UNIVERSITI TEKNOLOGI MARA

MODELING DAILY OZONE (O_3) POLLUTION AND THE PRECURSORS USING FUNCTIONAL REGRESSION

NUR AFIQAH
BINTI MOHD NAZARAN

MSc (Applied Statistics)

January 2020

UNIVERSITI TEKNOLOGI MARA

MODELING DAILY OZONE (O_3) POLLUTION AND THE PRECURSORS USING FUNCTIONAL REGRESSION

NUR AFIQAH BINTI MOHD NAZARAN

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

Faculty of Computer and Mathematical Sciences

January 2020

CONFIRMATION BY SUPERVISOR

APPROVED BY:

DR. NORSHAHIDA SHAADAN

Supervisor

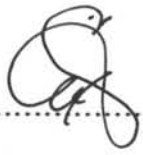
Faculty of Computer and Mathematical Sciences

Universiti Teknologi MARA

AUTHOR'S DECLARATION

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

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

Name of Student	:	Nur Afiqah Binti Mohd Nazaran
Student I.D. No.	:	2018450518
Programme	:	Master of Science in Applied Statistics
Faculty	:	Faculty of Computer and Mathematical Sciences
Dissertation Title	:	Modeling Daily Ozone (O3) Pollution and the Precursors Using Functional Regression
Signature of Student	:	
Date	:	January 2020

ABSTRACT

In Malaysia, ground level ozone (O_3) is considered as one of the most influential factors in air pollution due to the increasing sources of ozone precursors. Therefore, the concentration level of Ozone should get ample attention because of it gives negative effects to the environment, human health and also vegetation. In this study, daily hourly air pollutant data set (Ozone (O_3), Carbon Oxide (CO), Nitrogen Dioxide (NO_2)) and meteorological variables (temperature and humidity) for nine years' period (2009-2017) in Selangor and Penang state were selected for analysis in this study. Two monitoring stations were selected are Petaling Jaya and Perai. This study mainly has three objectives that is firstly to describe the diurnal and spatial behavior of Ozone (O3), the precursors (CO, NO2) and meteorological variables (temperature and humidity) at Petaling Jaya and Perai. Secondly, to investigate the diurnal inter - association pattern between O3 and the precursors as well as meteorological variables at Petaling Jaya and Perai.. Thirdly, to model the diurnal relationship between Ozone (O3) and the precursors (CO and NO2) as well as the meteorological variables (temperature and humidity) using Bivariate Functional Linear Regression at Petaling Jaya and PeraiBased on the all set curves obtained, the figures show a slight difference in concentration level of each variable but quite similar pattern for both stations. The same peak tends to appear at around 3 pm in the evening. Petaling Jaya station was observed to have higher peak for Ozone (O3), Carbon Oxide (CO), Nitrogen Dioxide (NO2) and temperature variable compared to Perai station The correlation coefficient of temperature variable with Ozone level has shown as the highest positive correlation for both Perai and Petaling Jaya station. Functional linear regression has shown that there is only slightly different for the result using with or without the outlier. Thus, it shows that the functional regression model is actually a robust model to Malaysia dataset as well as there is only a slight difference to the result whenever using with or without influencing outlier. The result does not give big changes or impact whenever using with or without outlier dataset.

ACKNOWLEDGEMENT

Firstly, I wish to thank Allah s.w.t for giving me the opportunity to embark on my master and for completing this long and challenging journey successfully. Without His bless and Mercifulness, this report may not be completed on time. Furthermore, I would like to express my sense of gratitude and sincere appreciation to my lovely supervisor Dr Norshahida Shaadan for her guidance, time, efforts, positive encouragement and patience. Without her guidance, I may not be complete this dissertation. Thank you for always giving positive words and encouragement during this two semester and also for providing countless ideas, support and comment while assisting me on my dissertation. There is no other lecturer that has beautiful and pure heart as you. You inspire and pick me up from the beginning until I finish the dissertation.

My appreciation goes to all my lecturers that dedicatedly teach us to be a better person than yesterday and also special thanks to my colleagues here who have helped me a lot continuously and persistently without a sigh during my journey to finish my research. A special place in my heart for all those never failed to put me up even during the stressful year of the study process.

Finally, this thesis is dedicated to my family for always being a truly supportive and patient with me throughout my master journey of completing postgraduate studies and through writing this thesis. This journey would not possible to come in this way without them. Alhamdulilah.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

CHAPTER ONE

INTRODUCTION

1.1 Background of Study

Nowadays, in the era of modernization and globalization, the quality of the air has become critical from day to days. All harmful effects of any sources that lead to the pollution in the atmosphere and the degradation of the ecosystem is called air pollution (Azam, Zanjani, & Mood, 2016). Based on the Natural Resources Defense Council (NRDC), pollutants that is released into the air that cause detrimental to health and the world is called as air pollution (Mackenzie, 2016). There are two classifications for the air pollution that is firstly visible air pollution and invisible air pollution. The smog is an example of air pollution that is visible while the example of invisible air pollution is nitrogen oxides (NOx) and carbon monoxide (CO). The air pollution issue has become a major concern as the increasing amount of air pollution in many large cities has given bad impact on the health and also to the environment. This situation has become alarming to the whole world because it's not only affecting individual quality of life, but it is actually affecting the public health situation. Based on the World Health Organization (WHO), air pollution is the biggest environmental risk to our health as for about one in every nine deaths annually and also that out of 10 people there are 9 people will breathe air containing high levels of pollutants (WHO, 2018). Generally, all the countries around the world are affected by the air pollution not consider the regions, age, ethnicity, or the socioeconomic groups (WHO, 2018). It also estimated that some geographical areas expose much higher levels of air pollutants than the other place, for example, citizens in Africa, Asia or the Middle East. World Health Organization highlights that this situation threatens the world, especially in the low-income cities would affect the most followed by other place (WHO, 2016).

Air pollution is one of the most environmental issues that are critically considered by the government of Malaysia. Overexpose to high air pollution level

negatively affect human's heath and if the level is too high it can even may cause human fatality. In The Star Online Newspaper on March 2019, there are 2775 people affected by chemical pollution that spread into the air and cause air pollution in Pasir Gudang Johor (Shah & Nordin, 2019). Air pollution levels in the Malaysia environment is described by the emission of five important air pollutants, including particulate matter of a size less than 10 micrometers called PM_{10}, Ozone (O_3), Sulfur Dioxide (SO_2), Nitrogen Oxide (NOx) and Carbon Oxide (CO) (M. B. Awang et al., 2001). This statement is also supported by another study where the same five types of air pollutant in Malaysia (Rani, Azid, Juahir, Khalit, & Samsudin, 2018). In Malaysia, the air pollution situation in the environment is measured with respect to air quality level by means of Air Pollution Index (API). PM_{10} and O_3 have been identified as the most dominant pollutant that contributes to the major API calculation. PM_{10} is often used as a proxy to assess the haze occurrence in Malaysia while O_3 is a secondary type of pollutants that is formed via the oxidation process between its precursors and climate variation defined by several meteorological variables. Other than CO and SO_2, the reactive Nitrogen Oxides (NOx) have been identified as the most important precursor of O_3 (Li et al., 2019). Several areas in Malaysia have frequently recorded high level of O_3. Not like PM_{10}, this particular pollutant is more dangerous, its existence is often neglected because this secondary gas pollution is invisible, cannot be seen by the naked eye.

The major source of NOx in Malaysia comes from transports and vehicles, where the higher daily flow of vehicles will increase also the level concentration of NOx (Yassen, Jahi, & Ahmad, 2005). Based on the Malaysia Automotive Association (MAA) info about summary of new passenger and commercial vehicles registered in Malaysia for the year 1980 to 2018 there were increment from 97,262 total units in 1980 to 598,714 total unit vehicles has been registered in 2018 ((MAA), 2019). Since there have increment in the number of vehicles registered in Malaysia, the level of air pollution cause from the vehicles would also increase. Table 1.1 shows the summary of new passenger and commercial vehicles registered in Malaysia for the year 1980 to 2018 from the Malaysian Automotive Association.

Table 1.1
Summary of New Passenger & Commercial Vehicles Registered in Malaysia for The Year 1980 To 2018

Year	Passenger Cars	Commercial Vehicles	4x4 Vehicles	Total Vehicles
1980	80,420	16,842	-	97,262
1985	63,857	26,742	4,400	94,999
1990	106,454	51,420	7,987	165,861
1995	224,991	47,235	13,566	285,792
2000	282,103	33,732	27,338	343,173
2005	416,692	97,820	37,804	552,316
2006	366,738	90,471	33,559	490,768
2007	442,885	44,291	-	487,176
2008	497,459	50,656	-	548,115
2009	486,342	50,563	-	536,905
2010	543,594	61,562	-	605,156
2011	535,113	65,010	-	600,123
2012	552,158	75,575	-	627,753
2013	576,640	79,104	-	655,744
2014	588,341	78,124	-	666,465
2015	591,298	75,376	-	666,674
2016	514,545	65,579	-	580,124
2017	514,679	61,956	-	576,635
2018	533,202	65,512	-	598,714

Note:

(i) Passenger Vehicle industry reclassified in January 2007 and includes all passenger carrying vehicles.
i.e. Passenger Cars, 4WD/SUV, Window Van and MPV models.

(ii) Commercial Vehicles also reclassified on 1 January 2007 and includes Trucks, Prime Movers, Pick-up, Panel Vans, Bus & Other

Sources (accessed 29th March 2019): http://www.maa.org.my/info_summary.htm

In relation to O_3 issue, it has been reported that in 2020, Malaysia would take the lead to become industrialized and develop the country as in Vision 2020 by Tun Dr Mahathir Mohamad. Therefore, the air quality level in Malaysia should become priority based on the vision to become a developed country. One of the least polluted

environments in Asia is Malaysia (Afroz, Hassan, & Ibrahim, 2003). Therefore, the air pollution monitoring should be tightened to reduce the air pollution thus make our atmosphere clean. One of the purposes of the monitoring activity is to increase understanding of the temporal and spatial formation and destruction behavior of the pollutants. The information can be explored by modeling the association between the pollutant and several possible factors.

This study aims to investigate the formation and destruction behavior of O_3 by exploring the association between O_3 precursors and several possible meteorological influences using regression models. In comparison to the pointwise regression, an alternative regression model, namely the functional regression will be employed to enhance the understanding of their dynamic association.

1.2 Problem Statement

Ground level Ozone (O3) has been identified as the most dominant pollutant after PM_{10} in the Malaysia environment. High O_3 concentration was observed at several locations in the country such as Shah Alam, Banting, and Petaling Jaya. O_3 is an invisible type of secondary gas pollutant and has many negative health impacts including the critical diseases such as pulmonary disease, lung cancer, premature death and the worst scenario would lead to sudden fatality. Ozone (O_3) is produced based on oxidization process between its precursor including NOx, CO, NO2 and other chemical components such as volatile organic compounds (VOCs) with the influence of weather variable particularly during high temperature level.

Malaysia has reported to experience an increasing number of registered vehicles on the road. Thus, this scenario would lead to the increase in the NOx emissions to the environment. Therefore, air quality monitoring activity needs to be tightened. Knowledge of the formation and destruction behavior of Ozone (O_3) need to be explored so that the information can be used as the input for mitigation and adaption purposes. **To provide the knowledge, a model to understand the association and interaction behavior between Ozone (O_3), the precursors and the meteorological variables need to be built.** However, in the practical application, the association models often being developed using average daily data to investigate the

influence of the precursor towards Ozone (O_3) level within a day process. In a particular point of view, the approach has an important drawback; the results obtained were found limited particularly in knowing the relationship between Ozone (O_3) and the precursors across the entire period or continuum of 24-hour time of the day process (i.e within the 24-hours). Additionally, the existence of Ozone (O_3) is naturally continuous with time, so as the precursors and the meteorological variables.

Thus, an alternative approach is needed to model the continuous relationship between Ozone (O_3) and the precursors over the entire period of time. This is conducted using functional data approach. This study will consider functional regression modelling to investigate the association between Ozone (O_3) and the precursors to overcome the problem. This study will give benefit to the Department of Environment Malaysia and also to the environmentalist in Malaysia. This research outcome can help as to determine the time and variables that influence Ozone (O_3) concentration in the Perai and Petaling Jaya station.

1.3 Theoretical Framework

This study has four independent variables that are Carbon Oxide (CO), Nitrogen Dioxide (NO_2), temperature and humidity. The dependent variable will be the Ozone (O_3). Figure 1.1 below shows the theoretical framework for this study.

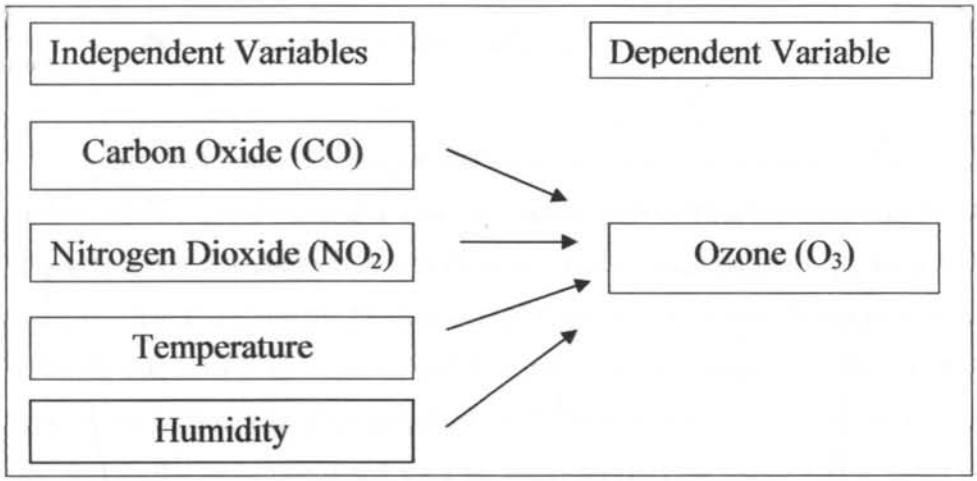


Figure 1.1 Theoretical Framework

1.4 Research Questions

1. What is the diurnal and spatial behavior of Ozone, the precursors and the meteorological variables between two industrial sites (Petaling Jaya and Perai) in peninsular Malaysia?
2. What is the diurnal inter - association pattern between Ozone and its precursors as well as the meteorological variables at the sites?
3. What is the temporal (diurnal) influence of precursors and meteorological variables towards Ozone concentration across 24 hours' period at the sites?

1.5 Research Objectives

1. To describe the diurnal and spatial behavior of Ozone (O_3), the precursors (CO, NO_2) and meteorological variables (temperature and humidity) at Petaling Jaya and Perai.
2. To investigate the diurnal inter - association pattern between O_3 and the precursors as well as meteorological variables at Petaling Jaya and Perai.
3. To model the diurnal relationship between Ozone (O_3) and the precursors (CO and NO_2) as well as the meteorological variables (temperature and humidity) using Bivariate Functional Linear Regression at Petaling Jaya and Perai.

1.6 Scope and Limitation of the Study

This study only focuses on the dynamic behavior of Ozone (O_3) at selected industrial locations in Selangor and Penang that are in Petaling Jaya and Perai. The selected station was chosen since both station have consistency in high level of Ozone concentration. Based on the Department of Statistics Malaysia, the population of total Malaysia has grown to 1.1% in 2018 which is 32.4 million (DOSM, 2018a). The Department of Statistics also stated that in 2018 with 6.47 million people in Selangor has been as the most populated state in Malaysia (DOSM, 2018b). The rapid growth of population in Selangor also leads to rapid development and traffic pollution. Based

on the previous studies, 70-75% of the total air pollution in Malaysia is caused from emission of mobile sources (Afroz et al., 2003). Both Petaling Jaya and Perai station were chosen as the study area due to heavily industrialized area that lead to high air pollution, with a high traffic density. Petaling Jaya is located in Selangor where known as the most populated state in Malaysia. Petaling Jaya also called as PJ is surrounded by the Malaysian capital that are Shah Alam capital of Selangor, Subang Jaya to the west, Puchong to the south, Kuala Lumpur to the east and Sungai buloh to the north. The Petaling Jaya station is located at Sek. Keb. Bandar Utama, Petaling Jaya with latitude 03°06.612' and longitude 101°42.274'. Meanwhile the second air quality monitoring station that is Perai is located at the southern bank of the Perai River and to the North of Butterworth. Perai is considered as industrial area since Perai's industries are major contributors in the town's port facilities that are shipments of coal and scrap metal. This second station is located at Sek. Keb. Seberang Jaya II, Perai with latitude 05°23.890' and longitude 100°24.194'. The major limitation to this study is due to data availability where the missing value may be appearing. The data may be in years, daily, hourly and we need to Figure out which one is suitable for this study. Lastly, the size of the data may be different since the initially study need data for 10 years but only 9 years are given.

1.7 Significance of The Study

Since the number of research in this study is still very low in Malaysia, the results of this study will provide a valuable information to the researchers to model the Ozone (O_3) behavior. The finding from this study may serve as a reference material to the future researchers in this environment area of study to enhance the knowledge and the model. Other than the researcher, the environmentalist also would get the benefit from this study as they will understand the Ozone (O_3) formation more dynamically. This study also will help the environmentalist who is governing the air pollution problem to give information to the people, enhance new policy and can take a proper action or new solution on this problem on how to maintain and to reduce the air pollution. Thus, the environmentalist can make our atmosphere cleaner so that the public health will not become an alarming hazard in the country anymore. Moreover, the finding of this study will be useful to improve the knowledge of photochemical

activity and the interaction between the meteorological factors and air pollution emission in Malaysia especially in urban and rural location in Selangor.

Next, the result also will tell us at what hour the Ozone (O_3) concentration level is high to increase the understanding on Ozone (O_3) formation and destruction and thus would provide insight on the relationship of the possible causes and sources of Ozone (O_3). Malaysian citizen will be able to know the appropriate time in daily life activity, particularly for the outdoor for them to be more alert on this Ozone pollution problem. This will help all people in Malaysia including the citizen and also the government will be a better knowledge about which hour in a day they should stay in the house to prevent the air pollution effect for the long term. Other than that, this study will compare between the conventional approach with the advanced analytical approach. Thus, this will help the government to choose the best statistical method to dig into more the air pollution issue.

Lastly, this study also will provide a better knowledge for all the people being more cautious and alert on the air pollution problem. Although much research on air pollution has been carried out in Malaysia and many statistical models have been proposed, the major importance in this study is in the functional regression method for the dynamic influence between Carbon Oxide (CO), Nitrogen Dioxide (NO_2), temperature and humidity level with Ozone (O_3) gas. This study is first to introduce the functional regression approach to determine the association in the variables and Ozone (O_3).

