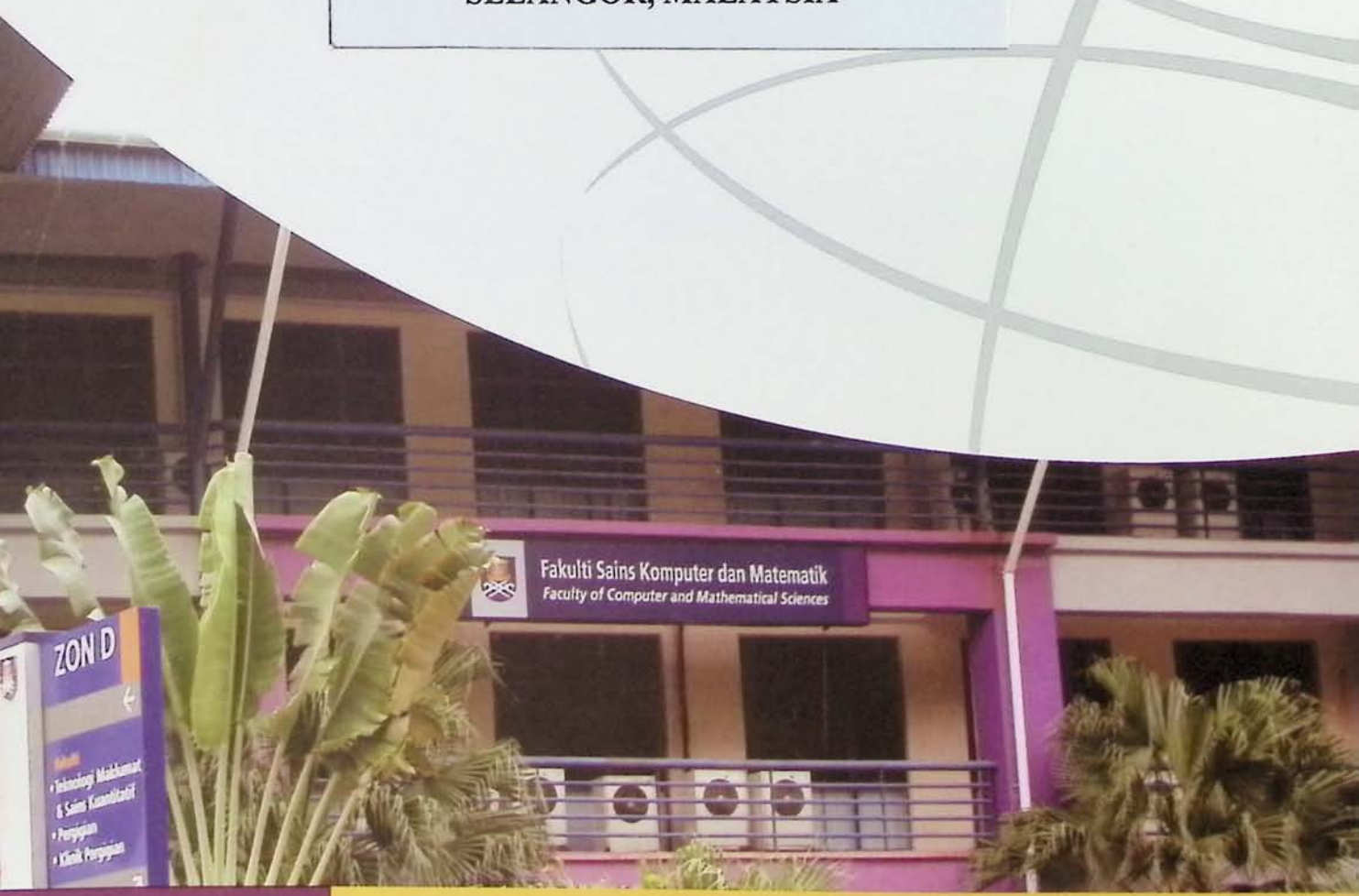
UNIVERSITI TEKNOLOGI MARA

Nombor Sijil: 404073

Fakulti Sains Komputer Dan Matematik

FINAL YEAR PROJECT
MODELLING OZONE AND ITS PRECURSORS TO UNDERSTAND OZONE CONCENTRATION IN SELANGOR, MALAYSIA



UNIVERSITI TEKNOLOGI MARA

FINAL YEAR PROJECT
MODELLING OZONE AND ITS PRECURSORS TO UNDERSTAND OZONE CONCENTRATION IN SELANGOR, MALAYSIA

EMILYA SHAHIRA BINTI AZLI
(2016692838)
NURAZRIN SYAFIQAH BINTI MAT ASRI
(2016692828)
SITI NURLIYANI BINTI AHMAD TAJUDDIN
(2016692902)

JULY 2019

REPORT

SUBMITTED TO

FACULTY OF COMPUTER AND MATHEMATICAL SCIENCES
UNIVERSITI TEKNOLOGI MARA

AS PART OF REQUIREMENT

FOR

BACHELOR OF SCIENCE (HONS.) STATISTICS

JULY 2019

SUPERVISOR'S APPROVAL:

DR. NORSHAHIDA SHAADAN
Faculty of Computer and Mathematical Sciences
Universiti Teknologi MARA
40450 Shah Alam
Selangor Darul Ehsan

UNIVERSITI TEKNOLOGI MARA

MODELLING OZONE AND ITS PRECURSORS TO UNDERSTAND OZONE CONCENTRATION IN SELANGOR, MALAYSIA

EMILYA SHAHIRA BINTI AZLI
NURAZRIN SYAFIQAH BINTI MAT ASRI
SITI NURLIYANI BINTI AHMAD TAJUDDIN

Report submitted in partial fulfillment
of the requirements for the degree of
Bachelor of Science (Hons.)
Statistics

Faculty of Computer and Mathematical Sciences

July 2019

ABSTRACT

Quality of air is vital to sustain life and prevent health problems to human beings. These days, air pollution has become one of the global environmental problems that need to be tackled. Tropospheric ozone is one of the air pollutions that has been recognized to be a threat to human health and have a deleterious impact on plants. This study is conducted to give an awareness and knowledge to the society about the ozone (O_3) pollution and the behaviour at two different locations in Selangor; Shah Alam and Petaling Jaya. In this study, Multiple Linear Regression (MLR) model is employed to understand the associations between the precursors (NO_X, NO, NO_2, SO_2, CO) in the formation of norm (average) O_3 concentration level. This information could provide knowledge on which important day-to-day contributing sources of O_3 pollution as well as the influence of the precursors towards this particular air pollution. Three years hourly data (2015-2017) which was obtained from the Department of Environment Malaysia (DoE) were analysed following a systematic MLR modelling procedure. The procedure involved with data randomization and outlier's treatment at both X and Y direction to achieve a valid and reasonable model since the data are time series and skewed. The results of the analysis have shown that Model 1 that followed the proposed systematic procedure has resulted into a valid MLR model. Based on the model, it is concluded that CO has been identified to be the most important precursors at both locations which indicates that the most important source comes from vehicles at both locations. It is also found that the contribution of SO_2 is significant in Petaling Jaya but not in Shah Alam. This study also provides the evidence that NOx and O_3 has negative association as shown in the previous studies. Lag variables is also shown needed to describe the formation of O_3 and to increase the capability of MLR model.

ACKNOWLEDGEMENT

The time that have been invested and a successful study on our final year project could hardly be possibly done without the helps, supports and guidance from our loved ones. We will take this opportunity to thank to all those people that help us a lot in completing our final year project.

First, we would like to express our deepest gratitude and respect to our supervisor, Dr. Norshahida Shaadan for her valuable guidance, continuous encouragement, and supports to us in completing this project. We appreciate her time and efforts that she had given to us besides giving important suggestions to build a better model in our final year project.

Next, we also want to express wholehearted thanks to the lecturers who had given a hand in showing their moral supports and guidance to us. We are grateful for their good intention and kindness towards us to complete this final year project successfully.

Moreover, we are also would like to say thank you to each member of this group which are ourselves for working hard in finishing this final year project. Even though it takes a lot of time, energy and cost we are still together through ups and downs for this project.

Finally, our special thanks go to those who have assisted us directly or indirectly starting from the day one of this project started until it is done. Your utmost cooperation and kindness are highly appreciated, and may Allah repay all those kindnesses that everyone had given to us.

Thank you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

Symbols

CH_4	Methane
CO	Carbon monoxide
HO_X	Hydrogen Oxide Radicals
NMHC	Non-Methane Hydrocarbon
NO	Nitrogen monoxide
NO_2	Nitrogen dioxide
NO_x	Nitrogen oxide
O_3	Ozone
PM_{10}	Particulate matter 10
RH	Functional Group/ Long Chain Hydrocarbon
SO_2	Sulphur dioxide
THC	Total Hydrocarbon
VOC	Volatile Organic Compounds
μ	Mean
σ^2	Variance

LIST OF ABBREVIATIONS

Abbreviations

ANN	Artificial Neural Network
DoE	Department of Environment
I.I.D	Identically and Independently Distributed
MAQG	Malaysia Air Quality Guideline
MLR	Multiple Linear Regression
MOT	Ministry of Transport
NPBL	Nocturnal Planetary Boundary Layer
SIA	Shuaiba Industrial Area
UV	Ultraviolet Ray
VIF	Variance Inflation Factor
WHO	World Health Organization

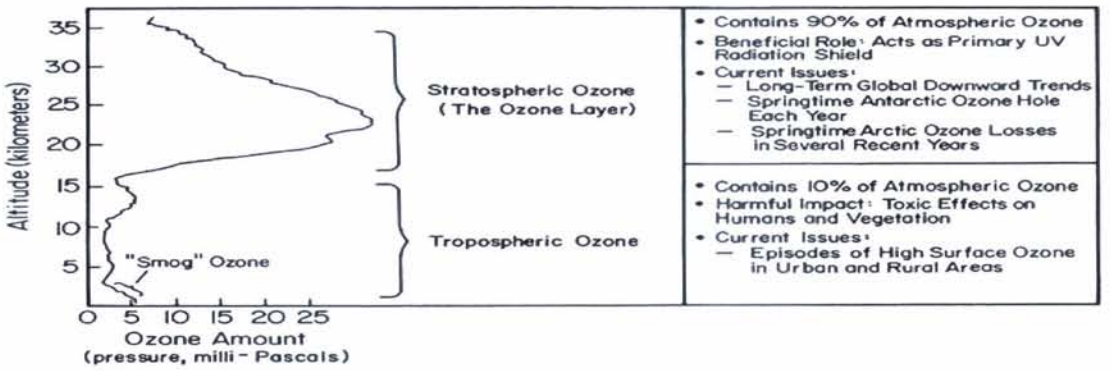
CHAPTER ONE
INTRODUCTION

1.1 Research Background

In Malaysia, ground level ozone (O_3) which is called tropospheric ozone has become one of the most significant air pollutants due to increasing sources of ozone precursors.

O_3 is type of a gas which composed of three atoms of oxygen that occurs both in Earth's upper and lower atmosphere (Ramli et al., 2010). It can be beneficial as good ozone and detrimental as bad ozone. The ozone that is well known by most people as good ozone is the upper ozone called stratospheric ozone. It is forms naturally in the upper atmosphere and protects us from harmful ultraviolet rays (UV). Unlike stratospheric ozone, tropospheric ozone is created through interactions between man-made emissions of volatile organic compounds (VOCs) and nitrogen oxides (NO_x) in the presence of heat and sunlight and this type is harmful if exposed beyond permissible limit.

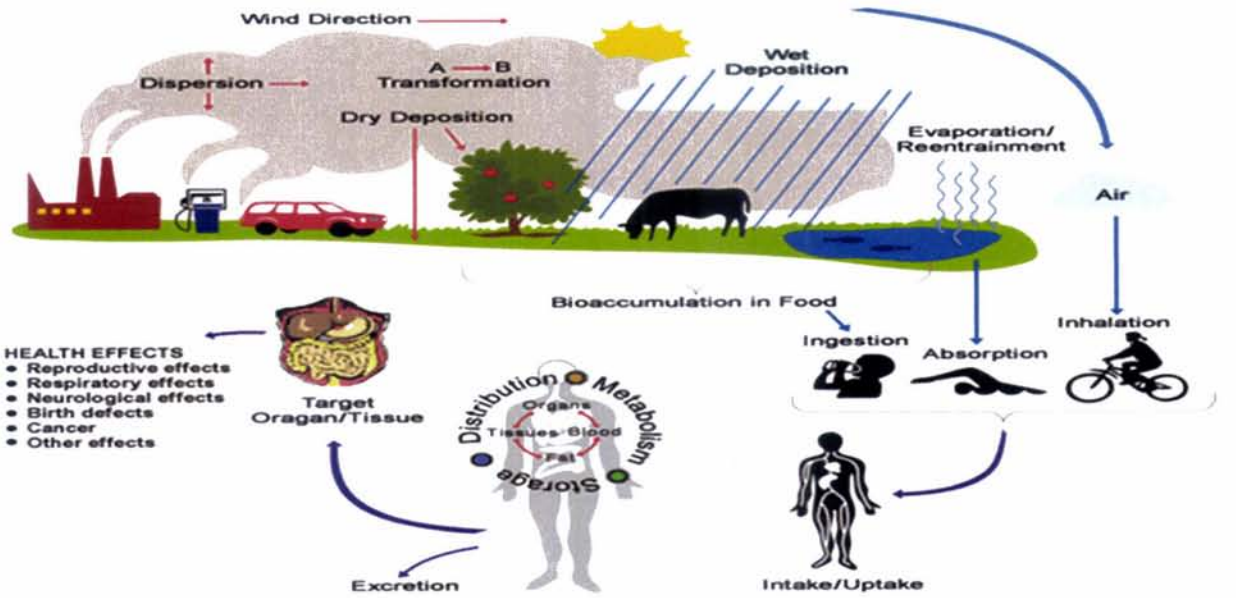
Other than man-made emission, tropospheric ozone is also formed when some of the ozone is transported down from the stratosphere (possibly as much as 50%). The ozone itself is created through the process photolysis of nitrogen dioxide (NO_2) to nitric oxide (NO) which in turn, reacts with carbon monoxide (CO) and hydrocarbons to form O_3. There are no significant primary emissions of ozone in the atmosphere because ozone is secondary pollutant that created by chemical reaction that occurs in air (Abdul-Wahab, Bakheit, & Al-Alawi, 2005).



(Source : World Meteorological Organization)

Figure 1.1 A cross-section of typical vertical ozone profile for the trophic

The influx of ozone from the stratosphere takes place mainly in middle and high latitudes and is most active in early spring. The generation of ozone in the troposphere is most active in summer, since it is caused by photochemical reactions involving nitrogen oxides (NOx), carbon monoxide (CO) and volatile organic compounds (VOCs) (Lelieveld & Dentener, 2000).



(Source : Tibbets, 2014)
Figure 1.2 Effects of ozone

This air pollutant with high concentrations deserves the special attention due to capability of causing adverse health effect towards human health and on the environment (WHO, 2003). Numerous studies were conducted on ozone trend and status in Malaysia environment. (Azmi, Latif, Ismail, Juneng, & Jemain, 2010) reported that ozone concentrations are predominantly related to regional tropical factors (biomass burning) and intensity of UV radiation. In Malaysia, many studies applied statistical distribution on particulate matter (PM_{10}) for the haze level issues but less in O_3 for the ozone concentration. So, we conduct this study to evaluate the most important precursors of ozone formation in two locations in Selangor, Malaysia which are Shah Alam and Petaling Jaya. Thus, the relationship between the study locations towards variability of ozone concentration is going to be explored.

1.2 Problem Statement

In Malaysia, it has been proven that ozone exposure has been linked to the increase in the hospital admissions and emergency room visits for asthma and other respiratory problems. Ozone pollution also reduces the body's resistance to infection. The exposure to high levels of ozone in a long term may lead to huge reductions in lung function, inflammation of the lung lining and more frequent and severe respiratory discomfort. The American Lung Association likens ozone exposure to "sunburn on the lung." Recent studies have also linked ozone to premature death, though further research continues this topic.

As ozone concentrations at several sites in Malaysia often exceeding the recommended Malaysia Air Quality Guidelines (MAQG) value (standard), health effects at the population level become increasingly numerous and severe (Awang et al., 2000). Such effects can occur in places where concentrations are currently high due to human activities or are elevated during episodes of very hot weather.

Air pollution monitoring is very important to inform public awareness of air quality level and the government on the needs to manage and control the source of pollution. This requires understanding on the sources of the emission of air pollutants as well as the interrelationship between the contributing factors.

Even though, Multiple Linear Regressions (MLR) has been considered as the most common model used in the literature to explain the relationship between a particular pollutant and the associated factors, however, noticeably, the application of MLR model as the average mean model for air quality time series data set with the existence of outliers and autocorrelation issue is often neglected. As a result, high potential of artificial and wrongly interpreted model and the coefficients was obtained. This result consequently provides wrong conclusion, thus very dangerous to the decision makers. Since air quality data are often in the form of time series record, its current concentration level often influenced by the previous concentration (auto correlated) and could be explained by multiple dependent variables, MLR modelling involved for air quality data set is supposed to be conducted with precaution. Therefore, to be able to understand the behaviour and the influence of ozone precursors towards O_3 concentration in Petaling Jaya and Shah Alam in Selangor Malaysia, a comprehensive procedure for MLR is going to be used in this study. Thus, the results would provide new insights on the modelling of O_3 for time series data set and its

precursors to understand O_3 concentration besides finding the way and suggestion to prevent its sources from getting worse.

1.3 Research Question

The research questions are as follows: -

a) What is the behaviour of O_3 concentration and its precursors between the study locations, i.e Petaling Jaya and Shah Alam respectively?

b) How to model the relationship between average (norm) O_3 level and its precursors at the study locations, i.e Petaling Jaya and Shah Alam respectively?

c) What is the type of association and how is the influence of the precursors towards average (norm) O_3 at the study locations, i.e Petaling Jaya and Shah Alam respectively?

1.4 Research Objectives

The objectives of this study are:

a) To describe the behaviours of ozone concentration and its precursors at the study locations.

b) To model a linear relationship between average (norm) O_3 and its precursors to understand their behaviour towards O_3 concentration at the study locations.

c) To assess the influence of the precursors towards average (norm) O_3 at the study locations respectively.

1.5 Theoretical Framework

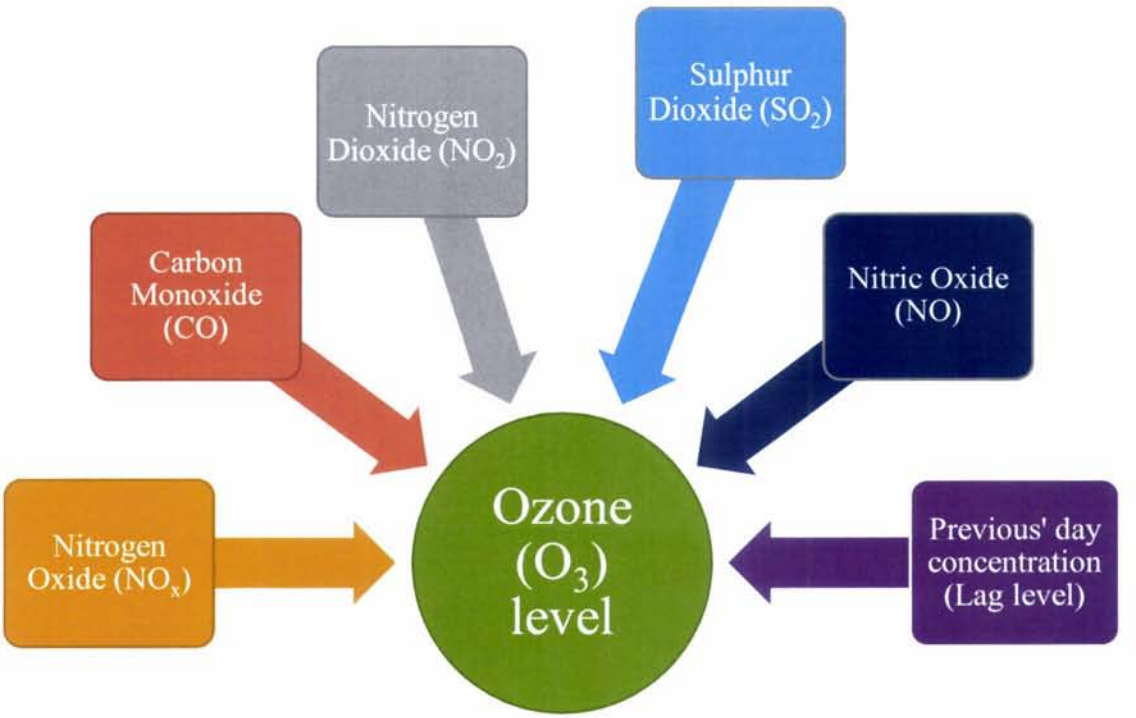


Figure 1.3 Theoretical Framework

Independent variables which are the precursors of the O_3 included in the research framework are nitrogen oxide (NOx), carbon monoxide (CO), nitrogen dioxide (NO_2), sulphur dioxide (SO_2) and nitric oxide (NO) as well as the O_3 lag level whereas the dependent variable include is the ozone level.

1.6 Significance of Study

The findings of this study will contribute to individual knowledge about the behaviour of tropospheric O_3 and help them to adapt with O_3 problems. This is because most of the people are not really exposed about the knowledge of O_3 gas. So, they did not know the real bad things behind the good things that they heard or see. So, by doing this research, it will help the researcher to reveal the importance source of ozone pollution and at the same time, give warnings to people to take precautions when dealing with O_3 problems. High concentration of O_3 may cause several health problems that may put our life on the line.

Besides that, the findings of this study will also give benefits to the Department of Environment Malaysia (DoE) for managing the air pollution issues and at the same time, to provide policy about this matter. The results will also help to identify the possible emission sources that significantly contribute in Petaling Jaya and Shah Alam.

1.7 Scope of Study

This study focuses on two locations in Selangor, Malaysia to find out what are the most important precursors of O_3 formation. The precursors including NO, NO_2, NO_X, SO_2, CO and the previous day level (lag level) is believed to contribute to O_3 formation as it is hazardous and may cause damage to the human being and may lead to death. Besides that, due to the geography of Malaysia which lies close to the Equator, it only experienced hot and humid seasons throughout the year. Therefore, this study aims to find out whether there are differences on behaviours of O_3 in different background of station. Petaling Jaya and Shah Alam may have different characteristics in their vulnerability towards the ozone concentration. Number of precursors emitted may differs in the both locations due to the physical of the area itself.

