UNIVERSITI TEKNOLOGI MARA

PREDICTION OF GROUND LEVEL OZONE CONCENTRATIONS USING WEIBULL AND GENERALIZED EXTREME VALUE (GEV) DISTRIBUTIONS

SHAFIQ NAIM BIN SHAHRUDIN

MSc

January 2019

UNIVERSITI TEKNOLOGI MARA

PREDICTION OF GROUND LEVEL OZONE CONCENTRATIONS USING WEIBULL AND GENERALIZED EXTREME VALUE (GEV) DISTRIBUTIONS

SHAFIQ NAIM BIN SHAHRUDIN

Dissertation submitted in partial fulfilment
of the requirements for the degree of
Master of Science Applied Statistics

Faculty of Computer and Mathematical Sciences

January 2019

APPROVED BY:

(DR HASFAZILAH BINTI AHMAT)

Supervisor

Faculty of Computer and Mathematical Sciences

CONFIRMATION BY PANEL OF EXAMINERS

I certify that a Panel of Examiners has met on 8th January 2019 to conduct the final examination of Shafiq Naim Bin Shahrudin in his Master of Science in Applied Statistics thesis entitled "Prediction Of Ground Level Ozone Concentrations Using Two And Three Parameter Weibull And Generalized Extreme Value (GEV)" in accordance with Universiti Teknologi MARA Act 1976 (Akta 173). The Panel of Examiner recommends that the student be awarded the relevant degree. The Panel of Examiners was as follows:

Sayang Binti Mohd Deni, PhD
Associate Professor
Faculty of Computer and Mathematical Sciences
Universiti Teknologi MARA
(Head of Internal Examiner)

Nik Arni Binti Nik Mohamad, Msc
Senior Lecturer
Faculty of Computer and Mathematical Sciences
Universiti Teknologi MARA
(Internal Examiner)

PROF TS DR HARYANI HARON
Dean Faculty of Computer and Mathematical Sciences
Date: 24 January 2019

AUTHOR'S DECLARATION

I declare that the work in this dissertation was carried out in accordance with the regulations of Universiti Teknologi MARA. It is original and is the results of my own work, unless otherwise indicated or acknowledged as referenced work. This thesis has not been submitted to any other academic institution or non-academic institution for any degree or qualification.

I, hereby, acknowledge that I have been supplied with the Academic Rules and Regulations for Post Graduate, Universiti Teknologi MARA, regulating the conduct of my study and research.

Name of Student	:	Shafiq Naim Bin Shahrudin
Student I.D. No.	:	2016940339
Programme	:	Master of Science In Applied Statistics – CS702
Faculty	:	Computer And Mathematical Sciences
Dissertation Title	:	Prediction Of Ground Level Ozone Concentrations Using Two And Three Parameter Weibull And Generalized Extreme Value (GEV)
Signature of Student	:	
Date	:	January 2019

ABSTRACT

Malaysia is one of the Asian counties which has witnessed rapid increase industrialization, modern service and modernization over the past few decades (Bekhet & Othman, 2017). This study was conducted with the objectives to examine the characteristics of ozone concentrations and its trends in Peninsular Malaysia, estimating the parameter to identify the best distribution between the two and three parameter Weibull and GEV for future prediction of ground level ozone exceedances. There is a few on the application of two and three- parameter Weibull and GEV in the study of ground level ozone concentrations in Malaysia. Monitoring records from only three stations and one background station in Peninsular Malaysia were selected for the period of 1st January 2007 to 31st December 2016. The best distribution with the smallest error measure and highest accuracy measure for Klang, Putrajaya and Shah Alam was found to be the two parameter Weibull. Hence, the two parameter Weibull distribution was the best distribution to assess high level of ground level ozone concentrations for the decision and policy maker to enhance current policy and implement effective policy to create healthy environment in these particular locations.

ACKNOWLEDGEMENT

IN THE NAME OF ALLAH, THE MOST GRACIOUS, THE MOST MERCIFUL

Firstly, thanks to Allah S.W.T for enabling and gave strength to me to complete this research thesis on time. Big thanks also to my respectful supervisor, Dr. Hasfazilah Binti Ahmat who had continuously given useful idea and monitoring my thesis progress. I am very grateful for having her as my supervisor.

I would also like to express my appreciation to my beloved mother, Puan Siti Hanim Binti Simarani, my beloved father, En Shahrudin bin Hassan and to my family for the motivation, moral and financial support in my journey to complete my thesis. In addition, thank you to all my friends and classmate who gave me helped and share knowledge during completing this thesis.

Lastly, I also would like to thanks to both panels Professor Madya Dr Sayang Binti Mohd Deni and Puan Nik Arni Binti Nik Mohamad for giving ideas during progress presentation to improve statistical analysis for my thesis.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE
INTRODUCTION

1.1 Background of Study

Malaysia is one of the Asian countries which has witnessed rapid increase in industrialization, modern service and modernization over the past few decades (Bekhet & Othman, 2017). The improvement and extension of urban regions has been the essential driver of the disintegration of air quality in urban (Sharma et al., 2014). Southeast Asia region also known as most heavily populated region in the world and has a vibrant mixture of cultures. Apart from that, these region become the major contributes to the global pollution in the world (Jasim et al., 2011).

Exhaust gases from vehicles, combustion in industries and from domestic purposes are the major sources for air pollution in urban areas. In fact, nitrogen oxides, carbon monoxide, Sulphur dioxide and ozone are the common pollutant contribute to the air pollution in urban areas. With the increasing number of industrializations, many developed countries, particularly large cities in Malaysia experiencing excessive level of these pollutions. Consequently, controlling air pollution has become the biggest challenge to many countries in order to comply with the environmental standards as well as to assess the effectiveness of control policies. Especially for ground level ozone, effective control policies are required to give insight to the part of physiochemical form in the troposphere (Banan, Latif, & Juneng, 2013).

In general, ozone is a good substance that prevent human from various serious health impact caused from the sun light. However, most people do not realize that ozone can cause one of the respiratory health problems and other external health effect such as asthma, caught, congestion, skin cancer and others. There are two layers of ozone, the first layer at ground level known as tropospheric ozone, while the second layer at the outer space known as stratospheric ozone. The ground level ozone found in troposphere is a harmful pollutant. In this level, ozone brings negative effects to human health, vegetation, and other

living organisms. The ozone alters molecules in the air and gradually destroys these entities (Shan, Yin, Zhang, & Ding, 2008). Inversely, stratospheric ozone is considered good for humans and other living forms because it protects the biosphere from harmful ultraviolet radiation (Bracher et al., 2005).

Approximately 90% of atmospheric ozone is contained in the "ozone layer". This layer protects us from the harmful ultraviolet radiation resulting from the sunlight. Stratospheric ozone absorbs all the UV–C, and a large amount of UV–B (Bian, Gettelman, Chen, & Pan, 2007). The ozone can be considered as a form of beneficial ultraviolet protection in the stratosphere; however, it remains harmful to human beings at the ground level (Lu & Wang, 2006).

Monitoring data and studies on ambient air quality show that some of the air pollutants in several large cities are increasing with time and are not always at acceptable levels according to the Malaysia Ambient Air Quality Guidelines (MAAQG). Most of the air prediction on ozone concentration using extreme value distribution were applied widely in foreign countries and Malaysia PM_{10} were common data to be used in extreme value distribution. In this regard, this study has two objectives in mind, to evaluate the performances of the classical approaches in estimating the parameters of the two and three parameter Weibull and Generalized Extreme Value (GEV) and secondly, to attain the best model to predict ozone concentrations level in industrial monitoring stations which are located in Peninsular Malaysia.

1.2 Extreme Value Distribution

A theory developed to address questions relating to the distribution of extremes is the Extreme Value Theory (EVT) (Finkenstadt & Rootzen, 2003). It develops techniques and models for describing the unusual (extremes) rather than the usual phenomenon (Samuel & Saralees, 2000). Where air pollution control is concerned, the rare event is typically more significant than the common event.

According to Finkenstadt and Rootzen (2003), Extreme Value Theory was used to solve problem relating to the distribution of extreme. In 1927, Frechet, while 1928, Fisher

and Tippet, were the pioneer in introducing an extreme value distribution works. During period 1920s until mid of 1930s, many theoretical developments was done, however, around 1940s, the first publication on extreme value distribution in environmental study was successfully done.

Weibull, a Swedish physicist, was used distribution admitted by Frichet to the analysis of material strength. In early 1951, the distribution was well known as a tool for modelling the statistical variations of the data until its was named as Weibull distribution after his name (Burry, 1999).

1.3 Problem Statement

Air pollution commonly occurs in any region in this world, especially in Malaysia. According to Department of Environment (DOE) Malaysia (2016), due to trans-boundary forest smoke throughout the dry period of June to September during the Southwest Monsoon, also increasing number of vehicles, expansion of industrial area and urbanization has been identified as a major pollution in Malaysia had occasionally exceeded the Malaysia Ambient Air Quality Guidelines (MAAQG) of 120 $\mu g/ m^3$for 8-hour record of ground level ozone.

Ozone at ground level become the main concern for the researcher because it has the capability to oxidize other gases present in the atmosphere after the process of photolysis because of its very nature (Duenas, Fernandez, S.Canete, Carretero, & Liger, 2004). The extreme ozone concentration means which where the records exceeded the air quality guideline. Moreover, when excessive vulnerable to highly ozone concentration can damage human health, vegetation, and other living organisms. Ozone alters molecules and gradually destroys these entities. Therefore, constant monitoring, analysis and enhanced model to assess the ground level ozone concentrations are vital to implement effective policies to create the cleaner environment which will eventually reduce the cost to maintain the public health in the long term.

Previous researches in Malaysia mostly concentrates on PM_{10}, which many statistical models have been proposed. There is a few on the application on two and three

parameter Weibull and Generalized Extreme Value (GEV) distribution used to predict the exceedances of high ozone concentration at ground level in Malaysia. According to Gwak, Goo, Choi, and Ahn (2016) Weibull distributions are widely used in environment study. Besides that, Generalized Extreme Value (GEV) was commonly used for describing the extreme value for environment data (Gwak et al., 2016). Also, the EVT or known as Generalized Extreme Value distribution, are widely used in finance, risk management, material sciences, economics, insurance, hydrology, telecommunications, and many other industries dealing with extreme events. Studies involving natural phenomena using EVT such as rainfall, the height of sea waves, floods, corrosion, and wind speed have been of great interest to scientist and researchers for a long period of time (Ozay & Celiktas, 2016).

Therefore, this study aims to determine the best distribution using two and three parameter Weibull and General Extreme Value (GEV) distribution to determine the suitable model to predict high concentrations of ozone at ground level. Meanwhile, the availability of appropriate statistical model in predicting future exceedances of awareness avoid the MAAQG restrict could beneficial for the environmentalist and strategists to devise proper movement and controlling strategies to triumph over the trouble

1.4 Research Questions

The research questions of this study involve of the following:

1. What are the characteristics and trend of ozone concentration at Peninsular Malaysia?
2. What are the estimated parameter of Weibull and GEV distribution?
3. What is the best distribution between two and three parameter of Weibull and GEV distribution?

1.5 Research Objectives

The objectives of this study involve of the following:

1. To examine the characteristics and trend of ozone concentration at Peninsular Malaysia .
2. To estimate the parameter of Weibull and GEV distribution.
3. To identify the best distribution between two and three parameter of Weibull and GEV distribution.

1.6 Scope of Study

This study focusses on one of the dominant air pollutions in Malaysia which is ozone (O^3) concentration at ground level. The average daily maximum concentrations of ozone were monitored for the study purposes.

The monitoring records in this study from the period of 2007 to 2016 from three (3) monitoring stations and one (1) background station: Putrajaya, Klang, Shah Alam and Jerantut furnished by the Department of Environment (DOE) were use. The locations were selected because of data availability and also based on the location characteristics which are urban and rural area. The preliminary analysis of descriptive statistics for monitoring records were conducted to confirm of the claim on the application Weibull and GEV distribution.

1.7 Significance of Study

Ozone are very important to many living things, especially human which are effect to individual productivity, government and public sector. Thus, the concentration of ground level ozone should be monitored in order to maintain the cleanliness of air quality for healthy life. This research will help to identify which distribution is the best to predict the high concentration of ground level ozone in Peninsular Malaysia especially in Jerantut, Klang, Ptrajaya and Shah Alam.

This study will give the significant information to Department of Environment

Malaysia about the sustainable of air quality in order to maintain the level of concentration ozone in below hazardous level that had been set by Malaysia Ambience Air Quality (MAAQ) index for ozone which is 120 μg/ $m^3$for 8-hour monitor records. A better prevention can be introducing to protect Malaysian citizen as well as Malaysian economy will burst in the next future years.

Though a number of researches on ozone (O3) concentrations has been carried out in Malaysia and many statistical models have been proposed, the major gap in this study is in the application of statistical modelling using two and three parameter Weibull and General Extreme Value (GEV) distribution. This study is the first to model ozone (O3) concentrations with the application of two and three parameter Weibull and General Extreme Value (GEV) approach. The enhance model can be used to assess high level of ozone (O3) concentrations which are vital to implement effective policies to create the cleaner environment.

CHAPTER TWO
LITERATURE REVIEW

2.1 Introduction

The section will provide insights of the researches that had been conducted to gain more knowledge and deeper understanding towards the research subject that is under investigation.

2.2 Air Pollution in Malaysia

Over the year there were 19 polluted days amid which the air contamination levels were between the Lower Moderate to Hazardous classifications (Othman, Sahani, Mahmud, & Ahmad, 2014). The average annual economic loss due to the inpatient health impact of polluted air was valued at MYR273,000 ($91,000 USD) (Othman et al., 2014). Research conducted by Azam, Mahmudul Alam, and Haroon Hafeez (2018) reveals that environmental pollution in Malaysia has a significant positive effect on tourism. Sustainable economic growth and development should be ensured by implementing prudent public policy where host governments must strive to promote socially and environmentally responsible tourism industries in their respective countries. Figure 2.1 below shows the cycle of sources, transport, transformation, and fate of atmospheric pollutants.

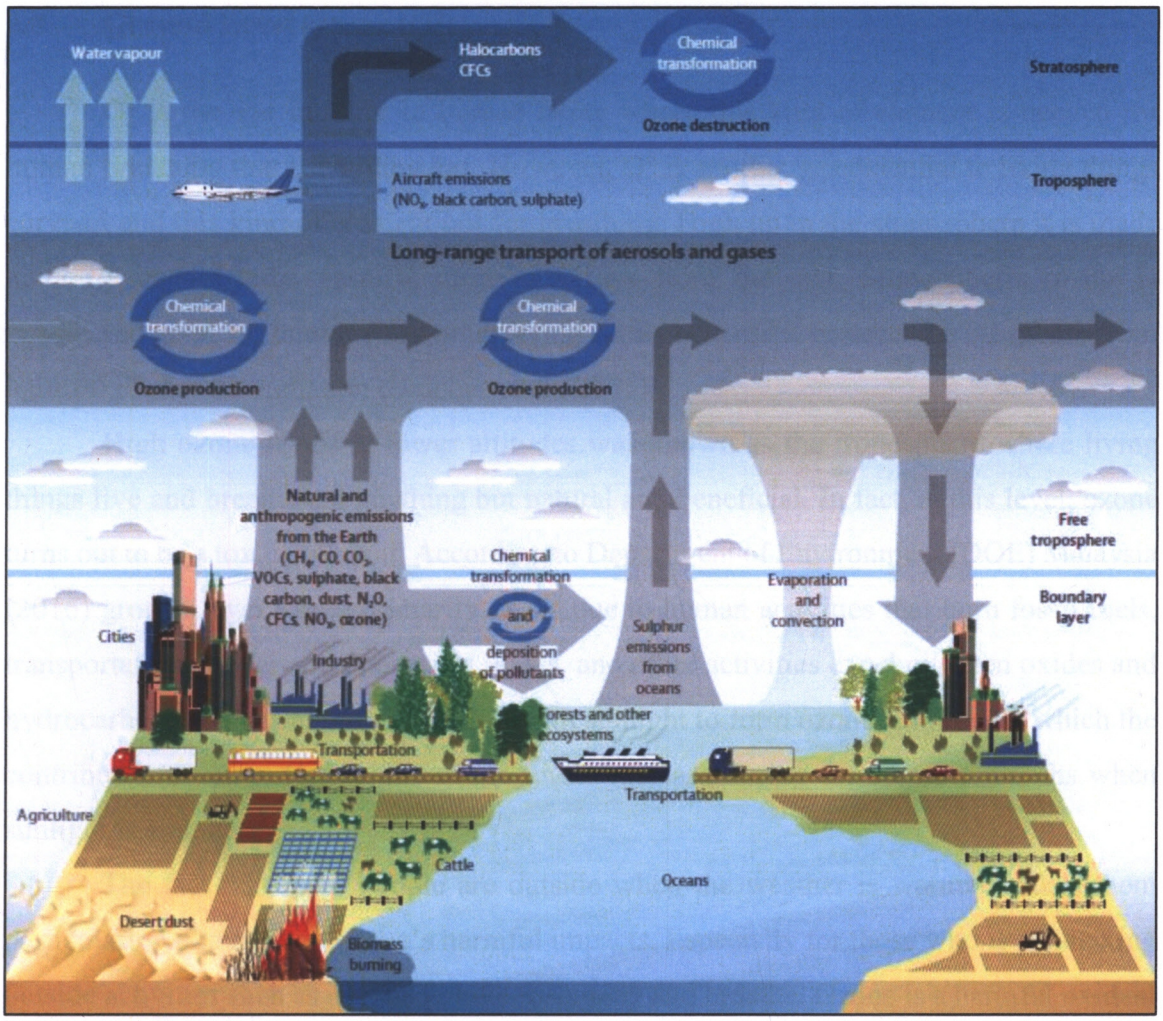


Figure 2.1 Sources, transport, transformation, and fate of atmospheric pollutants. (Source: Guarnieri and Balmes (2014))

The air pollution comes mainly from land transportation, industrial emissions, and open burning sources. Among them, land transportation contributes the most to air pollution. According to Afroz, Hassan, and Ibrahim (2003) the ambient air quality monitoring and studies related to air pollution and health impacts. This problem has contributed to the increasing of hospital visits for treatments related to chronic obstructive pulmonary diseases, upper respiratory infections, asthma and rhinitis. Besides that, respiratory mortality increased 19% due to air pollution. Almost all economic sectors also experienced losses, with the heaviest losses in the agriculture and tourism sectors (Latif et al., 2018).

2.3 Ground Level Ozone (O^3)

O^2 is the gas familiar to human being where this kind of element is needed for human breathing that life giving gas. However, O^3 is another gas essential to living things survival and this kind of element not for breathing. High up in the stratosphere it is made naturally and absorbs harmful ultraviolet rays from the sun. Stratospheric ozone is considered good for humans and other living forms because it protects the biosphere from harmful ultraviolet radiation (Bracher et al., 2005).

High ozone levels at lower altitudes was known as the troposphere where living things live and breathe are anything but natural and beneficial. In fact, at this level, ozone turns out to be a toxic pollutant. According to Department of Environment (DOE) Malaysia (2016) ground level ozone primarily exists due to human activities that burn fossil fuels, transportation, power and industrials plants, and other activities expel nitrogen oxides and hydrocarbons. This compound interacts with sunlight to form ozone compound, which the contributor to smog. Hence, the ozone levels increase during the summer months when sunlight is abundant.