CHAPTER TWO

LITERATURE REVIEW

2.1 Air Pollution in Worldwide

According to previous studies, it has been estimated that 4.3 million people would die every year that cause from the air pollution that is 3.3 million from Asia, and 3.7 million people die cause from ambient pollution that are 2.6 million comes from Asia (Lancet, 2016). Iran is the world's third main polluted country in the world. It has reached to harmful level of air pollution in Iran in some of the city such as Ahwaz, Ilam, Kermanshah, Khoramabad, Sanandaj, Uromiyeh and Yasouj since the countries are considered as developing countries (WHO, 2011). Based on the air quality index (AQI), air quality is classified into six different color codes and each of it represents to a different level of health implications. Figure 2.1 shows the world's air pollution real-time quality index for 21st March 2019.

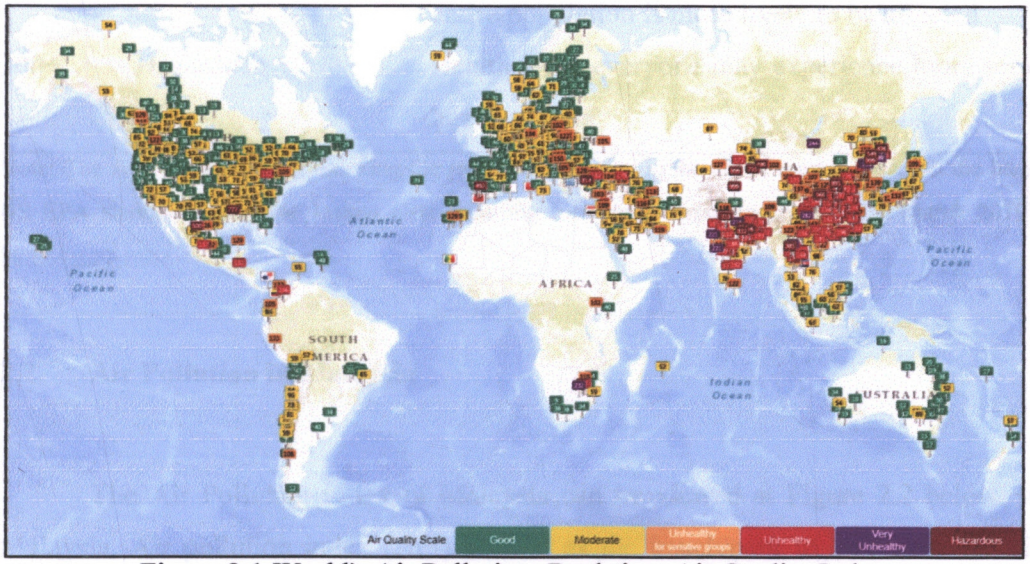


Figure 2.1 World's Air Pollution: Real-time Air Quality Index
Sources (accessed 21th March 2019): https://waqi.info/

The Air quality index (AQI) in the world has six different stages of air pollution level. Table 2.1 shows the air quality index with the air pollution level and health implication that will affect people.

Table 2.1
Air Quality Scale

AQI	Air pollution level	Health implications
0-50	Good	No risk
51-100	Moderate	Air quality is accepTable; however, for some pollutants there may be a moderate health concern for a very small number of people who are unusually sensitive to air pollution
101-150	Unhealthy for Sensitive Groups	Members of sensitive groups may experience health effects. The general public is not likely to be affected
151-200	Unhealthy	Everyone may begin to experience health effects; members of sensitive groups may experience more serious health effects
201-300	Very Unhealthy	Health warnings of emergency conditions. The entire population is more likely to be affected
300+	Hazardous	Health alert: everyone may experience more serious health effects

Note: The AQI scale used for indexing the real-time pollution in the above map is based on the latest US EPA standard, using the Instant Cast reporting formula. Sources (accessed 21th March 2019): https://waqi.info/

2.2 Air Pollution in Malaysia

The Air Pollutant Index of Malaysia can be viewed at Figure 2.2 below. In Malaysia, Air pollution index (API) is used as the benchmark to determine the air quality. The Malaysian Air Pollution Index (API) system based on the Pollutant Standard Index that created by United States Environment Protection Agency (US-EPA). Based on the Air Pollutant Index of Malaysia it can be considered as good in West and East Malaysia that is, the API level is between 0 to 50 and mostly moderate

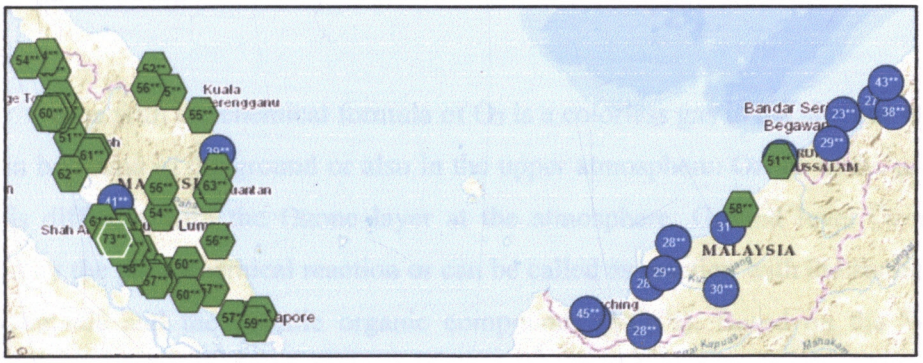
pollution index in the peninsular Malaysia where the API level is between 51 to 100. The air pollutant index flowchart in Malaysia can be viewed at Table 2.2 below.

Based on the Air Pollutant Index of Malaysia (APIMS), it has shown that the air pollution index API was unhealthy reading in Rompin, Pahang with 151-unit API reading and also in Shah Alam which is 105-unit API reading on 18 March 2019 (Ashaha, 2019).

Table 2.2
Air Pollution level of AQI benchmarks

AQI	Air pollution level
0-50	Good
51-100	Moderate
101-200	Unhealthy
201-300	Very Unhealthy
>301	Hazardous

Source: DOE (2019)



Sources (accessed 29th March 2019): http://apims.doe.gov.my/public_v2/home.html

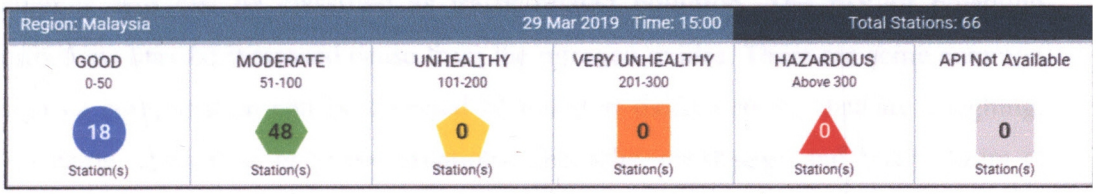


Figure 2.2 Air Pollutant Index of Malaysia, Department of Environment

From the Figure 2.2, it is shown that mostly the stations in Malaysia are in moderate. Based on the Department of Environment, the air pollution index level at Johan Setia Klang, showed the highest API level of 73 but still can be considered as moderate phase. While in Labuan, Sabah showed the lowest level of API since the API reading is 23 and considered as the best API. As time flies, the air quality of Malaysia has become more serious in these past few years. One of industrial place in Malaysia comes from Klang Valley. For the last decade of the twentieth century, Klang Valley was transforming rapidly into an urban region that is basically in an industrial city, has contributed to air pollution issue (Abdullah, Samah, & Jun, 2012).

There are many factors that contribute to the air pollution. Various sources such as factories, power plants, dry cleaners, vehicles and even windblown dust and wildfires can be the cause of air pollution (Abdullah et al., 2012). Sources that come from mobile, stationary and open burning are three major causes of air pollution in Malaysia that contribute up to 70-75% for mobile sources, 20-25% for stationary sources and 3-5% for open burning (Afroz et al., 2003).

2.3 Air Pollutants

Ozone with the chemical formula of O_3 is a colorless gas in the air. The Ozone gas can be found in the ground or also in the upper atmosphere. Ozone at the ground level is different with the Ozone layer at the atmosphere. Ground level Ozone is formed by the photochemical reaction or can be called as reaction with the sunlight of the pollutants and the volatile organic compounds (VOCs). Based on the WHO, Ozone pollution level is higher during sunny weather and can cause breathing problems, asthma, and lung diseases. Nitrogen oxide is mainly coming from motor engines also can be classified as traffic-related pollution. The risk of breathing infections may be increased cause from the nitrogen oxides. There are some common sign or complication can be detected of nitrogen oxide toxicity that are coughing, wheezing, eyes, nose or throat irritations, headache, chest pain, and fever (Azam et al., 2016). Many types of cancers are caused from the pollutants such as pollutant of the suspended area such as smokes, fumes, mists, dusts, hydrocarbons, volatile organic compounds (VOCs) (Kjellstrom et al.). Based on the previous study, many

developed countries and the WHO have their own legislated guideline for PM, NOx, Ozone levels (Table 3), while volatile organic compounds (VOCs) may not be legislated for (Kjellstrom et al.). In France, 48% emission of nitrogen oxide comes from the road traffic in 2002 followed by others factor like forestry, manufacturing and energy transformation (Morand & Maesano, 2004). Nitrogen Dioxide (NO2) also considered as gaseous pollutants in the air pollution that are formed in the combustion processes. Nitrogen Dioxide (NO2) tend to give people to have infections in respiratory such as flu and also react to allergens to give allergies. Carbon Oxides (CO) is generally coming from road vehicle exhausts and has potential to affect the oxygen ability to be carried out around the body. Based on (Mofijur et al., 2016), the development of tissue in young children and mortal growth in pregnant can be affected by Carbon Oxide (CO). Particulate matter (PM) also known as particle pollutants are major parts of air pollutants. It is a complex mixture of solid particle and liquid droplets that are found in the atmosphere. PM are varied in size, composition and concentration. Normally, $PM_{2.5}$ and PM_{10} usually discuss in many articles. Based on previous studies, PM_{10} that is big particles can survive in the atmosphere for minutes or hours and also can travel in range 100yards to 30 miles while the other one that is $PM_{2.5}$ which is fine particles can survive in the atmosphere for days or even weeks longer and also can travel farther (Morand & Maesano, 2004). Volatile organic compounds (VOCs) becomes more important to the source of air pollution especially in the urban area. Generally, the VOCs produced by the emission from burning coal, gasoline and oil, from solvents, cleaners and paint. All the VOCs sources contribute to different baseline level in the air. Based on previous studies, there are three main sources of VOC that are 24% of non-methane VOC emission in mainland France in 2002 which is road traffic, 31% of manufacturing industries and 22% comes from residential (Morand & Maesano, 2004). Table 2.3 below show the air quality limits for selected countries based on the World Health Organization recommended limits.

Table 2.3
Air Quality Limits for Selected Countries and World Health Organization Recommended Limits

Pollutant (μg/m3)	European Union	United States	Canada (2020 standards)	World Health Organization
$PM_{2.5}$	25	35*	8.8	10
PM_{10}	40	150*	-	20
NO_2	40	101.4**	-	40
SO_2	125	199.6***	-	20
O_3	120	139.7****	123	100

Note: PM, particulate matter; NO_2, Nitrogen Dioxide; SO_2, Sulphur dioxide; O_3, Ozone. All limits as average over 1 year except; Ozone: maximum daily hour mean. The pollutant which are in ppb are converted into μg/m3 using a conversion factor. Sources (accessed 29th March 2019):

https://www.epa.gov/criteria-air-pollutants/naaqs-Table
https://www.ccme.ca/en/resources/air/pm_Ozone.html
http://www.who.int/mediacentre/factsheets/fs313/en/
Conversion factors available at:
https://uk-air.defra.gov.uk/assets/documents/reports/cat06/0502160851_Conversion_Factors_Between_ppb_and.pdf

2.4 Impact of Air Pollutants

The effect of exposure to an air pollutant will produce different potential health effects. The air pollution can relate to the no communicable disease that is other than the lung and airway diseases (Schraufnagel et al., 2019). Based on the Schraufnagel et al. (2019) air pollution has contributed to 50000 lung cancer deaths, 1.6 million of chronic obstructive pulmonary disease, 19% of all cardiovascular deaths and 21% of all stroke deaths. Air pollution also would give big impact to the environmental issues such as the Ozone depletion, greenhouse effect, and acid rain that comes from the emission from fossil fuels such as carbon dioxide and Nitrogen Dioxide (Scoullos, 2013). Nitrogen oxide (NOx) would give both impacts to the long term exposure or also to the short term exposure. Based on the WHO, as the concentration of nitrogen oxide (NOx) increase, the bronchitis symptoms of asthmatic children would also increase and also as nitrogen oxide increase, the lung function growth in children would decrease (WHO, 2005). Based on previous studies, exposure to the Ozone (O_3) will give health effect such as decrement in pulmonary function, coughing and chest discomfort, and also increased asthma attacks (Vallero,

2008). Unwanted chemical pollutants which are VOCs are categorized as a carcinogen and mutagen compound and also responsible for the existence and development of cancer (Azam et al., 2016). Based on previous studies, PM_{10} would give effect on the environment where the smoke and dust can dirty and discolor structure, while on the health effects PM_{10} would cause nose and throat irritation. Lung damage, bronchitis, risk of cardiac arrest and also early death (Morand & Maesano, 2004). At high level concentration of Carbon Oxide, it will become toxic to the health that may cause nausea, headaches, reduced the thinking ability and may cause to death (Morand & Maesano, 2004). While it different effect to the Nitrogen Dioxide (NO2) where it may effect to respiratory system such as worsen the respiratory illness and decrease in lung function.

2.5 Sources of Air Pollution

Air pollution comes from a wide variety of sources. It can be categorized into local and transboundary or even stationary or mobile (Scoullos, 2013). Any sources that give destructive effects which give impact to the atmosphere pollution or the decline of the ecosystem is defined as air pollution (Azam et al., 2016). Based on the United States Environment Protection Agency's (EPA) definition of air pollution is the existence of contaminants are also can be called as pollutant substances in the atmosphere and are detrimental to the health, or produce other harmful environmental effects (Vallero, 2008).

The causes of air pollution are coming from harmful substances are introduced into the atmosphere. There are two main causes of air pollution into the atmosphere that are natural causes and also human intervention. Natural causes of air pollution mean that it may be caused by the change in temperature, regular cycle or also seasonal changes. For example, the dust can cause dust storm, wildfire that is a natural occurrence in a wooded area, animal digestion that are naturally can cause air pollution that comes from the methane releasing that would contribute greenhouse effect, volcanic eruptions that would produce large amounts of sulfur, chlorine, and ashes (Williams, 2016). Other than that, are resulting from human activity, for example the open burning, mining operations, burning of fossil fuels, and also exhausts from factories and industries.

Since the industry becomes the norm in the world, the air pollution from industrial processes becomes a major cause of air pollution. Basically, there are two types of pollutants that are primary pollutants and secondary pollutants. A direct result of the process from the pollutants can be called as primary pollutants. Sulfur dioxide that is emitted from factories can be considered as primary pollutant. While secondary pollutant is the intermingling and the reaction between the first pollutant. The simplest way to understand secondary pollutants is smog where it is created by the interactions of several primary pollutions.

The air pollution comes from various kinds of pollutants, including nitrogen oxides (NO_x), Ozone (O_3), sulphur dioxide (SO_2), carbon monoxide (CO), particulate matter (PM), and volatile organic compounds (VOCs) (Ritchie & Roser, 2017). Based on the World Health Organization (WHO), there are six major air pollutant that will harm the health and our ecosystem that is ground level Ozone, carbon monoxide (CO), Sulphur oxides (SO), Nitrogen oxides (NO) and also lead (Pb) (Azam et al., 2016).

Some pollutants can act as pre-cursors to others. Particulate matter (PM) compounds are an example where the interaction in the atmosphere between sulphur dioxide (SO_2) and nitrogen oxides (NO_x) (Ritchie & Roser, 2017). The reaction to sunlight or also can be called as the photochemical reaction of pollutants like nitrogen oxides and volatile organic compound will be formed the Ozone ground level (Brugha, Edmondson, & Davies, 2018). The Ozone ground level will give bad effect to health, unlike Ozone in the atmosphere that are naturally occurring to filter the ultraviolet ray that comes from the sun (Kjellstrom et al.). Based on the WHO, during sunny weather is the highest emission or production of Ozone pollution into the atmosphere (WHO, 2016). Figure 3 below shows the air pollution cycle.

2.6 Ozone (O_3) Process Formation and Its Theoretical Relationship with Precursor and The Meteorological Influence

Ozone (O_3) or also known as ground level Ozone is a natural occurring gas that has been reported as the most important pollutant for the air pollution in Malaysia (Shaadan, Nazeri, Jalani, Rahman, & Roslan, 2017). Ozone is formed caused by the

photochemical reactions between the precursor emission of volatile organic compounds (VOC) and also with the Nitrogen Oxides (NOx) (Banan, Latif, Juneng, & Ahamad, 2013). The emission of O_3, is mostly contributed by the chemistry of Nitrogen Oxides (NOx) especially in urban area during weekends (Sadanaga, Sengen, Takenaka, & Bandow, 2012). The meteorological influence also will give impact to the Ozone (O_3) formation where based on previous studies, the ground level Ozone concentration was peaking in the time between 1 p.m. and 3 p.m. where there are high in temperature (Awang, Elbayoumi, Ramli, & Yahaya, 2015). Ozone level variation of UVB, temperature, relative humidity, and wind speed or can be classified as meteorological variables that are higher than the Nitrogen Oxides (NOx), Carbon Monoxide (CO), Sulphur Dioxide (SO_2), PM_{10} or can be classified as a primary pollutant for Perai, Pasir Gudang, and Klang (Awang et al., 2015). Based on the previous study on Investigating a high Ozone episode in a rural mountain site, the main reason for the high concentration of Ozone at rural area coming from the transport of Ozone (O_3) and its precursor from the wind (Monteiro et al., 2012).

