UNIVERSITI TEKNOLOGI MARA

PREDICTION OF GROUND LEVEL OZONE (O_3) CONCENTRATIONS USING EXTREME VALUE DISTRIBUTIONS: BAYESIAN APPROACH

FARAH AMIRA BINTI
MOHD HANAFIAH

MSc

July 2019

UNIVERSITI TEKNOLOGI MARA

PREDICTION OF GROUND LEVEL OZONE (O_3) CONCENTRATIONS USING EXTREME VALUE DISTRIBUTIONS: BAYESIAN APPROACH

FARAH AMIRA BINTI MOHD HANAFIAH

Dissertation submitted in partial fulfillment
of the requirements for the degree of
Master of Science
(Applied Statistics)

Faculty of Computer and Mathematical Sciences

July 2019

CONFIRMATION BY SUPERVISOR

APPROVED BY:

(DR. HASFAZILAH AHMAT)

Supervisor

Faculty of

Computer and Mathematical Sciences

Universiti Teknologi Mara

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	:	Farah Amira binti Mohd Hanafiah
Student I.D. No.	:	2017965135
Programme	:	Master of Science in Applied Statistic – CS702
Faculty	:	Computer and Mathematical Sciences
Dissertation Title	:	Prediction of Ground Level Ozone Concentrations using Extreme Value Distributions : Bayesian Approach
Signature of Student	:	.....
Date	:	July 2019

ABSTRACT

Ground level ozone (O_3) has become one of the most significant air pollutants due to the increasing sources of ozone precursors. The O_3 has also been reported as the main pollutant for photochemical smog and also the main pollutant that degrades air quality. Therefore, this study was conducted to find the best Bayesian EVD model to predict the high exceedances of O_3 concentrations using Extreme Values Distribution (EVD). Daily maximum data of 6 locations in Peninsular Malaysia which are Petaling Jaya, Shah Alam, Cheras, Klang, Tanjung Malim and Jerantut given by the Department of Environment from 1st January 2008 to 31st December 2017 were used in this study. The parameters were estimated using the Bayesian approach to fit into the EVD, which is the Generalized Extreme Value distribution (GEV) and Weibull distribution. The distribution that has the smallest error and highest accuracy measurement will be the best distribution. In this study, it shows that the Bayesian approach showed a high consistency and accuracy result since all the monitoring stations show GEV was the best distribution. The probabilities of the exceedances concentration were calculated, and the return period for the coming year was predicted from the cumulative density function (cdf) obtained from the best-fit distributions. From the validation result using data 2016 and 2017, it is shown that Klang has the similar result of percent compliance with the overall data while other monitoring stations was far from the overall data. Therefore, this distribution was suitable for Klang but may not to other location in this study. The return period for Klang was show that there is one exceedances will occurred per 73 days. It is proved that the Bayesian approach is superior compared to the classical approach because of the consistency and accuracy result in Bayesian. This study can be used to assess the high level of O_3 concentrations for policymakers to implement effective policies to create a cleaner environment.

ACKNOWLEDGEMENT

In the Name of Allah, the Most Beneficent and the Most Merciful. All the praises and thanks be to Allah, the Lord of the *worlds* who showed us the straight way, the way of those on whom He have bestowed His Grace, not (the way) of those who earned His anger, nor of those who went astray.

Firstly, I wish to thank Allah for giving me the opportunity to embark on my master and for completing this challenging journey successfully. Without the support, patience, and guidance of a few significant persons in my life, this study would not have been completed.

My gratitude and thanks go to my supervisor Dr. Hasfazilah binti Ahmat, for giving ideas, information about new things, and her effort on helping me in success along the way. Without her guide in this study, this thesis will not been done properly. Many thanks for all the support and guidance while conducting this research in purpose to produce better finding for this research study.

Special thanks go to my colleagues and friends for helping me and supporting me in completing this study. Their ideas and comments improve this study in many ways.

Finally, this appreciation goes to my family, especially my father and my mother, who has been encouraged in pursuing this master, support me, especially in financial and believed in me along the way in this study. This piece of victory is dedicated to both of you. Alhamdulillah.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Abbreviations

API	Air Pollution Index
ASMA	Alam Sekitar Malaysia Sdn. Bhd.
BUGS	Bayesian Inference using Gibbs Sampling
CDF	Cumulative Distribution Function
DOE	Department of Environment
EVD	Extreme Value Distribution
EVT	Extreme Value Theory
GEV	Generalized Extreme Value
GPD	Generalized Pareto Distribution
IA	Index of Accuracy
KNN	K-Nearest Neighbors
MAAQG	Malaysia Ambient Air Quality Guidelines
MAE	Mean Absolute Error
MAQI	Malaysia Air Quality Index
MCMC	Monte Chain Monte Carlo
MK	Mann-Kendall's
MLE	Maximum Likelihood Estimator
MOM	Methods of Moment
NAE	Normalized Absolute Error

NO_X	Nitrogen Oxides
O_3	Ground Level Ozone
PA	Prediction Accuracy
PDF	Probability Density Function
$PM_{2.5}$	Particulate Matter with aerodynamic diameter less than 2.5μm
PM_{10}	Particulate Matter with aerodynamic diameter less than 10μm
R^2	Coefficient of Determination
RMSE	Root Mean Squared Error
SPSS	Statistical Package for the Social Sciences
VOCs	Volatile Organic Compounds
WHO	World Health Organisation

CHAPTER ONE
INTRODUCTION

1.1 Background of Study

Ground level ozone (O_3) has become one of the most significant air pollutants due to the increasing sources of ozone precursors. The O_3 has also been reported as the main pollutant for photochemical smog and also the main pollutant that degrades air quality. Unlike the ozone in the upper atmosphere that occurs naturally and useful to protect air qualities, the O_3 is human-made and can have harmful effects on humans and the environment (NHDES, 2015). The O_3 is a secondary air pollutant created by a combination of nitrogen oxides (NOx) and a variety of volatile organic compounds (VOCs) with the presence of sunlight (Hashim & Noor, 2017). Emissions of NOx are produced primarily by motor vehicle engines, power plants, industrial plants, boilers and burning of fossil fuels. VOCs emissions are motor vehicle emissions, gasoline vapours, and chemical solvents.

The O_3 is well known as a strong oxidant. It has a direct effect on human, vegetation, and materials. Inhalation of air mass containing 1 ppm of O_3 by volume causes severe irritation and headache. The high O_3 concentration also poses medical issues like irritating the nose, throat, and lungs and cause chest pain, coughing, and nausea. The O_3 bothering respiratory conditions such as allergies, asthma, and emphysema can have severe effects even on healthy people who work or play outdoors during the hot, sunny summer months when smog is usually at the highest level. The effect of O_3 is believed to harm forests and crops and may quicken the decay of elastic tires, paints and dyes in fabrics (Escarela, 2012; Mabahwi, Ling, Leh, & Omar, 2015).

Over 2 million unexpected losses each year can be credited to the impacts of open and indoor air contamination (World Health Organization, 2006). The populations of developing countries carry a more significant part of the illness. Heart diseases, lung problems, and lung cancer are all significantly higher in people breathing dirty air than in cleaner environments in groups (Mabahwi *et al.*, 2015). Hence, it becomes necessary to predict the O_3 concentration in order to ensure the O_3 concentration level does not

exceed the limit that has been stated by the Malaysia Ambient Air Quality Guideline (MAAQG).

High O_3 concentration levels is a particular interest for Extreme value theory (EVT), and most of the extreme value distributions studies in air quality used data that has a high occurrence of concentrations. For instance, daily maximum or data above a certain limit are used. The EVT develops techniques and models for describing the unusual (extremes) rather than the usual phenomenon (Kotz & Nadarajah, 2000). The EVT functions to express the extremes of the observed process in determining the likelihood of extreme events in the future. The EVT is a flexible model for entire conditional distribution since it has been characterised to focus on the upper tail. Thus, the EVT is assumed to be an appropriate method for the upper tail of the conditional distribution (Reich, Cooley, Foley, Napelenok, & Shaby, 2013). There are good reasons to choose the Bayesian procedure in determining the parameter of the distribution if an appropriate prior can be specified (Alam, Farnham, & Emura, 2018).

The most commonly used in fitting the parameter of the extreme value distribution is known as frequentist (or classical) methods, i.e., maximum likelihood estimator (MLE) or method of moments (MOM). The Bayesian approach offers a more coherent framework in incorporating all of the uncertainties involved in the prediction process using conventional methods (Coles, Pericchi, & Sisson, 2003). In general, the uncertainties of the parameter is as a result of a lack of data and that the selection of probability distribution does not describe the data correctly (Chung & Kim, 2013). The classical method based on MLE is unbiased for a large sample size but may be biased for limited data (Kery, 2010). Kery (2010) also stated that the Bayesian approach is exact for any sample size and will improve the quality of the inferences and tolerate a reduction in sample size.

Since the classical or conventional EVD methods are commonly used to analyse the concentration of air pollution in general, the Bayesian approach is frequently applied in hydrological studies compared to studies in air pollution (Ahmat, 2016). Hence, this study aims to find the best Bayesian EVD models and to predict the O_3 concentrations.

1.2 Problem Statement

There has been much publicity on Particulate Matter (PM_{10}), but ground-level ozone (O_3) is also an equally important air pollutant to be looked into since excessive levels of O_3 can be harmful. A location that has more transportation and industrialisation tends to have more exceedances of O_3, but due to factors of wind that transports O_3 concentrations, other locations may also have exceedances of O_3 concentrations. High exceedances of O_3 along with carbon dioxide, methane, water vapour, and nitrous oxide also can act as a greenhouse gas, and it is the third largest contributor to global warming (Ghazali, Yahaya, & Mokhtar, 2014). Therefore, predicting the high exceedances of O_3 using suitable distribution become essential in every location, since it can be a harmful pollutant for humans, materials, and also in the plantation.

Based on previous studies, the common practice in predicting O_3 concentration is using multiple linear regression, time series, and also a neural network for data mining. The statistical tools that are often used by researchers in estimating the parameters for distribution are the classical approach, which is maximum likelihood estimation and method of moment. However, maximum likelihood estimator can be unbiased for a large sample size but may be biased for limited data while the Bayesian approach is exact for any sample size and will improve the quality of the inferences and tolerate a reduction in sample size. Even though the alternative Bayesian approach is claimed to be increasingly popular nowadays, there is still a lack of literature reviews of estimating the parameters for extreme value distribution using the Bayesian approach in O_3 studies. In Ahmat (2016) that studied the high concentration of PM_{10}, it was stated that the Bayesian approach is more prevalent in hydrological studies but not commonly applied in air pollution studies.

This assessment helps to identify trends and how Bayesian methods have integrated into environment researches in the context of different statistical frameworks. The prediction of O_3 using Bayesian analysis also crucial because the uncertainty in O_3 is an important issue that needs to be considered. This method has been applied in a high concentration of PM_{10} but not on O_3 studies in Malaysia. This research aims to determine the best extreme value distribution and predict the O_3 concentration using a Bayesian approach.

1.3 Research Questions

The research questions in this study are:

i) What is the pattern and behaviour of Ground Level Ozone (O_3) concentrations in Peninsular Malaysia?

ii) What is the suitable prior to estimate the parameters for Bayesian model distributions?

iii) What is the best distribution for predicting future exceedances of high O_3 concentration using the Bayesian approach?

1.4 Research Objectives

The research objectives are:

i) To describe the pattern and behaviour of Ground Level Ozone (O_3) concentrations in Peninsular Malaysia and determine the occurrence of high concentrations.

ii) To determine the suitable prior to estimate the parameters for Bayesian model distributions.

iii) To determine the best Bayesian EVD for predicting future exceedances of high O_3 concentration.

1.5 Scope and Limitation of Research

This study only concentrates on one major pollutant in Malaysia, which is Ground Level Ozone (O_3). The air pollution records from 6 different monitoring stations; Jerantut, Cheras, Shah Alam, Klang, Petaling Jaya, and Tanjung Malim from the year 2008 until 2017 were used. The records were provided by the Department of Environment (DOE). The selection of the monitoring stations based on past literature to gauge the extreme concentrations of O_3. As there are sixty-five (65) monitoring stations, this study only considers locations which represent each category of urban, suburban and industrial areas to detect any significant changes in the air quality which may be harmful to human health and the environment.

The method that will be used to predict the exceedances of O_3 concentrations is the Bayesian approach using non-informative prior to the extreme value distribution in determining the parameters of the distribution. Ahmat (2016) applied the Bayesian approach to the study of PM_{10} and found that non-informative prior was the best distribution compared with informative prior. It was too tedious and time-consuming to identify the informative prior. Since this study will utilise ten years' worth of data, which has large amounts of data, it is not necessary to spend large amounts of time and energy in developing an informative prior distribution since the effect of the prior distribution on the posterior distribution tends to be weakened when dealing with large data (Chung & Kim, 2013). The Bayesian approach uses likelihood distributions of the GEV and two-parameter Weibull with the non-informative priors to estimate all parameters.

1.6 Significance of Research

Air pollution causes severe damage to health, environment, and property. There are hundreds of deaths which have been related to poor air quality in cities all over the world. The leading causes of this pollution are open fires, large numbers of motor vehicles with poor maintenance and industrial plants with weak regulation. High O_3 concentration poses medical issues such as irritating the nose, throat, and lungs and cause chest pain, coughing, and nausea. Long-term exposure may lead to permanent lung damage (Escarela, 2012). As O_3 concentrations increase above the guideline value, health effects at the population level become increasingly numerous and severe. The incidences of the high particulate event bring a negative effect on human health, environment, and country. Therefore, it is crucial for the government to raise public awareness relating to the effect of O_3 pollutant.

Even though the Bayesian approach gives many advantages to statistics such as to the sample size, ease of error propagation and intuitive appeal in the interpretation of probability, most researchers use classical statistics because of the resistance to the Bayesian philosophy with its perceived subjectivity of prior choice. Most applications of Bayesian statistics feature complex statistical models. The Bayesian statistics is hard to understand in the first place, and it may not be relevant to the majority. The Bayesian analysis process has typically involved custom-written code in general-purpose

computer languages. Therefore, for someone lacking a solid knowledge in statistics and computing, Bayesian analyses were mostly out of reach (Kery, 2010).

Moreover, the Bayesian approach is often claimed by its practitioners that it is more philosophically consistent and can tackle problems beyond the reach of classical statistics (Willink & White, 2011). The extreme value distribution using maximum likelihood method (MLE) is widely applied due to the estimator being less biased. However, MLE showed weaknesses in return period (Ouyang *et al.*, 2011; Zhang *et al.*, 2017). Thus, this study is to introduce the advantages of Bayesian estimation compared to the classical approach.

The Bayesian approach has been applied in the environmental field, which is PM_{10} in Malaysia but not in O_3. The major gap in this study is in the application of statistical modeling using the Bayesian approach in the O_3 study in Malaysia for the prediction of the high level of O_3 concentrations. It will be used to assess the high level of O_3 concentrations, which are vital for implementing effective policies to create a cleaner environment.

CHAPTER TWO
LITERATURE REVIEW

2.1 Introduction

This section provides a relevant topic on the research subject in gaining more knowledge and understanding of this study. This section will cover reviews on air pollution in Malaysia, ground-level ozone (O_3), past researches on extreme value distribution and the Bayesian approach.

2.2 Air Pollution in Malaysia

The haze phenomenon in Malaysia is a major and severe problem, especially in the central region. Haze in Malaysia contributes to the event of air pollution, including a high level of ground-level ozone (O_3) concentrations that exceed Malaysian standards and have been observed in some urban and industrial regions of Malaysia (Ghazali *et al.*, 2014). The main contributor to air pollution in urban areas is industrial and motor vehicle activities. In a decade, transboundary smoke haze pollution had become an annual phenomenon that is worsening on local air quality in an Asian country, including Malaysia. Haze has been attributed from forest and peat fires in Indonesia that contains a load of particulate matter, toxic, graphitic carbon that is hazardous to human health.

There are several main causes of air pollution in Malaysia, and it can be categorised into two, which is primary and secondary air pollution. Primary air pollution is where the pollutants are directly from the sources such as combustion of fuel in an engine releasing carbon monoxide and carbon dioxide out of the exhaust. Carbon monoxide comes from the smoke emitted from vehicles and open burning, and nitrogen oxides are produced by combustion of fossil fuels at high temperature. Meanwhile, secondary air pollution is air pollution that is not formed directly, which is a combination of other sources with the main pollution such as O_3. The O_3 is a combination of NO_x and VOC emitted from vehicles and industries with the presence of sunlight (Fadzly, Nordin, & Rashid, 2018).

According to the Department of Environment Malaysia, it is stated that there are six main pollutants which are the particulate matter with 10-micron size in diameter (PM_{10}), particulate matter with 2.5-micron size in diameter ($PM_{2.5}$), sulphur dioxide (SO_2), nitrogen dioxide (NO_2), carbon monoxide (CO), and ground-level ozone (O_3). Each of the pollutants can cause serious harm to human health, environment, and property. For instance, PM_{10} can cause lung cancer and cardiopulmonary deaths, while O_3 can reduce lung functions and induce coughing and choking. The presence of CO can cause mortal growth in pregnant women as well as affect tissue development of young children (Rani *et al.*, 2018). PM_{10} and O_3 are the main concerns in Malaysia since the two major air pollutants, particularly in the urban and suburban areas in Malaysia, have been recognised to have the high potential for deleterious effects on human health (Azid *et al.*, 2015). Figure 2.1 shows the health effects caused by the high exceedances of air pollution.

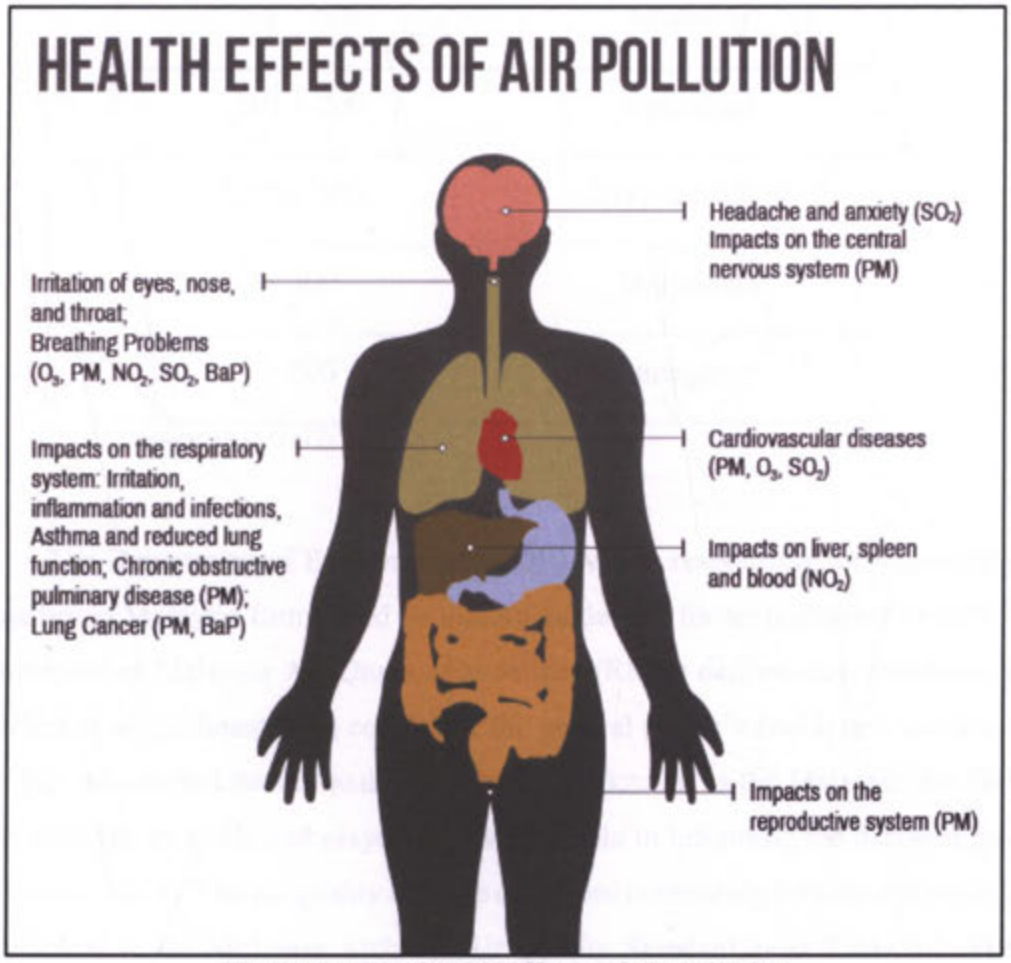


Figure 2.1 The health effects of air pollution

Therefore, DOE has monitored the country's ambient air quality through the establishment of 65 continuous monitoring stations throughout Malaysia. The monitoring stations are strategically located in different areas such as urban, suburban, and industrial to detect any significant changes in air pollution, which can bring danger to human health and the environment (Department of Environment, 2018). DOE also formulated the Air Pollution Index (API) to reflect the status of air quality in Malaysia. The API is calculated based on the concentrations of six primary pollutants, as stated in Table 2.1.

Table 2.1
Malaysia: Air Pollutant Index (API)

API	Air Quality Status
0 – 50	Good
51 – 100	Moderate
101 – 200	Unhealthy
201 – 300	Very unhealthy
> 300	Hazardous
> 500	Emergency

Source: DOE (2017)

The Department of Environment (DOE) who is responsible for monitoring the air quality in Malaysia formulated air quality guidelines for air pollutants in 1989. The Recommended Malaysia Air Quality Guidelines (RMG) defined concentration limits for selected air pollutants that could hurt the general public's health and welfare. The DOE has established the air quality index system, known as the Malaysia Air Quality Index (MAQI) in 1993, and played an essential role in informing the decision makers (Rani *et al.*, 2018). The air quality average data from monitoring stations were specified and referred to the Malaysia Ambient Air Quality Standard as in Table 2.2. The O_3 guideline was shown as highlighted, therefore the exceedances for O_3 for averaging time 1 hour is O_3 that above than 0.1 ppm and for 8 hours is O_3 that above 0.06 ppm.

Table 2.2
Malaysian Ambient Air Quality Standard

Pollutant	Averaging Time	Existing Guideline		Standard (2020)	
	Time	ppm	$\mu g/m^3$	ppm	$\mu g/m^3$
Particulate Matter with the size of less than 10 microns (PM_{10})	1 Year	-	50	-	40
	24 Hours	-	150	-	100
Particulate Matter with the size of less than 2.5 microns ($PM_{2.5}$)	1 Year	-	-	-	15
	24 Hours	-	-	-	35
Sulphur Dioxide (SO_2)	1 Hour	0.135	350	0.115	250
	24 Hours	0.040	105	0.035	80
Carbon Monoxide (CO)	1 Hour	30.6	35	30.6	30
	8 Hours	8.75	10	8.75	10
Nitrogen Dioxide (NO_2)	1 Hour	0.17	320	0.150	280
	24 Hours	0.04	75	0.037	70
Ground Level Ozone (O_3)	1 Hour	0.10	200	0.09	180
	8 Hours	0.06	120	0.05	100

Source: DOE (2017)

2.3 Ground Level Ozone

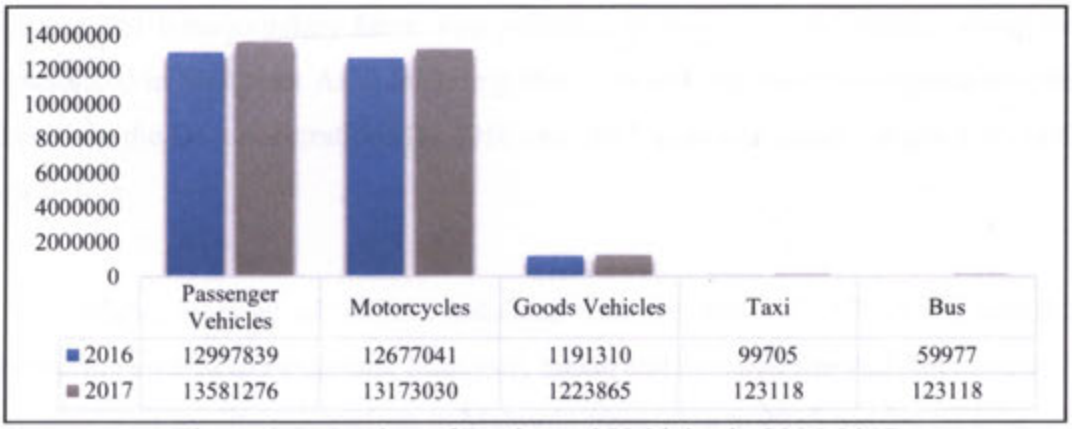
A series of Malaysia Environmental Quality Report (Department of Environment Malaysia, 2015, 2016, 2017, 2018) reported that ground-level ozone (O_3) continued to be the pollutant of concern due to favourable atmospheric condition and emission from motor vehicles particularly in urban areas that enhanced its formulation. The O_3 was determined as the main pollutant since it is shown to have a severe negative

impact on human health, crops, and other plants worldwide, including both developed and developing countries (Wahid, 2006). Mohammed (2012) stated that O_3 is the most harmful pollutant for humans since it plays an essential role in damaging vegetation and materials.

The O_3 is a pollutant that is not emitted directly by any source but is created by chemical reactions between nitrogen oxides (NO_x) and volatile organic compounds (VOCs) in the presence of sunlight and heat (World Health Organization, 2006). Urban environments on hot sunny summer days are very likely to reach unhealthy levels of O_3. Nitrogen oxides are emitted by motor vehicles, industrial areas, commercial and fuel combustion activities, and other miscellaneous activities like using lawn care equipment and vehicle refuelling. Meanwhile, VOCs are emitted by area of polluters (vehicle refuelling, solvent usage, etc.), motor vehicles, industrial and commercial processes, and also by non-road outdoor equipment.

The highest O_3 concentrations in Malaysia for air masses characterised by dry and warm conditions usually occur between June and September each year, especially during the Southwest monsoon and for air originating from Sumatra, Indonesia. The O_3 concentrations are at the lowest during the wet season, which is during the Northeast monsoon, and it occurs annually between November and March. From the back trajectory analysis that has been done, it showed that during the Northeast monsoon, the air masses are initiated from the South China Sea and Philippines (Abdullah, Ismail, Yuen, Abdullah, & Elhadi, 2017).

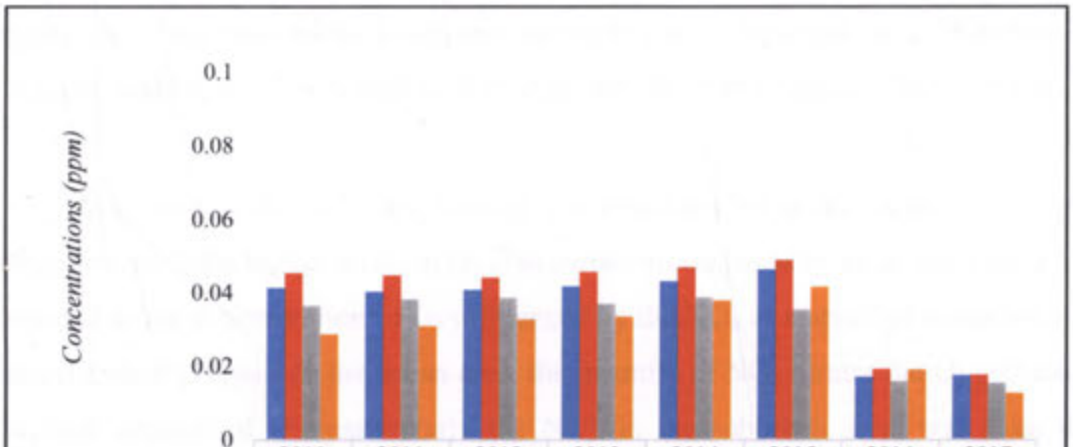
Since vehicles and gas powered equipment are the primary sources for NOx and VOC pollution, by reducing the numbers of cars on the road, fuelling car, or mowing the lawn late in the evening, it can reduce the potential of O_3 creation. However, the Department of Environment Malaysia (2018) stated that there is a growing trend in industrial sources and the number of motor vehicles. This can lead to air quality reduction if emissions, including smoke, are not adequately controlled by both sources. Motor vehicle emissions are the main source of air pollution, particularly in urban areas. In 2017, there was an overall increase in the number of motor vehicles registered compared to 2016, as shown in Figure 2.2.



