste Management Association.* 40 (10), p. 1378 – 1383

Ott, W. R. (1995) *Environmental Statistics and Data Analysis.* New york: Lewis Publishers

Pang, W. K., Hou, S. H., Yu, B. W. T. and Li, K. W. K. (2002) A Simulation Based Approach to the Parameter Estimation for the Three-Parameter Gamma Distribution. *European Journal of Operational Research.* 155(9), p.675 – 682

Pang, W. K., Leung, P. K., Huang, W. K. and Liu, W. (2003) On Interval Estimation of the Coefficient of Variation for the Three-Parameter Weibull, Lognormal and Gamma Distribution: A Simulation Based Approach. *European Journal of Operational Research.* 164 (7), p. 367 – 377

Peavy, H. S., Rowe, D. R. and Tchobanoglous, G. (1985) *Environmental Engineering.* Singapore: McGraw-Hill Co

Piegorsch W. W. & Bailer A. J. (1997) *Statistic for Environment Biology and Toxicity.* Australia: Chapman & Hall Publication. Pp. 491

Rao, C. R. (1973) *Linear Statistical Inference and its Applications: Second Edition.* New york: Wiley

Romano, D., Bernetti, A. and Lauretis, R. D. (2004) Different methodologies to Quantifyf Uncertainties of Air Emissions. *Environment International.* 30 (8), p. 1099 – 1107

Scheaffer, R. L. (1995) *Introduction to Probability and Its Applications, Second Edition.* United States: Duxbury Press

Schorp, M. K and Leyden, D. E. (2002) Distribution Analysis of Airborne Nicotine Concentrations in Hospitality Facilities. *Journal of Environment International.* 27 (34), p. 567 – 578

Seinfeld, J. H. and Pandis, S. N. (1998) *Atmospheric Chemistry and Physics. From Air Pollution to Climate Change.* John Wiley and Sons Publication

Singh, P. (2004) Simultaneous Confidence Intervals for the Successive Ratios of Scale Parameters. *Journal of Statistical Planning and Inference.* 36 (3), p. 1007 – 1019

Stanley, M., Amaral, L. Buldyrev, S., Havlin, S., Leschorn, H., Maas, P., Salinger, M. and Stanley, E. (1996) Scaling Behavior in the Growth of Companies. *Nature.* 379 (1996), p. 804 – 806

Tchobanoglous, G., Peavy, H. S. and Rowe, D. R. (1985) *Environmental Engineering.* Singapore: McGraw-Hill

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), p. 4383 – 4402

World Health Organization (1987) *Air Quality guidelines for Europe. European Series N.23* Denmark: WHO Regional Office for Europe

World health Organization (1998) *Report of the Bioregional Workshop on Health Impacts of Haze Relared Air Pollution.* Manila: WHO

LIST OF PUBLICATIONS

LIST OF PUBLICATIONS

1. Ahmad Shukri Yahaya, Nor Azam Ramli, Jannatul Naemah Mohd Sedek. Steady State Probabilities For Carbon Monoxide Concentration In Kuala Lumpur, Proceedings Of The International Conference On Quantitative Sciences And Its Applications (ICOQSIA2005), 2005, 6-8 December, Penang, Malaysia.

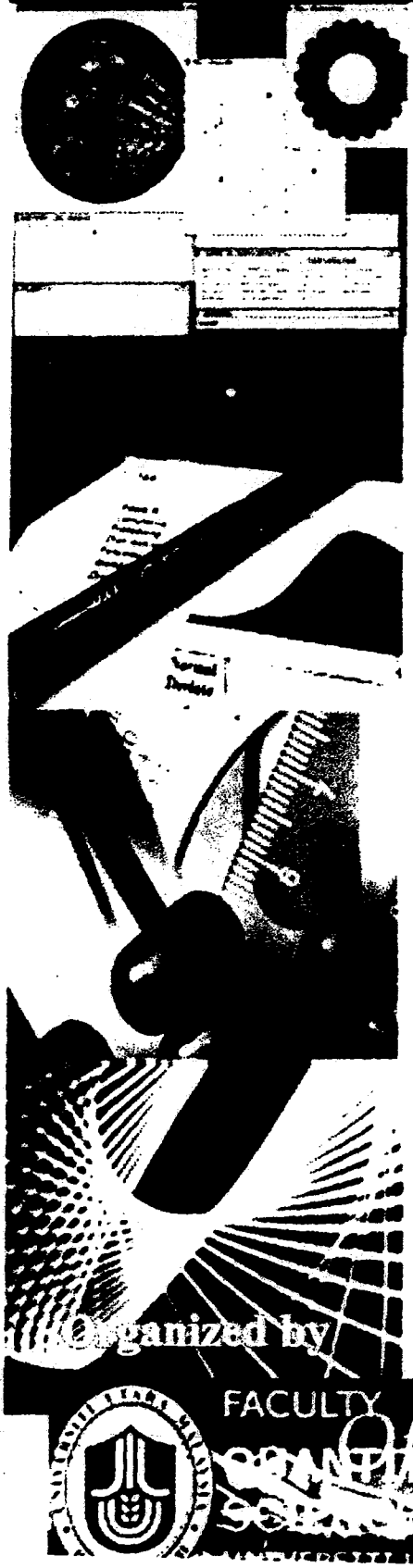
2. Ahmad Shukri Yahaya, Nor Azam Ramli, Noor Faizah Fitri Md Yusof . Fitting Extreme Value Distribution To CO Data, Proceedings of the 3rd Bangi World Conference on Environmental Management, 5-6 September 2006, Bangi, 324-330.

3. Ahmad Shukri Yahaya, Nor Azam Ramli, Hazrul Abdul Hamid. Review Of Fitting Distributions On Air Pollution Modelling, Prosiding Simposium Kebangsaan Sains Matematik Ke-15, 5-7 Jun 2007, Hotel Concorde, Shah Alam, Selangor, 449- 454.

4. Ahmad Shukri Yahaya, Norlida Mohd Noor, Chan Yin Yin, Lee Mun Yee. Ujian Kebagusan Penyuaian Bagi Taburan Gumbel, Prosiding Simposium Kebangsaan Sains Matematik Ke-15, 5-7 Jun 2007, Hotel Concorde, Shah Alam, Selangor, 461-466.

2005

EMPOWERMENT FOR ORGANIZATION INTELLIGENCE



INTERNATIONAL CONFERENCE

on

QUANTITATIVE SCIENCES

and

ITS APPLICATIONS

6-8 DECEMBER 2005,
THE GURNEY RESORT HOTEL
PENANG, MALAYSIA

Abstract

Organized by

FACULTY OF QUANTITATIVE

Multimedia Super Corridor

STEADY STATE PROBABILITIES FOR CARBON MONOXIDE CONCENTRATION IN KUALA LUMPUR

NOR AZAM RAMLI[+], AHMAD SHUKRI YAHAYA & JANNATUL NAEMAH MOHAMED SEDEK

School of Civil Engineering, Engineering Campus, Universiti Sains Malaysia, 14300 Nibong Tebal, Seberang Perai Selatan, Pulau Pinang, Malaysia.
shukri@eng.usm.my

School of Civil Engineering, Engineering Campus, Universiti Sains Malaysia, 14300 Nibong Tebal, Seberang Perai Selatan, Pulau Pinang, Malaysia.
ceazam@eng.usm.my

School of Civil Engineering, Engineering Campus, Universiti Sains Malaysia, 14300 Nibong Tebal, Seberang Perai Selatan, Pulau Pinang, Malaysia.

Abstract. The rapid economic growth of Kuala Lumpur has imposed costs in terms of industrial pollution and the degradation of urban degradation. Among them, air pollution is the major issue that has been affecting human health, agricultural crops, forest species and ecosystems. Thus this paper analyses one of the pollutants namely carbon monoxide (CO) in Kuala Lumpur. The data consist of 8760 observations taken every hour for a year. According to the Malaysian Ambient Air Quality Guideline, the concentration of CO for eight hours should not be more than nine parts per million (ppm). Thus three states were chosen that is when the CO concentration is (a) high (b) medium and (c) low. Using the transition matrix of the first order Markov Chain, the steady state distributions were obtained. The result shows that, at steady state, the probability that the CO concentration in Kuala Lumpur is high, medium and low, is 0.051, 0.231 and 0.718 respectively.