The fact that more people are outside when the weather is warmer makes them particularly vulnerable to ozone's harmful impacts. Especially for those who is conducting outside activities such as run, bike, hike, fish, play and breathe. Ozone is a harmful oxidant (Duenas et al., 2004). Surface O^3 is a very strong oxidizing agent. It is by-product of the photochemical reaction between carbon compounds such as Volatile Organic Compounds (VOCs), Carbon Monoxide (CO) and Methane (CH4) with NOx. The reactions are initiated in the presence of sunlight (H. Seinfeld & Pandis, 1998). The chemical reactions involve the production and the destruction of O^3 simultaneously. Photochemical smog is another by-product of the same reactions, irrespective of the degree of pollution in the atmosphere (Varshney & Sigh, 2003).

The reaction is heavily dependent on the concentration of Nitrogen Monoxide (NO) present in the atmosphere (Glavas, 1999). Other factors also have a significant impact on the process of photochemical ozone (O^3) production (Ghazali et al., 2010). These are mostly climate-related factors, such as temperature, cloudiness, sunlight, wind speeds and

directions, humidity and solar radiation. Consequently, alterations in the climatic condition affect the surface ozone concentrations. Change rapidly accompanying the change in wind speed and direction, temperature, humidity, and solar radiation.

Since it is harmful, it can be particularly serious for the young, old, active and those with respiratory conditions at any age. Vulnerability of ozone harm are not just to human, but also can harms plants, crop, and agricultural yield. Interfering with pretty important process such as photosynthesis and even economy. To make matters worse, ozone production increase with higher temperatures, which occur more frequently with climate change.

2.4 Malaysian Ground Level Ozone (O^3)

The Department of Environment (DOE) monitors ambient air quality throughout the country at peninsular area and Sabah and Sarawak. These monitoring stations are strategically located in urban, suburban and industrial areas to detect any significant change in the air quality which may be harmful to human health and the environment. The air quality status is reported in terms of Air Pollution Index (API). The API is calculated based on concentration of five major pollutants which are ground level ozone (O^3), carbon monoxide (CO), nitrogen dioxide (NO^2), sulphur dioxide (SO^2) and particulate matter of less than 10 microns in size (PM_{10}) (Department of Environment Malaysia, 2016). The API is categorized as good, moderate, unhealthy, very unhealthy and hazardous as presented in Table 2.1.

Table 2.1
Malaysia : Air Pollutant Index (API)
(Source : Department of Environment Malaysia 2016)

IPU / API	STATUS KUALITI UDARA / AIR QUALITY STATUS
0 – 50	Baik / Good
51 – 100	Sederhana / Moderate
101 – 200	Tidak Sihat / Unhealthy
201 – 300	Sangat Tidak Sihat / Very Unhealthy
> 300	Berbahaya / Hazardous

Ground level ozone (O^3) remained the pollutant of concern. Ozone pollutant was formed as a result of chemical reaction between Volatile Organic Compounds (VOCs) and nitrogen oxides (NOx) in the presence of sunlight. Formation of O^3 enhanced during hot and sunny day. Major sources of VOCs and NOx emissions were from industries and motor vehicles particularly in urban areas. These resulted in several unhealthy days recorded at various locations in the Klang Valley and in the States of Perak, Negeri Sembilan, Johor, Kedah and Pulau Pinang.

According to Department of Environment Malaysia (2016), occasionally, the daily maximum 1-hour concentration of O^3 exceeded the Malaysian Ambient Air Quality Guidelines at several stations in the Klang Valley, Perak, Negeri Sembilan and Kedah. These conditions led to a number of unhealthy days recorded in some areas especially those of central business with heavy traffic volumes. Table 2.2 below lists the new Malaysian ambient air quality standard.

Table 2.2
New Malaysia Ambient Air Quality Standard
(Source : Department of Environment Malaysia 2016)

Pollutants	Averaging Time	Ambient Air Quality Standard		
		IT-1 (2015)	IT-2 (2018)	Standard (2020)
		µg/m^3	**µg/m^3**	**µg/m^3**
Particulate Matter with the size of less than 10 micron (PM_{10})	1 Year	50	45	40
	24 Hour	150	120	100
Particulate Matter with the size of less than 2.5 micron ($PM_{2.5}$)	1 Year	35	25	15
	24 Hour	75	50	35
Sulfur Dioxide (SO^2)	1 Hour	350	300	250
	24 Hour	105	90	80
Nitrogen Dioxide (NO^2)	1 Hour	320	300	280
	24 Hour	75	75	70
Ground Level Ozone (O^3)	1 Hour	200	200	180
	8 Hour	120	120	100
Carbon Monoxide (CO)	1 Hour	35	35	30
	8 Hour	10	10	10

2.5 Extreme Value Distributions

According to Datsiou and Overend (2018); (Ozay & Celiktas, 2016; Seal & Sherry, 2016) two parameter Weibull distribution is a commonly used in probabilistic, wind speed measurements and determining wind energy potential. Other than that, Ozay and Celiktas (2016) stated that Weibull distribution was also applied for fire protection and insurance problems, some occurrences in Germany's stock index and the behaviour of solar proton peak flux.

Wais (2017) fitted the two and three parameter Weibull distribution in wind power analysis. Where the study knowledge of wind characteristics helps to define site requirements, choose a proper turbine design and estimate profits from the wind energy production. The findings indicated that the three-parameter Weibull distribution gave better results compared to the two-parameter Weibull distribution. This was consistent with the findings of Örkcü, Aksoy, and Dog˘an (2015) which applied three parameter Weibull

distribution to CPU system resources. The study suggested that differential evolution (DE) approach which requires significantly less CPU time and exhibits a rapid convergence to the maximum value of the likelihood function in less iteration, provides accurate estimates and is satisfactory for the parameter estimation of the three parameter Weibull distribution.

Marzano (2014) in his paper proved that the covariance in any GEV model was always expressed by a one-dimensional integral, whose integrand function is available in closed form as a function of the generating function of the GEV model. Study conducted by Kim, Shin, Joo, and Heo (2012) was used GEV distribution in order to determine the parameters of the plotting position formula. The distribution was estimated by using a genetic optimization method known as the real-coded genetic algorithm (RGA). The study said that the accuracy of the derived plotting position formula for the GEV distribution was examined on the basis of the root mean square errors and relative bias between the theoretical reduced varieties and those calculated from the derived and existing plotting position formulas.

2.6 Extreme Value Distributions in Environmental Study

Ilhan et al. (2018) introduced a new estimation approach that he used in calculating the Weibull parameter for the estimation of wind power in Turkey. The findings indicated that the multi-objective moments (MUOM) definitely provides more accurate estimate in estimating wind power based on the Weibull distribution. However, the findings of A. Tiaon et al. (2017) which compared different methods of determining Weibull parameter to the wind energy resources in Kiribati island. The study indicated that the most accurate method to obtain Weibull parameter was by using the method of moments.

In the application of GEV distribution in the study of Hanbeen et al. (2018), the study compared the performance indicators among Akaike's information criterion (AIC), corrected Akaike's information criterion (AICc), Bayesian information criterion (BIC), and likelihood ratio test (LRT) on the GEV to the rainfall data at Korean country, and the findings was in favour of the BIC. Other study on plotting position formula for the GEV distribution was carried out by Sooyoung et al. (2012) reveals that for various sample sizes

and shape parameter was derived by using the theoretical reduced variates of the GEV distribution.

Peiman et al., (2017) used Switzerland level of river discharges data at ungauged locations on a river network. The study estimated the high return levels based on regionalizing the parameters of GEV. The findings showed that the approach improves the estimation uncertainty for long return periods. Table 2.3 shows the past literatures on the Environmental study that review for this research paper.

Table 2.3
Summary of research Extreme value distributions in environmental Study

Authors	Tittle	Data(S)	Distribution in Analysis
Ilhan Usta et al. (2018)	A new estimation approach based on moments for estimating Weibull parameters in wind power applications	Wind Power	Weibull
Peiman Asadi et al. (2018)	Optimal regionalization of extreme value distributions for flood estimation	Flood	Generalized Extreme Value
Wonseon Gwaka et al. (2018)	Extreme value theory in mixture distributions and a statistical method to control the possible bias	Rainfall	Metastatistical Extreme Value (Mev)
Tiaon Aukitino et al. (2017)	Wind energy resource assessment for Kiribati with a comparison of different methods of determining Weibull parameters	Wind	Weibull
Hanbeen Kim et al. (2017)	Appropriate model selection methods for nonstationary generalized extreme value models	Rainfall Data	Generalized Extreme Value
Sooyoung Kim et al. (2012)	Development of plotting position for the general extreme value distribution	Rainfall Or Flood	Generalized Extreme Value

2.7 Extreme Value Distributions in Air Pollution Study

The study found the Gamma, Lognormal and Weibull distribution is the best fit model for air pollution by Muhammad Ismail Jaffar, Hazrul Abdul Hamid, Riduan Yunus, and Raffee (2018). In this study found that the log normal distribution as the best-fitted model to predict the PM_{10} concentration. While other distributions which are extreme value and logistic as the best distribution model to Sulphur dioxide, nitrogen dioxide and particulate matter. Other than that, the better fitting compare to Weibull and gamma, the study proved that the widely used distribution which is Weibull was not give better fit to the PM_{10} concentrations compared to gamma distribution.

However, study conducted by Ahmat and Yahaya (2018), found the different finding for the best fitted distribution on PM_{10} concentrations data in Malaysia. In this study, classical distribution which are Two-parameter Gumbel, Two and Three-parameter Weibull, Generalized Extreme Value (GEV), Two and Three-parameter Generalized Pareto Distribution, while, for Bayesian distribution which are Three-parameter Weibull, Generalized Extreme Value (GEV), Three-parameter Generalized Pareto Distribution was used to compare the best prediction to predict exceedances of PM_{10}. The study found that, the Bayesian approach is superior in prediction of PM_{10} concentration data compared to the other conventional method.

Other study form Brazil, conducted by Leila Droprinchinski Martins et al. (2017) were more on to compare the air quality between the two largest Brazilian urban areas (Metropolitan Area of Sao Paulo (MASP) and Metropolitan Area of Rio de Janeiro (MARJ)) and provide information for decision makers, government agencies and civil society. Generalize Extreme Value (GEV) and Generalized Pareto Distribution (GPD) was used on the air pollution data which are CO, NO, NO^2, PM_{10} and $PM_{2.5}$ and ozone. According to the study, these two distributions was give the same result event it's a different approach. However, the study found that when the regions are compared, MASP presented higher probabilities of extreme events for all analysed pollutants, except for NO; while O^3 and $PM_{2.5}$ are those with most frequent probabilities of presenting extreme episodes, in comparison other pollutants. Table 2.4 summaries literature in the air pollution studies.

Table 2.4

Summary of research in Air Pollution studies

Authors & Year	Area of Study	Distribution	Pollution
Muhammad Ismail et al. (2018)	Malaysia	Gamma, Lognormal and Weibull distribution	SO^2, NO^2 and PM_{10}
Ahmat and Yahaya (2018)	Malaysia	Classical (Two-parameter Gumbel, Two and Three-parameter Weibull, Generalized Extreme Value (GEV), Two and Three-parameter Generalized Pareto Distribution) and Bayesian (Three-parameter Weibull, Generalized Extreme Value (GEV), Three-parameter Generalized Pareto Distribution) approaches	PM_{10}
Leila Droprinchinski Martins et al. (2017)	Brazil	Generalized Extreme Value (GEV) and Generalized Pareto Distribution (GPD)	CO, NO, NO^2, PM_{10}, $PM_{2.5}$ and O^3

2.8 Extreme Value Distributions in Ozone Study

Mohd Talib et al. (2012) in his paper used the GEV and Generalized Pareto Distribution on four years' ozone data from year 2004 to 2008. Klang Valley was the targeted area in his research study. The paper found that there were distinct seasonal patterns in the of ozone surface across the Klang valley area in this study, suggested the most contributor to the concentration of ozone was may NMHCs substance and further investigation should be conducted in the future.

Other study conducted by Muqhlishah et al. (2015) applied different method on the ozone data. The paper used daily average data at Shah Alam area from 2002 until 2013. Multiple linear regression with a concept from Ordinary least square method were performed to find three days ahead prediction of daily 12-hour ozone (O^3) concentrations. In this study, the model was accessed by the accuracy measures (AI, PA and R2) and the error measures (RMSE, NAE) and the result showed that the result was close to zero hence, the model was classed as the best model to predict air pollutant of ozone concentration.

Trajectory analysis was used by Negar Banan et al. (2013) in his study which is characteristics of surface ozone concentration at stations with different background in peninsular Malaysia. Similar to other studies conducted by Muqlishah and Mohd Talib, Negar Bana also chose Klang Valley and Putrajaya as his location of interest. In this study, data from year 2005 to 2009 was used since in that year was recorded as the higher concentration of surface ozone in Peninsular Malaysia (Negar Banan et al. 2013). The analysis reveals that the positive sign of the highest surface ozone concentration was recorded in a suburban area which is Putrajaya. The study concluded that the ozone concentrations were influenced by the characteristics of nitrogen oxides, particularly the titration of NO.

Table 2.5
Summary of research Extreme Value distributions

Authors	Tittle	Area of study and period of data	Distribution in Analysis	Findings
Mohd Talib et al. (2012)	Variations of surface ozone concentration across the Klang Valley, Malaysia	Klang Valley 2004 - 2008	Backward Trajectory Analysis	NHMc substance is the potential sources contributing to O^3 during very high O^3 days.
Muqhlisah Muhamad at al. (2015)	Three Days Ahead Prediction of Daily 12 Hour Ozone (O^3) Concentrations for Urban Area in Malaysia	Shah Alam 2002 - 2013	Multiple Linear Regression (MLR)	The average accuracy (AI, PA and R^2) and the average error measures (RMSE, NAE) showed that the model is the best to predict the ozone concentrations.
Negar Banan, et al. (2013)	Characteristics of Surface Ozone Concentrations at Stations with Different Backgrounds in the Malaysian Peninsula	Petaling Jaya (S2) (urban), Putrajaya (S1) (suburban) and Jerantut (S3) (rural) 2005-2009	Trajectory Analyses	The results showed that the highest O^3 concentration was recorded in a suburban area which is Putrajaya.

CHAPTER THREE
METHODOLOGY

3.1 Introduction

This chapter provides the discussion about the method of analyzing ozone (O^3) concentrations in selected areas in peninsular Malaysia using extreme value theory approach. The area of research, and the analysis O^3 concentration will be explained further using two and three parameter Weibull and Generalized Extreme Value (GEV) approach in order to obtain the extreme concentration of O^3 in each selected area.

To determine trend of the O^3 concentrations, the R languages will be used. It can provide various range of graphical display, calculation and data manipulation. In addition, Matlab ver.17.1 was used to estimate the performance indicators and parameters of the distributions IBM SPSS version 25 was used to select data on certain entry and plotting time series.

Figure 3.1 illustrates the flow of methodology with the research gap in this study. In analyzing the O^3 concentrations, the study will be divided into few objectives. The research started with the selection of the records, and the analysis of O^3 using classical methods in order to fill the gap of research.

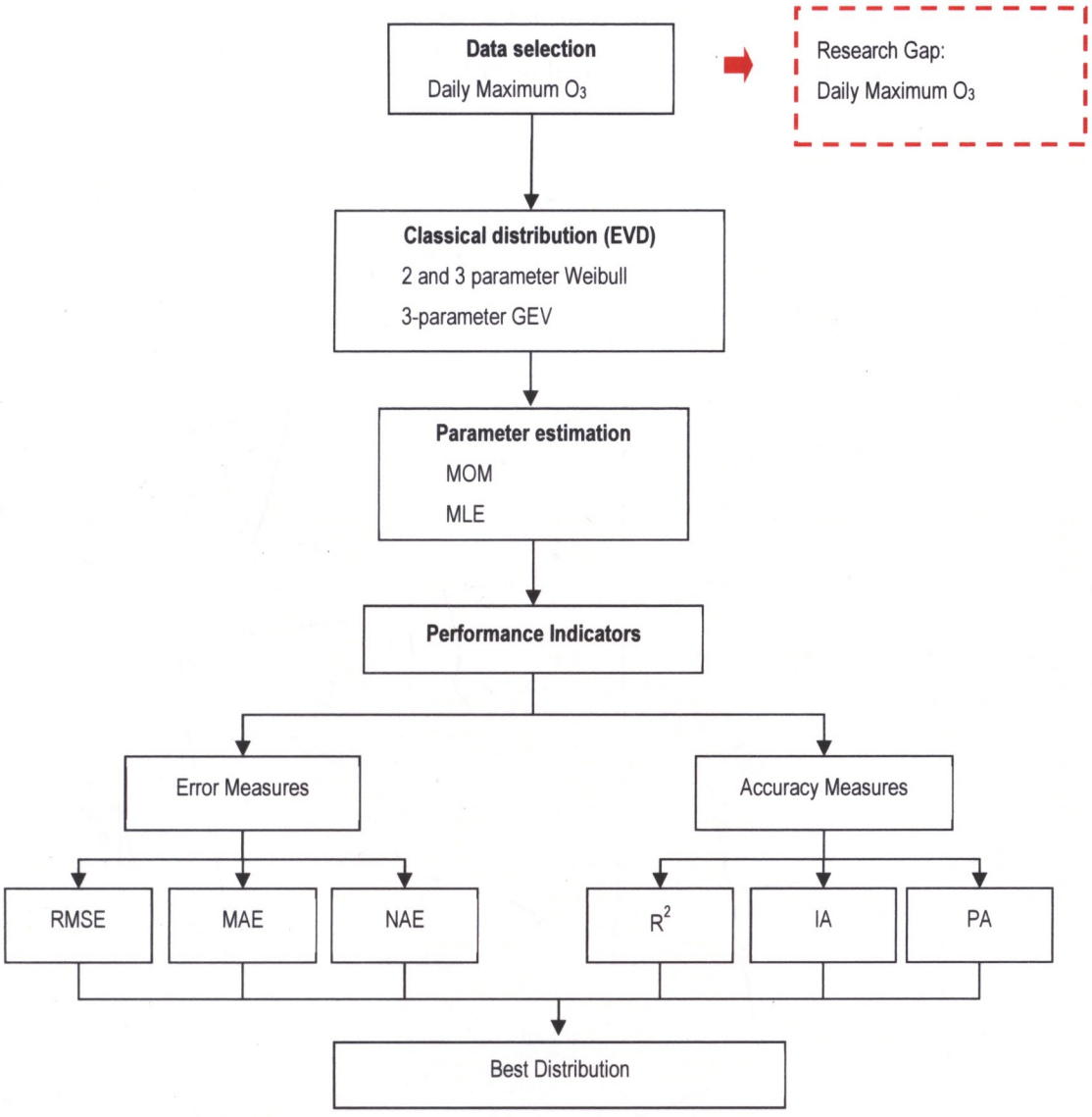


Figure 3.1 Flow of research methodology with indicator of research gap

Performance indicators will be used to evaluate the methods in order to obtain the best model. This performance indicator was measured the error measure and accuracy measure. Where, less value measure error and high value accuracy measure from performance indicator will reveal the best model to represent for each station.

3.2 Area of Research

Figure 3.2 Location of Continuous Air Quality Monitoring Stations in Peninsular Malaysia, 2016

Figure 3.2 showed the location of monitoring station. The investigation utilizes daily maximum concentrations of ozone (O^3) from three (3) monitoring stations and one (1) background stations in Peninsular Malaysia: Putrajaya, Klang, Shah Alam, and Jerantut furnished by the Department of Environment (DOE). The original data were the hourly concentrations per hour and the total observations for all locations were 3650 days.