	Passenger Vehicles	Motorcycles	Goods Vehicles	Taxi	Bus
2016	12997839	12677041	1191310	99705	59977
2017	13581276	13173030	1223865	123118	123118

Figure 2.2 Number of Registered Vehicles in 2016-2017
(Source: Department of Environment, 2018)

Due to high traffic with a favourable atmospheric condition, the urban area has recorded the highest level of the O3 concentration, as seen in Figure 2.3. The O3 was identified as the dominant pollution in some suburban and rural areas due to the downwind effect that transports the O3 pollutant (Department of Environment Malaysia, 2018).



	2010	2011	2012	2013	2014	2015	2016	2017
INDUSTRIAL	0.0412	0.0402	0.0407	0.0417	0.0431	0.0463	0.017	0.0175
URBAN	0.0453	0.0447	0.0441	0.0457	0.0469	0.0489	0.0192	0.0176
SUB URBAN	0.0364	0.0383	0.0386	0.037	0.0388	0.0354	0.0159	0.0155
BACKGROUND	0.0287	0.0309	0.0327	0.0335	0.0378	0.0416	0.019	0.0127

Figure 2.3 Annual Average Concentration of Ozone (O_3) by Land Use, 2010- 2017
(Source: Department of Environment Malaysia, 2018)

The Department of Environment (2017) reported that the API reading for overall air quality in 2017 was between good to moderate levels and there was a reduction in terms of a number of unhealthy days recorded in 2017 compared to 2016. In 2017, there was no haze episode in Malaysia compared to the past years that was mainly due to

regional and transboundary haze. The pollution is due to wetter weather conditions experienced in Southeast Asia, including Malaysia and Indonesia throughout the year. Therefore, the O_3 concentrations for 2016 and 2017 were decreased compared from the year before.

While, Rani *et al.* (2018) indicated that the trend of API at 50 sampling monitoring located in Peninsular Malaysia, Sabah and Sarawak showed that most of the continuous air monitoring stations in Malaysia from 2010 to 2015 were at an unhealthy level. However, some of the locations were detected at emergency level with API value higher than 500 from 2010 to 2015. In the study, it stated that the high API value might be due to the increase of the O_3 concentrations resulting from the sunrise coinciding with increasing solar radiation and other pollutants.

A study by Othman *et al.* (2014), assessed the economic value of health impacts of transboundary smoke haze pollution in Kuala Lumpur and adjacent areas in the state of Selangor, Malaysia. However, their study found that the Air Pollution Index (API) readings have not been useful to explain the variations in inpatient rates. Therefore, it is vital to study specific pollutant to determine the effect and causes of the pollutant.

In a study by Banan, Latif, Juneng, and Ahamad (2013), the result showed that suburban areas have higher levels of O_3 concentration compared to urban areas and rural areas due to the concentration of O_3 influenced by the NO_x as a result of oxidation and photochemical process. In the urban area, the quantity of NO_x titrates the O_3 pollutant. The high amount of O_3, particularly the NOx in the suburban area, was from the movement of air from the city center. Furthermore, the concentration of O_3 was influenced by the intensity of solar radiation and the wind pattern.

2.4 Extreme Value Theory

The extreme value theory had been used previously as a tool for studying air pollution problems (Smith, 1989). Extreme values are part of particular interest in monitoring the impact of high air pollution, and partly because air quality standards are formulated based on the highest level of permitted emissions. The EVT has been widely

found in most area of applications, especially in the environmental area where extreme levels can cause substantial damages. Examples of the case in extreme levels that have been of a concern in the research area are wind speeds, sea levels, pollutant concentrations, and river heights (Hurairah et al., 2005). Many areas of statistical applications are primarily interested in drawing inferences about the extreme (maximum or minimum) values.

The extreme value theory approach has the advantage over other methods in which the predicted meteorology, seasonality, temporal trend and outcome, and meteorological history, extreme quantiles can be directly calculated and predicted for a given day. As an example, a study by Escarela (2012) applied a new model in O_3 data using extreme value distribution and used the maximum likelihood estimation (MLE) in estimating the unknown parameter.

Smith (1989) has applied numerous different models based on Gumbel, Generalized Extreme Value (GEV) and Generalised Pareto Distribution (GPD) distribution to the study of a threshold value of 8 parts per 100 million of surface O_3 level in Houston, Texas for the period of April 1973 to December 1986. The study concluded that the application of GPD provides a useful approach for quantifying the return level of high O_3 pollution compared to GEV. However, the modeling approach only used the ideas of point-process theory to produce a general strategy. There is a need to repeat the analysis with data collected from other sites to get a reliable indication of the result.

In the study by Kuchenhoff and Thamerus (1995) that used GEV and GPD, the extreme value distribution was fitted to the maximum monthly concentration. The finding found that it was challenging to interpret the result compared to using a simple model since there were too many parameters that had to be included, and the models were complicated.

Hurairah *et al.* (2005) modified an extreme value distribution from Gumbel by introducing a new parameter, which was the shape parameter. The study used the new EVD in comparison with Weibull, Frechet, and Exponential using carbon monoxide

(CO). The parameter estimation for the new EVD used the maximum likelihood estimation method, and the results found out that the new model had higher accuracy compared to other EVD.

Yahaya *et al.* (2013) applied the Weibull and other non-EVD probability distributions (log-normal, gamma, Laplace, log-logistic, inverse Gaussian and Rayleigh) to the 2002 hourly CO concentration level in Kuala Lumpur. Maximum likelihood estimation method was used to estimate the parameters of the distribution. However, the finding was in favour of other probability distribution to represent the best fit for the data. The result came from the data use, which was the hourly data and not the extreme concentrations that are suited for EVD.

Ghazali, Yahaya, and Mokhtar (2014) used four probability distributions in their research, namely Weibull, Beta, lognormal, and Inverse Gaussian distribution in predicting O_3 concentration levels. The findings showed that the air quality status in Cheras was weak at all times, and the results concluded that Beta distribution fitted the distribution well as the performance indicator that gave the best result compared to others.

Based on a study by Ahmat, Yahaya, and Ramli (2015) on the prediction of PM_{10}, extreme concentration comparing six EVD which were Gumbel, 2 and 3-parameter Weibull, GEV and 2 and 3-parameter GPD showed that three-parameter GEV was the most appropriate distribution for the daily maximum concentration PM_{10}. The study showed that the predicted number of days in which PM_{10} concentration exceeding Malaysian Ambient Air Quality Guideline (MAAQG) were more than 85% in compliance with the actual number of days. The study concluded that the GEV could be used to estimate the extreme concentrations of PM_{10}.

A study was also conducted by Nasir, Ghazali, Mokhtar, and Suhaimi (2016), which used three distributions, Lognormal, Weibull, and Inverse Gaussian distribution. In their study, Weibull distribution was the best distribution to fit the O_3 concentration in Port Dickson and Port Klang. This study also found that the two-parameter Weibull distribution was widely used in data analysis because of its flexibility in modeling.

Martins *et al.* (2017) studied air pollution data using extreme value analysis comparing two large urban regions of South America. The findings showed that the GEV and GPD fitted the distribution well. Table 2.3 summarises the literature in air pollution studies with various approaches of EVD.

Table 2.3
Summary of research in Extreme Value Distribution studies

Citations	Area of study	Distribution	Pollutants
Martins *et al.* (2017)	South America	GEV and GPD	O_3, NO, NO_2, PM_{10}, CO, $PM_{2.5}$
Nasir *et al.* (2016)	Port Dickson and Port Klang, Malaysia	Lognormal, Weibull and inverse Gaussian	O_3
Ahmat *et al.* (2015)	Malaysia	Gumbel, 2 and 3-parameter Weibull, GEV and 2 and 3-parameter GPD	PM_{10}
Ghazali *et al.* (2014)	Cheras, Malaysia	Weibull, Beta lognormal and Inverse Gaussian	O_3
Yahaya, Ramli, *et al.* (2013)	Kuala Lumpur, Malaysia	Weibull and non-EVD	CO
Hurairah *et al.* (2005)	Malaysia	Modified Gumbel, Weibull, Frechet, and Exponential	CO
Kuchenhoff and Thamerus (1995)	Munich, Germany	GEV and GPD	O_3, and NO_2
R. L. Smith (1989)	Houston, Texas, USA	Gumbel, GEV, GPD	O_3

2.5 Bayesian Approach

Bayesian statistic consists of two main component which is likelihood distribution and prior distribution. "Prior" is the set of information about the data, but usually there is no prior information available. In this case, the posterior that will not influence the prior has to be specified. The distribution without prior information is known as non-informative prior (Ntzoufras, 2009).

Bayesian statistics are different from the classical approach since all the unknown parameters are considered as random variables. So, the prior distribution has to be defined initially. The prior distribution can be informative prior or non-informative prior regarding the information of the data. Since there is large data, non-informative prior will be used in this study. The prior distribution has the agreeable property with the posteriors of the same distributional family. Table 2.4 lists the family of conjugate prior used in the Bayesian analysis.

Table 2.4
Conjugate prior distributions for the commonly used algorithm

Likelihood Distribution	**Prior Distribution**
Poisson	Gamma
Binomial	Beta
Normal (known σ^2)	Normal
Normal	Inverse Gamma
Gamma (ν known)	Gamma
Exponential	Gamma
Negative binomial	Beta
Multinomial	Dirichlet
Exponential Family (ϕ known)	Exponential

(Source: Ntzoufras, 2009)

Data analysts did not prefer the Bayesian approach due to the difficulty with the calculation of posterior distribution (Ntzoufras, 2009). A simulation-based technique can solve the problem using Markov chain Monte Carlo (MCMC) to simulate realisations of the posterior distribution, and from the simulated sample, the parameter estimation of the posterior distribution can be obtained. Techniques of MCMC describe a method to simulate a complex distribution by simulating the target distributions from Markov chains as their stationary distributions (Alam et al., 2018).

Smith (2005) also stated that MCMC had provided a way for Bayesian techniques to be applied. Extreme value data using Bayesian analysis has many potential benefits over the classical approach such as sources of information other than the data which are likely to be scarce can be included. The predictive distribution also provides a natural way to estimate the probability of future events being extreme, and Bayesian analysis is not dependent on the regularity assumptions required by the asymptotic theory of maximum likelihood.

Fernández (2006) stated that in the perspective of Bayesian approach, the parameters of the distribution are not merely an unknown fixed quantity but rather a random variable that is characterised by some prior probability density function. The Bayesian approach allows sample and prior information to be incorporated into statistical analysis that will improve the quality of the inference and permit a reduction in the sample size. Bayesian approach is commonly used in the field of hydrology compared to air pollution.

In the study by Molitor *et al.* (2006), it examined the long-term effects of nitrogen dioxide (NO_2) exposure on children's lung function. It was used to compare the Bayesian and classical approach. It is claimed that the Bayesian framework utilises measurement error models to estimate missing exposure data in health effects assessment. The study also stated that various measures had been done, but there is no completely accurate measure of household-level for long-term nitrogen dioxide exposure available. Thus, the finding from this study shows that the method using the Bayesian approach improves the estimates compared with the classical approaches.

Eli (2012) used the Bayesian approach on the extreme rainfall data analysis and found that the method fitted well in estimating GEV parameters. Bayesian MCMC was found to be the most appropriate method in describing the annual maximum rainfall. The result also showed that the parameter estimation for GEV distribution was better using the Bayesian approach compared to the MLE method.

Chung and Kim (2013) studied the comparison of Bayesian analysis using GEV and Gumbel and concluded that the GEV distribution was applied effectively to the

data. Bayesian analysis is well suited in decision making when there is no information about the parameter, or there is uncertainty regarding the distribution. A study by Vidal (2014) utilised the Bayesian inference method for the Gumbel distribution using the annual rainfall maximum intensities data in the different region of Chile. The study applied the non-informative prior distribution of an inverse gamma for the location and scale parameter. The posterior distribution of future observations was obtained to predict and estimate the return periods and intensity-duration-frequency curves.

Chikobvu and Chifurira (2015) applied the MCMC techniques for Bayesian analysis to study the annual maximum and minimum rainfall data for Zimbabwe using the GEV distribution. The maximum likelihood estimation method was also used in this study to obtain the estimates of the parameters. The result showed that the GEV parameter estimates using the Bayesian approach were almost similar to the maximum likelihood estimates with smaller standard deviations. Chikobvu & Chifurira (2015) also mentioned that the use of expert priors might improve the precision of the parameters over the maximum likelihood estimates.

Martins, Sam, and David, (2015) suggested the Bayesian MCMC as the parameter estimation due to the advantage in reducing the parameter uncertainty. This study also mentioned that the MLE is a reliable principle to derive an efficient estimator for a model as sample size approaches infinity. However, by comparing the classical and Bayesian approach, this study proved that Bayesian MCMC is the better method to estimate the distribution parameters of extreme daily rainfall amount in Makurdi.

A study by Alam *et al.* (2018) applied Bayesian Modelling to the Flood Frequency Analysis Using Hamiltonian Monte Carlo Techniques. The method employed was to obtain the approximations of the posterior marginal distribution of the GEV model using annual maximum discharges for two major river basins in Bangladesh. The result concluded that the Bayesian MCMC could provide a more accurate description of flood risk and the parameter uncertainties with more accurate credible intervals.

Lima, Kwon, and Kim (2018) estimated rainfall intensity-duration-frequency (IDF) for future climate using Bayesian GEV. The EVD GEV to estimate the maximum rainfall and the estimates of the scaling parameters were used to estimate environmental parameters for shorter durations. The result suggests that the changes in the scale parameter drive the future changes in rainfall intensity. The study also found that Bayesian inference for the duration rainfall modeling provides a framework for estimating environmental parameters along with their uncertainties that can be used to build current and future climate IDF curves.

Cheong and Gabda (2018) studied using Bayesian MCMC to fit nine annual maximum river flows in Sabah for a period record of over 20-48 years into the GEV distribution. This study believed that Bayesian MCMC provides a more robust inference through prior and posterior distribution. By comparing the findings using maximum likelihood estimation and based on the Bayesian advantages, the Bayesian approach was chosen in this study to estimate the parameters.

Ahmat (2016) predicted PM_{10} concentration using extreme value distribution using the Bayesian and classical approach. She found that the Bayesian approach was preferred for its comprehensiveness in incorporating all the uncertainties involved in the prediction process compared to the classical approach. This study concluded that the Bayesian approach using GEV as the likelihood distribution with a non-informative uniform prior was the best distribution for maximum daily concentrations for Klang and Seberang Jaya while two-parameter Weibull was the best distribution for Perai. Table 2.5 explains the summary of the literature review using the Bayesian approach in environmental studies.

Table 2.5
Summary of the Bayesian approach in Environmental studies

Citations	Area	Likelihood Distribution	Field of Study
Alam *et al.* (2018)	Bangladesh	GEV	Flood
Cheong and Gabda (2018)	Sabah, Malaysia	GEV	River flow
Lima *et al.* (2018)	South Korea	GEV	Rainfall
Ahmat (2016)	Malaysia	GEV, Gumbel, Weibull, GPD	Air Pollution (PM_{10})
Martins, Sam, and David, (2015)	Nigeria	GEV	Rainfall
Chikobvu and Chifurira (2015)	Zimbabwe	GEV	Rainfall
Vidal (2014)	Chile	Gumbel	Rainfall
Chung and Kim (2013)	South Korea	Gumbel, GEV	Rainfall
Eli (2012)	Alor Setar, Malaysia	GEV	Rainfall

CHAPTER THREE
METHODOLOGY

3.1 Introduction

This study utilised the concentrations of ground-level ozone (O_3) in the unit of parts per million (ppm) furnished by the Department of Environment, Malaysia (DOE). The data was collected through continuous monitoring by Alam Sekitar Malaysia Sdn. Bhd. (ASMA), a private entity authorised by the DOE for sixty-five (65) monitoring stations throughout Malaysia. The monitoring stations are located in different areas such as in urban, suburban, and industrial to detect any changes that could be a danger to health and the environment. The concentrations for O_3 are recorded hourly. The Department of Environment Malaysia (2018) stated that the Malaysian ambient air quality guideline for O_3 is below 0.1 ppm based on one-hour concentration. The total number of exceedances were determined from the O_3 concentrations above 0.1 ppm.

In order to fulfil the objectives of this study, the flow of the methodology was illustrated in Figure 3.1. In achieving objective 1, which is to study the pattern and the behaviour of the O_3 concentrations was explore by using the descriptive statistics. The descriptive statistics were used to explore the pattern of the distribution of O_3 concentrations. To determine the occurrence of high concentrations will be presented using time series and Box-and-Whisker plot. The trend analysis for each monitoring stations will be used to determine if there is any significant increasing or decreasing trend of O_3 concentrations along the year. As for achieving objective 2 and objective 3, the Bayesian approach will be used to estimate the parameter of the distribution. There are two main components to fit into the Bayesian, which is likelihood distribution and prior distribution. Therefore, the likelihood is the EVD which is GEV and Weibull and the prior are selected from the family distribution of the EVD which is uniform for GEV and inverse gamma for Weibull are used to fit into the Bayes' theorem to estimate the posterior parameter. MCMC will be used in getting the suitable prior for the EVD that will explain further in Figure 3.4. The detailed explanation on estimation of parameter using Bayesian approach will be explained in Section 3.7.

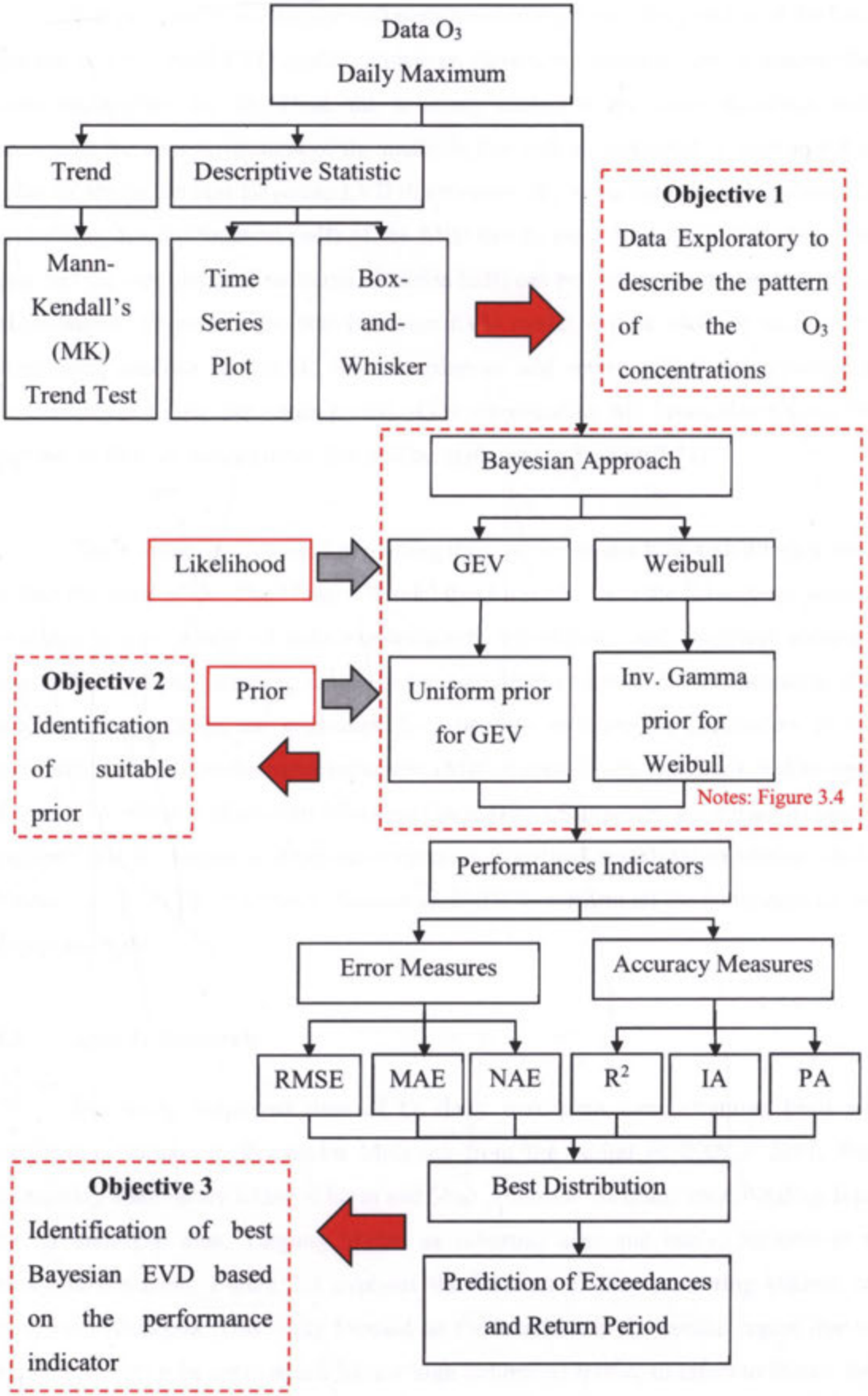


Figure 3.1 Flow of Methodology

The performance indicator will performed to determine the suitable of the prior and the best Bayesian EVD model. However, there is no consensus about which is the most appropriate for O_3. Thus, the accuracy measures and error measures shall determine the appropriateness of the methods that will be discussed in Section 3.8 in order to obtain the best Bayesian EVD distribution. By using the best distribution, the probability density function (pdf) of the EVD can be plot to compare with the actual data and the cumulative distribution function (cdf) can be used to get the probability of exceedances. Therefore, the best Bayesian EVD model will be used for each of the monitoring stations to predict the exceedances and return period of extreme O_3 concentration. Lastly, the return period of O_3 concentration will be validated using the number of days of exceedances that will be explained in Section 3.11.

The R language was used in plotting the time series and Box-and-Whisker plot, to find the trend of O_3. The Mann – Kendal trend test also used the R language since it provides a wide range of data manipulation, calculation, and graphical analysis. Matlab® is the programming language for numerical computation, visualisation, and programming package for engineers. It is used to estimate the parameters of the distribution and the performance indicators (MathWorks, 2015). OpenBUGS is an open source of WinBUGS (Bayesian Inference Using Gibbs Sampling), a software designed explicitly for the Bayesian Analysis of complex statistical models using Markov chain Monte Carlo (MCMC) methods (Ntzoufras, 2009) to perform all the simulation of the Bayesian model.

3.2 Area of Research

The study employed data of O_3 daily maximum concentrations from six monitoring stations in Peninsular Malaysia from the period of 2008 – 2017. The monitoring stations are Klang, Cheras and Shah Alam for the urban area, Petaling Jaya for an industrial area, Tanjung Malim as suburban area and lastly, Jerantut as a background station. Figure 3.2 explains the location of the monitoring stations in Peninsular Malaysia. This study focused on the locations in the central region due to the titration of O_3 in areas which have a high volume of traffic, in effect to reduce the amount of O_3 in city centers. The movement of O_3 precursors to suburban areas leads to a high concentration of O_3 in other areas (Banan *et al.*, 2013).

Most of the previous studies (Abdullah *et al.*, 2017; Ahamad *et al.*, 2014; Banan *et al.*, 2013; Nasir *et al.*, 2016) used Port Klang's monitoring station as it has exceedances of O_3. The Klang Valley is Malaysia's most developed region and is more susceptible to air pollution due to its geographical location, the development of large-scale industrial and commercial activities, densely populated areas and also high vehicular traffic. The air monitoring station is located at Port Klang as it lies within a port city with many heavy industries (Ahamad *et al.*, 2014).

Cheras is in the federal territory of Kuala Lumpur which is situated in the middle of Malaysia. The population estimated was high because of job offers. Cheras is a busy town with a lot of traffic. Therefore, due to the congestion of traffic, development, and population, Cheras has become the ideal place to study urban O_3 concentration (Ghazali *et al.*, 2014). Fadzly *et al.* (2018) also stated that Cheras was found as the haziest stated during the 10 days studied due to the smoke released by the vehicles and factories around the area.

The Shah Alam station is located in a residential and commercial area surrounded by extremely busy motorways. This station is also close to the industrial zone, which leads to high O_3 concentration. Majority of the urban stations has high daytime of O_3 concentration. Urban area such as Shah Alam frequently has a high O_3 concentration that surpasses the limits imposed by the MAAQG (Awang *et al.*, 2016).

Ahamad *et al.* (2014) stated that the Petaling Jaya station is an urban-traffic in nature as the monitoring station is located next to a busy intersection. The presence of NO_x from motor vehicle emissions is more likely to influence these stations by traffic conditions and the effect of the titration of O_3. Previous researchers studied air pollution in this area since the Klang Valley has been an environmental concern for researchers (Ahamad *et al.*, 2014; Mabahwi *et al.*, 2015; Mohamad, Ash'aari, & Othman, 2015).

The Tanjung Malim station is located at Universiti Pendidikan Sultan Idris (UPSI); the northernmost area, 80 km from Kuala Lumpur. The station is located about 5 km from the main road (North-south Highway), and it is surrounded by many residential and construction areas, including the main highway in Malaysia, which can

catalyse the formation of O_3 (Mokhtar *et al.*, 2016). Tanjung Malim was selected instead of Kuala Selangor, and Banting due to extreme O_3 concentration, which was low in Kuala Selangor and because of insufficient data in Banting since the monitoring stations started in 2010.

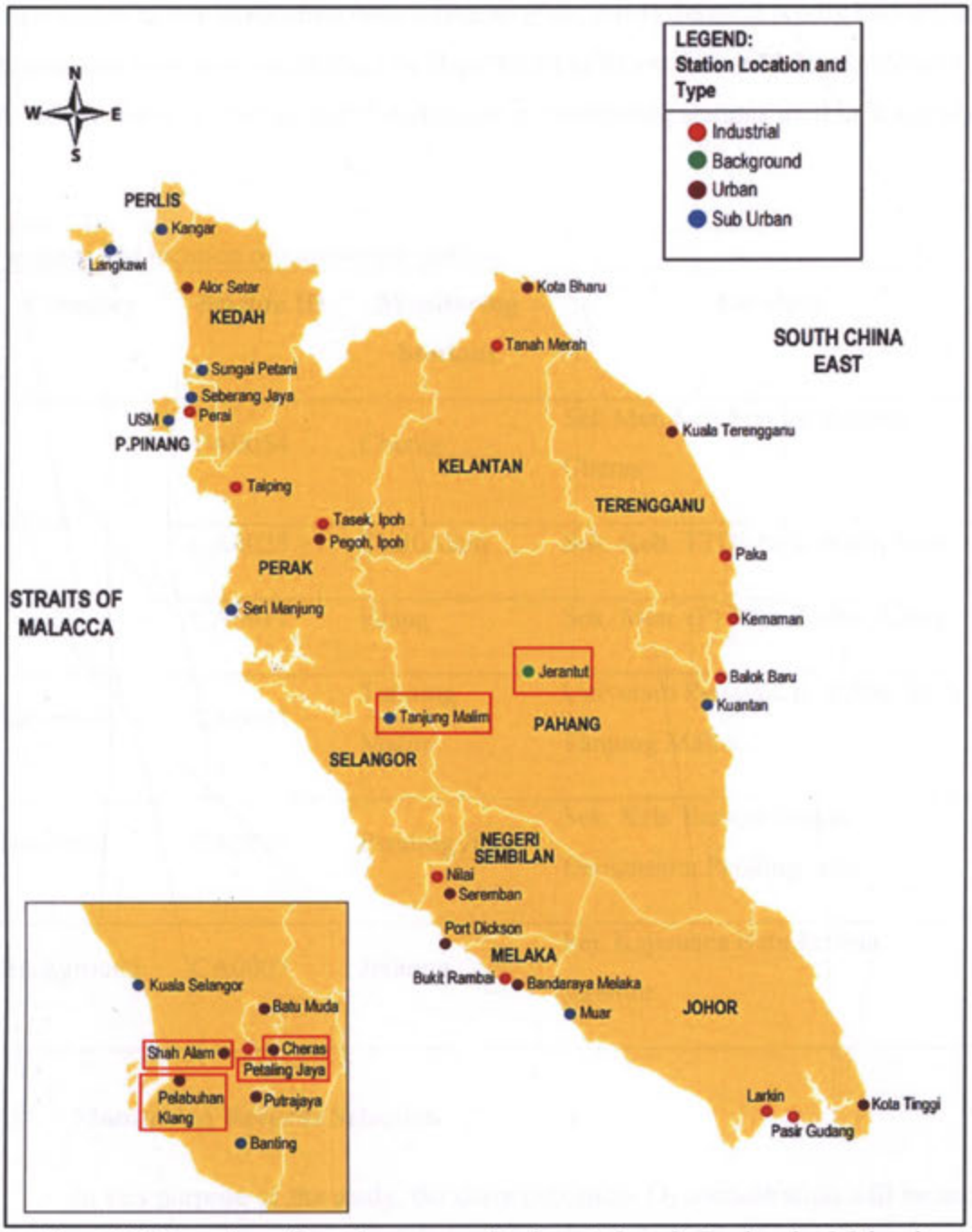


Figure 3.2 Location of continuous monitoring stations in Peninsular Malaysia (Source: Department of Environment Malaysia, 2017)

The last monitoring station for this study is Jerantut in Pahang. It is 200 kilometres away from Kuala Lumpur and 180 kilometres from Kuantan. This monitoring station is located at the Meteorological Department of Malaysia in Batu Embun, Jerantut, Pahang in the central region of Peninsular Malaysia. Natural forest, soil dust, open burning, and a small number of motor vehicles are expected to contribute to air quality at this monitoring station (Banan *et al.*, 2013). Jerantut is only background stations that have been established by Department of Environment Malaysia (Awang *et al.*, 2013). Table 3.1 details the classification of monitoring stations used in this study.