Keywords: Markov chain, steady state.

1. Introduction

Monitoring data and studies on ambient air quality show that some of the air pollutants in several large cities such as Kuala Lumpur, Malaysia are increasing with time and are not always at acceptable levels according to the national ambient air quality standards. The goal of achieving industrial country status by the year 2020 and the associated rapid economic growth have started to impose costs in terms of industrial pollution and the degradation of urban environment. Depletion of air and water quality, and contamination by industrial wastes has become more serious in Kuala Lumpur in recent years. Among them, air pollution is the major issue and need to be dealt with before it causes harm to human health and the environment.

From the geographical point of view, Kuala Lumpur which is situated in a valley, namely the Klang Valley is prone to serious air pollution compared to other parts of the country. This happen because the tendency of pollutants to be trapped in the mountain corridor is higher as mountains exist in the east and the Straits of Malacca on the west.

Gokhale and Khare (2004) review the different models that can be used in air quality studies. They classified air quality models into four different types' namely deterministic, statistical, statistical distributions and hybrid. Fourteen models were presented and all models can be used to predict CO concentrations with some models having higher predictive accuracy than others. Among the models used was the Markov-type model which was developed by North *et al.* (1984) based on up and down crossings of threshold concentrations of series of daily CO concentration in Madrid, Spain.

This paper discusses the use of the Markov chain to predict the concentration of CO when steady state is considered. Steady state is said to occur if all other factors are assumed to be constant.

2. The Data

The data used for this analysis is the hourly carbon monoxide (CO) data (measured in parts per million, ppm) for the year 1998 taken at a monitoring station in the Federal Territory of Kuala Lumpur. Kuala Lumpur was chosen because it has the highest growth rate in the field of transportation, utilities and manufacturing activities. As a result the deterioration of air quality is more serious in Kuala Lumpur compared to other parts of the country. Table 1 below describes the data.

Table 1. Descriptive statistics of CO data

Number of observations	7939
Number of missing observations	821
Mean	2.33
Standard deviation	1.88
Minimum	0
Maximum	14.03

From Table 1, it can be seen that there are 821 missing observations which accounts to about ten percent of the data. The missing data occurs mostly in lengths of about two or three. The mean imputation technique was used to replace all missing values. The method with the mean of one datum above the missing value and one datum below the missing value was used. Table 2 below describes the data after the mean imputation technique was used.

Table 2. Descriptive statistics of CO data after the mean imputation technique was used

Number of observations	8760
Number of missing observations	-
Mean	2.27
Standard deviation	1.86
Minimum	0
Maximum	14.03

3. Markov chain

Consider a system that can be in any of a finite number of states, and assume that it moves from state to state according to some prescribed probability law. The system, could record the concentration of pollution from day to day, with the possible states being highly concentrated, medium concentrated and low concentrated. Observing the condition over a long period of time would allow one to find the probability of being highly concentrated tomorrow given that it is low concentrated today. This system is called Markov Chain (Scheaffer, 1995). Hines *et al.*, (2003), defined the Markov Chain as a stochastic process that exhibits

$$P\{X_{t+1} = j / X_t = i\} = P\{X_{t+1} = j / X_t = i, X_{t-1} = i_1, X_{t-2} = i_2, ..., X_0 = i_t\} \quad (1)$$

for $t = 0,1,2,...$ and every sequence $j, i, i_1, ..., i_t$.

The conditional probabilities

$$P\{X_{t+1} = j / X_t = i\} = P_{ij} \quad (2)$$

are called one-step transition probabilities, and are said to be stationary if

$$P\{X_{t+1} = j / X_t = i\} = P\{X_1 = j / X_0 = i\} \quad \text{for } t = 0,1,2,... \quad (3)$$

so that the transition probabilities remain unchanged through time. They are denoted as the transition matrix $P = \lfloor p_{ij} \rfloor$, called the one-step transition matrix. The matrix P has $m+1$ rows and $m+1$ columns, and the sum of probabilities in each row of the transition matrix is one.The existence of the one-step stationary transition probabilities implies that

$$P_{ij}^{(n)} = P\{X_{t+n} = j / X_t = i\} = P\{X_n = j / X_0 = i\} \quad (4)$$

for all $t = 0,1,2,....$ The values $P_{ij}^{(n)}$ are called n-step transition probabilities, and they may be displayed in an n-step transition matrix, $P^{(n)} = \lfloor p_{ij}^n \rfloor$.

The Malaysian ambient air quality guideline (Department of Environment, Malaysia (1998)) states that for CO, the concentration for 8 hours should not be more than 9 ppm. Thus the data set for CO have been recorded as shown in Table 3.

Table 3 : Recoded CO data

Recoded data	Concentration (ppm)
High	> 6.01
Medium	3.01 – 6.0
Low	0.0 – 3.0

4. Results and Discussion

Figure 1 below shows bar chart for the hourly CO concentrations for the year 1998 in Kuala Lumpur. It can be seen that about 73% of the time the CO concentration in Kuala Lumpur is low and only about 4.7% of the time that the CO concentration in Kuala Lumpur is high.

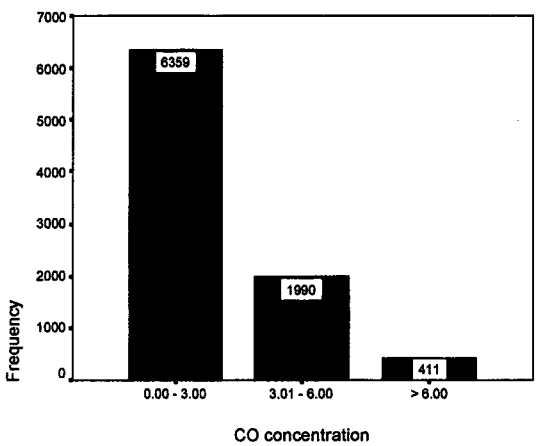


Figure 1. Bar chart for hourly CO concentration in Kuala Lumpur (1998)

Based from the definition of the Malaysian ambient air quality guideline, the transition matrix of the first order Markov Chain model was obtained and is given by

$$P = \begin{array}{c} \\ H \\ M \\ L \end{array} \begin{array}{c} \begin{array}{ccc} H & M & L \end{array} \\ \begin{pmatrix} 0.760 & 0.210 & 0.030 \\ 0.290 & 0.532 & 0.177 \\ 0.078 & 0.388 & 0.534 \end{pmatrix} \end{array} \tag{5}$$

From the first order Markov Chain model, the probability that the CO concentration is high today given that it is high yesterday is 0.760 and the probability that the CO concentration is low today given that it is high yesterday is 0.078.

The *n*-step transition matrix will provide the steady state probabilities for CO concentration and the time taken to reach the steady state. Steady state is assumed to occur if the CO concentrations are constant over time. From the analysis, it was found that steady state occurs at $n = 14$ days and with the *n*-step transition matrix given by

$$P^{14} = \begin{bmatrix} 0.048 & 0.227 & 0.727 \\ 0.048 & 0.227 & 0.727 \\ 0.048 & 0.227 & 0.727 \end{bmatrix} \tag{6}$$

Therefore, the concentration of CO in Kuala Lumpur will reach steady state after 14 days and the steady state probabilities are as given in the *n*-step transition matrix above.

7. Conclusion

This paper discusses a stochastic method called Markov Chain to predict the probability of occurrences of high, medium and low CO concentrations for the year 1998 in Kuala Lumpur. The CO concentrations were recoded according to the Malaysian ambient air quality guideline which states that for CO, the concentration for 8 hours should not be more than 9 ppm. From the first order Markov Chain model, it can shown that the probability that the CO concentration is high today given that it is high yesterday is 0.760 showing the dependency of CO concentration on what happens on the previous day. It was also found that steady state would be reached after 14 days and the steady state probabilities are 0.048, 0.227 and 0.727 for high concentration of CO, medium concentration of CO and low concentration of CO respectively. Thus, it can be concluded that the CO concentration in Kuala Lumpur is still low.