2.7 Technique to Assess Association- Air Quality

The strength and direction of the relationship between variables can be measured by measuring of association. The measure of association can be analyzed through many methods of analysis such as visualization approach, correlation analysis that are include Pearson correlation coefficient, Spearman rank-order correlation coefficient, chi-square test, relative risk and also odds ratio. In air quality, visualization approach is quite well used to show the association of the variables. A visual analysis is important which allow rapid processing and also a multi perspective exploration of the air pollution data to expose the basic relationship among variables and spatio-temporal patterns. For example, in previous studies of A Visual Analytics Approach for Station-Based Air Quality Data, three types of visual analysis approach that are used in the study are map-based views, calendar views and trends views (Du et al., 2017). In this study, the spatial distribution at the station and the situation of different areas are shown using map-based view. Next, in sight of the cyclical situation of air quality was showed using calendar view.

In another study of "A Visualization Approach to Air Pollution Data Exploration - A Case Study of Air Quality Index ($PM_{2.5}$) in Beijing, China", a basic visualization analysis and visual analytics was used to analysis the air pollution condition in the Beijing China (H. Li, Fan, & Mao, 2016). The visualization that has been used in the study are in tabular form, heat matrix, line chart, bar chart and others. While in the visual analytics, the visual analysis that use was heat map, parallel coordinates, geovisualization and others. Based on the visualization approach, the result showed that a correlation exists among the $PM_{2.5}$, PM_{10}, and NO_2, and wind speed using scatter plots. Other than that, a regular time pattern exists for air pollutants was shown using circle heat maps, and also a calendar view. Lastly, a geovisualization approach was used for illustrating the geographical distribution on air contaminant.

Next, in "Visualizing the intercity correlation of $PM_{2.5}$ time series in the Beijing-Tianjin-Hebei region using ground-based air quality monitoring data" study used time series and a visualization framework to visualize the intercity correlation of $PM_{2.5}$ (Liu, Li, Wu, & Liu, 2018). This visualization on this study has shown that there is correlation between $PM_{2.5}$ time series of different cities in all these regions and more significant in colder months. Moreover, this visualization from this study also has proved that, the meteorological variations are the cause the intercity correlations of $PM_{2.5}$ series.

2.8 Techniques and Model to Investigate Association Using Air Quality Data Set

The air quality association modeling in worldwide use many types of analysis to visualize the air quality condition. The analysis may be in ordinary least square method, multiple linear regression, neural network, principal component analysis and many more. Based on the previous study by Shaadan et al. (2017) in the journal title "Data Visualization Of Temporal Ozone Pollution Between Urban And Sub-Urban Locations In Selangor Malaysia", the variables that are used Ozone (O_3) exceedances in urban and sub-urban area. The objective of this study is to assess and visualize the occurrence of potential Ozone pollution severity in Banting and Shah Alam. This study used principal component analysis (PCA) as the method to estimate the Ozone

exceedances. The findings showed that there is an increase pattern of Ozone pollution occurrence with several modes of distinct diurnal variability in the location, and Ozone pollution was higher in Banting compared to Shah Alam.

Another previous study of Ozone pollution in Malaysia also used principal component analysis. The studies are about diurnal variety of ground-level Ozone in three port cities in Malaysia by Awang et al. (2015), used the Ozone concentrations, its precursor and meteorological parameters as the variables. The findings of this study showed that the concentration of Ozone in the three ports (Klang, Perai, and Pasir Gudang) was still below the maximum permissible values prescribed by the Malaysian Ambient Air Quality Guidelines (MAAQG). The second finding of this study was the highest average concentration of Ozone is in Klang. It also has found that the diurnal cycle of Ozone concentration highest at 1.00 p.m. to 3.00 p.m. Furthermore, meteorological conditions and prevailing levels of precursors (NOx) and CO is strongly influenced the diurnal pattern of Ozone concentration. The last finding was during January until May showed that the concentration of Nitrogen Oxide (NOx), Carbon Oxide (CO) and Ozone (O_3) were high at that time.

Another study from the journal of "Characterization of VOCs and their related atmospheric processes in a central Chinese city during severe pollution periods" used Ozone (O_3) concentration, volatile organic compounds (VOCs), Nitrogen Oxide (NOx) as the variables in this study (Li et al., 2019). The aim of this study was to characterize the volatile organic compounds (VOCs) at four site that had shown an increasing trend in Ozone concentration that are in Zhengzhou, Henan Province, China. In addition, the method that was used are chemical analysis, positive matrix factorization (PMF) and the potential source contribution function (PSCF). The results show that the Ozone formation was more sensitive to VOCs than NOx formation in Zhengzhou. Other than that, it was found that, vehicle exhaust, coal and biomass burning and solvent usage were the major sources of ambient VOCs at all four sites. Lastly, this study also shows that the strong emission from coal and biomass burning and solvent usage were concentrated in the southwest of Shanxi and Henan province.

In other studies by Monteiro et al. (2012) in their journal of "Investigating a high Ozone episode in a rural mountain site" has used an hourly concentration of

Ozone (O_3), Nitrogen Dioxide (NO_2), Sulphur Dioxide (SO_2), and PM_{10} as the variables. The aim of this was to investigate the origin of the very high O_3 concentrations that occurred in July 2005 in a rural mountain area at the Lamas d'Olo (LOL) monitoring site. There are two analyses that use in this study that are synoptic and backtrajectory analysis and also cross spectrum analysis. Based on this study, it was found that measured Ozone peaks at the LOL station in July 2005 were produced from the transport phenomena by winds and not produced locally.

A previous study of "Characteristics of Surface Ozone Concentrations at Stations with Different Backgrounds in the Malaysia Peninsula" by Banan et al. (2013) that aim to identify and describe the variations in Ozone concentrations recorded in Petaling Jaya, Putrajaya, and Jerantut monitoring backgrounds. This study also aimed to investigate the relationship between Ozone distribution and its association with Nitrogen Oxides (NO and NO_2) and non-methane hydrocarbon (NMHC). The interested variable in this study is Ozone concentration, Nitrogen Oxides, and non-methane hydrocarbon. Statistical analysis like normal P-P plot, normal Q-Q plot, one sample Kolmogorov-Smirnov test, analysis of Variance and Bonferroni is used as to analysis in this study. Other than that, a trajectory analysis was also used for backward trajectory analyses. It was found that suburban area that is Putrajaya showed the highest concentration of Ozone that was the influence by the characteristics of Nitrogen Oxides, particularly the titration of NO. Lastly, it was found that the surface of Ozone level was to be influenced by meteorological that are solar radiation and wind direction from the busy area especially in urban Kuala Lumpur.

2.9 The Application of Regression Analysis in Air Quality Modelling in Malaysia

Regression analysis in statistical analysis is a method or technique for estimating, measure the relationships among variables. Regression have many techniques for analyzing depends on the objectives and also the relationship between dependent variable and independent variables. In other words, regression is a method where, when the dependent variable changes when any one of the independent is varied, the other remaining independent variables are held fixed. There are many

models that can be classified into the regression such as linear regression, polynomial, general linear, logistic, multinomial and many more. In regression there are few of classical assumptions that need to be fulfilled such as the error is random variable, the independent variables are linearly independent, the errors are uncorrelated, there are homoscedasticity of variance error.

2.9.1 Importance of Outlier in Air Quality Data

Outlier is defined as the extreme value either extremely high or extremely low values where it is an abnormal distance compared to other points. Outlier data are considered as important to the air quality data as it will provide meaningful value to the environment. This will explain as for example outlier of Ozone (O_3) level will show the value where Ozone considered as harmful to health or welfare. Air quality data with outlier will show dynamic influence that has variation across the day. While the clean air quality data is associate at normal level, where it is not meaningful to the environmentalist since it gives the data at normal level that are not harmful to the health or welfare. There are many studies of outlier detection for air quality data. This show that outlier is important in air quality studies. In a study of by (Torresa, Nietob, Alejanoc, & Reyesc, 2011) entitle detection of outlier in gas emission from urban areas using functional data analysis shows that the outliers that comes from Carbon Oxide emission in 3 months are corresponded to low temperature during the whole day. Other than that, a study from (Martínez et al., 2014) is also using air quality data for outlier detection based on the Carbon Oxide (CO) emission, Nitrogen Dioxide (NO_2) emission and Sulphur Dioxide (SO_2) emission in the Northern Spain area. Thus, it is clear that outlier is important in the air quality data study because it will reveal the variety of data in a day which level is dangerous to the health and welfare.

2.10 Functional regression as a robust model

Nowadays, the modern statistics method that is functional data analysis method has become popular in its theory and also its application including the functional regression. However, in normal practice, the existence of outlier or also called as influential cases may affect the goodness of fit of a model that may cause misleading results. Based on previous study, a simulation data has used and resulted that the robust functional regression protects against outlier, reasonable efficiency, get a better result of estimation and well performed compared to normal practice (Kalogridis & Aelst, 2018). In another study, a simulation study also showed that the presence of outlier in the regression behavior can be capture effectively using robust functional linear regression model (Hullait, Leslie, Pavlidis, & King, 2019). Another prove showed by a simulation study using robust regression has given a good result as it is a good outlier resistance properties (Gervini, 2012). This showed that, functional regression is a robust model even with outlier dataset.

CHAPTER 3

METHODOLOGY

3.1 Introduction

This chapter will discuss and briefly explain about the methodology that will be used to conduct this study. This includes the source of data, description of variable, location of the study, validation tools, data analysis, and research framework. There will be two part of data analysis where the first one using pointwise regression and the second is using functional regression.

3.2 Study Location

In Malaysia, there are 52 monitoring stations to monitor the air quality in the atmosphere. Figure 3.1 shows the location of continuous air quality monitoring stations in peninsular Malaysia. The study areas selected are comprised of two air quality monitoring stations that are in Selangor and Penang which is under the supervision of the Department of Environment (DOE). These two station comprise from industrial area which is in Perai and Petaling Jaya. Figure 3.2 show map of Petaling Jaya and Perai, within the state of Selangor and Penang in Peninsular Malaysia that is interested in this research. The details of the stations used in this study was shown in the Table 3.1 below.

Petaling Jaya is the commercial and also an industrial hub of Malaysia where it is a highly populated area in Selangor. The Petaling Jaya monitoring station is located in a school area. More precisely, this station located at Sekolah Kebangsaan Bandar Utama, Petaling Jaya, Selangor (N03°06.612', E101°42.274'). Petaling Jaya also called as PJ is surrounded by the Malaysian capital that are Shah Alam capital of Selangor, Subang Jaya to the west, Puchong to the south, Kuala Lumpur to the east and Sungai buloh to the north.

The second monitoring station which is in Perai is also an industrial area place where it is near to Butterworth in the north and Perai river at the South. Besides that, Banting also well known as a port major contributors in the town's port facilities that are shipments of coal and scrap metal. The monitoring station of Perai is located at school area which is at Sekolah Kebangsaan Sebarang Jaya II, Perai (N05°23.890', E100°24.194').

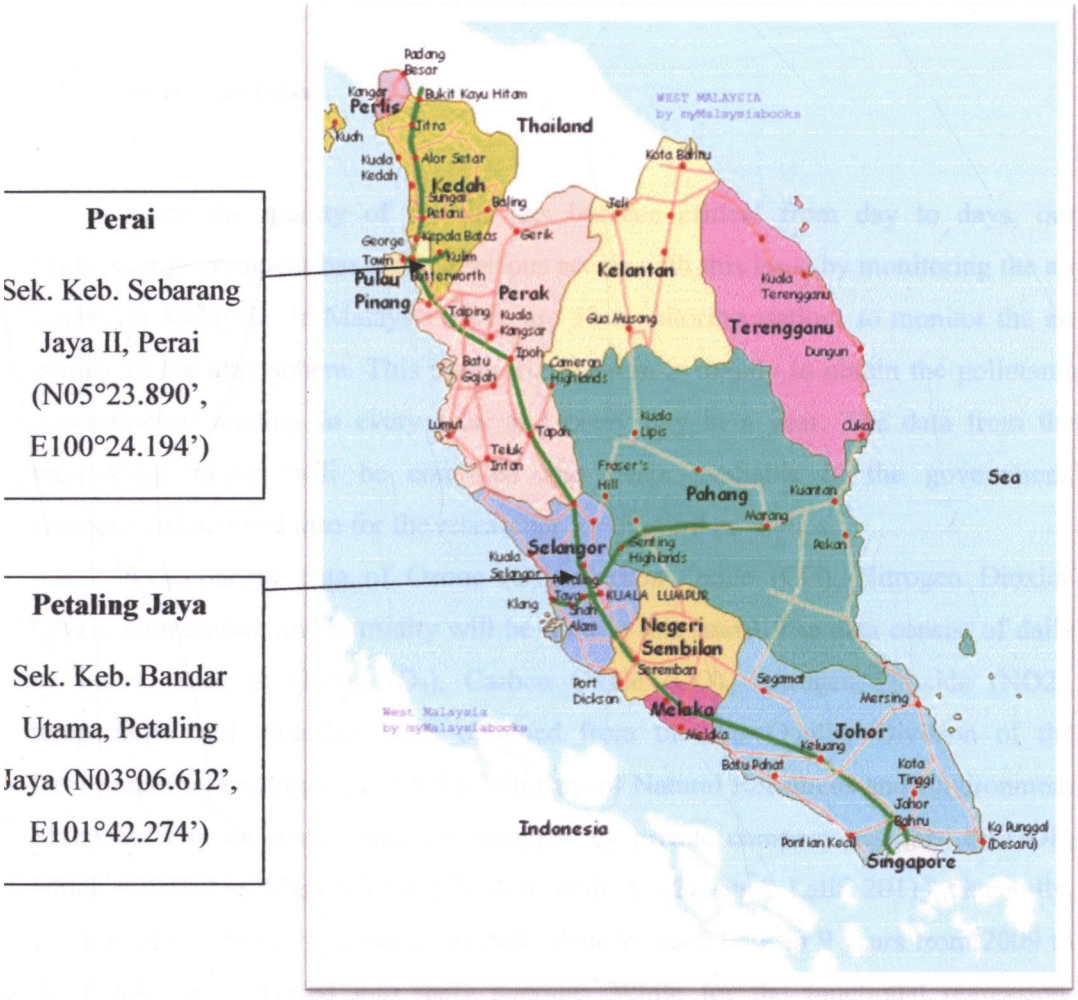


Figure 3.1 Location of Continuous Air Quality Monitoring Stations, Peninsular Malaysia

Table 3.1
Details of Air Quality Monitoring Stations

Station Name	Coordinate	Background
Petaling Jaya	(N03°06.612', E101°42.274')	Industrial area
Perai	(N05°23.890', E100°24.194')	Industrial area

3.3 Source of Data

Since the quality of the air has become critical from day to days, our Malaysian government has taken a serious action with this issue by monitoring the air quality in Malaysia. In Malaysia, there are 52 monitoring stations to monitor the air quality in the atmosphere. This monitoring station is mainly to obtain the pollutants concentration reading in every hour and every day in a year. The data from the monitoring station will be compiled and make available to the government, environmentalist and also for the researcher for the further studies.

A secondary data of Ozone (O_3), Carbon Oxide (CO), Nitrogen Dioxide (NO_2), temperature and humidity will be used in this study. The data consist of daily by hourly data of Ozone (O_3), Carbon Oxide (CO), Nitrogen Dioxide (NO2), temperature and humidity were obtained from the Air Quality Division of the Department of Environment (DOE), Ministry of Natural Resources and Environment Malaysia. The air quality data are managed by private company assigned by DOE, which is called as Alam Sekitar Sdn Bhd (ASMA) (Banan & Latif, 2011). Generally, for normal pointwise regression, the daily data for each hour in 9 years from 2009 to 2017 will be converted into daily average. While for the functional regression, temporal data on each 24 hourly points of each day for 9 years' data will be used.

Based on the (Mohammed, Ramli, & Yahya, 2013) a UV Absorption Ozone Analyzer Model 400A is used for collecting the samples of Ozone concentrations by measuring the low ranges of Ozone in the atmosphere. The analyzer utilized a system based on the Beer-Lambert law, which is the model 400A UV Absorption Ozone Analyzer microproprocess or-controlled analyze for measuring low ranges of Ozone

concentration in the ambient air (Mohammed et al., 2013). The hourly nonmethane hydrocarbon (NMHC) concentration was measured by a flame ionization detector which is called as Teledyne Model 4020 (Ahamad et al., 2014). Other than O3, NOx and VOC, data of temperature (°C) are also included in this research. The measurement of temperature was used the Met One 062 sensor (Ahamad et al., 2014).

In Malaysia, hourly air pollutant index (API) is systematically reported the current status of air pollution from air quality monitoring system all over the country. The API readings need a one-hour complete cycle before can be retrieved. The data that will be obtained will follow the standard quality control processes. The procedures that are used for the continuous monitoring data are followed in accordance with the standard as outlined by internationally recognized environmental organizations such as the United State Environmental Protection Agency (USEPA) (Latif et al., 2014). The instruments used to monitor the changes of ambient of NO2 are Teledyne API Model 200A/200E (Latif et al., 2014). This instrument applies the chemiluminescence detection principles are used for the detection of NO2 in the ambient of air. The sensor of the chemiluminescence will provide stability and ease of use needed for ambient air. Volatile organic compound (VOC) will be monitored using a direct air quality (AQ) meter and also a Photo Ionization Detector (PID) (Madhoun, Ramli, & Yahaya, 2012). This AQ meter and PID can detect potential air quality problem before it becomes worse and it also has high accuracy respond to complaints. Meanwhile, Teledyne API Model 100A/100E and Teledyne API Model 300/300E was used to monitor Carbon Oxide (CO). Lastly, for meteorological variables that is temperature, a device called Met One 062 sensor was used to monitor the concentration level of each variable meawhile Met One 083D sensor to monitor the concentration level of humidity (Awang et al., 2015).

3.4 Methodological Framework

The study that will be conducted will follow several stages that is important as shown in the conceptual framework Figure 3.2. The study starts with data acquisition of also called as data collected from the department of environment Malaysia (DOE), data management in which the data will undergo cleaning process, data preparation where the data will be rearranged based on the analysis. Lastly the data analysis.

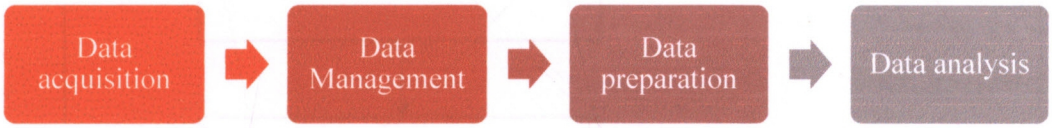


Figure 3.2 Research Framework

The first step is the data arrangement and data management of the observed discrete data. The observed discrete data will be converted into daily average. The second step is the correlation analysis where we will determine which independent variable that will associate with the dependent variable before can proceed with the modeling. Third step is the modeling phase where functional linear regression will be done. Figure 3.3 shows the research analysis that will be carried out to answer the objectives in the study for functional regression. The first step is the data conversion which this step enables the discrete data to be converted thus the data will be in the form of the function. Second step is the functional descriptive analysis where it is the method for descriptive analysis in functional analysis. Step three is modeling the association using functional regression method.