In this study, Multiple Linear Regression (MLR) is used to investigate the behaviour and influence of precursors towards ozone concentration. In modelling the MLR, precursors such as NO, NO_2, NO_X, SO_2, CO and the lag level act as independent variables whereas ozone concentration level is the dependent variable. There will be several precursors that can be dropped from the model if multicollinearity exists. From the obtained model, the important precursors which lead to the ozone formation can be determined.

1.8 Limitation of Study

Although this study aims to investigate and assess the association and influence of O_3 and the precursors using MLR, there is limitation on the amount and quality of data that were encountered as the data obtained for the findings are secondary data. Since the data include several precursors and it is hourly data for three years, it is also found that the distribution of missing values for each precursor varies. Thus, a prior data cleaning stage needs to be conducted. An important precursor such as VOC that contain high percentage of missingness (>30%) will be discarded from the data set for analysis.

CHAPTER TWO
LITERATURE REVIEW

2.1 Ozone Scenario in Malaysia

Ozone is known to be a secondary pollutant, photochemical oxidant and the main component of smog (Finlayson-Pitts & Pitts Jr, 1999). (Azmi et al., 2010) said that based on the previous study, the results on land use and the level of compliance mentioned that some problems in air quality status in the Malaysian peninsular only exist in highly spot on the urbanization areas. According to the Malaysian department of environment and others study stated that motor vehicles get 82%, power stations 9%, industrial fuel burning 5%, industrial concentration processes 3% domestic and commercial 0.2% and open burning at solid waste disposal site 0.8% to the total air emission load in the peninsular of Malaysian. In addition, the development of the Klang Valley, Malaysia has many types of physical activities such as urban commercial, industrial area, settlement area and others, which has increased the risk of atmospheric pollution (Sulaiman et al., 2017).

2.2 Effects of Ozone

Tropospheric ozone is also recognized to be a threat to human health (WHO, 2003) (Tan et al., 2013) and have a deleterious impact on vegetation (Monks et al., 2015), and through plant damage it impedes the uptake of carbon into the biosphere (Sitch, Cox, Collins, & Huntingford, 2007) as well as impacting built infrastructure (Kumar & Imam, 2013). High level of O_3 has a detrimental effect on plants, as it enters the plant leaves through stomata generates odd species which can oxidize plant tissue resulting in changes in gene expression causing impaired photosynthesis. Ozone concentrations greater than 40 ppbv may be harmful to the crop yield, biomass concentration, vitality and stress tolerance of forest trees (Verma, Satsangi, Lakhani, & Kumari, 2015).

(Fuhrer, Skärby, & Ashmore, 1997) states that excessively high levels of O_3 may be an obstacle to a forests' capacity to seize carbon should there be an excess of carbon dioxide in the future (Karnosky et al., 2003). Moreover, long-term human

exposure to ozone causes chest pains, persistent coughing, respiratory irritation, impaired lung function, sense of dry throat, worsening of previous respiratory diseases like asthma and even irreversible damage to the lungs (Moustris, 2014).

2.3 Formation of Ozone

Ozone is central to the chemistry of the troposphere owing to its role in the initiation of photochemical oxidation processes via direct reaction, photolysis and the subsequent reactions of the photoproducts to form the hydroxyl radical (Monks et al., 2015). Ozone formation by the photochemical reactions of precursors requires at least one-hour time (Marathe, 2018). Early in the 1970s, Crutzen (1973) and Chamei- des and Walker (1973) suggested that tropospheric ozone originated mainly from concentration within the troposphere by photochemical oxidation of CO and hydrocarbons catalysed by HOx and NOx. Tropospheric ozone is largely dependent on meteorology (Ghazali et al., 2010). In an urban atmosphere, short lived VOCs are responsible for O_3 formation rather than CO and CH_4 as their atmospheric lifetime is very high (Seinfeld et al., 2006). According to the Environmental Protection Agency (2007), ground level ozone is formed by a series of reactions, under the influence of sunlight, involving volatile organic compounds (VOCs) combined with a group of air pollutants known as nitrogen oxides (NOx). Volatile organic compounds are emitted by automobiles and various commercial and industrial sources. Nitrogen oxides are by-products of burning fuel in automobiles and heavy industries. Collectively, NOx and VOCs are referred to as ozone precursors (Abdul-Wahab et al., 2005).

2.3.1 Ozone Pollution in Industrial Area

PM_{10} is the main component of dust fall, which it potentially comes from the industrial activities and construction sites, the transportation exhaust emission and soil dust, and open burning activity around the study area (Arsene, Olariu, & Mihalopoulos, 2007). According to the Malaysian Ministry of Transport (MOT), the total amount of new registered motor vehicles in Malaysia was increased 4.42% from 934,367 in 2004 to 1,160,082 in 2010. Based on this information, motor vehicles in Malaysia are one of the major factors that contribute to the deterioration of atmospheric conditions. Moreover, petrochemical processes and fuel or oil burning along with transportation

have been the major causes of atmospheric HCs and NOx. Most of such pollutants are measured by emission rates that are counted with the help of activities going on in urban and industrial areas. In addition, methane (CH_4), non-methane hydrocarbon (NMHC), total hydrocarbon (THC), O_3 and particulate matter under 10 microns (PM_{10}) are responsible for air quality variations in the industrial area (Mohd Kalkausar et al., 2019).

2.3.2 Ozone Pollution in Urban Area

Analysis of plume characteristics and air-mass back trajectories show the highest concentrations of O_3 (and other trace gases) were mainly due to emissions from the Beijing urban area. The highest O_3 concentrations were due to transport of the Beijing plumes superimposed on pollution contributed by regional sources (Lu & Wang, 2006). Most of the NO_2 in urban air results from the rapid oxidation of NO which is the major nitrogenous product emitted from combustion (Safieddine et al., 2013). Moreover, the O_3 concentration in both urban and rural areas with maxima around spring-summer. The O_3 variations in the urban areas are larger than those at the rural sites. The O_3 concentrations are similar between weekend and weekdays in contrast to the findings in many other urban areas in the world. The average ozone concentrations are lower in urban areas compared to the sites outside urban centres (Zheng, Zhong, Wang, Louie, & Li, 2010).

2.4 Previous Study of Ozone in Malaysia

Ozone is a secondary pollutant resulting from photochemical reaction of variety of natural and anthropogenic precursors mainly volatile organic compound (VOCs) and oxides of nitrogen (NOx). There is a positive correlation between concentration of ozone and solar radiation (J/m2. hr) (Abdullah, Ismail, Yuen, Abdullah, & Elhadi, 2017). The concentration of ozone reaches its peak during the middle of the day while during night-time, the concentration lowers (Rahim et al, 2010). The peak occurs at 2 pm (Azmi et al., 2010). During the day, pollutants would be diluted when mixing layer rises while during night-time, it will be limited to only inside nocturnal planetary boundary layer (NPBL). NO and NO_2 are emitted pollutants that are kept beneath the inversion, which may cause the hourly NOx concentration to increase during night (Han

et al., 2011). According to Abdullah et al. (2017), it shows that the daily maximum ozone concentration is higher in industrial area compared to urban area. This indicates that the precursors of ozone mainly NOx and VOC are produced by motor vehicles and smoke from the industrial area. Muhamad et al. (2015) stated that the maximum amount of O_3 concentration was 0.097 ppm, caused mainly by open burning and smokes emitted from vehicles. According to Department of Environment (DoE) 2011, Klang Valley experienced the unhealthy days in year 2012 was due to high concentration level of O_3. Shah Alam was recorded as having the highest number of unhealthy days.

2.5 Statistical Model and Analysis to Investigate Ozone Pollution

The relationships among O_3 precursors, meteorological parameters have been examined by several studies using various statistical techniques, multiple linear regression and principal component analysis (Abdul-Wahab et al., 2005). There are also other several studies that using fuzzy logic, artificial neural network (ANN), graphical analysis and time-series analysis (Sebald et al., 2000). Among these methods, multiple linear regression (MLR) has provided successful results in O_3 modelling studies (Sousa et al., 2007). Specifically, Ziomas et al.,(1995) presented analytical models relating maximum ozone concentrations in Athens area with various meteorological variables. (Spellman, 2000) developed ANN models to be able to forecast surface ozone concentration in five different locations in London, the result indicate that these models surpassed in comparison with the regression models.

2.6 Outliers in Regression Analysis

An outlier is an observation with large residual. In other words, it is an observation where the dependent variable value is unusual given its value on the independent variable. Outliers are often caused by human error or mistakes such as in data collection, recording and entry of data (Obsborne & Overbay, 2013). There are a few numbers of methods to diagnose the outliers. One of them is Cook's Distance (or Cook's D). As a result of the outliers, the values of regression coefficients or summary statistics such as F statistics, R^2 and the residual become very sensitive to these outliers hence the estimation gives very misleading result (Dey, Hossain, & Das, 2015). There are many people who like to see a prior method for identifying outliers usually by

removing or down weighting them by means of robust approach and this method have been frequently applied with apparent success (Schutte & Violette, 1994).

2.7 Using Multiple Linear Regression to Model Ozone Pollution

(Verma et al., 2015) said that multiple linear regression is based on linear and additive association of independent explanatory variables. The following assumptions should be met when MLR is applied :

i) The variables must be independent

ii) The residual errors must be independent and normally distributed with almost zero mean and constant variance (. i.id assumptions).

As the result of their study, O_3 shows negative correlation with NO, NO_2 and RH. As NO and NO_2 act as precursor gases for the photochemical formation of surface O_3 therefore rise in O_3 concentration is associated with a drop in the level of NO and NO_2. O_3 concentration correlated positively with temperature, wind speed and solar radiation.

(Samadianfard, Delirhasannia, Kisi, & Agirre-Basurko, 2013) stated that the goal of the regression analysis is to determine the coefficient values of the parameters of the regression equation and then to quantify the goodness of the fit in respect of the dependent variable Y. The observations $\{X_{i1}, X_{i2}, \ldots, X_{ip}, Y_i\}$, i = 1, 2,…,n is helpful in the estimation of the parameters β and they form the calibration set. The least square method is the usual technique used to estimate the parameters.

There are three wide areas in terms of meteorology that need to be focused for ozone's concentration through statistical methods. Every area of approach is considerably unique from others: first one is regression-based method, the second is extreme value method, and third one is Space-Time approach (Al-Shammari, 2017). In predicting the levels of ozone from meteorological conditions and precursor concentrations at (SIA) Shuaiba Industrial Area of Kuwait during the daylight hours was achieved by using step wise multiple regression modelling (Wahab et Al., 2000).

Moreover, among the most frequently methods used are regression, time series, and neural network models. Multiple regression analysis is one of the most widely used methodologies for expressing the dependence of a response variable on several predictor variables (AbdulWahab et al., 2005). The application of multiple linear

regression techniques allows formulation of explicit equations that are simple and can be used to better understand the processes involved in O_3 formation (Barrero et al. 2005). However, noticeably, the previous studies had ignored the influence of autocorrelation issues in the modelling process, but air quality data are often recorded as time series. On top of that, pollutant concentration is naturally continuous, and the formation process is dependent on time. Thus, its current concentration level often associated with its previous level (Musa & Ibrahim, 2012). There exists a long memory in time series ozone in the Peninsular Malaysia environment. However, based on literature review, many of the studies conducted does not consider the contribution of lag level in the modelling of ozone. One of the studies by (Yahaya et al., 2009) has shown that previous hour ozone level influences the next hour level.

2.8 The Study of the Association between Ozone and its Precursor using MLR

O_3 in the troposphere is generally increasing because of increasing emissions of precursors such as NOx and volatile organic compounds (VOC). NOx and O_3 levels in urban air and amounts of nitrate in the aerosol particles over South Asia indicate that NOx levels are not negligible (Suresh Kumar Reddy et al., 2012). Based on study at semi-arid rural site, Anantapur in southern India, during January-December 2010 the monthly mean variation of tropospheric ozone is similar to that tropospheric NO2, with a correlation coefficient of +0.80.

On other observational study that takes place in Kannur, India from November 2009 to December 2011, the diurnal cycle for surface O_3 had a peak during afternoon and declined in the night-time. The diurnal variations of mixing ratios for NO_x and O_3 were anti-correlated. A positive correlation was found between O_3, NO and NO_2 which revealed the chemistry of O_3 concentration from NO_2 and titration with NO at this location. The net effect of NO_x on O_3 concentrations was negative with a decaying exponential correlation indicating a possible VOC sensitive location, which showed the prominent role of abundant biogenic VOCs on the concentration of O_3 even at lower levels of NO_2 (Nishanth, Praseed, Satheesh Kumar, & Valsaraj, 2014).

To explore the precursor's potential relationship based on application of an observation-based model (OBM), ambient ozone (O_3) and its precursors were simultaneously measured at two sites in the Pearl River Delta, namely, Wan Qing Sha

(WQS) in Guangzhou and Tung Chung (TC) in Hong Kong, started from 23 October to 01 December 2007 in order. According to Hai Guo et. Al, (2009) a large increment in both simulated HO_2 and O_3 concentrations was achieved with additional input of hourly carbonyl data. This suggested that apart from hydrocarbons, carbonyls might significantly contribute to the O_3 concentration in the Pearl River Delta. It also stated that NO was negatively correlated with O_3 concentration, indicating that reducing NO would lead to an increase in O_3 concentration (Cheng et al., 2010).

(Han et al., 2011) found that there is an inverse relationship between O_3, NO, NO_2 and NO_X. In addition, there is also a linear relationship between NO_2 and NO_X, as well as NO and NO_X, and a polynomial relationship between O_3 and NO_2/NO. Due to photochemical O_3 formation, the ozone concentration slowly rise after the sun rises, reaching a maximum during the daytime and then decreases until the next morning.

CHAPTER THREE
RESEARCH METHODOLOGY

3.1 Introduction

This chapter explains the method that will be adopted by this research and gives an outline of the research methods that is followed by this study. It provides information about every component involved and describes the research design that was chosen for the purpose of this study. This chapter also going to discuss the methods used to analyse the data and the procedures that were followed to make sure the appropriate study is being constructed to achieve the study objectives. Software such as R, Statistical Package for Social Sciences and Microsoft Excel were used throughout the analysis process.

3.2 Data and Study Locations

Data consist of ozone concentration (O_3) and the precursors such as NO_2, NO, SO_2, NO_x and CO as well as the lag level which were obtained from Department of Environment (DoE), Federal Territory of Putrajaya, Malaysia. The variables are recorded by hourly basis. Data collected are ranged from January 2015 until December 2017 with total of 26304 hourly data.

The location of study includes Petaling Jaya and Shah Alam. Petaling Jaya was categorized as industrial location while Shah Alam was categorized as urban location with high number of vehicles.

Figure 3.1 Map of Study Location

3.3 Research Framework

The research framework for this study is shown in Figure 3.2

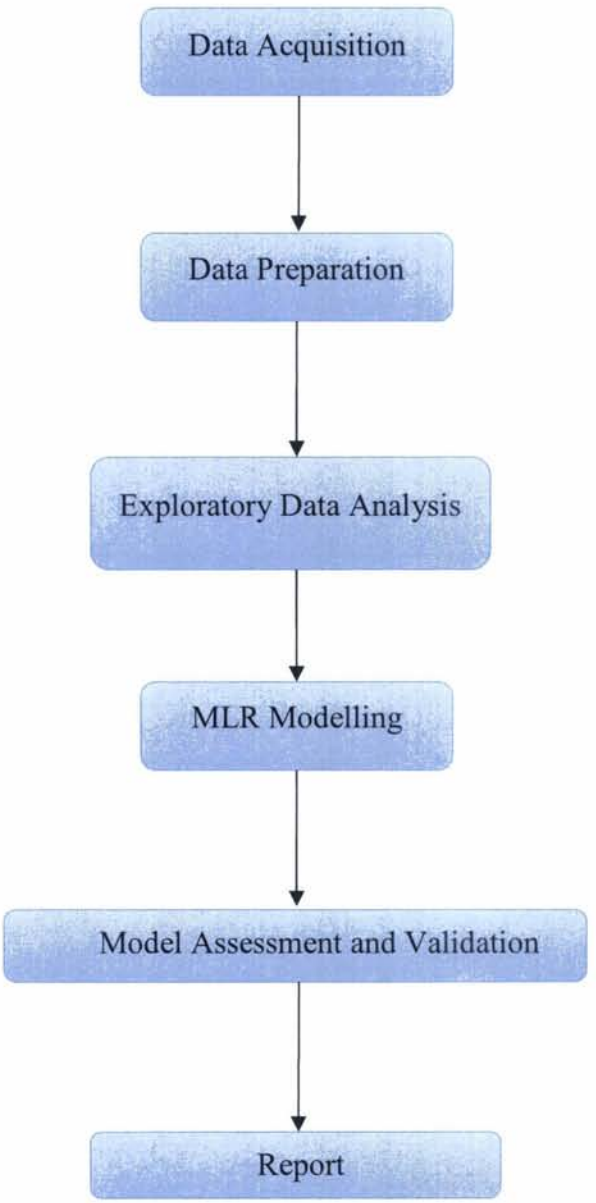


Figure 3.2 Research Framework flowchart

To achieve the research objectives, the activities need to be involved including data acquisition, data preparation, data exploration, modelling the association of O_3 and the precursors using MLR, next the validity and the performance of the model is assessed and finally the activity ends with report writing.

3.4 Multiple Linear Regression Modelling Process

The following Figure 3.4 shows the framework of how MLR model is constructed and established to investigate the association and influence of several precursors towards O_3 level to increase understanding on O_3 concentration or formation at Petaling Jaya and Shah Alam locations.

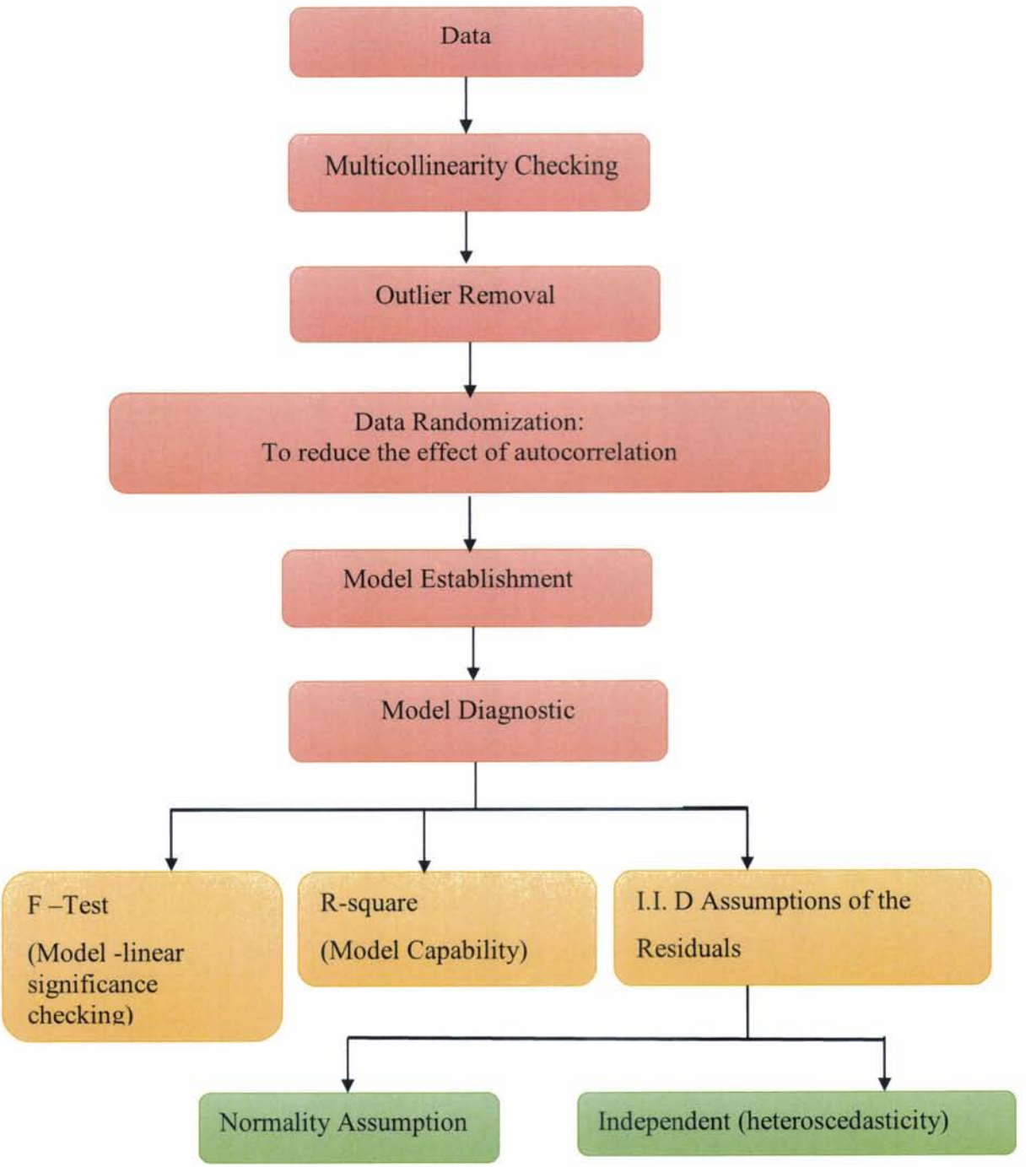


Figure 3.3 Process of MLR Modelling

3.5 Data Preparation

Data preparation describes the process of activities that take place in order to transform a source data into format, quality and suitable structure for further analysis. One of it is dealing with unstandardized, unstructured and or inconsistent data. Data preparation is important to get good quality of data. The better the quality the more likely will be able to turn the data into usable information.

In this project, the data were obtained from Department of Environment, Putrajaya, Malaysia. Data were given from year 2015 until 2017 where it has a record of hourly recorded ozone (O_3) and its precursors. The data used for this project were categorised as secondary data and consists of a very large number. The data obtained were very raw and not treated so there were many missing values. The percentages of missing values were approximately 13.3% in Petaling Jaya and 8.4% in Shah Alam. Since the percentage of missing values in both locations are less than 30%, it is appropriate to use the data. The imputation method is used by imputing the average values of its own variable. Moreover, since our data is in time series format lag variable is added during this step. This is because according to M.A Barrero et. Al, (2005) lag variable is needed to increase the efficiency of the model. Their study findings have shown that the future ozone formation is actually dependent on the previous ozone level.