Table 3.1
Category and location of monitoring stations

Category	Station ID	Monitoring Stations	Location
Urban	CA0054	Cheras	Sek.Men.Keb.Seri Permaisuri, Cheras
	CA0025	Shah Alam	Sek. Keb. TTDI Jaya, Shah Alam
	CA0011	Klang	Sek. Men. (P) Raja Zarina, Klang
Sub-urban	CA0045	Tanjung Malim	Universiti Pendidikan Sultan Idris, Tanjung Malim
Industrial	CA0016	Petaling Jaya	Sek. Keb. Bandar Utama Damansara, Petaling Jaya
Background	CA0007	Jerantut	Pej. Kajicuaca Batu Embun, Jerantut

3.3 Monitoring Records Selection

In this purpose of the study, the daily maximum O_3 concentration will be used for the monitoring record selection for the maximum value. The straightforward method of the daily maximum data is selected from the maximum of the series of 24-hour monitoring record from 00:00 to 23:00 hours.

3.4 Missing Value Treatment

There are various reasons why missing values in O_3 data occur, i.e., a malfunction of equipment, human error, and calibration process. Complete data are required in performing time series analysis, and trend analysis in this study. Therefore, this study had to overcome this problem. There are several methods in treating the missing values, and one of the most popular approaches to overcoming these missing values is by using listwise deletion method. Listwise deletion method is simply removing the missing data and uses the remaining data set for analysis. Even though this method is easy to implement, but this method introduces a substantial bias in this study (Little & Rubin, 2002). Another method to overcome this problem is the adoption of imputation techniques (Junninen, Niska, Tuppurainen, & Ruuskanen, 2004).

After trying a few techniques of imputation such as interpolation, daily mean, and a few other techniques, K-nearest neighbor (KNN) was the most suitable imputation for the O_3 data in this study. A hot-deck imputation that implies the nearest – neighbour method was used where the missing values were filled by the mean value of the corresponding column of the nearest neighbour of the corresponding row that has no missing values. The nearest neighbour can be defined in terms of Euclidean distance.

Based on Ahmat Zainuri, Jemain, and Muda (2015), the KNN method is one of the best methods that perform consistently superior compared to the other method in the study. The method used in the study was tested using a different set of ground-level ozone (O_3) data from eight monitoring stations located at the central, south and north region of Peninsular Malaysia with 5%, 10%, 15%, 20%, 25% and 30% missing. Three performance indicators used in the study agree that KNN is one of the best methods with a higher correlation coefficient and index agreement and with a lower minimum absolute error compared to other methods used in the study. This study also recommended KNN as the preferred method to be used in air quality data.

3.5 Data Exploration

The O_3 daily maximum concentration will be presented in descriptive statistics, time series plot, and Box-and-Whisker plot. While, the trend analysis will be tested

using Mann – Kendall (MK) trend test to test if there is significant decreasing or increasing test in the monitoring stations from 2008 – 2017.

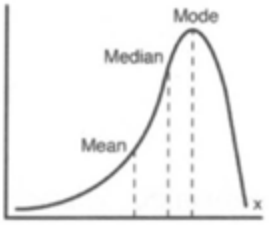
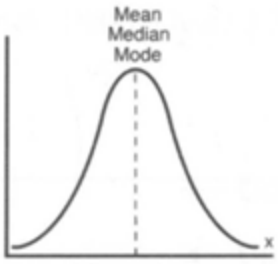
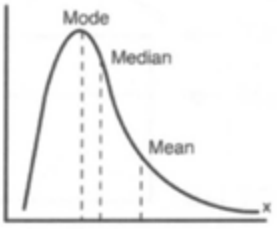
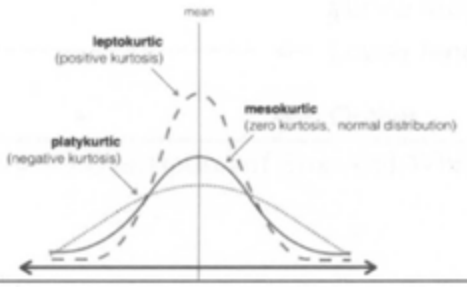
3.5.1 Descriptive Statistics

The descriptive statistics is the discipline of quantitatively describing the main features of a collection of information or the quantitative description in the form of numerical or graphical techniques. Descriptive statistics, in general, represent three types of information about the data, namely: location, dispersion, and shape. The mean, median, and mode represent the location of the data. Dispersion is to determine the spread of the data from its mean. The range, standard deviation, coefficient of variation, percentile, and interquartile range are the measurement of dispersion. The shape of the data distribution is presented by variance, skewness, and kurtosis. These components are classified as the measures of central tendency (Fisher & Marshall, 2009).

Table 3.2
Measurement used in Descriptive statistics

Category	Notation	Description	Formula
Location	Mean, $\bar{x}$	The average set of data	$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
	Median	The middle score of rank-ordered values	If n is odd, $Median = x_{(n+1)/2}$ If n is even, $Median = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}$
	Mode	The numerical value with the most frequency	
Dispersion	Range	The difference between the highest and lowest value	$x_n - x_i$
	Standard Deviation, s	The average difference between each data to the mean	$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$
	Interquartile range, IQR	The difference between quartile 3 and quartile 1	$Q_3 - Q_1$

Table 3.2
Continued

Shape	Skewness, C_s	The measure of the asymmetry of the distribution of data about its mean Negatively (left) skewed Data mostly lie on the right of the mean  (a) Negatively Skewed Normal (skewness = 0) Data evenly distributed on the left and right of the mean $C_s = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^3$  (b) Normal (no skew) Positively (right) skewed Data mostly lie on the left of the mean  (c) Positively skewed (Durkhure & Lodwal, 2014)
	Kurtosis	The measure of the peakedness of the distribution of data.  $Kurtosis = \frac{1}{n-1}\sum_{i=1}^{n}\frac{x_i - \bar{x}}{s}$

3.5.2 Time Series Plot

A time series is a plot of the sequence of data points, usually consisting of successive measurements made over a time interval. The monitoring records used in this study are from 2008 – 2017. Since the monitoring records for O_3 concentrations were in series, the time series plot can be applied.

3.5.3 Box-and-Whisker Plot

Box-and-Whisker plot visually summarises and compares the groups of data. The presentation of the Box-and-Whisker plot utilises the median, quartiles, the lowest and highest data points to express the dispersion and symmetry of the distribution of the data. Box-and-Whisker plot can easily demonstrate the existence of outliers in the data (Williamson, Parker, & Kendrick, 1989). Figure 3.3 shows the description of the Box-and-Whisker plot.

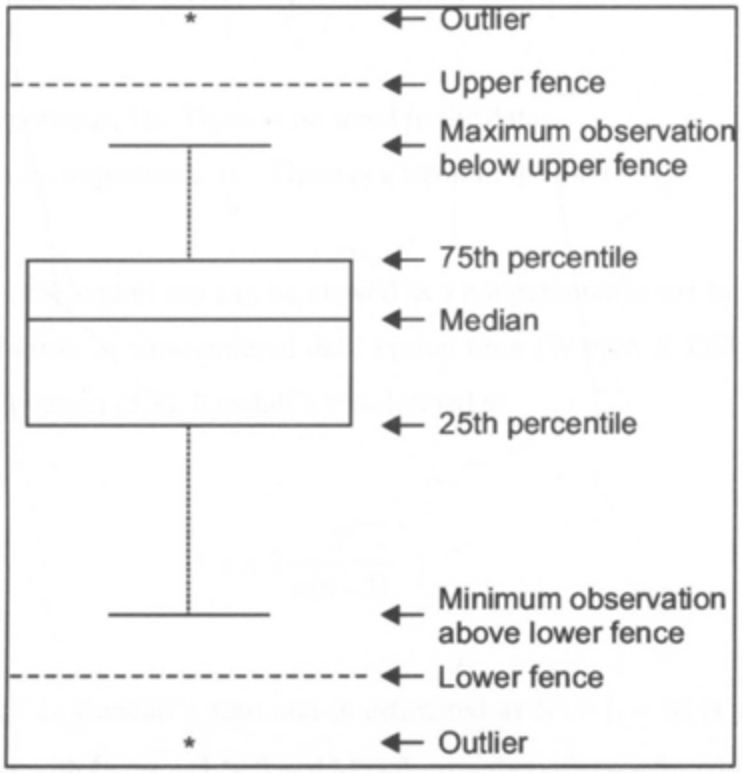


Figure 3.3 Description of Box-And-Whisker Plot

3.5.4 Mann-Kendall's (MK) Trend Test

Mann initiated the test in 1945 (McLeod, 2011), which statistically assessed the existence of an increasing or decreasing trend within a time series. The non-parametric Mann-Kendall test for the trend can be applied over ten years to the daily maximum of O_3 values for each subregion. The procedure entails listing the time-ordered medians from 1980 to 1989 and computing the n (n - 1)/2 possible differences $x_i - x_j$ where i > j. The sign ($x_i - x_j$) becomes an indicator function that takes on the value 1, 0, or -1 according to the sign $x_i - x_j$. The Mann- Kendall statistic is written in equation (3.1):

$$S = \sum_{i=2}^{n} \sum_{j=1}^{i-1} sign(x_i - x_j) \tag{3.1}$$

where *S* is the number of positive differences minus the negative differences. If S is a large negative (positive) number, ambient measurements taken later tend to be smaller (larger) than those taken earlier. The hypothesis of this test is:

Null hypothesis, H_0: There is no trend in the data
Alternative hypothesis, H_1: There is a trend in the data

The Mann-Kendall test can be viewed as a nonparametric test for zero slopes of the linear regression of time-ordered data versus time (Warren & Gilbert, 1988). As mentioned in Equation (3.2), Kendall's τ is defined as

$$\tau = 2\frac{S'}{n(n-1)} \tag{3.2}$$

where S' is Kendall's sum and is 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 (3.3) and Equation (3.4)

$$Z = \begin{cases} \dfrac{S-1}{Var(S)'}, & S > 0 \\ 0, & S = 0 \\ \dfrac{S+1}{Var(S)'}, & S < 0 \end{cases} \tag{3.3}$$

And

$$Var(S) = \frac{n(n-1)(2n+5) - \sum_{p=1}^{q}(t_p - 1)(2t_p + 5)}{18} \tag{3.4}$$

where t_p is the number of ties for the p^{th} values, and q is the number of tied values. For the two-sided test for trend, the null hypothesis, H_0 is rejected when $|Z| > Z_{\alpha/2}$, where $\alpha = 0.05$ is the significance level.

Hirsch, Slack, and Smith (1982) stated that the trend magnitude could be estimated using Kendall's slope estimator, and it is an unbiased estimator of the linear trend slope. It provides higher accuracy than a regression estimator where data is highly skewed but lowers accuracy when data is normal.

3.6 Extreme Value Distribution

The quality of procedures used in statistical analysis relies significantly on the assumed probability model or distribution. As a result, the development of large sets of standard probability distributions together with appropriate statistical methodologies has increased considerable effort to provide statistical models for a wide range of real-world phenomena. The standard model would not be a useful probability density function (pdf) for the study of each phenomenon. One of the statistical procedures is the measures of central tendency. The measure of central tendency is a single value that attempts to explain a set of data by recognizing the central position within that set of data. On this note, it is not suitable to analyse the set of extreme data (Aryal & Tsokos, 2009).

Nasir *et al.* (2016) used the probability density function (pdf) and cumulative density function (cdf) to estimate the probability of the exceedances of O_3 based on the

Malaysian Ambient Air Quality Guideline (MAAQG). Therefore, this study used the pdf plot to identify the skewness and plotting using the parameter estimation value compared with the observed data. On the other hand, the cdf was used to determine the probability of O_3 concentration.

3.6.1 Generalized Extreme Value (GEV) Distribution

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 popularised by Jenkinson in 1955 (Bali, 2003), has three parameters, namely: location, shape, and scale. The location parameter, μ determines 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. The shape parameter, λ on the other hand, affects the shape of distribution and tails of the distribution. It governs the skewness as to where the majority of data are concentrated, thus, creates the tail(s) of distribution (Millington, Das, & Simonovic, 2011). The pdf is written in Equation (3.5)

$$f(x;\sigma,\lambda,\mu) = \frac{1}{\sigma}\left[1+\lambda\left(\frac{x-\mu}{\sigma}\right)^{-\frac{1}{\lambda}-1}\right]\exp\left\{-\left[1+\lambda\left(\frac{x-\mu}{\sigma}\right)^{-\frac{1}{\lambda}}\right]\right\} \qquad (3.5)$$

where $x > \mu - \frac{\sigma}{\lambda}, -\infty < \mu < \infty$ 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 is when $\lambda = 0$. Type II – Frechet has $\lambda > 0$ and it has bounded lower side when $x > 0$ (Aryal & Chris, 2011). 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 (3.6)

$$F(x;\lambda,\sigma,\mu)=\begin{cases}\exp\left\{-\left[1+\lambda\left(\frac{x-\mu}{\sigma}\right)^{-\frac{1}{\lambda}}\right]\right\} & \lambda\neq 0 \\ & and\ for\ 1+\lambda\left(\frac{x-\mu}{\sigma}\right)>0 \\ \exp\left\{-\exp\left(-\frac{x-\mu}{\sigma}\right)\right\} & \lambda=0\end{cases} \quad (3.6)$$

3.6.2 Weibull Distribution

The EVD of Type III was formulated by Waloddi Weibull (1887-1979), a Swedish engineer and scientist well-known for his work on the strength of materials and fatigue analysis (Georgopoulos & Seinfeld, 1982). The pdf and cdf for two Weibull distributions are written in Equation (3.7) - (3.8).

$$f(x;\sigma,\lambda)=\frac{\lambda}{\sigma}\left(\frac{x}{\sigma}\right)^{\lambda-1}\exp\left[-\left(\frac{x}{\sigma}\right)^{\lambda}\right] for\ x\geq 0;\sigma,\lambda>0 \quad (3.7)$$

$$F(x;\sigma,\lambda)=1-\exp\left[-\left(\frac{x}{\sigma}\right)^{\lambda}\right] for\ x\geq 0;\sigma,\lambda>0 \quad (3.8)$$

where λ is the shape parameter and σ is the scale parameter (Rinne, 2009).

3.7 Bayesian Approach

The Bayesian approach using EVT as prior distribution is more familiar in the hydrology field (Chung & Kim, 2013; Eli, 2012; Smith, 2005). Regardless of that, Bayesian approach using EVT for prior distribution is not common in air pollution data and in obtaining the posterior distribution for a complex model can be problematic because of the difficulty in calculating the integral techniques (Coles, 2001).

There are three objectives stated by Kruschke (2010) in the inference from data which are: (i) Estimation of parameter values, (ii) Prediction of data values and (iii) Model selection. From the given set of beliefs, the posterior belief of the parameters

and values are estimated. One of the principals used in a mathematical model is the ability to make a specific prediction. It stated that complex models usually fit data better than the simple model.

For this study, the Bayesian approach used OpenBUGS – computing-language-oriented software that utilises Bayesian inference using Gibbs sampling. OpenBUGS has become popular in recent years because it can estimate the posterior distribution of the parameters of interest using MCMC in a variety of models. It only specifies the model code in which the model likelihood and the prior distribution are defined. In this approach, the Markov Chain Monte Carlo (MCMC) method is employed in estimating the unknown parameters. The MCMC method was initiated by Nick Metropolis and his associates in the 1950s and (re)discovered by Gelfand and Smith in 1990. The selection of prior distribution is essential for better convergence towards the equilibrium distribution. At equilibrium, the MCMC generates dependent random values from the corresponding stationary distribution (Ntzoufras, 2009). The sets of Bayes' rule is as written in Equation (3.9)

$$\underbrace{p(\theta \mid D)}_{posterior} = \underbrace{p(D \mid \theta)}_{likelihood} \times \underbrace{p(\theta)}_{prior} / \underbrace{p(D)}_{evidence} \tag{3.9}$$

where the evidence is as stated in Equation (3.10)

$$p(D) = \int d\theta p(D \mid \theta) p(\theta) \tag{3.10}$$

The prior, $p(\theta)$ is the initial set of beliefs in θ without the data, D. The posterior, $p(\theta|D)$ is the set belief in θ when the data D is taken into consideration. The likelihood, $p(D|\theta)$ is the probability that the data could be generated by the model with parameter values θ (Kruschke, 2010). In this study, the concept of Bayesian is represented in Equation (3.11).

$$Posterior = \underbrace{likelihood}_{distribution\ for\ O_3} \times \underbrace{prior}_{distribution\ for\ parameters} \tag{3.11}$$

3.7.1 Non-Informative Prior

There are two types of prior distributions which are informative and non-informative prior. When a prior distribution is derived using data through objective analysis, it is referred as the informative prior distribution. Otherwise, it is referred as a non-informative prior distribution when derived from subjective judgments or theoretical considerations. When the non-informative prior distribution is used, the posterior distribution only reflects the information in the sample (Chung & Kim, 2013).

Chung and Kim (2013) stated that the Bayesian analysis does not always provide the reduction of the uncertainty, especially when the information is a large sample size in defining the unknown parameters, the influence of the uncertainty is weak in determining a specific decision. It is not needed to spend large amounts of time and energy for the development of an informative prior distribution. However, if there is little information, the analysis of the uncertainty has a strong influence on the final selection of the parameters. Therefore, in this study, the non-informative prior distribution such as conjugate prior distribution can be used since there is a large amount of data existing in O_3. However, if the sample size is small or only indirect information about the known parameters is available, it becomes more important to choose a reasonable prior distribution such as informative distribution.

Ahmat (2016) conducted a study on Bayesian comparing informative and non-informative prior in predicting the high concentration of PM_{10} using extreme value distribution by using the daily maximum data of PM_{10} and the EVD used were GEV, GPD, Weibull, and Gumbel. As for non-informative prior, uniform and normal distribution were used for the GEV, Inv. Gamma distribution was used for the Weibull distribution, and normal distribution was used for the GPD. The finding was in favour of non – informative prior in which the non – informative prior fitted the observation better compared to the informative prior for the eight (8) monitoring stations used in the study using GEV with a uniform prior.

The choice of the uniform distribution is because the prototype of the GEV Type III is the uniform distribution over some interval. It suggests that the GEV is in the same

family of the uniform distribution. The Weibull distribution interpolates between the exponential distribution, and thus, the choice of non-informative prior will be chosen from within the exponential family (Rinne, 2009) which is inverse gamma. The parameter was obtained for each location, as stated in Figure 3.4.

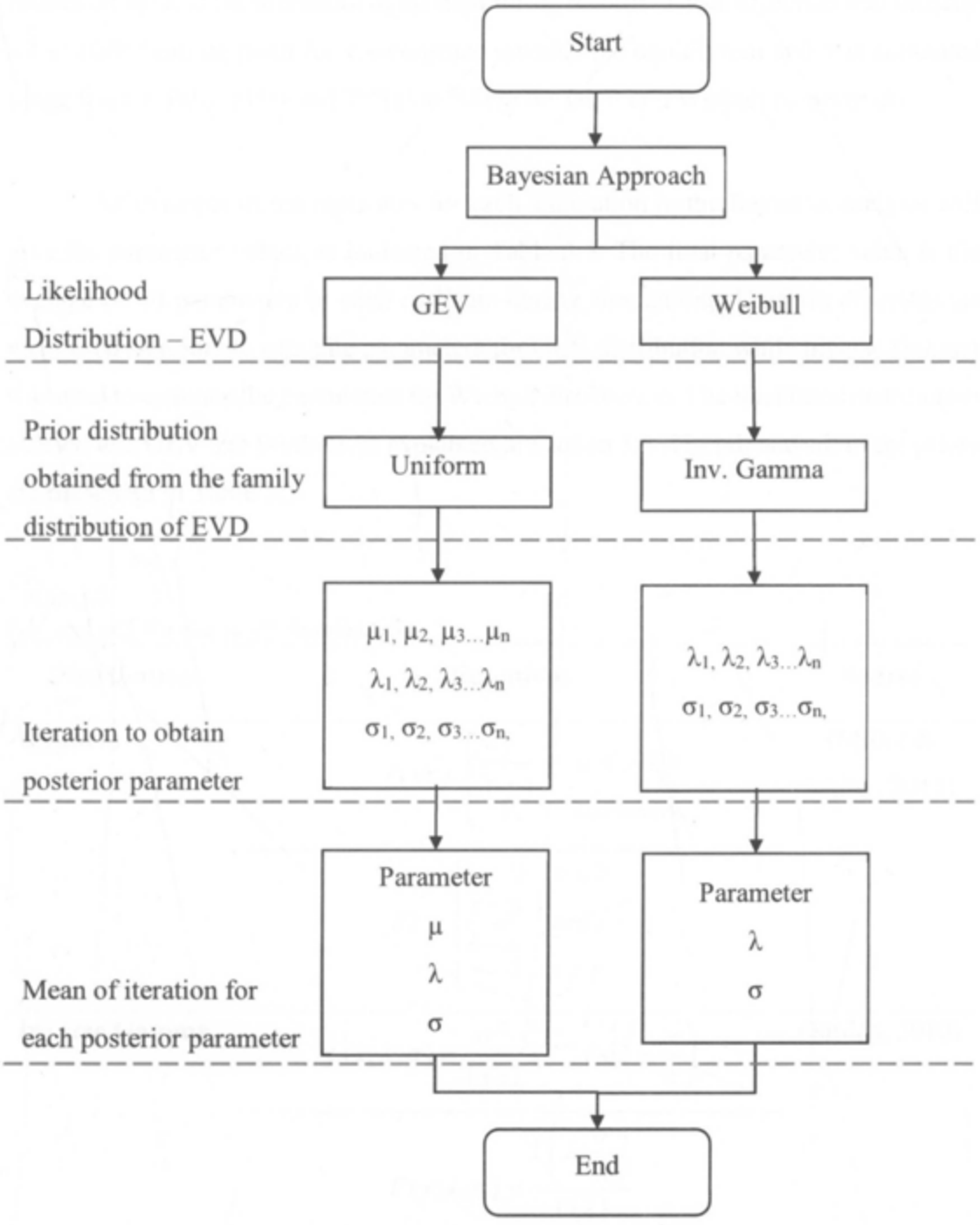


Figure 3.4 Flowchart of obtaining parameter

As stated in Table 3.4, for GEV distribution, the location, scale and shape parameters were estimated using uniform distribution with means equal to 0, 0 and -1, and variances equal to 200, 50 and 1. As per Weibull distribution, the scale and shape parameters were estimated using Inverse Gamma with means equal to zero and variances equal to 100 and 10, respectively. In regards to the location parameter, the estimated value is the minimum of the monitoring records. The distribution was initially set at 1000 burning point for convergence towards the equilibrium and was simulated range from 4000 to 8000 and 20000 to 30000 for GEV and Weibull respectively.

An example of ten replicates for each simulation in the Bayesian analysis will give the parameter values as indicated in Table 3.5. The final parameter value is the average of all parameters in each replicate during simulations. Uniform distributions were used as priors to estimate parameters for GEV distribution while Inverse Gamma was used to estimate the parameters for Weibull distribution. The likelihood distribution utilised was GEV and Weibull, as explained in Section 3.6. The pdf and cdf of the priors are presented in Table 3.3

Table 3.3
Pdf and cdf for the prior distribution

Distribution	Equations	Source
Uniform	$f(x) = \begin{cases} \frac{1}{b-a}, & a \le x \le b \\ 0, & elsewhere \end{cases}$	(Miller & Miller, 2013)
	$Fx = \begin{cases} 0 & x < 0 \\ \frac{x-a}{b-a} & a \le x < b \\ 1 & x \ge b \end{cases}$	
Inverse Gamma	$f(x;\lambda,\sigma) = \frac{\sigma^{\lambda}}{\Gamma(\lambda)} x^{-\lambda-1} \exp\left(-\frac{x}{\sigma}\right)$	(Jordan, 2010)
	$F(x;\lambda,\sigma) = \frac{\Gamma\left(\lambda, \frac{\sigma}{x}\right)}{\Gamma(\lambda)}$	
	Where σ is the scale parameter, λ is the shape parameter and Γ is the gamma distribution	

Table 3.4
The non-informative prior distribution for each monitoring stations with iterations

Monitoring Stations	EVD	Non-informative Prior	Iterations
Petaling Jaya	GEV	$\mu \sim$ uniform (0,200) $\sigma \sim$ uniform (0,50) $\lambda \sim$ uniform (-1,1)	5000
	Weibull	$\sigma \sim$ invgamma (0,100) $\lambda \sim$ invgamma (0,10)	30000
Shah Alam	GEV	$\mu \sim$ uniform (0,200) $\sigma \sim$ uniform (0,50) $\lambda \sim$ uniform (-1,1)	5000
	Weibull	$\sigma \sim$ invgamma (0,100) $\lambda \sim$ invgamma (0,10)	20000
Cheras	GEV	$\mu \sim$ uniform (0,200) $\sigma \sim$ uniform (0,50) $\lambda \sim$ uniform (-1,1)	4000
	Weibull	$\sigma \sim$ invgamma (0,100) $\lambda \sim$ invgamma (0,10)	30000
Klang	GEV	$\mu \sim$ uniform (0,200) $\sigma \sim$ uniform (0,50) $\lambda \sim$ uniform (-1,1)	5000
	Weibull	$\sigma \sim$ invgamma (0,100) $\lambda \sim$ invgamma (0,10)	25000
Tanjung Malim	GEV	$\mu \sim$ uniform (0,200) $\sigma \sim$ uniform (0,50) $\lambda \sim$ uniform (-1,1)	8000
	Weibull	$\sigma \sim$ invgamma (0,100) $\lambda \sim$ invgamma (0,10)	30000
Jerantut	GEV	$\mu \sim$ uniform (0,200) $\sigma \sim$ uniform (0,50) $\lambda \sim$ uniform (-1,1)	5000
	Weibull	$\sigma \sim$ invgamma (0,100) $\lambda \sim$ invgamma (0,10)	25000

Table 3.5
Determination of parameter using ten replicates

Replicates	Parameters		
	Location, μ	**Scale, σ**	**Shape, λ**
1	μ_1	σ_1	λ_1
2	μ_2	σ_2	λ_2
3	μ_3	σ_3	λ_3
4	μ_4	σ_4	λ_4
5	μ_5	σ_5	λ_5
6	μ_6	σ_6	λ_6
7	μ_7	σ_7	λ_7
8	μ_8	σ_8	λ_8
9	μ_9	σ_9	λ_9
10	μ_{10}	σ_{10}	λ_{10}
Average	$\bar{\mu}$	$\bar{\sigma}$	$\bar{\lambda}$

3.8 Performance Indicators (PI)

The fitting of the model is best carried out using a systematic optimisation routine that estimates the parameters by several distributions indicated in Section 3.6 and Section 3.7 using the O_3 concentrations data and then compares how these estimated distributions fit the data using various criteria of "goodness of fit". Three common error measures mainly; the Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and Normalized Absolute Error (NAE) and three accuracy measures : Prediction Accuracy (PA), coefficient of Determination (R^2) and Index of Accuracy (I.A) were used in this study (Ahmat *et al.*, 2015). For more understanding of the calculation of goodness-of-fit, Table 3.6 lists the notations needed in the calculation of goodness-of-fit, which are used in Table 3.7.