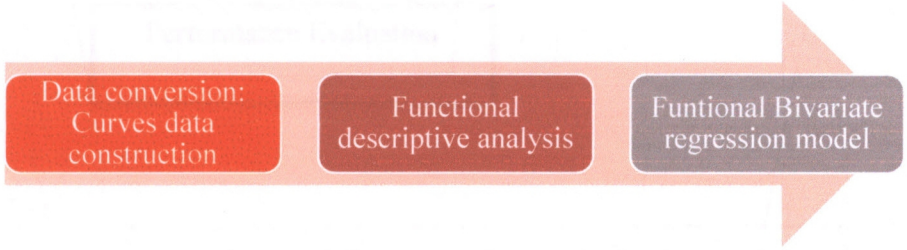


Figure 3.3 Steps of Functional Data Analysis

The full conceptual research framework has shown at Figure 3.4 in more details. The step by step from the data collection method is described in the Figure below.

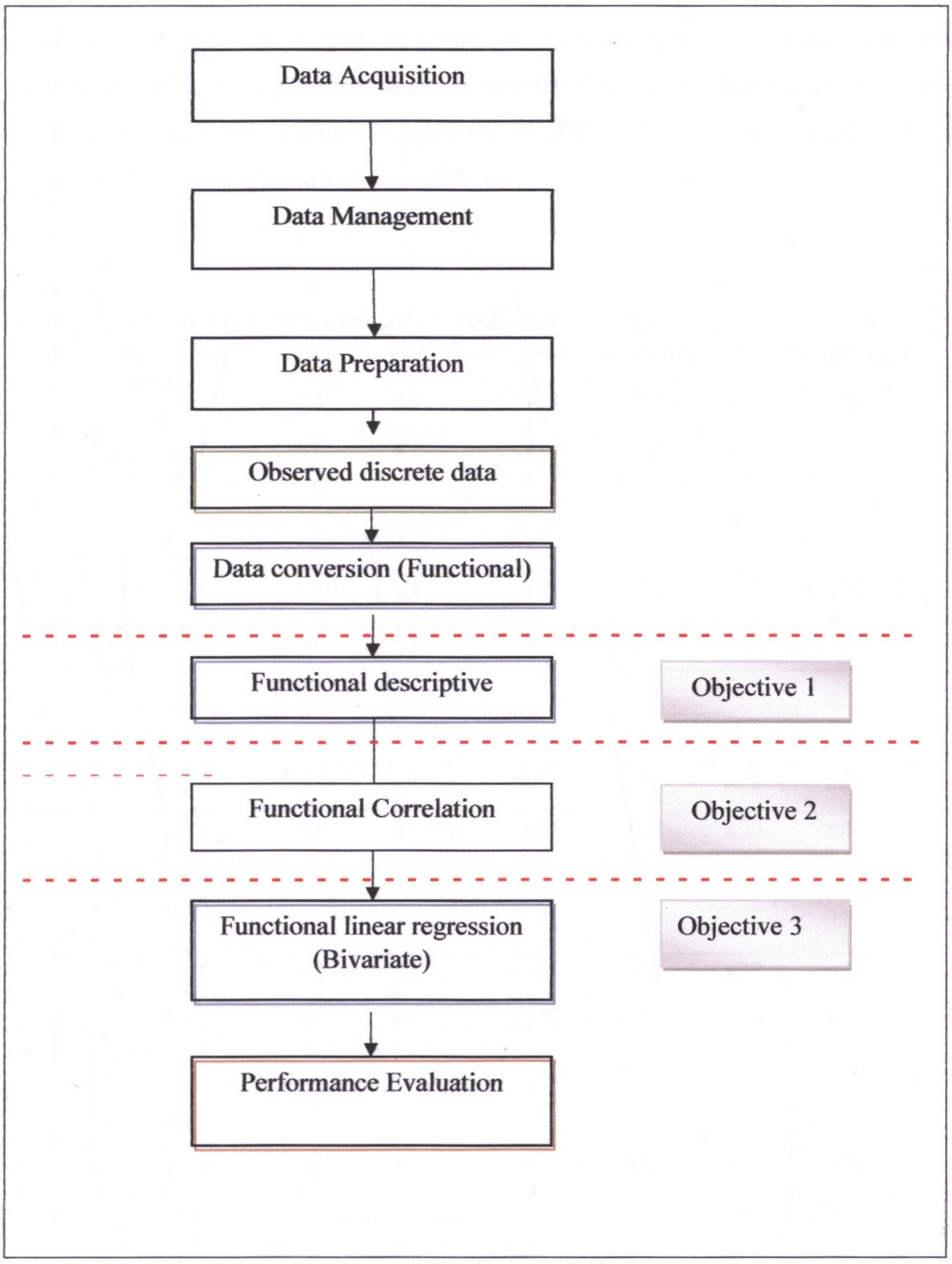


Figure 3.4 Conceptual Framework

3.5 Data Analysis

3.5.1 Data Arrangement

Firstly, data would be obtained from the Department of Environment Malaysia (DOE). After receiving the data, an arrangement according to variable interested and with respect to the objective will be done. The data then will be reorganized according to the daily hourly data within a 9-year period (2009-2017). The representation of the observed data that will be used is as follows:

Table 3.2
Example the Arrangement of Data for a Particular Pollutant

Year	Day	Hour1	Hour2	Hour3	...	Hour21	Hour22	Hour23	Hour24
2015	1	x1	x2	x3	...	x21	x22	x23	x24
	2	x1	x2	x3	...	x21	x22	x23	x24
	. .								
	365	x1	x2	x3	...	x21	x22	x23	x24
2016	1	x1	x2	x3	...	x21	x22	x23	x24
	. .								
	365	x1	x2	x3	...	x21	x22	x23	x24
2017	1	x1	x2	x3	...	x21	x22	x23	x24
	. .								
	365	x1	x2	x3	...	x21	x22	x23	x24
2018	1	x1	x2	x3	...	x21	x22	x23	x24
	. .								
	365	x1	x2	x3	...	x21	x22	x23	x24

Data used for Functional Modelling

3.5.2 Data Quality

In order to maintain the quality of the data, a data cleaning process will be done after reorganizing data in the previous step. Data cleaning process will include the checking on the missing values, checking on the existence outlier and many more. If there is any missing value problem is detected, a suitable imputation will be used for the treatment method. Imputation is a common method used to assign values to the missing items. There are six types of imputation that is commonly used which are deductive imputation, cell mean imputation, hot deck imputation, regression imputation, cold deck imputation, and multiple imputation. Imputation method will reduce the bias due to item nonresponse. A multiple imputation is suitable for this research since it this method is suitable for complex incomplete data problems in which missing values occurs in more than one variable (Febrero-Bande & Fuente, 2012). Multiple imputation will replace the missing value with a set of plausible values that have natural variability and uncertainty of right values (Kang, 2013). The multiple imputation started with the existing data from other variables will be used as a prediction for the missing data (Sinharay & Russell, 2001).

A full data set that called the imputed data set will be created by replace the missing values using the predicted values. The term of multiple imputation is generated from the repeatability process of iteration to make it multiple imputed data set. Then, the multiple imputed data set will be analyzed using standard statistical analysis procedure for a complete data set to give multiple analysis results. Afterwards, the result analysis will be combined together to produce a single overall analysis result. It is clear that there are important benefits from multiple imputation where a valid statistical inference will be produce that are reflecting the uncertainty related to the estimation of missing data, and also it will come appropriate results either in small sample size or high number of missing values since multiple imputation is robust to the violation of the normality assumptions (Kang, 2013).

A package that called mice package was used for this study. Mice package helps to deal with missing value issue. This mice package solves the missing values problem by creating multiple imputation for multivariate missing data as portrayed in Table 3.2. The name mice package comes from multivariate imputations by chained

equations. This imputation package gives more advantageous compared to the other imputation package which is mixes continuous, binary, unordered categorical and ordered categorical can be impute using MICE algorithm. MICE is preferable since its properties that are can be used in broad setting and very flexible in nature. MICE algorithm is a multiple imputation that will create multiple prediction for each missing values. The uncertainty imputations also will be take into the account as it is important and then will generate accurate standard errors. if there is least information provided in the observed data cause from the missing values, the imputations will be very variable, that will lead to high standard errors in the analyses. Meanwhile, if the observed data are many information provided, the imputations will be more precise and consistent along imputations, that will result in smaller and accurate, standard errors (Greenland and Finkle, 1995). (Azur, Stuart, Frangakis, & Leaf, 2011). While, for removing the outliers, a package called MVN helps to perform multivariate normality test, graphical approaches and implements multivariate outlier detection and univariate normality of marginal distributions through plots and test (Korkmaz, Goksuluk, & Zararsiz, 2014). This package comes with two multivariate outlier detection methods that are based on robust Mahalanobis distance. Mahalanobis distance is a measure that calculates how far each observation to the joint distribution center, that can be thought of as the centroid in multivariate space. Meanwhile, robust distance is determined from the minimum covariance determinant estimators. This MVN functions will return up a new data set where the outliers are removed.

3.5.3 Functional method

Figure 3.5 below shows the sub-sections of functional data analysis, which are functional descriptive, and functional regression.

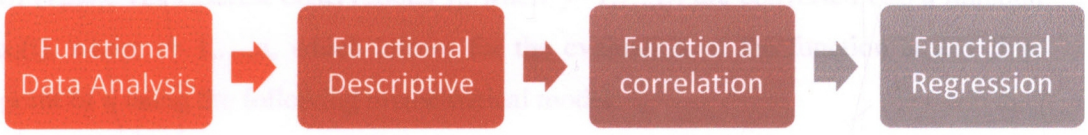


Figure 3.5 Sub-Section of Functional Data Analysis

i. Functional Data Analysis

Functional Data Analysis (FDA) is a studied branch of statistics that analyze curves data. The observed discrete data of time points can be shown in functions to represent information from a collection of all the functions. Analysis and theory that are in type of functions, images and shapes, or more general objects are called as functional data analysis (FDA) (Wang, Chiou, & M¨uller, 2015). Based on previous studies, functional data analysis (FDA) is a statistical method to analyze curves or functional data (Shaadan, Jemain, & Deni, 2014). According to Ullah and Finch (2013), the usage of functional data analysis has been increasing for better analyze, model and predict the time series data. The first step in functional data analysis (FDA) is to covert the discrete observed data into a functional data or curve data. This method is called data conversion in which for smoothing the data. The objective of data conversion is to recover the underlining curve for the discrete recorded data. There are two types for data conversion either using regression smoothing or roughness penalty smoothing approach. The regression and roughness penalty smoothing approach are most popular smoothing methods by means of basis expansion. Basis function expansion is determined by combining a set of coefficients and a k number of suitable basis functions. The easiest way to convert data is by using data interpolation, but it will result to a rough curve that will affect a limitation in the usage. (Shaadan et al., 2014). Thus the smooth curve is much more preferable as it can provide more information and analysis.

Spline and Fourier basis are the most popular in the bases. B-spline smoothing is much more preferable due to its flexibility properties. The choice of smoothing technique is dependent upon the underlying behavior of the data being analyzed (Ramsay, Hooker, & Graves, 2009). Generally, the smoother should reflect the features that match with the data. Based on previous studies by Shaadan, Deni, and Jemain (2015) stated that a daily curve is actually a function defined over an interval of (1,24). The discrete observations y_j, where j= 1,…,24 are converted into a function $x_i(t)$, where i= 1,…,n, which allows for the evaluation of the function at any time point of t_j using the following mathematical model:

$$y = x(t) + \varepsilon \tag{3.1}$$

Where notation ε is called as random error or random noise in the data. The estimation of daily curves represented by function $x_j(t)$ is conducted by means of a system of basis function expansion, which is a linear combination of K independent basis function $\varphi_k(t)$ whereas the term β_k refers to the basic coefficient as follows:

$$x_i(t) = \sum \beta_k \varphi_k(t) \tag{3.2}$$

There are many types of basis function such as spline, Fourier, quadratic, polynomial, constant and many more. The basis that will be chosen are based on the underlying pattern of data. Generally, periodic data are much more suitable for Fourier basis while the spline is used for non-periodic data. In this study, a linear combination of a K number of B-spline basis is used to represent the curve for a more flexible fitting. Splines are polynomial segments that joined end to end. Knots are the name of the points at which the segments are joining (Hooker, 2017). Splines are piecewise polynomial, therefore, to define a spline basis, a set of knots information or the K number of basis and the degree of the polynomial is needed. Before a daily curve can be constructed, a degree of polynomial is used in this study. The coefficients β_k will be determined using the least square method of minimizing the sum of squared residuals (SSE) as follows:

$$SSE = \sum_{j=1}^{24} (y_j - x(t_j))^2 = \sum_{j=1}^{24} (y_j - \sum_{k=1}^{K} \beta_k \varphi_k(t_j))^2 \tag{3.3}$$

Regression smoothing is a method for smoothing a curve using the least square method (Ramsay et al., 2009). Another alternative way for curve conversion is the roughness penalty approach. Appropriate K is estimated to minimize an appropriate criterion function (Huang & Shen, 2004). The Bayesian Information Criteria (BIC) is used to be an indicator for choosing the best appropriate k in this study. Let m denotes the number of recorded data points (in this case m=24), K is the

number of basis and SSE is the sum of the square or error variance of the estimated mean curve. The BIC formula is given by:

$$BIC = \log\left(\frac{SSE}{m}\right) + \frac{k}{m}\log(m) \tag{3.4}$$

ii. **Functional Descriptive Analysis**

After conversion data, functional descriptive analysis method will help in determining the functional mean, functional median and functional standard deviation. Functional mean gives the definition of the sum of the portion level at each time period divide by point hours. Functional median x_i at each time form.

$$Functional\ mean\left(\bar{x}(t)\right) = N^{-1}\sum_{i=1}^{N}[x_i(t) - \bar{x}(t)]^2 \tag{3.5}$$

$$Functional\ median = \sum median\ of\ x_i(t) \tag{3.6}$$

$$Standard\ deviation = \sqrt{\left[(N-1)^{-1}\sum_{i=1}^{N}[x_i(t) - \bar{x}(t)]\right]^2} \tag{3.7}$$

iii. Cross Correlation Functions

Cross correlation function is a measure of association or also can be called as a relationship between two variables. In this study, cross correlation is used to determine the association between Ozone and the variable (Carbon Oxide (CO), Nitrogen Dioxide (NO2), temperature and humidity). Based on the Functional Data Analysis book by Ramsay and Silverman (2006) the cross correlation formula are shown below:

$$cov_{X,Y}(t_1, t_2) = (N-1)^{-1} \sum_{i=1}^{N} \{x_i(t_1) - \bar{x}(t_1)\}\{y_i(t_2) - \bar{y}(t_2)\} \quad (3.8)$$

$$corr_{X,Y}(t_1, t_2) = \frac{cov_{X,Y}(t_1, t_2)}{\sqrt{var_X(t_1) var_Y(t_2)}} \quad (3.9)$$

iv. Functional Regression

Functional linear models (FLM) are defined as the model construction that describe the relationship between a dependent variable and independent variable (Ullah & Finch, 2013). The approach of functional regression based on whether the responses or covariates are vector data or functional data that have three types which is (i) functional responses with functional covariates, (ii) vector responses with functional covariates and lastly (iii) functional response with vector covariates (Wang et al., 2015). The assumption that need to be fulfilled are the ε_i are assumed to be independently and identically distributed (M¨uller & Stadtm¨uller, 2005). The classic linear regression models are:

$$y_i = \propto + \sum_{j=0}^{p} x_{ij}\beta_j + \varepsilon_i, \qquad i = 1, \ldots, N \quad (3.10)$$

Steps to Be Taken in Order to Conduct the Functional Regression

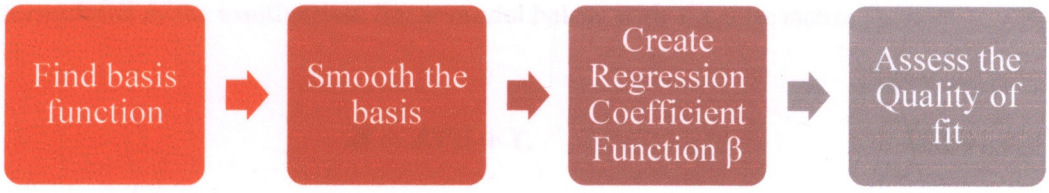


Figure 3.6: Steps to be Taken in the Functional Regression

A scalar response and a functional independent variable

Assume that $y_i, i = 1, \ldots, n$ are the scalar responses, and $x_{ij}(t), j = 1, \ldots p$ are the functional regressors. Let the functional regression model be:

$$y_i = \propto + \int \beta(t)x_i(t)dt + \varepsilon_i \tag{3.11}$$

Where constant $\propto$ is the usual intercept term that adjusts for the origin of the precipitation variable (Ramsay & Silverman, 2006). The regression coefficient function β is basically for the functional parameter of interest. The $\varepsilon_1, \ldots, \varepsilon_n$ are the unobservable independent random errors, that is assumed to be normally distributed with zero expectation and has unknown variance σ^2 (Smaga & Matsui, 2018). In general, the functional data model is the multivariate model with the sums replaced by the integrals.