References

1. Department of Environment, Malaysia, 1998. *Malaysia Environment Quality Report 1998*, Kuala Lumpur: Department of Environment, Ministry of Sciences, Technology and the Environment, Malaysia.
2. Gokhale, S., Khare, M., 2004. A Review of Deterministic, Stochastic and Hybrid Vehicular Exhaust Emission Models, *International Journal of Transport Management*, 2, 59-74.
3. Hines, W.W., Montgomery, D.C., Goldsman, D.M., Borror, C.M., 2003. *Probability and Statistics in Engineering*, Fourth Edition, 21, United States, Wiley.
4. Scheaffer, R.L., 1995. *Introduction to Probability and its Applications* , Second Edition, United States, Duxbury Press.

MANAGING CHANGES

Proceedings 3rd Bangi World Conference On Environmental Management

Bangi, 5th – 6th September, 2006

Editors:

Jamaluddin Md. Jahi
Kadir Arifin
Azahan Awang
Muhammad Rizal Razman

Environmental Management Programme
Centre for Graduate Studies
Universiti Kebangsaan Malaysia, Bangi
and
Environmental Management Society (EMS) Malaysia
Malaysian Institute of Historical and Patriotism Studies (IKSEP)

Bangi 2006

34

FITTING EXTREME VALUE DISTRIBUTION TO CO DATA

Ahmad Shukri Yahaya, Nor Azam Ramli & Nor Faizah Fitri Md Yusof

School of Civil Engineering,
Engineering Campus,
Universiti Sains Malaysia,
14300 Nibong Tebal,
Seberang Perai Selatan, Pulau Pinang, MALAYSIA.

E-mail: shukri@eng.usm.my, ceazam@eng.usm.my

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. Among the pollutants that are present is carbon monoxide (CO). Recent studies have shown that three distributions are usually used to fit the whole CO data namely Weibull, gamma and log-normal. However extremes drawn from such distributions tend to a type I distribution. This paper fits the Gumbel distribution which is a Type I distribution to the CO data which was collected at a station in Seberang Perai, Penang. The first step in the analysis is to determine the threshold level of the CO concentration using available data. This was done using the threshold method. Subsequently two methods were used to estimate the parameters of the Gumbel distribution namely the method of moments and the maximum likelihood method. Two performance indicators namely the index of agreement and the root mean square error were used to determine the goodness of fit of the distribution. The results show that the estimates using method of moments for the Gumbel distribution fits well the extreme values when threshold method was used.

INTRODUCTION

Seberang Perai is one of the major towns in Malaysia which is situated not far from a large industrial area and is experiencing rapid development in the industrial and economic sector. As a developing town, Seberang Perai cannot avoid the occurrences of air pollution. Based on the Department of Environment data of air quality status in for the west coast of Peninsular Malaysia in 2003 (Department of Environment 2003), the air pollution index (API) scale which can be categorized as unhealthy (API values between 101-200) occurs for 25 days in Seberang Perai. Therefore, research on ways to reduce the air pollution concentrations is very important. From the geographical point of view, Seberang Perai is strategically located on the north-western coast of Peninsular Malaysia.

Extreme value distributions have been used widely in storm, flood, wind, sea waves and estimation (Kottegoda and Rosso 1998). Sharma et al. (1999) stated that extreme air pollution event, that is, the maximum air pollution concentration is governed by many complex and interrelated factors. As such, deterministic models fail in general to predict extreme event adequately. They used four methods to estimate parameters of the extreme value distributions for CO concentration in India and found that the least square fit and Gumbel's method gave the best fit. Lu and Fang (2003) used extreme value theory to fit the monthly maximum data and high concentration data of air pollutants concentration over a specific percentile. Then the cumulative probability extremes and return period can be estimated.

This paper discusses the use of Gumbel distribution to predict the concentration of CO when two methods of parameters are used. Two different performance indicators are used to obtain the best estimator.

THE DATA

The data used for this analysis is the hourly carbon monoxide (CO) data (measured in parts per billion, ppb) for the year 2002 taken at a monitoring station in Seberang Perai, Pulau Pinang. Pulau Pinang was chosen because it has experienced a rapid growth of population and is a highly industrialised town which is accompanied by a growing number of vehicles that contribute to air pollution. As a result the deterioration of air quality is quite serious in Seberang Perai. Table 1 below describes the data.

Table 1. Descriptive statistics of CO data (ppb)

Number of observations	8232
Number of missing observations	524
Mean	614.82
Standard deviation	384.27
Minimum	10
Maximum	3130
Skewness	1.73
Kurtosis	4.63

From Table 1, it can be seen that the annual mean for CO concentration is 614.82 ppb which is below the Malaysian Ambient Air Quality Guidelines (MAAQG) which is 900 ppb. The concentrations are skewed to the right indicating that high CO concentrations do occur. The maximum CO concentration is 3130 ppb. There are 524 missing observations in the data. These missing observations were ignored from the analyses which are carried out in this paper.

GUMBEL DISTRIBUTION

The Gumbel distribution is an example of a family of extreme value distributions. The tails of distributions are better described by these type of distributions rather than the normal or lognormal distributions (McBean & Rovers 1998).

The probability density function (PDF) of the Gumbel distribution (Kottegoda & Rosso 1998) is given by

$$f(x;\mu,\sigma)=\frac{1}{\sigma}\exp\left[-\frac{x-\mu}{\sigma}-\exp\left(-\frac{x-\mu}{\sigma}\right)\right],0<x<\infty,-\infty<\mu<\infty,\sigma>0 \quad (1)$$

and the cumulative distribution function (CDF) is given by

$$F(x,\mu,\sigma)=\exp\left[-\exp\left(\frac{x-\mu}{\sigma}\right)\right],0<x<\infty,-\infty<\mu<\infty,\sigma>0 \quad (2)$$

where μ and σ are the location and scale parameters which needs to be estimated.

PARAMETER ESTIMATION

The parameters of the Gumbel distribution can be estimated using a few methods. Two commonly used methods are the method of moments and the method of maximum likelihood estimators. The accuracy of the prediction made by the distribution is quite sensitive to the estimated parameters. A small difference in the estimation of the parameters may result in a poor prediction. Thus, the two estimation methods stated above were compared to determine the best estimate of the parameters. The two methods are described below.

Method of moments

By using the method of moments, the estimates for μ and σ are given by

$$\sigma=\frac{\sqrt{6}}{\pi}s \quad (3)$$

$$\mu=\bar{x}-\frac{n_e\sqrt{6}}{\pi}s \quad (4)$$

where $\bar{x}$ is the sample mean, s is the standard deviation for the concentration and $n_e=0.5772$ is the Euler constant.

Method of maximum likelihood estimators

By using the method of maximum likelihood estimators (MLE), the estimates for μ and σ are given by

$$\sigma=\bar{x}-\frac{\sum_{i=1}^{n}x_i\exp(-x_i/\sigma)}{\sum_{i=1}^{n}\exp(-x_i/\sigma)} \quad (5)$$

$$\mu=-\sigma\ln\left(\frac{1}{n}\sum_{i=1}^{n}\exp\left(-\frac{x_i}{\sigma}\right)\right) \quad (6)$$

Equation (5) can solved using mathematics software such as Matlab.

EXTREME DATA

The analysis of extreme value distributions requires the use of extreme data to be selected from the original observations. The threshold method was used for this purpose. The threshold method is given as follows (Madsen et al. 1984):

$$q_0 = \overline{x} + 2s \tag{7}$$

where $\overline{x}$ is the sample mean and s is the sample standard deviation. Therefore the concentrations of CO greater than the threshold value, q_0 are selected to be the extreme data. These extreme data will be used to fit the Gumbel distribution.