The locations can be categorized as follows: Sah Alam (urban area), Klang (urban area), Putrajaya (rural area) and Jerantut will be the background station.

3.3 Analysis of ozone (O^3) Concentrations

The daily maximum concentrations will be presented in descriptive statistics and time series plot. Missing value issue might due to machine error or human error at most of data collection process. N et al. (2011) said that the missing value can affect the entire data if not treated well. However, for this study, missing value were omitted from the analysis. There are few researches in environmental study that omitted the missing value for their analysis of the data (Ahmat, Yahaya, & Ramli, 2016; Awang, Ramli, Mohammed, & Yahaya, 2013; Junninen, Niska, Tuppurainen, Ruuskanen, & Kolehmainen, 2004; Svensson, Clarke, & Jones, 2007; Yahaya et al., 2013).

3.3.1 Descriptive Statistics

The descriptive statistics is the discipline of qualitatively portraying the fundamental natures of a quantitative illustrative as numerical or graphical strategies. The main role of descriptive statistics is to conclude a sample compare than use the data to drill the information about the population. Generally, three type of information about the data represented by descriptive statistics, are, location, dispersion and shape. The range, standard deviation, coefficient of variation, percentile and interquartile range are the measurement of dispersion. The mean, median and mode represent the location of the data to determine the spreadness of the data from its mean is known as dispersion. Variance, skewness and kurtosis will show the shape of data distribution. The components are

classified as the measures of central tendency (M. J. Fisher & Marshall, 2009)

3.3.2 Box-and Whisker Plot

Box-and-Whisker plot visually summarizes and compares the group of data. The median, quartiles, the lowest and highest data points to express the dispersion and symmetry of a distribution of the data were utilize by the presentation of Box-and-Whisker plot. Moreover, demonstrate the existence of outliers in the data were more easily in the form of Box-and-Whisker plot (Williamson et al., 1989). Figure 3.3 gives full description of Box-and-Whisker plot.

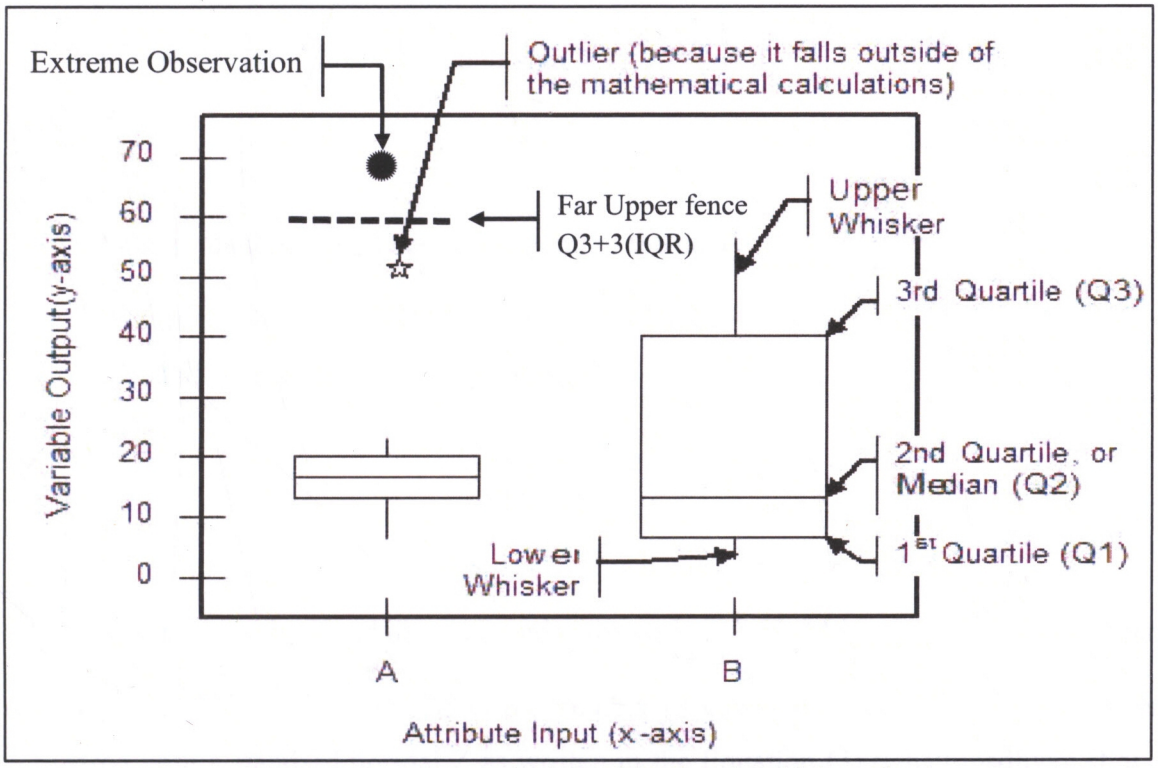


Figure 3.3 Description of Box-and-Whisker plot
Source : (Blattenberger, 2018)

3.3.3 Mann-Kendall's (MK) Trend Test

The Mann-Kendall's test which statistically assesses any abrupt change in trend in the time series (Rinner & Hussain, 2011). Mann-Kendall has been widely used in climatological and environment studies where it is rank-based non-parametric test to test the significance of a trend (Chattopadhyay, Chakraborthy, & Chattopadhyay, 2012)

The MK trend tests use the statistics S with the formula in Equation (1).

$$S = \sum_{i=2}^{n} \sum_{j=1}^{i-1} sign(x_i - x_j) \quad (1)$$

where x_j is the sequential data values, n is the length of the time series and $sign\,(x_i - x_j) = -1$ if $(x_i - x_j) < 0$ and sign $(x_i - x_j) = 1$ if $(x_i - x_j) > 0$. The hypothesis of this is:

Null hypothesis H_0: There is no trend in the data

Alternative hypothesis H_1: There is trend in the data

Kendall's τ is defined as in Equation (2)

$$\tau = 2\frac{S'}{n(n-1)} \quad (2)$$

Where S' is the Kendall's sum and estimated as S'=L-M is where L is the number of cases with $(x_i - x_j) > 0$ and M is the number of cases for which $(x_i - x_j) < 0$. It is compared with a standard normal Z as written in the Equation (3) (Chattopadhyay et al., 2012).

$$Z = \begin{cases} \dfrac{S-1}{Var(S)}, & S > 0 \\ 0, & S = 0 \\ \dfrac{S+1}{Var(S)}, & S < 0 \end{cases} \qquad (3)$$

And $Var(S) = \dfrac{n(n-1)(2n+5) - \sum_{p=1}^{q} t_p(t_p-1)(2t_p+5)}{18}$ where t_p is the number of ties for the pth value and q is the number of tied values. For two-sided test for trend, the null hypothesis H0 is rejected when $|Z| > Z_{\alpha/2}$ where $\alpha = 0.05$ is the significance level.

Hirsch (1982) stated that using Kendall slope estimator and it's unbiased estimator of the slope of a linear trend can estimate magnitude. It gives higher exactness than regression estimators where data are profoundly skewed however bring down accuracy where the information is ordinary.

3.4 Monitoring Records Selection

This study engages one method in selecting the records of ozone concentrations. Various studies used the method of daily maximum method in selecting the record. This study also interested to explore the ozone concentrations which are above the Malaysia ambient air Quality Guidelines (MAAQG) levels of 120 μg/ m^3 for daily concentrations.

The straight-forward method of the daily maximum data is selected from the maximum of series of 24-hour monitoring record from 00:00 to 23:00 hours.

3.5 Extreme Value Distribution (EVD)

The nature of the method utilized as a part of statistical analysis depends extraordinarily on the expected of probability model or distributions. As a result, extensive effort has been expanded in the advancement of huge arrangements of standard probability

distributions along with relevant statistical methodologies, designed to provide statistical models would not be a valuable probability density function (PDF) for studying every phenomenon. One of the statistical procedures is the measures of central tendency. Measure of central tendency is a single value that attempts to clarify a set of data by perceiving the central position within that set of data. I. Elbatal (2014) stated central tendency can be used to study the most important features and characteristics of a distribution.

Among the objectives of this study are to apply a probability distribution, which are, extreme value distribution (EVD) which provides a theoretical framework to model and analyze air pollution extreme concentration. In 1928, Fisher and Tippet initiated Extreme Value Theory (EVT) unfolding the behaviour of maximum of independent and identically distributed random variable (IID). The EVT or also known as Generalized Extreme Value distribution, are widely used in finance, risk management, material sciences, economics, insurance, hydrology, telecommunications, and many other industries dealing with extreme events. Studies involving natural phenomena using EVT such as rainfall, the height of sea waves, floods, corrosion, and wind speed have been of great interest to scientist and researchers for a long period of time (Ozay & Celiktas, 2016).

Figure 3.4 illustrate the flow of methodology in obtaining the best EVD to represent each station. However, there is no consensus about which the most appropriate methods from the several methods is to estimate parameters for each EVD. Performance indicators or error measures are the indicators to determine appropriateness of the methods. Two methods of estimates were presented in this study, namely: The Method of Moment (MOM), and method of Maximum Likelihood Estimator (MLE).

The Method of Moment (MOM), are widely used and known as the simplest method, involves the calculation of the first few sample moments of the observed data. Estimates for the values of parameters are obtained from the observed data, and the equations for the moments solve for the parameters in different distribution (Jasim et al., 2011). The formula for MOM is different in every distribution, hence, in this paper the formula for MOM was write below the explanation of two and three parameter Weibull and GEV distribution explanation.

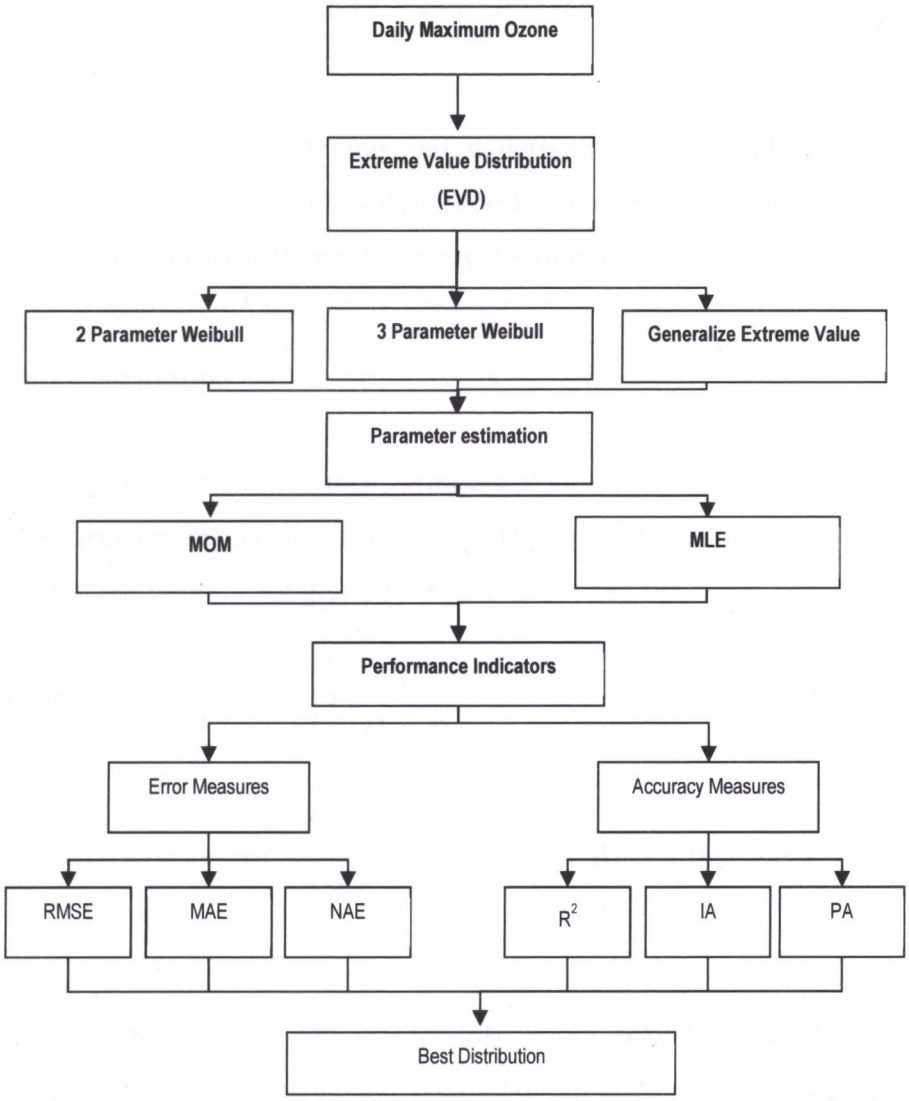


Figure 3.4 Flow of methodology to obtain the best EVD

The second method is the MLE which optimized the estimates however it may require more complex numerical calculation than the MOM. The method is to maximize the likelihood function of the parameters $\theta_1, \theta_2, ..., \theta_k$ of a k-parameter distribution defined as the joint probability density function (PDF) of the observations in random sample size n

as shown in Equation (4).

$$L(\theta_1,\theta_2,...,\theta_k)=\prod_{t=1}^{n} f_x(x_i\,|\theta_1,\theta_2,...,\theta_k) \tag{4}$$

The MLE is obtained by taking partial derivatives of L with respect to each parameter, set them equal to zero and solving the k equations simultaneously (Chung & Kim, 2013). The formulae are different for every distribution.

3.5.1 Weibull Distribution

A Swedish engineer and scientist, Waloddi Weibull (1887-1979) formulated type III EVD, for his work on quality of materials and fatigue analysis. The pdf and cdf of two-parameter Weibull distribution are given by Equation (5) and (6).

$$f(x;\sigma,\lambda)=\frac{\lambda}{\sigma}\left(\frac{x}{\sigma}\right)^{\lambda-1}\exp\left[-\left(\frac{x}{\sigma}\right)^{\lambda}\right] \qquad for\, x\geq 0;\sigma,\lambda>0 \tag{5}$$

$$F(x;\sigma,\lambda)=1-\exp\left[-\left(\frac{x}{\sigma}\right)^{\lambda}\right] \qquad for\, x\geq 0,\sigma>0,\lambda>0 \tag{6}$$

where λ is the shape parameter, σ is the distribution scale (Rinne, 2008).

The pdf and cdf for three-parameter Weibull distribution are written in Equation (7) and (8).

$$f(x;\mu,\sigma,\lambda)=\frac{\lambda}{\sigma}\left(\frac{x-\mu}{\sigma}\right)^{\lambda-1}\exp\left[-\left(\frac{x-\mu}{\sigma}\right)^{\lambda}\right] \qquad for\, x\geq \mu;\sigma,\lambda>0 \tag{7}$$

$$F(x;\mu,\sigma,\lambda)=1-\exp\left[-\left(\frac{x-\mu}{\sigma}\right)^{\lambda}\right] \qquad for\, x\geq 0,\sigma>0,\lambda>0 \tag{8}$$

where λ is the shape parameter, σ is the scale parameter and μ is the location parameter (Rinne, 2008).

The shape parameter determines s the appearance of the Weibull density while the scale parameter determines the appearance of the Weibull density while the scale parameter represents the spreadness of distribution. The shape parameter $0<\lambda<1$, the density approaches ∞ as x approaches to 0 from the right. In the case of $\lambda=1$, the density approaches 1 as x=0 and the curve follows $1/\lambda$ as x increases.

3.5.1.1 Method of Moments (MOM)

The MOM estimates the two parameter and three parameter Weibull using the formula in Equation (9), (10), (11) and (12) respectively.

$$\lambda = cv^{-1.0852} \tag{9}$$

$$cv = \frac{s}{\bar{x}} \tag{10}$$

$$\sigma = \frac{\bar{x}}{\Gamma\left(1+\frac{1}{\lambda}\right)} \tag{11}$$

$$\frac{s}{\bar{x}} = \sqrt{\frac{\Gamma\left(1+\frac{2}{\lambda}\right)-\Gamma^2\left(1+\frac{1}{\lambda}\right)}{\frac{\mu}{\sigma}+\Gamma\left(1+\frac{1}{\lambda}\right)}} \quad for\ \lambda > 2 \tag{12}$$

Where $\bar{x}$ the sample is mean, s is the standard deviation and Γ is the Gamma function (Bury, 1999).

3.5.1.2 Method of maximum likelihood (MLE)

Two-parameter and three-parameter estimates for Weibull distribution are written in Equation (13), (14), (15), (16) and (17) respectively (Rinne, 2008).

$$\sigma = \left[\frac{1}{n} \sum_{i=1}^{n} (x_i)^{\lambda} \right]^{1/\lambda} \tag{13}$$

$$\frac{1}{\lambda} - \frac{\sum_{i=1}^{n} x_i^{\lambda} \ln x_i}{\sum_{i=1}^{n} x_i^{\lambda}} + \frac{1}{n} \sum_{i=1}^{n} \ln x_i = 0 \tag{14}$$

$$\sigma = \left[\frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^{\lambda} \right]^{1/\lambda} \tag{15}$$

$$\frac{1}{\lambda} - \frac{\sum_{i=1}^{n} (x_i - \mu)^{\lambda} \ln(x_i - \mu)}{\sum_{i=1}^{n} (x_i - \mu)^{\lambda}} + \frac{1}{n} \sum_{i=1}^{n} \ln(x_i - \mu) = 0 \tag{16}$$

$$\frac{\lambda - 1}{\lambda} \sum_{i=1}^{n} (x_i - \mu)^{-1} - n \frac{\sum_{i=1}^{n} (x_i - \mu)^{-1}}{\sum_{i=1}^{n} (x_i - \mu)^{\lambda}} = 0 \tag{17}$$

3.5.2 Generalized Extreme Value (GEV)

A single GEV distribution is a result of a combination of a family of continuous probability distribution which was proposed for statistical stability. The GEV which was popularized by Jenkinson in 1955 (Bali, 2003) have three parameters, namely: location, scale and shape. Location parameter, μ determining the shifting of a distribution in a specified direction on the horizontal axis. The dispersion of the distribution is measured by the scale parameter, σ and it indicates where the concentration of the distribution lies. Reducing the value of σ will cause an expansion of the density and vice-versa. On the other

hand, λ is the shape parameter, affects the shape of distributions and tails of the distribution. It organize the skewness as to where the greater part of data are concentrated, in this way, makes the tail(s) of distribution (Millington, Das, & Simonovic, 2011).

The pdf is written in Equation (18)

$$f(x;\lambda,\sigma,\mu)=\frac{1}{\sigma}\left[1+\lambda\left(\frac{x-\mu}{\sigma}\right)^{-1/\lambda-1}\right]\exp\left\{-\left[1+\lambda\left(\frac{x-\mu}{\sigma}\right)^{-1/\lambda}\right]\right\} \tag{18}$$

where $x>\mu-\frac{\sigma}{\lambda}$, $-\infty<\lambda<\infty$ is the location parameter, $\sigma>0$ is the scale parameter and $-\infty<\lambda<\infty$ is the shape parameter. The shape parameter, λ establishes the type of extreme value distribution. Gumbel distribution –Type I distribution in when $\lambda=0$. Type II – Frèchet has $\lambda>0$ and it has bounded lower side when $x>0$ (Aryal & Tsokos, 2009). Type III distribution, which is commonly known as "Reversed Weibull", is the result of a negative shape parameter, $\lambda<0$. The upper ends of both Type I and Type II are unbounded, however, Type I has a thinner tail than Type II. Type III has a finite upper limit.