Model Parameter Estimation

Assuming that X and Y belong to finite dimension spaces spanned by two basis $\{\vartheta_p : p = 1, \ldots, P\}$ and $\{\phi_q : q = 1, \ldots, Q\}$,

$$x_w(t) = \sum_{p-1}^{P} a_{wp}\,\vartheta_p(t) \quad (3.12) \qquad y_w(s) = \sum_{q=1}^{Q} b_{wq}\,\phi_q(s), \quad (3.13)$$

The parameter function is $\beta(t,s) = \sum_{p=1}^{P}\sum_{q=1}^{Q}\beta_{pq}\,\vartheta_p(t)\phi_q(s)$. Equation 3.11 can be formulated as the multivariate linear model below with Υ a noise matrix

$$\boldsymbol{B} = \boldsymbol{A}\Psi\beta + \Upsilon, \tag{3.14}$$

$$\boldsymbol{B} = (b_{wq})_{n\times Q,}\boldsymbol{A} = (a_{wp})_{n\times P,}\Psi = (< \vartheta_p, \vartheta_{p'} > L^2(t))_{P\times P,} \tag{3.15}$$

Where the Least squares estimation is

$$\hat{\beta} = \left((A\Psi)'(A\Psi)\right)^{-1}(A\Psi)'B \tag{3.16}$$

Assessing Goodness of Fit in Functional Regression

i. Squared correlation function (R^2)

Squared correlation function or also called as R-squared (R^2) is a statistical measure to determine the proportion that can be converted to a percentage of the variance for dependent variable that is explained by an independent variable. This method comes from conventional method to determine the R^2.

$$R^2(t) = 1 - \sum_i \{\hat{y}_i(t) - y_i(t)\}^2 / \sum_i \{y_i(t) - \bar{y}(t)\}^2 \tag{3.17}$$

A complementary method to determine the goodness of fit is by considering each individual functional of R^2 that define by Ramsay and Silverman (2006) is

$$R_i^2 = 1 - \int \{\hat{y}_i(t) - y_i(t)\}^2 / \int \{y_i(t) - \bar{y}(t)\}^2 dt \tag{3.18}$$

ii. F-Ratio

Another method for assessing goodness of fit is F-Ratio that comes from F-distribution. F-Ratio approach is used only as an approximation to the F distribution for FRatio(t) for each *t*. A F-Ratio function is constructed by using formula:

$$FRatio(t) = \frac{\sum_i \{\hat{y}_i(t) - \bar{y}(t)\}^2/(K_0 - 1)}{\sum_i \{y_i(t) - \hat{y}_i(t)\}^2/(n - K_0)} \quad (3.19)$$

with degrees of freedom, v1= k-1 and v2=n-k. Where n= is the no observations and k is the number of basis.

3.6 Further Investigation Using the Influence of Outliers On the Model

An additional investigation was also conducted to assess whether outliers in the data set could influence Functional Regfression model obtained. To conduct the analysis, another data set without outlier were used. The outliers were detected using Cooks Distance.

3.6.1 Cooks Distance

There are many statistical technique or method is used to detect outliers in a dataset. In this study, Cook's Distance is used for detect influence of multivariate outlier in the dataset. The Cook's Distance is based from R.D. Cook. Outlier in the dataset can disturb the efficiency and accuracy of a result. Sometimes, treating or removing the influencing outlier is important before we can proceed with the analysis. The cook's distance is determined by the effect of each row in the data point on the outcome of the predicted value. Basically, to determine the influential outlier in the cook's distance is by determine the observation that have the cook's distance value bigger than mean values by 4 times. A cook's distance is constructed by using formula:

$$D_i = \frac{\sum_{j=1}^{n} (\hat{Y}_j - \hat{Y}_{j(i)})^2}{p \times MSE} \quad (3.20)$$

Where,

- $\hat{Y}_j$ is the value of j_{th} fitted response when all the observations are included.
- $\hat{Y}_{j(i)}$ is the value of jth fitted response, where the fit does not include observation i.
- MSE is the mean squared error
- p is the number of coefficients in the model

3.7 Software

In order to perform the data analysis in Chapter Four, a software named RStudio has been utilized. The RStudio software is a free software environment for statistical computing and graphics which compiles and run on a wide variety of UNIX platforms, Windows and MacOS. RStudio is an enhancement of R software that are more convenient and interactive. This RStudio software has a one site for console, one site for viewing history, also included syntax-highlighting, one site for editor that can be directed to code execution, many types of tool for plotting, debugging button and the best thing about RStudio is that it is much easier for managing the workspace. The R language is widely used among researchers, statisticians and data miners for developing statistical software and data analysis. Furthermore, R software is a data analysis software for data scientist, statistician, analysts and others who need to make sense of data use R for statistical analysis, data visualization and predictive modeling.

3.8 Summary of Analysis

There are three analyses that will be done in this study in order to achieve the research objectives. Functional descriptive analysis will be used to determine the first objective that is to describe the diurnal and spatial behavior of Ozone (O_3), the precursors (CO, NO_2) and meteorological variables (temperature and humidity) at industrial sites in peninsular Malaysia. In this analysis, fda package will be used to determine the functional median, functional standard deviation and functional mean. The second analysis that include for this study is functional correlation where to investigate the diurnal inter - association pattern between O_3 and the precursors as well as meteorological variables. Cor.fd function will be used in this analysis. The last analysis that is included in this study is functional regression analysis to achieve the third research objective that is to model the diurnal relationship between Ozone and the precursors (CO and NO_2) as well as the meteorological variables (temperature and humidity) using Bivariate Functional Linear Regression. In this analysis fregress function will be used that comes from the fda package. This analysis will use the F-ratio and R squared multiple correlation.

CHAPTER 4
FINDING AND ANALYSIS

4.1 Introduction

In this chapter, it will comprise of several sections such as data arrangement and cleaning. This will follow from functional data analysis, including the dynamics of overall Ozone curves, the mean, median and standard deviation, the correlation and the functional regression analysis.

4.2 Data Processing and Exploration

Data that was collected from the Department of Environment (DOE) Malaysia was reorganized suitably according to the objectives of the study. In statistics, missing data or also can be called as missing values will be occurring when there is no value stored in the observation of interest. It always arises in almost all serious statistical analysis. The existence of missing values will give many problems such as reducing statistical power, cause bias in the estimation of parameters, reduce the representativeness of samples and complicate the analysis of the study (Kang, 2013). Since there are missing values presence for our data, an imputation method was applied to overcome this problem by using the imputation method. Imputation method is used by fill in or substitute values rather than removing the variables or observations. This method gives advantages where it will keep the full size of sample size and maintain the precision and avoid biases. In this study, multivariate imputation by a chained equations method that also called as MICE is used since it is very useful for large imputation procedures. Based on (Rubin, 1987) multiple imputation method is suitable for complex, incomplete data problem. There will be a special challenge and effort when the missing values occur in more than one variable. The MICE package in R helps to impute the missing values with a set of plausible data values that are drawn from a distribution that are specifically designed for each missing data

point. The percentage of missing value and descriptive analysis are summarized in the Table 4.1 below. Descriptive statistic is a categorized as measures of central tendency and measure of variability spread in which mean, median, mode, standard deviation, variance, kurtosis and skewness are included.

The results indicate that the maximum Ozone level was observed in Petaling Jaya with 0.216. Thus, Petaling Jaya is the most severe Ozone pollution and most severe alarming station compared to Perai. This means that Petaling Jaya will need a comprehensive study by researchers in the future. Meanwhile, both the station has recorded the minimum Ozone level at zero level. Based on the Petaling Jaya and Perai station, the highest percentage number of missing values is located in Perai station with 12.71%. The missing values need to be imputed to replace the missing value using multiple imputation. Both stations give same mean which is 0.015 where this value is the average or the central position in the dataset that used to derive the central tendency.

Table 4.1:
Summary of Ozone Data and the Percentage of Missing Values

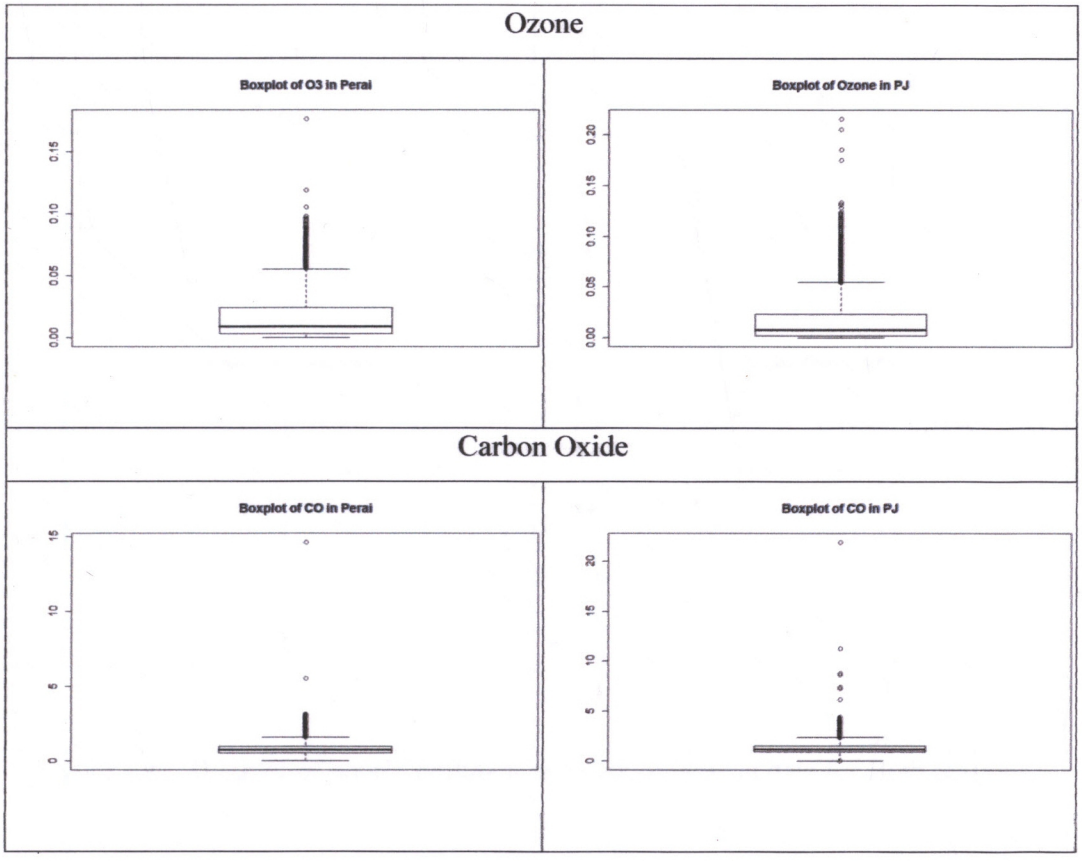
Locations	Percentage of missing (%)	Descriptive Analysis				
		Mean	Standard Deviation	Maximum	Minimum	Median
Petaling Jaya	9.64	0.015	0.0177	0.216	≈(0.000)	0.008
Perai	12.71	0.015	0.0165	0.176	≈(0.000)	0.009

Note: (≈) the value is too small (i.e, near zero)

4.2.1 Boxplot of Hourly Pollutant and Meteorological Data for Both Stations

As seen in Figures 4.1, it shows the summary of the boxplot of hourly pollutant and meteorological data for Perai and Petaling Jaya station. Generally, a boxplot or also called as box and whisker plot is a plot where to determine the spread that are interquatile range of a dataset and also to determine the centers of a dataset which is the median of a dataset. The minimum and maximum also can be showed from the boxplot. Overall, all variables have similar pattern when compared to the same variable in the both stations. Both stations tend to have similar range but a

slightly different range at upper whisker, lower whisker, first quartile, median and third quartile. For boxplot of Ozone (O_3), the upper whisker shows the maximum of Ozone in Perai is slightly higher than the maximum in the Petaling Jaya station. Both stations of Ozone (O_3) variable show quite similar pattern of boxplot where the lower and the upper whisker shows at similar range. The median of Ozone (O_3) in Perai also shows slightly higher compared to the Petaling Jaya station. The boxplot of Carbon Oxide (C0) shows very short box compared to the other variables. This means that the spread of the Carbon Oxide (CO) variable is very small where the minimum, first quartile median, third quartile and maximum is close to each other. Things differently to boxplot of Nitrogen Dioxide (NO_2) variable where the maximum at the upper whisker for Petaling Jaya shows a higher value compared to the Perai station. Next, the boxplot of temperature where the boxplot is comparatively short in the Petaling Jaya station compared to Perai station. Lastly, in the humidity boxplot it shows that Perai station has more outlier compared Petaling Jaya station with similar maximum and minimum level.



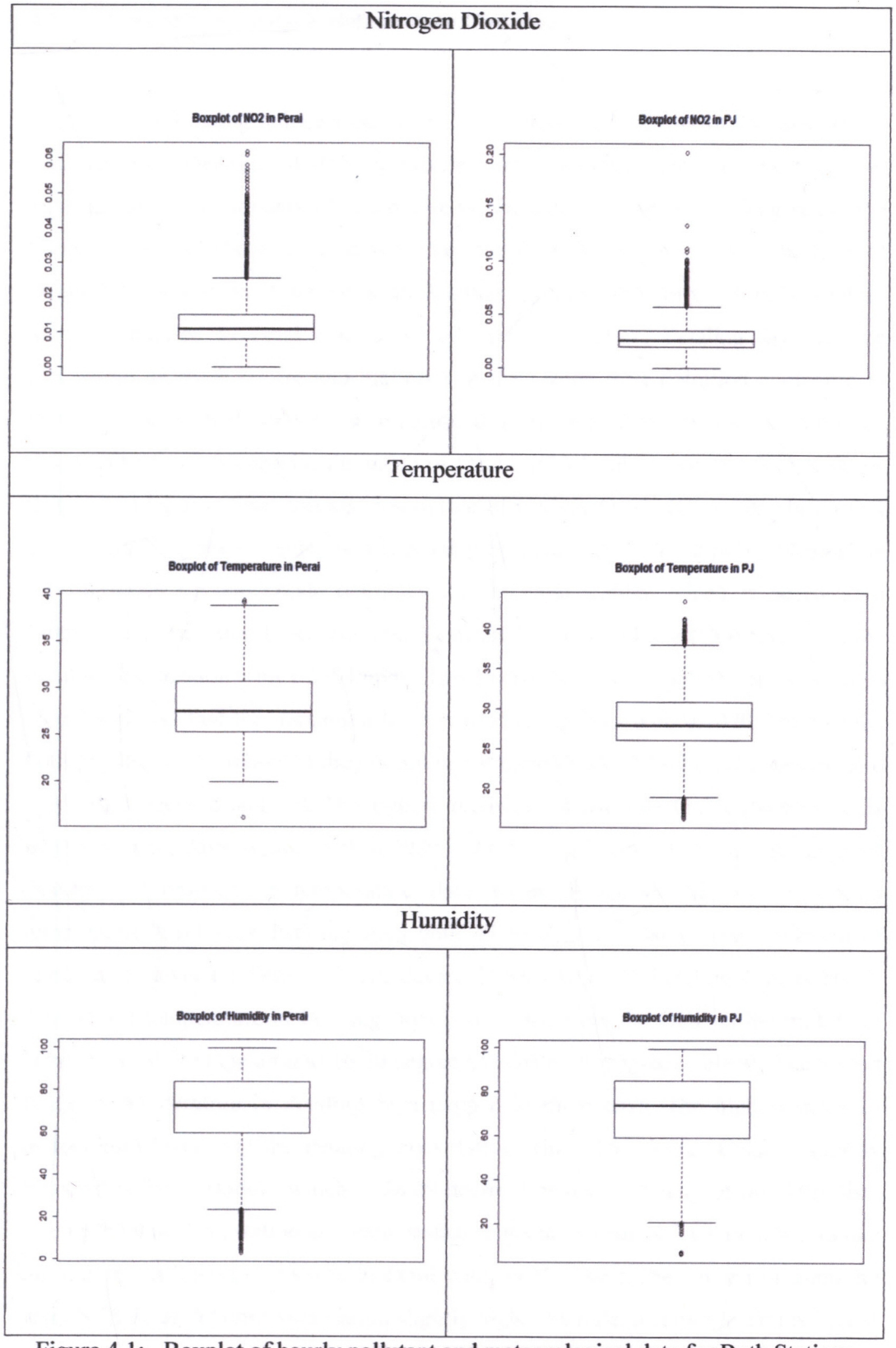


Figure 4.1: Boxplot of hourly pollutant and meteorological data for Both Stations

4.3 Descriptive Analysis Before Data Cleaning

The following sub-sections will discuss about the results of the descriptive statistics after cleaning of data by multiple imputations in order to substitute the missing values. A summary of descriptive Ozone data for both stations is given in the Table 4.2 below. The result of descriptive after data cleaning shows that the level of Ozone (O_3) was showed maximum in Petaling Jaya station with 0.216 ppbv slightly higher compared to Perai that is 0.1763 ppbv. Therefore, Petaling Jaya can be considered as more severe and alarming in the Ozone (O_3) pollution compared to Perai. While the both stations have recorded nearly zero (≈0.000) for the minimum Ozone (O_3) level. Moreover, the mean of Ozone (O_3) level is same for both stations that is 0.015 ppbv. Furthermore, descriptive of Carbon Oxide (CO) data shows that the maximum Carbon Oxide level is at Petaling Jaya with 21.9510 ppbv. Meanwhile the minimum Carbon Oxide (CO) level is in Perai station, which is nearly zero (≈0.000) concentration level. The highest mean of Carbon Oxide (CO) level is at the Petaling Jaya station with 1.2394 ppbv. The descriptive of Nitrogen Oxide (NOx) data where it shows that the maximum level is in Petaling Jaya with 0.2020. Meanwhile, both the station has recorded the minimum Nitrogen Oxide (NOx) level to nearly zero (≈0.000) concentration level. The highest maximum of Nitrogen Oxide (NOx) level is in the Petaling Jaya station with 0.2020 ppbv level of concentration. The range of descriptive statistics of temperature data where it shows that the maximum temperature level is at Petaling Jaya with 43.50 degree Celsius. The minimum of temperature level for Perai is 16.10 degree Celsius while in Petaling Jaya is 16.30. The lowest temperature in Petaling Jaya is recorded from the DOE in the middle of November 2009 range around 16.30 degree Celsius to 20 degrees Celsius. The lowest range of temperature in Petaling Jaya happen in three days. The highest mean of temperature level is at the Petaling Jaya station with 28.56 degree Celsius, slightly higher than Perai station, which is 28.09 degree Celsius. Summary of humidity data for the Petaling Jaya station and Perai station is given in Table 4.2 where it shows that the maximum humidity level is in Perai with 99.70. The highest mean of humidity level is 71.78 at Petaling Jaya station slightly higher than Perai station which is 71.13.