Table 3.6
Notations used in obtaining goodness-of-fit

	Past					Present Time
Observed Value	O_1	O_2	...	O_{t-2}	O_{t-1}	O_t
Period t	1	2	...	$t-2$	$t-1$	t
Estimated Values	P_1	P_2	...	P_{t-2}	P_{t-1}	P_t
Error Values	e_1	e_2	...	e_{t-2}	e_{t-1}	e_t

Table 3.7
Notations used in defining performance indicators

Notation	Meaning
N	Number of observed records
e_t	Forecast error, $O_t - P_t$
O_t	Observed records
$\overline{O}$	Mean of observation, $\frac{1}{n}\sum_{t=1}^{n} O_t$
P_t	Predicted records
$\overline{P}$	Mean of predicted records $\frac{1}{n}\sum_{t=1}^{n} P_t$
S_O	Standard deviation of Observed records $S_o = \sqrt{\frac{1}{n-1}\sum_{t=1}^{n}\left(O_t - \overline{O}\right)^2}$
S_P	Standard deviation of Predicted records $S_P = \sqrt{\frac{1}{n-1}\sum_{t=1}^{n}\left(P_t - \overline{P}\right)^2}$

3.8.1 Error Measures

The value of error measures 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 value of error measures have the lowest values using formulae (3.12)-(3.14) in Table 3.8. The error measures in this study are scale and unit-dependant (Ji & Gallo, 2006).

Table 3.8
Description of Error Measures and their formulae

Error Measures	Description	Formulae	Equation
RMSE	The average difference between the observed and predicted values	$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(O_t - P_t)^2}{n}}$	(3.12)
MAE	The average absolute difference between the observed and the predicted values	$MAE = \frac{\sum_{t=1}^{n}\lvert(O_t - P_t)\rvert}{n}$	(3.13)
NAE	The error between the predicted and observed values of a model or estimator and the calculation	$NAE = \frac{\sum_{t=1}^{n}\lvert(P_t - O_t)\rvert}{\sum_{t=1}^{n}O_t}$	(3.14)

(Source: Ji & Gallo, 2006)

3.8.2 Accuracy Measures

The formulae for each accuracy measures shown in Table 3.9 will be calculated for the selected model. The accuracy value fluctuates between 0 and 1, and as the value approaches 1, the model is appropriate (Ahmat, Yahaya, & Ramli, 2014). The closer the value to 1, the more appropriate the model to simulate the data.

Table 3.9
Description of Accuracy Measures and their formulae

Accuracy Measures	Description	Formulae	Equation
R^2	The statistical indicators to measure the accuracy of models or estimators	$R^2 = 1 - \frac{\sum_{t=1}^{n}(O_t - P_t)^2}{\sum_{t=1}^{n}(O_t - \overline{O})^2}$	(3.15)
PA	The proximity of predicted to the observed values	$PA = \sum_{t=1}^{n} \frac{(P_t - \overline{P})(O_t - \overline{O})}{(n-1)S_p S_o}$	(3.16)
IA	Dimensionless statistical indicator that expresses the difference between predicted and observed values	$IA = 1 - \frac{\sum_{t=1}^{n}(P_t - O_t)^2}{\sum_{t=1}^{n}(\lvert P_t - \overline{O}\rvert - \lvert O_t - \overline{O}\rvert)^2}$	(3.17)

(Source: Ji & Gallo, 2006)

3.9 The Exceedances

The estimation of the exceedances is calculated from the probability of concentrations exceeding 0.1 ppm which is obtained from the cumulative distribution function (cdf) plot multiplied with the number of days (Fitri *et al.*, 2011). Equation (3.18) was used in getting the probability of exceedances.

In calculating the exceedances, Equation (3.19) were used in getting the estimated days of exceedances of the distribution. Percent of Compliance in Equation (3.20) was used in order to check the similarities between the actual and estimated numbers.

$$Probability\ of\ exceedances = P(X > 0.1) = 1 - P(X \leq 0.1) = 1 - F(0.1) \quad (3.18)$$

$$Exceedances = P(X > 0.1)\ x\ number\ of\ days \quad (3.19)$$

$$\%\ Compliance = (1 - \frac{|\ actual - estimated\ |}{actual})\ x\ 100 \quad (3.20)$$

3.10 The Return Period

According to Coles (2003), Bayesian analysis is preferable due to the prediction of return level, which is based on the predictive distribution as it can be estimated easily. The value of the return period for every location is the reciprocal of the probability of exceedances (Selaman, Said, & Putuhena, 2007).

In this study, the probability of concentrations exceeding 0.1 ppm was obtained from the cumulative distribution function (cdf) plot of the best distribution chosen. The formula used is as indicated in Equation (3.21), and the number of exceedances can be predicted using the Equation (3.22) (Noor, *et al.*, 2011).

$$Return\ Period = \frac{1}{Probability\ of\ exceedances} \tag{3.21}$$

$$Number\ of\ exceedances = \frac{Number\ of\ days}{Probability\ of\ exceedances} \tag{3.22}$$

3.11 Validation

For the validation process, only two years of data were used, which is 2016 and 2017 for every monitoring stations selected. This process is to verify the estimated number of exceedances and the actual number of exceedances. Thus, the percent compliance that has been discussed in Section 3.9 were used.

CHAPTER FOUR
RESULT AND ANALYSIS

4.1 Introduction

The findings for this study are presented in Section 4.1 through Section 4.8. Objective 1, which is to describe the characteristics for each monitoring stations, is discussed in this section. Then, the results for objective 2 – the findings of the suitable prior are discussed in Section 4.5. The findings are based on the performance indicators along with the results for objective 3, which is to determine the best distribution. The exceedances and the return period in this study are discussed in Section 4.7 and Section 4.8, respectively.

4.2 Missing Value Treatment

From the discussion in Section 3.4, the missing value occurs in every location, which was due to various reasons. Therefore, this study imputed all the missing values using KNN hot-deck imputation. As shown in Table 4.1, there are more than 10% of missing values in the monitoring stations at Shah Alam and Klang. There are more than 5% of missing values for Cheras and Jerantut. As for Petaling Jaya and Tanjung Malim, the missing values are less than 3%. Daily time series plot with the missing values for all monitoring stations and after the imputation is illustrated in Figure 4.1.

Table 4.1
Percentage of missing values

Monitoring Stations	Missing values	Percent missing
Petaling Jaya	96	2.63%
Shah Alam	655	17.93%
Cheras	209	5.72%
Klang	419	11.47%
Tanjung Malim	73	2.00%
Jerantut	258	7.06%

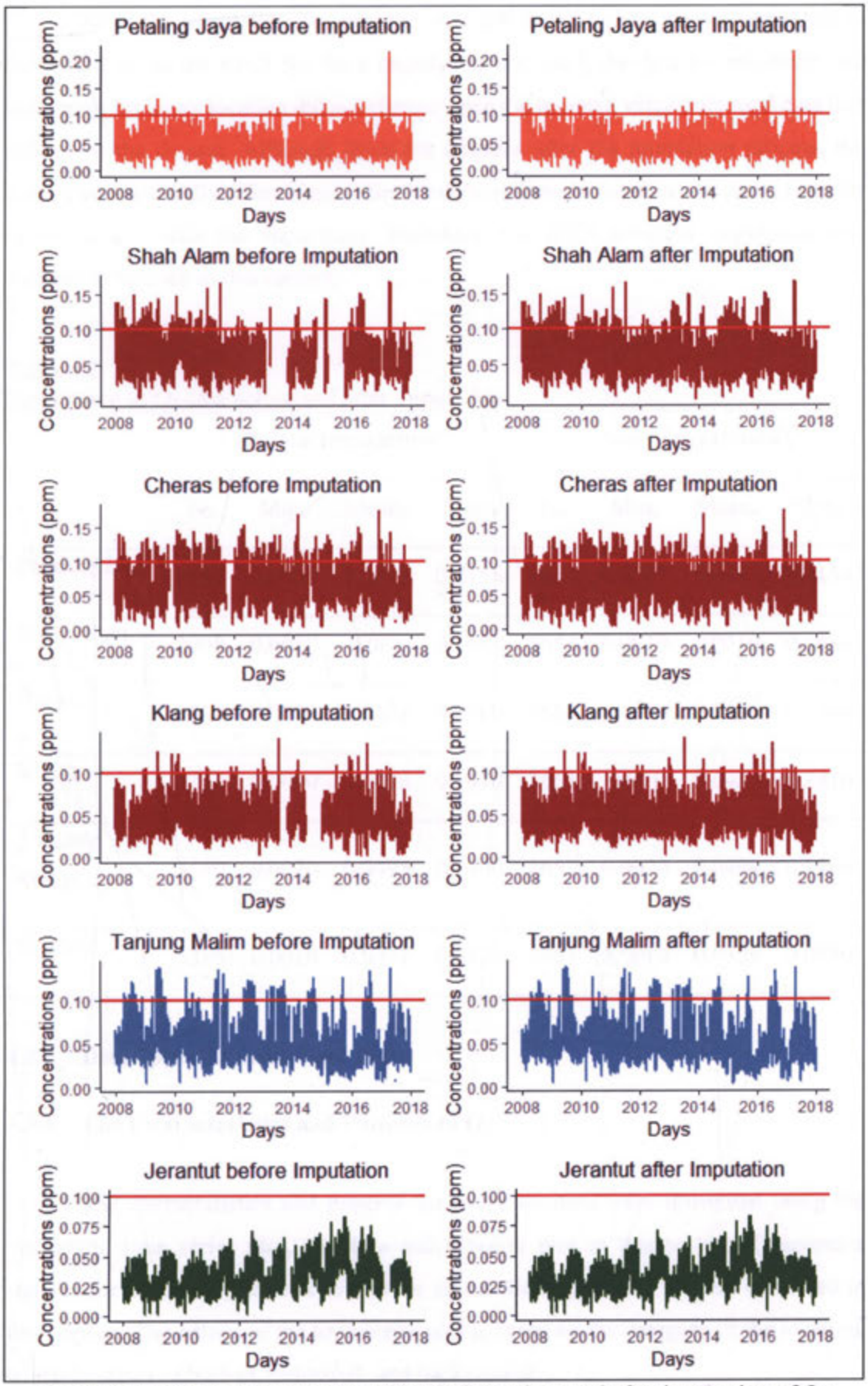


Figure 4.1 Daily time series plot with missing values and after imputation of O_3 concentrations

After the imputation process, the descriptive of O_3 data were summarised in Table 4.2. Since the KNN hot deck imputation was used, the data for minimum and maximum for every location did not change since the imputed value only used data that existed in the dataset. Although there are changes after the imputation process, the dataset was not really affected since the mean before the imputation was almost similar to the dataset after the imputation. Therefore, this KNN hot-deck imputation was suitable to be used in this dataset.

Table 4.2
Descriptive of O_3 data before and after imputation

	Before Imputation				**After Imputation**			
	N_b	Min_b	$Mean_b$	Max_b	N_a	Min_a	$Mean_a$	Max_a
Petaling Jaya	3557	0.0003	0.0495	0.2156	3653	0.0003	0.0497	0.2156
Shah Alam	2998	0.0010	0.061	0.1681	3653	0.0010	0.0619	0.1681
Cheras	3444	0.0010	0.0658	0.1760	3653	0.0010	0.0656	0.1760
Klang	3234	0.0010	0.0486	0.1410	3653	0.001	0.0483	0.1410
Tanjung Malim	3580	0.0020	0.0495	0.1390	3653	0.0020	0.0494	0.1390
Jerantut	3395	0.0010	0.0331	0.0830	3653	0.0010	0.0329	0.0830

4.3 Descriptive Statistics and Plots

4.3.1 The Characteristics and Patterns of O_3

The characteristics and patterns for every location were illustrated using the histogram, time series plot, and Box-and-Whisker plot in this section. Descriptive statistic for an annual year and overall for all monitoring stations are also discussed in this section. The colour of the time series plot is based on the category of the location, which is urban, suburban, industrial, and background.

i) Petaling Jaya

The distribution of O_3 concentration in Petaling Jaya shown in Figure 4.2 was nicely shaped with a slight skew to the right. The distribution shows the existence of extreme values. The maximum daily concentrations for Petaling Jaya are mostly below 0.1 ppm, although the extreme value occurs every year, as shown in the Box-and-Whisker Plot in Figure 4.3.

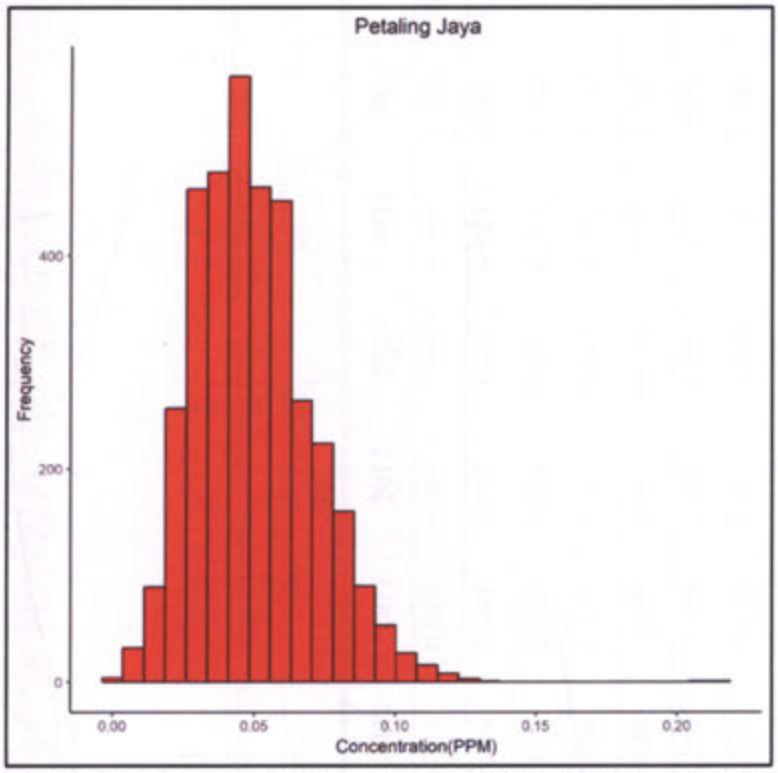
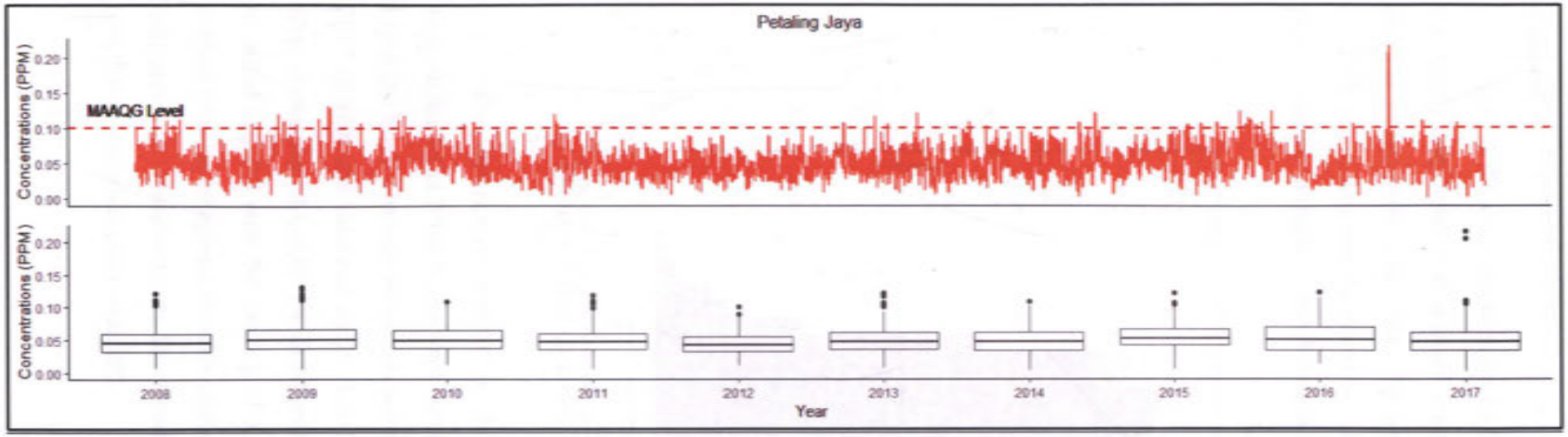


Figure 4.2 Histogram of O_3 concentrations for Petaling Jaya

Based on Figure 4.3, the annual averages of concentrations in Petaling Jaya were around 0.05 ppm, which were below than MAAQG level. The concentrations were all positively skewed with the highest value in 2017 (1.908), an indication of more extreme concentrations recorded in 2017. All the annual maximum concentrations were above the MAAQG for daily level. The highest maximum concentrations were recorded in 2017 with the readings of 0.216 ppm. Distribution for 2017 registered the highest peak as compared to other years with the kurtosis value of 1.951. The analysis indicates low variability in the monitoring records every year with the overall coefficient of variation value of 0.417.



	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	Overall
Mean	0.048	0.052	0.051	0.048	0.044	0.049	0.049	0.054	0.053	0.049	0.050
Median	0.045	0.050	0.049	0.047	0.042	0.047	0.047	0.052	0.05	0.046	0.048
Min	0.006	0.004	0.010	0.004	0.009	0.006	0.002	0.004	0.012	0.0003	0.0003
Max	0.121	0.131	0.108	0.119	0.101	0.121	0.108	0.122	0.124	0.216	0.216
Std. Deviation	0.020	0.022	0.020	0.019	0.017	0.019	0.020	0.020	0.024	0.024	0.021
Coefficient of Variations	0.422	0.418	0.389	0.396	0.378	0.390	0.419	0.368	0.450	0.486	0.417
Skewness	0.729	0.687	0.387	0.523	0.492	0.408	0.424	0.345	0.542	1.908	0.768
Kurtosis	0.424	0.749	-0.450	0.399	-0.042	0.329	-0.229	0.032	-0.380	9.877	1.951

Figure 4.3 Time-Series plot, Box-and-Whisker plot and Descriptive Statistics for Petaling Jaya

ii) Shah Alam

As for the O_3 concentration in Shah Alam, as shown in Figure 4.4, the shape was nicely distributed with a slight skew to the right. The data distribution shows that extreme values exist in the data. The annual average of maximum daily concentration for Shah Alam was mostly below 0.1 ppm, although the extreme value occurs every year, as shown in the Box-and-Whisker Plot in Figure 4.5.

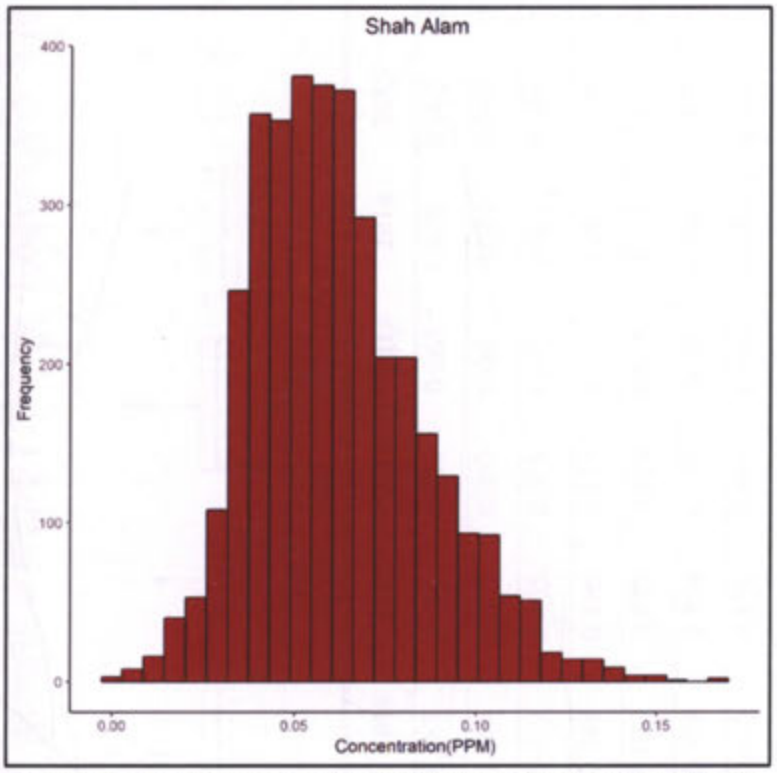
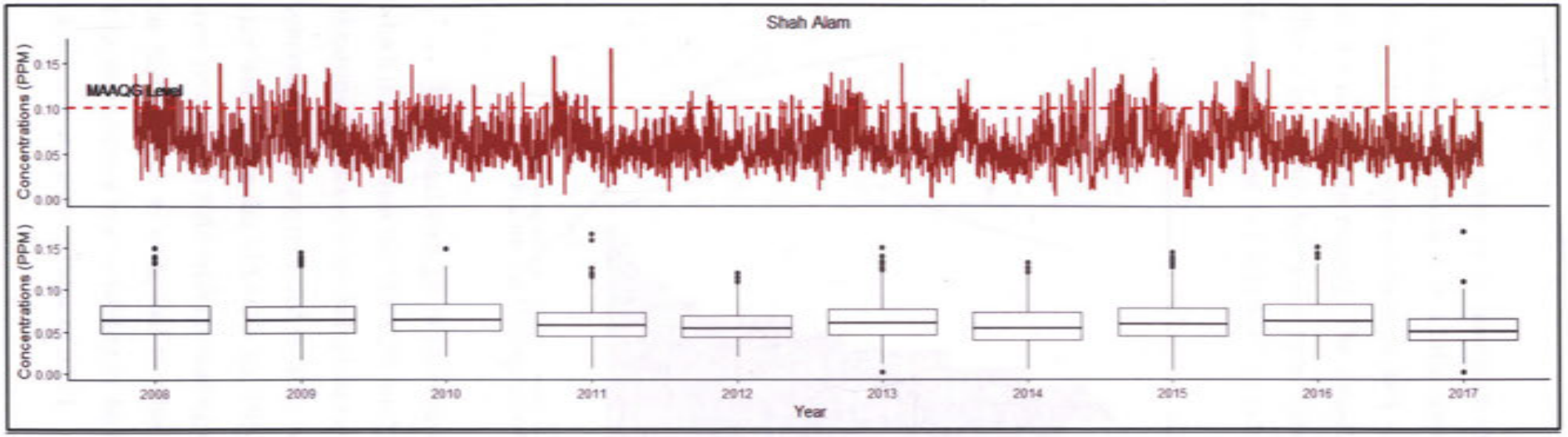


Figure 4.4 Histogram of O_3 concentrations for Shah Alam

The descriptive statistic is shown in Figure 4.5, the annual averages of concentrations in Shah Alam were recorded above 0.05 ppm, which is still below the MAAQG level. The concentrations were all positively skewed with the highest value in 2017 (0.919) and followed by 2011 (0.819). All the annual maximum concentrations were above the MAAQG for daily level. The highest maximum concentrations were recorded in 2017 with the reading of 0.168 ppm. Distribution for 2017 registered the highest peak as compared to other years with the kurtosis value of 3.030. The analysis indicates low variability in the monitoring records every year with the overall coefficient variation value of 0.383.



	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	Overall
Mean	0.066	0.067	0.067	0.061	0.057	0.063	0.058	0.063	0.065	0.053	0.062
Median	0.063	0.064	0.064	0.057	0.053	0.06	0.054	0.059	0.063	0.049	0.059
Min	0.004	0.015	0.019	0.005	0.019	0.001	0.004	0.003	0.015	0.001	0.001
Max	0.149	0.145	0.148	0.166	0.119	0.149	0.132	0.145	0.151	0.168	0.168
Std. Deviation	0.025	0.025	0.023	0.023	0.019	0.025	0.023	0.027	0.023	0.019	0.024
Coefficient of Variations	0.381	0.372	0.338	0.384	0.340	0.390	0.395	0.427	0.361	0.364	0.383
Skewness	0.527	0.673	0.412	0.819	0.671	0.550	0.634	0.531	0.690	0.919	0.667
Kurtosis	0.050	0.214	-0.259	1.432	0.101	0.252	0.276	0.253	0.437	3.030	0.531

Figure 4.5 Time-Series plot, Box-and-Whisker plot and Descriptive Statistics for Shah Alam

iii) Cheras

Next, as seen in the histogram of the O_3 concentration in Cheras in Figure 4.6, it is nicely distributed with a little skew to the right. The data distribution shows the existence of extreme values in the data, which is expected to fit well with the distribution of the monitoring records. The annual average daily maximum concentrations for Cheras are mostly below 0.1 ppm, although the extreme value occurs every year, as shown in the Box-and-Whisker Plot in Figure 4.7.

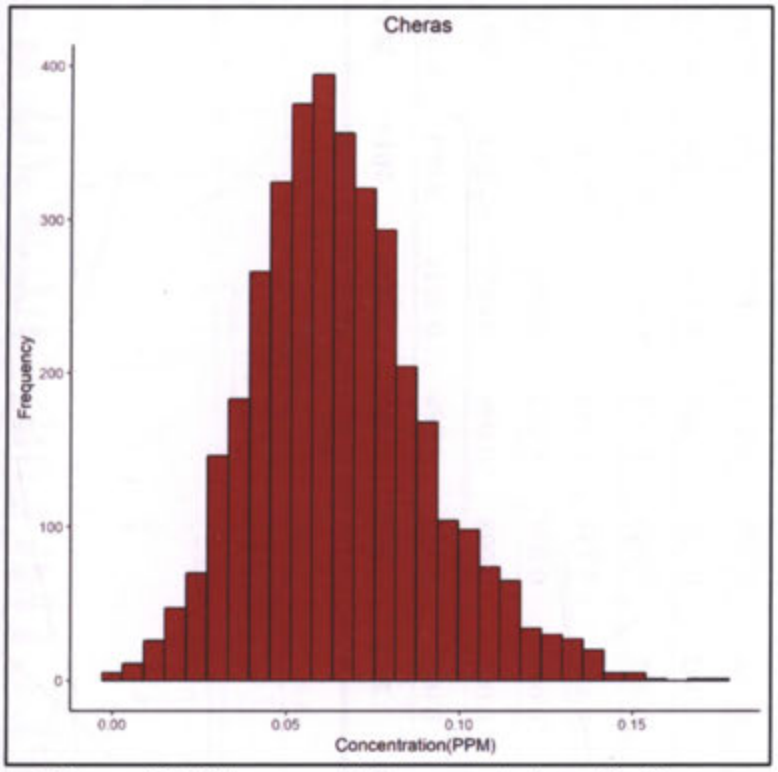
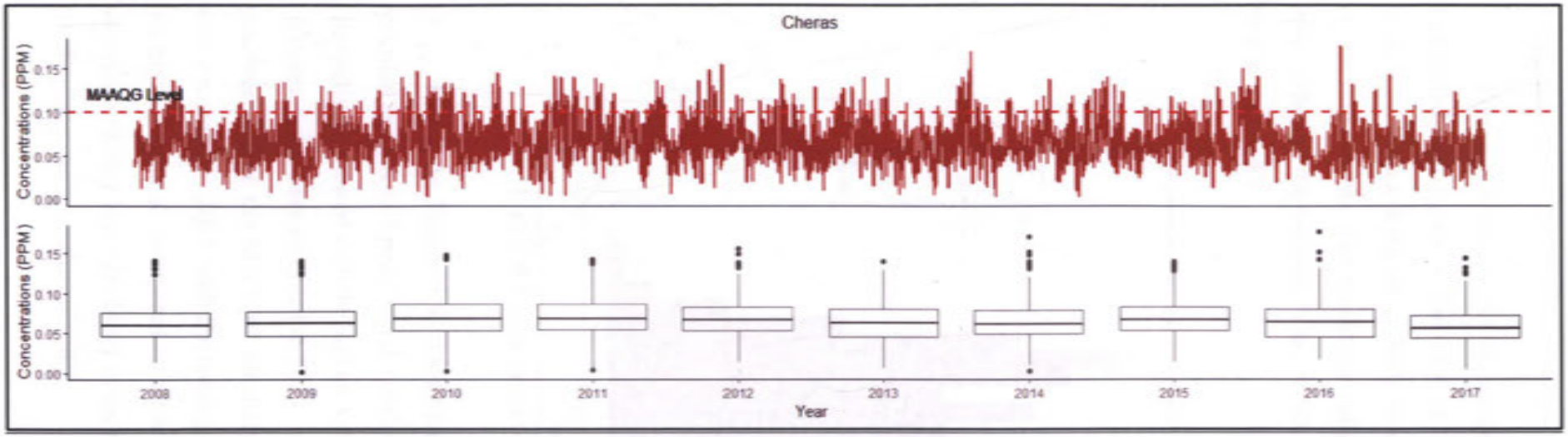


Figure 4.6 Histogram of O_3 concentrations for Cheras

The annual averages of concentrations in Cheras have recorded above 0.05 ppm, which is below than the MAAQG level shown in Figure 4.7. The concentrations were all positively skewed with the highest value in 2016 (0.732), meaning that more extreme concentrations were recorded in 2016. All the maximum concentrations recorded each year were above the MAAQG for daily level. The highest maximum concentrations were recorded in 2016 with the reading of 0.176 ppm. Distribution for 2014 registered the highest peak as compared to other years with the kurtosis value of 1.255. The analysis indicates low variability in the monitoring records every year with the overall coefficient of variation value of 0.381.