The corresponding cdf is written in Equation (19)

$$F(x;\lambda,\sigma,\mu)=\begin{cases}\exp\left\{-\left[1+\lambda\left(\frac{x-\mu}{\sigma}\right)^{-1/\lambda}\right]\right\} & \lambda\neq 0 \\ & and\ for\ 1+\lambda\left(\frac{x-\mu}{\sigma}\right)>0 \\ \exp\left\{-\exp\left(-\frac{x-\mu}{\sigma}\right)\right\} & \lambda=0\end{cases} \tag{19}$$

3.5.2.1 Method of Moments (MOM)

Equation (20), (21) and (22) were used to estimate parameter of the GEV distributions (Martins & Stedinger, 2000).

$$\mu=\bar{x}-\frac{\sigma}{\lambda}\left[1-\Gamma(1+\lambda)\right] \tag{20}$$

$$\sigma = \frac{s|\lambda|}{\left\{\Gamma(1+2\lambda)-[\Gamma(1+\lambda)]^2\right\}^{3/2}} \tag{21}$$

$$C_s = sign(\lambda)$$

$$\frac{-\Gamma(1+3\lambda)+3\Gamma(1+\lambda)\Gamma(1+2\lambda)-2[\Gamma(1+\lambda)]^3}{\left\{\Gamma(1+2\lambda)-2[\Gamma(1+\lambda)]^2\right\}^{3/2}} \tag{22}$$

where Cs is skewness.

3.5.2.2 Method of maximum likelihood (MLE)

Parameter estimation using this method utilizes Equations (23), (24) and (25) (Martins & Stedinger, 2000).

$$\frac{1}{\sigma}\sum_{i=1}^{n}\left(\frac{1-\lambda-(1-(\lambda/\sigma)(x_i-\mu))^{1/\lambda}}{(1-(\lambda/\sigma)(x_i-\mu))}\right)=0 \tag{23}$$

$$-\frac{n}{\sigma}+\frac{1}{\sigma}\sum_{i=1}^{n}\left[\frac{1-\lambda-(1-(\lambda/\sigma)(x_i-\mu))^{1/\lambda}}{(1-(\lambda/\sigma)(x_i-\mu))}\left(\frac{(x_i-\mu)}{\sigma}\right)\right]=0 \tag{24}$$

$$-\frac{1}{\lambda^2}\sum_{i=1}^{n}\left[\begin{array}{l}\ln(1-(\lambda/\sigma)(x_i-\mu))\left\{1-\lambda-[1-(\lambda/\sigma)(x_i-\mu)]^{1/\lambda}\right\} \\ +\dfrac{1-\lambda-[1-(\lambda/\sigma)(x_i-\mu)]^{1/\lambda}}{(1-(\lambda/\sigma)(x_i-\mu))}\lambda\left(\dfrac{x_i-\mu}{\sigma}\right)\end{array}\right]=0 \tag{25}$$

3.6 Performance Indicators (PI)

The best method for fitting the model is using a systematic optimization routine that estimates the parameters by several distributions indicated in Section 3.5 and 3.6 using the $PM_{2.5}$ concentrations records and then compares how these estimated distributions fit the data using various criteria of "goodness of fit". Three common error measures; the Mean

Absolute Error (MAE), Root Mean Square Error (RMSE) and Normalized Absolute Error (NAE) and three accuracy measures; Prediction Accuracy (PA), and Coefficient of Determination (R2) will be used in this study (Ahmat, Yahaya, Ramli, Japeri, & Hamid, 2015). For more understanding of calculation of goodness-of-fit, Table 3.1 lists the notations needed in the calculation of goodness-of-fit which are used in Table 3.2.

Table 3.1
Notations used in obtaining goodness-of-fit

	Past					Present time
Observed Value	O1	O2	… ..	Ot-2	Ot-1	Ot
Period t	1	2	… ..	t-2	t-1	t
Estimated Values	P1	P2	… ..	Pt-2	Pt-1	Pt
Error Values	e1	e2	… ..	et-2	et-1	et

3.6.1 Error Measures

The value of error measure fluctuates from 0 to $+\infty$. The errors measure the average magnitude of the errors. The model is deemed to be the best model as the values of error measures have the lower values using formulae (26) - (28) in Table 3.2 (Jasim et al., 2011). The error measures in this study are unit-dependent and scale (Ji & Gallo, 2006).

Table 3.2
Notations used in defining performance indictaors

Notation	Meaning
n	Number of observed records
et	Forecast error, Ot - Pt
Ot	Observed records
$\overline{O}$	Mean of observation $\frac{1}{n}\sum_{t=1}^{n} O_t$
P_t	Prediction records
$\overline{P}$	Mean of predicted records $\frac{1}{n}\sum_{t=1}^{n} P_t$
S_O	Standard deviation of Observed records $S_O = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(O_i - \overline{O})^2}$
S_P	Standard deviation of Predicted records $S_p = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(P_i - \overline{P})^2}$

Table 3.3
Description of Error Measures and their formulae

Error Measures	Description	Formulae	Equation no.
RMSE	The normal distinction between the observed and predicted values	$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(O_t - P_t)^2}{n}}$	(26)
MAE	The normal outright distinction between predicted values and observed.	$MAE = \frac{\sum_{t=1}^{n}\lvert(O_t - P_t)\rvert}{n}$	(27)
NAE	The error between the predicted and observed values of an estimator and the calculation or model	$NAE = \frac{\sum_{t=1}^{n}\lvert(P_t - O_t)\rvert}{\sum_{t=1}^{n} O_t}$	(28)

3.6.2 Accuracy Measures

Table 3.4 shows the formulae (29) - (31) will be used to calculate the accuracy of the selected model. The accuracy value fluctuates between 0 and 1 and as the value approaches 1, the model is appropriate (Jasim et al., 2011). When the value is closed to 1, the model is more suitable to simulate the experimental data. All the accuracy measures are dimensionless, that is independent of the unit of data (Ji & Gallo, 2006).

Table 3.4
Description of Accuracy Measures and their formulae

Accuracy Measures	Description	Formulae	Equation no.
R2	Statistical indicators to measure the accuracy of estimators or models.	$R^2 = 1 - \frac{\sum_{t=1}^{n}(O_t - P_t)^2}{\sum_{t=1}^{n}(O_t - \overline{O})^2}$	(29)
PA	The proximity of predicted to the observed values	$PA = \sum_{t=1}^{n}\frac{(P_t - \overline{P})(O_t - \overline{O})}{(n-1)S_P S_O}$	(30)
IA	Dimensionless statistical indicator that communicates the contrast amongst predicted and observed values	$IA = 1 - \frac{\sum_{t=1}^{n}(P_t - O_t)^2}{\sum_{t=1}^{n}(\lvert P_t - \overline{O}\rvert - \lvert O_t - \overline{O}\rvert)^2}$	(31)

3.7 The Exceedances

The Estimation of the exceedances is calculated from the probability of concentrations exceeding 120 μg/m^{3}which is obtained from the cumulative distribution function (CDF) plot multiply with the number of days (Ahmat et al., 2015).

CHAPTER FOUR
RESULT AND DISCUSSION

4.1 Introduction

Section 4.2 through Section 4.6 will discuss on the research findings for this study. The descriptive of monitoring records presented in Section 4.2 consist of descriptive, time-series plot and MK test for daily maximum ground level ozone concentration and method will be deliberated in this section. The result of Objective 2 the classical approach where also known as the Extreme Value Distribution (EVD) which including two and three parameter Weibull and GEV distribution are available in Section 4.3. The findings are presented in organize way where the parameter estimation will identify from Objective 2 and will be used in Objective 3 to examine the performance indicators to identify the best distribution of EVD.

4.2 Daily maximum

Daily maximum method is well known as common method in various environmental pollution studies.

4.2.1 The Characteristics and Pattern of Daily Maximum

Histogram as indicated in Figure 4.1 shows that there was one peak existed for the concentration in 2007 - 2016 and those situations may additionally influence the determination of model this is delegate to the distribution of the concentrations in Jerantut.

The histogram plot of the ozone concentration in Klang Putrajaya and Shah Alam as shown in Figure 4.1, shows that the concentration were nicely distributed with light right tail. The model obtain in this study is expected to fit well the distribution of the monitoring records.

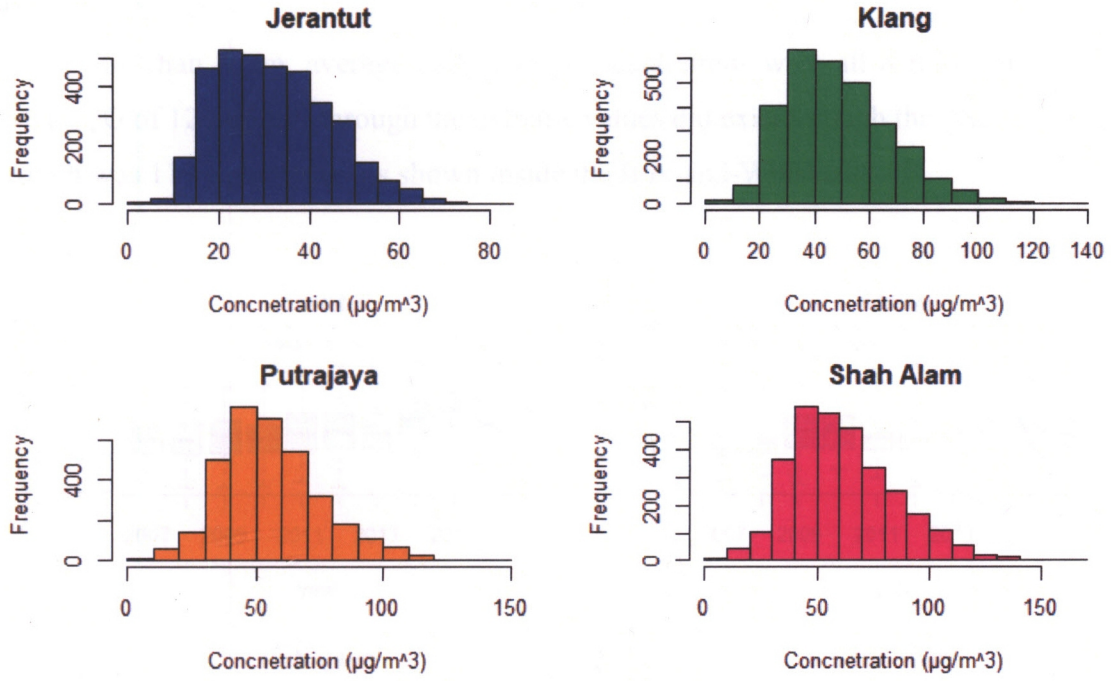


Figure 4.1 Histogram plot showing the distribution of ozone concentrations at four location in Peninsular Malaysia

Figure 4.2 showed Box-and-Whisker plot for Jerantut, as a reference station, the daily most concentrations were all beneath the everyday MAAQG of 120 µg/m3, through the extreme values did not exist in year 2002 - 2016, however not exceed the MAAQG of 120 µg/m^3 level. All the ozone concentrations records through 2007 – 2016 indicated the non-existence of extreme concentrations in Jerantut. This shows that the ozone concentration in Jerantut is under control.

Different in Klang, average daily most concentrations was beneath the 8-hour MAAQG of 120 µg/m3, through the extreme values did exist through the year except 2009,2010, 2015 and 2016 which above MAAQG of 120µg/m^3 level as shown inside the Box-and-Whisker.

Figure 4.2 shows that there were two peaks existed for the concentration in 2007 – 2016 in Putrajaya. As categorical as rural area, the daily most concentrations for Putrajaya were all beneath the everyday MAAQG of 120 µg/m^3, through the extreme values did exist in year 2014 and 2016 as shown inside the Box-and-Whisker which is exceed the MAAQG

of 120μg/m^3 level.

In Shah Alam, average daily most concentrations were all beneath the everyday MAAQG of 120 μg/m^3, through the extreme values did exist through the year except 2012 which was 119 μg/m^3 level as shown inside the Box-and-Whisker.

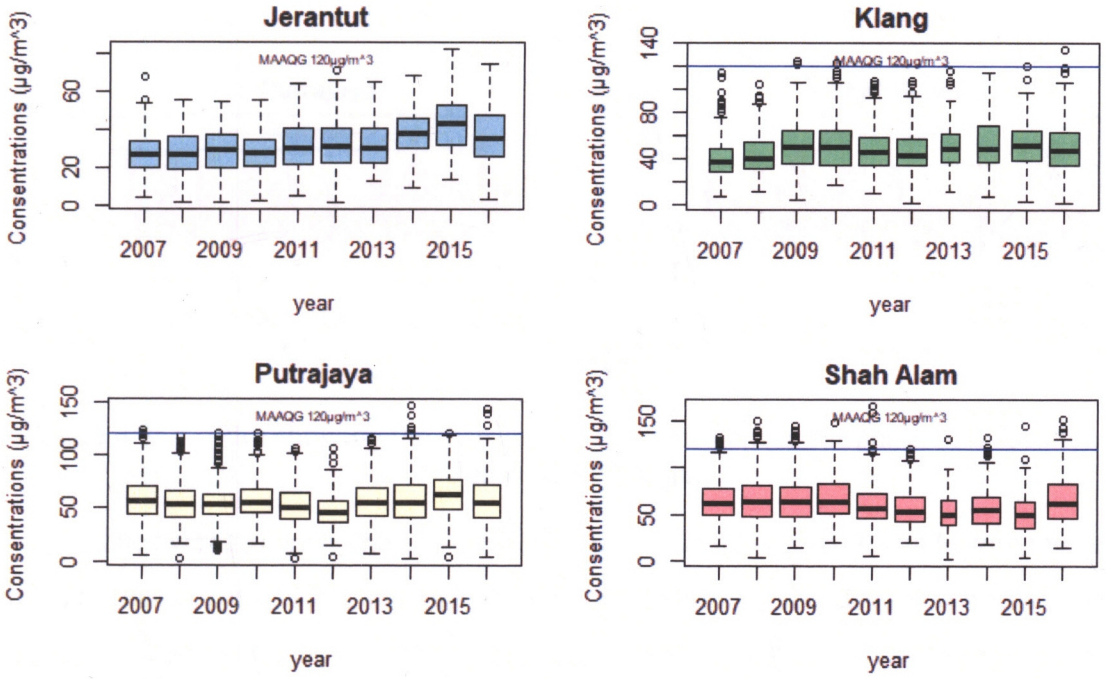


Figure 4.2 Box-Plot showing the distribution of ozone concentrations at four location in Peninsular Malaysia

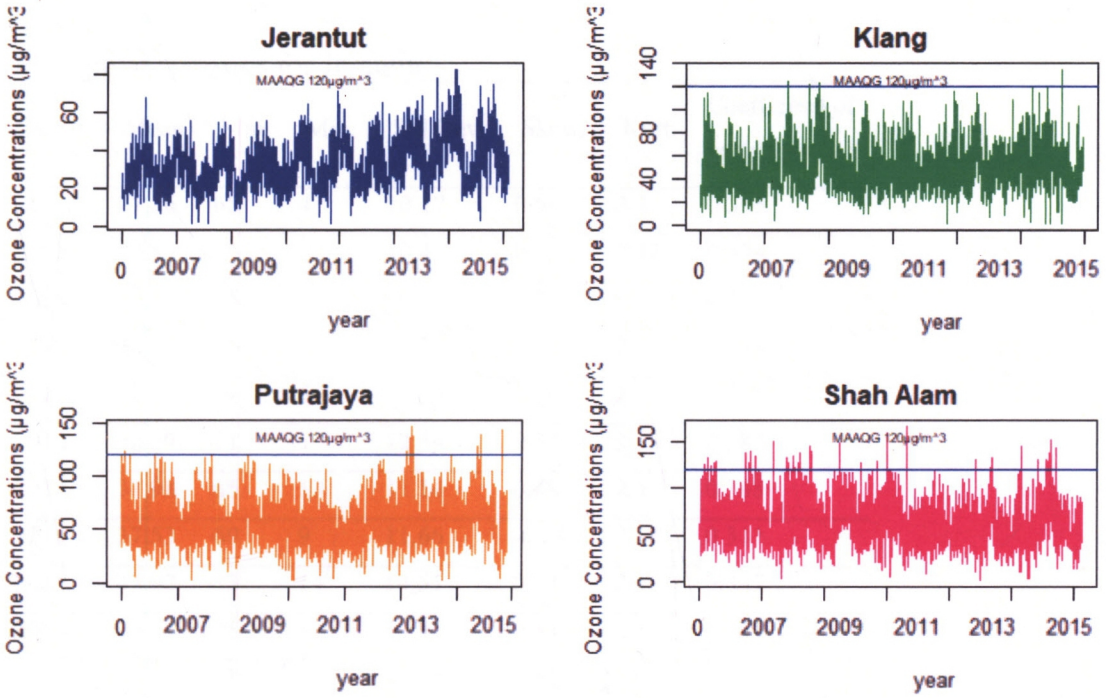


Figure 4.3 Plot showing the distribution of ozone concentrations at four location in Peninsular Malaysia

The average concentrations during the period 2007 – 2016 fluctuate with the yearly averages inside permissible annual MAAQG of 120 µg/m^3 .The average concentrations in Jerantut in 2015 is slightly increased as compared to period in 2007 – 2014 as shown in Figure 4.2 and decrease to 36.84 µg/ µg/m^3 for the year 2016. Overall the total average of ozone concentrations in Jerantut was levelled around 32.69 µg/m^3 Table 4.1.

Overall mean was 32.69 µg/m^3 which was below the MAAQG of 120 µg/m^3 for Jerantut. The highest peak of concentration was recorded in 2007. The analysis indicates the variability is low in the monitoring records since the coefficient of variation for all years are less than an equal to 0.38.

Table 4.1
Descriptive statistics for Jerantut

YEAR	Mean	Max	Min.	Std Dev.	Skew.	Kurt.	Coefficient of variations
2007	27.86	68	4	10.17	0.56	3.2	0.365
2008	27.89	56	1	10.44	0.22	2.42	0.374
2009	29.04	55	1	10.61	0.2	2.28	0.365
2010	28.53	56	2	9.21	0.23	2.58	0.323
2011	31.76	64	5	11.99	0.26	2.18	0.378
2012	32.29	71	1	12.34	0.39	2.53	0.382
2013	31.98	65	13	11.14	0.48	2.4	0.348
2014	37.53	69	9	11.66	-0.01	2.58	0.311
2015	43.02	83	14	14.51	0.24	2.46	0.337
2016	36.84	75	3	12.91	0.34	2.4	0.351
All	32.69	83	1	12.55	0.51	2.94	0.384

Table 4.2 the annual averages of concentration in Klang were recorded lower than 120 $\mu g/m^3$ through 2007 – 2016 which was below the MAAQG level of 120 μg/m3. However, by looking at the annual maximum shows that the fluctuation trend from 2007 – 2014 which are the concentration were below the MAAQG level of 120 $\mu g/m^3$ and the trend increased by 4% for the year 2015 and 13% for the year 2016 were showed that the concentrations of ozone for both years in Klang were above the MAAQG level of 120 $\mu g/m^3$.

Taken as whole, maximum concentrations were 135 $\mu g/m^3$ while average concentrations were 49.15 $\mu g/m^3$ in Klang. The analysis indicates low variability in the monitoring record every year with coefficient of variations value is 0.4.