Table 4.2:
Summary of Descriptive Statistics from original data (before imputation for complete data set)

Variables	Descriptive Statistics	Location	
		Perai	Petaling Jaya
Ozone (O_3)	Mean	0.01500	0.01500
	Standard Deviation	0.01648	0.01772
	Maximum	0.17630	0.216
	Minimum	≈(0.0000)	≈(0.0000)
	Median	0.0090	0.00800
Carbon Oxide (CO)	Mean	0.76700	1.24300
	Standard Deviation	0.33310	0.48199
	Maximum	14.6340	21.9510
	Minimum	≈(0.0000)	0.01600
	Median	0.73200	1.18900
Nitrogen Dioxide (NO_2)	Mean	0.01200	0.02800
	Standard Deviation	0.00681	0.01226
	Maximum	0.06200	0.20200
	Minimum	≈(0.0000)	≈(0.0000)
	Median	0.01100	0.02600
Temperature	Mean	28.09	28.56
	Standard Deviation	3.3436	3.354
	Maximum	39.40	43.50
	Minimum	16.10	16.30
	Median	27.50	28.03
Humidity	Mean	71.20	71.78
	Standard Deviation	15.40	15.76
	Maximum	99.70	99.60
	Minimum	2.90	6.30
	Median	72.60	73.93

Note: (≈) the value is too small (i.e, near zero)

4.4 Descriptive Analysis After Data Cleaning

The following sub-sections will discuss about the results of the descriptive statistics after cleaning of data by multiple imputations in order to substitute the missing values. A summary of descriptive Ozone data for both stations is given in the Table 4.3 below. The result of descriptive after data cleaning shows that the level of Ozone (O_3) was showed maximum in Petaling Jaya station with 0.2156 ppbv slightly higher compared to Perai that is 0.1763 ppbv. Therefore, Petaling Jaya can be considered as more severe and alarming in the Ozone (O_3) pollution compared to Perai. While the both stations have recorded nearly zero (≈0.000) for the minimum Ozone (O_3) level. Moreover, the mean also gives the same Ozone (O_3) level with 0.01535 ppbv for Petaling Jaya and slightly higher with 0.0154 in Perai station. Furthermore, descriptive of Carbon Oxide (CO) data shows that the maximum Carbon Oxide level is in Petaling Jaya with 21.9510 ppbv. Meanwhile the minimum Carbon Oxide (CO) level is in Perai station, which is nearly zero (≈0.000) concentration level. The highest mean of Carbon Oxide (CO) level is at the Petaling Jaya station with 1.2394 ppbv. The descriptive of Nitrogen Oxide (NOx) data where it shows that the maximum level is in Petaling Jaya with 0.2020. Meanwhile, both the station has recorded the minimum Nitrogen Oxide (NOx) level to nearly zero (≈0.000) concentration level. The highest maximum of Nitrogen Oxide (NOx) level is at the Petaling Jaya station with 0.2020 ppbv level of concentration. The range of descriptive statistics of temperature data where it shows that the maximum temperature level is in Petaling Jaya with 43.50 degree Celsius. The minimum of temperature level for Perai is 16.10 degree Celsius while for Petaling Jaya is 16.30. The lowest temperature in Petaling Jaya is recorded from the DOE in the middle of November 2009 range around 16.30 degree Celsius to 20 degrees Celsius. The lowest range of temperature in Petaling Jaya happen in three days. The highest mean of temperature level is at the Petaling Jaya station with 28.48 degree Celsius, slightly higher than Perai station, which is 28.09 degree Celsius. Summary of humidity data for the Petaling Jaya station and Perai station is given in Table 4.3 where it shows that the maximum humidity level is in Perai with 99.70. The highest mean of humidity level is 71.78 at Petaling Jaya station slightly higher than Perai station which is 71.13.

Table 4.3:

Summary of Descriptive Statistics After Imputation

Variables	Descriptive Statistics	Location	
		Perai	Petaling Jaya
Ozone (O_3)	Mean	0.0154	0.01535
	Standard Deviation	0.01649	0.0177
	Maximum	0.17630	0.2156
	Minimum	≈(0.0000)	≈(0.0000)
	Median	0.0090	0.00800
Carbon Oxide (CO)	Mean	0.7675	1.2394
	Standard Deviation	0.3366	0.4927
	Maximum	14.6340	21.9510
	Minimum	≈(0.0000)	0.0162
	Median	0.7312	1.1857
Nitrogen Dioxide (NO_2)	Mean	0.01217	0.02808
	Standard Deviation	0.00638	0.01220
	Maximum	0.06200	0.20200
	Minimum	≈(0.0000)	≈(0.0000)
	Median	0.01100	0.02600
Temperature	Mean	28.09	28.48
	Standard Deviation	3.3436	3.345
	Maximum	39.40	43.50
	Minimum	16.10	16.30
	Median	27.50	28.00
Humidity	Mean	71.13	71.78
	Standard Deviation	15.40	15.76
	Maximum	99.70	99.60
	Minimum	2.90	6.30
	Median	72.60	74.00

Note: (≈) the value is too small (i.e, near zero)

4.4 Functional Data Analysis

This following sub-sections includes functional data analysis (FDA) technique that have been conducted will be discussing the results in order to come up with answers to the research questions and to achieve the research objectives.

4.4.1 Constructing Functional (Curve) Data

As previously mentioned, in order to start the functional data analysis (FDA), the first step that needs to be taken is data conversion. These need to be taken into account by converting or also can be called as constructing functional data from a set recorded discrete point of observations. A few steps need to be done before constructing functional curve data. The steps are as follows;

I. Determining the appropriate number of K with minimum BIC value

Before started functional data analysis, a suitable number of basis, K was determined from the average daily curve. Table 4.4 shows a summary of RSS value with respect to different number of K. As have been discussed earlier, in order to get the appropriate number of K, a B-spline method is used instead for the other basis method due to its flexibility. This step is essential for making the method feasible for number of bases, K where the basis has got by using a trial an error method where the value of K is set from 5 to 20. The most suitable number of K was chosen by selecting the lowest BIC value in between the K value from 5 to 20. An example is shown below where the appropriate number of bases K for Petaling Jaya station air quality monitoring station for Ozone level. Given that RSS for this average curve is 0.0006389835, with BIC value for K=5 is done such that from the previous equation:

$$BIC = log\left(\frac{RSS}{m}\right) + \frac{k}{m}\log(m)$$

$$BIC = log\left(\frac{0.0006389835}{24}\right) + \frac{5}{24}\log(24)$$

$$= -9.871591$$

Table 4.4
Summary of RSS Value with Respect to Different Number of K

Locations	K	RSS	BIC
Petaling Jaya	5	0.0006389835	-9.871591
	6	0.0001811462	-10.99975
	7	0.0001523581	-11.0404
	8	3.977789e-05	-12.2509
	9	9.102629e-06	-13.59323
	10	1.300127e-05	-13.10433
	11	9.984262e-07	-15.53853
	12	1.634186e-06	-14.91339
	13	1.209022e-06	-15.08231
	14	1.756784e-07	-16.8788
	15	2.408923e-07	-16.43069
	16	2.928305e-07	-16.10302
	17	7.743189e-08	-17.3008
	18	1.70118e-07	-16.38129
	19	1.049248e-07	-16.73212
	20	6.778507e-08	-17.0366

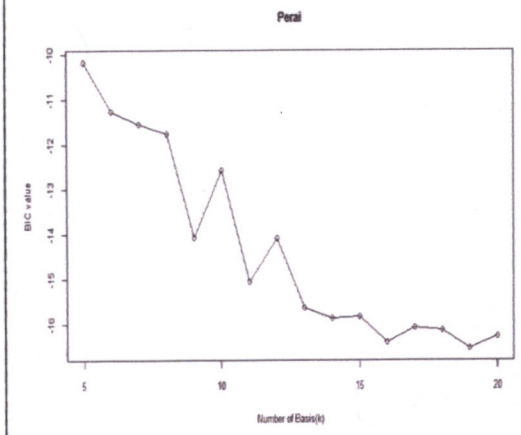
A summary of RSS value with respect to different number of K is given in Table 4.4. Based on the Table, the most suitable number of K at Petaling Jaya is 17 since the BIC value shows the most minimum compared to the others. This method is different from the previous researcher where a judgmental technique of observation is being used to decide the suitable number of K. Antecedently, there was no appropriate technique as a guideline to determine the number of K. The summary of the number of bases K for other air quality monitoring stations are shown in the Table 4.5 below.

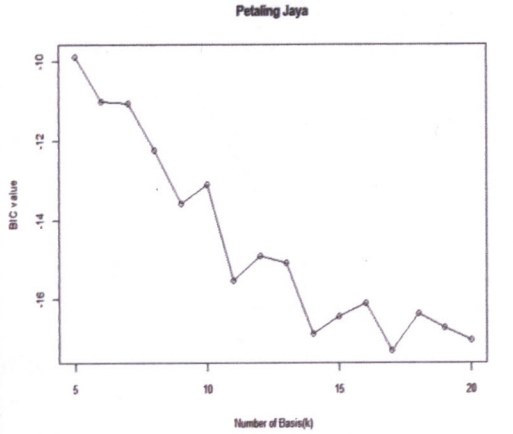
Table 4.5:
Summary of Number of Bases, K for Each Air Monitoring Stations

Location	Variables	Number of Bases, K
Perai	Ozone (O_3)	19
	Carbon Oxide (CO)	19
	Nitrogen Dioxide (NO_2)	19
	Temperature	19
	Humidity	19
Petaling Jaya	Ozone (O_3)	17
	Carbon Oxide (CO)	18
	Nitrogen Dioxide (NO_2)	18
	Temperature	19
	Humidity	19

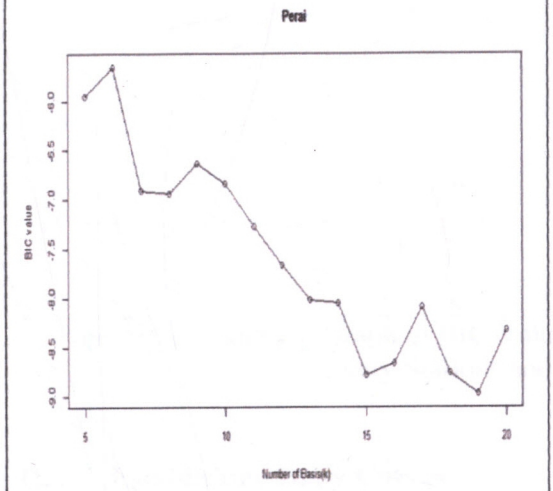
Table 4.5 above summarizes the number of bases, K for Perai and Petaling Jaya air monitoring station with respect to Figure 4.2. The most suitable number of bases, K for Perai station is 19 for all the variables which are Ozone (O_3), Carbon Oxide (CO), Nitrogen Dioxide, temperature and also humidity. Meanwhile, for Petaling Jaya air monitoring station, the number of bases, K for Ozone is 17 followed by the number of bases for Carbon Oxide and Nitrogen Dioxide variables which is 18. While for the variable temperature and humidity, the number of bases, K for Petaling Jaya air monitoring station is 19. From Figure 4.2, the number of basis is chosen based on the minimum and tolerate values between K equal 5 to 20 in which the number of basis is not too small and not too large.

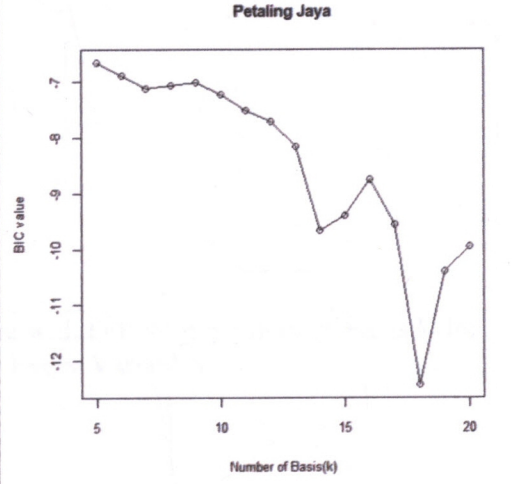
Ozone (O_3)



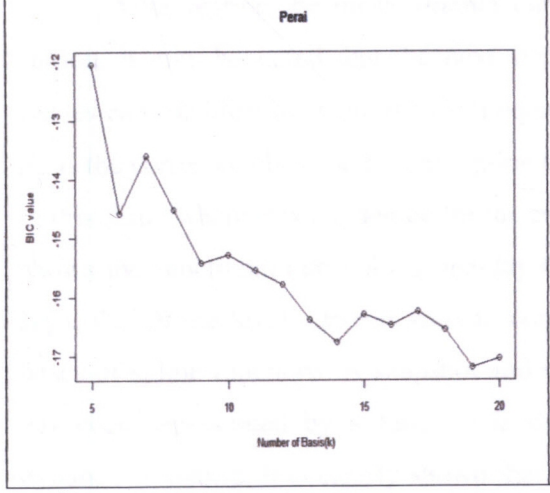


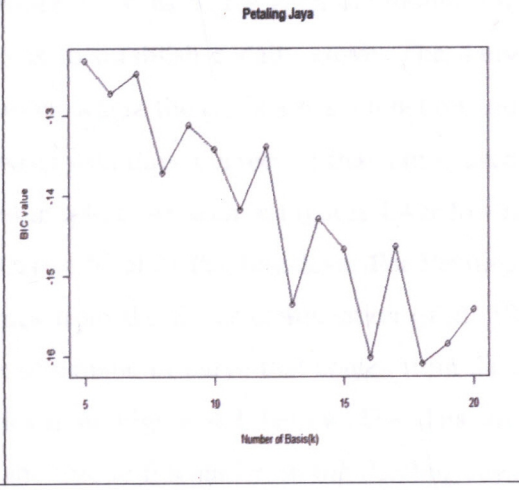
Carbon Oxide (CO)





Nitrogen Dioxide (NO_2)





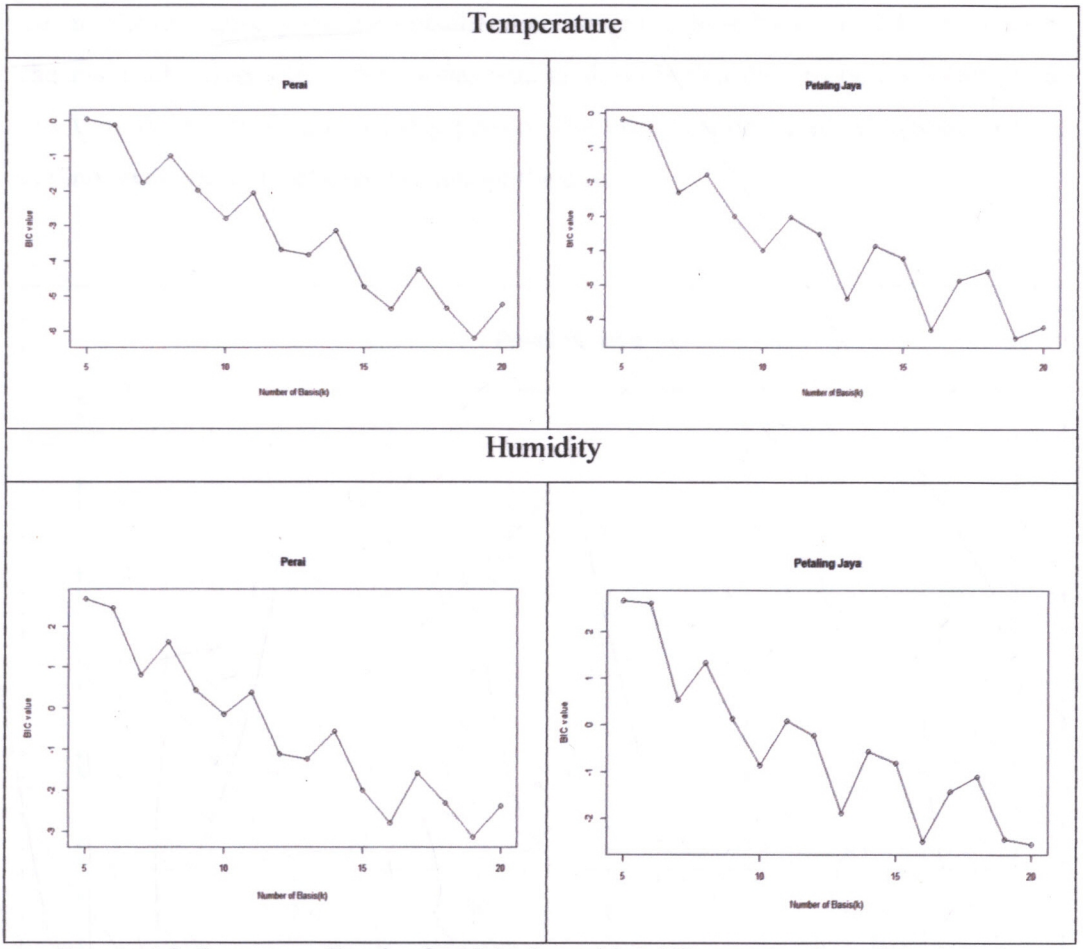


Figure 4.2: Summary Graph of BIC Value with Different Number of Basis K for Every Stations and Every Variables.

II. Establishing Daily Curves

After getting the most suitable number of bases K for each air monitoring station, it may be noted that the next step is to establish a daily curve. The daily curves can establish by using the (3.2) equation; where the φ_k is a basis function and β_k is the corresponding coefficient. Spline basis with three degrees of freedom is used in this study where it is the degree for the cubic spline. As seen in Figures 4.4 below it shows the functional curve for a one-day Ozone level in Petaling Jaya. For Petaling Jaya, the Ozone level curve is smooth comes from the linear combination of all 19 bases of spline functions. A snapshot and a schematic of curve that comes from data has been represented by a function is shown in Figure 4.3 below. The data are dynamic in nature. It is clearly shown that the Ozone fluctuation in the Petaling Jaya

air monitoring quality can be visualized according to time basis in 24 hours. One of the main advantages from functional data analysis is that the Ozone level can get at any time frame within the one-day period. The visualization for the fluctuation of the Ozone level also can be seen throughout the day.