3.6 Data Randomization

Randomization is a process where subjects are placed or given equal chance to be treated in some treatment or controlled group. The process has already growing into a very fundamental aspect of research methodology, especially in scientific research area. (Kang, Ragan, & Park, 2008)

In this project, data randomization process is done to reduce the effect of autocorrelation in the data set. Data obtained from the Department of Environment is a raw data set in which the data is a time-series data. Time-series data could cause an autocorrelation effect in modelling the multiple linear regression.

3.7 Multicollinearity Checking

The problem of multicollinearity occurs when there are two or more predictor variables, X were analysed simultaneously in a regression model. When the predictor variables are highly correlated, the interpretation of the expected value in the response variable may be incorrect or not valid and lead to misleading conclusions. (Vatcheva, Lee, Mccormick, & Rahbar, 2016).

The way in detecting the problems for multicollinearity is by checking the value of Variance Inflation Factor (VIF) that acts as a tool to measure how the variance of the data are reduced. (Ferdaos, 2017)

This step is needed to check if there is any multicollinearity exist between the precursor's variables, to prevent the statistical inferences made about the data may not be reliable. The result from the VIF values shown that there is multicollinearity exists between variables NO and NO_x in Shah Alam and Petaling Jaya with value more than 10. Variable NO has been removed since the variable is less harmful toward health compared to NO_x. Nitric oxide (NO) is not considered to be hazardous to health at typical ambient concentrations, but nitrogen dioxide does.

3.8 Outlier Removal

In a statistical analysis, an outlier is a data point where its values differs the sample mean (De & De, n.d.). In this project, removing the outliers are done after the results of analysis using boxplot were observed. Boxplot indicates where it shows points that further from the mean. Outliers points tells which of the data point are flawed and may not be valid for further analysis.

3.9 Describing the Behaviour of Ozone and its Precursors

To achieve the above objective of this study, the understandings of the variables are needed. Descriptive analysis is going to be used to describe the basic information and features of the data in this study. The descriptive statistics to represents the data distribution pattern are histogram, box plot and line plot. Other descriptive statistics are used to measure the central of tendency, measures of dispersion and to check the normality of data.

3.10 Modelling the Relationship of Ozone and Its Precursors using MLR

Multiple Linear Regression (MLR) analysis is one of the most widely used of all statistical methods. MLR attempts to model the relationship between two or more explanatory variables, X's and one response variable, Y by fitting a linear equation to the observed data. The population regression line for p independent variables are: -

$Yi = \beta_0 + \beta_i X_i + \epsilon_i$

where:

Y_i – value of response variable in the i^{th} trial

β_0– the intercept of the population regression line

β_i– regression coefficients (parameters) of the explanatory variables

X_i– the value of explanatory variable in the i^{th} trial

ϵ_i– random error term with mean $E\{\epsilon_i\} = 0$ and variance $\sigma^2\{\epsilon_i\} = \sigma^2$

This line describes the mean response Y_i changes with the explanatory variables.

3.10.1 Normality Checking

For the analysis to be reliable and valid, we need to do normality checking. In multiple linear regressions, this assumption applies to the disturbance term, not to the independent variables. Simply put, it is to check the random error in the relationship between the dependent and independent variables. There are several statistics available to examine the normality such as skewness, kurtosis and normal probability plot.

3.10.2 Diagnostic Checking

Diagnostic checking plays a vital role in finding and validating the relationship between the dependent and independent variables. To proceed to the next stage analysis of Multiple Linear Regression (MLR), the assumptions of MLR needs to be satisfied.

This assessment may be an exploration of the model underlying the statistical assumptions. The assumptions of MLR are there is linear relationship between the dependent variable and each dependent variable, the errors are independent, the errors at each set of values of the independent variables are normally distributed and the errors

at each set of values of the independent variables have equal variances (denoted σ^2). If all these assumptions were followed, then we can say that the model that we propose is significant.

For diagnostic checking, we run ANOVA test to check for coefficient of multiple R-squared and significance of each precursors. Multiple R-squared tells the total variation explained by each precursor on ozone level.

4.9.3 Assessment and Validation of MLR Model Obtained

It is important to know whether different background having different level of average ozone concentration. Since every location is in different background activity, so the ozone concentration was expected to be different. The association and influence of the precursors towards O3 concentration will be evaluated based on the obtained model. Hence the performance and the validity of the established model must be assessed.

To assess the performance of the established MLR model given name as Model 1, the obtained model was also compared with another model that is conducted without following the framework proposed in Figure 3.3 given name as Model 2. The purpose of comparison is to highlight the consequence of modelling Time Series Ozone data without considering the autocorrelation effect and when outliers are in the data set.

In this study, for each study location, it has two models; Model 1 is the model with randomized data set and without the outliers and Model 2 is the model with non-randomized data with outliers. For both models, the performances will be checked by using standardized residuals and the multiple R-squared. The standardized residuals of the model, multiple R-squared and i.i.d assumptions checking were used to assess the better performance of the establish model. As an information, outlier's removal involved with removing outliers from both at Xs and Y direction.

CHAPTER FOUR
RESULTS AND DISSCUSSION

4.1 Introduction

In this chapter we are going to analyse two stations which are Petaling Jaya as and Shah Alam. This chapter will cover the data analysis, which includes data cleaning, descriptive analysis, correlation analysis, regression tests and regress by using randomization method. Descriptive analysis describes the behaviour of dependent variable, O_3 and independent variables which are NO_x, NO, SO_2, NO_2 and CO. This important to evaluate whether the samples collected are valid and consistent. Furthermore, Cronbach's alpha technique is used for reliability test. Correlation and multiple regression tests are used to further analyse and explore the relationship between the variables.

4.2 Behaviours of Ozone and its Precursors at Both Study Locations

Petaling Jaya and Shah Alam are known as locations that currently going through a lot of process city development. These areas experiences air pollutions in Klang Valley. Both areas are also surrounded by the residential area which also contributes to the air pollution from vehicular exhaust. Hence, it is important to understand the behaviour of the ozone and its precursors to monitor the air quality and the ozone level in these areas.

4.2.1 Descriptive Analysis of Data

To describe the basic feature in the data of this study, descriptive analysis is used to understand the association between the O_3 and its precursors at both study locations.

a) Petaling Jaya

Table 4.1 shows the descriptive measure of O_3 and the precursors. The maximum values are among O_3, CO and Lag within range 3. The minimum values of all the precursors including O_3 are zero except NO with the value of nearly zero (two small level). For precursor SO_2, it seems that it has outliers because the range between the mean and maximum value is far from one another.

Table 4.1
Descriptive Analysis of O_3 and its Precursors in Petaling Jaya

	Ozone (O_3) Precursor's Variables						
Statistics	NO_X	NO	SO_2	NO_2	O_3	CO	Lag
Min	0.00000	-0.0007	0.00000	0.00000	0.00000	0.00000	0.00000
Median	0.06020	0.03181	0.003694	0.02600	0.02000	1.08000	0.02000
Mean	0.05768	0.03126	0.003694	0.02596	0.23820	0.96480	0.23820
Max	0.33600	0.29670	0.165800	0.13400	3.82800	3.60000	3.82800
Std. Dev	0.02903	0.02275	0.01062	0.05348	0.53481	0.57967	0.53481

b) Shah Alam

Table 4.2 shows the descriptive measure of O_3 and its precursors. The maximum values for SO_2, NO_2, O_3 and lag are within the range 0.1 while for NO is 0.27200 and the maximum value for NO_X is 1.0000. The minimum values for all variables are zero. It seems that for precursor SO_2 it has outliers since the range between mean and maximum is far from one another.

Table 4.2
Descriptive Analysis of O_3 and its Precursors in Shah Alam

	Ozone (O_3) Precursor's Variables						
Statistics	NO_X	NO	SO_2	NO_2	O_3	CO	Lag
Min	0.00000	0.00000	0.00000	0.00000	0.00000	0.0000	0.00000
Median	0.03784	0.01480	0.00200	0.02071	0.02025	0.7552	0.02025
Mean	0.03868	0.01639	0.00259	0.02196	0.02106	0.8193	0.02106
Max	1.00000	0.27200	0.10130	0.11100	0.16810	2.8930	0.16810
Std. Dev	0.02477	0.01705	0.00178	0.01209	0.01937	0.37351	0.01937

4.2.2 Histograms of Ozone and its Precursors

A histogram is the common graph used to graphically summarize and display the frequency distributions of the variables. It can also show the normality of distributions either it is symmetric or not.

a) Petaling Jaya

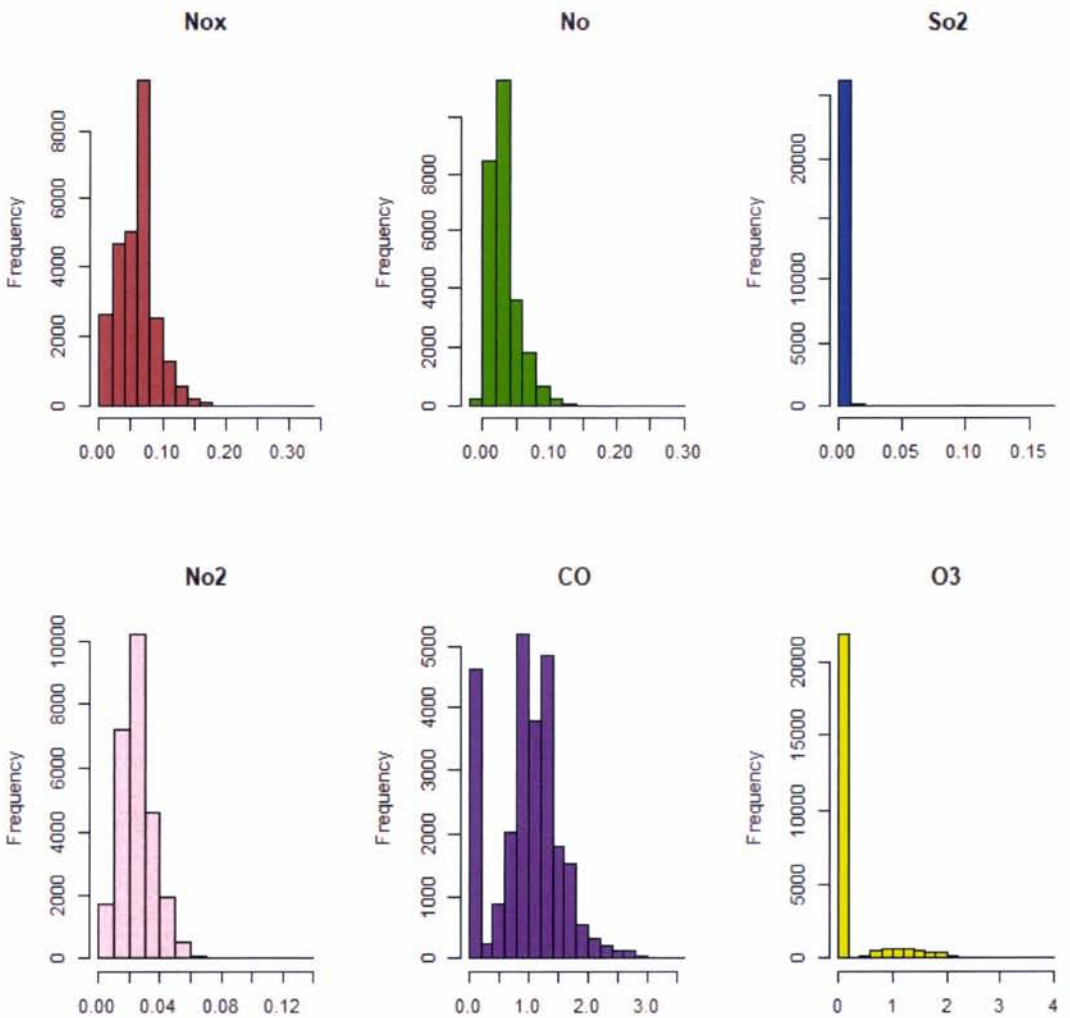


Figure 4.1 Histogram of Ozone and its Precursors in Petaling Jaya

From Figure 4.1, it shows that the distribution of SO_2 is not in symmetric shape. The range within SO2 values between the mean and maximum are quite large. Thus, it leads to produce only one bar. While for NO, NO_X and NO_2, all of them seem to follow normal distribution with a slight skewed to the right. For O_3 and CO, both show

unrecognizable pattern at initial value and followed by the pattern of normal distribution. But for CO, it obviously showed the criteria of the outliers on its graph.

b) Shah Alam

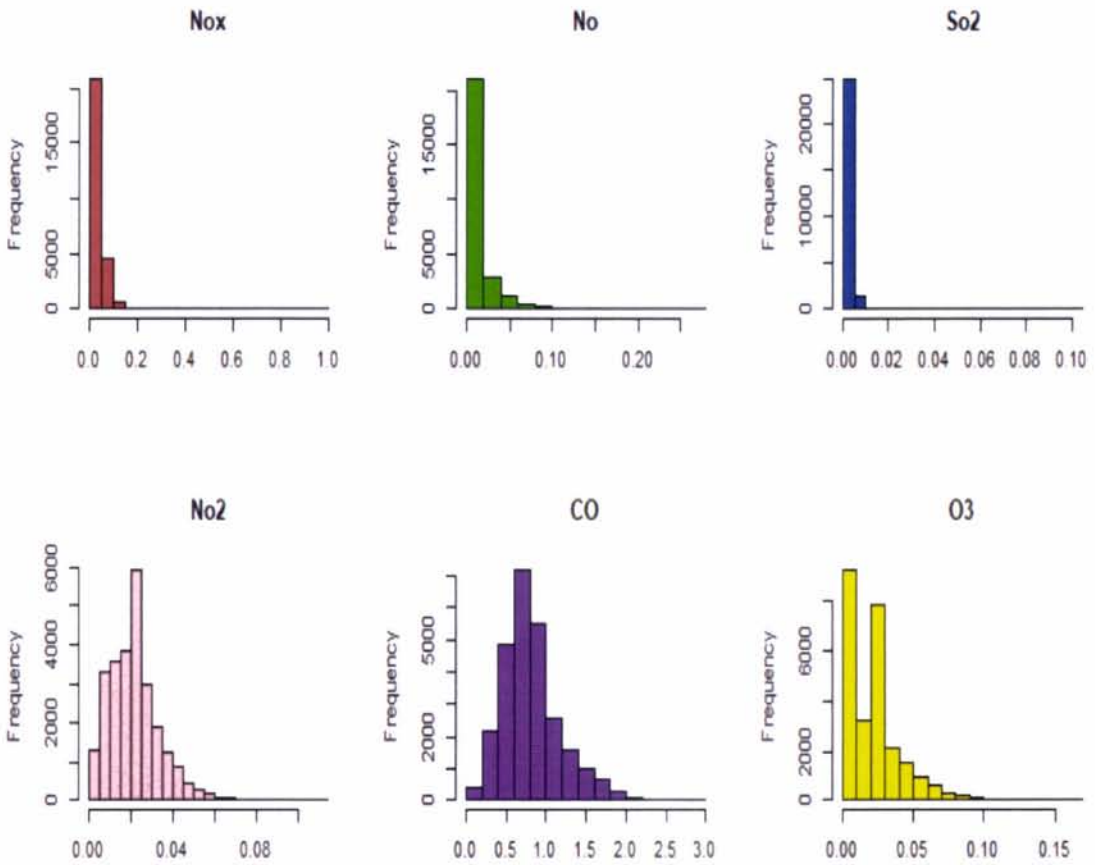


Figure 4.2 Histogram of Ozone and its Precursors in Shah Alam

From Figure 4.2, it shows that NO_X, NO and SO_2 are not normally distributed. This is due to the presence of outliers and the range within NO_X, NO and SO_2 are too small and its lead to produce only several bars for the histogram. The distribution for NO_2 and CO is normal which is with slightly skewed to the right. The histogram for O_3 shows that it does not follows normal distribution and skewed to the right.

4.2.3 Box Plot of the Ozone and its Precursors

These box plots were produced to see the graphical view of the outliers more precise. Moreover, it is also being carried out as the reference to look at the cut-off points to remove the outliers later in the step of building the model in MLR.

a) Petaling Jaya

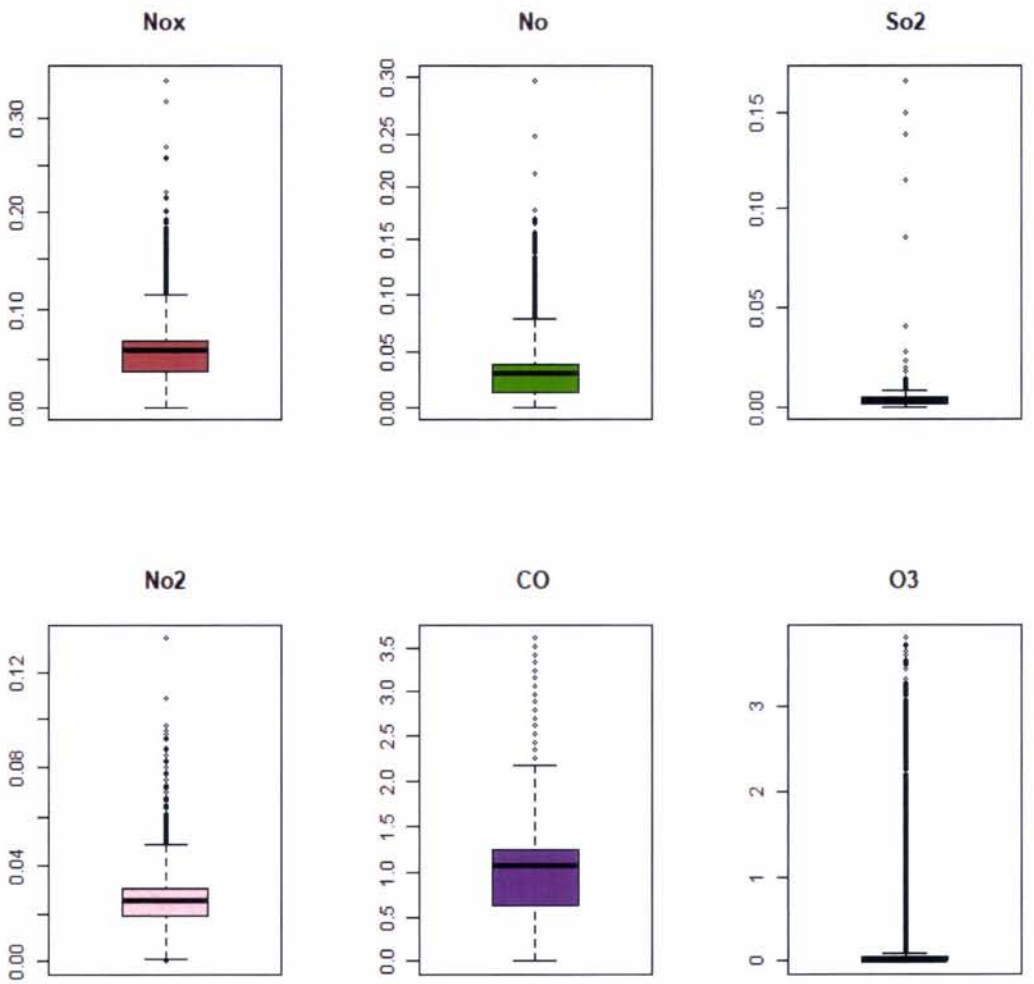


Figure 4.3 Box Plot of Ozone and its Precursors in Petaling Jaya

Figure 4.3 shows the box plots for each of the precursors and the O_3. Each of the variables above has outliers. As same as the result in figure 4.1 before that showed SO2 and CO has outliers, this time they are truly and clearly showed in this box plot graph. Both have the worst box plot measures compared to others. NO_X, NO and NO tend to have a quite similar pattern of box plot while CO shows a wider range of box plot compared to other precursors.

b) Shah Alam

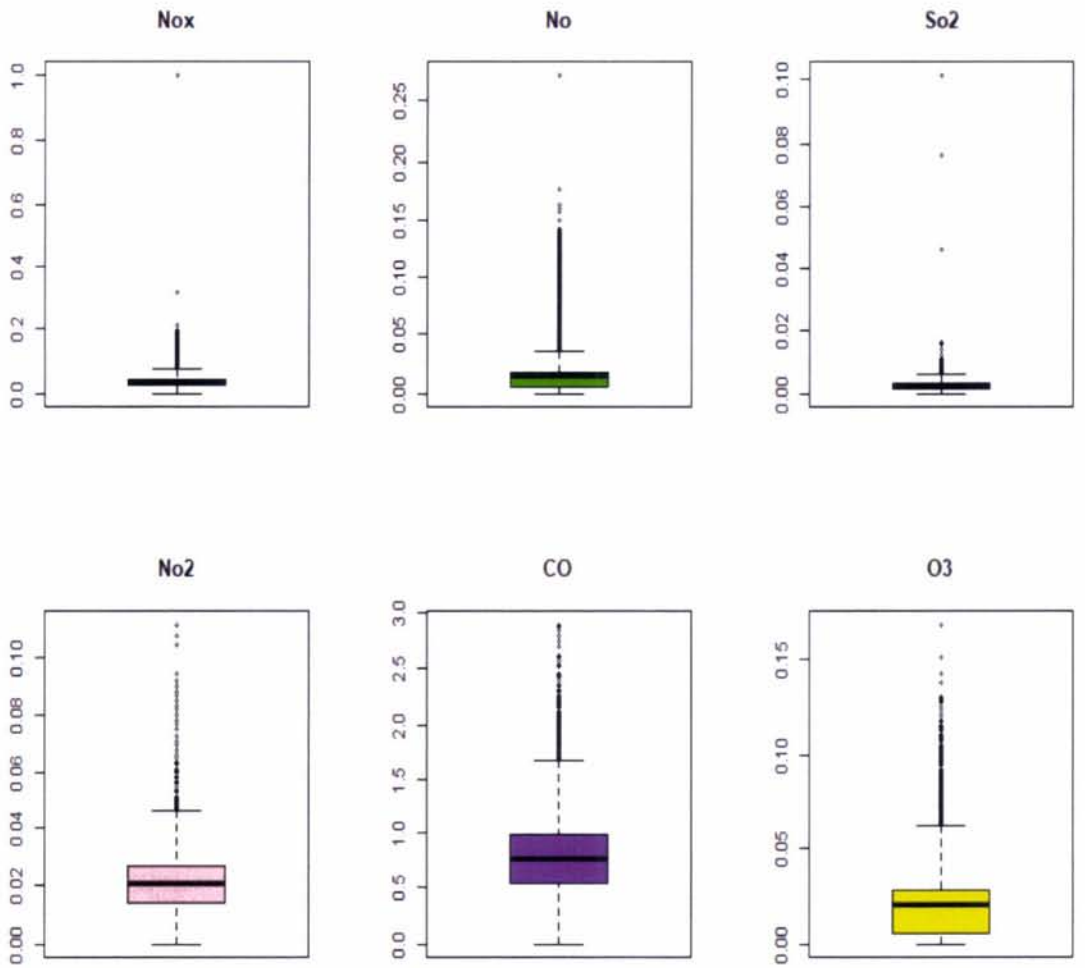


Figure 4.4 Box Plot of Ozone and its Precursors in Shah Alam

Figure 4.4 shows the box plots for each of the precursors and the O_3. All the precursors have outliers. Like the result in Figure 4.3 that showed NO_X, NO and SO_2 have outliers with extreme values and now are clearly showed in box plot graph. These three precursors have the worst box plot measures compared to others. Box plot for NO_2 and CO shows that both are normally distributed with wider range of box plot while for O_3 seem not follow normal distribution.