PERFORMANCE INDICATORS

Two performance indicators (Junninen et al. 2002) will be used to describe the goodness-of-fit for the Gumbel distribution. The first indicator is the index of agreement (IA) which is defined as

$$PA = \sum_{i=1}^{N} \frac{\left[\left(P_i - \overline{P}\right)\left(O_i - \overline{O}\right)\right]}{(N-1)\sigma_P \sigma_O} \tag{8}$$

where N is the number of imputations, O_i the observed data points, P_i the imputed data point, $\overline{P}$ is the average of imputed data, $\overline{O}$ is the average of observed data, σ_P is the standard deviation of the imputed data and σ_O is the standard deviation of the observed data.

The prediction accuracy (PA) values range from 0 to 1, with higher values of PA indicating a better fit.

The root mean square error $(RMSE)$ method is the most common indicator and is given by

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(P_i - O_i\right)^2} \tag{9}$$

For a good model, the $RMSE$ method must approach zero. Therefore, a smaller *RMSE* value means that the model is more appropriate.

RESULTS AND DISCUSSION

By using the threshold method, observations greater than 1383 ppb were chosen. Thus there are 382 observations which have values greater than 1383 ppb. The minimum value is 1390 ppb (Table 2). The concentrations are skewed to the right indicating occurrence of extreme concentrations.

Table 2. Descriptive statistics of CO data (ppb)

Number of observations	382
Mean	1778.61
Standard deviation	352.94
Minimum	1390
Maximum	3130
Skewness	1.50
Kurtosis	2.40

Table 3 presents the values of the location parameter, μ and the scale parameter, σ when the Gumbel distribution was fitted using the maximum likelihood method and method of moments were used to estimate the parameters.

Table 3. Parameter values for the Gumbel distribution

Estimation methods	σ	μ
MLE	275.186	1620
Moments	234.747	1629

From Table 3, there is a big difference between the scale parameters using the two methods while the difference is small for the location parameter. These differences are the results of estimating the missing values through different methods.

Figure 1 show the probability density functions obtained from the MLE method and the method of moments. The PDF obtained from the MLE shows a higher peak than the method of moments. The PDF from the method of moments have a better fit at the right tail of the distribution.

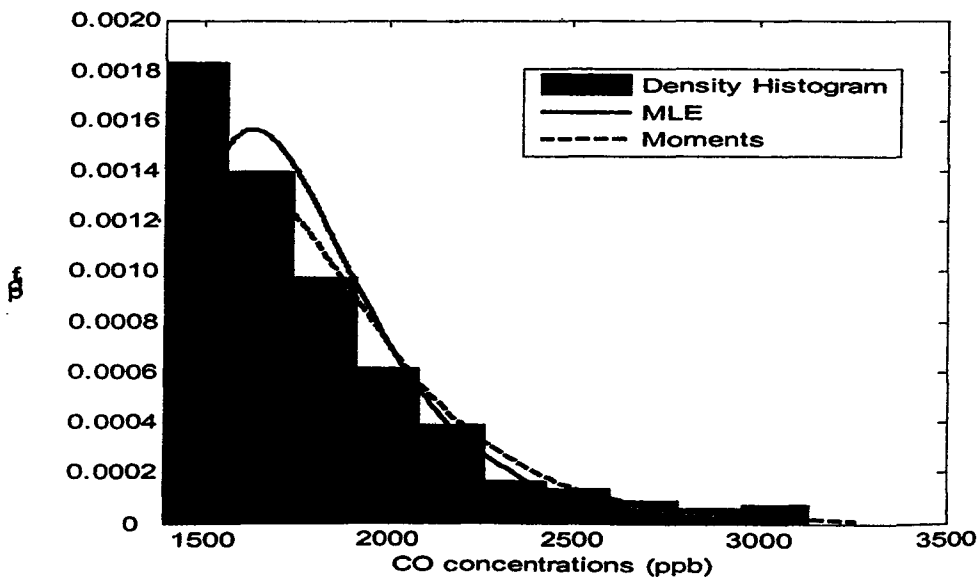


Figure 1. Comparisons between MLE and method of moments for Gumbel pdf

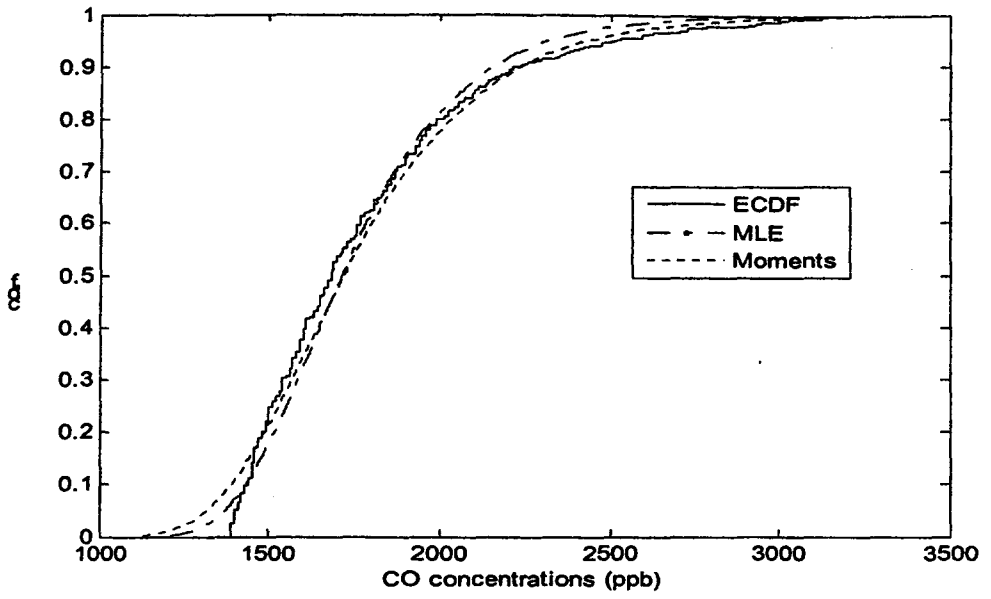


Figure 2. Comparisons between MLE and method of moments for the Gumbel cdf with the empirical cumulative distribution

Figure 2 compares the CDF of the Gumbel distributions for the MLE and the method of moments with the empirical cumulative distribution function (ECDF). Again the method of moment provides a better fit for the distribution.

Table 4 give the values of the prediction accuracy and root mean square error for the two methods. The results show that the method of moments give a better fit for the Gumbel distribution than the MLE.

Table 4. Performance indicators

Estimation methods	*IA*	*RMSE*
MLE	0.98	79.74
Moments	0.99	56.97

CONCLUSION

This paper discusses the technique to fit an extreme value distribution namely the Gumbel distribution to the right tail of the CO concentrations for the year 2002 in Seberang Perai, Pulau Pinang. The threshold method was used to obtain the extreme data which will be used to fit the Gumbel distribution. By using the threshold method, 382 observations were selected. To find the best fit between the method of moments and MLE, graphical techniques and performance indicators were used. From these techniques, it was found that the method of moments give the best fit. Thus, it can be concluded that the parameter estimates using the method of moments for the Gumbel distribution can be used to predict extreme concentrations of CO in Seberang Perai, Pulau Pinang.

REFERENCES

Department of Environment. 2003. *Malaysia Environment Quality Report 2003.* Kuala Lumpur: Department of Environment, Ministry of Sciences, Technology and the Environment, Malaysia.

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.

Kottegoda, N.T. & Rosso, R. 1998. *Statistics, probability and reliability for civil and environmental engineers.* Singapore: McGraw-Hill.

Lu, H.C. & Fang, G.C. 2003. *Predicting the exceedences of a critical PM_{10} concentration-a case study in Taiwan. Journal of Atmospheric Environment* 37: 3491-3499.

Madsen, H., Rosbjerg, D. & Harremoes, P. 1994. PDS – Modelling and regional bayesian estimation of extreme rainfalls. *Journal of Nordic Hydrology* 25: 279 - 300.

McBean, E.A. & Rovers F.A. 1998. *Statistical procedures for analysis of environmental monitoring data and risk assessment.* New Jersey: Prentice Hall.