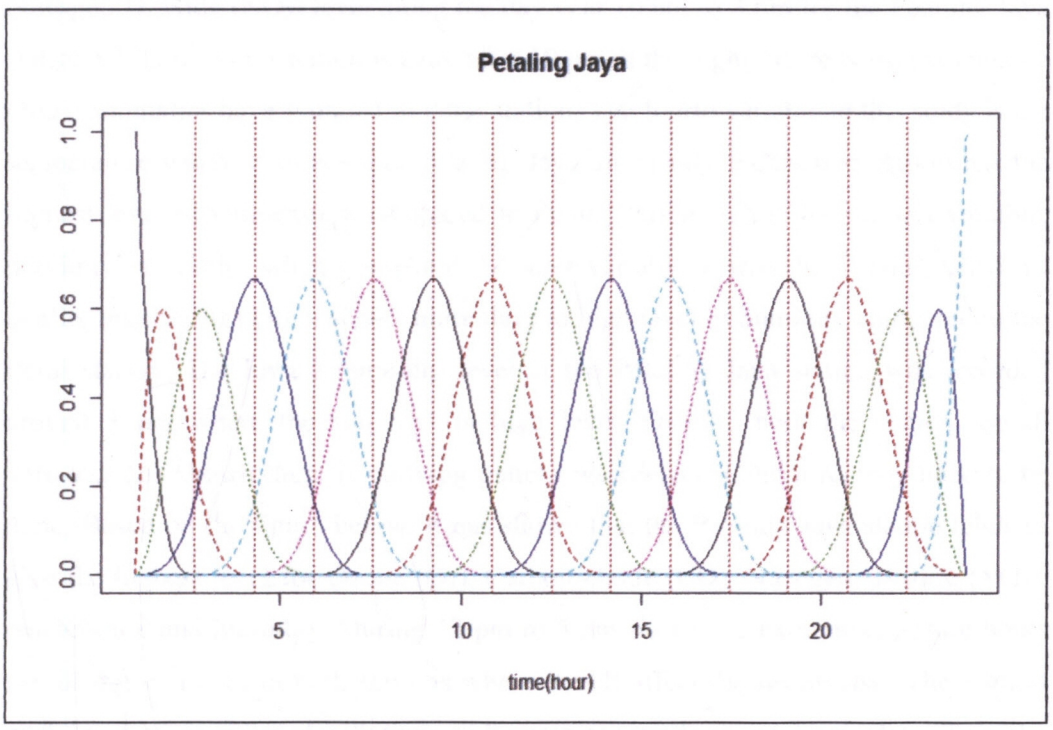
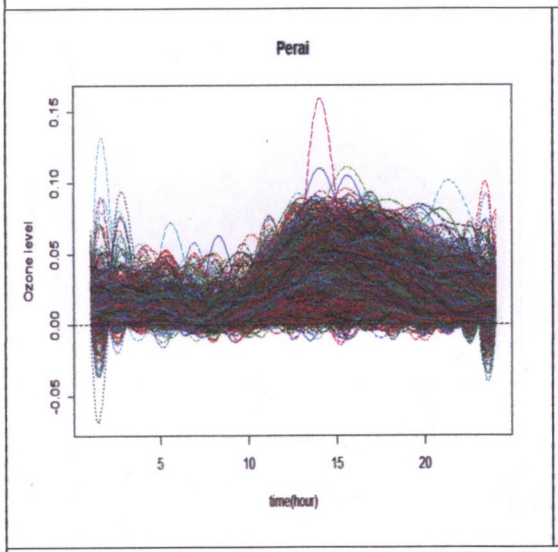


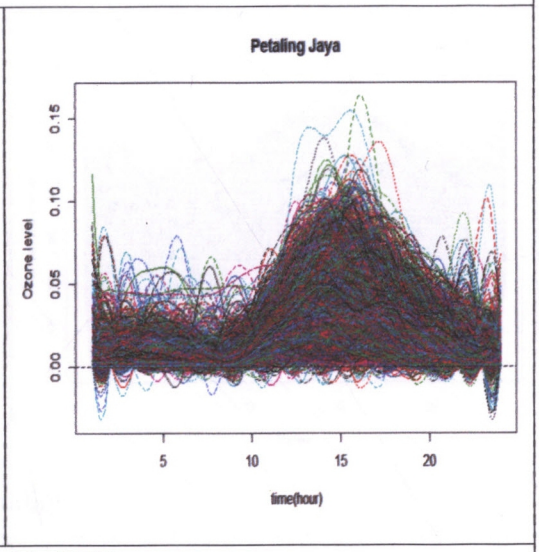
Figure 4.3: Graph of Spline Basis

In the meantime, as seen in Figure 4.4 it shows functional diurnal for each variable (Ozone, Carbon Oxide, Nitrogen Dioxide, Temperature and Humidity) daily curves for all 3287 days which is equivalent to 9 years (2009 to 2017) at both air quality monitoring station (Perai and Petaling Jaya). The number of curves shown in the graph is interrelated to the number of days that was being sampled which is 3287 days. Overall, the dynamic Ozone (O_3) level at both stations is about the same level, which is at 3 p.m. in the evening and lower in night time for the along 9 years. As shown in the Figure 4.4 below, all the curves do not lines in the same each other. For the Ozone (O_3) level, the highest peak was shown in the Petaling Jaya station. Petaling Jaya has record more number of exceedance of Ozone concentration compared to the Perai station since the station having more curves with peak that exceeding the standard level (>0.1ppbv). Furthermore, for the second variable which

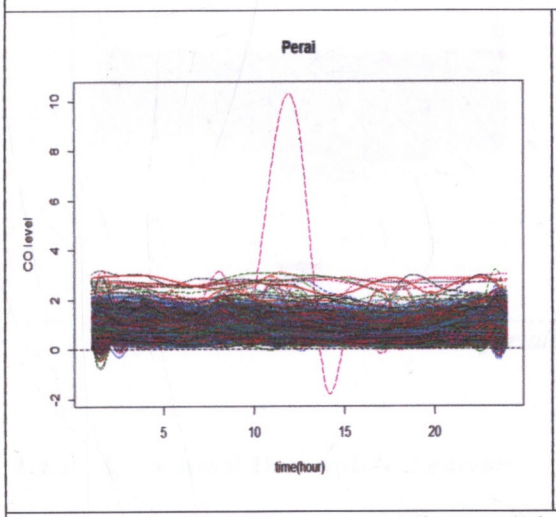
is Carbon Oxide showed the dynamic Carbon Oxide level at Perai station was peak in between 10 am to 3 pm, while for Petaling Jaya station the level of Carbon Oxide (CO) was peak in between 1 pm to 3 pm in a day. The third variables are Nitrogen Dioxide where it can be observed from the Figures below that Petaling Jaya station has the highest level of Nitrogen Dioxide for the along the years. The highest peak of Nitrogen Dioxide (NO_2) level along the day is at 10 am to 3 pm for the Petaling Jaya station while for Perai station is peak after 10 pm in the night. More Nitrogen Dioxide (NO_2) anomalies have founded in Perai station. The fourth variable in this study is the temperature where it shows that Petaling Jaya air quality monitoring station has the highest level of temperature compared to Perai` station. While for the last variable, humidity is similar when compared to other variable where the Petaling Jaya air quality monitoring station shows more days in high level of humidity compared to the Perai station. The lowest humidity level at the Petaling Jaya station was recorded around 3 pm where the sun rays in high level. Overall, both the station for all variables has shown there is outlying pattern was detected and a remedy need to be done. Based on the figure below, it has shown that the Petaling Jaya station tends to give the highest level for Ozone (O_3), Carbon Oxide (CO), Nitrogen Dioxide (NO_2), temperature and humidity. During 12 pm to 5 pm shows the extreme exposure hours for all the variables in both stations where it will affect the health risk. The highest peak level of air polluted with these pollutants occurred around 3 pm where it similar to the previous study that during these time the sun rays is strongest that may dangerous for our health. All sites having a similar bimodal shape for average 24 hours of all variables.

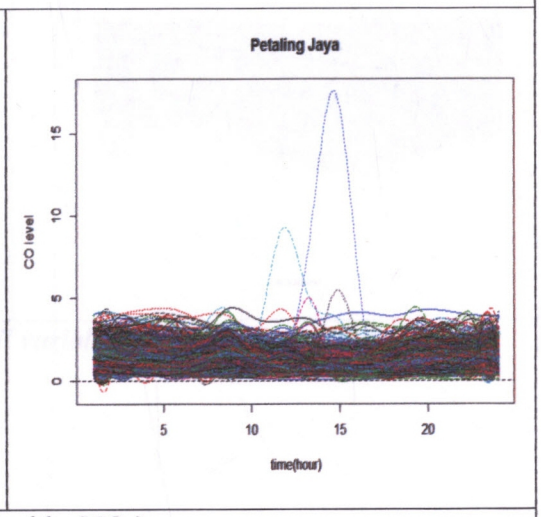
Ozone (O_3)



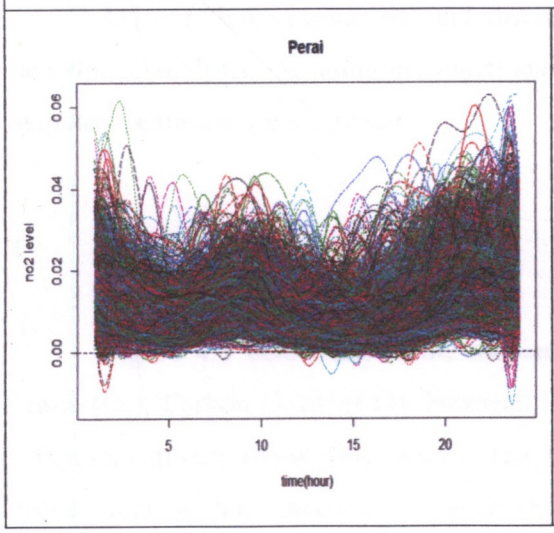


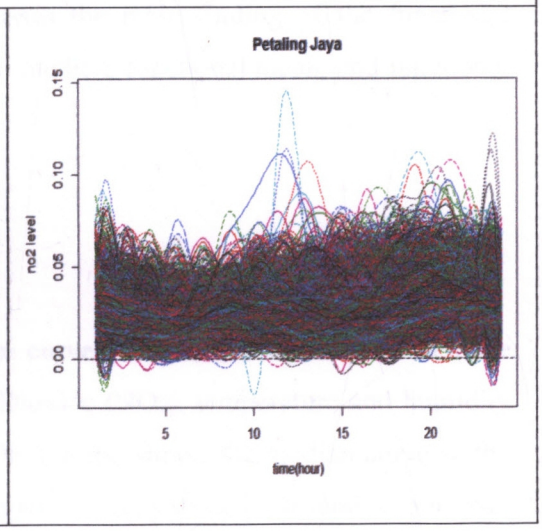
Carbon Oxide (CO)





Nitrogen Dioxide (NO_2)





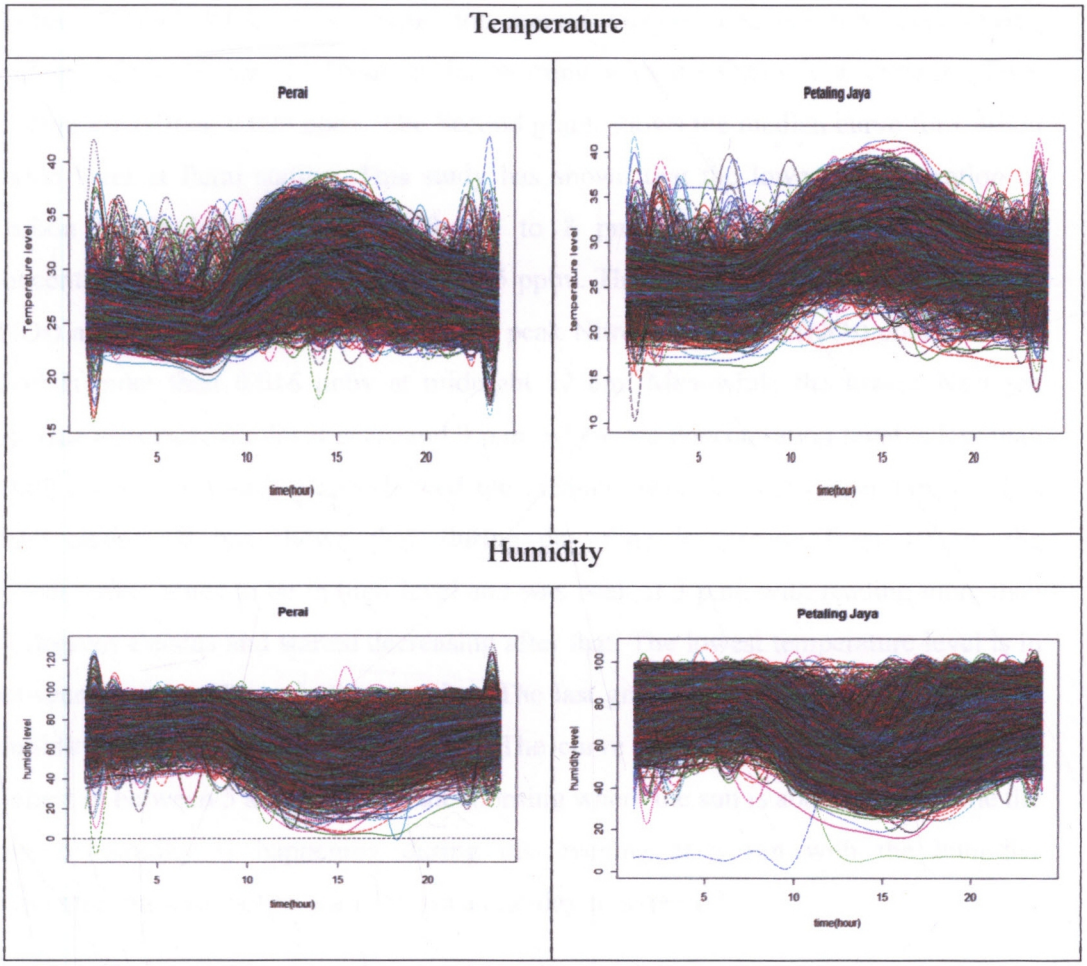


Figure 4.4: Functional Diurnal for all variables for 3287 Days (9 years)

4.4.2 Functional Descriptive Analysis

This section summarizes and discusses the main finding of the functional descriptive analysis, including the functional median, functional mean, and functional standard deviation are discussed.

I. Median Curve

Figure 4.5 below shows the median curves for all five variables which are Ozone (O_3), Carbon Oxide (CO), Nitrogen Dioxide (NO_2), temperature and humidity in Perai air quality monitoring station. The first graph shows the median curve in the Ozone level at Perai station. The peak Ozone concentration is around 3 pm with

reading around 0.035 ppbv while the lowest Ozone concentration level during midnight from 12 am till 10am in the morning with the Ozone concentration level reading lower than 0.005 ppbv. The Second graph shows the median curve for Carbon Oxide level at Perai station. This study has shown that the lowest concentration of Carbon Oxide level is between 3 pm to 8 pm with the Carbon Oxide (CO) concentration level reading lower than 0.5 ppbv. The median curve for Nitrogen level (NO_2) at Perai station has shown that the peak Nitrogen Dioxide (NO_2) concentration level is more than 0.016 ppbv at midnight 12 am. Meanwhile the lowest Nitrogen Dioxide concentration level is around 3 p.m. where the concentration level is less than 0.008 ppbv. The fourth graph showed the median curve for the temperature level at Perai station. It has shown that, during day time in starting from 10 am the concentration tends to be in high level and was peak at 3 p.m. with reading more than 32 degrees Celsius and started decreasing after that. The lowest temperature level is in between 5 am to 10 am in the morning. The last graph of the Figure 4.5 showed the humidity median curve at Perai station. The curve shows that the humidity level is highest in between 5 am to 8 am in the morning where the sun is about to rise while the lowest humidity is happening during the evening at 3 pm with the humidity concentration level below than 55% in a one-day time period.

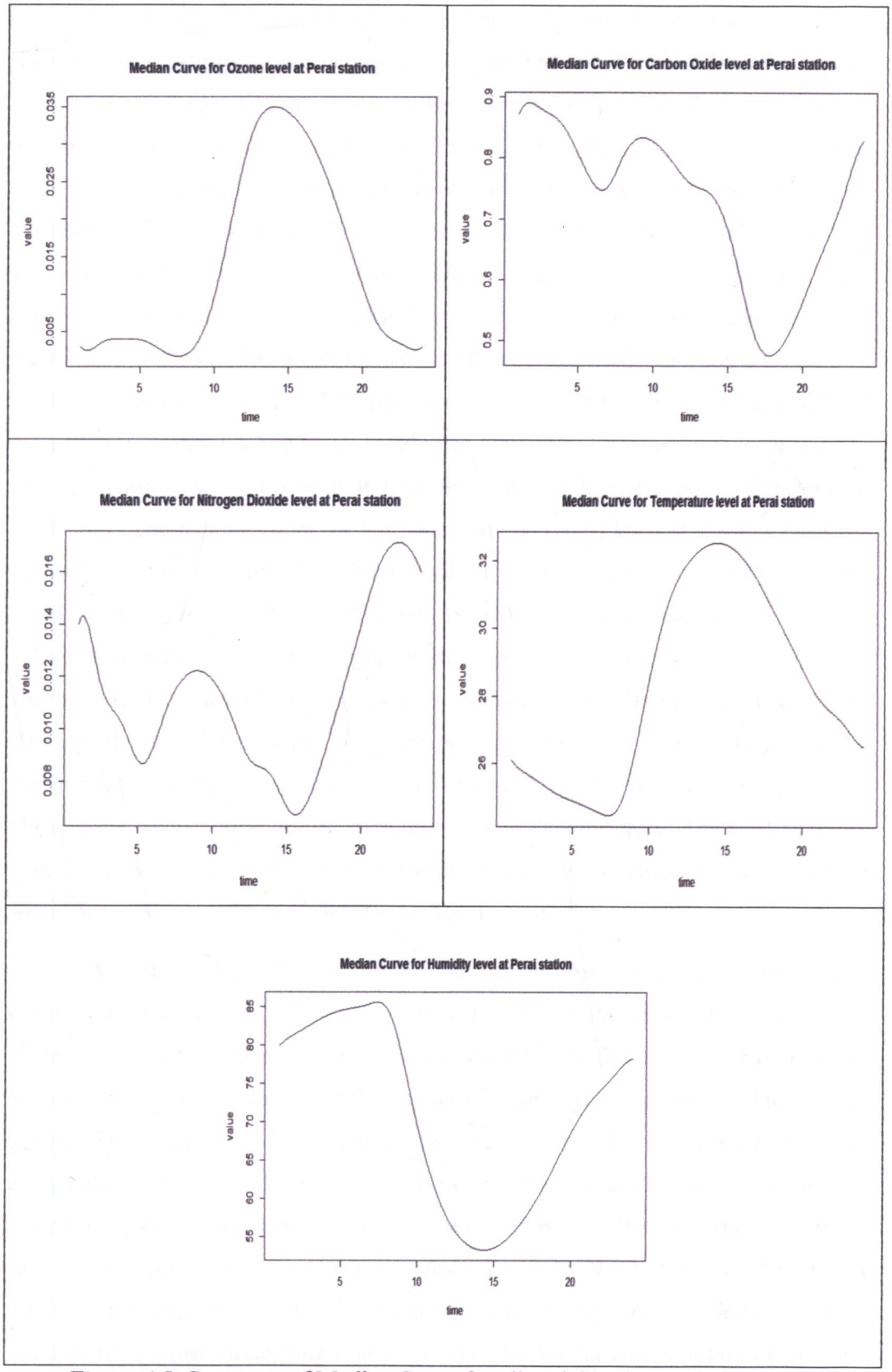


Figure 4.5: Summary of Median Curve for all variables in Perai Air Quality Monitoring Station

A snapshot and a schematic of the median curve of the variables (Ozone, Carbon Oxide, Nitrogen Dioxide, temperature and humidity) level in Petaling Jaya air quality monitoring station are shown in the Figure 4.6 below. The first graph showed the median curve of Ozone (O_3) in the Petaling Jaya station. During 3 pm it was shown at the highest concentration level with more than 0.03 ppbv of Ozone concentration level. Meanwhile, after 3 pm the Ozone (O_3) level started to decrease and remain to be lower than 0.01 ppbv of Ozone concentration level until 10am. Next, the second graph showed the median curve of the Carbon Oxide variables at the Petaling Jaya station. The graph has showed that during 12am the Carbon Oxide is at peak with reading more than 1.35 ppbv Carbon Oxide (CO) concentration level and start to decrease the Carbon Oxide level after 12am. The Carbon Oxide level starts to rise up continuously in between 3pm to 8pm with the lowest concentration level at 1.05 ppbv. The median curve for Nitrogen Dioxide (NO_2) level at Petaling Jaya has showed in the third graph where the peak of the level is at 10am and 8pm before start to fluctuate again. The highest Nitrogen Dioxide (NO_2) concentration level was at 8pm with more than 0.03 ppbv concentration level of Nitrogen Dioxide (NO_2). The fourth graph has showed the temperature's median curve at the Petaling Jaya station. The graph showed that it has only one peak at 3 pm with more than 32 degrees Celsius. The lowest temperature level is in between 5am to 8am. The last graph has showed the median curve for the humidity level in Petaling Jaya. The least humidity level at 3pm with less than 55% level concentration of humidity and started to rise afterwards and reach the peak in between 5am to 10am.