Table 4.2
Descriptive statistics for Klang

YEAR	Mean	Max	Min.	Std Dev.	Skew.	Kurt.	Coefficient of variations
2007	40.94	114	7	18.23	1.28	5.26	0.45
2008	43.7	105	12	17.06	0.83	3.43	0.39
2009	51.71	125	5	19.93	0.6	3.34	0.39
2010	52.22	123	18	20.2	0.64	3.11	0.39
2011	48.27	108	10	18.71	0.7	3.27	0.39
2012	46.26	107	2	16.67	0.71	3.42	0.36
2013	51.41	116	12	19.17	0.72	3.61	0.37
2014	53.47	115	7	21.88	0.56	2.61	0.41
2015	52.24	120	3	18.98	0.1	3.31	0.36
2016	50.53	135	1	21.19	0.67	3.52	0.42
All	49.15	135	1	19.5	0.67	3.38	0.4

Table 4.3 showed, in all parts, mean for all years was 56.3 $\mu g/m^3$ which was below the MAAQG of 120 $\mu g/m^3$ in Putrajaya. The highest peak of concentration distribution was recorded in 2014. The analysis indicates the variability were low in the monitoring records since the coefficient of variation for all years were less than an equal to 0.35.

Table 4.3
Descriptive statistics for Putrajaya

YEAR	Mean	Max	Min.	Std Dev.	Skew.	Kurt.	Coefficient of variations
2007	58.96	123	6	20.3	0.61	3.36	0.34
2008	55.87	117	2	19	0.7	3.54	0.34
2009	54.97	120	11	17.03	0.73	4.41	0.31
2010	57.97	121	16	17.69	0.79	3.95	0.31
2011	52.74	107	2	19.1	0.34	2.99	0.36
2012	47.67	106	5	14.42	0.49	3.81	0.3
2013	56.95	115	7	19.54	0.41	2.96	0.34
2014	58.87	147	3	22.93	0.72	3.75	0.39
2015	63.72	121	4	20.54	0.18	2.75	0.32
2016	56.83	144	0	24.52	0.32	3.57	0.43
All	56.3	147	0	19.91	0.59	3.70	0.35

Total average for year 2007 - 2016 was 62.4 which was below the MAAQG of 120 $\mu g/m^3$ as showed in Table 4.4. The highest peak of concentration distribution was recorded in 2011. The analysis indicates the variability were low in the monitoring records since the coefficient of variation for all years were less than an equal to 0.38.

The annual averages of concentration in Shah Alam were recorded lower than 120 $\mu g/ m^3$through 2007 – 2016 which was below the MAAQG level of 120 $\mu g/ m^3$. However, the annual maximum concentrations showed an increasing trend from 2007 – 2011 and drastically drop below the MAAQG level in year 2012 and return above 120 $\mu g/ m^3$in year 2013 and onwards.

Table 4.4
Descriptive statistics for Shah Alam

YEAR	Mean	Max	Min.	Std Dev.	Skew.	Kurt.	Coefficient of variations
2007	64.61	132	16	22.2	0.54	2.97	0.34
2008	65.87	149	4	25.1	0.53	3.09	0.38
2009	66.39	145	15	24.51	0.65	3.20	0.37
2010	66.95	148	19	22.64	0.41	2.76	0.34
2011	60.56	166	5	23.29	0.83	4.47	0.38
2012	56.24	119	0	20.09	0.44	3.37	0.36
2013	51.89	131	1	21.78	0.49	4.00	0.42
2014	56.83	132	17	21.57	0.68	3.27	0.38
2015	50.09	145	0	24.67	0.55	3.83	0.49
2016	64.98	151	15	23.68	0.7	3.42	0.36
All	62.04	166	0	23.52	0.59	3.42	0.38

4.2.2 Trend of the concentrations

Figure 4.4 illustrates the trend of annual average daily maximum for Jerantut. The graph shows an increasing trend of the annual average of daily maximum during 2007 – 2016. The analysis of Mann-Kendall's (MK) trend indicate that (τ = 0.822, p – value = 0.0012822 < 0.05). There was significant evidence that the annual average of daily average of ozone concentrations increase during the period.

The trend of annual average daily maximum in Klang is demonstrated in Figure 4.4. The ozone concentrations in Klang were relatively low during 2007 – 2016 but shows increasing trend through the year. The analysis of MK's trend indicated that the increasing trend was not statistically significant (τ = 0.422, p – value = 0.1074).

Besides that, the trend of annual average of daily maximum in Putrajaya during the period of 2007 -2016. The decreasing of annual increasing of daily average as illustrated in the graph was not significant follows the analysis of MK's trend test (τ = 0.111, p – value = 0.72051). While, in Shah Alam the increasing trend of annual average of daily average showed during the year 2007 – 2016. However, there was not enough statistical evidence to indicate the significant inclination trend exist in Putrajaya (τ = -0.422, p – value =

0.1074).

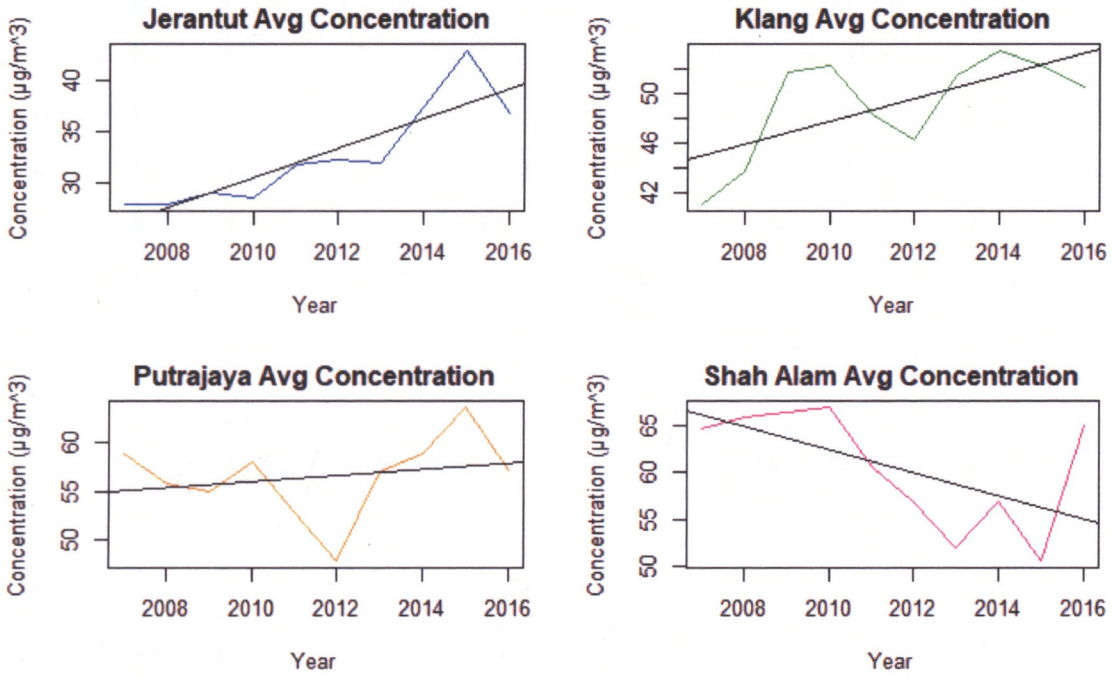


Figure 4.4 Trend of annual average daily maximum ozone concentrations at four location in Peninsular Malaysia

The analysis of trend of annual maximum in Jerantut during 2007 – 2016 as shown in Figure 4.5 indicates increasing trend. The analysis from MK's trend test (τ = 0.582, p – value = 0.024764) was proving that there was significant evidence that the annual average of daily maximum ozone concentrations increases during the period.

Other part of analysis is annual maximum in Klang is demonstrated in Figure 4.5. The trend in Klang was slightly increase during the period of 2007 – 2016 in comparison with the annual average of daily maximum. Meanwhile, the increasing trend was not significantly significant as indicated in the MK's trend analysis (τ = 0.289, p – value = 0.28313).

As depicted in Figure 4.5, the annual maximum was almost all year above the annual MAAQG of 120 µg/m3. Through the graph shows an increasing trend in the annual maximum during the period of 2007 – 2016 in Putrajaya, the trend was not statistically

proven using the MK's trend test to conclude the downward trend during this period Putrajaya (τ = 0.135, p – value = 0.65342).

Figure 4.5 shows another perspective of observing the trend of annual maximum. Various factors such as the industrial activities, vehicle emission and open burning contributed to the high concentrations in Shah Alam particularly in 2010 and 2011. Through the linear equation of the concentrations indicates the decreasing trend, the analysis of MK's trend test showed that it was not enough evidence statistically to conclude the downward trend in Shah Alam during the year 2007 – 2016 (τ = 0.0682, p – value = 0.85689).

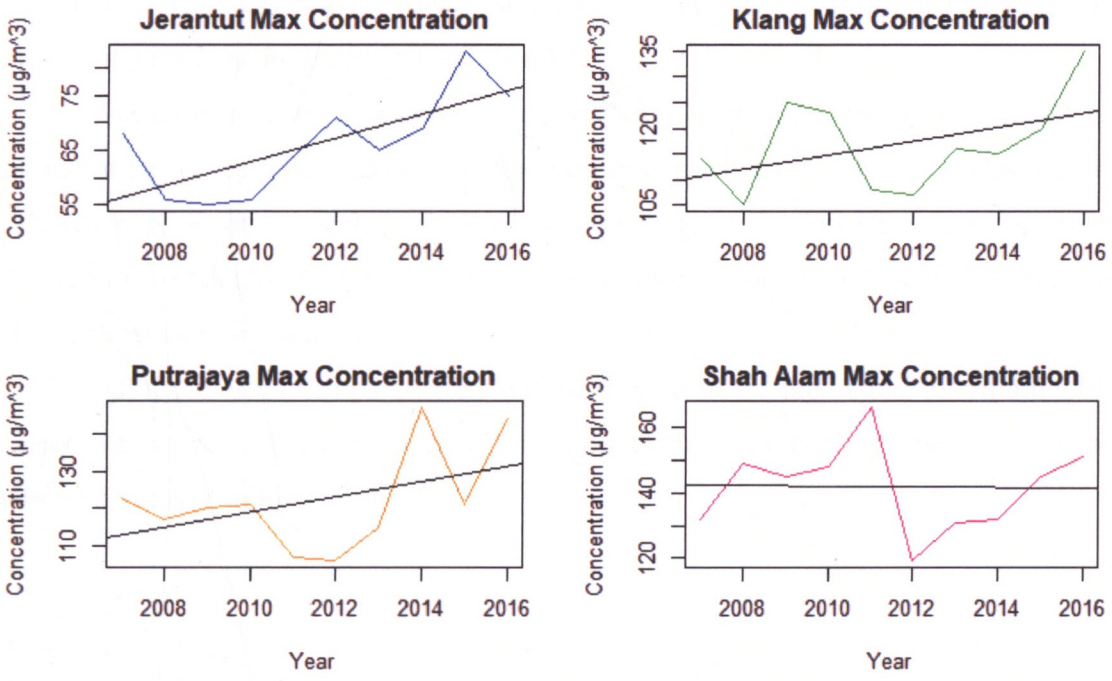


Figure 4.5 Trend of annual maximum ozone concentrations at four location in Peninsular Malaysia

4.3 Fitting of extreme value distribution (EVD)

The best EVD can be conclude after a few phases involve in Objective two. MOM and MLE had been used to estimate the parameter for each of data selection – Daily Maximum. After the estimation parameter, the analysis of six performance indicators was

done for methods, distributions and stations. The performance indicators were then ranked to attain the best distribution for each station according to data selection.

4.3.1 Parameter estimation

The location parameter, μ is generally the mean or level of concentration while the scale, σ parameter is relative to the sources emission strength. According to Lu and Wang (2006) the higher the value of scale parameter, the more polluted the monitoring station would be Shape parameter, λ, is determined by the meteorological factors independently of the sources strength of the pollutant. The estimation of location, scale and shape parameters using different methods for different data selection and distributions for all monitoring stations are listed in the Table 4.5.

The shape and scale for two-parameter Weibull were higher in Putrajaya using MLE estimator as compared to the other method estimator, MOM. Different estimators for the estimation of location, scale and shape parameter gave the lowest values for all three distributions in Jerantut while the other two monitoring locations Klang and Shah Alam gave almost similar values.

All these estimated parameters will be used to fit the theoretical distribution of the respective monitoring stations. The distribution of the theoretical against the observation will be discussed in Section 4.3.2.

Table 4.5
Parameter estimation for all selected location

		2- parameter Weibull		3 – parameter Weibull			3 – parameter GEV		
Monitoring Locations	Method	σ	λ	μ	σ	λ	μ	σ	λ
Jerantut	MLE	36.69	2.84	7.24	28.73	2.13	27.35	11.22	.01
	MOM	36.69	2.82	32.08	2.25	0.21	27.48	11.29	0.13
Klang	MLE	55.21	2.78	13.58	40.09	1.90	40.60	16.62	0.07
	MOM	55.25	2.73	48.56	5.61	0.19	40.82	16.93	0.09
Putrajaya	MLE	62.90	3.19	18.06	43.20	2.05	47.93	17.22	0.10
	MOM	62.99	3.09	55.69	8.51	0.19	47.93	17.48	0.11
Shah Alam	MLE	69.65	2.95	18.33	49.42	1.96	52.02	20.15	0.08
	MOM	69.71	2.89	61.53	11.03	0.18	52.23	20.54	0.10

4.3.2 Performance indicator

Following the estimation of parameters, performance indicator for each of distribution using different methods was obtained. The performance indicator was compared and then ranked to obtain the best distribution to represent the monitoring stations.

Table 4.6 and Table 4.7 presents the performance indicators for daily maximum of ozone concentration in Jerantut, Klang, Putrajaya and Shah Alam and the best model was selected based on the smallest error measures and the largest accuracy measures as highlighted in bold. For MOM data selection for all monitoring stations, the best distribution with the best performance indicators were the GEV. On the other hand, for MLE, the best distribution which the best performance indicators was the two parameter Weibull.

4.3.3 The best distribution

Table 4.8 lists the best distribution for every monitory station for every records selection. The findings show that the best distributions same for different stations and different methods of data selections since different method of data selection produce different sample size available for the analysis.

Table 4.6
Performance indicators for daily maximum of ozone concentrations for Jerantut and Klang

Locations		Distribution	PERFORMANCE INDICATORS					
			NAE	PA	RMSE	R^2	IA	MAE
Jerantut	MOM	2Weibull	1.5589	0.987	52.307	0.973	0.383	50.941
		3Weibull	0.376	0.942	12.783	0.888	0.809	12.278
		GEV	0.092	0.937	8.386	0.878	0.924	3.021
	MLE	2Weibull	0.027	0.995	1.189	0.995	0.998	0.874
		3Weibull	0.376	0.942	12.783	0.887	0.809	12.278
		GEV	0.079	0.949	7.233	0.810	0.941	2.581
Klang	MOM	2Weibull	0.702	0.982	35.025	0.964	0.641	34.480
		3Weibull	0.590	0.954	31.655	0.910	0.705	28.992
		GEV	0.066	0.965	9.050	0.930	0.959	3.244
	MLE	2Weibull	0.036	0.993	2.262	0.986	0.997	1.789
		3Weibull	0.307	0.954	15.811	0.910	0.868	15.079
		GEV	0.574	0.967	22.310	0.934	0.706	18.777

Table 4.7
Performance indicators for daily maximum of ozone concentrations for Putrajaya and Shah Alam

Locations		Distribution	PERFORMANCE INDICATORS					
			NAE	PA	RMSE	R^2	IA	MAE
Putrajaya	MOM	2Weibull	0.484	0.985	27.875	0.970	0.727	27.293
		3Weibull	0.566	-0.942	25.663	0.887	0.081	18.491
		GEV	0.072	0.958	10.593	0.918	0.949	4.0479
	MLE	2Weibull	0.037	0.990	2.766	0.980	0.995	2.087
		3Weibull	0.315	0.950	18.420	0.902	0.833	17.746
		GEV	0.064	0.963	9.508	0.928	0.957	3.625
Shah Alam	MOM	2Weibull	0.346	0.985	21.896	0.969	0.831	21.484
		3Weibull	0.978	0.951	72.971	0.903	0.528	60.772
		GEV	0.069	0.961	11.751	0.923	0.953	4.275
	MLE	2Weibull	0.037	0.993	2.801	0.985	0.996	2.276
		3Weibull	0.311	0.951	20.187	0.903	0.853	19.322
		GEV	0.054	0.969	9.626	0.9389	0.967	3.345

4.3.4 Daily maximum

The best EVD for all the monitoring stations were the two parameter Weibull MLE.

Table 4.8
The best distribution for each location and data selection

Estimator	MONITORING STATIONS			
	JERANTUT	KLANG	PUTRAJAYA	SHAH ALAM
n	3542	2985	3448	3056
MOM	GEV	GEV	GEV	GEV
MLE	2W	2W	2W	2W
OVERALL	2W	2W	2W	2W
	mle	mle	mle	mle

All the estimated parameters were used to fit the theoretical distribution to compare with the distribution of the observations. Figure 4.6, Figure 4.7, Figure 4.8 and Figure 4.9 illustrates the probability density function (pdf) of concentrations in all monitoring stations. The pdf plot using the two parameter Weibull MLE was almost identical to the observations in all monitoring station. The GEV probability distribution plot clearly not a good fit to the observation in all monitoring station.