Sharma, P., Khare, M. & Chakrabarti, S.P. 1999. Application of extreme value theory for predicting violations of air quality standards for an urban road intersection. *Transportation Research* Part D: 201-216.

Simposium
KEBANGSAAN SAINS MATEMATIK KE-15

Pusat Penerbitan Universiti (UPENA)
Universiti Teknologi MARA • SHAH ALAM • 2007

REVIEW OF FITTING DISTRIBUTIONS ON AIR POLLUTION MODELLING

Ahmad Shukri Yahaya, Nor Azam Ramli and Hazrul Abdul Hamid

ABSTRACT

Air pollution emissions degrade air quality whether in urban or rural settings. An issue of great concern has been the detrimental effect of low air quality onto human health, chronically or acutely. Understanding the behaviour of air pollution statistically would allow predictions to be made accurately. Many researches conducted on air pollution circulate within the scope of descriptive statistics while the more pressing needs are to understand the distributions that fits the collected data which can further be used for predictions of exceedences. This paper reviews several results of fitting distributions studies on air pollution modelling. The specific distribution was used to predict the mean concentration and probability of exceeding a critical concentration. The probability model may initiate a basis for estimating the parameters to meet the evolving information needs of environmental quality management.

Keywords: *Distribution Fitting, Air Pollution Modelling*

1. INTRODUCTION

Air quality is highly correlated with our everyday lives. This is because air pollution gives a big impact especially to human health. Some health impacts that are correlated with air pollutant levels are asthma, chronic bronchitis, sore throat, dry and wet cough, and hay fever (World Health Organisation, 1998). In Malaysia, Department of Environment is fully responsible in environmental management including monitoring the air quality level. There are 50 monitoring stations to monitor the air quality level belonging to the Department of Environment Malaysia. The parameters monitored include Particulate Matter (PM_{10}), Sulphur Dioxide (SO_2) and several airborne heavy metals. Besides that, other air pollutants that are normally measured are Carbon Monoxide (CO) and Nitrogen Dioxide (NO_2). Most of the air pollutants are toxic and dangerous.

Air pollutants can be divided into two categorise which are primary pollutants and secondary pollutants. Primary pollutants are those that have the same form (state and chemical composition) in the ambient atmosphere as when emitted from the sources. Secondary pollutants are those that have changed in form after leaving the source due to oxidation or decay or reaction with other primary pollutants (Stander, 2000). There are many sources of air pollution such as mobile sources, stationary sources and open burning sources (Afroz, et al., 2003). Mobile sources include personal vehicles, commercial vehicles and motorcycles. Stationary sources refer to factories, power stations, industrial fuel burning processes, and domestic fuel burning while open burning sources refer to burning of solid wastes and forest fires.

In Malaysia, air pollution index (API) is used as a standard to categorise the level of air pollution. Table 1 below shows the air pollution index for Malaysia.

Table 1: The Malaysia Air Pollution Index

Air Pollution Index	Diagnosis
0 – 50	Good
50 – 100	Moderate
101 – 200	Unhealthy
201 – 300	Very Unhealthy
301 – 500	Hazardous

Source: Department of Environment, Malaysia (1996)

2. PROBABILITY DENSITY FUNCTION

The probability density function of concentration in an atmospheric plume is an important quantity used to describe and discuss environmental diffusion (Yee and Chan, 1997). The concentrations of air pollutants are usually correlated with the emission levels and meteorological conditions. Selecting appropriate probability models for the data is an important step in environmental data analysis. These probability models may become the basis for estimating the parameters to meet the evolving information needs of environmental quality management.

The extreme value theory has also been used in air pollution study. For example, the extreme value theory can be used to fit the monthly maximum data and high concentration data of air pollutants concentration over a specific percentile (Lu, 2003). By this, the cumulative probability extremes and return period can be computed. Leong *et al.*, (2001) said that Tipett laid the theoretical foundations in 1928 when he showed that there could be only three possible types of extreme value limit distribution that are Gumbel distribution, Frechet distribution and Weibull distribution. The usual approach to distribution fitting is to fit as many distributions as possible and use goodness-of-fit tests to determine the best fit. This method, the empirical method, is subjective and is not always conclusive.

3. DISTRIBUTIONS FITTING

3.1 Estimating of Parameters

For all types of parent distributions, ± is the shape parameter that determines the form of the distribution and ² is the scale parameter that determines the skewers of the distribution. To estimate the parameters several methods can be used such as method of maximum likelihood, probability plot, method of moments and method of percentiles. Probability plot is a visual method for presentation of data in the form of graph. This method can be used to estimates how well a theoretical distribution fits a sequence of data. For the method of percentile, the value of ± lies approximately at the 63.2th percentile of the data set while for the method of maximum likelihood and method of moments, it is depend on the distributions used.

3.2 Fitting of Distributions

A distribution describes the frequency or probability of possible events. There are many distributions that can be use to fit to the air pollutants data. Georgopoulus and Seinfeld (1982) present methodologies and limitations in describing air quality through statistical distributions of pollutant concentration. This paper also explains the use of extreme statistics in the evaluation of different forms of air quality standards. The useful probability density functions in representing atmospheric concentrations include two-parameter

distributions, three-parameter distributions and four-parameter distributions.

For two-parameter distributions, the useful probability density functions are lognormal, Weibull and Gamma. For three-parameter distributions, the useful probability distribution functions are lognormal, Gamma, Weibull and Beta distribution while for four-parameter distributions, Beta distribution is very useful. Table 2 below shows the probability density functions that is useful in representing atmospheric concentrations. This paper also discuss about estimation of parameters for each distributions.

Table 2: The Probability Density Functions That Useful in Representing Atmospheric Concentrations

Distribution	Probability density function
Log-normal	$\frac{1}{x\sigma(2\pi)^{1/2}}\exp\left[-\frac{(\ln x-\mu)^2}{2\sigma^2}\right]$ $x>0\,;\sigma>0\,;-\infty<\mu<\infty$
Weibull	$\frac{\lambda}{\sigma}\left(\frac{x}{\sigma}\right)^{\lambda-1}\exp\left[-\left(\frac{x}{\sigma}\right)^{\lambda}\right]$ $x\geq 0\ ;\ \sigma,\lambda>0$
Gamma	$\frac{1}{\sigma\Gamma(\lambda)}\left(\frac{x}{\sigma}\right)^{\lambda-1}\exp\left(-\frac{x}{\sigma}\right)$ $x\geq 0\ ;\ \sigma,\lambda>0$
Three-parameter log-normal	$\frac{1}{(x-\gamma)\sigma(2\pi)^{1/2}}\exp\left[-\frac{[\ln(x-\gamma)-\mu]^2}{2\sigma^2}\right]$ $x>\gamma\,;\sigma>0\,;-\infty<\mu<\infty$
Three-parameter gamma	$\frac{1}{\sigma\Gamma(\lambda)}\left(\frac{x-\gamma}{\sigma}\right)^{\lambda-1}\exp\left(-\frac{x-\gamma}{\sigma}\right)$ $x>\gamma\,;\sigma,\lambda>0$

Contd..

Three-parameter Weibull	$\frac{\lambda}{\sigma}\left(\frac{x-\gamma}{\sigma}\right)^{\gamma-1}\exp\left[-\left(\frac{x-\gamma}{\sigma}\right)^{\lambda}\right]$ $x > \gamma\ ;\ \sigma,\lambda > 0$
Three-parameter beta	$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{1-\alpha-\beta}x^{\alpha-1}(\theta-x)^{\beta-1}$ $0 \le x \le \theta$
Four-parameter beta	$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{1-\alpha-\beta}(x-\gamma)^{\alpha-1}(\theta-x)^{\beta-1}$ $\gamma \le x \le \theta\ ;\ \alpha,\beta > 0$

Source : Georgopoulus and Seinfeld (1982)

Lu (2002) studies the statistical characters of PM_{10} concentration in Taiwan area. In this study, utilizes three distributions which are lognormal, Weibull and type V Pearson distribution to simulate PM_{10} concentration distribution in Taiwan. Three monitoring stations that he chooses for this study are L Chu, Sha-Lu and Gian-Jin where Lu takes the air quality data to compare the characters of PM_{10} concentra from 1995 to 1999. In order to estimate the parameters, Lu uses two methods which are method of mo and method of least squares. The results of this study show that the lognormal is the best distributi represent the PM_{10} daily average concentration. For the comparison of these two parametric estim methods, the method of least squares has more accurate results than the moment method.