	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	Overall
Mean	0.062	0.063	0.071	0.071	0.068	0.063	0.064	0.069	0.065	0.058	0.066
Median	0.06	0.062	0.068	0.068	0.066	0.062	0.061	0.067	0.064	0.056	0.063
Min	0.013	0.001	0.003	0.004	0.013	0.005	0.003	0.013	0.016	0.004	0.001
Max	0.141	0.141	0.148	0.142	0.155	0.140	0.170	0.140	0.176	0.143	0.176
Std. Deviation	0.023	0.025	0.028	0.026	0.024	0.025	0.025	0.023	0.026	0.021	0.025
Coefficient of Variations	0.377	0.393	0.392	0.376	0.350	0.390	0.395	0.328	0.396	0.364	0.381
Skewness	0.664	0.468	0.295	0.286	0.517	0.302	0.717	0.558	0.732	0.605	0.525
Kurtosis	0.534	0.306	-0.115	0.018	0.538	-0.041	1.255	0.275	0.709	0.986	0.439

Figure 4.7 Time-Series plot, Box-and-Whisker plot and Descriptive Statistics for Cheras

iv) Klang

Figure 4.8 shows the O_3 concentration in Klang. The histogram is nicely distributed with a skew to the right; in which the data distribution shows the presence of extreme values in the data which are expected to fit well with the distribution of the monitoring records. The maximum daily concentration for Klang is mostly below 0.1 ppm, although the extreme value occurs every year, as shown in the Box-and-Whisker Plot in Figure 4.9.

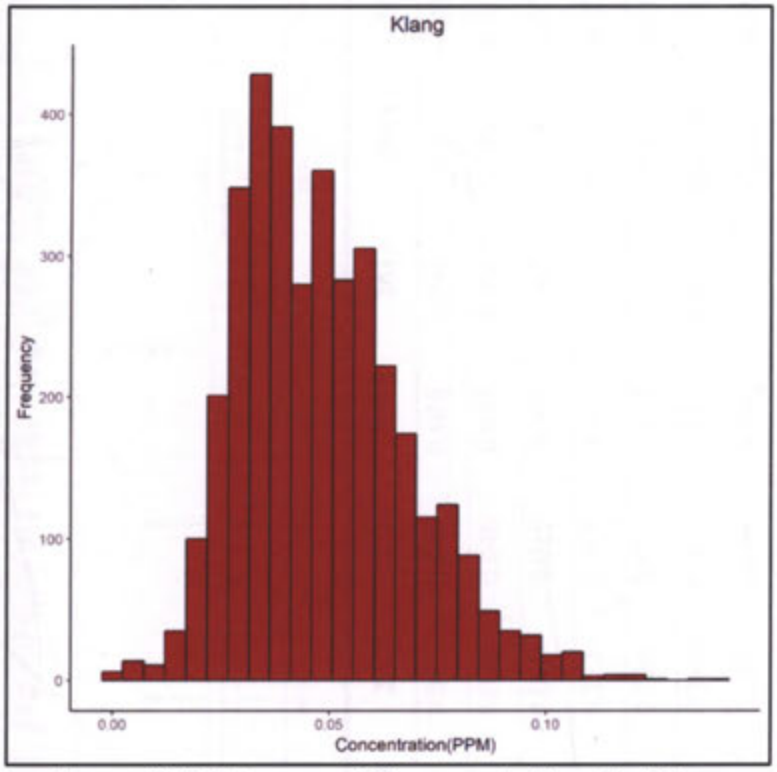
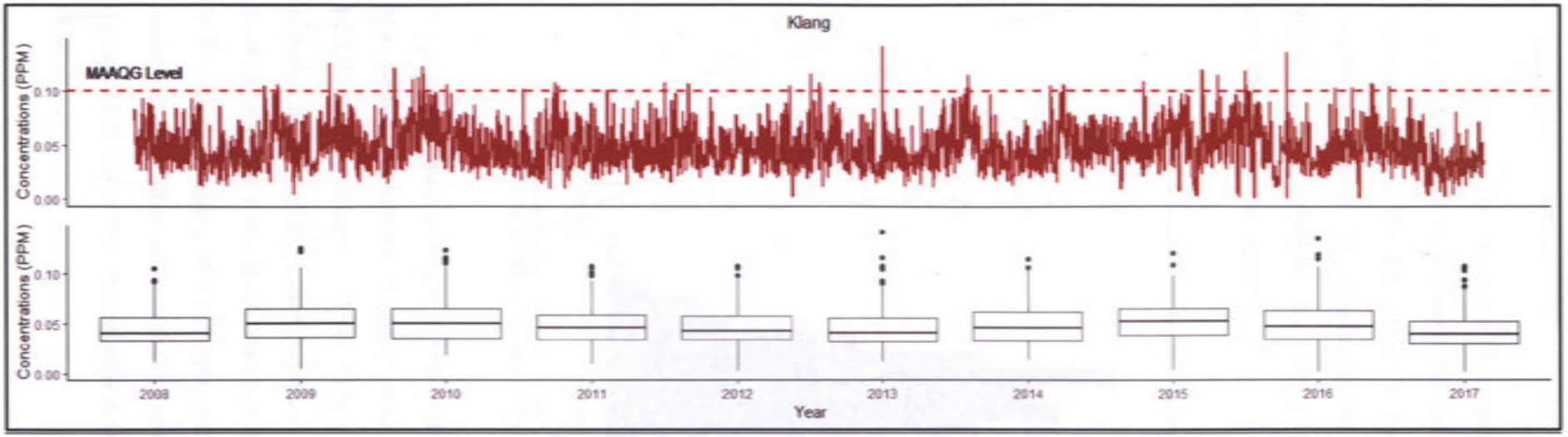


Figure 4.8 Histogram of O_3 concentrations for Klang

Based on Figure 4.9, the annual averages of concentrations in Klang were recorded around 0.05 ppm; which is below the MAAQG level. The concentrations were all positively skewed with the highest value in 2013 (1.143); meaning that more extreme concentrations were recorded in 2013. All the maximum concentrations recorded each year were above the MAAQG for daily level. The highest maximum concentrations were recorded in 2013 with the readings 0.141 ppm. Distribution for 2013 registered the highest peak as compared to other years with the kurtosis value of 2.321. The analysis indicates low variability in the monitoring records every year with the overall coefficient variation value of 0.402.



	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	Overall
Mean	0.045	0.052	0.052	0.048	0.046	0.045	0.050	0.052	0.050	0.042	0.048
Median	0.041	0.050	0.050	0.046	0.043	0.041	0.046	0.052	0.047	0.040	0.046
Min	0.012	0.005	0.018	0.010	0.002	0.012	0.013	0.003	0.001	0.001	0.001
Max	0.105	0.125	0.123	0.108	0.107	0.141	0.115	0.120	0.135	0.107	0.141
Std. Deviation	0.018	0.020	0.020	0.019	0.017	0.018	0.021	0.019	0.021	0.018	0.019
Coefficient of Variations	0.3947	0.3841	0.3869	0.3877	0.3597	0.4084	0.4213	0.3648	0.4240	0.4322	0.4023
Skewness	0.740	0.591	0.633	0.696	0.708	1.143	0.739	0.088	0.658	0.629	0.680
Kurtosis	0.049	0.316	0.096	0.253	0.415	2.321	-0.042	0.263	0.504	0.730	0.474

Figure 4.9 Time-Series plot, Box-and-Whisker plot and Descriptive Statistics for Klang

v) Tanjung Malim

Figure 4.10 shows the O_3 concentration in Tanjung Malim. The histogram is nicely distributed with a skew to the right. The data distribution shows the existence of extreme values in the data, which is expected to fit well with the distribution of the monitoring records. The maximum daily concentration for Tanjung Malim is mostly below 0.1 ppm, although the extreme value occurs every year, as shown in the Box-and-Whisker Plot in Figure 4.11.

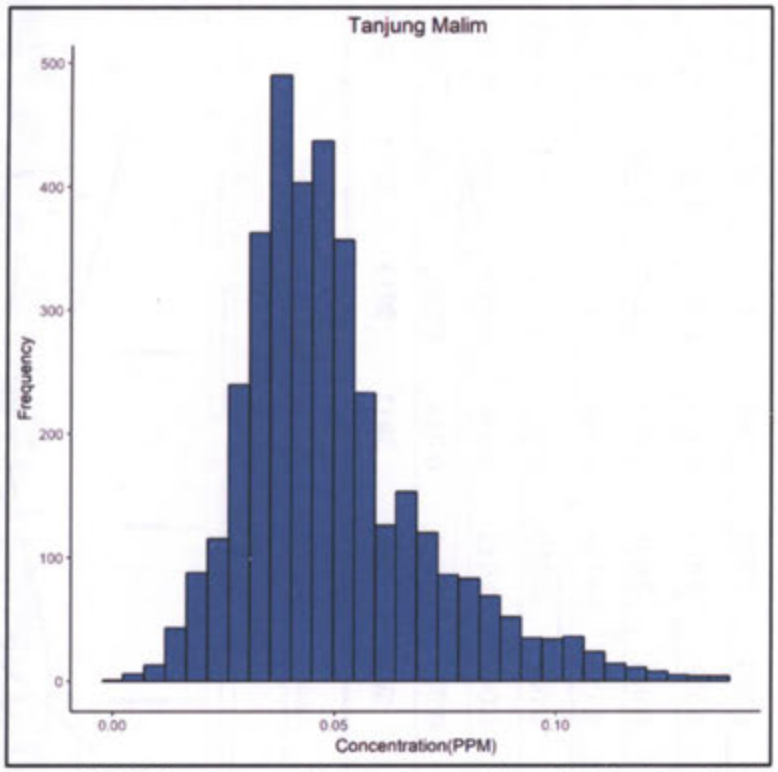
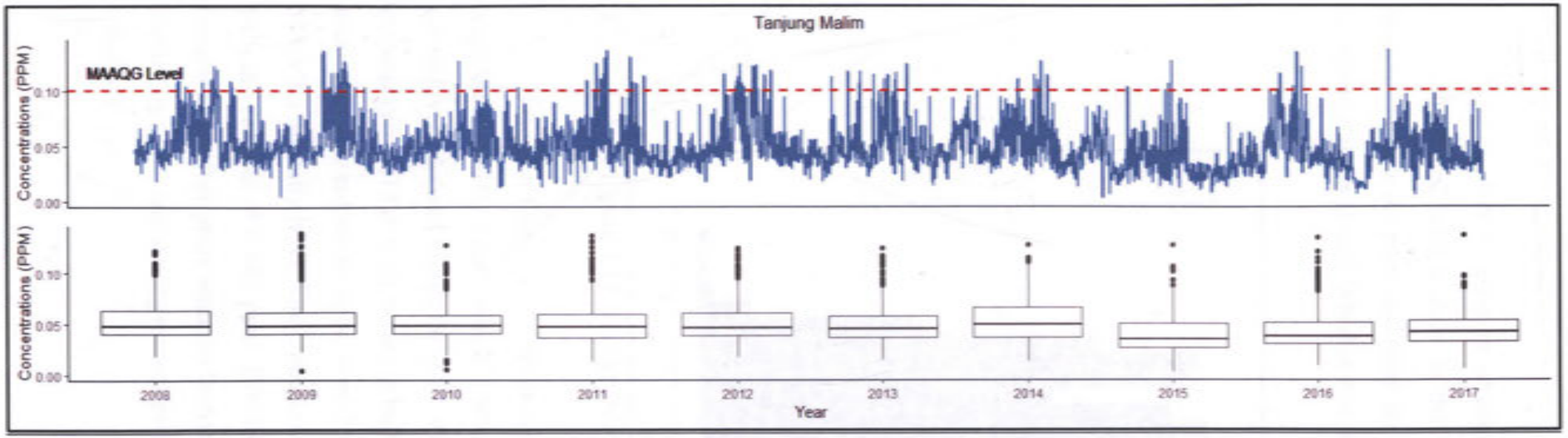


Figure 4.10 Histogram of O_3 concentrations for Tanjung Malim

The annual averages of concentrations in Tanjung Malim recorded around 0.05 ppm, which is below the MAAQG level. The concentrations were all positively skewed with the highest value in 2016 (1.440), meaning that more extreme concentrations were recorded in 2016. All the maximum concentrations are recorded each year were above the MAAQG for daily level. The highest maximum concentrations were recorded in 2009 with the reading of 0.139 ppm. Distribution for 2016 registered the highest peak as compared to other years with the kurtosis value of 2.300. The analysis indicates low variability in the monitoring records every year with the overall coefficient of variation value of 0.419.



	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	Overall
Mean	0.055	0.055	0.052	0.051	0.053	0.050	0.052	0.039	0.043	0.044	0.049
Median	0.048	0.048	0.048	0.047	0.046	0.045	0.049	0.035	0.038	0.041	0.045
Min	0.018	0.004	0.006	0.013	0.017	0.014	0.013	0.002	0.009	0.005	0.002
Max	0.122	0.139	0.127	0.136	0.124	0.124	0.127	0.127	0.134	0.137	0.139
Std. Deviation	0.020	0.022	0.017	0.021	0.022	0.020	0.020	0.019	0.020	0.019	0.021
Coefficient of Variations	0.371	0.404	0.320	0.416	0.414	0.401	0.381	0.477	0.471	0.429	0.419
Skewness	1.166	1.386	0.892	1.339	1.264	1.323	0.795	1.078	1.440	0.628	1.112
Kurtosis	0.727	1.864	1.680	2.051	0.853	1.597	0.433	1.489	2.300	1.292	1.511

Figure 4.11 Time-Series plot, Box-and-Whisker plot and Descriptive Statistics for Tanjung Malim

vi) Jerantut

The O_3 concentration in Jerantut is shown in Figure 4.12. The histogram is nicely distributed with a skew to the right. The maximum daily concentration for Jerantut is below 0.1 ppm, and there are outlier data but not extreme in 2012 and 2013, as shown in the Box-and-Whisker Plot in Figure 4.13.

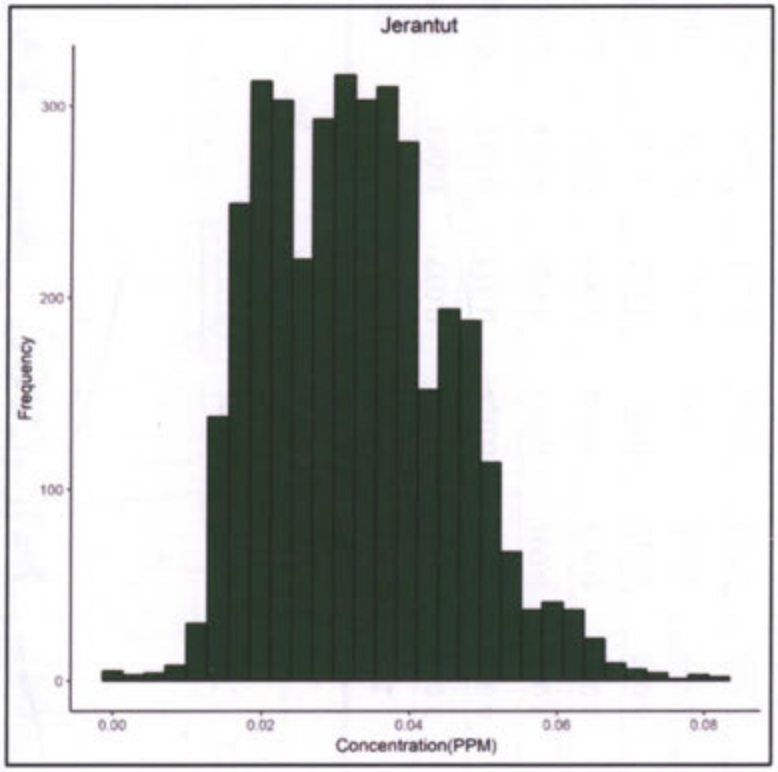
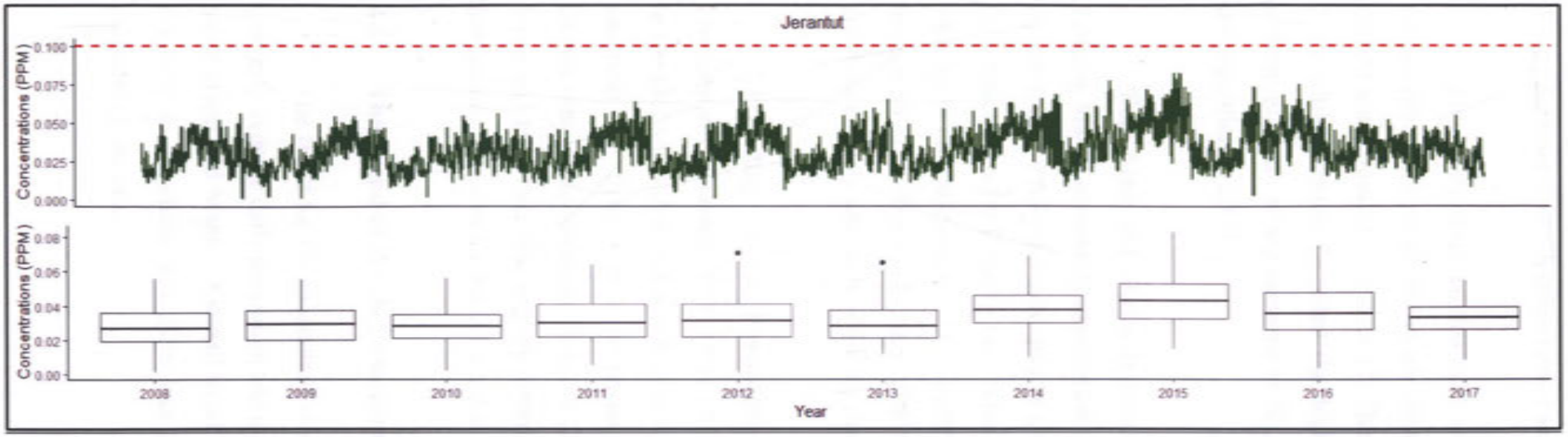


Figure 4.12 Histogram of O_3 concentrations for Jerantut

Based on Figure 4.13, the annual averages of concentrations in Jerantut were recorded around 0.03 ppm, which is below the MAAQG level. The concentrations were all positively skewed except for 2017, where it negatively skewed. The highest value of extreme is in 2013 (0.606), meaning high concentrations were recorded in 2013. All the maximum concentrations in between 2008 – 2017 recorded each year were below the MAAQG for daily level. The highest maximum concentrations were recorded in 2015 with the reading of 0.083 ppm. Distribution for 2013 registered the highest peak as compared to other years with the kurtosis value of - 0.392. The analysis indicates low variability in the monitoring records every year with the overall coefficient of variation value of 0.376.



	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	Overall
Mean	0.028	0.029	0.029	0.032	0.032	0.030	0.037	0.043	0.037	0.032	0.033
Median	0.027	0.029	0.028	0.030	0.031	0.028	0.037	0.043	0.036	0.033	0.032
Min	0.001	0.001	0.002	0.005	0.001	0.011	0.009	0.014	0.003	0.008	0.001
Max	0.056	0.055	0.056	0.064	0.071	0.065	0.069	0.083	0.075	0.055	0.083
Std. Deviation	0.010	0.011	0.009	0.012	0.012	0.011	0.012	0.015	0.013	0.009	0.012
Coefficient of Variations	0.374	0.365	0.323	0.377	0.382	0.374	0.312	0.338	0.351	0.280	0.376
Skewness	0.217	0.202	0.227	0.265	0.394	0.606	0.015	0.245	0.343	-0.238	0.478
Kurtosis	-0.596	-0.735	-0.437	-0.827	-0.474	-0.392	-0.444	-0.556	-0.615	-0.486	-0.028

Figure 4.13 Time-Series plot, Box-and-Whisker plot and Descriptive Statistics for Jerantut

4.3.2 Summary of Characteristics and Pattern of O_3 Concentrations

This study observed that there were no specific years or months of extreme O_3 concentrations occurred for every location. For Petaling Jaya and Shah Alam, the highest concentration was in 2017. The highest concentration for Cheras occurred in 2016 while Jerantut recorded the highest concentration in 2015, and the highest concentration for Klang occurred in 2013. Lastly, Tanjung Malim recorded the highest concentration in 2009.

Shah Alam and Cheras experienced the highest O_3 concentrations. However, Tanjung Malim commonly experiences a high concentration of O_3 between the second quarter (April – Jun) and third quarter (July – September) of the year. The annual mean O_3 concentration for Petaling Jaya, Klang, and Tanjung Malim are almost similar. This was due to the seasonality of O_3 concentration, which occurs in that particular location but not for Shah Alam and Cheras. The monitoring stations for Shah Alam and Cheras are close enough, and both of the monitoring stations are located in an urban area.

According to the Environmental Quality Report 2017 (Department of Environment Malaysia, 2018), there was no haze episode in Malaysia in 2017 compared to the past years that was mainly due to regional and transboundary haze. This study observed that there was a decreasing trend of O_3 concentrations in 2017. The Department of Environment Malaysia (2018) also reported that there were numbers of forest and bush fires that slightly deteriorated local air quality in the country but were not prolonged due to the humid weather all year round.

4.4 The Trend of the O_3 Concentrations

The trend for O_3 concentration is illustrated in two ways, which are by using quarterly average daily maximum and quarterly maximum data. The data are illustrated using graph and Mann – Kendall trend test was used to support if there were enough evidence to conclude that there was an increasing or decreasing trend for every monitoring stations.

4.4.1 The Mann – Kendal Trend Test

i) Petaling Jaya

Figure 4.14 illustrates the trend of the quarterly average of daily maximum in Petaling Jaya from 2008 – 2017. The graph shows that there is a slightly increasing trend, but the analysis of Mann-Kendall's show (MK) trend indicated that the $\tau = 0.0436$ (p-value = 0.7006 > 0.05). There was no significant evidence that the quarterly average of the daily maximum of O_3 concentrations increased during the period.

Figure 4.14 Trend of quarterly average daily maximum for Petaling Jaya

Another trend plot using quarterly maximum for Petaling Jaya from 2008 – 2017 is as illustrated in Figure 4.15. The graph also indicates that there is a slightly decreasing trend over the year. However, the trend is also not significant statistically following the MK's trend test ($\tau = -0.0349$, p-value = 0.7617 > 0.05).

Figure 4.15 Trend of quarterly maximum for Petaling Jaya

ii) Shah Alam

The trend for the quarterly average of daily maximum in Shah Alam from 2008 – 2017 is illustrated in Figure 4.16. The figure clearly shows there is a decreasing trend along with the year and the analysis of Mann-Kendall's shows the result that the τ = -0.269 (p-value = 0.01489 < 0.05). Thus, it concludes that there was significant evidence that the quarterly average of the daily maximum of O_3 concentrations decreased during the period.

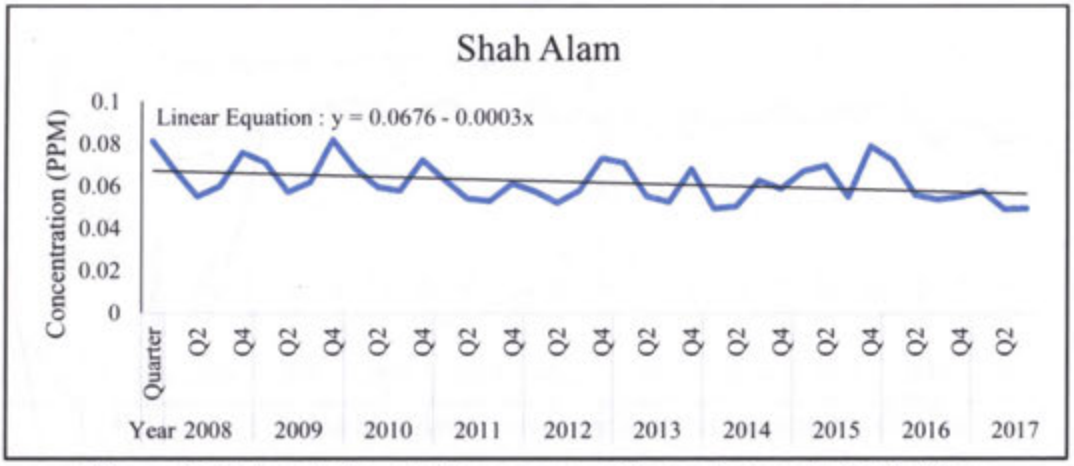


Figure 4.16 Trend of quarterly average daily maximum for Shah Alam

By using the quarterly maximum for Shah Alam from 2008 – 2017, as illustrated in Figure 4.17, the graph also shows a decreasing trend for this location. However, the trend is not significant statistically following the MK's trend test (τ = -0.164, p-value = 0.1416 > 0.05).

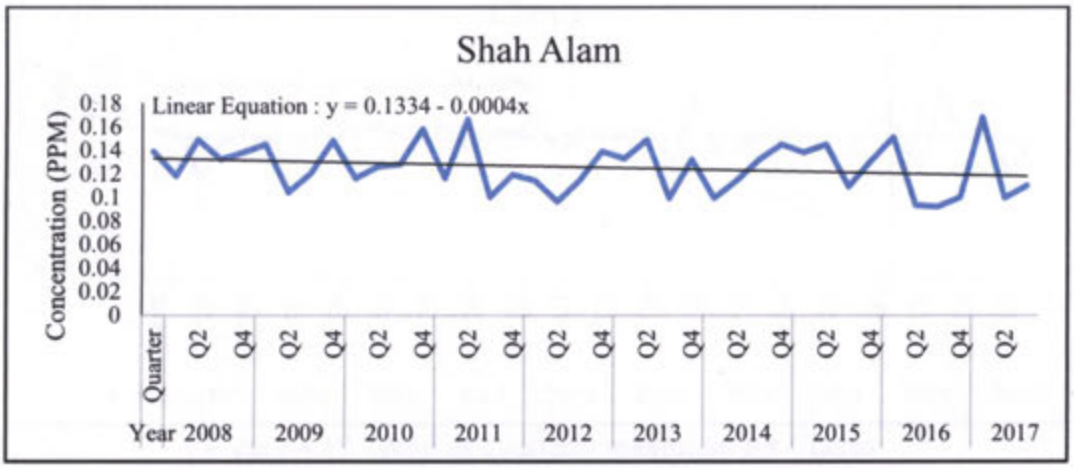


Figure 4.17 Trend of quarterly maximum for Shah Alam

iii) Cheras

Figure 4.18 illustrates the trend of the quarterly average of daily maximum in Cheras from 2008 – 2017. The graph shows that there is a decreasing trend, but the analysis of Mann-Kendall's show (MK) trend indicated that the $\tau = -0.128$ (p-value = 0.2487> 0.05). There was no significant evidence that the quarterly average of the daily maximum of O_3 concentrations decreased during the period.

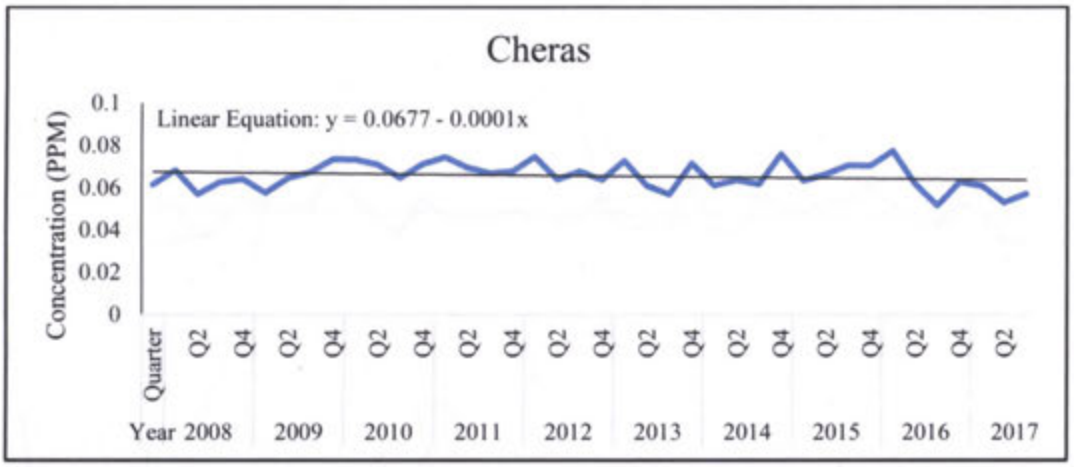


Figure 4.18 Trend of quarterly average daily maximum for Cheras

Another trend test using the quarterly maximum for Cheras from 2008 – 2017 is illustrated in Figure 4.19. The graph shows there is a slightly decreasing trend and the MK's trend test confirmed it ($\tau = -0.0272$, p-value = 0.8156> 0.05) that there is no significant decreasing trend for this location.