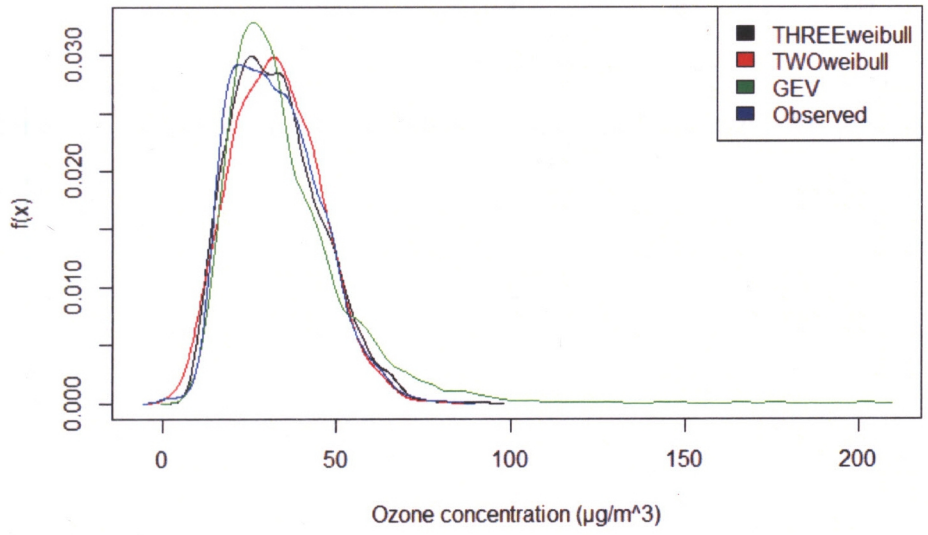


Figure 4.6 Probability density function using daily maximum for Jerantut station

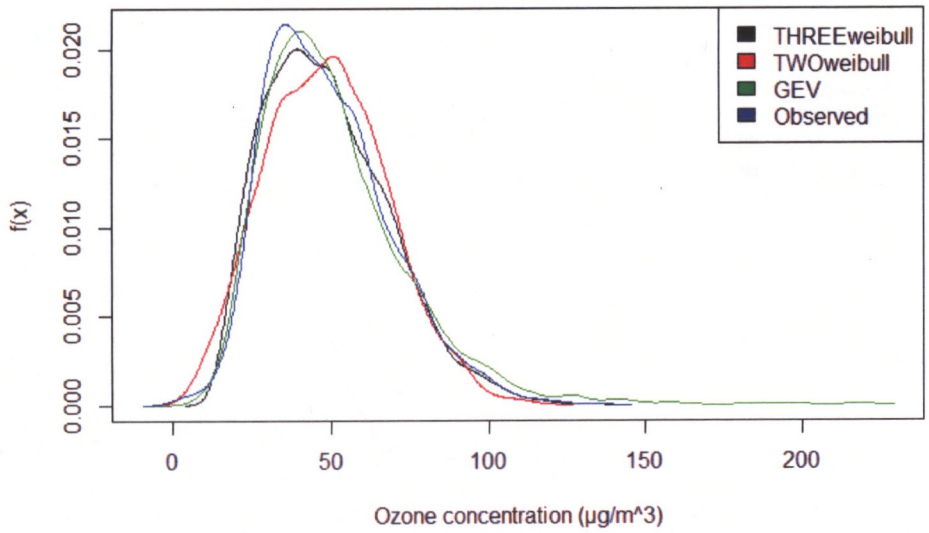


Figure 4.7 Probability density function using daily maximum for Klang station

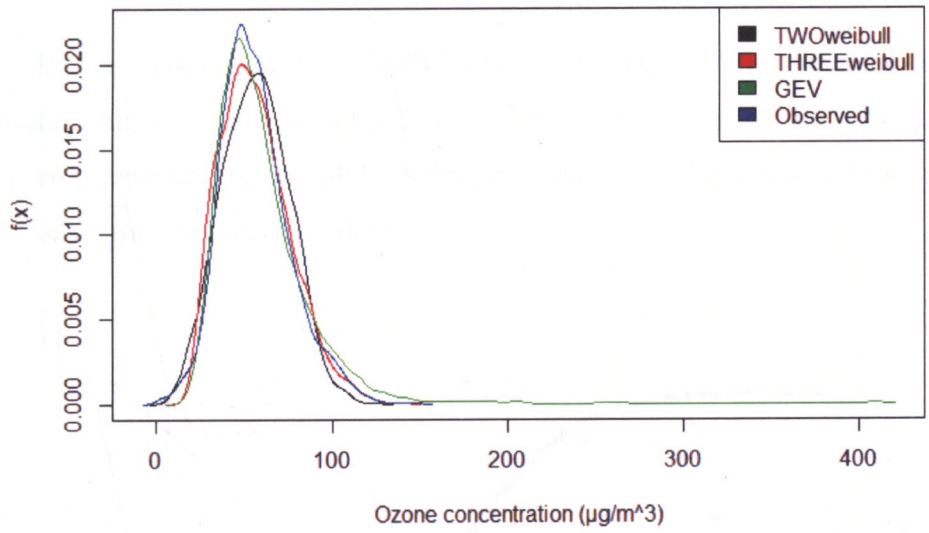


Figure 4.8 Probability density function using daily maximum for Purajaya station

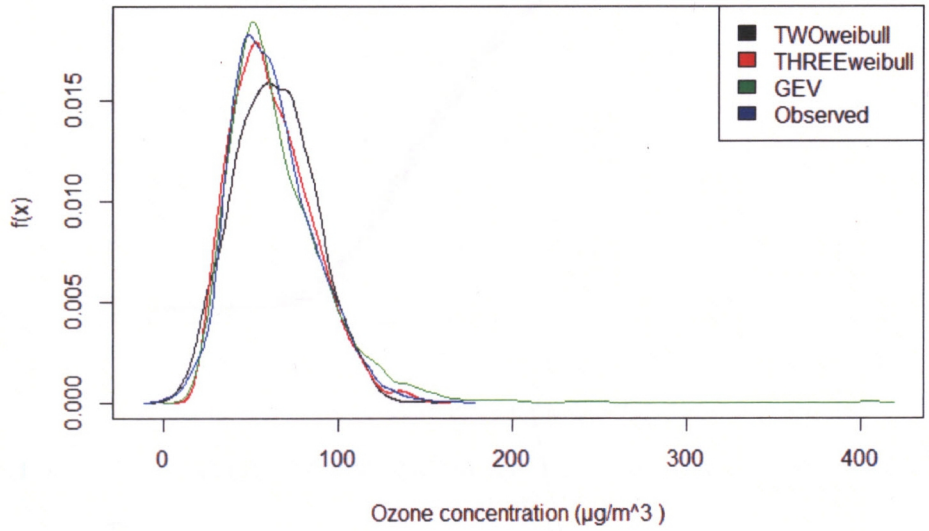


Figure 4.9 Probability density function using daily maximum for Shah Alam station

Figure 4.10, Figure 4.11, Figure 4.12 and Figure 4.13 illustrates the cumulative distribution function (cdf) of concentrations in all monitoring stations. The plot of cdf were used to estimate the probabilities and exceedances of the concentrations above 120 µg/ $m^3$for all the monitoring stations.

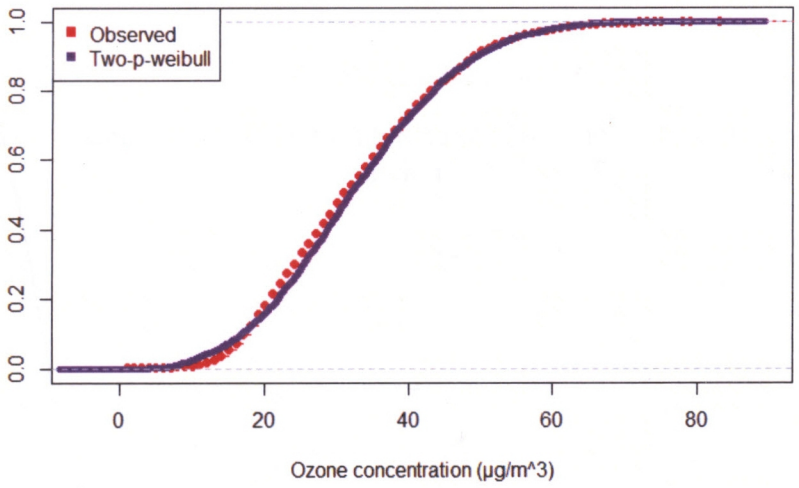


Figure 4.10 Cumulative distribution function using daily maximum for Jerantut station

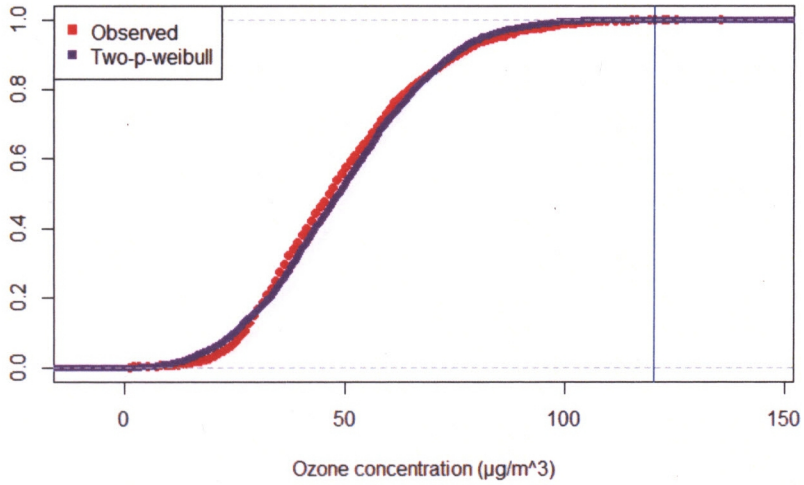


Figure 4.11 Cumulative distribution function using daily maximum for Klang station

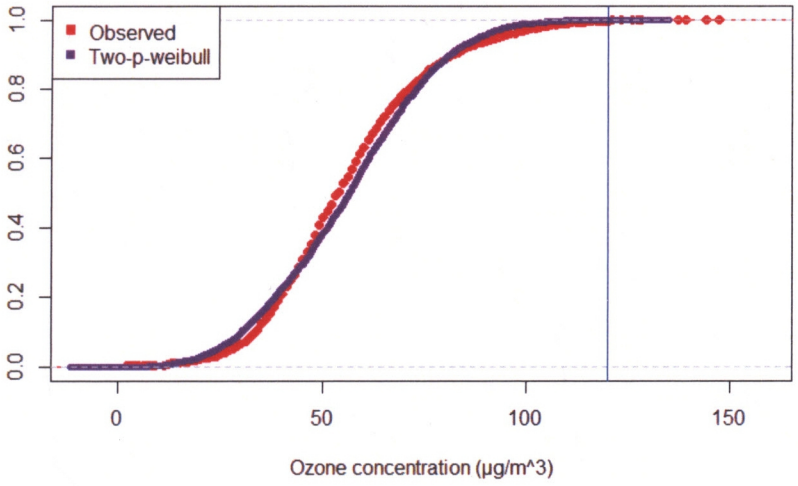


Figure 4.12 Cumulative distribution function using daily maximum for Putrajaya station

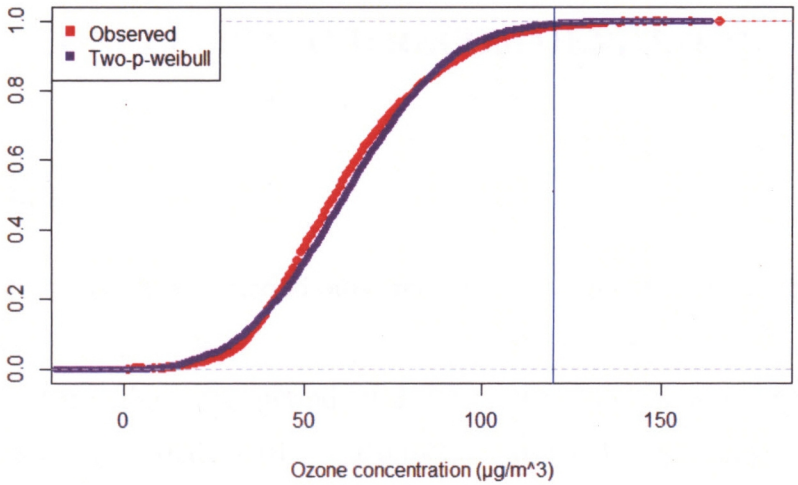


Figure 4.13 Cumulative distribution function using daily maximum for Shah Alam station

CHAPTER FIVE
CONCLUSION AND RECOMMENDATION

5.1 Conclusion

The study which was carried out three strategic objectives can be concluded as follows:

i. Throughout the period of 2000 – 2012, the highest level of ozone concentrations was recorded in 2011. The only least affected monitoring station was the Putrajaya which was located in rural area of Peninsular Malaysia and also Jerantut as a background station.

ii. Based on the parameter estimation indicated that the two parameter Weibull has the higher estimated value compare to the other distribution. The parameter scale and the shape of the distribution from the estimation parameter is significantly supporting the performance indicator to determine the best distribution for each monitoring locations.

iii. The EVD model with the best performance indicators for Jerantut, Klang, Putrajaya and Shah Alam was the two-parameter Weibull with the MLE parameter estimator.

5.2 Limitations

i. The study only limited to four monitoring stations, hence, the proposed distribution only restricted for the respective monitoring location. There is no specific model that best fit for all locations.

ii. The access of new records must be made to be had to researchers in order that the cutting-edge and up-to –date version may be proposed for the prediction of the extreme exceedances at some point of Malaysia considering the occurrence of excessive particulate activities that constantly occurs each 12 months in Malaysia.

v. The records series for the reason of verification cannot be attended to the alternative monitoring stations in this look at. Thus, the verification of the model acquired within the have a look at handiest confined to one collection simplest.

5.3 Recommendations

The Malaysian Environmental Quality Report posted every year, by way of the Department of Environment Ministry of Natural Resources and Environmental, Malaysia report that the floor stage ozone is still the main pollutant of difficulty in Malaysia. Hence, the availability of appropriate statistical model in predicting future exceedances of awareness avoid the MAAQG restrict could beneficial for the environmentalist and strategists to devise proper movement and controlling strategies to triumph over the trouble.

REFERENCES

Afroz, R., Hassan, M. N., & Ibrahim, N. A. (2003). Review of air pollution and health impacts in Malaysia. *Environmental Research, 92*(2), 71-77. doi:10.1016/s0013-9351(02)00059-2

Ahmat, H., & Yahaya, A. S. (2018). The analysis of PM10 concentrations using the generalized extreme value (GEV) and generalized pareto distribution (GPD) in the Bayesian approach. *AIP Conference Proceedings, 1974*(1), 040019. doi:10.1063/1.5041693

Ahmat, H., Yahaya, A. S., & Ramli, N. A. (2016). Prediction Of Pm10 Concentrations Using Extreme Value Distribution (EVD): Clasical And Bayesian Approach. *ESTEEM Academic Journal, 12*(1), 1-10.

Ahmat, H., Yahaya, A. S., Ramli, N. A., Japeri, A. Z. u.-S. M., & Hamid, H. A. (2015). Analysis of PM10 Using Extreme Value Theory. *ESTEEM Academic Journal, 11*(1), 135-143.

Aryal, G. R., & Tsokos, C. P. (2009). On the transmuted extreme value distribution with application. *Nonlinear Analysis: Theory, Methods & Applications, 71*(12), e1401-e1407. doi:10.1016/j.na.2009.01.168

Awang, N. R., Ramli, N. A., Mohammed, N. I., & Yahaya, A. S. (2013). Time series evaluation of ozone concentration in malaysia based on location of monitoring station. *International Journal of Engineering and Technology, 3*(3).

Azam, M., Mahmudul Alam, M., & Haroon Hafeez, M. (2018). Effect of tourism on environmental pollution: Further evidence from Malaysia, Singapore and Thailand. *Journal of Cleaner Production, 190*, 330-338. doi:10.1016/j.jclepro.2018.04.168

Bali, T. G. (2003). The generalized extreme value distribution. *Economics Letters, 79*(3), 423-427. doi:10.1016/s0165-1765(03)00035-1

Banan, N., Latif, M. T., & Juneng, L. (2013). An Assessment of Ozone Levels in Typical Urban Areas in the Malaysian Peninsular. *International Journal of Environmental and Ecological Engineering, 7*(2).

Bekhet, H. A., & Othman, N. S. (2017). Impact of urbanization growth on Malaysia CO2 emissions: Evidence from the dynamic relationship. *Journal of Cleaner Production, 154*, 374-388. doi:https://doi.org/10.1016/j.jclepro.2017.03.174

Bian, J., Gettelman, A., Chen, H., & Pan, L. L. (2007). Validation of satellite ozone profile retrievals using Beijing ozonesonde data. *Journal of Geophysical Research, 112*(D6). doi:10.1029/2006jd007502

Bracher, A., Lamsal, L. N., Weber, M., Bramstedt, K., Coldewey-Egbers, M., & Burrows, J. P. (2005). Global satellite validation of SCIAMACHY O^3 columns with GOME WFDOAS. *Atmospheric Chemistry and Physics, 5*, 2357–2368.

Bury, K. (1999). *Statistical Distributions in Engineering*. Cambridge: Cambridge University Press.

Chattopadhyay, G., Chakraborthy, P., & Chattopadhyay, S. (2012). Mann–Kendall trend analysis of tropospheric ozone and its modeling using ARIMA. *Theoretical and Applied Climatology, 110*(3), 321-328. doi:10.1007/s00704-012-0617-y

Chung, E.-S., & Kim, S. U. (2013). Bayesian rainfall frequency analysis with extreme value using the informative prior distribution. *KSCE Journal of Civil Engineering, 17*(6), 1502-1514. doi:10.1007/s12205-013-0189-0

Datsiou, K. C., & Overend, M. (2018). Weibull parameter estimation and goodness-of-fit for glass strength data. *Structural Safety, 73*, 29-41. doi:10.1016/j.strusafe.2018.02.002

Duenas, C., Fernandez, M. C., S.Canete, Carretero, J., & Liger, E. (2004). Analyses of ozone in urban and rural sites in Malaga (Spain). *Chemosphere, 56*, 631–639. doi:10.1016/j.chemosphere.2004.04.013

Finkenstadt, B., & Rootzen, H. (2003). *Extreme Values in Finance, Telecommunications, and the Environment*: Taylor & Francis.

Fisher, M. J., & Marshall, A. P. (2009). Understanding descriptive statistics. *Australian Critical Care, 22*(2), 93-97. doi:https://doi.org/10.1016/j.aucc.2008.11.003

Fisher, R. A., & Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Mathematical Proceedings of the Cambridge Philosophical Society, 24*(2), 180-190. doi:10.1017/S0305004100015681

Ghazali, N. A., Ramli, N. A., Yahaya, A. S., Yusof, N. F. F. M., Sansuddin, N., & Al Madhoun, W. A. (2010). Transformation of nitrogen dioxide into ozone and prediction of ozone concentrations using multiple linear regression techniques. *Environmental Monitoring and Assessment, 165*(1), 475-489. doi:10.1007/s10661-009-0960-3

Glavas, S. (1999). Surface ozone and NOx concentrations at a high altitude Mediterranean site, Greece. *Atmospheric Environment, 33*(23), 3813-3820. doi:https://doi.org/10.1016/S1352-2310(98)00393-8

Guarnieri, M., & Balmes, J. R. (2014). Outdoor air pollution and asthma. *Lancet, 383*(9928), 1581-1592. doi:10.1016/S0140-6736(14)60617-6

Gwak, W., Goo, H., Choi, Y. H., & Ahn, J. Y. (2016). Extreme value theory in mixture distributions and a statistical method to control the possible bias. *Journal of the Korean Statistical Society, 45*(4), 581-594. doi:10.1016/j.jkss.2016.04.003

H. Seinfeld, J., & Pandis, S. (1998). *Atmospheric Chemistry and Physics: From Air Pollution to Climate Change* (Vol. 51).

Hirsch, R. M., J. R. Slack, and R. A. Smith. (1982). Techniques of trend analysis for monthly water quality data. *Water Resources Research, 18*(1), 107-121. doi:doi:10.1029/WR018i001p00107

I. Elbatal, G. A., A Vincent Raja. (2014). Transmuted Exponentiated Frechet Distribution: ˆ

Properties and Applications. *Journal of Statistics Applications & Probability, 3*, 379-394. doi:10.12785/jsap/030309

Jasim, M.R, Tan, K.C, Lim, H.s, . . . M.Z. (2011). Investigation on the Carbon Monoxide Pollution over Peninsular Malaysia Caused by Indonesia Forest Fires from AIRS Daily Measurement in Advanced Air Pollution.pdf. 115-137.

Ji, L., & Gallo, K. (2006). *An Agreement Coefficient for Image Comparison* (Vol. 73).

Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J., & Kolehmainen, M. (2004). Methods for imputation of missing values in air quality data sets. *Atmospheric Environment, 38*(18), 2895-2907. doi:10.1016/j.atmosenv.2004.02.026

Kim, S., Shin, H., Joo, K., & Heo, J.-H. (2012). Development of plotting position for the general extreme value distribution. *Journal of Hydrology, 475*, 259-269. doi:10.1016/j.jhydrol.2012.09.055

Latif, M. T., Othman, M., Idris, N., Juneng, L., Abdullah, A. M., Hamzah, W. P., . . . Jaafar, A. B. (2018). Impact of regional haze towards air quality in Malaysia: A review. *Atmospheric Environment, 177*, 28-44. doi:10.1016/j.atmosenv.2018.01.002

Leila Droprinchinski Martins, Caroline Fernanda Hei Wikuats, Mauricio Nonato Capucim, Daniela S. de Almeida , Silvano Cesar da Costa, Taciana Albuquerque , . . . Martins, J. A. (2017). Extreme value analysis of air pollution data and their comparison between two large urban regions of South America. *Weather and Climate Extremes, 18*, 44-54.

Lu, W. Z., & Wang, X. K. (2006). Evolving trend and self-similarity of ozone pollution in central Hong Kong ambient during 1984-2002. *Sci Total Environ, 357*(1-3), 160-168. doi:10.1016/j.scitotenv.2005.03.015

Malaysia, D. o. E. (2015). *Malaysia Environmental Quality Report 2015*.

Martins, & Stedinger, J. R. (2000). Generalized maximum-likelihood generalized extreme-value quantile estimators for hydrologic data. *Water Resources Research, 36*(3), 737–744.

Marzano, V. (2014). A simple procedure for the calculation of the covariances of any Generalized Extreme Value model. *Transportation Research Part B: Methodological, 70*, 151-162. doi:10.1016/j.trb.2014.08.011

Millington, N., Das, S., & Simonovic, S. P. (2011). The Comparison of GEV Log-Pearson Type 3 and Gumbel Distribution in the Upper Thames River Watershed under Global Climate Models.

Muhammad Ismail Jaffar, Hazrul Abdul Hamid, Riduan Yunus, & Raffee, A. F. (2018). Fitting Statistical Distribution on Air Pollution: an Overview. *International Journal of Engineering & Technology, 7*, 40-44. doi:10.14419/ijet.v7i3.23.17256

N, N. M., Abdullah, M. M. A., Tan, C.-y., Ramli, N. A., Yahaya, A. S., & Fitri, N. F. M. Y. (2011). Modelling of PM10 concentration for industrialized area in Malaysia: A case study in Shah Alam. *Physics Procedia, 22*, 318-324. doi:10.1016/j.phpro.2011.11.050

Örkcü, H. H., Aksoy, E. r., & Dog˘an, M. İ. (2015). Estimating the parameters of 3-p Weibull distribution through differential evolution. *Applied Mathematics and Computation, 251*, 211-224. doi:10.1016/j.amc.2014.10.127

Othman, J., Sahani, M., Mahmud, M., & Ahmad, M. K. (2014). Transboundary smoke haze pollution in Malaysia: inpatient health impacts and economic valuation. *Environ Pollut, 189*, 194-201. doi:10.1016/j.envpol.2014.03.010

Ozay, C., & Celiktas, M. S. (2016). Statistical analysis of wind speed using two-parameter Weibull distribution in Alaçatı region. *Energy Conversion and Management, 121*, 49-54. doi:10.1016/j.enconman.2016.05.026

Rinne, H. (2008). The Weibull Distribution. New York: Chapman and Hall/CRC.

Rinner, C., & Hussain, M. (2011). *Toronto's Urban Heat Island—Exploring the Relationship between Land Use and Surface Temperature* (Vol. 3).

Samuel, K., & Saralees, N. (2000). *Extreme Value Distributions*: World Scientific Publishing Company.

Seal, C. K., & Sherry, A. H. (2016). Weibull distribution of brittle failures in the transition region. *Procedia Structural Integrity, 2*, 1668-1675. doi:10.1016/j.prostr.2016.06.211

Shan, W., Yin, Y., Zhang, J., & Ding, Y. (2008). Observational study of surface ozone at an urban site in East China. *Atmospheric Research, 89*(3), 252-261. doi:10.1016/j.atmosres.2008.02.014

Sharma, A. P., Kim, K. H., Ahn, J. w., Shon, Z. H., Sohn, J. R., Lee, J. H., . . . Brown, R. J. C. (2014). Ambient particulate matter (PM10) concentrations in major urban areasof Korea during 1996–2010. *Atmospheric Pollution Research, 5*, 161-169. doi:10.5094/APR.2014.020

Svensson, C., Clarke, R. T., & Jones, D. A. (2007). An experimental comparison of methods for estimating rainfall intensity-duration-frequency relations from fragmentary records. *Journal of Hydrology, 341*(1-2), 79-89. doi:10.1016/j.jhydrol.2007.05.002

Varshney, C. K., & Sigh, A. P. (2003). Passive samplers for NOx monitoring: A Critical Review.pdf. *The Environmentalist, 23*, 127–136.

Wais, P. (2017). Two and three-parameter Weibull distribution in available wind power analysis. *Renewable Energy, 103*, 15-29. doi:10.1016/j.renene.2016.10.041

Yahaya, A. S., Ramli, N. A., Ul-Saufie, A. Z., Hamid, H. A., Ahmat, H., & Mohtar, Z. A. (2013). Predicting CO Concentrations Levels Using Probability Distributions. *International Journal of Engineering and Technology, 3*(3).

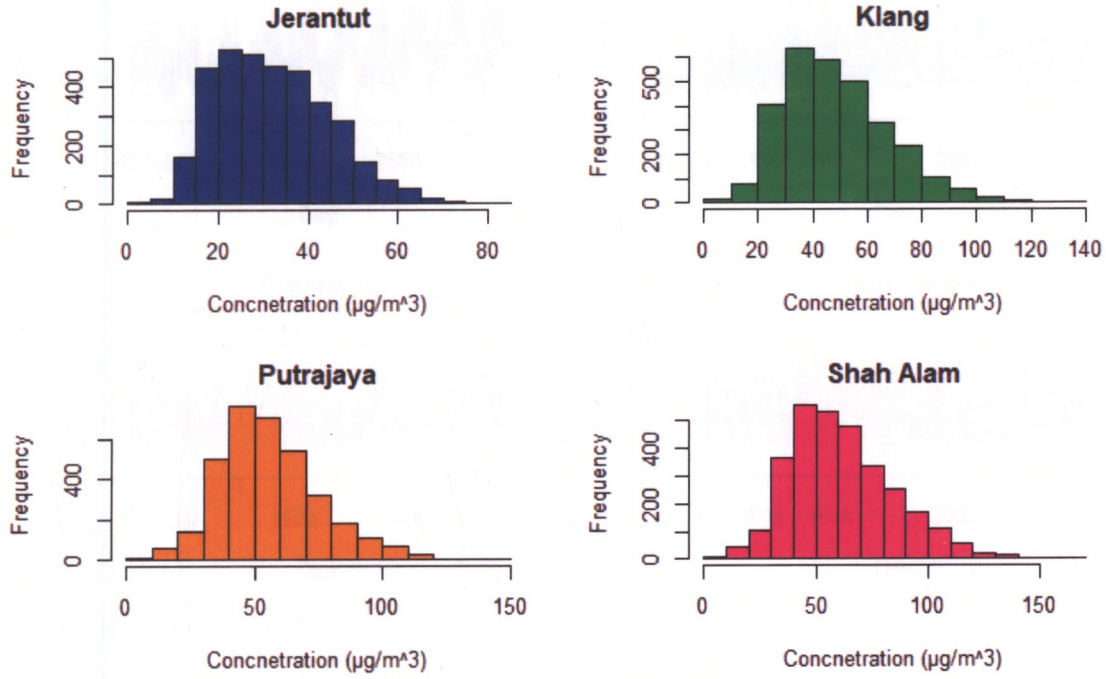
APPENDICES

APPENDIX 1

Example coding for characteristics of ozone concentration and MK trend analysis

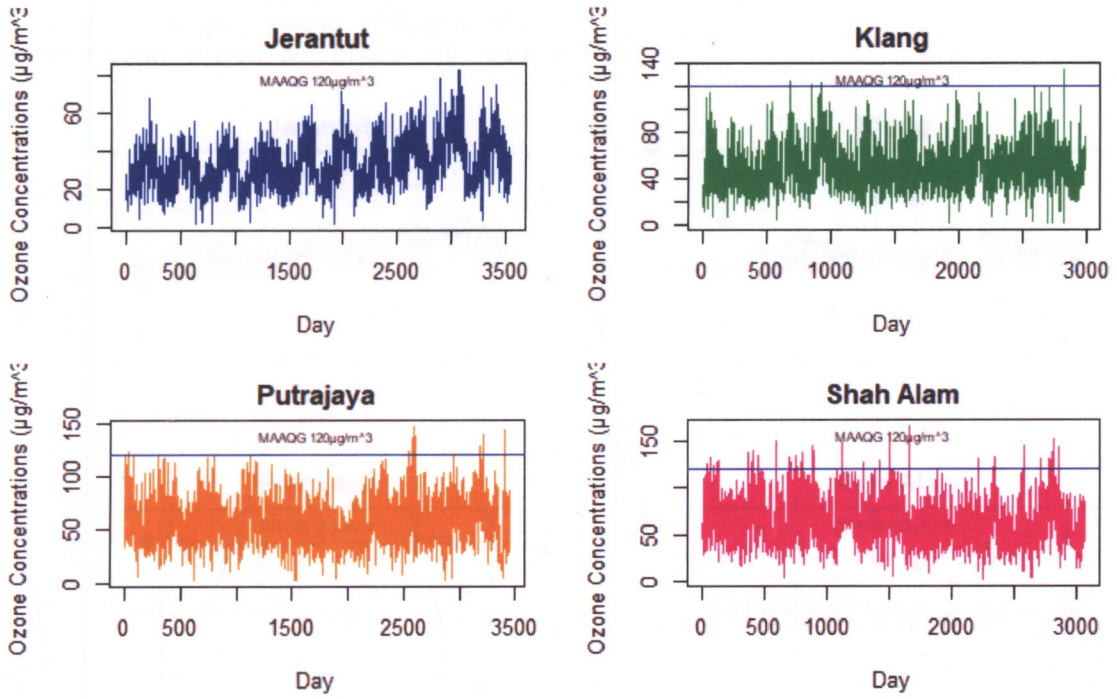
```
> library(tidyverse)
> library(plotly)
> library(ggplot2)
> library(ggfortify)
> myfile1 <- 'E:/Along punya fail/Master Applied Statistics/2 - RP STA7
98/My Proposal/4. Data/R/csv file/Jerantut.csv'
> Jerantut <- read.csv(myfile1)
> head(Jerantut)
       id     date year time O3_um3 O3concppm_max
1 Jerantut 1/1/2007 2007    0     26         0.026
2 Jerantut 2/1/2007 2007    0     28         0.028
3 Jerantut 3/1/2007 2007    0     26         0.026
4 Jerantut 4/1/2007 2007    0     20         0.020
5 Jerantut 5/1/2007 2007    0     19         0.019
6 Jerantut 6/1/2007 2007    0     21         0.021
> myfile2 <- 'E:/Along punya fail/Master Applied Statistics/2 - RP STA7
98/My Proposal/4. Data/R/csv file/Kelang.csv'
> Klang <- read.csv(myfile2)
> head(Klang)
      id     date year time O3_um3 O3concppm_max
1 Kelang 1/1/2007 2007    0     29         0.029
2 Kelang 2/1/2007 2007    0     35         0.035
3 Kelang 3/1/2007 2007    0     32         0.032
4 Kelang 4/1/2007 2007    0     27         0.027
5 Kelang 5/1/2007 2007    0     16         0.016
6 Kelang 6/1/2007 2007    0     16         0.016
> myfile3 <- 'E:/Along punya fail/Master Applied Statistics/2 - RP STA7
98/My Proposal/4. Data/R/csv file/Putrajaya.csv'
> Putrajaya <- read.csv(myfile3)
> head(Putrajaya)
       id      date year time O3_um3 O3concppm_max
1 Putajaya 30/1/2007 2007    0     40         0.040
2 Putajaya 31/1/2007 2007    0    120         0.120
3 Putajaya  1/2/2007 2007    0     34         0.034
4 Putajaya  2/2/2007 2007    0     50         0.050
5 Putajaya  3/2/2007 2007    0     39         0.039
6 Putajaya  4/2/2007 2007    0     56         0.056
> myfile4 <- 'E:/Along punya fail/Master Applied Statistics/2 - RP STA7
98/My Proposal/4. Data/R/csv file/Shah Alam.csv'
> Shah_Alam <- read.csv(myfile4)
> head(Shah_Alam)
        id     date year time O3_um3 O3concppm_max
1 Shah Alam 1/1/2007 2007    0     50         0.050
2 Shah Alam 2/1/2007 2007    0     51         0.051
3 Shah Alam 3/1/2007 2007    0     48         0.048
4 Shah Alam 4/1/2007 2007    0     46         0.046
5 Shah Alam 5/1/2007 2007    0     61         0.061
6 Shah Alam 6/1/2007 2007    0     53         0.053
> par(mfrow=c(2,2))
```

```
> hist (Jerantut$O3_um3, xlab = "Concnetration (µg/m^3)", col = "blue",
        main="Jerantut")
> hist (Klang$O3_um3, xlab = "Concnetration (µg/m^3)", col = "green",
        main="Klang")
> hist (Putrajaya$O3_um3, xlab = "Concnetration (µg/m^3)", col = "orang
e",     main="Putrajaya")
> hist (Shah_Alam$O3_um3, xlab = "Concnetration (µg/m^3)", col = "magen
ta",    main="Shah Alam")
```



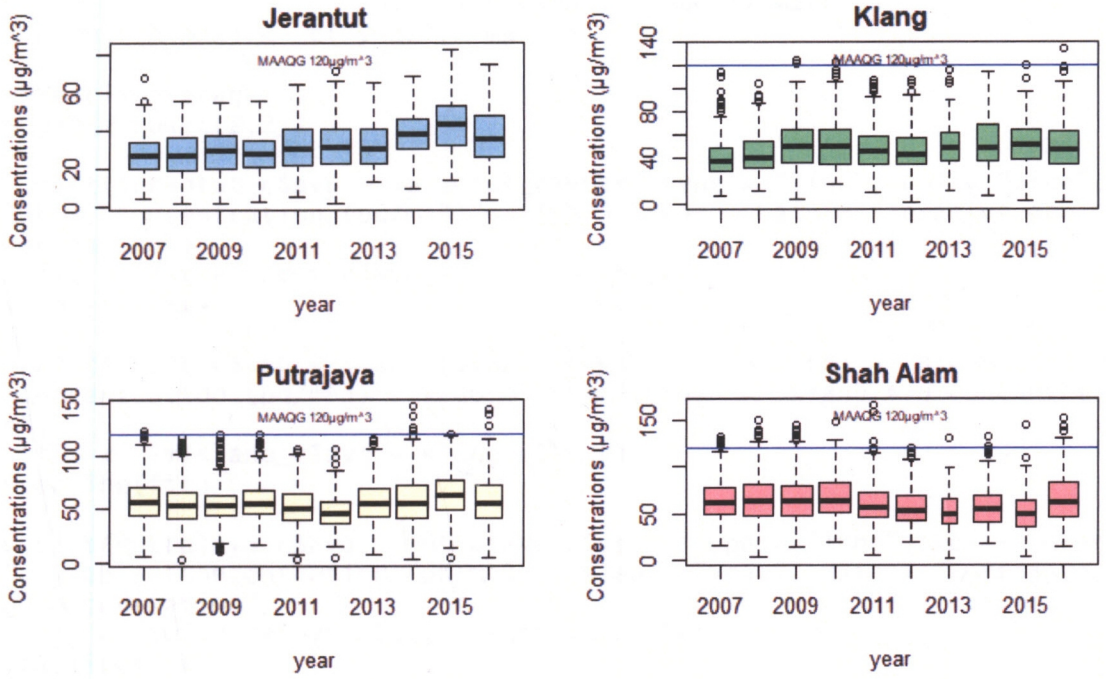

```
> par(mfrow=c(2,2))
>
> plot(Jerantut$O3_um3, type = "line", col = "blue", xlab = "Day", ylab
= "Ozone Concentrations (µg/m^3) ", main = "Jerantut")
> abline(h = 120, col=c("blue", "red"), lty=c(1,2), lwd=c(1, 3))
> eq <- paste0("MAAQG 120µg/m^3")
> mtext(eq,side= 3, line= -1, cex = 0.5)
>
> plot(Klang$O3_um3, type = "line", col = "green", xlab = "Day", ylab =
"Ozone Concentrations (µg/m^3) ", main = "Klang")
> abline(h = 120, col=c("blue", "red"), lty=c(1,2), lwd=c(1, 3))
> eq <- paste0("MAAQG 120µg/m^3")
> mtext(eq,side= 3, line= -1, cex = 0.5)
>
> plot(Putrajaya$O3_um3, type = "line", col = "orange", xlab = "Day", y
lab = "Ozone Concentrations (µg/m^3)", main = "Putrajaya")
> abline(h = 120, col=c("blue", "red"), lty=c(1,2), lwd=c(1, 3))
> eq <- paste0("MAAQG 120µg/m^3")
> mtext(eq,side= 3, line= -1, cex = 0.5)
>
> plot(Shah_Alam$O3_um3, type = "line", col = "magenta", xlab = "Day",
ylab = "Ozone Concentrations (µg/m^3)", main = "Shah Alam")
> abline(h = 120, col=c("blue", "red"), lty=c(1,2), lwd=c(1, 3))
```

```
> eq <- paste0("MAAQG 120µg/m^3")
> mtext(eq,side= 3, line= -1, cex = 0.5)
```