In 2003, Lu compares the statistical characteristic of air pollutants in Taiwan by frequency distrib Similar with the previous research, Lu again selects the lognormal, Weibull and type V Pearson distribu to fit the concentration frequency distributions of particulate matter and SO_2 in Taiwan. Lu fits and com air quality data with the data that he measures. In this study, Lu also obtains the parameters of unimo using the maximum likelihood method and obtains bimodal fitted distributions using the methods of mo least squares. Besides that, Lu also uses the root mean square error (RMSE), index of agreeme Kolmogorov-Smirnov test as criteria to judge the goodness-of-fit of these three distributions. Result that the frequency distributions of PM_{10} concentration at two stations are unimodal and the distribu another one station is bimodal but for SO_2 concentration distribution, the distributions were all uni The results also show that to represent PM_{10} distribution, lognormal distribution is more appropria Weibull distributions are more suitable to represent the SO_2 distribution.

Rumburg, et al. (2001) also did a study on statistical distributions of particulate matter and th associated with sampling frequency. Rumburg resamples daily particulate matter data from Sp Washington to simulate common sampling schedules and later on computes the sampling error for reg and distribution statistics. For annual daily data, Rumburg fits probability distribution functions to de the shape of $PM_{2.5}$ and PM_8 concentration distributions. Results show that for the $PM_{2.5}$ concentrati the use of a three-parameter lognormal distribution would give the best fit whereas for PM_8 concen data the use of generalized extreme value distribution would fit best.

Lu in 2004 has done a separate study to estimate the emission source reduction of PM_{10} in central Taiwan. In this study, Lu uses three distributions to fit the complete set of PM_{10} data in central Taiwan. The distributions are lognormal, Weibull and gamma. Lu finds that to represent the performance of high PM_{10} concentration, the gamma distribution is the best one. However, the parent distribution sometimes diverges in predicting the high PM_{10} concentrations. Thus, to fit the high PM_{10} concentration distribution more accurately, Lu uses two predicting methods. Method I is known as two parameter exponential distribution and Method II is known as asymptotic distribution of extreme value. The results fitted by the two-parameter exponential distribution are better matched with the actual high PM_{10} data. Method I and Method II can successfully predict the return period exceedences over a critical concentration in the future year. By using Method I and Method II, the estimated emission source reductions of PM_{10} required meets the air quality standard very closely.

Hadley and Toumi (2003) investigate whether there has been any change in the concentration probability distribution of sulphur dioxide for over 40 years at ten monitoring sites in United Kingdom. For this study, Hadley and Toumi use the lognormal probability plot, correlation coefficient and a test for significance to fit and access how well a two-parameter lognormal distribution describe the data. The study finds that for daily data, the lognormal is good and robust to fit a variety of conditions.

4. CONCLUSION

Air pollution modelling is important because the measurement of air pollution is done at certain places only or in another word, we can't measure air pollution in every place where it occurs. So, models are used to simulate the dispersion of air pollutants away from emission sources, and to estimate ground level pollution concentrations. Since fitting distribution is one of the most important steps in air pollution modelling, there are many studies done related to this area. From several studies that has been discuss in this paper, the lognormal distribution, gamma distribution and Weibull distribution are widely used in fitting distribution on air pollutants data. However, the use of type V Pearson distribution is also suitable to fit the distribution for PM_{10} and SO_2. Generally, the best distributions to fit air pollutant data is depend on the studies area and it is not unique. In Malaysia, there are only few studies related to air pollution especially in fitting of distributions on air pollutants data. Therefore, many studies can be conducted in Malaysia to develop models by fitting distributions on the air pollution concentrations and propose the strategies to improve the air quality management.

REFERENCES

Afroz, R., Hassan, M.N and Ibrahim, N.A. (2003) Review of Air pollution and Health Impacts in Malaysia. Environment Research, 92(2), p. 71 – 77.

Department of Environment (1996). Annual Report.

Georgopoulus, P.G. and Seinfeld, J.H. (1982) Statistical Distributions of Air Quality Concentrations. Environmental Science and Technology, 16, p. 401A – 416A

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, p 1461 – 1474

Leong, Y., Sleigh, P. and Torrance, J.M. (2001) Extreme Value Theory Applied to Posporative Breathing Patterns, British Journal of Anaesthesia, 88, p. 61 – 64

Lu, H.C. (2002) The Statistical characters of PM_{10} concentration in Taiwan area. Atmospheric Environment, 36, p. 491 – 502

Lu, H.C. (2003) Comparison of statistical characteristic of air pollutants in Taiwan by frequency distribution. Journal of the Air & Waste Management Association, 53(5), p 608 – 616

Lu, H.C. (2004) Estimating the Emission Source Reduction of PM_{10} in Central Taiwan. Journal of Chemosphere, 54(7), p. 805 – 814.

Rumburgh, B., Richard, A. and Claiborn, C. (2001) Statistical Distributions of particulate matter and the error associated with sampling frequency. Atmospheric Environment, 35, p 2907 – 2920

Stander, L.H., (2000) Regulatory Aspects of Air Pollution Control in the United States. Air & Waste Management Association, p. 8 – 21.

World Health Organization (1998) Report of the Bioregional Workshop on Health Impacts of Haze Related Air Pollution. Manila: WHO.

Yee, E. and Chan, R. (1997) A Simple Model for the Probability Density Function of Concentration Fluctuations in Atmospheric Plumes. Atmospheric Environment, 31(7), p. 991 – 1002.

---

[1]Ahmad Shukri Yahaya, [2]Nor Azam Ramli, [3]Hazrul Abdul Hamid
[1, 2] School of Civil Engineering
Universiti Sains Malaysia Engineering Campus
14300, Nibong Tebal
PULAU PINANG
[1]shukri@eng.usm.my [2]ceazam@eng.usm.my
[3]Department of Mathematics and Informatics Sciences
Penang Matriculation College
13200, Kepala Batas
PULAU PINANG
[3]hazrul@kmpp.matrik.edu.my

Simposium
KEBANGSAAN SAINS MATEMATIK KE-15

Pusat Penerbitan Universiti (UPENA)
Universiti Teknologi MARA • SHAH ALAM • 2007

UJIAN KEBAGUSAN PENYUAIAN BAGI TABURAN GUMBEL

Ahmad Shukri Yahaya, Norlida Mohd Noor,
Chan Yin Yin dan Lee Mun Yee

ABSTRAK

Taburan Gumbel banyak digunakan untuk memodelkan peristiwa ekstrim seperti paras maksimum atau kadar aliran sungai, kadar kelajuan angin serta analisis terhadap pencemaran udara. Ujian kebagusan penyuaian biasanya digunakan untuk menentukan samada sesuatu taburan itu boleh mewakili peristiwa-peristiwa ekstrim itu. Terdapat beberapa ujian kebagusan penyuaian yang boleh digunakan. Oleh itu lima jenis ujian statistik berasaskan fungsi taburan empirik untuk menjalankan ujian kebagusan penyuaian bagi taburan Gumbel dibincangkan. Lima statistik ujian yang dikaji ialah statistik-statistik ujian Anderson-Darling (A^2), Cramer-Von-Mises (W^2), Kolmogorov-Smirnov (D), Kuiper (V) dan Watson (U^2). Kuasa bagi setiap statistik ujian ini teleh diperolehi menggunakan kaedah simulasi Monte Carlo dengan menjanakan 100000 sampel rawak masing-masingnya bersaiz 5, 10, 20, 30, 40, 50, 60, 100, 200 dan 300. Kajian ini mengandaikan bahawa suatu sampel rawak yang bertabur secara taburan Gumbel telah dikutip apabila semua parameter tidak diketahui nilainya. Tiga taburan alternatif iaitu taburan eksponen, taburan log-normal dan taburan Pareto telah digunakan. Apabila taburan eksponen digunakan sebagai taburan alternatif, statistik A^2 memberikan kuasa ujian yang terbaik tetapi apabila taburan log-normal dan taburan Pareto digunakan, statistik D memberikan kuasa ujian yang terbaik.