4.2.4 Scatter Plot of the Precursors that Associated with the Ozone Level

The relationship between the variables can be seen through scatter plot matrix. It depicts a relationship between each of the variables, but it does not indicate a cause and effect relationship.

a) Petaling Jaya

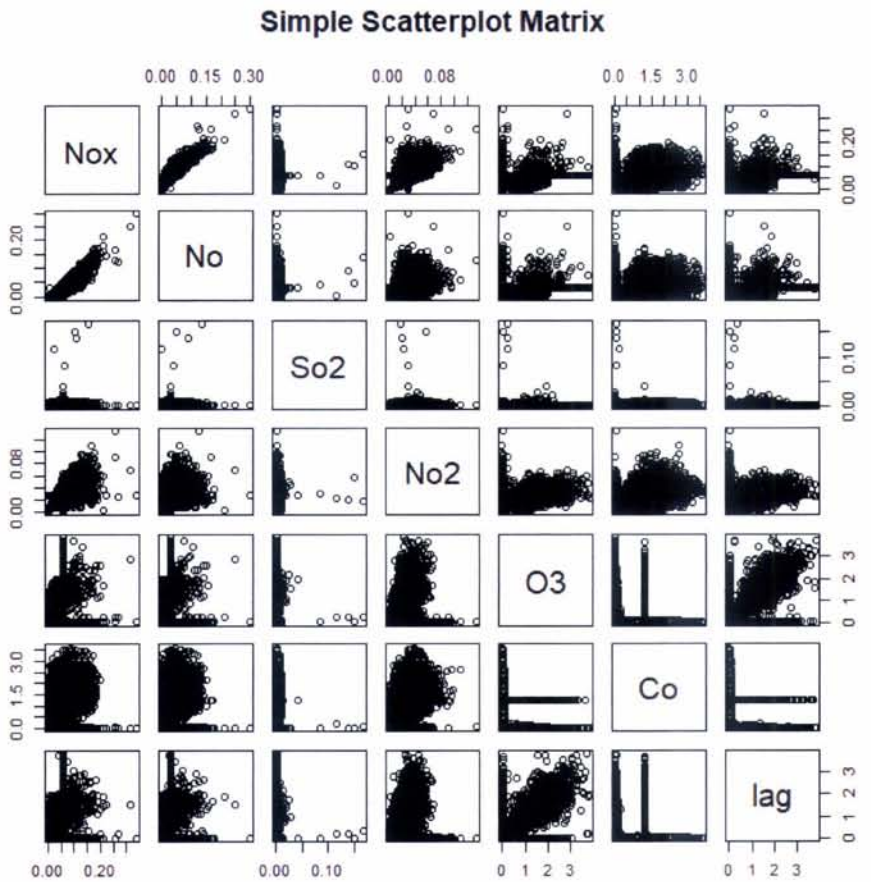


Figure 4.5 Scatter Plot of Ozone and Its Precursors in Petaling Jaya

From Figure 4.5, the scatter plot between O_3 and NO_X seems to show a positive linear relationship. This pattern is similar with the relationship between O_3 and NO. This is because both precursors which are NO_X and NO tend to have same criteria and pattern since the beginning of this study. While for SO_2 and CO, both have no clear pattern of relationship towards the O_3 concentration, thus there may be no linear relationship exists between these two precursors with the O_3 concentration. Other than that, the scatter plot between NO_2 and O_3 shows drastic uphill pattern of distribution.

This may indicate that there is exists a linear relationship between the precursor NO_2 and O_3 concentration.

b) Shah Alam

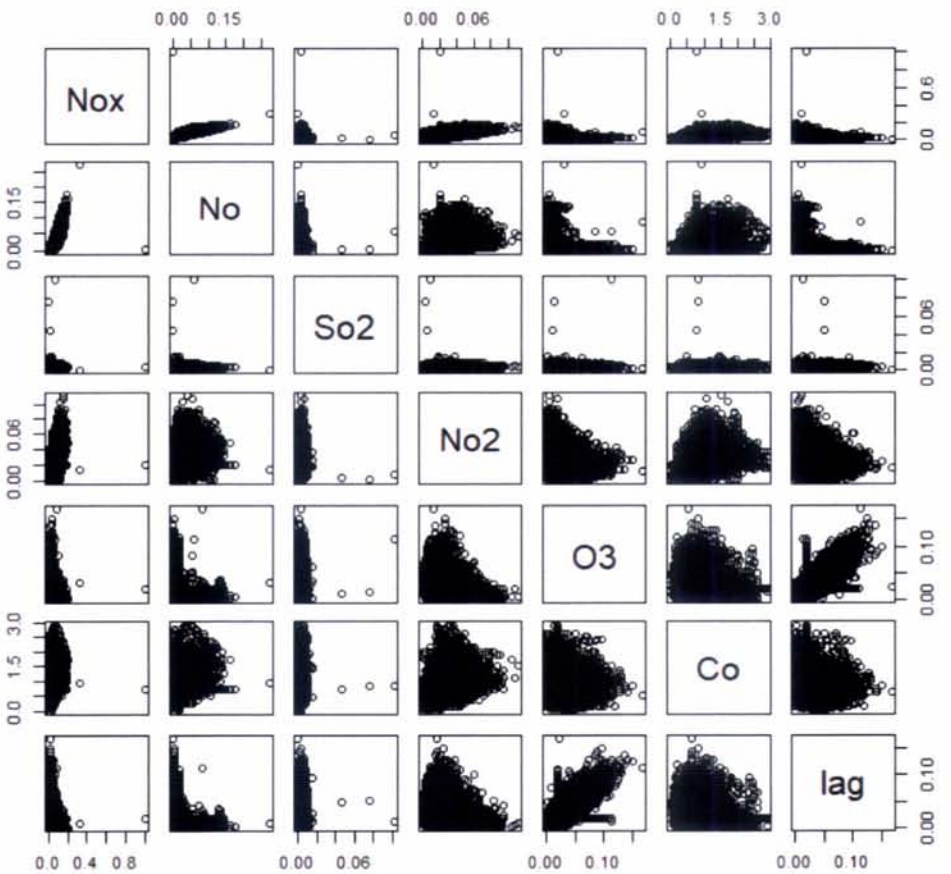


Figure 4.6 Scatter Plot of Ozone and Its Precursors in Shah Alam

From Figure 4.6, the scatter plot between O_3 and NO_2 seems to show a positive linear relationship. This pattern is similar with the relationship between O_3 and CO. While for NO_X and NO both have a correlation with each other since they have a positive linear relationship. Same goes to NO_X and NO_2, they also have a correlation. This is because the precursors NO_X, NO and NO_2 tend to have same criteria and pattern since the beginning of this study. For SO_2 and O_3 they have no clear pattern of relationship towards the O_3 concentration, thus there may be no linear relationship exists between these three precursors with the O_3 concentration. Other than that, the scatter plot between NO_2 and O_3 shows drastic uphill pattern of distribution. This may indicate that there is exists a linear relationship between the precursor NO_2 and O_3 concentration.

4.2.5 General Pattern of Ozone and Its Precursors Based on Monthly Average for 3 Years (2015-2017)

To study the general pattern of O_3 and its precursors at both study locations which are Petaling Jaya and Shah Alam, the average values of O_3 and the precursors for 3 years are considered.

a) Petaling Jaya

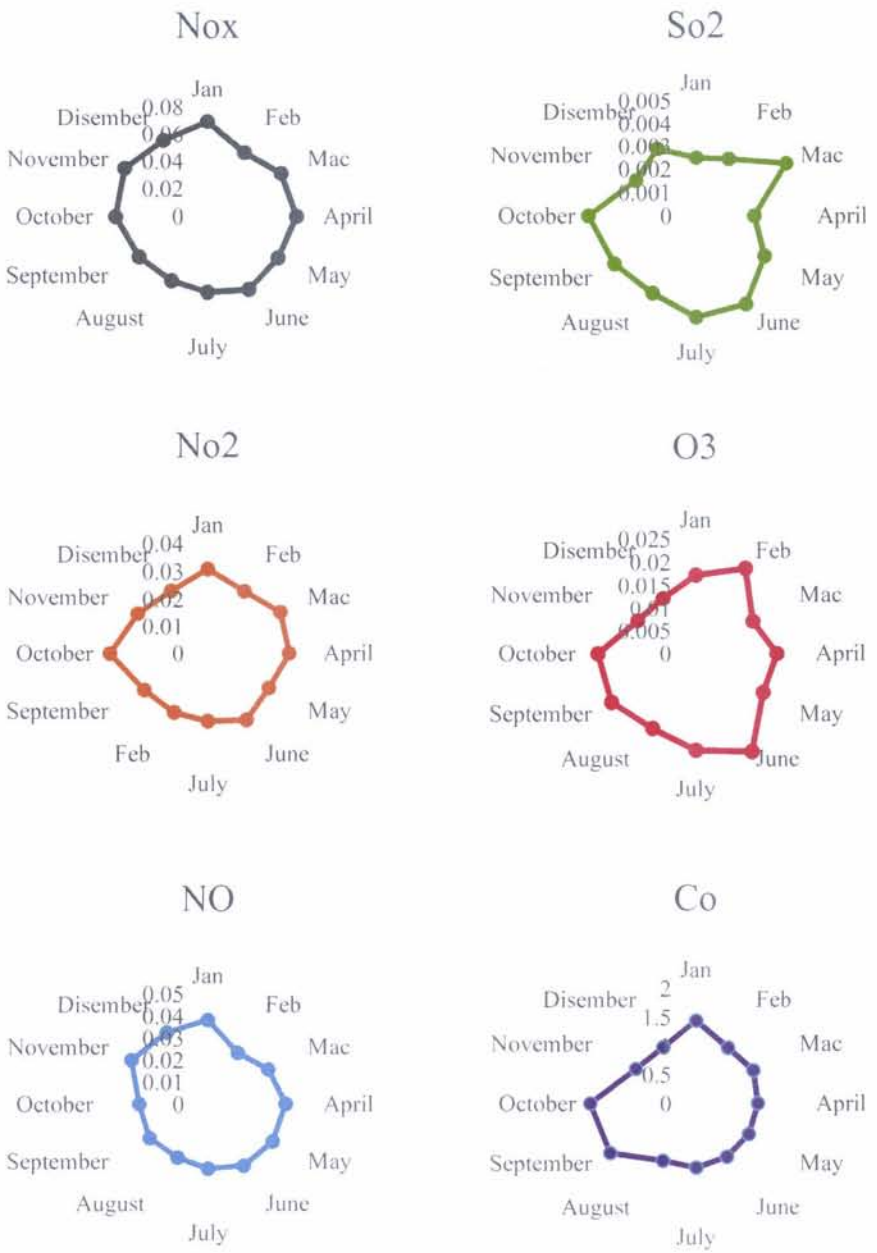
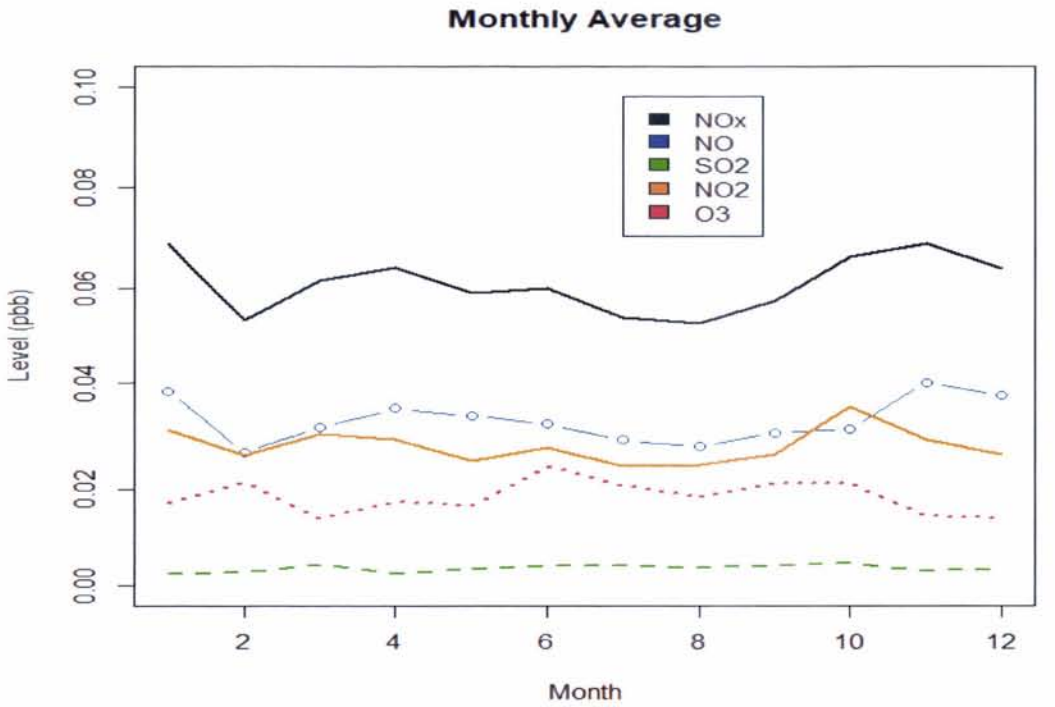


Figure 4.7 Monthly Average Reading of O_3 and its Precursors in Petaling Jaya

Based on Figure 4.7 above, NO_X has high concentrations within the end of the year until January. For the other months, it is at moderate level. A similar cycle is recorded for NO concentration in Petaling Jaya. However, for NO_2 concentration it reaches its highest peak in October and starts to decline until December before it is rise in January. This rise and down cycle are alternate until June before it is started to stable in moderate level. For SO_2, the behaviour is quite extreme because it can be seen clearly that it has high concentration in March and October while the lowest levels are in January and April. Unlike the precursors whom have low level in February, the O_3 reaches its highest level in February and June. While the low level of O_3 concentration is recorded in December and Mac. For CO, start from February until August the level for CO level remains at moderate level and start to rise to its peak in October and decline moderately until December.

(a)



(b)

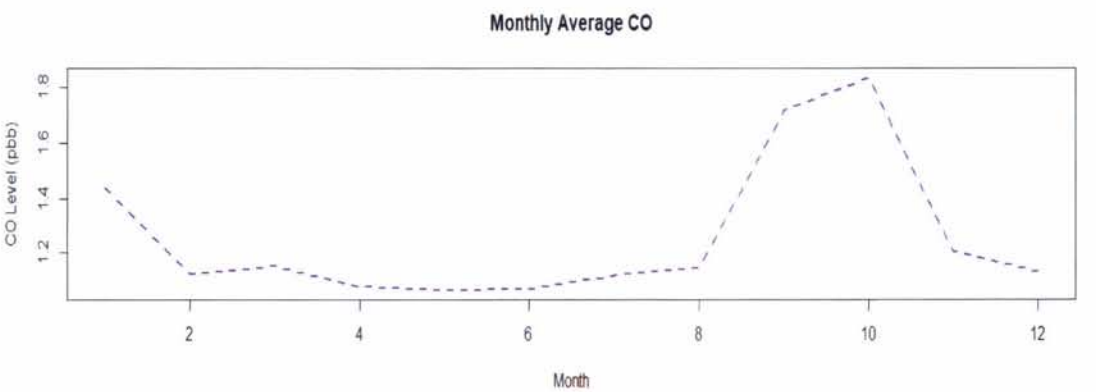
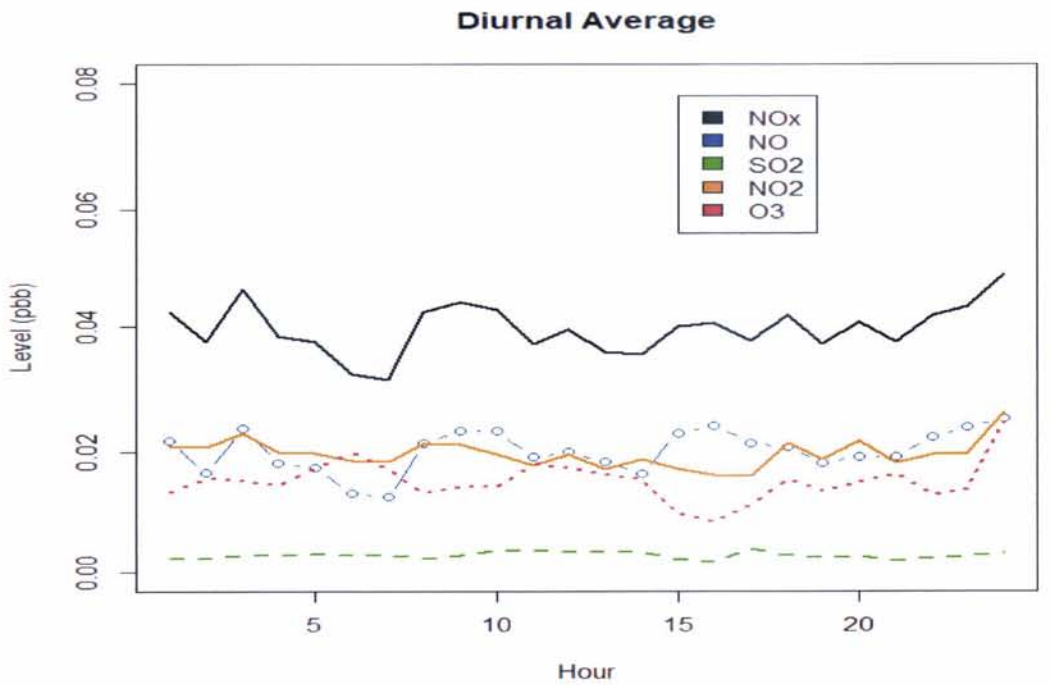


Figure 4.8 (a) Monthly Average level (Pbb) of NO_X, NO, SO_2, NO_2 and O_3 and (b) Monthly average level (Pbb) of CO in Petaling Jaya

The summarization for the Figure 4.8 (a) above, the O_3 level has shown some relationship with its precursors. When NO_X, NO and NO_2 are at their lowest concentration in February, O_3 shows itself by reaching its highest level. This is similar with the expectation that O_3 has negative correlation with NO_X, NO and NO_2. According to Nidhi Verma et. al, (2015), O_3 shows negative correlation with NO2 and NO. Therefore, the rises in the level of O_3 concentration is associated with a drop level of NO_2 and NO this is due to the photochemical activity and reaction with other chemical gases (i.e. the titrate effect).

From Figure 4.8 (a), SO_2 has the lowest range compared to other precursors and the highest level goes to CO that shown in figure 4.8 (b). For NO_X, NO and NO_2 they seem to have similar pattern of concentration throughout the monthly average for 3 years. When the precursors level for NO and NO_X increases, the level of O_3 tends to decrease. It is shown that most of the months have the maximum level of O_3 less than 0.03 which indicates that the level of O_3 is still under control which is lower than the Malaysian standard of 0.1 pbb. However, the exposure of O_3 to the environment is still health risk and harmful.