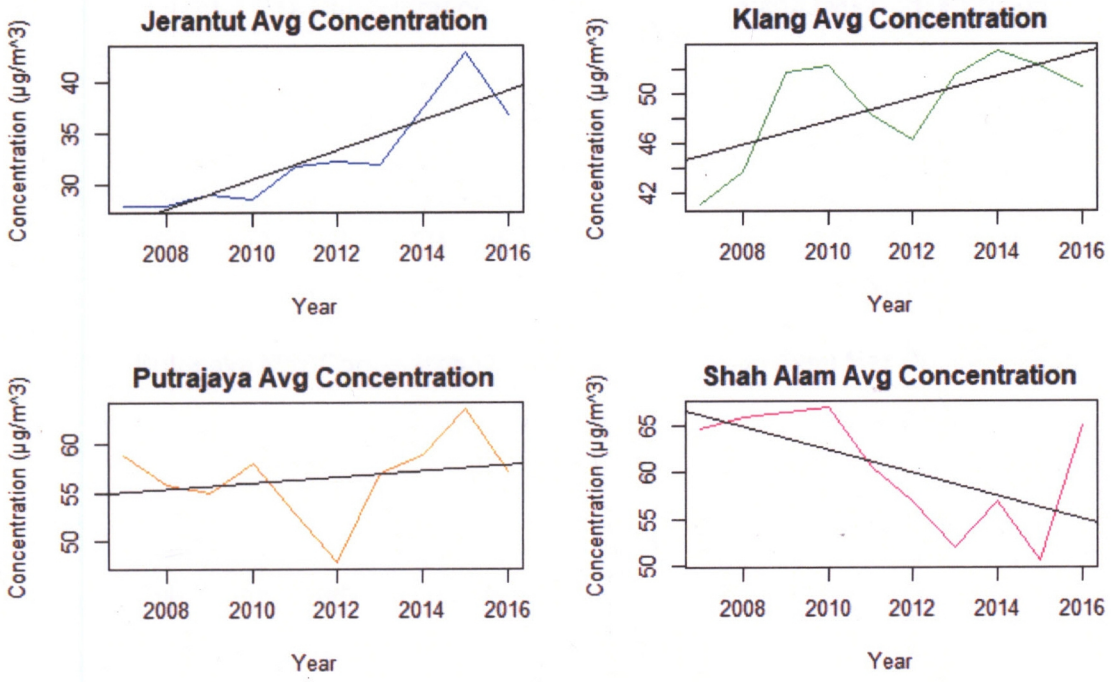

```
> par(mfrow=c(2,2))
>
> boxplot(Jerantut$O3_um3 ~ Jerantut$year, data = mydata, col = "lightb
lue", varwidth = TRUE,
+         ylab = "Consentrations (µg/m^3) ",xlab = "year", main = "Jeran
tut")
> abline(h = 120, col=c("blue", "red"), lty=c(1,2), lwd=c(1, 3))
> eq <- paste0("MAAQG 120µg/m^3")
> mtext(eq,side= 3, line= -1, cex = 0.5)
>
> boxplot(Klang$O3_um3 ~ Klang$year, data = mydata, col = "lightgreen",
varwidth = TRUE,
+         ylab = "Consentrations (µg/m^3) ",xlab = "year", main = "Klang
")
> abline(h = 120, col=c("blue", "red"), lty=c(1,2), lwd=c(1, 3))
> eq <- paste0("MAAQG 120µg/m^3")
> mtext(eq,side= 3, line= -1, cex = 0.5)
>
> boxplot(Putrajaya$O3_um3 ~ Putrajaya$year, data = mydata, col = "ligh
tyellow", varwidth = TRUE,
+         ylab = "Consentrations (µg/m^3) ",xlab = "year", main = "Putra
jaya")
> abline(h = 120, col=c("blue", "red"), lty=c(1,2), lwd=c(1, 3))
> eq <- paste0("MAAQG 120µg/m^3")
> mtext(eq,side= 3, line= -1, cex = 0.5)
>
```

```
> boxplot(Shah_Alam$O3_um3 ~ Shah_Alam$year, data = mydata, col = "ligh
tpink", varwidth = TRUE,
+         ylab = "Consentrations (µg/m^3) ",xlab = "year", main = "Shah
Alam")
> abline(h = 120, col=c("blue", "red"), lty=c(1,2), lwd=c(1, 3))
> eq <- paste0("MAAQG 120µg/m^3")
> mtext(eq,side= 3, line= -1, cex = 0.5)
```

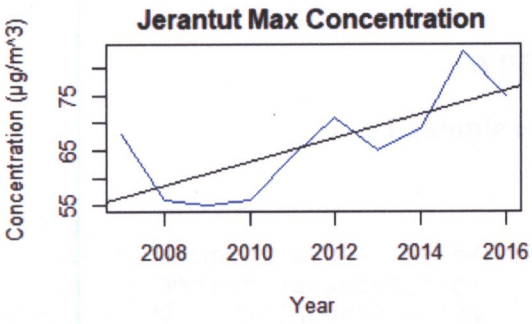


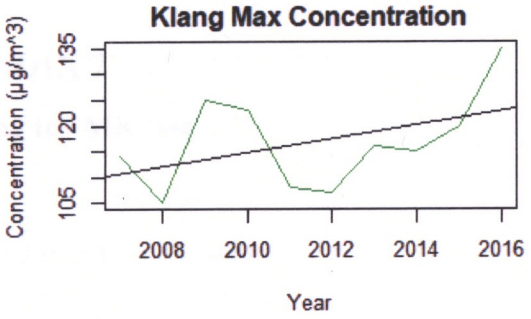

```
##Aveage
> Jerantut.mean <- aggregate(x = Jerantut$O3_um3, FUN = mean,
+                            by = list(Year = Jerantut$year))
> names(Jerantut.mean) <- c("year", "avg")
>
> Klang.mean <- aggregate(x = Klang$O3_um3, FUN = mean,
+                         by = list(Year = Klang$year))
> names(Klang.mean) <- c("year", "avg")
>
> Putrajaya.mean <- aggregate(x = Putrajaya$O3_um3, FUN = mean,
+                             by = list(Year = Putrajaya$year))
> names(Putrajaya.mean) <- c("year", "avg")
>
> SA.mean <- aggregate(x = Shah_Alam$O3_um3, FUN = mean,
+                      by = list(Year = Shah_Alam$year))
> names(Jerantut.mean) <- c("year", "avg")

##Max
> Jerantut.max <- aggregate(x = Jerantut$O3_um3, FUN = max,
+                           by = list(Year = Jerantut$year))
> names(Jerantut.max) <- c("year", "max")
>
> Klang.max <- aggregate(x = Klang$O3_um3, FUN = max,
```

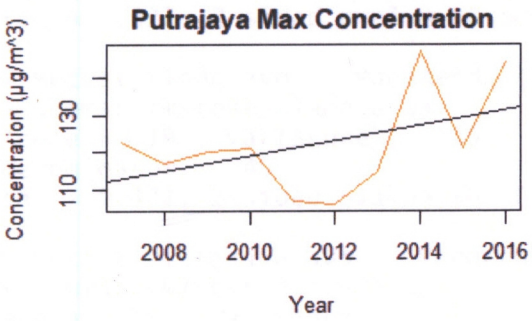
```
+                       by = list(Year = Klang$year))
> names(Klang.max) <- c("year", "max")
>
> Putrajaya.max <- aggregate(x = Putrajaya$O3_um3, FUN = max,
+                       by = list(Year = Putrajaya$year))
> names(Putrajaya.max) <- c("year", "max")
>
> SA.max <- aggregate(x = Shah_Alam$O3_um3, FUN = max,
+                       by = list(Year = Shah_Alam$year))
> names(SA.max) <- c("year", "max")

##Yearly Average
> par(mfrow=c(2,2))
>
> plot(Jerantut_c$avg ~ Jerantut_c$year, type = "line", col = "blue", y
lab = "Concentration (µg/m^3)", xlab = "Year", main = "Jerantut Avg Con
centration")
> fit <- lm(Jerantut_c$avg ~ Jerantut_c$year)
> abline(fit)
>
> plot(Klang_c$avg ~ Klang_c$year, type = "line", col = "green", ylab =
"Concentration (µg/m^3)", xlab = "Year", main = "Klang Avg Concentratio
n")
> fit <- lm(Klang_c$avg ~ Klang_c$year)
> abline(fit)
>
> plot(Putrajaya_c$avg ~ Putrajaya_c$year, type = "line", col = "orange
", ylab = "Concentration (µg/m^3)", xlab = "Year", main = "Putrajaya Av
g Concentration")
> fit <- lm(Putrajaya_c$avg ~ Putrajaya_c$year)
> abline(fit)
>
> plot(SA_c$avg ~ SA_c$year, type = "line", col = "magenta", ylab = "Co
ncentration (µg/m^3)", xlab = "Year", main = "Shah Alam Avg Concentrati
on")
> fit <- lm(SA_c$avg ~ SA_c$year)
> abline(fit)
```

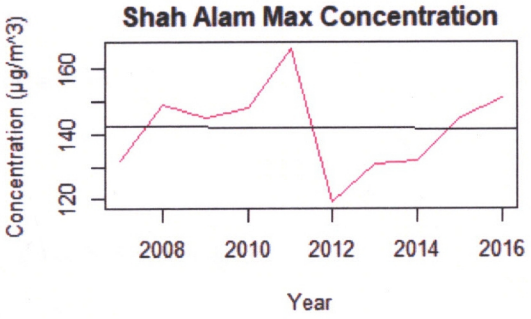



```
##Yearly Max
> par(mfrow=c(2,2))
>
> plot(Jerantut_c$max~ Jerantut_c$year, type = "line", col = "blue", yl
ab = "Concentration (µg/m^3)", xlab = "Year" , main = "Jerantut Max Con
centration")
> fit <- lm(Jerantut_c$max~ Jerantut_c$year)
> abline(fit)
>
> plot(Klang_c$max~ Klang_c$year, type = "line", col = "green", ylab =
"Concentration (µg/m^3)", xlab = "Year" , main = "Klang Max Concentrati
on")
> fit <- lm(Klang_c$max~ Klang_c$year)
> abline(fit)
>
> plot(Putrajaya_c$max~ Putrajaya_c$year, type = "line", col = "orange"
, ylab = "Concentration (µg/m^3)", xlab = "Year" , main = "Putrajaya Ma
x Concentration")
> fit <- lm(Putrajaya_c$max~ Putrajaya_c$year)
> abline(fit)
>
> plot(SA_c$max~ SA_c$year, type = "line", col = "magenta", ylab = "Con
centration (µg/m^3)", xlab = "Year" , main = "Shah Alam Max Concentrati
on")
> fit <- lm(SA_c$max~ SA_c$year)
> abline(fit)
```


Jerantut Max Concentration
Concentration (µg/m^3)
55
65
75
2008
2010
2012
2014
2016
Year


Klang Max Concentration
Concentration (µg/m^3)
105
120
135
2008
2010
2012
2014
2016
Year


Putrajaya Max Concentration
Concentration (µg/m^3)
110
130
2008
2010
2012
2014
2016
Year


Shah Alam Max Concentration
Concentration (µg/m^3)
120
140
160
2008
2010
2012
2014
2016
Year

APPENDIX 2

Example coding for MK test

```
##Avg
> MKtest_Jerantut_avg <- MannKendall(Jerantut_c$avg)
> summary(MKtest_Jerantut_avg)
Score =  37 , Var(Score) = 125
denominator =  45
tau = 0.822, 2-sided pvalue =0.0012822
>
> MKtest_Klang_avg <- MannKendall(Klang_c$avg)
> summary(MKtest_Klang_avg)
Score =  19 , Var(Score) = 125
denominator =  45
tau = 0.422, 2-sided pvalue =0.1074
>
> MKtest_Putrajaya_avg <- MannKendall(Putrajaya_c$avg)
> summary(MKtest_Putrajaya_avg)
Score =  5 , Var(Score) = 125
denominator =  45
tau = 0.111, 2-sided pvalue =0.72051
>
> MKtest_SA_avg <- MannKendall(SA_c$avg)
> summary(MKtest_SA_avg)
Score =  -19 , Var(Score) = 125
denominator =  45
tau = -0.422, 2-sided pvalue =0.1074


##Max
> MKtest_Jerantut_max <- MannKendall(Jerantut_c$max)
> summary(MKtest_Jerantut_max)
Score =  26 , Var(Score) = 124
denominator =  44.49719
tau = 0.584, 2-sided pvalue =0.024764
>
> MKtest_Klang_max <- MannKendall(Klang_c$max)
> summary(MKtest_Klang_max)
Score =  13 , Var(Score) = 125
denominator =  45
tau = 0.289, 2-sided pvalue =0.28313
>
> MKtest_Putrajaya_max <- MannKendall(Putrajaya_c$max)
> summary(MKtest_Putrajaya_max)
Score =  6 , Var(Score) = 124
denominator =  44.49719
tau = 0.135, 2-sided pvalue =0.65342
>
> MKtest_SA_max <- MannKendall(SA_c$max)
> summary(MKtest_SA_max)
Score =  3 , Var(Score) = 123
```

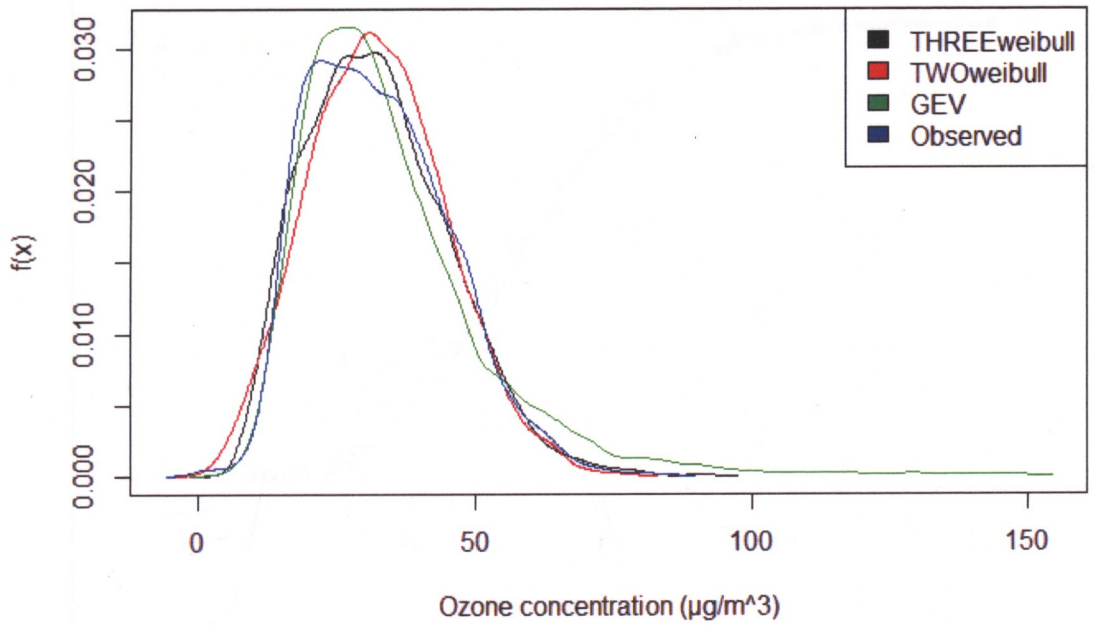
```
denominator =  43.98863
tau = 0.0682, 2-sided pvalue =0.85689
```

APPENDIX 2

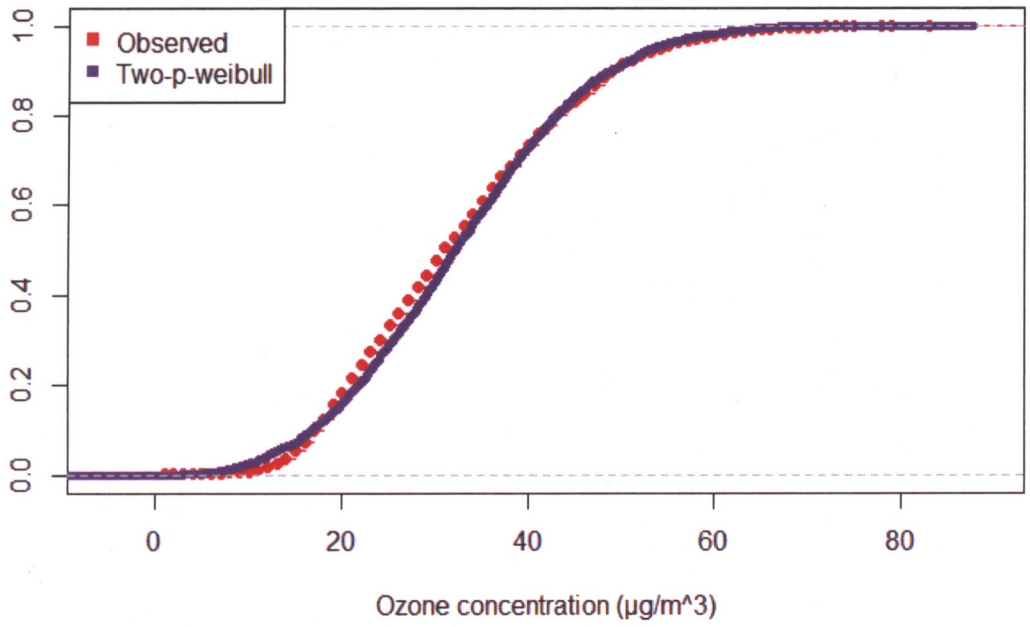
Example coding for distribution plot

```
> library(WeibullR)
> library(plotly)
> library(MASS)
> library(evd)
> library(reshape2)
> library(ggplot2)
> library(FAdist)
> library(stats)

> FileJ <- '. . . . 4. Data/R/csv file/Jerantut.csv'

> Jerantut <- read.csv(FileJ)

> head(Jerantut)

       id     date year time O3_um3 O3concppm_max
1 Jerantut 1/1/2007 2007    0     26         0.026
2 Jerantut 2/1/2007 2007    0     28         0.028
3 Jerantut 3/1/2007 2007    0     26         0.026
4 Jerantut 4/1/2007 2007    0     20         0.020
5 Jerantut 5/1/2007 2007    0     19         0.019
6 Jerantut 6/1/2007 2007    0     21         0.021

> summary(Jerantut)

       id             date           year           time           O3_
um3      O3concppm_max
 Jerantut:3542   1/1/2007:   1   Min.   :2007   Min.   :0.000   Min.
: 1.00   Min.   :0.00100
                 1/1/2008:   1   1st Qu.:2009   1st Qu.:2.000   1st Qu.
:23.00   1st Qu.:0.02300
                 1/1/2009:   1   Median :2011   Median :4.000   Median
:31.00   Median :0.03100
                 1/1/2010:   1   Mean   :2011   Mean   :4.462   Mean
:32.69   Mean   :0.03269
                 1/1/2011:   1   3rd Qu.:2014   3rd Qu.:7.000   3rd Qu.
:41.00   3rd Qu.:0.04100
                 1/1/2012:   1   Max.   :2016   Max.   :9.000   Max.
:83.00   Max.   :0.08300
                 (Other) :3536

> myplot1 <- data.frame(THREEweibull=rweibull3(3542, thres = 7.24, scal
e = 28.73, shape = 2.13),
+                       TWOweibull=rweibull(3542, scale = 36.69, shape =
2.84),
+                       GEV = rgev(3542, loc = 27.35, scale = 11.22, sha
pe = 0.11),
+                       Observed =Jerantut$O3_um3)>
```

```
> head(myplot1)
  THREEweibull TWOweibull      GEV Observed
1     32.95905   23.62628 25.25952       26
2     38.78215   35.76678 44.86286       28
3     45.24790   51.82065 47.48124       26
4     72.21125   40.90012 37.13984       20
5     49.38119   32.94147 25.44969       19
6     20.98188   55.04493 44.86419       21

> myplot1 <- data.frame(THREEweibull=rweibull3(3542, thres = 7.24, scal
e = 28.73, shape = 2.13),
+                       TWOweibull=rweibull(3542, scale = 36.69, shape =
2.84),
+                       GEV = rgev(3542, loc = 27.35, scale = 11.22, sha
pe = 0.11),
+                       Observed =Jerantut$O3_um3)

> head(myplot1)

  THREEweibull TWOweibull      GEV Observed
1     42.14307   33.65861 38.02297       26
2     26.21302   36.00623 40.22888       28
3     32.69696   33.22637 38.41499       26
4     28.63587   28.34463 61.30288       20
5     13.43798   26.00931 16.87073       19
6     27.66128   45.84211 27.10242       21

> dens <- apply(myplot1, 2, density)

> plot('f(x)', xlim=range(sapply(dens, "[", "x")), ylim=range(sapply(de
ns, "[", "y")), xlab = 'Ozone concentration (µg/m^3)', ylab = 'f(x)')

> mapply(lines, dens, col=1:length(dens))

> legend("topright", legend=names(dens), fill=1:length(dens))
```




```
> plot(ecdf(Jerantut$O3_um3),xlab = 'Ozone concentration (µg/m^3)', yla
b = '', main = '', col = "red" )

> lines(ecdf(myplot1$TWOweibull), col = "purple",lwd = 5)

> legend('topleft',
+        legend=c("Observed","Two-p-weibull"),  # text in the legend
+        col=c("red","purple"),  # point colors
+        pch=15)
```


Observed
Two-p-weibull
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
Ozone concentration (µg/m^3)