Kata Kunci: *Taburan Gumbel, Ujian Kebagusan Penyuaian, Kuasa Ujian*

1. PENGENALAN

Ujian kebagusan penyuaian merupakan ujian yang digunakan untuk menguji samada suatu sampel rawak sepadan dengan fungsi taburan yang telah ditetapkan (D' Agostino *et al.*, 1986). Ujian tertua dan paling banyak digunakan ialah ujian khi-kuasadua (Davis dan Stephens, 1989). Walau bagaimanapun ujian khi-kuasadua memerlukan sampel yang agak besar dan tidak menggunakan sepenuhnya maklumat daripada sampel. Oleh itu lima jenis statistik ujian berasaskan fungsi taburan empirik telah digunakan. Statistik ujian ini telah dibincangkan dengan mendalam oleh Davis dan Stephens (1989).

Kajian ini tertumpu kepada taburan Gumbel iaitu salah satu daripada bentuk taburan nilai ekstrim. Taburan nilai ekstrim sangat penting dan sering digunakan dalam kawalan mutu dan dalam bidang sains persekitaran seperti memodelkan banjir, hujan, kelajuan angin dan perubahan cuaca (Kattegoda dan Rosso, 1998).

Kertas kerja ini membincangkan serta membandingkan kuasa bagi lima statistik ujian yang lazim digunakan untuk menguji kebagusan penyuaian taburan Gumbel. Untuk mendapatkan nilai kuasa ujian, tiga taburan alternatif telah digunakan iaitu taburan eksponen, taburan log-normal dan taburan Pareto.

2. STATISTIK UJIAN

Secara amnya, ujian kebagusan penyuaian menguji hipotesis berikut:

H_0 : Suatu sampel rawak bersaiz n dikutip daripada taburan $F(x;\theta)$

H_1 : Suatu sampel rawak bersaiz n dikutip daripada taburan lain

dengan $F(x;\theta)$ ialah fungsi taburan longgokan dan θ ialah vektor parameter taburan terse... Pengujian terhadap hipotesis nol dan alternatif seperti dalam persamaan (1) dinamakan seba... kebagusan penyuaian. Lima statistik ujian kebagusan penyuaian yang boleh digunakan diberik... Jadual 1 di bawah.

Jadual 1: Statistik Ujian Kebagusan Penyuaian

Statistik ujian	Rumus
Anderson-Darling	$A_n^2 = -\dfrac{\sum_{i=1}^{n}(2i-1)\{\log z_i + \log(1-z_{n+1-i})\}}{n} - n$
Cramer-Von Mises	$W_n^2 = \sum_{i=1}^{n}\left(z_i - \left(\dfrac{2i-1}{2n}\right)\right)^2 + \dfrac{1}{12n}$
Kolmogorov-Smirnov	$D_n = \underset{1\le i\le n}{maks}\left(D_n^+, D_n^-\right)$ $D_n^+ = \underset{1\le i\le n}{maks}\left(\dfrac{1}{n} - z_i\right)$ $D_n^- = \underset{1\le i\le n}{maks}\left(z_i - \dfrac{(i-1)}{n}\right)$
Kuiper	$V_n = D_n^+ + D_n^-$
Watson	$U_n^2 = W_n^2 - n(\bar{z} - 0.5)^2$ dengan $\bar{z} = \sum_{i=1}^{n}\dfrac{z_i}{n}$

3. ANALISIS

Dalam kajian ini, fungsi taburan longgokan yang digunakan ialah taburan Gumbel (Evans *et al.*, 2000). Fungsi taburan longgokan bagi taburan Gumbel ialah

$$F(x;\mu,\sigma) = 1 - \exp\left[-\exp\left(\frac{x-\mu}{\sigma}\right)\right] \tag{2}$$

dengan penganggar bagi parameter lokasi μ dan skala σ masing-masingnya dianggarkan oleh

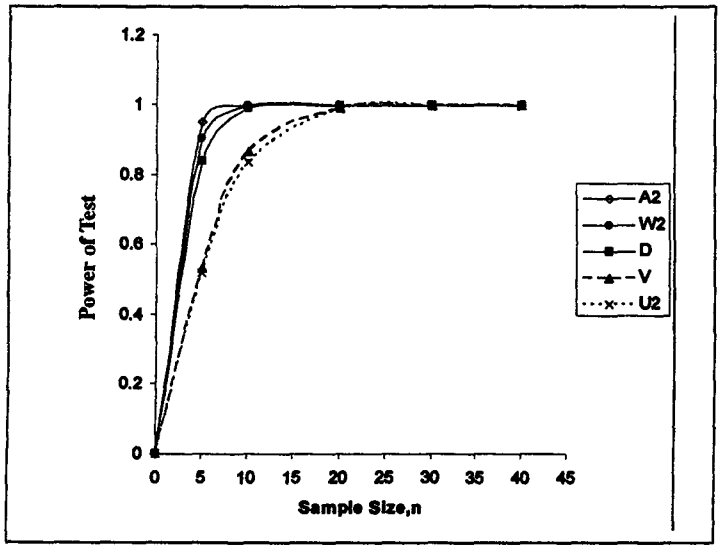
$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \text{ dan } \hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{3}$$

Kuasa bagi suatu ujian, ditandakan dengan $1-\beta$, ialah keupayaan ujian tersebut untuk menghasilkan ujian yang betul manakala β ialah kebarangkalian berlakunya ralat jenis kedua. Kuasa bagi setiap statistik ujian diperolehi dengan membandingkan kecekapan statistik ujian berkenaan.

Kuasa bagi setiap statistik ujian ini teleh diperolehi menggunakan kaedah simulasi Monte Carlo dengan menjanakan 100000 sampel rawak masing-masingnya bersaiz 5, 10, 20, 30, 40, 50, 60, 100, 200 dan 300. Untuk menjalankan simulasi ini, kaedah pendarab kongruential (Ahmad Shukri Yahaya, 2002) telah digunakan. Kuasa bagi kelima-lima statistik ujian ini telah dibandingkan dengan menggunakan tiga taburan alternatif iaitu taburan eksponen, taburan log-normal dan taburan Pareto.

4. KEPUTUSAN DAN PERBINCANGAN

Rajah 1 menunjukkan perbandingan kuasa bagi setiap statistik ujian apabila taburan alternatif ialah taburan eksponen dengan min 5 pada paras keertian lima peratus.

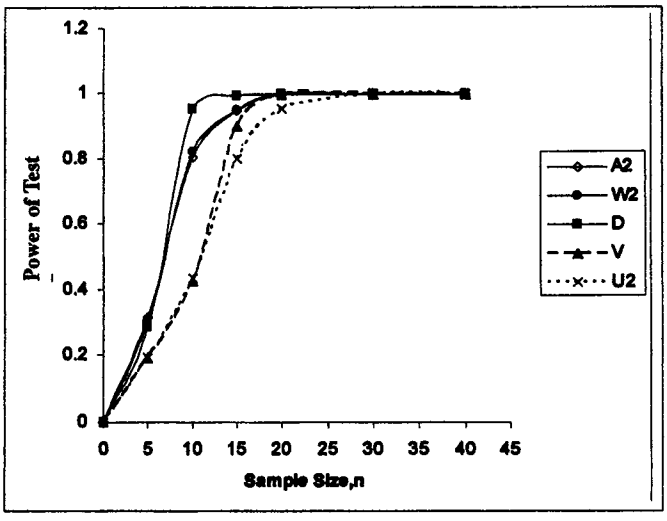


Rajah 1 : Perbandingan Kuasa Ujian Statistik apabila Taburan Alternative adalah Eksponen

Daripada Rajah 1, didapati bahawa statistik ujian Anderson-Darling $\left(A_n^2\right)$ adalah paling sensitif dan ikuti oleh statistik ujian Cramer-Von Mises $\left(W_n^2\right)$, Kolmogorov-Smirnov $\left(D_n\right)$, Kuiper $\left(V_n\right)$ dan Watson $\left(U_n^2\right)$. Kuasa bagi ujian Anderson-Darling, Cramer-Von Mises dan Kolmogorov-Smirnov menghampiri satu abila $n=10$ manakala statistik ujian Kuiper dan Watson menghampiri satu apabila $n=20$.