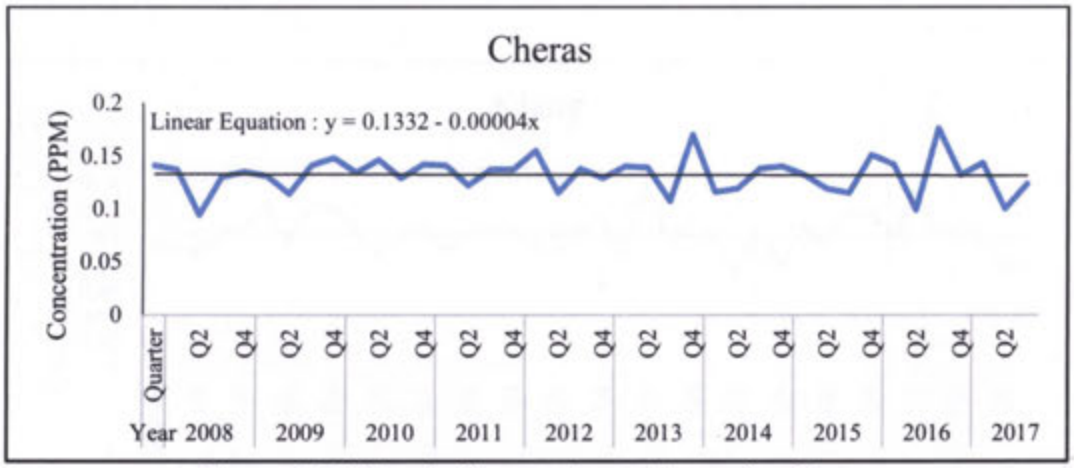


Figure 4.19 Trend of quarterly maximum for Cheras

iv) Klang

The trend graph using average quarterly is illustrated in Figure 4.20. The graph shows that there is a decreasing trend. However, the analysis of Mann-Kendall's show (MK) trend indicated that the $\tau = -0.108$ (p-value = 0.3335> 0.05). Thus, it is concluded that there was no significant evidence that the quarterly average of the daily maximum of O_3 concentrations decreased during the period.

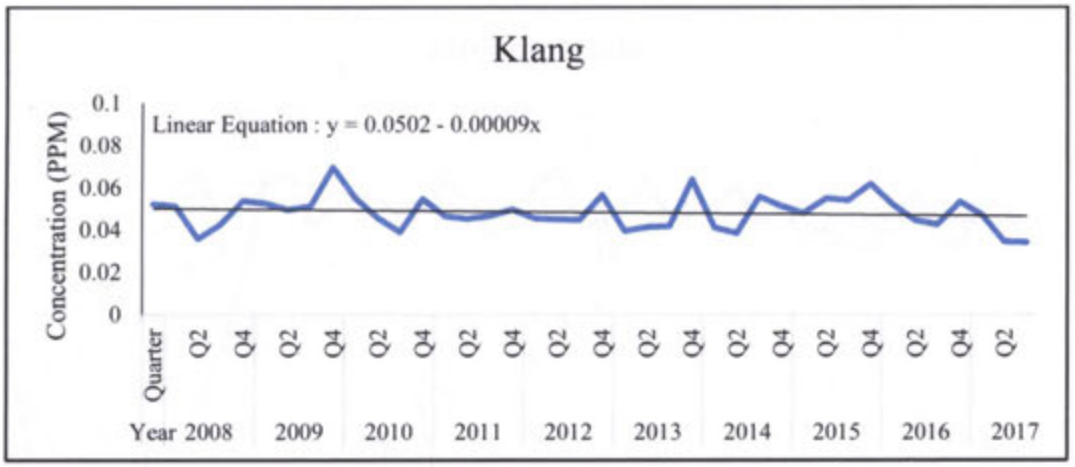


Figure 4.20 Trend of quarterly average daily maximum for Klang

By using the quarterly maximum for Klang from 2008 – 2017, as illustrated in Figure 4.21, the graph also shows a slightly decreasing trend. However, the MK's trend test concluded that the trend is not significant ($\tau = -0.0065$, p-value = 0.9628 > 0.05). This concluded that there is no decreasing trend in the Klang station for O_3 concentrations.

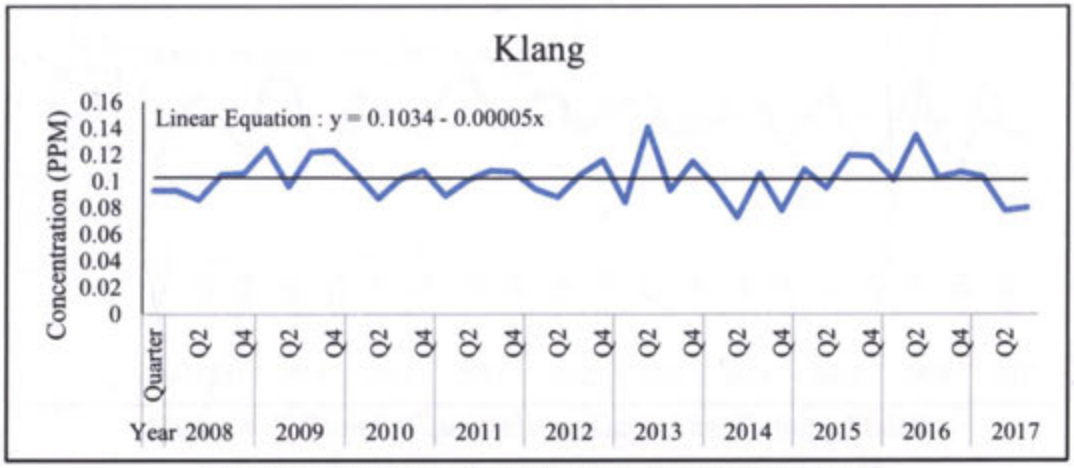


Figure 4.21 Trend of quarterly maximum for Klang

v) Tanjung Malim

Figure 4.22 illustrates the trend of the quarterly average of daily maximum in Tanjung Malim. The figure shows a decreasing trend from 2008 – 2017 with support from the analysis of Mann-Kendall that shows MK's test statistic of $\tau = -0.273$(p-value = 0.0135< 0.05). Therefore, it concludes that there is significant evidence that there is a decreasing trend for a quarterly average daily maximum of O_3 during the period.

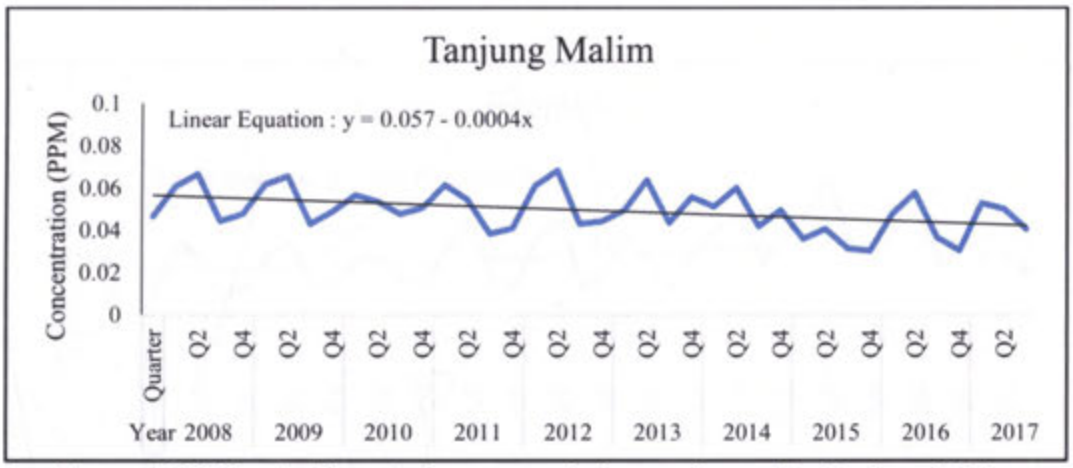


Figure 4.22 Trend of quarterly average daily maximum for Tanjung Malim

Meanwhile, Figure 4.23 illustrates the quarterly maximum for Tanjung Malim from 2008 – 2017. It shows that the graph has a slightly decreasing trend. But, there is not enough evidence from MK's test that concluded there is a decreasing trend for this location ($\tau = -0.0541$, p-value = 0.6327 > 0.05).

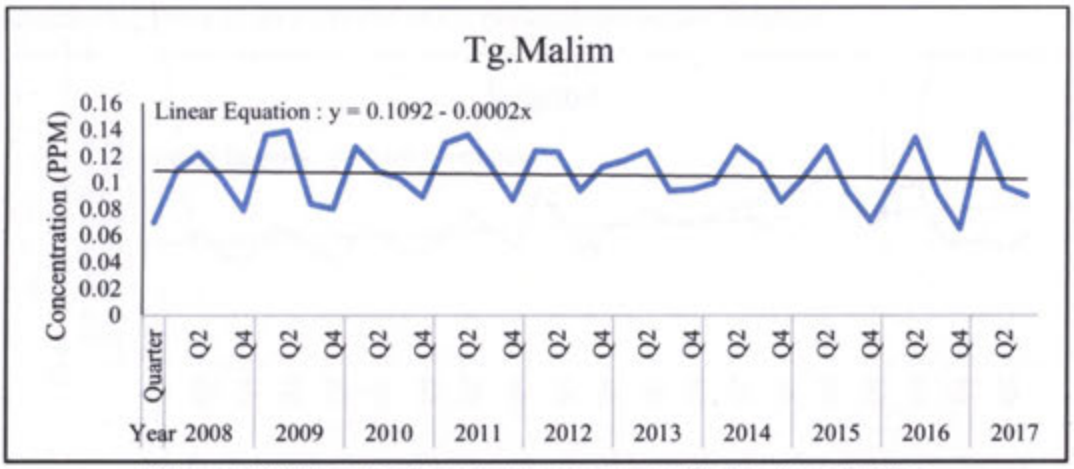


Figure 4.23 Trend of quarterly maximum for Tanjung Malim

vi) Jerantut

As for the background station, Figure 4.24 is illustrated using the quarterly average of daily maximum in Jerantut from 2008 – 2017. The graph shows that there is an increasing trend, with the support from the analysis of Mann-Kendall's (MK) trend that indicated the $\tau = 0.267$ (p-value = 0.01588< 0.05). There was significant evidence that the quarterly average of the daily maximum of O_3 concentrations increased during the period.

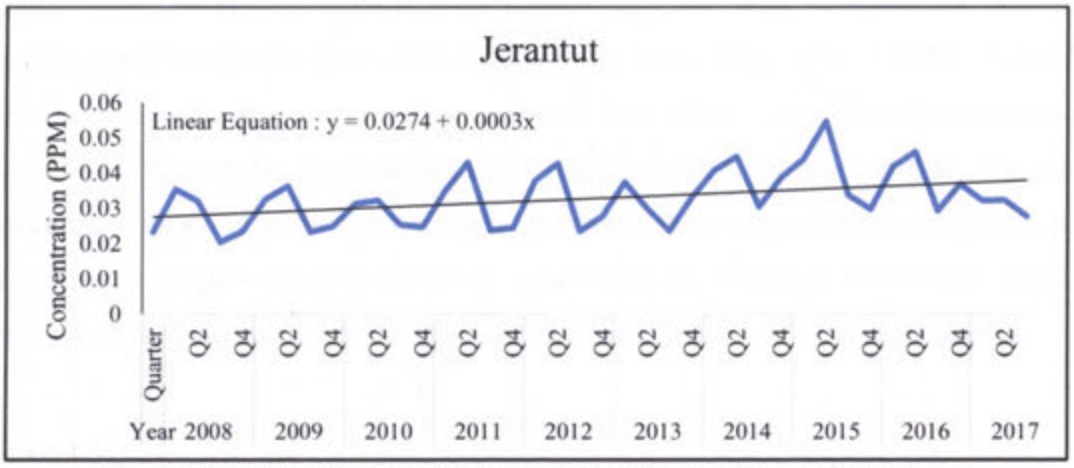


Figure 4.24 Trend of quarterly average daily maximum for Jerantut

By using the quarterly maximum for Jerantut from 2008 – 2017, as illustrated in Figure 4.25, the graph also shows an increasing trend. The trend is also significant statistically following the MK's trend test ($\tau = 0.349$, p-value = 0.00176< 0.05). The results obtained in this study have to be given extra precaution since both of the figures indicate that there is an increase of O_3 concentrations for Jerantut.

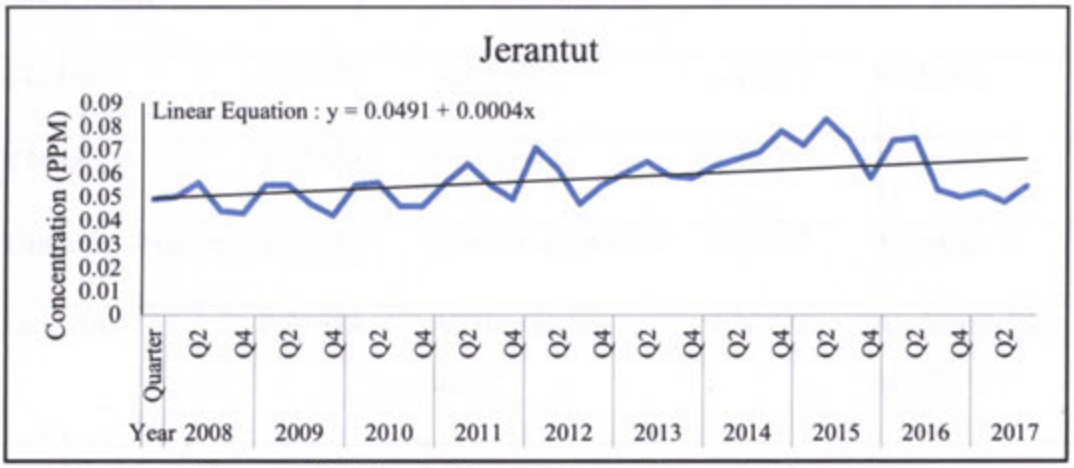


Figure 4.25 Trend of quarterly maximum for Jerantut

4.4.2 Summary of the Trend for All Monitoring Stations

From the results discussed in Section 4.4, overall, the locations showed that there was no trend for O_3 concentrations except for Shah Alam, Tanjung Malim, and Jerantut. The summary table for the Mann Kendal trend test for quarterly average daily maximum and the quarterly maximum is shown in Table 4.3.

The discussion shows that the main concern is not for Shah Alam and Tanjung Malim since the results show a good performance for both locations for quarterly average daily maximum data with a decreasing trend from 2008 – 2017. It can be concluded that the O_3 concentration declined from 2008 – 2017 for these locations. However, there was no trend for quarterly maximum data for this location. The main concern in this study is that the trend test shows that there was significant evidence that Jerantut had an increasing trend for O_3 concentration. Although the overall data for Jerantut did not exceed the MAAQG, this location needs to take extra precaution.

Table 4.3
Summary of Mann Kendal trend test

Monitoring Stations	**Quarterly Average Daily Maximum Data**		**Quarterly Maximum Data**	
	p-value	**Result**	**p-value**	**Result**
Petaling Jaya	0.70062	No trend	0.76172	No trend
Shah Alam	0.01489	A decreasing trend	0.14164	No trend
Cheras	0.24872	No trend	0.81557	No trend
Klang	0.33353	No trend	0.96280	No trend
Tanjung. Malim	0.01350	A decreasing trend	0.63267	No trend
Jerantut	0.01588	An increasing trend	0.00176	An increasing trend

4.5 The Bayesian Approach

The Bayesian approach that has been discussed in Section 3.7, and this Section discusses the outcomes of the Bayesian approach in achieving objective 2 and 3. The result of the parameter estimation for the GEV and Weibull is discussed in Section 4.5.1, while the suitable prior and the best distribution will be discussed in Section 4.5.2.

4.5.1 Parameter Estimation

From Figure 3.4 in Section 3.7.1, the process of getting the parameter of the posterior was using the MCMC. Thus, Figure 4.26 shows the iterative history of the MCMC approach for the GEV likelihood and uniform distribution as prior for all the parameters in Petaling Jaya. The iteration for Petaling Jaya was stabilised at 5000 of iterations. The mean for each parameter of the 5000 iterations was determined as the parameter to fit into the distribution shown in Section 4.6.

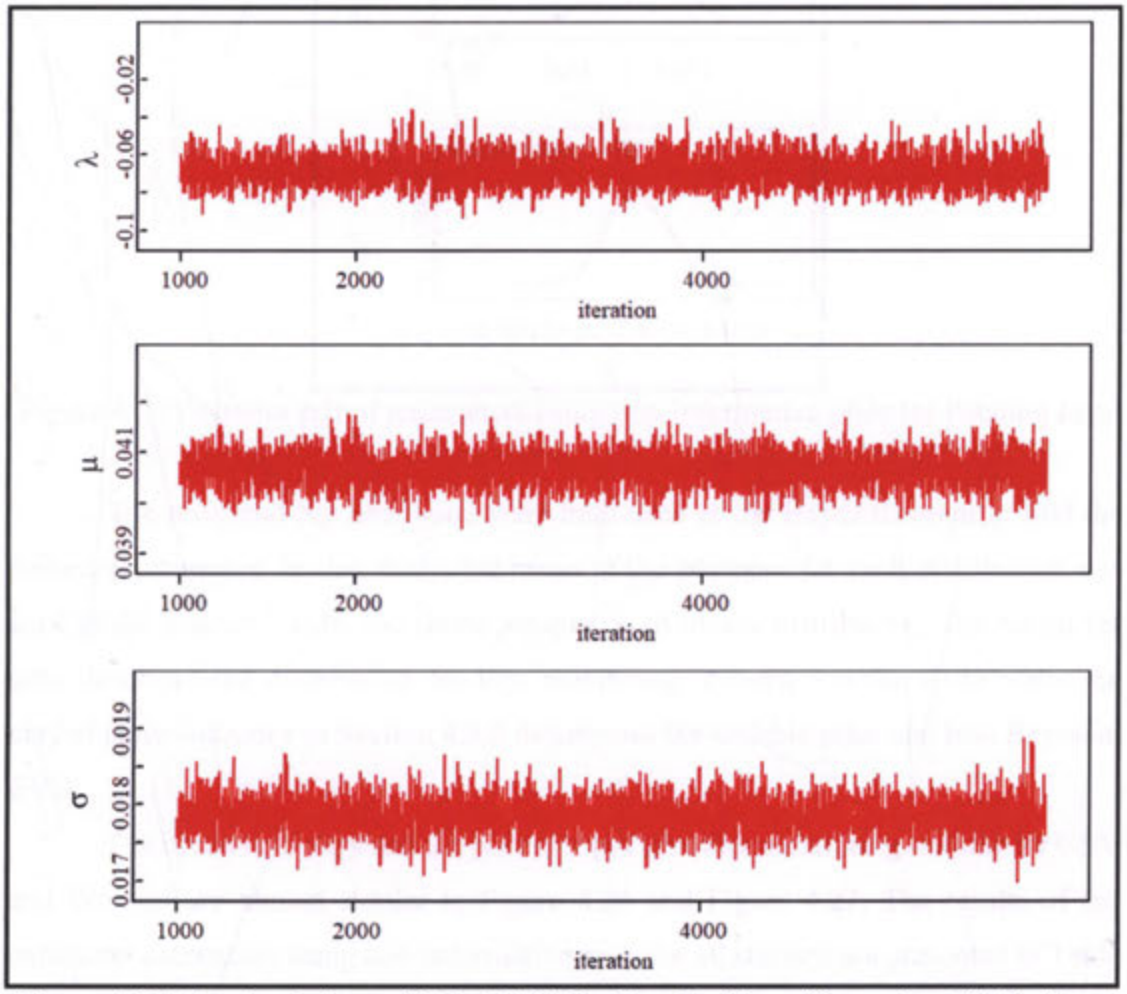


Figure 4.26 The simulation with the GEV likelihood with uniform prior distributions for Petaling Jaya

The estimation using the Bayesian approach was appropriate because the simulation for all location, μ, scale, σ and shape, λ parameters of the GEV distribution were stabilised around the mean shown in Figure 4.26. The posterior densities of each parameter, as shown in Figure 4.27, appear to be in conformity with the given data. Thus, it can be concluded that the estimation using the Bayesian approach was adequate.

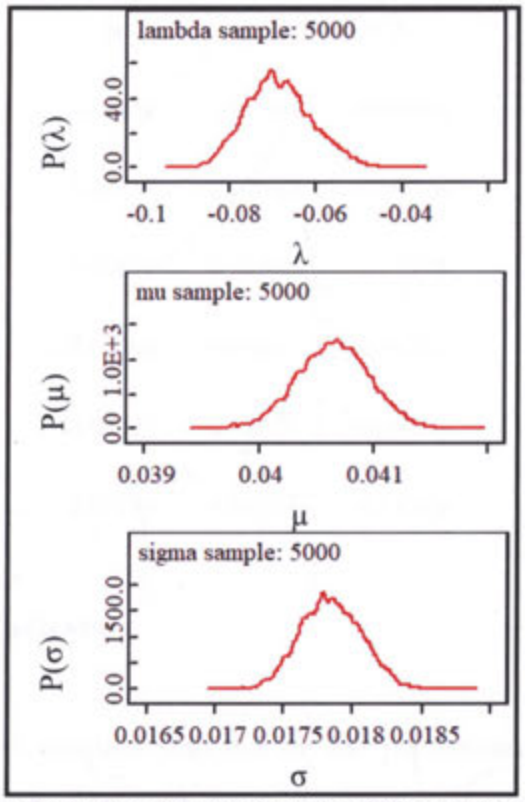


Figure 4.27 Posterior pdf of parameters using non-informative prior for Petaling Jaya

The prior and the likelihood were both used in the Bayes theorem to find the posterior parameter. In this study, the mean of the posterior for each distribution was used as the location, scale, and shape parameters to fit into distribution. The parameter only describes the distribution for that monitoring stations and the EVD while the performance indicator in Section 4.5.2 determines the suitable prior and best Bayesian EVD.

The iterative history and the posterior pdf for other monitoring stations for GEV and Weibull are almost similar to Figure 4.26 and Figure 4.27. The results of the parameter estimation using non-informative prior for all stations are presented in Table 4.4. The scale parameter for GEV is higher than the Weibull distribution. However, the shape parameter shows that the Weibull distribution is higher than the GEV distribution.

Table 4.4
Parameter estimation for all locations and extreme value distribution

Monitoring Stations	Extreme Value Distribution				
	Generalized Extreme Value			Weibull	
	Uniform Prior			Inverse Gamma Prior	
	μ	σ	λ	σ	λ
Petaling Jaya	0.04066	0.01785	-0.06856	0.01032	1.595
Shah Alam	0.05181	0.02063	-0.09776	0.01007	1.773
Cheras	0.05534	0.02265	-0.13770	0.01054	1.754
Klang	0.03995	0.01667	-0.08393	0.00945	1.609
Tanjung Malim	0.04022	0.01609	-0.00403	0.00990	1.607
Jerantut	0.02769	0.01099	-0.12300	0.00693	1.515

4.5.2 Performance Indicator

Table 4.5 shows detailed statistics on the performance indicators for every distribution and location. The best error measures and accuracy measures were highlighted based on the smallest error and the accuracy measures which were closest to 1. GEV shows the smallest error measurement and high accuracy measurement compared to Weibull distribution. The Bayesian approach appears to perform consistently regardless of a different location, and all locations showed that GEV is the best distribution in this study. Thus, it is concluded that uniform prior is suitable to be used in GEV distribution.

Table 4.5
Performance indicator for all locations and extreme value distribution

Monitoring Stations	EVD	Prior	Accuracy Measure			Error Measure		
			R^2	PA	IA	RMSE	MAE	NAE
Petaling Jaya	**GEV**	**Uniform**	**0.9940**	**0.9973**	**0.9985**	**0.0016**	**0.0007**	**0.0131**
	Weibull	Inv.Gamma	0.9864	0.9935	0.4503	0.0431	0.0405	0.8138
Shah Alam	**GEV**	**Uniform**	**0.9974**	**0.9990**	**0.9995**	**0.0011**	**0.0008**	**0.0125**
	Weibull	Inv.Gamma	0.9933	0.9969	0.4097	0.0560	0.0529	0.8551
Cheras	**GEV**	**Uniform**	**0.9977**	**0.9991**	**0.9995**	**0.0011**	**0.0008**	**0.0128**
	Weibull	Inv.Gamma	0.9872	0.9939	0.4057	0.0595	0.0562	0.8569
Klang	**GEV**	**Uniform**	**0.9970**	**0.9988**	**0.9994**	**0.0010**	**0.0007**	**0.0143**
	Weibull	Inv.Gamma	0.9911	0.9958	0.4397	0.0422	0.0398	0.8246
Tanjung Malim	**GEV**	**Uniform**	**0.9874**	**0.9939**	**0.9970**	**0.0023**	**0.0014**	**0.0274**
	Weibull	Inv.Gamma	0.9863	0.9934	0.4332	0.0433	0.0405	0.8204
Jerantut	**GEV**	**Uniform**	**0.9948**	**0.9977**	**0.9988**	**0.0008**	**0.0007**	**0.0208**
	Weibull	Inv.Gamma	0.9761	0.9882	0.4435	0.0279	0.0266	0.8097

4.6 The Probability Density Function and the Cumulative Distribution Function of the Monitoring Stations

From the discussion at Section 4.5.1 and 4.5.2, GEV was the best distribution to represent the distribution for all the monitoring stations. The parameter of GEV for each monitoring stations, which is the location, scale, and shape parameters were fit into the distribution using the same number of observation. Therefore the pdf and cdf were used to compare the observation and the theoretical of this study. If the theoretical distribution were closed enough with the observation, it is shown that the theoretical was a good distribution to explain the monitoring stations.

Figure 4.28 - Figure 4.33 shows the pdf and cdf for all the monitoring stations. The best distribution to represent the O_3 concentrations was the GEV distribution. From the graph, it can be seen that all the theoretical distributions fit the observations very well with the trace almost identical to that of the observations. The pdf had a heavy long tail to the right, the indication of extreme concentrations existed in all locations, except Jerantut. The probability of concentrations exceeding 0.1 ppm can be obtained from the cdf plot for all the monitoring stations, and thus, the estimation of the exceedances can be calculated using the probability of exceedances.