(a)



(b)

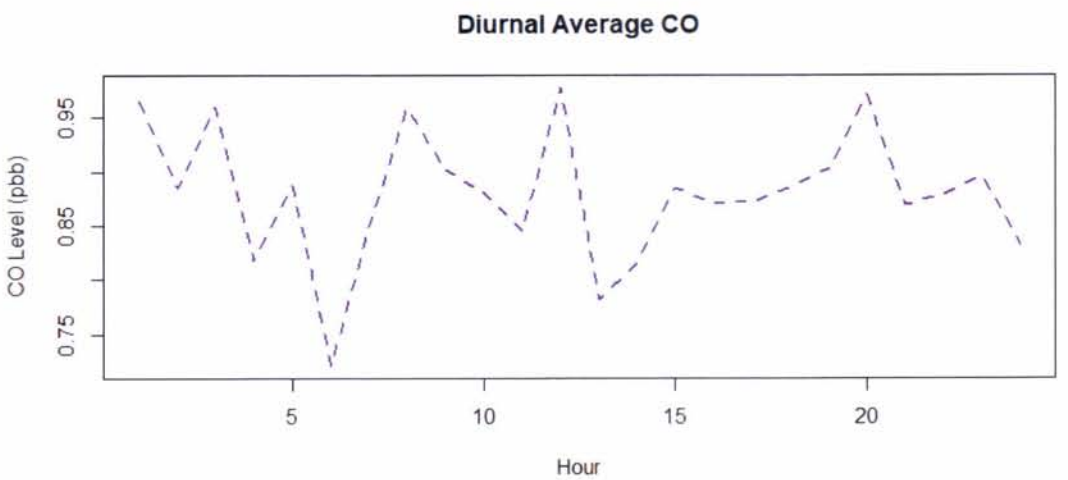


Figure 4.9 (a) Diurnal Average Level (pbb/h) for NO_X, NO, SO_2, NO_2 and O_3 and (b) Diurnal Average Level (pbb/h) for CO in Petaling Jaya

Figure 4.9 shows the diurnal average for O_3 and its precursor in the rate change of per hour (pbb/hour). Generally, over Petaling Jaya region SO_2 concentration is quite flat compared to other precursors. However, for NO_X and NO both tends to have similar pattern. Every time NO_X and NO reach their high peak, O_3 concentration tend to become lower. O_3 concentration reaches its high amount of concentration at dawn, afternoon and middle of the night which close to 1am. This is because at dawn and middle of the night, the existence of vehicles on the road are lesser compared at other hours. This makes lesser NO_X and NO emissions which cause the O_3 level increases. At afternoon, O_3 level increases because the O_3 needs sunlight for photochemical reactions, it reaches a maximum during daytime then slowly to decrease and fluctuates again (Nur Izzah et. al ,2018).

b) Shah Alam

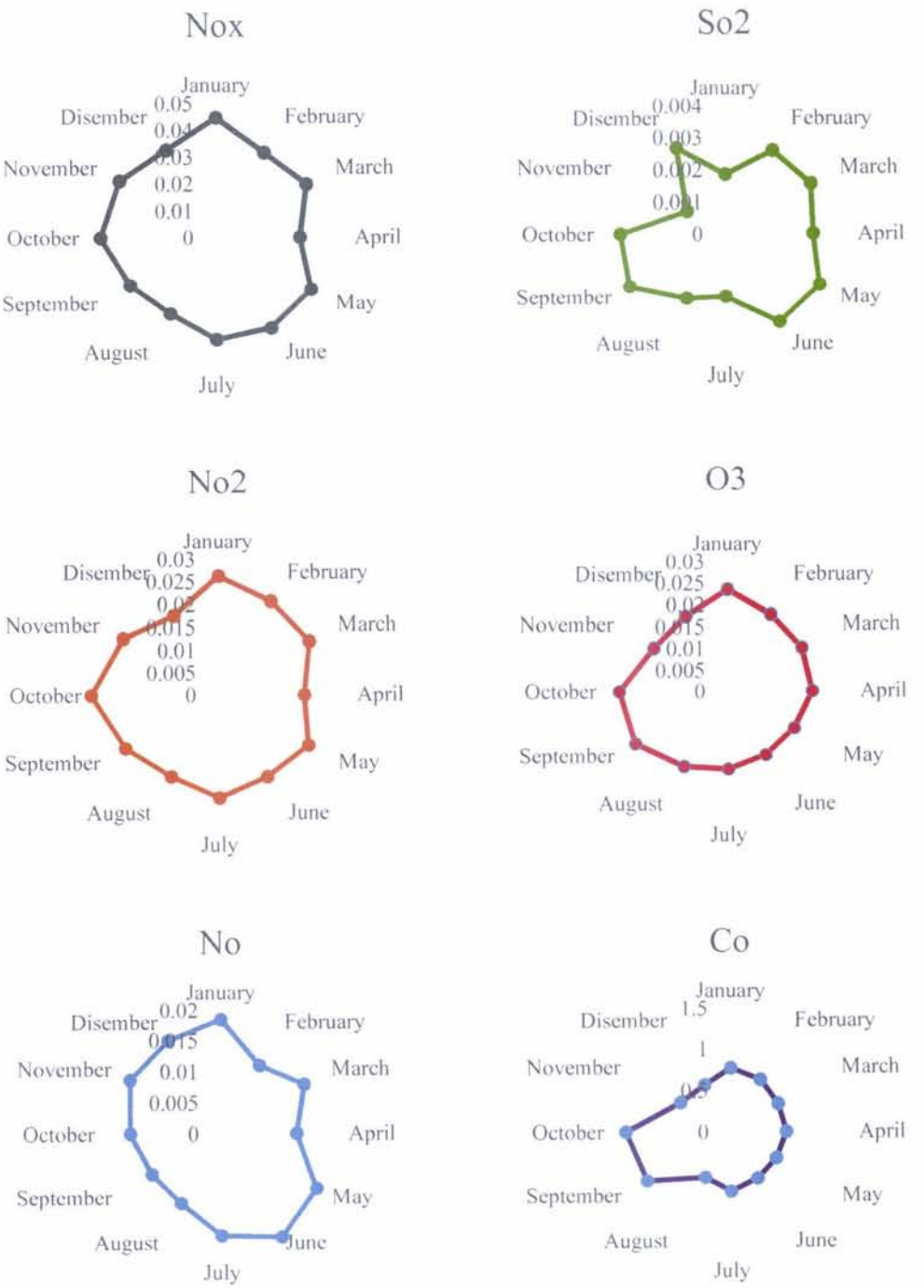
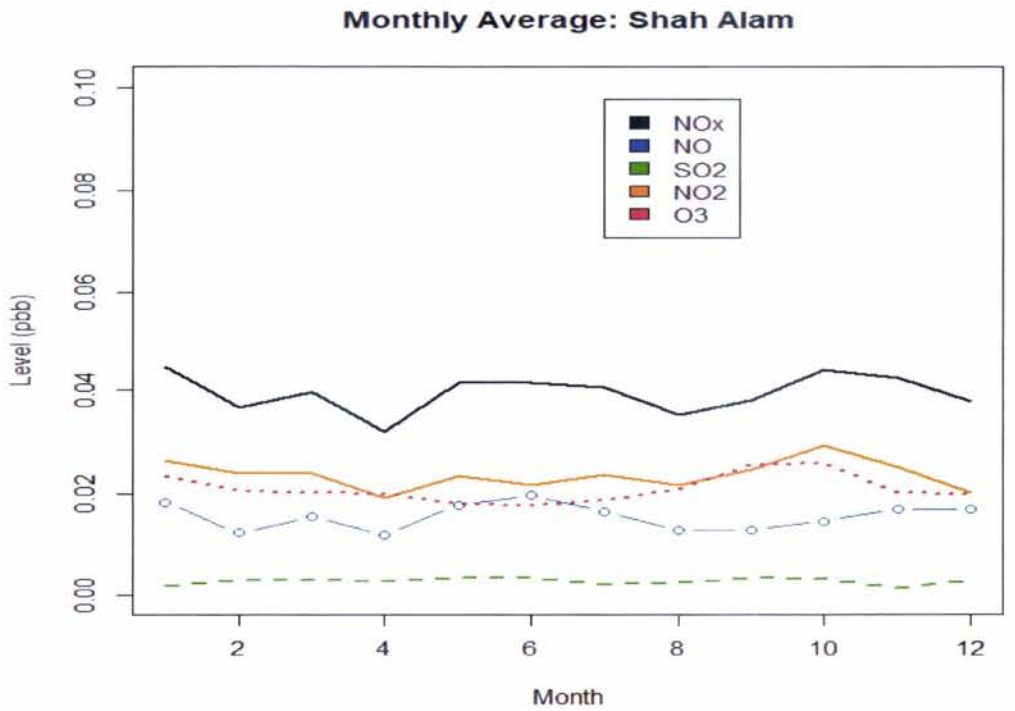


Figure 4.10 Monthly Average Reading of O_3 and its Precursors in Shah Alam

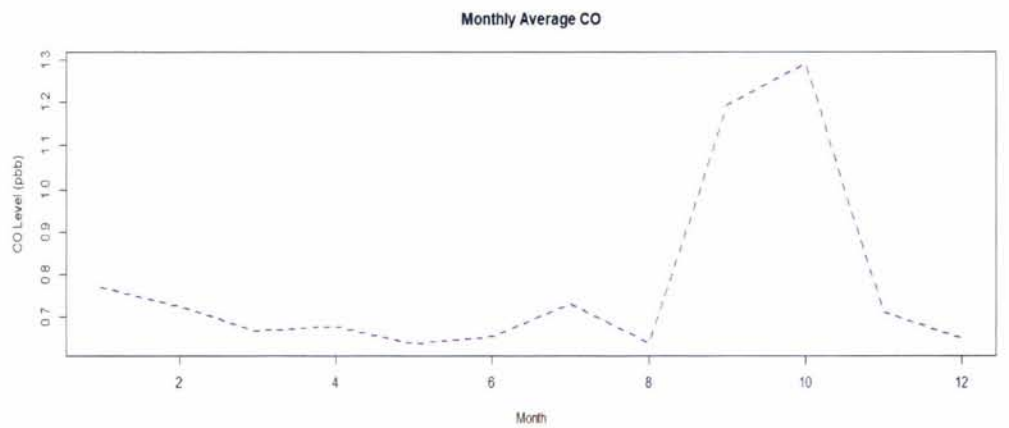
Referring to Figure 4.10, NO_X recorded moderate level of concentration for almost every month. There is only slight lower concentration recorded in April and December. It is also quite similar cycle recorded for NO_2 concentration in Shah Alam. A bit different for NO, the concentration level reaches its highest peak in January. There are fluctuations starting from the decline in February until NO concentration level rise again in May. After the rise and down cycle, the NO reading starts to become stable in moderate concentration level. For SO_2, it can be seen clearly that the behaviour is quite extreme when it has sudden decline and rise of SO_2 concentration level. The reading is low in July before it starts to rise in August to September then become stable until October. SO_2 concentration level having a sudden decline in November and suddenly recorded high concentration in December then the concentration becomes low again in January. Compared to its precursor, O_3 concentration level is stable which is in moderate level from January until August then the reading is slightly increase in September then stable until October. The concentration of O_3 then slightly decreases from November to December. For CO, from January until July the concentration level remains moderate and starts to decrease in August before suddenly recorded high concentration level in September and October then decreases in November and December.

To summarize the figure above, O_3 level does shows some relationship with its precursors. This obviously can be seen in NO_X, NO and NO_2 when low concentration recorded in these three precursors which is in February and April, the reading of concentration level of O_3 is high.

(a)



(b)



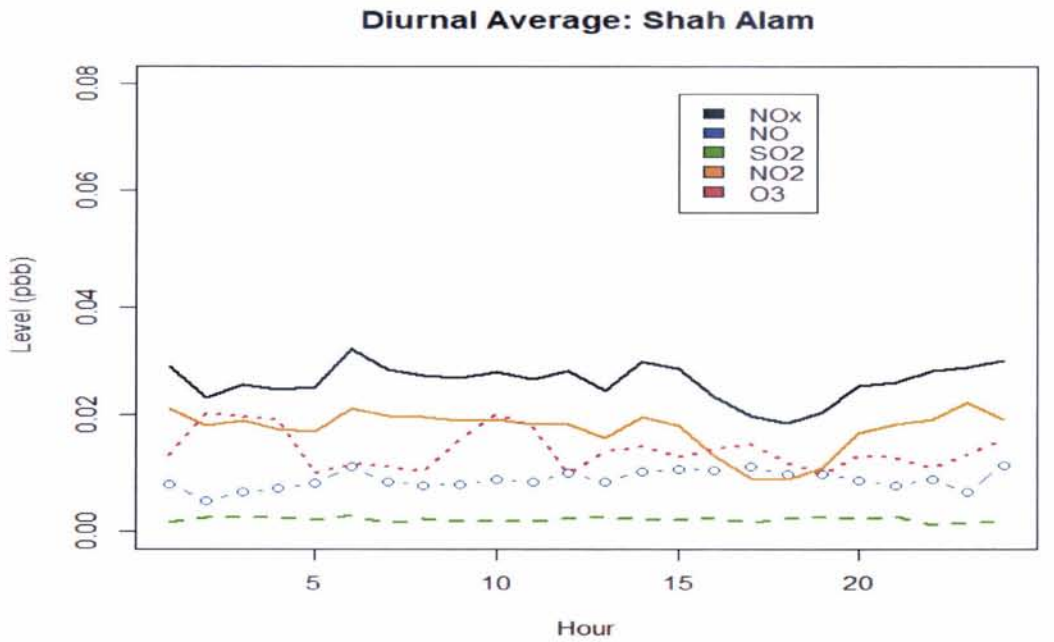
(b)

Figure 4.11 (a) Monthly Average level (Pbb) of NO_x, NO, SO_2, NO_2 and O_3 and (b) Monthly Average level (Pbb) of CO

Figure 4.11 shows SO_2 has the smallest range compared to other precursors while CO has the widest range compared to other precursors. The highest peak in O_3 occurred in October. The concentration level of O_3 decreases from April until August while the concentration of NO_X increases at the same period. This shows that O_3 and NO_X have a negative relationship. Both NO_X and NO have the same pattern. There are same fluctuations since January until May. SO_2 has the same trend with O_3 when both

variables rise and decline at the same period. This shows that O_3 and SO_2 have a positive relationship.

(a)



(b)

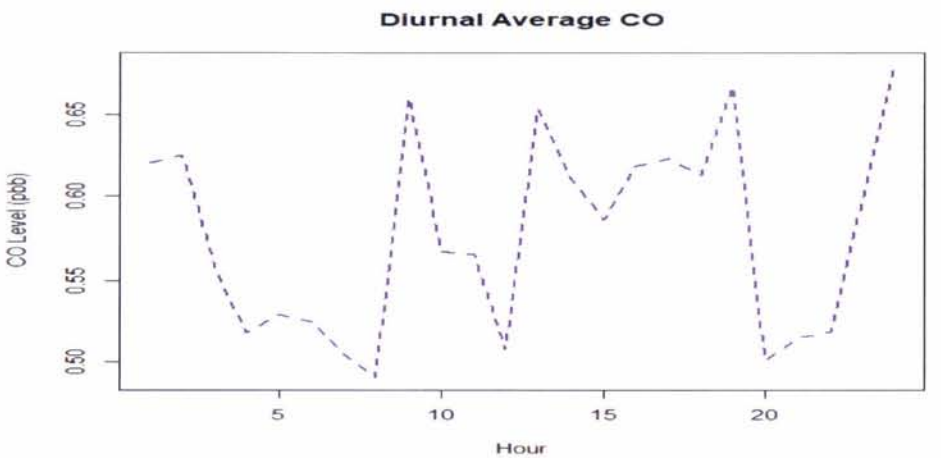


Figure 4.12 (a) Diurnal Average Level (pbb/h) for NO_X, NO, SO_2, NO_2 and O_3 and (b) Diurnal Average Level (pbb/h) for CO

Figure 4.12 shows the diurnal average for O_3 and its precursor in the rate change of per hour(pbb/hour). Form the figure, NO and NO_X have the same pattern. The concentration of SO_2 is quite low compared to another precursor in Shah Alam region. When NO and NO_X reach their high peak, the O_3 concentration tend to become lower. O_3 concentration seems to reach its high concentration at dawn, afternoon and midnight. This is because at dawn and midnight, the number of vehicles on the road is lesser compared to other hours. It can be said that with a lesser emission of NO and NO_X cause the O_3 level increases. At afternoon, O_3 level increases because the O_3 needs sunlight for photochemical reactions. Izzah et. al, (2017) stated that O_3 reaches its maximum during daytime then slowly to decrease and fluctuates again.

4.3 Statistical Model Building

In this section, there are two models of MLR that were going to be built. The first model (Model 1) is the model where the outliers are discarded and data randomization was conducted while the other model (Model 2), these two steps of procedures were excluded. The performance of Model 1 is assessed and will be validated with comparison with Model 2 performance and will be discussed later.

4.3.1 Randomization of Time Series Data

In order to build a significant and useful model, the errors terms over time need to be free from autocorrelation problems. This is because when the error terms are correlates with one another, there are some problems will arise such as the standard error of the regression coefficients may seriously underestimate the true standard deviation of estimated regression coefficients. For our data, randomization of the data needs to be done because the value of the current O_3 is depends on the value of O_3 a day or an hour before. Hence, randomization of the data is used in both study locations for Model 1 to reduce the autocorrelation effect.

4.3.2 Multicollinearity Checking for the Precursors Variables

Multicollinearity exists when the independent variables have high correlation among one another. This can be seen through the value of Pearson correlation coefficient (r value). If r value is more than 0.8, then that indicates that the two independent variables are correlate. If that case happened, then one of the variables needs to be removed.

a) Petaling Jaya

Table 4.3 shows the correlation matrix for the industrial area. The result shows the precursors for NO_X and NO have high collinearity among them which is more than 0.8. Thus, one of the variables need to be dropped and that is NO. This is because NO is considered as not to be hazardous to health at typical ambient concentrations as compared to NO_X (IcopalUK, 2015).

Table 4.3
Correlation Matrix for Petaling Jaya

Variables	NO_X	NO	SO_2	NO_2	O_3	CO	Lag
NO_X	1.0000	0.9405	-0.0267	0.6299	-0.2621	0.2118	-0.2073
NO	0.9405	1.0000	-0.0328	0.4419	-0.3323	0.1994	-0.2869
SO_2	-0.0267	-0.0328	1.0000	0.0709	-0.2700	0.3235	-0.2740
NO_2	0.6299	0.4419	0.0709	1.0000	-0.0785	0.2188	-0.0298
O_3	-0.2621	-0.3323	-0.2700	-0.0785	1.0000	-0.4847	0.8960
CO	0.2118	0.1994	0.3235	0.2188	-0.4847	1.0000	-0.5236
Lag	-0.2073	-0.2869	-0.2740	-0.0298	0.8960	-0.5236	1.0000

b) Shah Alam

Table 4.4 shows the correlation matrix table for Shah Alam. The results show that precursor that has high collinearity are NO and NO_X and also NO_2 and NO_X. Thus, we have decided to exclude NO and NO_2 in the modelling process. This is because NO is considered as not to be hazardous to health at typical ambient concentrations as compared to NO_X (IcopalUK, 2015). NO_2 is also can be exclude because the correlation is quite strong.

Table 4.4
Correlation Matrix for Shah Alam

	NO_X	NO	SO_2	NO_2	O_3	CO	Lag
NO_X	1.00000	0.85357	0.06958	0.73310	-0.46601	0.37574	-0.45917
NO	0.85357	1.00000	0.00713	0.42965	-0.47939	0.35104	-0.50355
SO_2	0.06958	0.00713	1.00000	0.15479	0.05066	0.08629	0.04503
NO_2	0.73310	0.42965	0.15479	1.00000	-0.35511	0.35305	-0.30848
O_3	-0.46601	-0.47939	0.05066	-0.35511	1.00000	-0.23508	0.85354
CO	0.37574	0.35104	0.08629	0.35305	-0.23508	1.00000	-0.29806
Lag	-0.45917	-0.50355	0.04503	-0.30848	0.85354	-0.29806	1.00000

4.3.3 Outlier Removing

Outliers are extreme values that fall far away from others or common observations. As shown in histogram and box plot figures in earlier analysis, the outliers are existing for the O_3 and its precursors. When the outliers are said to exist, the normality of the data might get interrupted. For this study, outlying values analysis is used to filter out the outliers from the data. This is guided by the results of boxplot analysis. The model is assessed until the best performance of data is obtained. This method is applied on both study locations, Petaling Jaya and Shah Alam from both X's and Y direction.

4.3.4 Model Establishment

To develop the first model (Model 1), the dataset that have been cleared from the outliers is used. This is to make sure that the information obtained about understanding the association of the precursors in the formation of O_3 level is significant. Moreover, lag variable that is developed from the O_3 values is also included in this model to increase the efficiency and the relationship between the precursors and the O_3. In addition, when doing the multicollinearity checking, NO is removed from the entry to any model in Petaling Jaya while NO and NO_2 are removed in Shah Alam. This is because it has high collinearity as shown in correlation matrix for both locations.

The second model (Model 2) is going to be obtained by using the same criteria variables as in the first model (Model 1). The only thing that distinguishes these two models is Model 2 does not discard any outliers or do any treatments to their outliers and the regression was conducted using non-randomized data set.

4.3.5 Diagnostic Checking for Both Models

This section may be an exploration of the model underlying the statistical assumptions. This is to ensure that the model adequately describe about the O_3 association with its precursors as well as the influence in the O_3 concentration.

4.3.5.1 Model Significance

a) Petaling Jaya

i) Model 1 : Outliers are discarded in both X and Y directions and data are randomized.

Table 4.5 below shows the p-value for Model 1 in Petaling Jaya. The value of α used is 0.05 while the decision rule is rejecting H_o if p-value < α. Since p-value = 2.2e-16 < α = 0.05, H_0 is rejected. At α = 0.05, we conclude that the model is significant.

H_0: The model is not significant

H_1: The model is significant

Table 4.5
P-value for Model 1 in Petaling Jaya

P- value of model:	2.2e-16

Table 4.6 shows the coefficient table for Petaling Jaya. This formula is used to run a coefficient table in R software where O_3 acts a dependent variable while NO_X, SO_2, NO_2, CO and lag act as the independent variables.

Table 4.6
Coefficient Table for Model 1 in Petaling Jaya

	Estimate	Std. Error	t-value	P-value
(Intercept)	0.0005793	0.0004054	1.429	0.153
NO_X	-0.0705566	0.0077352	-9.122	2e-16
SO_2	0.4755807	0.0846523	5.618	2.04e-08
NO_2	0.0870808	0.0209760	4.151	3.36e-05
CO	0.0029999	0.0003489	8.599	2e-16
Lag	0.8106473	0.0084366	96.086	2e-16

Based on Table 4.6, the important information is taken out from the table to decide either the variable have a significant linear relationship with the O_3 or not. The result for summarization is shown in Table 4.7 below. It shows that all precursors and the Lag variable are significant to the model. This is because the p-value for all the variables is less than α which is 0.05. Hence, H_0 is rejected for each variable and conclude that the variable has a significant linear relationship with the O_3.

H_0: The variable does not have a significant linear relationship with O_3

H_1: The variable has a significant linear relationship with O_3

Table 4.7
Summarization Table of Variable Significances in Model 1 in Petaling Jaya

Variable	P-value	Decision	Conclusion
NO_X	2e-16	Reject H_0	Significant
SO_2	2.04e-08	Reject H_0	Significant
NO_2	3.36e-05	Reject H_0	Significant
CO	2e-16	Reject H_0	Significant
Lag	2e-16	Reject H_0	Significant

ii) Model 2 : Outliers are not discarded in both X and Y direction and data set was not randomized

Table 4.8 below shows the p-value for Model 2 in Petaling Jaya. The value of α used is 0.05 while the decision rule is rejecting H_0 if p-value $<\alpha$. Since p-value = 2.2e-16 $< \alpha = 0.05$, H_o is rejected. At $\alpha = 0.05$, we conclude that the model is significant.

H_0: The model is not significant

H_1: The model is significant

Table 4.8
P-value for Model 2 in Petaling Jaya

P- value of model:	2.2e-16

Table 4.9 shows coefficient table for Petaling Jaya the formula used to run a coefficient table R software is O_3 acts a dependent variable while NO_X, SO_2, NO_2, CO and lag act as the independent variables.

Table 4.9
Coefficient Table for Model 2 in Petaling Jaya

	Estimate	Std. Error	t-value	P-value
(Intercept)	0.161635	0.0142733	11.325	2e-16
NO_X	-0.309788	0.187538	-1.652	0.098624
SO_2	-6.489043	1.751437	-3.705	0.000214
NO_2	2.300577	0.513913	4.477	7.75e-06
CO	-0.124555	0.008925	-13.956	2e-16
Lag	0.749708	0.009267	80.900	2e-16

Based on Table 4.9, important information is taken out from the table to decide either the variable have a significant linear relationship with the O_3 or not. The result for summarization is shown in Table 4.10 below. It shows that all the variables are significant to the model except for precursor NO_X. This is because the p-value for all variables is less than α which is 0.05. Hence, H_0 is rejected for those variables that are significant and conclude that the variable has a significant linear relationship with the O_3. For precursor NO_X, H_0 is accepted as it has no significant linear relationship with the O_3.