Rajah 2 menunjukkan perbandingan kuasa bagi setiap statistik ujian apabila taburan alternatif ialah uran log-normal dengan parameter lokasi bernilai sifar dan parameter skala bernilai satu pada paras rtian lima peratus.

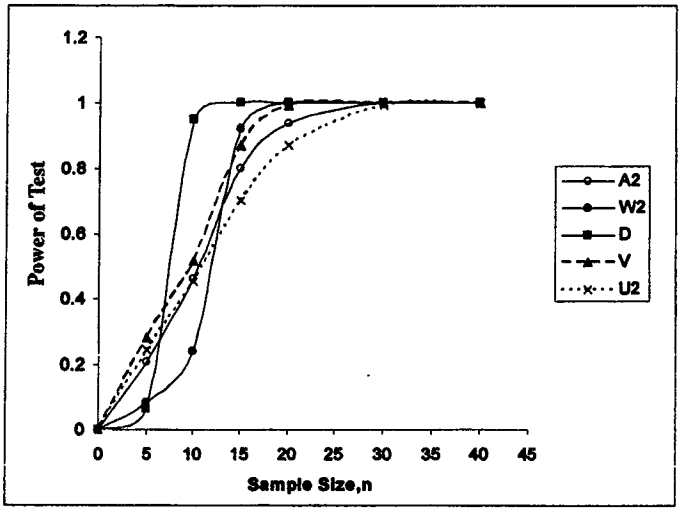
Daripada Rajah 2, didapati bahawa statistik ujian Kolmogorov-Smirnov mempunyai kuasa paling gi dengan statistic ujian Watson mempunyai kuasa terendah. Kuasa ujian bagi statistik ujian Anderson-ling dan Cramer-Von Mises adalah agak sama.



Rajah 2: Perbandingan Kuasa Ujian Statistik apabila Taburan Alternatif adalah Lognormal

Rajah 3 menunjukkan perbandingan kuasa bagi setiap statistik ujian apabila taburan alternatif ialah taburan Pareto dengan parameter lokasi bernilai 0.1 dan parameter skala bernilai 0.5 pada paras keertian lima peratus.

Daripada Rajah 3, didapati bahawa statistik ujian Kolmogorov-Smirnov masih lagi mempunyai kuasa paling tinggi dengan statistik ujian Watson mempunyai kuasa terendah.



Rajah 3 : Perbandingan Kuasa Ujian Statistik Apabila Taburan Alternatif Adalah Pareto

Secara keseluruhannya, kuasa bagi sesuatu ujian itu akan meningkat apabila saiz sampel meningkat. Ini boleh diperhatikan daripada hasil keputusan yang dipaparkan dalam Rajah 1 hingga Rajah 3.

5. KESIMPULAN

Kertas kerja ini membandingkan serta membincangkan kuasa ujian bagi lima statistik ujian yang boleh digunakan untuk menguji kebagusan penyuaian. Lima statistic ujian yang digunakan adalah statistic ujian Anderson-Darling $\left(A_n^2\right)$, Cramer-Von Mises $\left(W_n^2\right)$, Kolmogorov-Smirnov $\left(D_n\right)$, Kuiper $\left(V_n\right)$ dan Watson $\left(U_n^2\right)$. Kuasa bagi ujian telah diperolehi menggunakan simulasi Monte Carlo dengan mengambil sebanyak 100000 sampel bersaiz 5, 10, 20, 30, 40, 50, 60, 100, 200 dan 300. Tiga taburan alternatif telah digunakan iaitu taburan eksponen, taburan log-normal dan taburan Pareto.

Hasil kajian mendapati bahawa statistik ujian Kolmogorov-Smirnov paling berkuasa apabila taburan alternatif adalah adalah log-normal dan Pareto nmanakala statistik ujian Anderson-Darling paling berkuasa apabila taburan alternatif ialah eksponen.

RUJUKAN

Ahmad Shukri Yahaya (2002), Tinjauan Terhadap Beberapa Penjana Nombor Rawak. *Prosiding Simposium Kebangsaan Sains Matematik Ke 9*, 33-37.

D'Agostino, R.B. and Stephens, M.A. (1986), *Goodness-of-fit Techniques*, Vol. 68, New York: Dekker.

Davis, C.S. and Stephens, M.A. (1989), *Empirical Distribution Function Goodness-of-fit Tests, Appl. Statist.* 38, No. 3, pp. 535-582.

Evans, M., Hastings, N., Peacock, B. (2000), *Statistical Distribution*, Third Edition, New York: John Wiley & Sons, Inc.

Kottegoda, N.T. and Rosso, R. (1998) *Statistics, Probability and Reliability for Civil and Environmental Engineers,* Singapore: McGraw-Hill Book Co.

Ahmad Shukri Yahaya[1], Norlida Mohd Noor[2], Chan Yin Yin[2] dan Lee Mun Yee[2]
[1]Pusat Pengajian Kejuruteraan Awam, Universiti Sains Malaysia,
Kampus Kejuruteraan, 14300 Nibong Tebal,
Seberang Perai Selatan, PULAU PINANG.
shukri@eng.usm.my
[2] Pusat Pengajian Sains Matematik,
Universiti Sains Malaysia, 11800 Minden,
PULAU PINANG.

STATEMENT OF ACCOUNT

PROF MADYA AHMAD SHUKRI YAHAYA

304.PAWAM.6035161

JUMLAH GERAN :- 15,480.00

NO PROJEK :-

PANEL :- J/PENDEK

PENAJA :- JANGKA PENDEK

UNIT KUMPULAN WANG AMANAH
UNIVERSITI SAINS MALAYSIA
KAMPUS KEJURUTERAAN
SERI AMPANGAN

PENYATA KUMPULAN WANG

TEMPOH BERAKHIR 31/01/2008

Tempoh Projek:01/10/2005 - 30/09/2007

MODELLING OF CARBON MONOXIDE COCENTRATION IN MAJOR TOWN IN MALAYSIA

Vot		*Peruntukan* (a)	*Perbelanjaan sehingga 31/12/2007* (b)	*Tanggungan semasa 2008* (c)	*Perbelanjaan Semasa 2008* (d)	*Jumlah Perbelanjaan 2008* (c + d)	*Jumlah Perbelanjaan Terkumpul* (b+c+d)	*Baki Peruntukan Semasa 2008* (a-(b+c+d)
11000	GAJI KAKITANGAN AWAM	3,340.00	2,853.40	0.00	0.00	0.00	2,853.40	486.60
21000	PERBELANJAAN PERJALANAN DAN SARAHI	5,000.00	1,795.00	0.00	0.00	0.00	1,795.00	3,205.00
23000	PERHUBUNGAN DAN UTILITI	540.00	0.00	0.00	0.00	0.00	0.00	540.00
26000	BAHAN MENTAH & BAHAN UNTUK PENYELE	0.00	270.00	0.00	0.00	0.00	270.00	(270.00)
27000	BEKALAN DAN ALAT PAKAI HABIS	1,300.00	8,066.53	0.00	0.00	0.00	8,066.53	(6,766.53)
28000	PENYELENGGARAAN & PEMBAIKAN KECIL	300.00	0.00	0.00	0.00	0.00	0.00	300.00
29000	PERKHIDMATAN IKTISAS & HOSPITALITI	5,000.00	2,004.57	0.00	0.00	0.00	2,004.57	2,995.43
		15,480.00	14,989.50	0.00	0.00	0.00	14,989.50	490.50
	Jumlah Besar	15,480.00	14,989.50	0.00	0.00	0.00	14,989.50	490.50