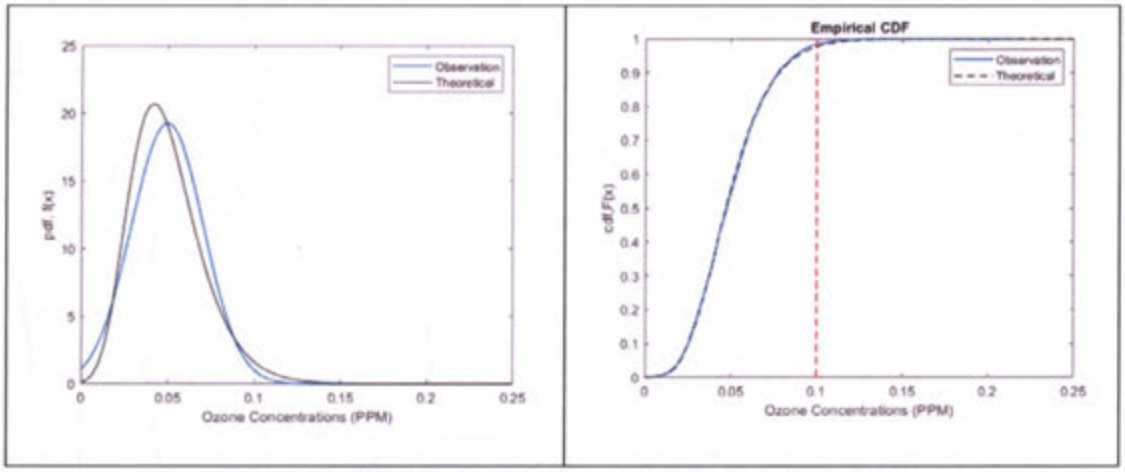


Figure 4.28 Pdf and cdf using GEV for Petaling Jaya

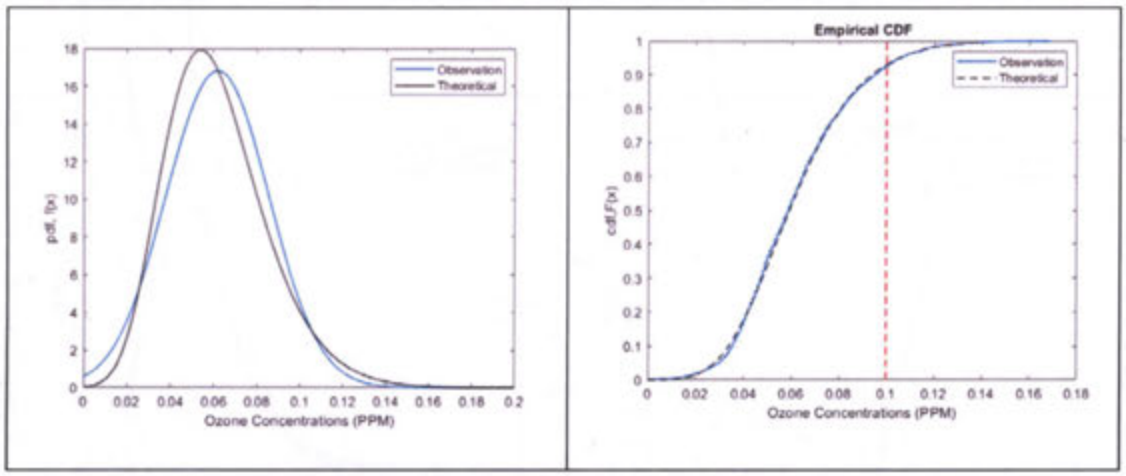


Figure 4.29 Pdf and cdf using GEV for Shah Alam

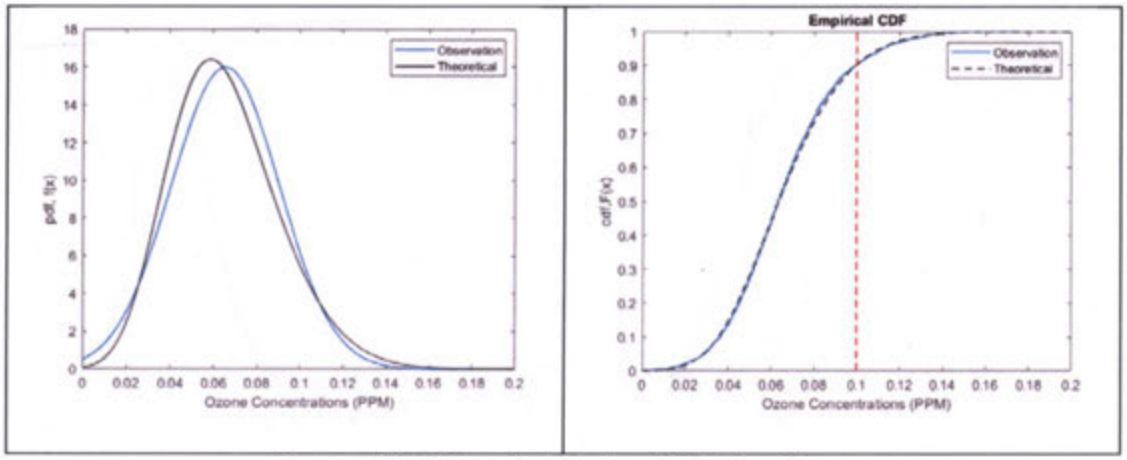


Figure 4.30 Pdf and cdf using GEV for Cheras

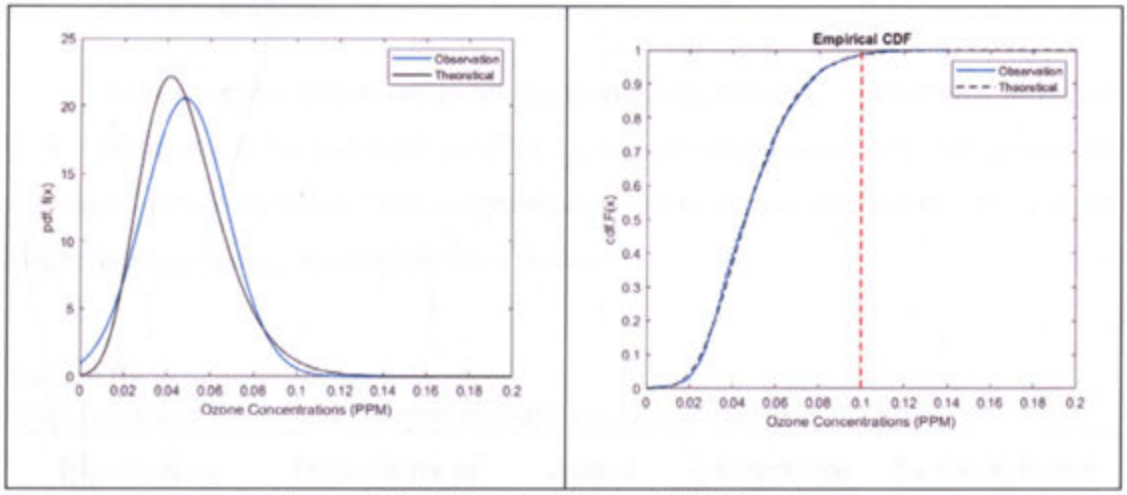


Figure 4.31 Pdf and cdf using GEV for Klang

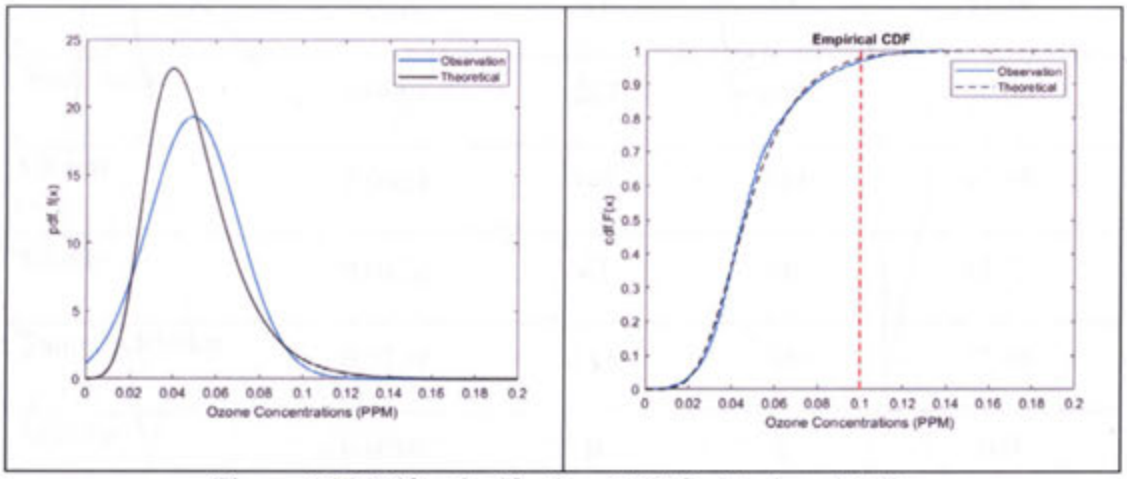


Figure 4.32 Pdf and cdf using GEV for Tanjung Malim

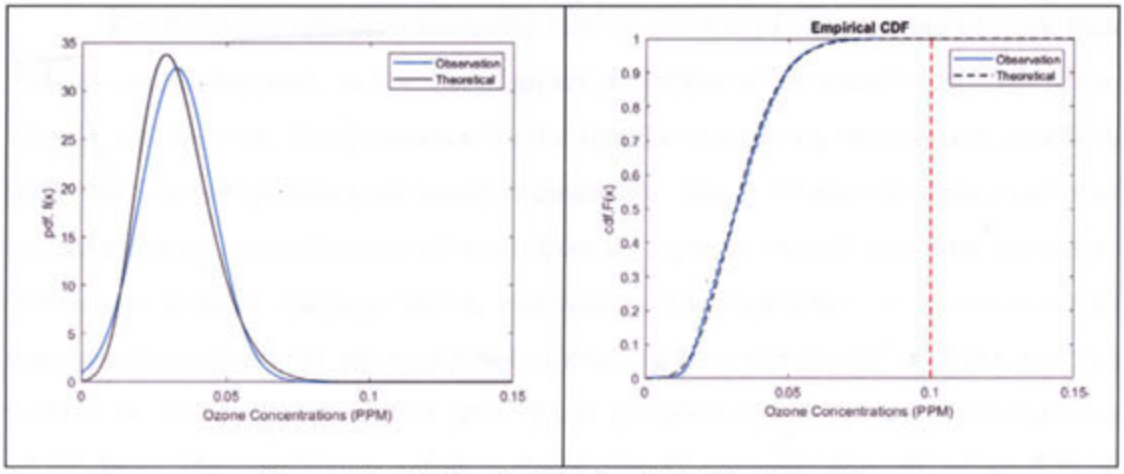


Figure 4.33 Pdf and cdf using GEV for Jerantut

4.7 The Exceedances

Following the plot of cdf in all the monitoring stations, as illustrated in Figure 4.28 – Figure 4.33, the probabilities of the O_3 concentrations exceeding 0.01 ppm were calculated for all locations. The comparison of the estimated number of days and the actual number of days are depicted in Table 4.6.

Table 4.6
Probability and Estimated Number of Days vs. Actual Number of Days

Monitoring Stations	Probability of Exceedances	Actual	Estimated	% Compliance
Petaling Jaya	0.0227	57	83	54.39
Shah Alam	0.0681	263	249	94.68
Cheras	0.0954	341	349	97.65
Klang	0.0136	43	50	83.72
Tanjung Malim	0.0234	111	86	77.48
Jerantut	0.0000	0	0	100

The Bayesian approach estimated that the number of exceedances of more than 94% in compliance with the number of an actual number of exceedances in Shah Alam, Cheras, and Jerantut. The estimation for the Jerantut monitoring stations was excellent, with 100% in compliance with actual exceedances. The EVD estimated the number of exceedances of concentrations of more than 0.1 ppm in compliance with the actual number of days in Tanjung Malim and Klang. The probability of exceedances of concentrations above 0.1 ppm obtained in Petaling Jaya was 0.0227, and that is equal to 83 days. The predicted number was 54% in compliance with the actual exceedances of 57 days. The compliance result is due to the O_3 concentration of actual data in Petaling Jaya. It shows a result close enough to 0.1 ppm but not exceeding 0.1 ppm. Thus, it shows that the estimator for Petaling Jaya was not good enough in explaining the prediction and return for that location. Since Jerantut does not have data that exceed 0.1 ppm, therefore no exceedance and return period need to be considered.

4.8 The Return Period

By using the probability of the exceedances for that 4 locations that can be used to predict the high exceedances of O_3 concentrations, the return period and number of exceedances was calculated. The result for the return period and the number of days that exceed 0.1 ppm for that 4 location were shown in Table 4.7.

Table 4.7
The exceedances probability and return period

Monitoring Stations	Probability of Exceedances	Return Period	Number of Exceedances (Days)			
			1 year	2 year	5 year	10 year
Shah Alam	0.0681	15	24	49	122	243
Cheras	0.0954	10	37	73	183	365
Klang	0.0136	73	5	10	25	50
Tanjung Malim	0.0234	43	8	17	42	85

In summary, based on the result as listed in Table 4.7, it can be estimated that Tanjung Malim has a return period of one occurrence per 43 days. Meanwhile, Shah Alam and Cheras will be expected to experience the exceedances of one occurrence per 15 and 10 days, respectively. Then, Klang will have the least number of days in a year since it has the return period of one occurrence per 73 days. Table 4.7 also shows the average number of exceedances that can occur in a year, 2 years, 5 years, and 10 years. Klang has a minimum number of exceedances compared to other monitoring stations while Cheras has the highest number of exceedances, followed by Shah Alam.

4.9 Validation Result

The daily maximum O_3 concentrations data for 2016 and 2017 were used to determine the validity of using Bayesian estimator in the GEV distribution. By using the monitoring stations that has high compliance that has been discussed in Section 4.7 except for Jerantut since it has no data that exceed 0.1 ppm. Therefore, the prediction from the number of exceedances using two years of data in Table 4.7 and the actual number of exceedances for that years are shown in Table 4.8.

Table 4.8
Probability and Estimated Number of Days vs. Actual Number of Days for validation

Monitoring Stations	Probability of Exceedances	Actual	Estimated	% Compliance
Shah Alam	0.0681	33	49	51.52
Cheras	0.0954	48	73	47.92
Klang	0.0136	12	10	83.33
Tanjung Malim	0.0234	8	17	12.5

From the validation result in Table 4.8 using two years of data, the percentage compliance for Klang was similar to the percent compliance using overall data. This means that the distribution is suitable for that particular location. It is also indicated that Klang shows a regular number of exceedances for every year compared to other

locations. However, the percentage compliance for Shah Alam, Cheras, and Tanjung Malim was overestimated, and the compliance was far from the compliance using overall data in Table 4.6.

Therefore, it is likely that the distribution was not suitable for that location. There are many factors explaining the validation result are far from the discussion in Section 4.7. One of the factors is based as explained in Figure 2.3 in Section 2.3, and discussion in Section 4.3.2 by the Department of Environment Malaysia (2018), the O_3 for the distribution in the year of 2016 and 2017 were low compared to the other years. Thus, the distribution may not be suitable and differ from the other years.

4.10 Comparison of Bayesian Approach and Classical Approach

Based on Shahrudin (2019), studies on O_3 concentrations use the classical approach to determine the parameter for GEV and Weibull. The study used daily maximum O_3 for four locations, which were Putrajaya, Shah Alam, Klang, and Jerantut in the year of 2007 – 2016. This is almost similar to this very study. From the finding in Shahrudin (2019), it shows that the performance indicator did not show a consistent result for MLE and MOM. The best distribution for MLE was Weibull distribution, but for MOM, it was GEV distribution. Differently, from using the Bayesian approach, all the six locations gave the same results in this study, which is GEV distribution. Based on the performance indicator in the classical approach, studies indicate that the error measure was higher, and the prediction accuracy was lower compared to the Bayesian approach.

CHAPTER FIVE
CONCLUSION AND RECOMMENDATION

5.1 Conclusion

This study was conducted using two extreme value distributions to predict the high exceedances of O_3 concentrations, and the Bayesian approach was applied in this study to introduce the application using O_3. There are three objectives carried out in this study which are to explore the descriptive statistic and trend of the monitoring stations; to determine the suitable prior for Bayesian EVD model and lastly, to determine the best Bayesian EVD for predicting the high exceedances.

The descriptive statistic shows that missing values exist for each monitoring stations since the missing values are common in almost all dataset, which is due to many reasons such as machine error or human error. Thus, this study used the K- Nearest Neighbour imputation method since the method was consistently superior irrespective of the station. The imputation also shows a similar result in descriptive statistic with data before the imputation. This concluded that the imputation does not affect the data and is suitable to be used in this study.

From 2008 to 2017, Tanjung Malim recorded the highest peak of O_3 concentration in 2009. Then, 2013 was the highest peak of O_3 concentration for Klang while the highest peak for Jerantut was in 2015. This is followed by the year 201, which recorded the highest peak of O_3 concentration for Cheras and Tanjung Malim. As for Shah Alam and Petaling Jaya, the highest peak of O_3 concentration were in 2017. There is not much specific particulate event that increases the O_3 concentrations compared to PM_{10} concentrations, and the O_3 concentrations rarely have a long-term of high exceedances based on the time – series plot for each monitoring stations.

Based on the parameter estimation in achieving objectives 2 and 3, this study indicated that the GEV distribution has a higher estimated value compared to the Weibull distribution. This is because the GEV distributions fit the O_3 concentrations well compared to that of the Weibull distribution. Since the GEV distribution fitted

well, this indicated that uniform prior is suitable to be used in this study. The performance indicator also indicates that the GEV distribution is the best distribution for every monitoring stations in this study since the error measure is the smallest, and the prediction accuracy is closest to 1. The Bayesian analysis also shows a consistent result since GEV was the best distribution for every location. Thus, GEV distribution was used to find the probability of exceedances based on the cdf plot and to predict the return period for each of the monitoring stations in this study.

The percent compliance for using overall data for monitoring stations at Klang and the validation was the same. It is shown that distribution is suitable for Klang monitoring stations to predict the high exceedances of O_3 concentrations. The return period predicted that there would be one occurrence of high exceedance occurring per 73 days. This prediction could be beneficial to be used for government in controlling air pollution. This study was also compared with the study that has been done by previous researchers who used the classical approach. The classical approach by Shahrudin, 2019 showed that the best distribution for O_3 concentration was the Weibull distribution. However, the performance indicator using the classical approach was lower compared to the Bayesian approach in this study.

Although GEV Bayesian was suitable in this analysis of concentrations for these 6 locations in Peninsular Malaysia, nonetheless, it might be otherwise for other locations. No single and definite model can be claimed as the best estimation method in the field of extreme value distributions that fit all the monitoring stations in Malaysia. It is recommended that the most current data from all monitoring stations be considered for future studies to get an insight into the appropriate model that can represent each location.

5.2 Limitations

The study was only limited to 6 locations. Hence the proposed distribution is only restricted to the particular monitoring stations. There is no specific model that fits best for all locations. Even though Bayesian is a good approach to determine the distribution parameter; the distribution used must be suitable with the data to get an accurate result. Hence, the selection of distribution is essential in the study. This study

only used two distribution due to the limited time. The access of new record must be made in order for the cutting–edge and up–to–date version may be proposed for the prediction of extreme exceedances at some point in Malaysia, considering the occurrence of excessive particulate activities regularly occur every 12 months.

5.3 Recommendations

The Malaysia Environmental Quality report posted every year by the Department of Environment in the Ministry of Natural Resources and Environment, reported that the O_3 is still the main pollutant in Malaysia. Hence, the availability of an appropriate statistical model in predicting future exceedances of awareness avoiding the MAAQG restrict could be beneficial for environmentalists and strategists to devise proper movements and control strategies to triumph over the trouble.

The O_3 concentration in Jerantut has not exceeded the 1- hour of MAAQG, Department of Environment should be aware of that location since that location shows a significant increasing trend of O_3 concentration along the year of 2008 – 2017. A comprehensive study that covers more monitoring records from other locations should be engaged in order to obtain suitable distributions to fit other locations as well. The selected monitoring records will be subjected to the availability of extreme concentrations recorded at the respective stations. Despite the long – term data that has been used, it is also suggested to extend this study to obtain a suitable distribution for a different year since this study finds out that the distribution for a different year could give a different result since not all of that long – term data have high concentrations.

The validation of this study comparing the actual days of exceedances with the estimated days of exceedances for 2016 and 2017 gives different results with the overall distribution. It is suggested that this research be extended by using another sort of distribution to determine the appropriate model to predict exceedances of O_3 concentrations. This research can also extend the validation of the model from the start by dividing the training and validation set of the data to validate the best Bayesian EVD and then using the validated model to estimate the exceedances of O_3 levels.

REFERENCES

Abdullah, A. M., Ismail, M., Yuen, F. S., Abdullah, S., & Elhadi, R. E. (2017). The relationship between daily maximum temperature and daily maximum ground level ozone concentration. *Polish Journal of Environmental Studies*, *26*(2), 517–523. https://doi.org/10.15244/pjoes/65366

Ahamad, F., Latif, M. T., Tang, R., Juneng, L., Dominick, D., & Juahir, H. (2014). Variation of surface ozone exceedance around Klang Valley, Malaysia. *Atmospheric Research*, *139*, 116–127. https://doi.org/10.1016/j.atmosres.2014.01.003

Ahmat, H. (2016). *Prediction of PM10 Concentrations using Extreme Value Distributions (EVD): Classical and Bayesian Approaches* (Doctoral Dissertation). Universiti Sains Malaysia.

Ahmat, H., Yahaya, A. S., & Ramli, N. A. (2014). Prediction of PM 10 Extreme Concentrations using Three Parameters Extreme Value Distributions (EVD). *Sixth International Conference on Postgraduate Education (ICPE-6)*.

Ahmat, H., Yahaya, A. S., & Ramli, N. A. (2015). PM10 analysis for three industrialized areas using extreme value. *Sains Malaysiana*, *44*(2), 175–185. https://doi.org/10.17576/jsm-2015-4402-03

Alam, M. A., Farnham, C., & Emura, K. (2018). Bayesian modeling of flood frequency analysis in bangladesh using Hamiltonian Monte Carlo techniques. *Water (Switzerland)*, *10*(7), 1–21. https://doi.org/10.3390/w10070900

Aryal, G. R., & Tsokos, C. P. (2011). Transmuted Weibull Distribution : A Generalization of the Weibull Probability Distribution. *European Journal of Pure and Applied Mathematics*, *4*(2), 89–102.

Aryal, G., & Tsokos, C. (2009). *On the transmuted extreme value distribution with application. Nonlinear Anal. Theory, Methods Appl.* https://doi.org/10.1016/j.na.2009.01.168

Awang, N. R., Elbayoumi, M., Ramli, N. A., & Yahaya, A. S. (2016). Diurnal variations of ground-level ozone in three port cities in Malaysia. *Air Quality, Atmosphere and Health*, *9*(1), 25–39. https://doi.org/10.1007/s11869-015-0334-7

Awang, N. R., Ramli, A., Mohammed, N. I., & Yahaya, A. S. (2013). Time Series Evaluation of Ozone Concentrations in Malaysia Based on Location of

Monitoring Stations. *International Journal of Engineering and Technology*, *3*(3), 390–394. https://doi.org/10.1016/j.earscirev.2014.02.005

Azid, A., Juahir, H., Toriman, M. E., Endut, A., Kamarudin, M. K. A., Rahman, M. N. A., … Yunus, K. (2015). Source apportionment of air pollution: A case study in Malaysia. *Jurnal Teknologi*, *72*(1), 83–88. https://doi.org/10.11113/jt.v72.2934

Bali, T. G. (2003). The generalized extreme value distribution. *Economics Letters*, *79*(3), 423–427. https://doi.org/10.1016/S0165-1765(03)00035-1

Banan, N., Latif, M. T., Juneng, L., & Ahamad, F. (2013). Characteristics of Surface Ozone Concentrations at Stations with Different Backgrounds in the Malaysian Peninsula. *Aerosol and Air Quality Research*, *13*(3), 1090–1106. https://doi.org/10.4209/aaqr.2012.09.0259

Cheong, R. Y., & Gabda, D. (2018). Frequency Analysis of Annual Maximum River Flow by Generalized Extreme Value Distribution with Bayesian MCMC. *Journal of Computer Science & Computational Mathematics*, *8*(4), 77–81. https://doi.org/10.20967/jcscm.2018.04.004

Chikobvu, D., & Chifurira, R. (2015). Modelling of extreme minimum rainfall using generalised extreme value distribution for Zimbabwe. *South African Journal of Science*, *111*(9–10), 1–8. https://doi.org/10.17159/sajs.2015/20140271

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. https://doi.org/10.1007/s12205-013-0189-0

Coles, S. (2001). An Introduction to Statistical Modeling of Extreme Values. https://doi.org/10.1007/978-1-4471-3675-0

Coles, S., Pericchi, L. R., & Sisson, S. (2003). A fully probabilistic approach to extreme rainfall modeling. *Journal of Hydrology*, *273*(1–4), 35–50. https://doi.org/10.1016/S0022-1694(02)00353-0

Department of Environment Malaysia. (2015). *Malaysia Environmental Quality Report 2014*.

Department of Environment Malaysia. (2016). *Malaysia Environmental Quality Report 2015*.

Department of Environment Malaysia. (2017). *Malaysia Environmental Quality Report 2016*.

Department of Environment Malaysia. (2018). *Malaysia Environmental Quality Report 2017*.

Durkhure, P., & Lodwal, A. (2014). Fault diagnosis of ball bearing using time domain analysis and fast fourier transformation.

Eli, A. (2012). Preliminary Study on Bayesian Extreme Rainfall Analysis : A Case Study. *Sains Malaysiana*, *41*(11), 1403–1410.

Escarela, G. (2012). Extreme value modeling for the analysis and prediction of time series of extreme tropospheric ozone levels: A case study. *Journal of the Air and Waste Management Association*, *62*(6), 651–661. https://doi.org/10.1080/10962247.2012.665414

Fadzly, M. K., Nordin, F., & Rashid, I. (2018). Impact of temperature to the air pollution in Malaysia: A case study in 10 days cycle, *020170*, 020170. https://doi.org/10.1063/1.5066811

Fernández, A. J. (2006). Bayesian Estimation Based on Trimmed Samples from Pareto Populations. *Comput. Stat. Data Anal.*, *51*(2), 1119–1130. https://doi.org/10.1016/j.csda.2005.11.010

Fisher, M., & Marshall, A. (2009). *Understanding descriptive statistics. Australian critical care : official journal of the Confederation of Australian Critical Care Nurses* (Vol. 22). https://doi.org/10.1016/j.aucc.2008.11.003

Fitri, N., Azam Ramli, N., Yahaya, A. S., Sansuddin, N., Ghazali, N. A., & Al Madhoun, W. (2011). *Central fitting distributions and extreme value distributions for prediction of high PM10 concentration. 2011 International Conference on Multimedia Technology, ICMT 2011.* https://doi.org/10.1109/ICMT.2011.6003204

Georgopoulos, P., & H. Seinfeld, J. (1982). *Statistical distributions of air pollution concentrations. Environ. Sci. Technol.; (United States)* (Vol. 16:7). https://doi.org/10.1021/es00101a002

Ghazali, N. A., Yahaya, A. S., & Mokhtar, M. I. Z. (2014). Predicting Ozone Concentrations Levels Using Probability Distributions. *Journal of Engineering and Applied Sciences*, *9*(11). Retrieved from www.arpnjournals.com

Hashim, N. I. M., & Noor, N. M. (2017). Variations of Ground-level Ozone Concentration in Malaysia: A Case Study in West Coast of Peninsular Malaysia. *MATEC Web of Conferences*, *97*(January), 01048. https://doi.org/10.1051/matecconf/20179701048

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

Hurairah, A., Ibrahim, N. A., Daud, I. Bin, & Haron, K. (2005). An application of a new extreme value distribution to air pollution data. *Management of Environmental Quality: An International Journal*, *16*(1), 17–25. https://doi.org/10.1108/14777830510574317

Ji, L., & Gallo, K. (2006). *An Agreement Coefficient for Image Comparison. Photogrammetric Engineering and Remote Sensing.* (Vol. 73). https://doi.org/10.14358/PERS.72.7.823

Jordan, M. (2010). The Conjugate Prior for the Normal Distribution. *Bayesian Modeling and Inference*, *2*(3), 1–6. https://doi.org/10.1016/j.schres.2006.05.009

Junninen, H., Niska, H., Tuppurainen, K., & Ruuskanen, J. (2004). Methods for imputation of missing values in air quality data sets, *38*, 2895–2907. https://doi.org/10.1016/j.atmosenv.2004.02.026

Kery, M. (2010). *Introduction to WinBugs for Ecologists. Academic Press.*

Kotz, S., & Nadarajah, S. (2000). *Extreme Value Distributions*. Published by Imperial College Press and Distributed by World Scientific Publishing Co. https://doi.org/doi:10.1142/p191

Kruschke, J. K. (2010). Doing Bayesian Data Analysis: A Tutorial with R and BUGS. https://doi.org/10.1128/AAC.03728-14

Kuchenhoff, H., & Thamerus, M. (1995). Extreme value analysis of Munich air pollution data. *Sonderforschungsbereich*, *4*, 1–24.