H_0: The variable does not have a significant linear relationship with O_3

H_1: The variable has a significant linear relationship with O_3

Table 4.10
Summarization Table of Variable Significances in Model 2 in Petaling Jaya

Variable	P-value	Decision	Conclusion
NO_X	0.098624	Do not reject H_o	Not significant
SO_2	0.000214	Reject H_o	Significant
NO_2	7.75e-06	Reject H_o	Significant
CO	2e-16	Reject H_o	Significant
Lag	2e-16	Reject H_o	Significant

b) Shah Alam

i) Model 1 : Outliers are discarded in both X and Y directions and data are randomized.

Table 4.11 below shows the p-value for Model 1 in Shah Alam. The value of α used is 0.05 while the decision rule is rejecting H_0 if p-value < α. Since p-value = 2.2e-16 < $\alpha = 0.05$, H_0 is rejected. At $\alpha = 0.05$, we conclude that the model is significant.

H_0: The model is not significant

H_1: The model is significant

Table 4.11
P-value for Model 1 in Shah Alam

P-value of model	2.2e-16

Table 4.12 shows coefficient table for Shah Alam. The formula used to run the coefficient table in R software is O_3 acts as an independent variable while NO_X, SO_2, CO and lag act as the independent variables.

Table 4.12
Coefficient Table for Model 1 in Shah Alam

	Estimate	Std. Error	t-value	P-value
Intercept	0.00382	0.00036	10.555	2e-16
NO_X	-0.05832	0.00719	-8.108	6.42e-16
SO_2	-0.10838	0.08619	-1.257	0.209
CO	0.00266	0.00040	6.575	5.36e-11
Lag	0.73103	0.00988	73.997	2e-16

Based on Table 4.12, to decide either the variable have a significant linear relationship with the O_3 or not, important information is taken out from the table. The summarization of the result is shown in the Table 4.13 below. It shows that all precursors and the lag variable are significant except for SO_2. This is because the p-value for all variables is less than $\alpha = 0.05$. Thus, H_0 is rejected for those variables that are significant and can be conclude that the variables have a significant linear relationship with the O_3. For precursor SO_2, H_0 is accepted thus it has no significant linear relationship with the O_3.

H_0: The variable does not have a significant linear relationship with O_3

H_1: The variable has a significant linear relationship with O_3

Table 4.13
Summarization Table of Variable Significance in Model 1 in Shah Alam

Variable	P-value	Decision	Conclusion
NO_X	6.42e-16	Reject H_o	Significant
SO_2	0.209	Do not reject H_o	Not Significant
CO	5.36e-11	Reject H_o	Significant
Lag	2e-16	Reject H_o	Significant

ii) Model 2 : Outliers are not discarded in both X and Y directions and data are not randomized

Table 4.14 below shows that p-value for Model 2 in Shah Alam. The value of α used is 0.05 while the decision rule is rejecting H_0 if p-value < α. Since p-value = 2.2e-16 < α = 0.05, H_0 is rejected. At α = 0.05, we conclude that the model is significant.

H_0 : The model is not significant

H_1 : The model is significant

Table 4.14
P-value for Model 2 in Shah Alam

P-value of model	2.2e-16

Table 4.15 shows the coefficient table for Shah Alam. The formula used to run a coefficient table in R software is O_3 acts as an independent variable while NO_X, SO_2, CO and lag act as the independent variables.

Table 4.15
Coefficient Table for Model 2 in Shah Alam

	Estimate	Std. Error	t-value	P-value
Intercept	0.00314	0.00019	15.837	2e-16
NO_X	-0.0648	0.00271	-23.878	2e-16
SO_2	0.17383	0.03355	5.181	2.22e-07
CO	0.00264	0.00017	15.360	2e-16
Lag	0.84543	0.00329	256.255	2e-16

Based on Table 4.15, to decide either the variable have a significant linear relationship with the O_3 or not, important information is taken out from the table. The summarization of the result is shown in the Table 4.16 below. It shows that all precursors and the Lag variable are significant except for SO_2. This is because the p-value for all variables is less than $\alpha = 0.05$. Thus, H_0 is rejected for those variables that are significant and can be conclude that the variables have a significant linear relationship with the O_3. For precursor SO_2, H_0 is accepted thus it has no significant linear relationship with the O_3.

H_0: The variable does not have a significant linear relationship with O_3

H_1: The variable has a significant linear relationship with O_3

Table 4.16
Summarization Table of Variable Significance in Model 2 in Shah Alam

Variable	P-value	Decision	Conclusion
NO_X	2e-16	Reject H_o	Significant
SO_2	2e-16	Reject H_o	Significant
CO	2.22e-07	Reject H_o	Significant
Lag	2e-16	Reject H_o	Significant

4.3.5.2 Model Capability Measures

To measure how many percentages of the independent variable variation explained by a linear model, R-squared value is used. In general, the higher the value of R-squared, the better the model fits the data. However, in this study the capability of the models is measured by the multiple R-squared. This is because this study implies multiple independent variables.

a) Petaling Jaya

Table 4.17 shows the value of multiple R-squared for both models in Petaling Jaya. There is a slight difference in the value of multiple R-squared between the two models where Model 2 has high value compared to Model 1. For Model 1, there is 69.8% variation in O_3 that is explained by the independent variables. While for Model 2, it has 72.1% of variation in O_3 that is explained by the independent variables.

Table 4.17
Multiple R-squared Values for Petaling Jaya

Model 1	Model 2
0.698	0.7212

b) Shah Alam

Table 4.18 shows the value of multiple R-squared for both models in Shah Alam. There is a slight difference in the value of multiple R-squared between the two models where Model 2 has a higher value compare to Model 1. For Model 1, there is 54.39% variation in O_3 that is explained by the independent variables. For Model 2, it has 75.08% of variable in O_3 that is explained by the independent variables.

Table 4.18
Multiple R-squared Values for Shah Alam

Model 1	Model 2
0.5439	0.7508

4.3.5.3 I.I.D Assumptions for Residual Variables

I.I.D means that every residual is independent and identically distributed across the model. This is to make sure that the result obtained can represent the actual understanding about the formation of O_3.

a) Petaling Jaya

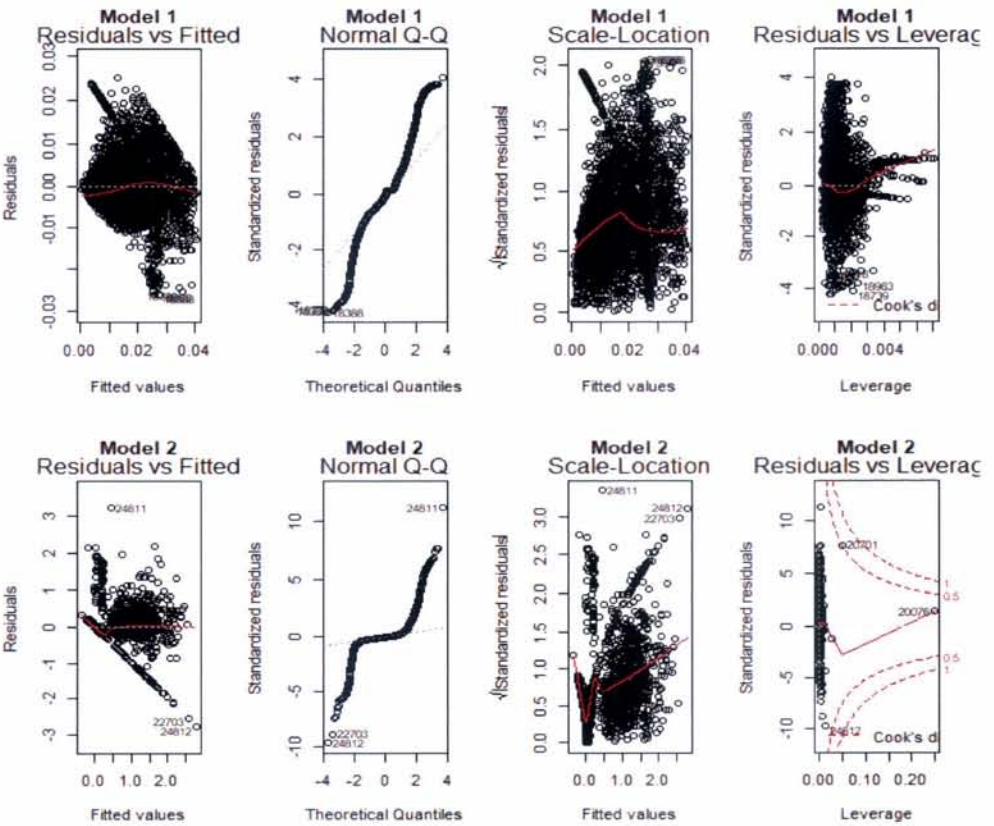


Figure 4.13 Residuals Plots for Petaling Jaya

i) Residual vs Fitted

The first column in Figure 4.13 shows the plot of residuals vs fitted. From this plot, non-linear patterns can be checked by looking at the pattern of the graph. Model 1 does not show any specific pattern in their graph. Moreover, the residuals are also equally spread around the horizontal line. This means Model 1 has a random pattern while Model 2 has non-random pattern.

ii) Normal Q-Q Plot

The second column in Figure 4.13 shows the Q-Q plot of the residuals. From this plot, normality of the data can be checked by looking at the points that lies close to the fitted line. For Model 1, all the points are lies and merge close to the fitted line including their tails. But for Model 2, its tails seem to be far from the fitted line and its pattern does not seem to follow normal distribution.

iii) Scale-Location

The third column in Figure 4.13 shows scale-location plot that is also called as spread-location plot. From this plot, equal variance which is homoscedasticity can be checked by looking at the scattered points of the residuals along the ranges of the independent variables. For Model 1, its horizontal line is surrounded equally by the spread points. While for Model 2, the residuals begin to spread wider along the x-axis as it is passes around 2. Since the residuals spread wider, the red line is not horizontal and shows a steep angle in Model 2.

iv) Residual vs Leverage

The fourth column in Figure 4.13 shows the residuals vs leverage plot. From this plot, influential outliers can be detected by looking for cases outside of a dashed line, Cook's distance. For Model 1, there is no influential case because all cases are well inside of the Cook's distance lines. However, for Model 2 there are few cases lies far beyond from the Cook's distance lines which indicates that they are influential outliers. This can be seen from reading the value of multiple R-squared for Model 1 and Model 2. Model 2 that has influential outliers has multiple R-squared of 0.7212 but for model 1 when there are no influential outliers the value is only 0.698. This shows that the influential outliers give an impact to the analysis of models.

b) Shah Alam

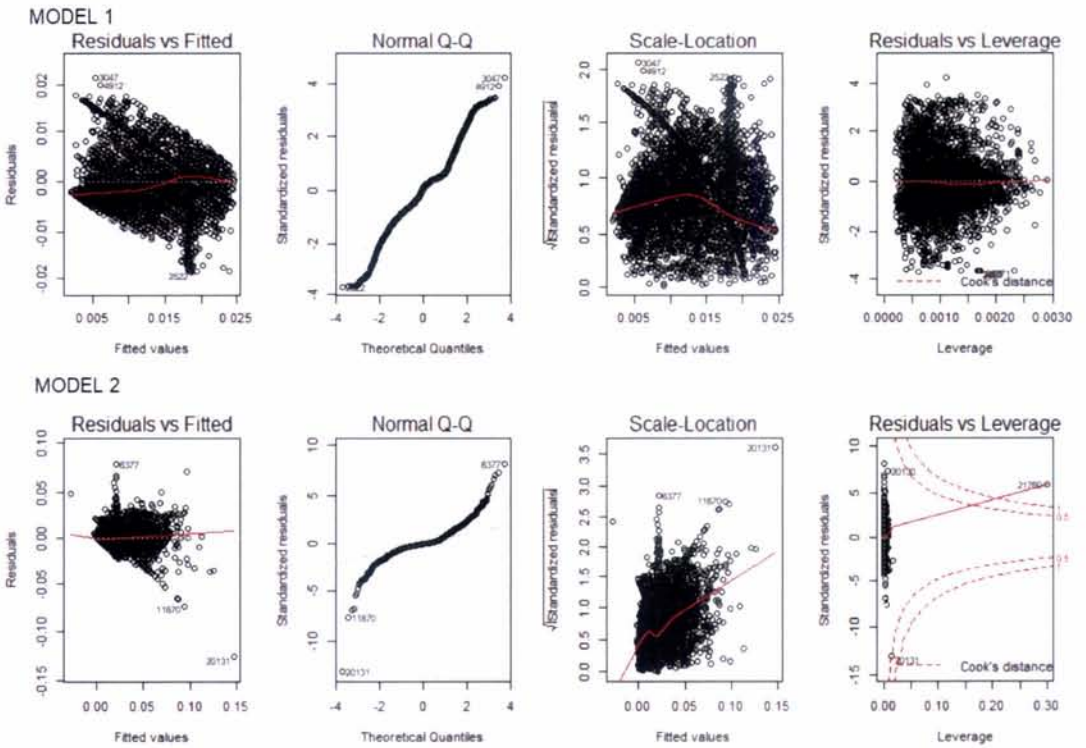


Figure 4.14 Residuals Plots for Shah Alam

i) Residual vs Fitted

In Figure 4.14, the first column shows the plots of residuals vs fitted. From this plots, non-linear pattern can be checked by looking at the pattern of the graph. Model 1 does not show any specific pattern in their graph. The residuals are also equally spread around the horizontal line. This means Model 1 has a random pattern while Model 2 has non-random pattern.

ii) Normal Q-Q

The second column in Figure 4.14 shows the Q-Q plot of the residuals. From this plot, normality of the data can be checked by looking at the points that lies close to the fitted line. For Model 1, all the points seem lies and merge close to the fitted line. Different for Model 2 where most of the point does lies and merge close to the fitted line but the tails seem to lie far from the fitted line and its pattern does not seem to follow normal distribution.

iii) Scale-Location

From the scale-location plot in Figure 4.14 the equal variance which is homoscedasticity can be checked by looking at the scattered points of the residuals along the ranges of independent variables of the ranges of the independent variables. For Model 1 its horizontal line is surrounded equally by the spread points. While for Model 2, the residuals begin to spread wider along the x-axis. Since the residuals spread wider, the red line is not horizontal with steep angle shown in Model 2.

iv) Residual vs Leverage

The last column in Figure 4.14 shows the residuals vs leverage plot. From this plot, influential outliers can be detected by looking for points that is outside of a dashed line, Cook's distance. For Model 1, there is no influential case because all cases are well inside the Cook's distance lines. However, there are few cases lies far from the Cook's distance lines in Model 2 which indicates that they are influential outliers. This can be seen from reading the value of multiple R-squared for Model 1 and Model 2. The value of multiple R-squared in Model 2 with outliers is 0.7508 but for Model 1 when there are no influential outliers the value is only 0.5439. This shows that the influential outliers give an impact to the analysis of model.

4.3.5.4 Model Validation Results

To verify that the models are performing as expected, model validation is performed. In this study, simple validation approach is used by plotting the observed O_3 with their fitted O_3. From the plot, the fitness of the data can be seen and interpret.

a) Petaling Jaya

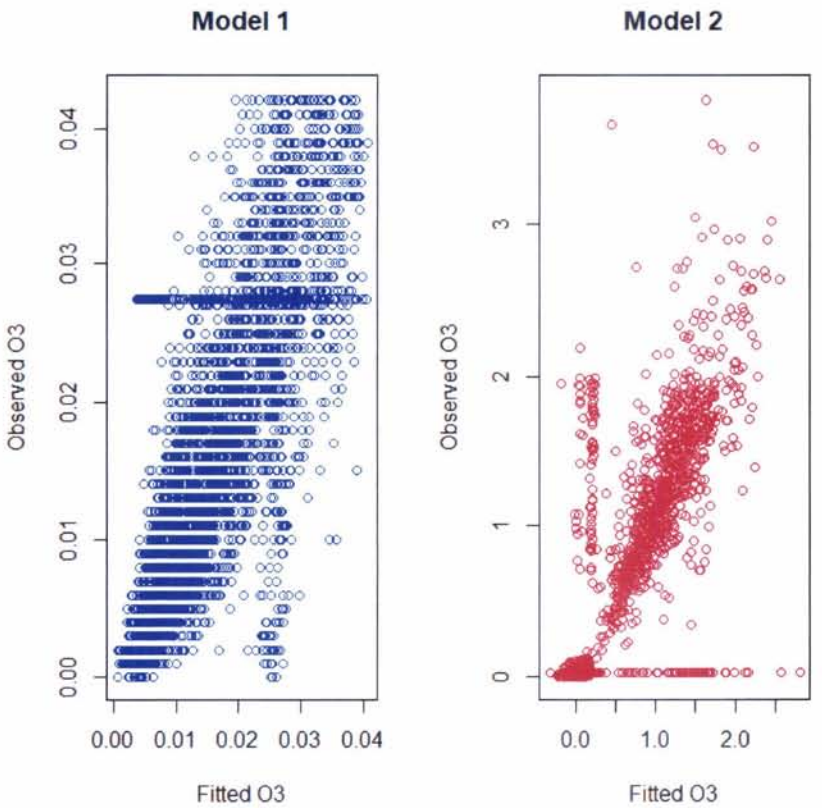


Figure 4.15 Model Validation Results for Both Models in Petaling Jaya

From Figure 4.15, Model 1 has better validation result compared to Model 2. This can be seen through the scattered plot of the model. Model 1 plots scattered more since the range values of observed and fitted O_3 is small which is in range 0 until 0.04. While for Model 2, their range is quite big which is from 0 until around 3. This makes their plots narrower and focused at one place.

b) Shah Alam

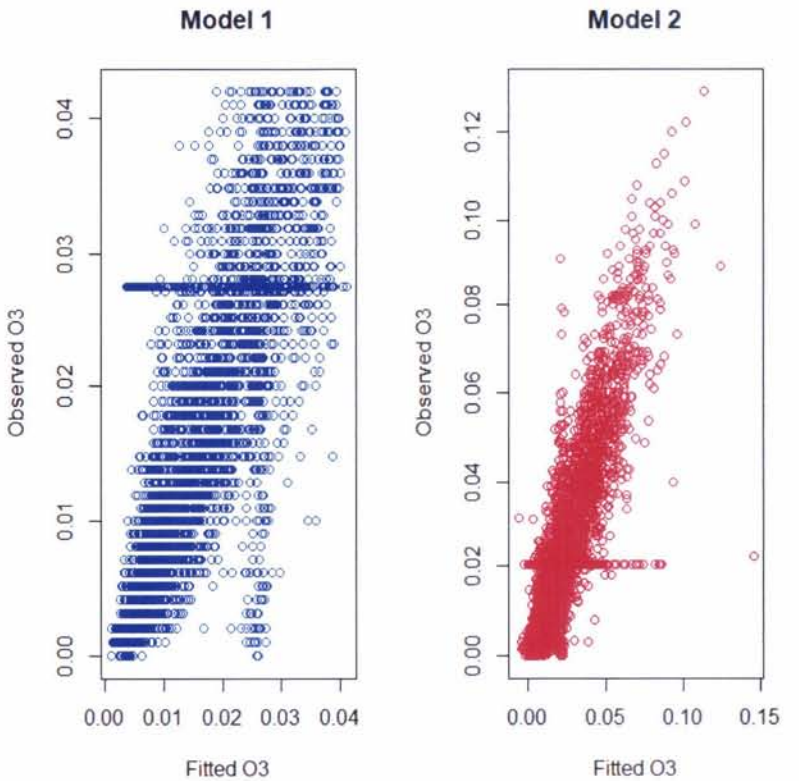


Figure 4.16 Model Validation Results for Both Models in Shah Alam

From Figure 4.16, Model 1 has better validation result compared to Model 2. This can be seen through the scattered plot of the model. Model 1 plots scattered more since the range values of observed and fitted O_3 is small which is in range 0 until 0.04. While for Model 2, their range is quite big which is from 0 until around 0.12. This makes their plots narrower and focused at one place

4.4 Model Comparisons and Findings

In this section, the performances of both models discussed earlier will be compared at both study locations which are Petaling Jaya and Shah Alam. Hence, from this comparison, the best model is going to be selected to understand the association between the precursors in the formation of O_3 in two locations in Selangor.

a) Petaling Jaya

Table 4.19 shows model comparison of performances for both models in Petaling Jaya. Based on the table below, Model 1 has better performance than Model 2. Since Model 1 follows all the assumptions, value of R-squared and residual standard error in Model 1 is better compared to Model 2. When the model does not meet any i.i.d assumptions, the results of the output could be misinterpreted and not accurate. Even though the multiple R-squared is high in Model 2. However, this model is shown not valid because the required assumptions for MLR are not satisfied. Other than that, accuracy of the model can be seen through the value of residual standard error for both models. Residual standard error for Model 2 is 0.7212 compared to model 1 is only 0.00628. These differences give a high impact to understand the association of O_3 and its precursors. Thus, the interpretation to understand the association between the O_3 and the precursors is based on Model 1.

Table 4.19
Model Comparison of Performances in Petaling Jaya

	Model 1	Model 2
Regression Equation	$O_3 = 0.0005793 - 0.0705566*NO_X + 0.4755807*SO_2 + 0.0870808*NO_2 + 0.0029999*CO + 0.8106473*Lag$	$O_3 = 0.161635 - 6.489043*SO_2 + 2.300577*NO_2 - 0.124555*CO + 0.749708*Lag$
F- statistics	Significant	Significant
Multiple R-squared	0.698	0.7212
Significant Independent Variable	$-NO_X$, $+SO_2$, $+NO_2$, + CO and +Lag	$-SO_2$, + NO_2,-CO and +Lag
Residual standard error	0.006228	0.2885
I.I.D Assumptions for residuals	**Satisfied**	**Does not satisfied**
1. Residual vs Fitted	1. Random pattern	1. Non-random pattern
2. Normal Q-Q Plot	2. Normally distributed	2. Not normally distributed
3. Scale-Location	3. Random pattern (homoscedasticity)	3. Non-random pattern (heteroscedasticity)
4. Residual vs Leverage	4. No influential outliers	4. Has influential outliers

b) Shah Alam

Table 4.20 shows model comparison of performances for both models in Shah Alam. Based on the table below, Model 1 has better performance than Model 2. Since Model 1 follows all the assumptions, value of R-squared and residual standard error in Model 1 is better compared to Model 2. When the model does not meet any i.i.d assumptions, the results of the output could be misinterpreting and not accurate. Even though the multiple R-squared is high in Model 2, this may be the result of influential outliers that have been discussed. Other than that, accuracy of the model can be seen through the value of residual standard error for both models. Residual standard error for Model 2 is 0.009714 compared to Model 1 is only 0.005064. These differences give a high impact to understand the association of O_3 and its precursors.

Table 4.20
Model Comparison of Performances in Shah Alam

	Model 1	Model 2
Regression Equation	$O_3 = 0.00382 - 0.05832*NO_X - 0.00266*CO + 0.73103*Lag$	$O_3 = 0.00314 - 0.0648*NO_X + 0.17383*SO_2 + 0.00264*CO + 0.84543*Lag$
F- statistics	Significant	Significant
Multiple R-squared	0.5439	0.7508
Significant Independent Variable	$-NO_X$, +CO and Lag	$-NO_X$, $+SO_2$, +CO and +Lag
Residual standard error	0.005064	0.009714
I.I.D Assumptions for residual	**Satisfies**	**Does not satisfies**
1. Residual vs Fitted	1. Random pattern	1. Non-random pattern
2. Normal Q-Q Plot	2. Normally distributed	2. Not normally distributed
3. Scale-Location	3. Random pattern (homoscedasticity)	3. Non-random pattern (heteroscedasticity)
4. Residual vs Leverage	4. No influential outliers	4. Has influential outliers

CHAPTER FIVE
CONCLUSION AND RECOMMENDATION

5.1 Introduction

This chapter contains the conclusion and summary for this study. Some suggestions and recommendations for future studies are also included.

5.2 Conclusions

This study was conducted to describe the behaviours of ozone concentration and its precursors at the study locations between Petaling Jaya and Shah Alam, to model a linear relationship of O_3 and its precursors to understand the O_3 concentration and lastly to evaluate the suitability of multiple linear regression (MLR) in understanding the association between precursors in ozone formation in Petaling Jaya and Shah Alam. The data used in this study was obtained from Department of Environment Malaysia in form of hourly data from 2015 until 2017 for Petaling Jaya and Shah Alam.

The first objective which is to describe the behaviours of ozone concentration and its precursors at the study locations, which is Petaling Jaya and Shah Alam, Selangor was done by conducting histogram, boxplot, scatter plot, spider web plot and line graph using Microsoft Excel. Based on the histogram, the pattern of the concentration of O_3 and its precursors in Petaling Jaya were quite similar. For Petaling Jaya and Shah Alam, both histograms show that the distribution of NO_2 is normal. In Petaling Jaya, other precursors that are normally distributed are NO_X and NO. While in Shah Alam, other normally distributed precursor is CO. Based on boxplot, both Petaling Jaya and Shah Alam contain outliers for all variables and SO_2 has the most extreme outliers. Next, the scatter plot for Petaling Jaya shows that precursors that have positive linear relationship is NO_X while NO_2 in Shah Alam. The scatter plot also shows that NO_X, NO and NO_2 have the same criteria and pattern. Based on spider web plot the maximum value of concentration level is in the month of March, June, July and October. The low concentration recorded in the month of February and August. The line graph shows the monthly trend of O_3 and its precursors. The O_3 level will decreases when the level on NO_X and NO increases in both Petaling Jaya and Shah Alam. For the

diurnal average level, the O_3 concentration is high at dawn afternoon and in the middle of the night. This is because at that time, the number of vehicles on the road is lesser compared to other hours. It can be said that with a lesser emission of NO and NO_X cause the O_3 level increases. At afternoon, O_3 level increases because the O_3 needs sunlight for photochemical reactions.

Another objective for this study is to model a linear relationship of ozone and its precursors to understand the ozone concentration. To develop the first model, the dataset that have been cleared from the outliers is used to make sure that the information obtained about understanding the association of the precursors in the formation of average O_3 level is significant. Moreover, lag variable that is developed from the O_3 values is also included in this model to increase the efficiency and the relationship between the precursors and the O_3. In addition, when doing the multicollinearity checking, NO is removed from the entry to any model in Petaling Jaya while NO and NO_2 are removed in Shah Alam. This is because it has high collinearity as shown in correlation matrix for both locations. This is because it has high collinearity as shown in correlation matrix for both locations.

The second model was developed by using the same criteria as the first model except the outlier treatment and with non-randomized data. Thus, the precursors that included in the model for Petaling Jaya were NO_X, NO_2, SO_2, and CO while only NO_X, SO_2, and CO for Shah Alam with variable lag in model 1 and 2 for both locations. For Petaling Jaya, all variables are found to be significant in both model except for Model 2 which is NO_X precursors is not available anymore. Same goes for Shah Alam, all variables are found to be significant in both models while only variable SO_2 is not significant in Model 1.

To fulfil the last objective, the model capability was measured in Petaling Jaya, the multiple R-squared for Model 2 is higher with value 0.7212 compared to in Model 1 which is 0.698. Similar for Shah Alam, model 2 is shown with higher value in Model 2 compared to Model 1 with values 0.7542 and 0.539. The higher the value means that higher percentage of independent variables explains the O_3. Other than that, the i.i.d assumptions for residual variables also were considered. The results have shown that Model 1 has a linear relationship while Model 2 has a non-linear relationship. By looking at the normal Q-Q plot, it is shown that Model 1 has normal distribution while Model 2 not normally distributed. When look at the scale-location result, the plots for both locations; Petaling Jaya and Shah Alam show that Model 1 has random pattern

which is homoscedasticity while Model 2 has non-random pattern which is heteroscedasticity. This means that Model 1 has a constant variance while Model 2 not. Lastly, it is shown that Model 1 do not has any influential outliers while Model 2 has influential outliers. All assumptions are met for Model 1. Thus, this provide the evidence that the established Model 1 which is developed based on randomized data and without outliers is a valid model and should be chosen to understand the linear association between precursors in O_3 formation.

5.3 Recommendation

The analysis of this study is not powerful enough to predict the O_3 level due to several factors of the model performances. Thus, for future studies, it is recommended to use the 'real data' that tell a compelling story which make use all the outliers and lessen in missing values. Hence, it can predict the association of the precursors in the O_3 formation which opposed to normal predictions that many have been done. For that, non-linear regression model may be needed to run the analysis.

Due to the limitation of data, the analysis produced can only make to understand the association of the precursors in the formation of O_3. So, this study would like to suggest another factor that can be evaluate in the formation of O_3 which meteorological factors such is as wind, season and temperature. Hence, with these additional suggestions the paper of analysis that is going to produce in future becomes more remarkable in addressing a global environment problem.

REFERENCES

Abdul-Wahab, S. A., Bakheit, C. S., & Al-Alawi, S. M. (2005). Principal component and multiple regression analysis in modelling of ground-level ozone and factors affecting its concentrations. *Environmental Modelling & Software*, *20*(10), 1263–1271.

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

Arsene, C., Olariu, R. I., & Mihalopoulos, N. (2007). Chemical composition of rainwater in the northeastern Romania, Iasi region (2003–2006). *Atmospheric Environment*, *41*(40), 9452–9467.

Awang, M. Bin, Jaafar, A. B., Abdullah, A. M., Ismail, M. Bin, Hassan, M. N., Abdullah, R., … Noor, H. (2000). Air quality in Malaysia: Impacts, management issues and future challenges. In *Respirology* (Vol. 5). https://doi.org/10.1046/j.1440-1843.2000.00248.x

Azmi, S. Z., Latif, M. T., Ismail, A. S., Juneng, L., & Jemain, A. A. (2010). Trend and status of air quality at three different monitoring stations in the Klang Valley, Malaysia. *Air Quality, Atmosphere and Health*, *3*(1), 53–64. https://doi.org/10.1007/s11869-009-0051-1

Cheng, H., Guo, H., Wang, X., Saunders, S. M., Lam, S. H. M., Jiang, F., … Ho, K. F. (2010). On the relationship between ozone and its precursors in the Pearl River Delta: Application of an observation-based model (OBM). *Environmental Science and Pollution Research*, *17*(3), 547–560. https://doi.org/10.1007/s11356-009-0247-9

De, L. O., & De, L. O. (n.d.). *LO DE S : Lo ca l De n s i t y M ee ts S p ec tr a l O ut li e r De t ec t ion.* (v), 171–179.

Dey, A. K., Hossain, M. A., & Das, K. P. (2015). Regression Analysis for Data Containing Outliers and High Leverage Points. *Alabama Journal of Mathematics*, *39*(March).

Ferdaos, N. A. (2017). *Multicollinearity and Regression Analysis Multicollinearity and Regression Analysis.*

Finlayson-Pitts, B. J., & Pitts Jr, J. N. (1999). *Chemistry of the upper and lower*

atmosphere: theory, experiments, and applications. Elsevier.

Fuhrer, J., Skärby, L., & Ashmore, M. R. (1997). Critical levels for ozone effects on vegetation in Europe. *Environmental Pollution*, *97*(1–2), 91–106.

Ghazali, N. A., Ramli, N. A., Yahaya, A. S., Yusof, N. F. F. M., Sansuddin, N., & Al Madhoun, W. A. (2010). Transformation of nitrogen dioxide into ozone and prediction of ozone concentrations using multiple linear regression techniques. *Environmental Monitoring and Assessment*, *165*(1–4), 475–489. https://doi.org/10.1007/s10661-009-0960-3

Han, S., Bian, H., Feng, Y., Liu, A., Li, X., Zeng, F., & Zhang, X. (2011). Analysis of the Relationship between O3, NO and NO2 in Tianjin, China. *Aerosol Air Qual. Res*, *11*(2), 128–139.

Kang, M., Ragan, B. G., & Park, J. H. (2008). Issues in outcomes research: An overview of randomization techniques for clinical trials. *Journal of Athletic Training*, *43*(2), 215–221. https://doi.org/10.4085/1062-6050-43.2.215

Karnosky, D. F., Zak, D. R., Pregitzer, K. S., Awmack, C. S., Bockheim, J. G., Dickson, R. E., … Kopper, B. J. (2003). Tropospheric O3 moderates responses of temperate hardwood forests to elevated CO2: a synthesis of molecular to ecosystem results from the Aspen FACE project. *Functional Ecology*, *17*(3), 289–304.

Kumar, P., & Imam, B. (2013). Footprints of air pollution and changing environment on the sustainability of built infrastructure. *Science of The Total Environment*, *444*, 85–101.

Lelieveld, J., & Dentener, F. J. (2000). What controls tropospheric ozone? *Journal of Geophysical Research: Atmospheres*, *105*(D3), 3531–3551.

Lu, W.-Z., & Wang, X.-K. (2006). Evolving trend and self-similarity of ozone pollution in central Hong Kong ambient during 1984–2002. *Science of the Total Environment*, *357*(1–3), 160–168.

Marathe, S. A. (2018). Multiple Regression Analysis of Ground level Ozone and its Precursor Pollutants in Coastal Mega City of Mumbai, India. *MOJ Ecology & Environmental Sciences*, *2*(6). https://doi.org/10.15406/mojes.2017.02.00041

Mohd Kalkausar, K., Yusof, K., Azid, A., Shirwan, M., Sani, A., Samsudin, S., … Jamalani, A. (2019). The evaluation on artificial neural networks (ANN) and multiple linear regressions (MLR) models over particulate matter (PM10) variability during haze and non-haze episodes: A decade case study. / *Malaysian*

Journal of Fundamental and Applied Sciences, *15*(2), 164–172.

Monks, P. S., Archibald, A. T., Colette, A., Cooper, O., Coyle, M., Derwent, R., … Williams, M. L. (2015). Tropospheric ozone and its precursors from the urban to the global scale from air quality to short-lived climate forcer. *Atmospheric Chemistry and Physics*, *15*(15), 8889–8973. https://doi.org/10.5194/acp-15-8889-2015

Musa, M., & Ibrahim, K. (2012). Existence of long memory in ozone time series. *Sains Malaysiana*, *41*(11), 1367–1376.

Nishanth, T., Praseed, K. M., Satheesh Kumar, M. K., & Valsaraj, K. T. (2014). Observational study of surface O3, NOx, CH4 and total NMHCs at Kannur, India. *Aerosol and Air Quality Research*, *14*(3), 1074–1088. https://doi.org/10.4209/aaqr.2012.11.0323

Obsborne, J. W., & Overbay, A. (2013). The power of outliers (and why researchers should always check for them). *Journal of Chemical Information and Modeling*, *53*(9), 1689–1699. https://doi.org/10.1017/CBO9781107415324.004

Safieddine, S., Clerbaux, C., George, M., Hadji-Lazaro, J., Hurtmans, D., Coheur, P. F., … Hao, N. (2013). Tropospheric ozone and nitrogen dioxide measurements in urban and rural regions as seen by IASI and GOME-2. *Journal of Geophysical Research Atmospheres*, *118*(18), 10555–10566. https://doi.org/10.1002/jgrd.50669

Samadianfard, S., Delirhasannia, R., Kisi, O., & Agirre-Basurko, E. (2013). Comparative analysis of ozone level prediction models using gene expression programming and multiple linear regression. *Geofizika*, *30*, 43–74.

Schutte, J. M., & Violette, D. M. (1994). The Treatment of Outliers and Influential Observations in Regression-Based Impact Evaluation. *American Council for an Energy Efficient Economy Summer Study*, 8.171-8.176.

Sitch, S., Cox, P. M., Collins, W. J., & Huntingford, C. (2007). Indirect radiative forcing of climate change through ozone effects on the land-carbon sink. *Nature*, *448*(7155), 791.

Sulaiman, A., Ab Rahman, A. A., Abdul Maulud, K. N., Latif, M. T., Ahmad, F., Abdul Wahid, M. A., … Abdul Halim, N. D. (2017). Distribution ozone concentration in Klang Valley using GIS approaches. *Journal of Physics: Conference Series*, *852*(1). https://doi.org/10.1088/1742-6596/852/1/012021

Suresh Kumar Reddy, B., Raghavendra Kumar, K., Balakrishnaiah, G., Rama Gopal,

K., Reddy, R. R., Sivakumar, V., … Shyam Lal, S. (2012). Analysis of diurnal and seasonal behavior of surface ozone and its precursors (NOx) at a semi-arid rural site in southern India. *Aerosol and Air Quality Research*, *12*(6), 1081–1094. https://doi.org/10.4209/aaqr.2012.03.0055

Tan, K. C., Lim, H. S., & Mat Jafri, M. Z. (2013). Analysis of total column ozone in Peninsular Malaysia retrieved from SCIAMACHY. *Atmospheric Pollution Research*, *5*(1), 42–51. https://doi.org/10.5094/apr.2014.006

Vatcheva, K. P., Lee, M., Mccormick, J. B., & Rahbar, M. H. (2016). *Multicollinearity in Regression Analyses Conducted in Epidemiologic Studies*. *6*(2). https://doi.org/10.4172/2161-1165.1000227

Verma, N., Satsangi, A., Lakhani, A., & Kumari, K. M. (2015). *Journal of Chemical , Biological and Physical Sciences Prediction of Ground level Ozone concentration in Ambient Air using Multiple Regression Analysis*. *5*(4), 3685–3696.

Yahaya, A. S., Yusof, N. F. F. M., Sansuddin, N., Ghazali, N. A., Al Madhoun, W. A., & Ramli, N. A. (2009). Transformation of nitrogen dioxide into ozone and prediction of ozone concentrations using multiple linear regression techniques. *Environmental Monitoring and Assessment*, *165*(1–4), 475–489. https://doi.org/10.1007/s10661-009-0960-3

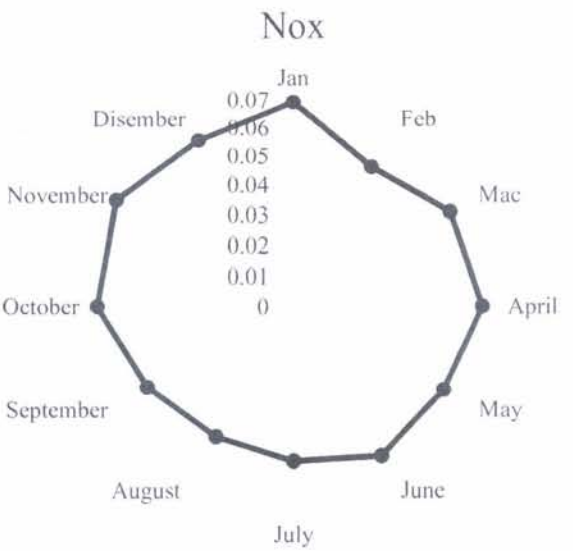
Zheng, J., Zhong, L., Wang, T., Louie, P. K. K., & Li, Z. (2010). Ground-level ozone in the Pearl River Delta region: Analysis of data from a recently established regional air quality monitoring network. *Atmospheric Environment*, *44*(6), 814–823.

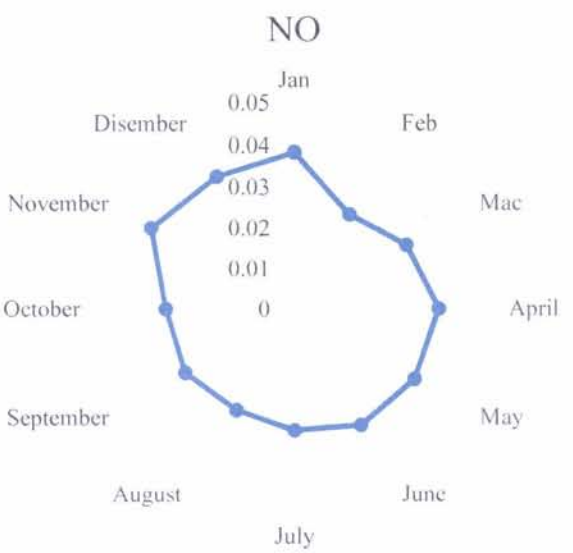
APPENDICES

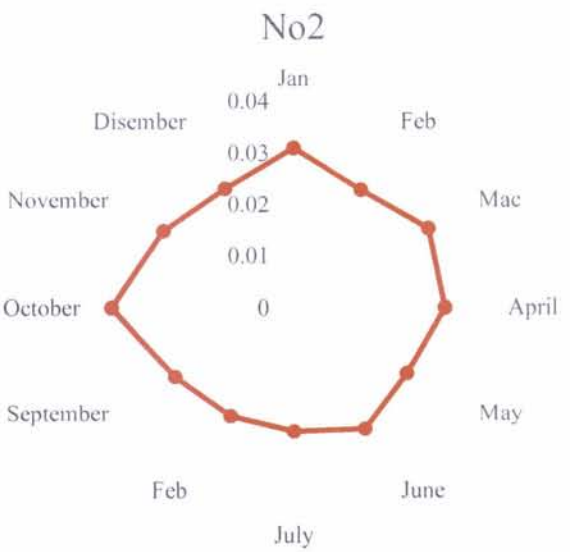
APPENDIX 1

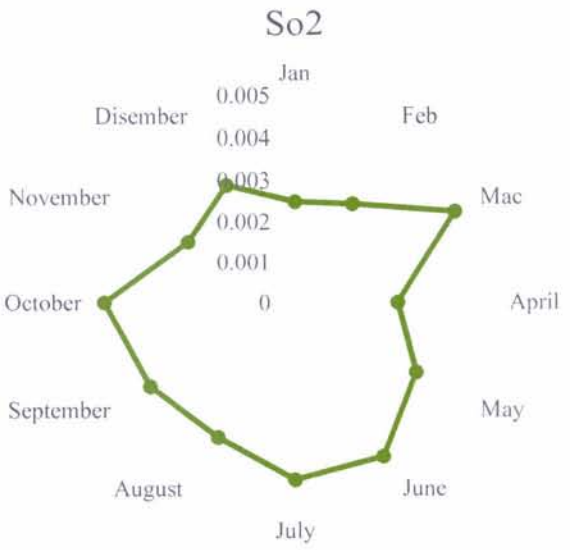
Result of Analysis from Microsoft Excel

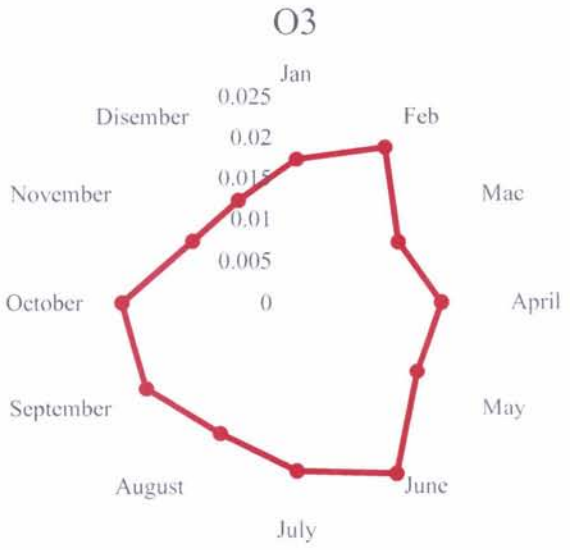
a) Petaling Jaya

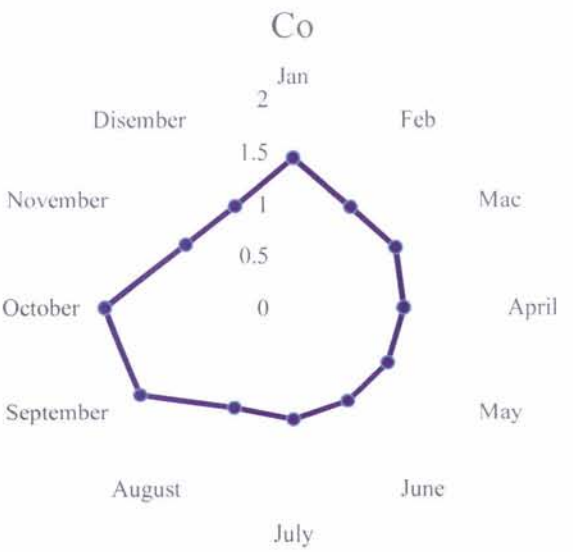


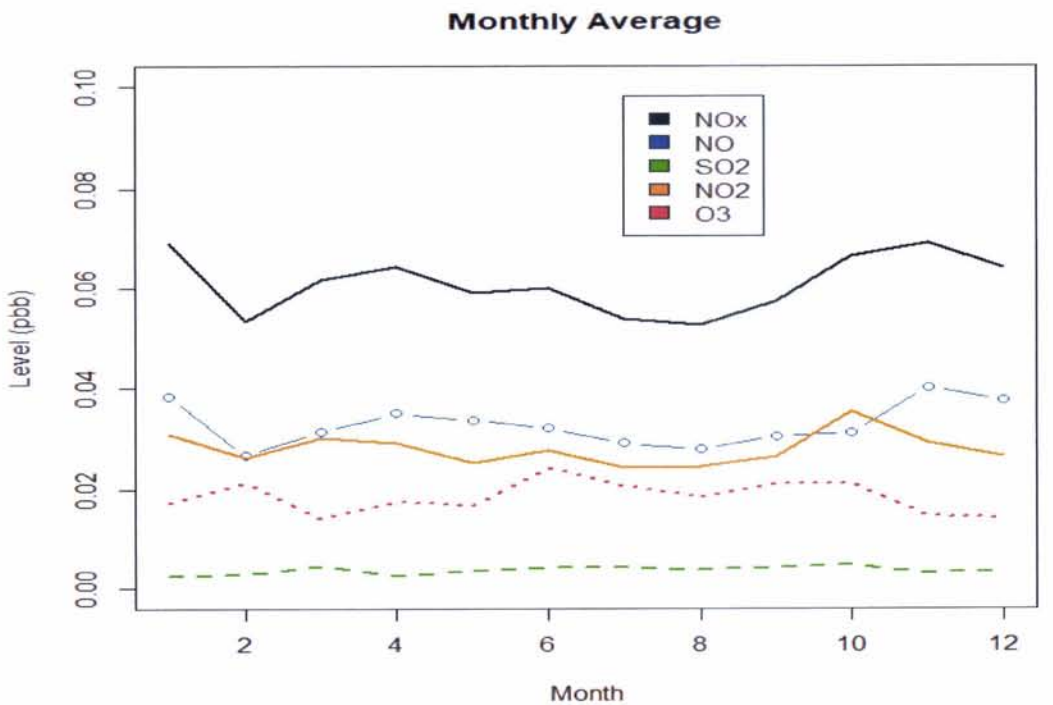


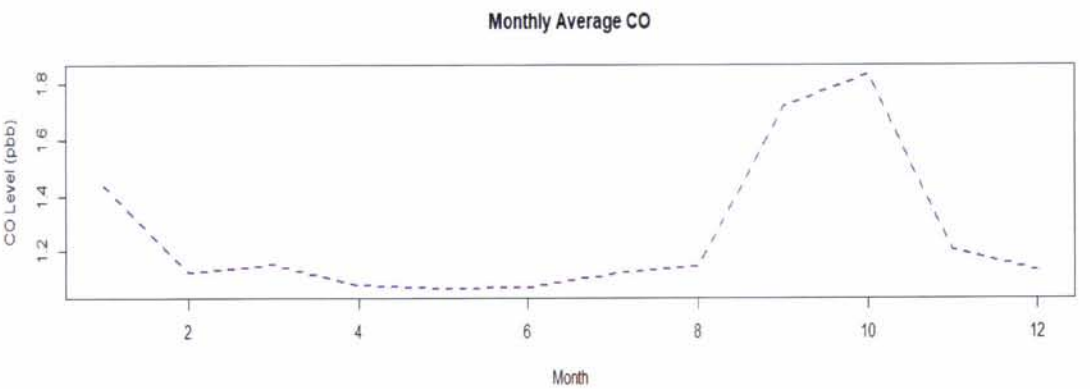

No2
Jan
Feb
Mac
April
May
June
July
Feb
September
October
November
Disember
0.04
0.03
0.02
0.01
0