Lima, C. H. R., Kwon, H.-H., & Kim, Y.-T. (2018). A local-regional scaling-invariant Bayesian GEV model for estimating rainfall IDF curves in a future climate. *Journal of Hydrology*, *566*(August), 73–88. https://doi.org/10.1016/j.jhydrol.2018.08.075

Little, R. J. A., & Rubin, D. B. (2002). *Statistical Analysis with Missing Data. John Wiley & Sons, Inc.* (Second Edi). https://doi.org/10.2307/3172915

Mabahwi, N. A., Ling, O., Leh, H., & Omar, D. (2015). Urban Air Quality and Human Health Effects in Selangor , Malaysia. *Procedia - Social and Behavioral Sciences*, *170*, 282–291. https://doi.org/10.1016/j.sbspro.2015.01.038

Martins, L. D., Wikuats, C. F. H., Capucim, M. N., de Almeida, D. S., da Costa, S. C., Albuquerque, T., … 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*(September), 44–54. https://doi.org/10.1016/j.wace.2017.10.004

Martins, O. I., Sam, B. O., & David, S. N. (2015). Classical and Bayesian Markov Chain Monte Carlo (MCMC) modeling of extreme rainfall (1979-2014) in Makurdi, Nigeria. *International Journal of Water Resources and Environmental Engineering*, *7*(9), 123–131. https://doi.org/10.5897/ijwree2015.0588

MathWorks. (2015). *MATLAB® Programming Fundamentals*. https://doi.org/10.1201/9781420048162.ch38

McLeod, A. . (2011). Kendall rank correlation and Mann-Kendall trend test. *Https://CRAN.R-Project.Org/Package=Kendall*.

Miller, I., & Miller, M. (2013). *John E. Freund's Mathematical Statistics with Applications: Pearson New International Edition*. Pearson Education Limited. Retrieved from https://books.google.com.my/books?id=urmpBwAAQBAJ

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

Mohamad, N. D., Ash'aari, Z. H., & Othman, M. (2015). Preliminary Assessment of Air Pollutant Sources Identification at Selected Monitoring Stations in Klang Valley, Malaysia. *Procedia Environmental Sciences*, *30*, 121–126. https://doi.org/10.1016/j.proenv.2015.10.021

Mohammed, Y. A. . M. and S. J. . (2012). Measurement of Ground Level Ozone at Different Locations. *Environmental Sciences*, *8*(3), 311–321.

Mokhtar, M. I. Z., Ghazali, N. A., Nasir, M. Y., & Suhaimi, N. (2016). Modelling Distribution Function of Surface Ozone Concentration for Selected Suburban Areas in Malaysia. *Malaysian Journal of Analytical Science*, *20*(4), 863–869. https://doi.org/10.17576/mjas-2016-2004-21

Molitor, J., Molitor, N. T., Jerrett, M., McConnell, R., Gauderman, J., Berhane, K., & Thomas, D. (2006). Bayesian modeling of air pollution health effects with missing exposure data. *American Journal of Epidemiology*, *164*(1), 69–76. https://doi.org/10.1093/aje/kwj150

Nasir, M. Y., Ghazali, N. A., Mokhtar, M. I. Z., & Suhaimi, N. (2016). Fitting Statistical Distributions Functions on Ozone Concentration Data at Coastal Areas.

Malaysian Journal of Analytical Science, *20*(3), 551–559. https://doi.org/10.17576/mjas-2016-2003-13

NHDES. (2015). Smog and Ground-Level Ozone. *New Hampshire Department for Environmental Service*. Retrieved from https://www.des.nh.gov/organization/commissioner/pip/factsheets/ard/documents/ard-13.pdf

Noor, N. M., Tan, C., Ramli, N. A., Yahaya, A. S., & Yusof, N. F. F. M. (2011). Assessment of Various Probability Distributions to Model Pm 10 Concentration for Industrialized Area in Peninsula Malaysia : A Case Study in Shah Alam and Nilai. *Australian Journal of Basic and Applied Sciences*, *5*(12), 2796–2811.

Ntzoufras, I. (2009). *Bayesian Modeling Using WinBUGS. A JOHN WILEY & SONS.*

Othman, J., Sahani, M., Mahmud, M., & Ahmad, M. K. S. (2014). Transboundary smoke haze pollution in Malaysia: Inpatient health impacts and economic valuation. *Environmental Pollution*, *189*, 194–201.

Ouyang Zisheng, & Yang Xiangqun. (2011). Extreme value flood frequency analysis at water gauging station near the Dongting lake. In *2011 International Symposium on Water Resource and Environmental Protection* (pp. 592–594). IEEE. https://doi.org/10.1109/ISWREP.2011.5893076

Rani, N. L. A., Azid, A., Khalit, S. I., Juahir, H., & Samsudin, M. S. (2018). Air pollution index trend analysis in Malaysia, 2010-15. *Polish Journal of Environmental Studies*, *27*(2), 801–808. https://doi.org/10.15244/pjoes/75964

Reich, B., Cooley, D., Foley, K., Napelenok, S., & Shaby, B. (2013). Extreme value analysis for evaluating ozone control strategies. *Annals of Applied Statistics*, *7*(2), 739–762. https://doi.org/10.1214/13-AOAS628

Rinne, H. (2009). *The Weibull Distribution A Handbook*. https://doi.org/10.1201/9781420087444

Selaman, O. S., Said, S., & Putuhena, F. J. (2007). Flood Frequency Analysis for Sarawak using Weibull,Gringorten and L-Moments Formula. *Journal The Institution of Engineers,Malaysia*, *68*(1), 43–52.

Shahrudin, S. N. (2019). *Prediction of Ground Level Ozone Concentrations using Weibull and Generalized Extreme Value (GEV) Distributions* (Master Dissertation). Universiti Teknologi Mara.

Smith, E. (2005). *Bayesian Modelling of Extreme Rainfall*. University of Newcastle.

Smith, R. L. (1989). Extreme Value Analysis of Environmental Time Series: An Application to Trend Detection in Ground-Level Ozone. *Statistical Science*, *4*(4), 367–377. https://doi.org/10.1214/ss/1177012400

Vidal, I. (2014). A Bayesian analysis of the Gumbel distribution: an application to extreme rainfall data. *Stochastic Environmental Research and Risk Assessment*, *28*(3), 571–582. https://doi.org/10.1007/s00477-013-0773-3

Wahid, A. (2006). Influence of atmospheric pollutants on agriculture in developing countries: a case study with three new wheat varieties in Pakistan. *The Science of the Total Environment*, *371*(1–3), 304–313. https://doi.org/10.1016/j.scitotenv.2006.06.017

Warren, J., & Gilbert, R. O. (1988). Statistical Methods for Environmental Pollution Monitoring. *Technometrics*, *30*(3), 348. https://doi.org/10.2307/1270090

Williamson, D. F., Parker, R. A., & Kendrick, J. S. (1989). *The box plot: A simple visual method to interpret data. Annals of internal medicine* (Vol. 110). https://doi.org/10.1059/0003-4819-110-11-916

Willink, R., & White, R. (2011). Disentangling Classical and Bayesian Approaches to Uncertainty Analysis, 1–19. Retrieved from http://www.bipm.org/cc/CCT/Allowed/26/Disentangling_uncertainty_v14.pdf

World Health Organization. (2006). WHO Air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide, *22*(9), 2070–2071. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/0004698188901096

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

Zainuri, N. A., Jemain, A. A., & Muda, N. (2015). A Comparison of Various Imputation Methods for Missing Values in Air Quality Data. *Sains Malaysiana*, *44*(3), 449–456.

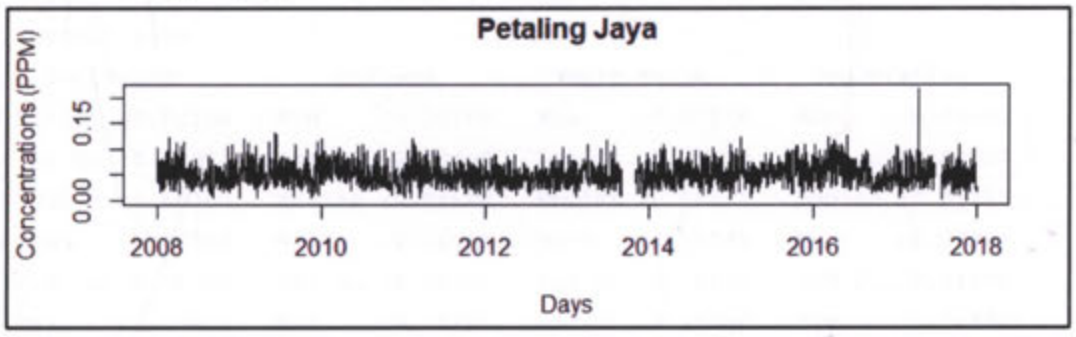
Zhang, W., Cao, Y., Zhu, Y., Wu, Y., Ji, X., He, Y., … Wang, W. (2017). Flood frequency analysis for alterations of extreme maximum water levels in the Pearl River Delta. *Ocean Engineering*, *129*, 117–132. https://doi.org/10.1016/J.OCEANENG.2016.11.013

APPENDICES

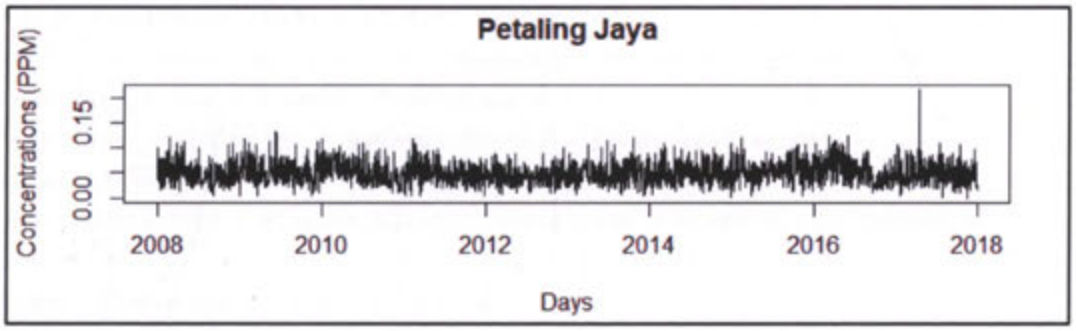
APPENDIX 1

Example R coding for data imputation

```
##############################################################################
##----------------------------IMPORT DATA-------------------------------##
##############################################################################
data=read.table("C:/Users/HP/OneDrive - Universiti Teknologi
MARA/Dissertation July 2019/1. Data/OriginalData.csv", header=TRUE,sep=",")
date=seq.Date(as.Date("2008-01-01"), as.Date("2017-12-31"), "days")
ozone=cbind(date,ozone)

summary(ozone)
#       date                  PJ                 SA                Klang
# Min.   :2008-01-01   Min.   :0.00030   Min.   :0.0010   Min.   :0.0010
# 1st Qu.:2010-07-02   1st Qu.:0.03400   1st Qu.:0.0440   1st Qu.:0.0340
# Median :2012-12-31   Median :0.04700   Median :0.0580   Median :0.0460
# Mean   :2012-12-31   Mean   :0.04953   Mean   :0.0609   Mean   :0.0486
# 3rd Qu.:2015-07-02   3rd Qu.:0.06200   3rd Qu.:0.0740   3rd Qu.:0.0600
# Max.   :2017-12-31   Max.   :0.21560   Max.   :0.1681   Max.   :0.1410
#                      NA's   :96        NA's   :655      NA's   :419
#      Cheras          Tanjung.Malim     Jerantut
# Min.   :0.00100   Min.   :0.00200   Min.   :0.00100
# 1st Qu.:0.04900   1st Qu.:0.03600   1st Qu.:0.02300
# Median :0.06325   Median :0.04500   Median :0.03200
# Mean   :0.06577   Mean   :0.04948   Mean   :0.03307
# 3rd Qu.:0.08000   3rd Qu.:0.05800   3rd Qu.:0.04100
# Max.   :0.17600   Max.   :0.13900   Max.   :0.08300
# NA's   :209       NA's   :73        NA's   :258


##------------------------- Example Plotting---------------------------##

attach(ozone)
plot(date,PJ,type="l",ylab="Concentrations(PPM)",xlab="Days",
main="Petaling Jaya")
```



```
##############################################################################
##----------------------Data Cleaning - KNN IMPUTATION-------------------##
##-------------------------Example for Petaling Jaya----------------------##
##############################################################################
library(HotDeckImputation)
library(forecast)

PJ=as.data.frame(ozone$PJ)
impPJ=impute.NN_HD(PJ)
plot(imp.ozone$date, imp.ozone$impPJ, type="l", ylab="Concentrations (PPM)",
xlab="Days", main="Petaling Jaya")
```

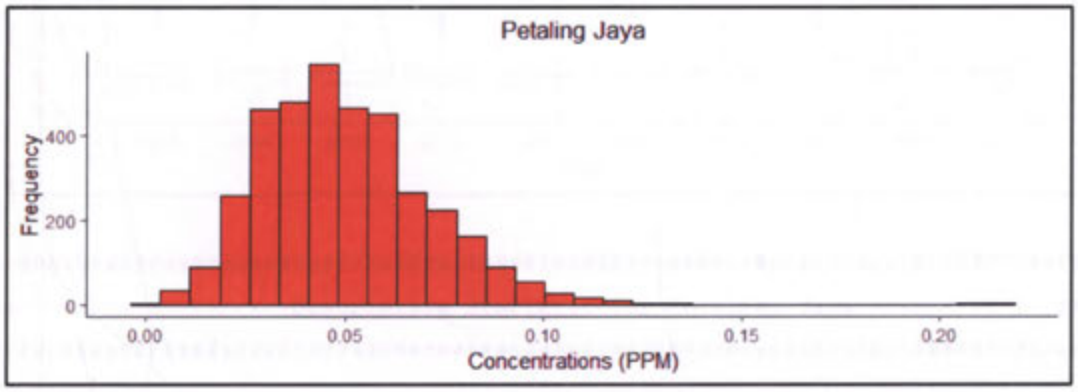



```
##############################################################################
##--------------------Summary for all monitoring stations-----------------##
##############################################################################

DATA=as.data.frame(cbind(impPJ,impSA,impCheras,impKlang,impTg.Malim,impJeran
tut))
Year=ozone$Year
imp.ozone=cbind(Year,date,DATA)
summary(imp.ozone)
#       Year           date                impPJ             impSA
# 2008    : 366   Min.   :2008-01-01   Min.   :0.00030   Min.   :0.00100
# 2012    : 366   1st Qu.:2010-07-02   1st Qu.:0.03500   1st Qu.:0.04500
# 2016    : 366   Median :2012-12-31   Median :0.04800   Median :0.05900
# 2009    : 365   Mean   :2012-12-31   Mean   :0.04974   Mean   :0.06187
# 2010    : 365   3rd Qu.:2015-07-02   3rd Qu.:0.06200   3rd Qu.:0.07500
# 2011    : 365   Max.   :2017-12-31   Max.   :0.21560   Max.   :0.16810
# (Other):1460
#    impCheras          impKlang         impTg.Malim       impJerantut
# Min.   :0.00100   Min.   :0.00100   Min.   :0.00200   Min.   :0.00100
# 1st Qu.:0.04900   1st Qu.:0.03400   1st Qu.:0.03600   1st Qu.:0.02300
# Median :0.06300   Median :0.04600   Median :0.04500   Median :0.03200
# Mean   :0.06559   Mean   :0.04831   Mean   :0.04942   Mean   :0.03287
# 3rd Qu.:0.08000   3rd Qu.:0.06000   3rd Qu.:0.05800   3rd Qu.:0.04100
# Max.   :0.17600   Max.   :0.14100   Max.   :0.13900   Max.   :0.08300
```

APPENDIX 2

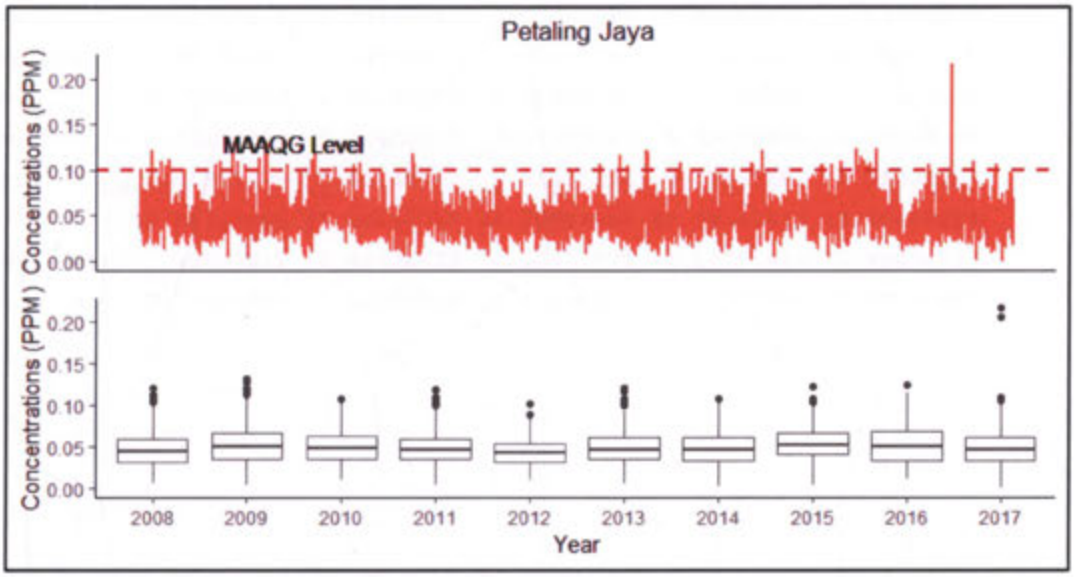
Example R coding for time series, box-plot and descriptive statistic

```
##############################################################################
##----------------------Histogram plot for Petaling Jaya-----------------##
##############################################################################
library(gridExtra)
library(ggplot2)
min=as.Date("2009-10-01")
plot=ggplot(data=imp.ozone, aes(x=Petaling.Jaya))+
geom_histogram(color="black",fill = "tomato")+
labs(title="Petaling Jaya", y="Frequency", x="Concentrations (PPM)")+
theme(plot.title = element_text(hjust = 0.5))+
theme(panel.grid.major = element_blank(), panel.grid.minor =
element_blank(),
panel.background = element_blank(), axis.line = element_line(colour =
"black"))
ggsave("Histogram.PJ.png", plot)
plot
```




```
##############################################################################
##-------------Box-plot and Time series plot for Petaling Jaya----------##
##############################################################################
plot1= ggplot(data=imp.ozone, aes(x=date,y=Petaling.Jaya))+
geom_line(color = "tomato", size = 1)+
labs(title="Petaling Jaya", x="Day", y="Concentrations (PPM)")+
theme(plot.title = element_text(hjust = 0.5))+
theme(axis.title.x=element_blank(),
axis.text.x=element_blank(),axis.ticks.x=element_blank())+
geom_hline(yintercept = 0.1,color="red",size=1, linetype="dashed")+
geom_text(aes(min,0.1,label = "MAAQG Level", vjust = -1))+
theme(panel.grid.major = element_blank(), panel.grid.minor =
element_blank(),
panel.background = element_blank(), axis.line = element_line(colour =
"black"))
```

```
plot2= ggplot(imp.ozone) + geom_boxplot(aes(Year, Petaling.Jaya, group =
Year))+
labs(y="Concentrations (PPM)")+
theme(plot.title = element_text(hjust = 0.5))+
theme(panel.grid.major = element_blank(), panel.grid.minor =
element_blank(),
panel.background = element_blank(), axis.line = element_line(colour =
"black"))

grid.arrange(plot1, plot2, nrow=2)
```




```
##############################################################################
##------------------Descriptive Statistic for Petaling Jaya-------------##
##############################################################################

library(raster)
library(fBasics)

mean=sapply(dataPJ,mean,na.rm=TRUE)
median=sapply(dataPJ,median,na.rm=TRUE)
min=sapply(dataPJ,min,na.rm=TRUE)
max=sapply(dataPJ,max,na.rm=TRUE)
sd=sapply(dataPJ,sd,na.rm=TRUE)
coef.var=sapply(dataPJ,cv,na.rm=TRUE)
skewness=sapply(dataPJ,skewness,na.rm=TRUE)
kurtosis=sapply(dataPJ,kurtosis,na.rm=TRUE)

result=rbind(mean,median,min,max,sd,coef.var,skewness,kurtosis)
```

```
result
#               X2008       X2009       X2010       X2011       X2012
#mean       0.04779781  0.05223562  0.05145479  0.04844110  0.04379235
#median     0.04500000  0.05000000  0.04900000  0.04700000  0.04200000
#min        0.00600000  0.00400000  0.01000000  0.00400000  0.00900000
#max        0.12100000  0.13100000  0.10800000  0.11900000  0.10100000
#sd         0.02014734  0.02183302  0.02003818  0.01919421  0.01656488
#coef.var 42.15116730 41.79718300 38.94328054 39.62380593 37.82595860
#skewness   0.72855627  0.68745363  0.38723604  0.52297015  0.49177360
#kurtosis   0.42359666  0.74908461 -0.44968265  0.39922950 -0.04179814
#               X2013       X2014       X2015       X2016       X2017
#mean       0.04900274  0.04889041  0.05389589  0.05262295  0.04928877
#median     0.04700000  0.04700000  0.05200000  0.05000000  0.04600000
#min        0.00600000  0.00200000  0.00400000  0.01200000  0.00030000
#max        0.12100000  0.10800000  0.12200000  0.12400000  0.21560000
#sd         0.01909498  0.02048382  0.01981730  0.02367161  0.02395762
#coef.var 38.96716086 41.89741715 36.76959458 44.98343316 48.60665224
#skewness   0.40805011  0.42426419  0.34477075  0.54151443  1.90830823
#kurtosis -0.32908983  0.22934023  0.03192501 -0.37973366  9.87740471
```

APPENDIX 3

Example R coding for Mann-Kendall trend test

```
##############################################################################
##--------------Mann-Kendal for Quarterly Average Daily Maximum----------##
##############################################################################

library(Kendall)
PJ.mean=aggregate(ozone1$Petaling.Jaya,FUN=mean,by=list(Quarter=ozone1$Quart
er,Year=ozone1$Year))
colnames(PJ.mean)=c("Quarter","Year","Average")
head(PJ.mean)
#  Quarter Year    Average
#1      Q1 2008 0.05705495
#2      Q2 2008 0.04940659
#3      Q3 2008 0.04076087
#4      Q4 2008 0.04408696
#5      Q1 2009 0.05775556
#6      Q2 2009 0.05295604
summary(PJ.mean)
# Quarter      Year       Average
# Q1:10   2008   : 4   Min.   :0.03742
# Q2:10   2009   : 4   1st Qu.:0.04611
# Q3:10   2010   : 4   Median :0.04951
# Q4:10   2011   : 4   Mean   :0.04977
#         2012   : 4   3rd Qu.:0.05311
#         2013   : 4   Max.   :0.06542
#         (Other):16

MKPJ=MannKendall(PJ.mean$Average)
summary(MKPJ)
#Score =  34 , Var(Score) = 7366.667
#denominator =  780
#tau = 0.0436, 2-sided pvalue =0.70062
```

```
##############################################################################
##-----------------Mann-Kendal for Quarterly Maximum data---------------##
##############################################################################

PJ.max=aggregate(ozone1$Petaling.Jaya,FUN=max,by=list(Quarter=ozone1$Quarter
,Year=ozone1$Year))
colnames(PJ.max)=c("Quarter","Year","Max")
head(PJ.max)
#  Quarter Year   Max
#1      Q1 2008 0.121
#2      Q2 2008 0.112
#3      Q3 2008 0.101
#4      Q4 2008 0.109
#5      Q1 2009 0.119
#6      Q2 2009 0.131
summary(PJ.max)
# Quarter      Year          Max
# Q1:10   2008   : 4   Min.   :0.07900
# Q2:10   2009   : 4   1st Qu.:0.09675
# Q3:10   2010   : 4   Median :0.10640
# Q4:10   2011   : 4   Mean   :0.10766
#         2012   : 4   3rd Qu.:0.11650
#         2013   : 4   Max.   :0.21560
#         (Other):16

MKPJ=MannKendall(PJ.max$Max)
summary(MKPJ)
#Score =  -27 , Var(Score) = 7352.333
#denominator =  773.4727
#tau = -0.0349, 2-sided pvalue =0.76172
```

APPENDIX 4

Example Matlab coding for Performance Indicator

```
n=length(impPJ);
obs=sort(impPJ);

meu = 0.04066;
sigma = 0.01785;
lambda = -0.06868;

p=linspace(0.00001,0.99999,n);
prediction=sort(gevinv(p,lambda, sigma, meu));
pred=transpose(prediction);

%%-------------------------The Bias statistic,B-------------------------%%
B = (sum(pred - obs))/n;

%%--------------------Normalised Absolute Error, NAE --------------------%%
NAE = (sum(abs(pred - obs)))/(sum(obs));

%%-------------------------Prediction Accuracy-------------------------%%
pbar = mean(pred);
obar = mean(obs);
stddevpbar = std(pred);
stddevobar = std(obs);
p2 = pred-pbar;
obs2 = obs-obar;
numPA = p2'*obs2;
denoPA = (n-1)*stddevpbar*stddevobar;
PA = numPA/denoPA;

%%-------------------Coefficient of Determination, rsq------------------%%
%numersq = sum((pred - obar)'*(pred - obar));
%denorsq = sum((obs - obar)'*(obs - obar));
numer = (pred - pbar)'*(obs - obar);
denor = stddevpbar*stddevobar;
RSQ = (numer/(n*denor))^2;

%%----------------------Root Mean Square Error, RMSE--------------------%%
RMSE1 = (pred - obs)'*(pred - obs);
RMSE2 = sum(RMSE1);
RMSE3 = RMSE2/(n-1);
RMSE = sqrt(RMSE3);
```

```
%%-------------------------Index of Agreement (d)-----------------------%%
numd = (pred-obs)'*(pred-obs);
deno1d = abs(pred-obar);
deno2d = abs(obs-obar);
sumdeno3d = deno1d + deno2d;
sqsumdeno4d = sumdeno3d'*sumdeno3d;
IA = 1-(numd/sqsumdeno4d);

%%-------------------------Mean Absolute Error -------------------------%%
MAE = (sum(abs(pred - obs)))/n;


fprintf('Petaling Jaya %F\n')

fprintf('Value of Bias statistic is                          %f \n',B)
fprintf('Value of Normalised Absolute Error statistic is     %f \n',NAE)
fprintf('Value of Prediction Accuracy statistic is           %f \n',PA)
fprintf('Value of Coefficient of Determination statistic is %f \n',RSQ)
fprintf('Value of Root Mean Square Error statistic is        %f \n',RMSE)
fprintf('Value of Index of Agreement statistic is            %f\n',IA)
fprintf('Value of Mean Absolute Error statistic is           %f\n',MAE)

%PI_GEV_PJ
%Value of Bias statistic is                          0.000092
%Value of Normalised Absolute Error statistic is     0.013065
%Value of Prediction Accuracy statistic is           0.997272
%Value of Coefficient of Determination statistic is  0.994007
%Value of Root Mean Square Error statistic is        0.001609
%Value of Index of Agreement statistic is            0.998526
%Value of Mean Absolute Error statistic is           0.000650
```

APPENDIX 5

Example Matlab coding for plotting CDF and PDF

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%----------Cumulative Distribution Function (CDF) for Petaling Jaya-----%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
meu = 0.04066;
sigma = 0.01785;
lambda = -0.06856;

rv = impPJ';
rvsort = sort(rv);
xtranspose = rvsort';
n = length(xtranspose)
meanxtranspose = mean(xtranspose);
stdxtranspose = std(xtranspose);

ii = 1:n;
xx1 = linspace(0,0.25,n);
g = gevcdf(xx1,lambda,sigma,meu);

cdfplot(rvsort)
hold on
plot(xx1,g,'k--')
grid off
line([0.1 0.1], [0 1], 'Color','red','LineStyle','--' )
xlabel ('Ozone Concentrations (PPM)');
ylabel ('cdf,F(x)');
legend('Observation','Theoretical');

%%-----------------------Probability of Excedances ----------------------%%
cdfP=gevcdf(0.1,lambda,sigma,meu)
prob = 1 - cdfP
%%-------------------------------Return Period----------------------------%%

RP = 1/prob

%%-----------------------------Days of Exceedances-----------------------%%
RPyear = 365/RP
RP2year = (365*2)/RP
RP5year = (365*5)/RP
RP10year = (365*10)/RP
```

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%---------Probability Distribution Function (PDF) for Petaling Jaya------%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

meu = 0.04066;
sigma = 0.01785;
lambda = -0.06856;

x = impPJ;
n = length(x);
xx1 = linspace(0,0.25,n);
h = gevpdf(xx1,lambda,sigma,meu);
pd = fitdist(x,'Normal');

y = pdf(pd,xx1);
plot(xx1,y)
hold on

plot(xx1,h,'k-')
xlabel ('Ozone Concentrations (PPM)');
ylabel ('pdf, f(x)');
legend('Observation','Theoretical');
```