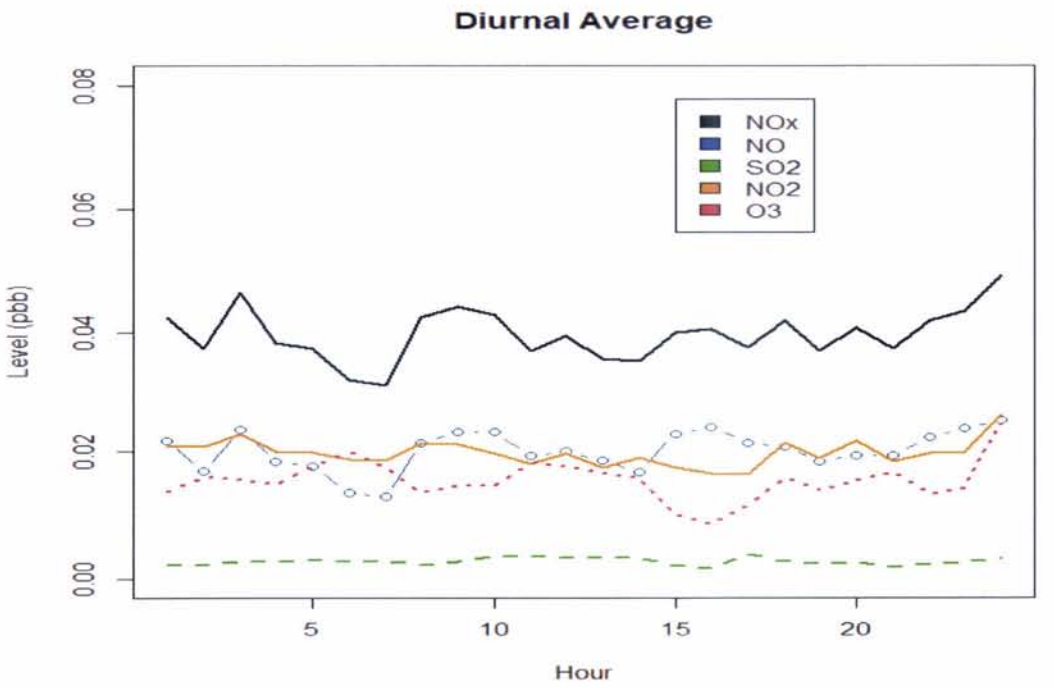

So2
Jan
Feb
Mac
April
May
June
July
August
September
October
November
Disember
0.005
0.004
0.003
0.002
0.001
0

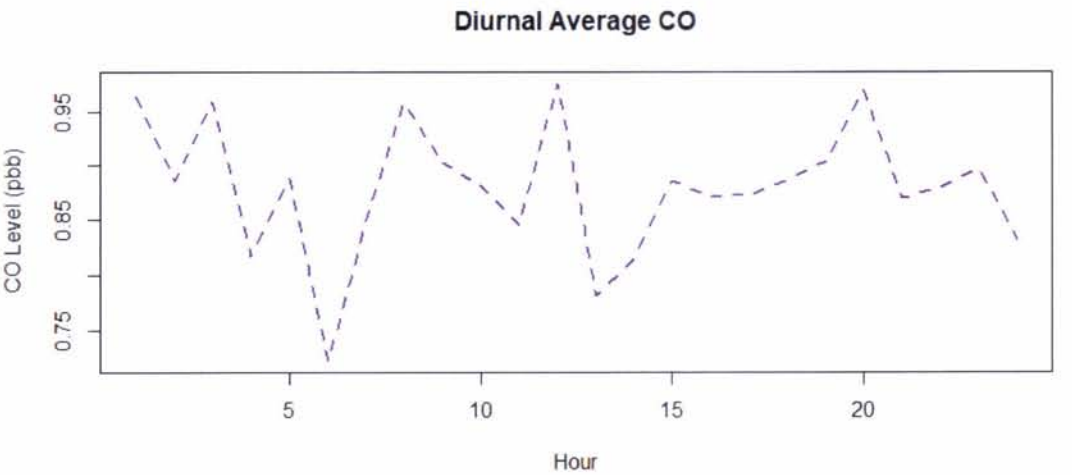

O3
Jan
Feb
Mac
April
May
June
July
August
September
October
November
Disember
0.025
0.02
0.015
0.01
0.005
0


Co
Jan
Feb
Mac
April
May
June
July
August
September
October
November
Disember
2
1.5
1
0.5
0

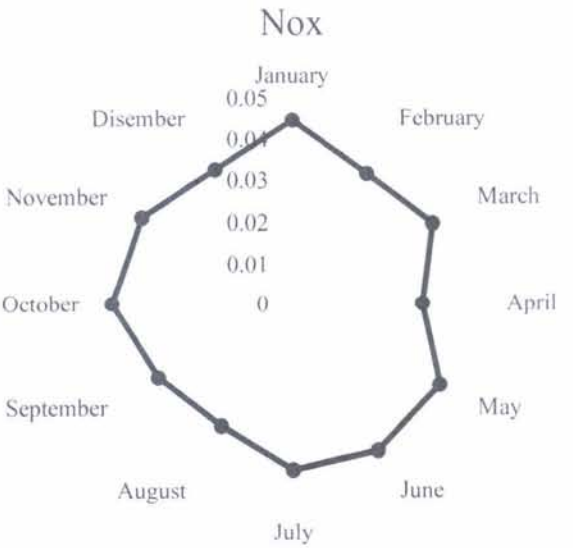

Monthly Average
NOx
NO
SO2
NO2
O3
Level (pbb)
0.00
0.02
0.04
0.06
0.08
0.10
2
4
6
8
10
12
Month

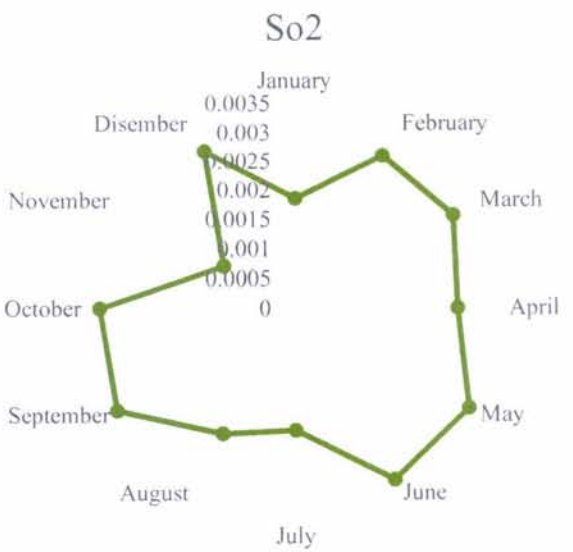

Monthly Average CO
CO Level (pbb)
1.2
1.4
1.6
1.8
2
4
6
8
10
12
Month

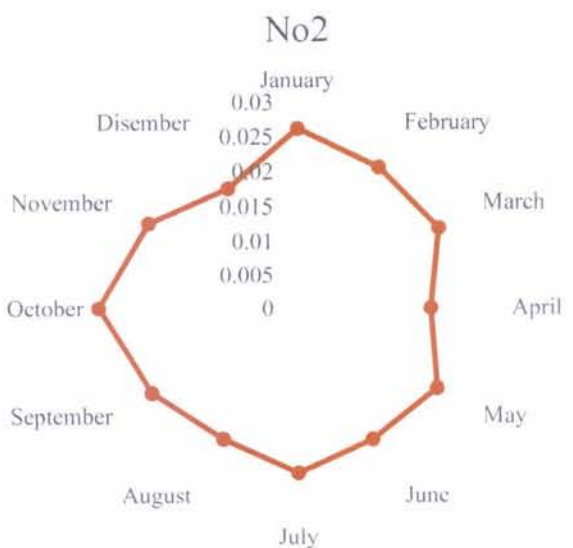

Diurnal Average
NOx
NO
SO2
NO2
O3
Level (pbb)
0.00
0.02
0.04
0.06
0.08
5
10
15
20
Hour

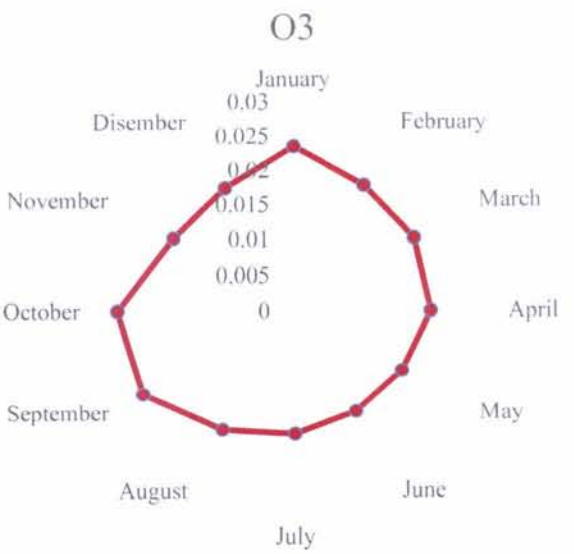

Diurnal Average CO
CO Level (pbb)
0.75
0.85
0.95
5
10
15
20
Hour

b) Shah Alam

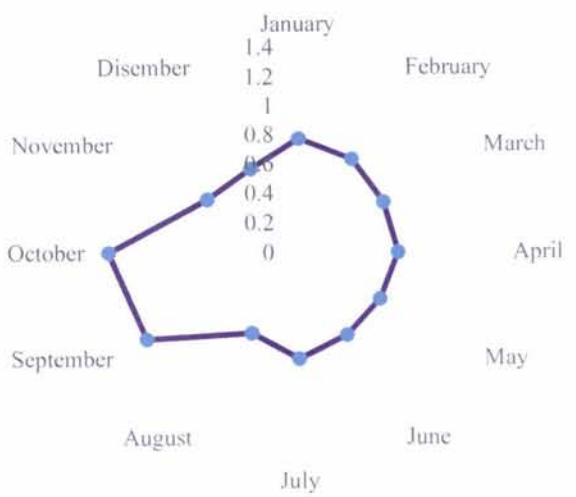

Nox
January
February
March
April
May
June
July
August
September
October
November
Disember
0.05
0.04
0.03
0.02
0.01
0

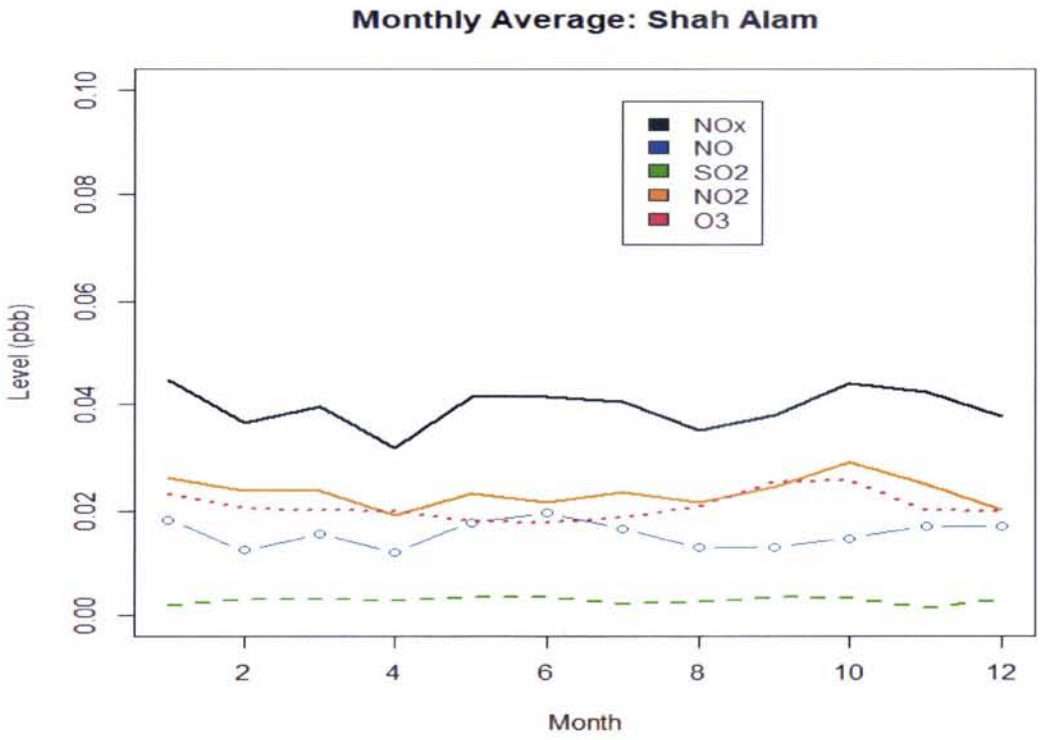

So2
January
February
March
April
May
June
July
August
September
October
November
Disember
0.0035
0.003
0.0025
0.002
0.0015
0.001
0.0005
0

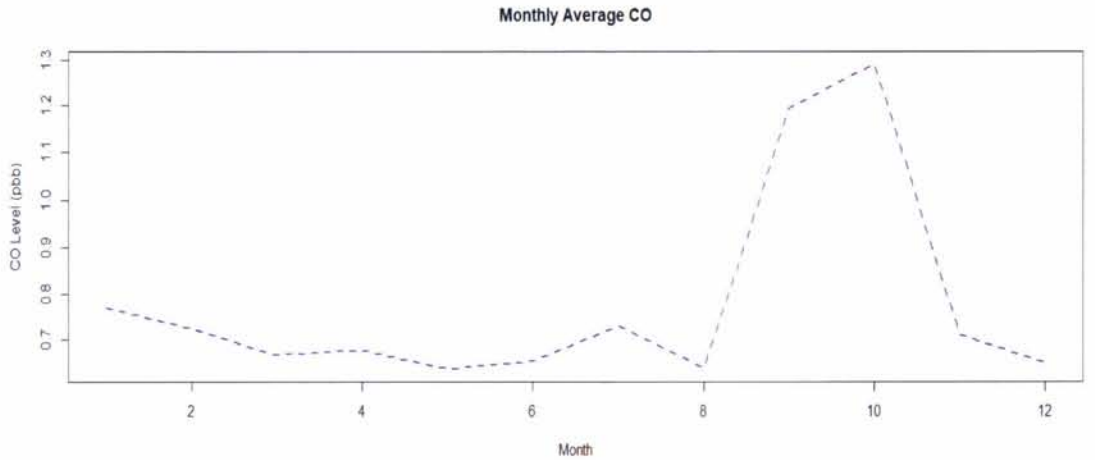

No2
January
February
March
April
May
June
July
August
September
October
November
Disember
0.03
0.025
0.02
0.015
0.01
0.005
0

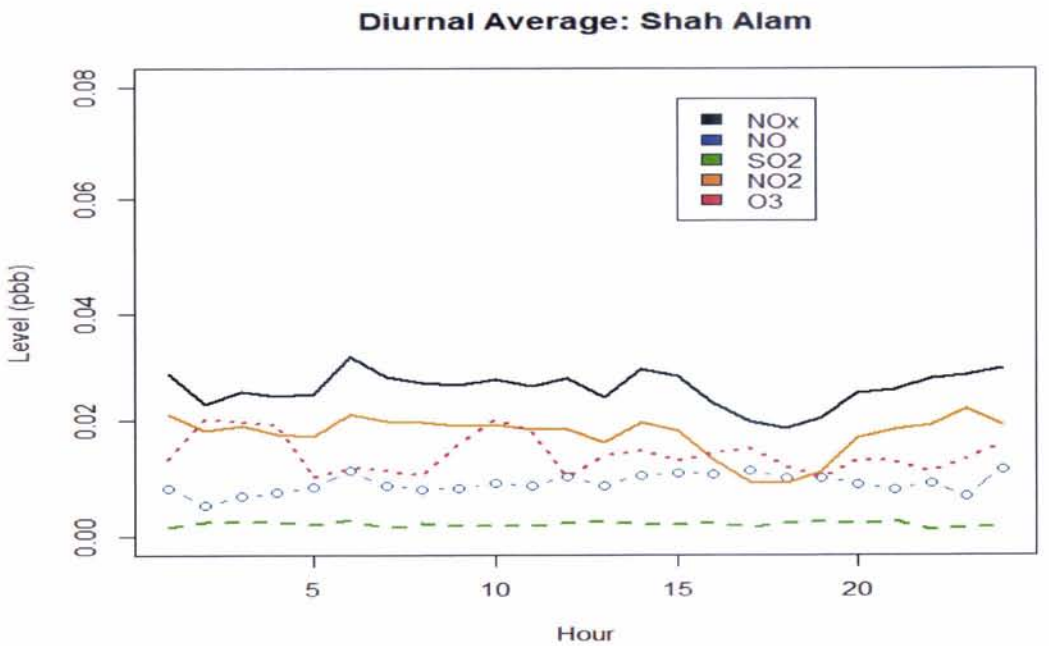

O3
January
February
March
April
May
June
July
August
September
October
November
Disember
0.03
0.025
0.02
0.015
0.01
0.005
0

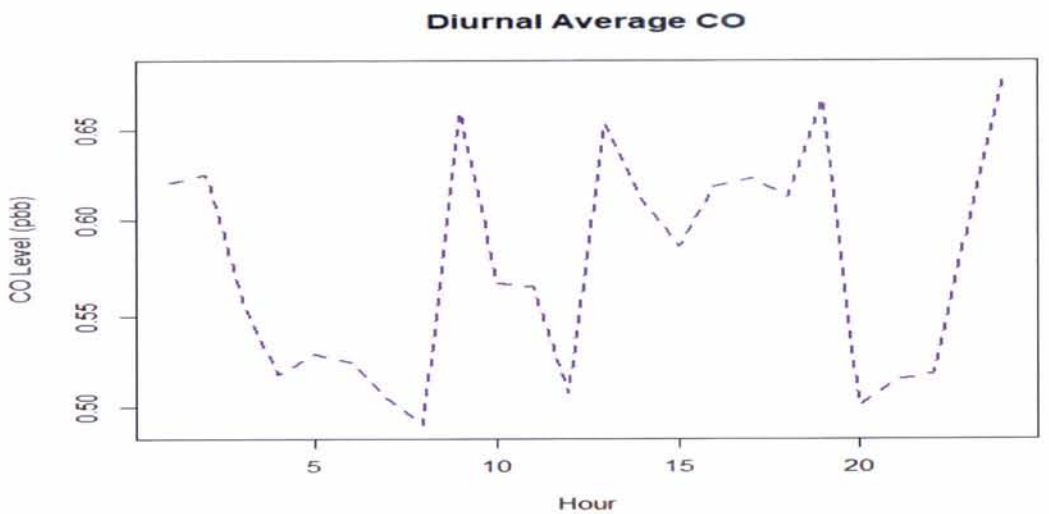

No
January
February
March
April
May
June
July
August
September
October
November
Disember
0.02
0.015
0.01
0.005
0


Co
January
February
March
April
May
June
July
August
September
October
November
Disember
1.4
1.2
1
0.8
0.6
0.4
0.2
0


Monthly Average: Shah Alam
NOx
NO
SO2
NO2
O3
Level (pbb)
0.00
0.02
0.04
0.06
0.08
0.10
2
4
6
8
10
12
Month


Monthly Average CO
CO Level (pbb)
0.7
0.8
0.9
1.0
1.1
1.2
1.3
2
4
6
8
10
12
Month


Diurnal Average: Shah Alam
NOx
NO
SO2
NO2
O3
Level (pbb)
0.00
0.02
0.04
0.06
0.08
5
10
15
20
Hour


Diurnal Average CO
CO Level (pbb)
0.50
0.55
0.60
0.65
5
10
15
20
Hour