As seen in Figure 4.5 and 4.6, the fluctuation time in a day of both Perai and Petaling Jaya air quality monitoring station has shown the same pattern. The only difference is the concentration of each variable either higher or lower based on the station background. 3 pm during the daytime has showed as the main influence time for all the variable as it showed as the highest peak for the Ozone (O_3) and temperature variables. Meanwhile, early in the morning after 5 am to 10 am the temperature level, having lowest mean concentration since the production of Ozone and temperature is depends on the intensity of sunlight. While for the rest variables which are Carbon Oxide (CO), Nitrogen Dioxide (NO_2), and humidity, 3 pm is considered as most influencing time as it gives the lowest concentration of each the variables.

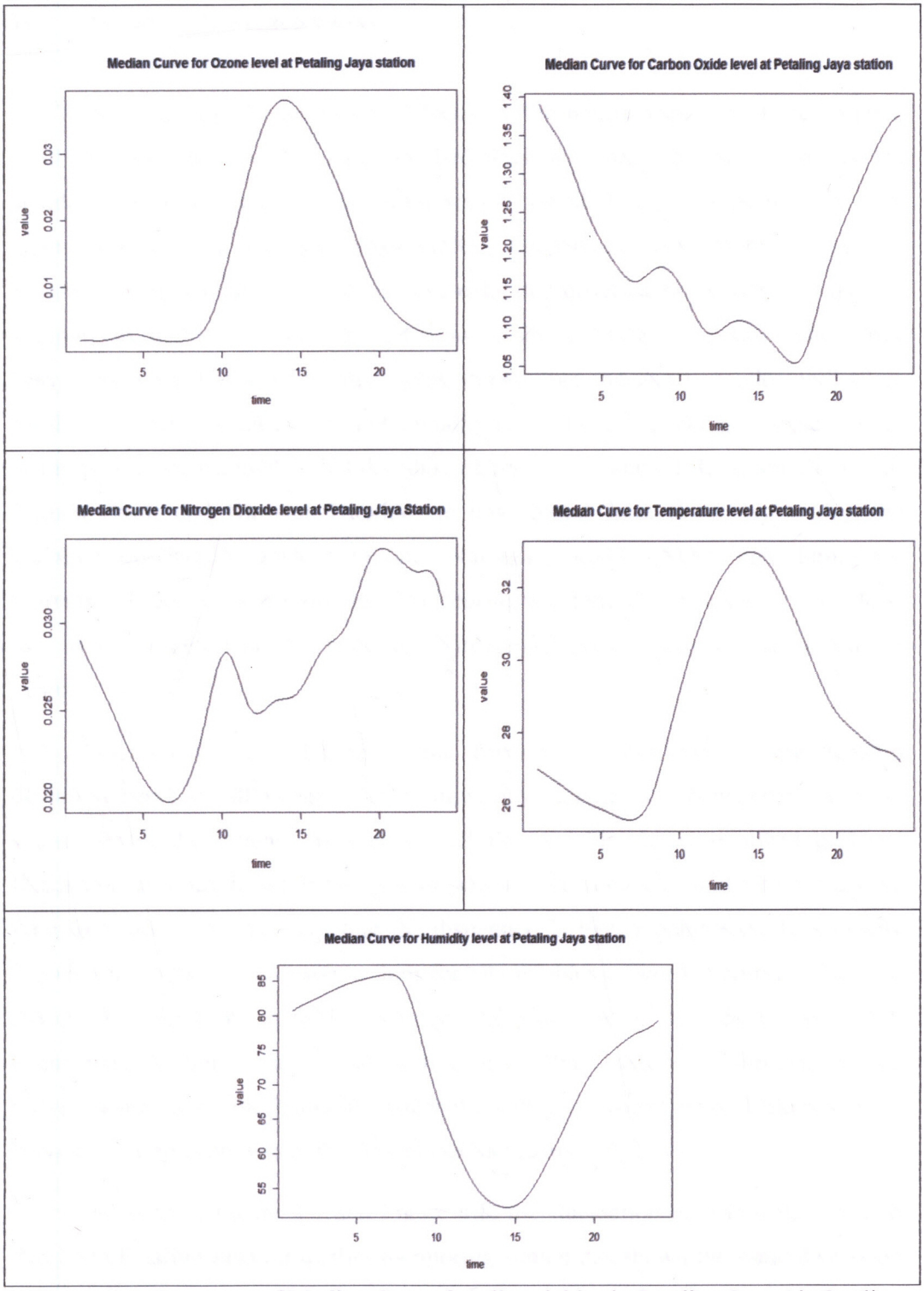


Figure 4.6: Summary of Median Curve for all variables in Petaling Jaya Air Quality Monitoring Station

II. Standard Deviation Curve

Next descriptive analysis will focus on the hourly variation of the variable level or also can be called as a standard deviation curve. Figure 4.7 shows the standard deviation curves at the Perai station for all five variables (Ozone (O_3), Carbon Oxide (CO), Nitrogen Dioxide (NO_2), temperature, and humidity). From the graphs, we can see all the variables have exhibited different fluctuation in a day. In comparison, in within 24hours period shows a similar dispersion pattern where they start to rise up at 10am and having a peak around 3pm and start to decline until 12am for the Ozone (O_3), temperature and humidity level. It was happening because at 3pm, at the present of sunlight is if fully showed up. Both standard deviation curves for Carbon Oxide and Nitrogen Dioxide are low around 3pm. From day to day, the variation of Carbon Oxide (CO) and Nitrogen Dioxide (NO_2) high during the midnight till early in the morning where during this time the presence of sunlight is low. Both variable Carbon Oxide and Nitrogen Dioxide having a sharper peak at 12am.

As seen in Figure 4.8 below, the dispersion or also called as the standard deviation curve of all variables in Petaling Jaya air quality monitoring station is shown. From the graphs, we can see all the variables have exhibited different fluctuation in a day in the Petaling Jaya station. The common similar things can be seen from all the variables is that the dispersion is high around 3pm. It is clearly showing the peak of the graph at 3pm for all variables except for Nitrogen Dioxide (NO_2). While for the variables Nitrogen Dioxide (NO_2), it tends to be a high dispersion after 8pm. There are similarities for variable Ozone (O_3), Nitrogen Dioxide (NO_2), temperature and humidity where the valley or lowest level dispersion is in between 5am to 10am where the sunlight existence are very low.

As seen in Figure 4.7 and Figure 4.8, the fluctuation time in a day of both Perai and Petaling Jaya air quality monitoring station has shown the same dispersion pattern when compared to same variable for the Ozone (O_3), Carbon Oxide (CO), temperature and humidity variables. It was different things happen to the Nitrogen Dioxide variable for both stations were around 5am to 10am, Perai station was having a slight peak during that time while for Petaling Jaya station tend to have valley

during that hour. In conclusion, for a standard deviation curve, it has shown that Petaling Jaya has a higher dispersion compared to the Petaling Jaya station.

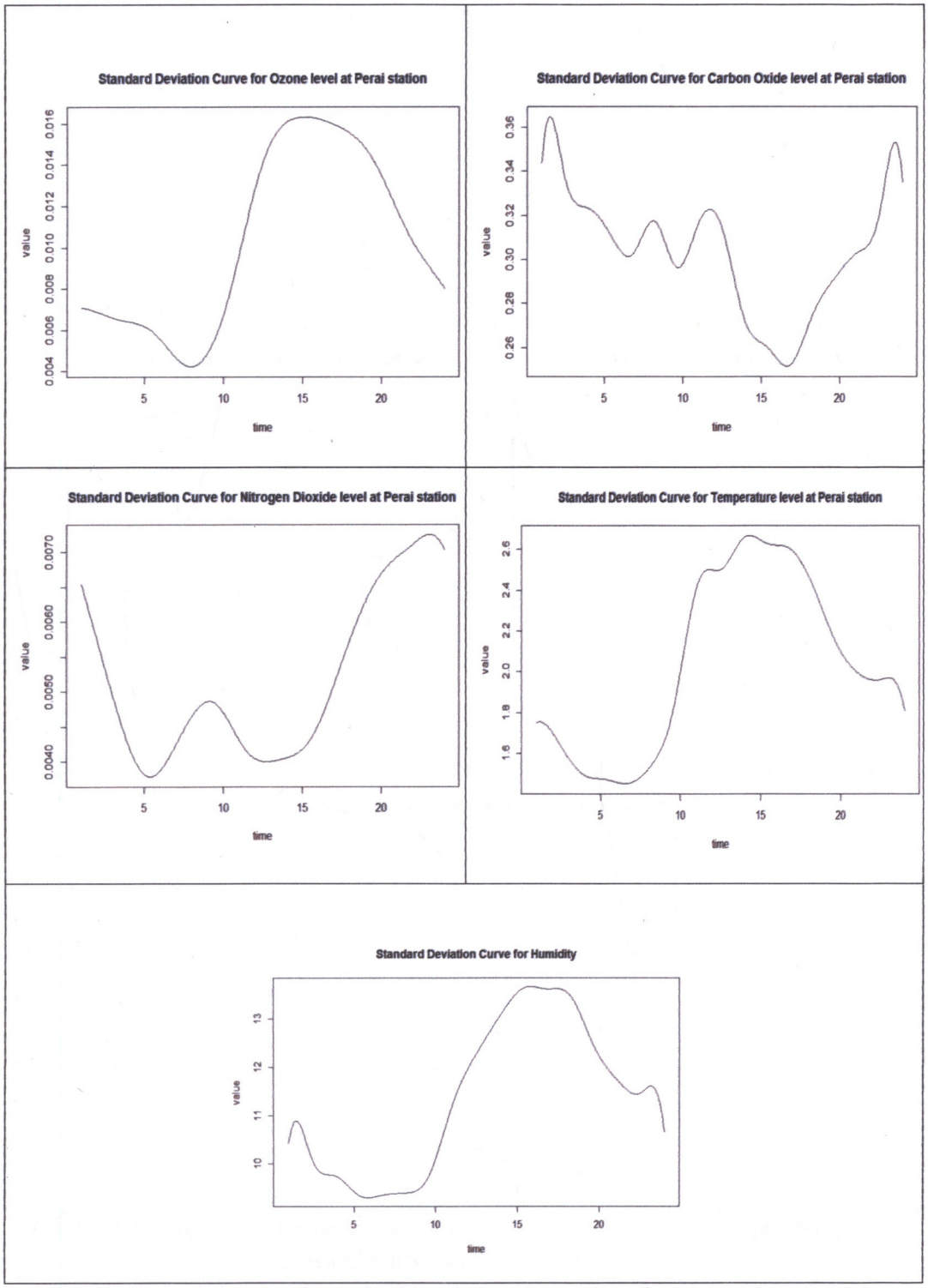


Figure 4.7: Summary of Standard Deviation Curve for all variables in Perai Air Quality Monitoring Station

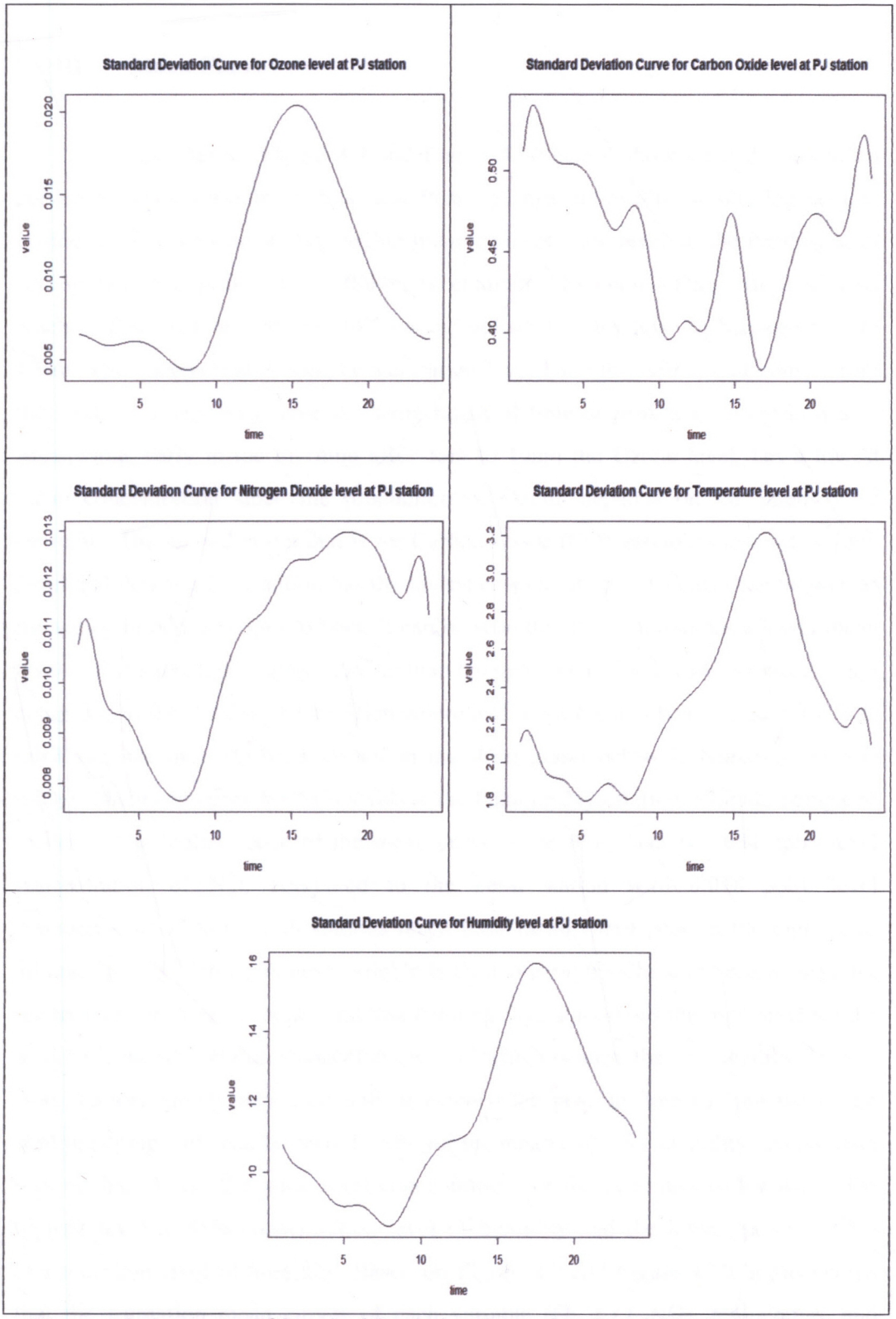


Figure 4.8: Summary of Standard Deviation Curve for all variables in Petaling Jaya Air Quality Monitoring Station

III. Mean Curve

As seen below, Figure 4.9 and Figure 4.10 below shows the diurnal mean curves for each variable at Perai and Petaling Jaya air quality monitoring station. Ozone level was found slightly higher mean concentration level at the Petaling Jaya station with the peak near to 0.04ppbv compared to Perai station. Both stations reached their peak around 2pm to 3pm and got their valley level in between 5am to 10am. This means that, Ozone concentration level has same effect that comes from the sunlight where they have the same range of time at peak and valley in a day. Meanwhile, early in the morning after 5am to 10am the Ozone level, has a lowest mean concentration since the production of Ozone depends on the intensity of sunlight. The second graph that is for Carbon Oxide (CO) variable showed that both Perai and Petaling Jaya station has the lowest concentration of CO also can be seen as the valley in between 3pm to 8pm. It can be seen that Perai station has a lower mean level of CO since it the valley is lower than 0.6 ppbv mean levels of CO concentration compared to the Petaling Jaya station where its lowest point is below 1.10 ppbv. The third variable that has been shown in the third graph below is Nitrogen Dioxide (NO_2). The mean curve for NO_2 levels at the Petaling Jaya station is higher compared to Perai. The highest peak of the mean curve at Petaling Jaya is 0.034 ppbv level concentration of NO_2 compared to the Perai station with 0.018 ppbv level concentration of NO_2. Both stations reach their mean curve peak in the night hour around 8pm to 12am. The next variable is the temperature where the mean curve for the temperature level at both Perai and Petaling Jaya station has the high level for the peak of mean temperature concentration level which is more than 32 degrees Celsius. Both stations slowly increased until it reached the peak at 2pm to 3pm where the sunlight completely can be seen. Lastly, for the mean curve for humidity level at both stations has shown the same level concentration for the peak and valley where the highest level is 85% concentration level of humidity and the lowest point is 65% concentration level of humidity. Based on Figure 4.9 and Figure 4.10, it has shown that the formation mean curves of each variable (O_3, CO, NO_2, temperature and humidity) have a similar pattern in both stations.

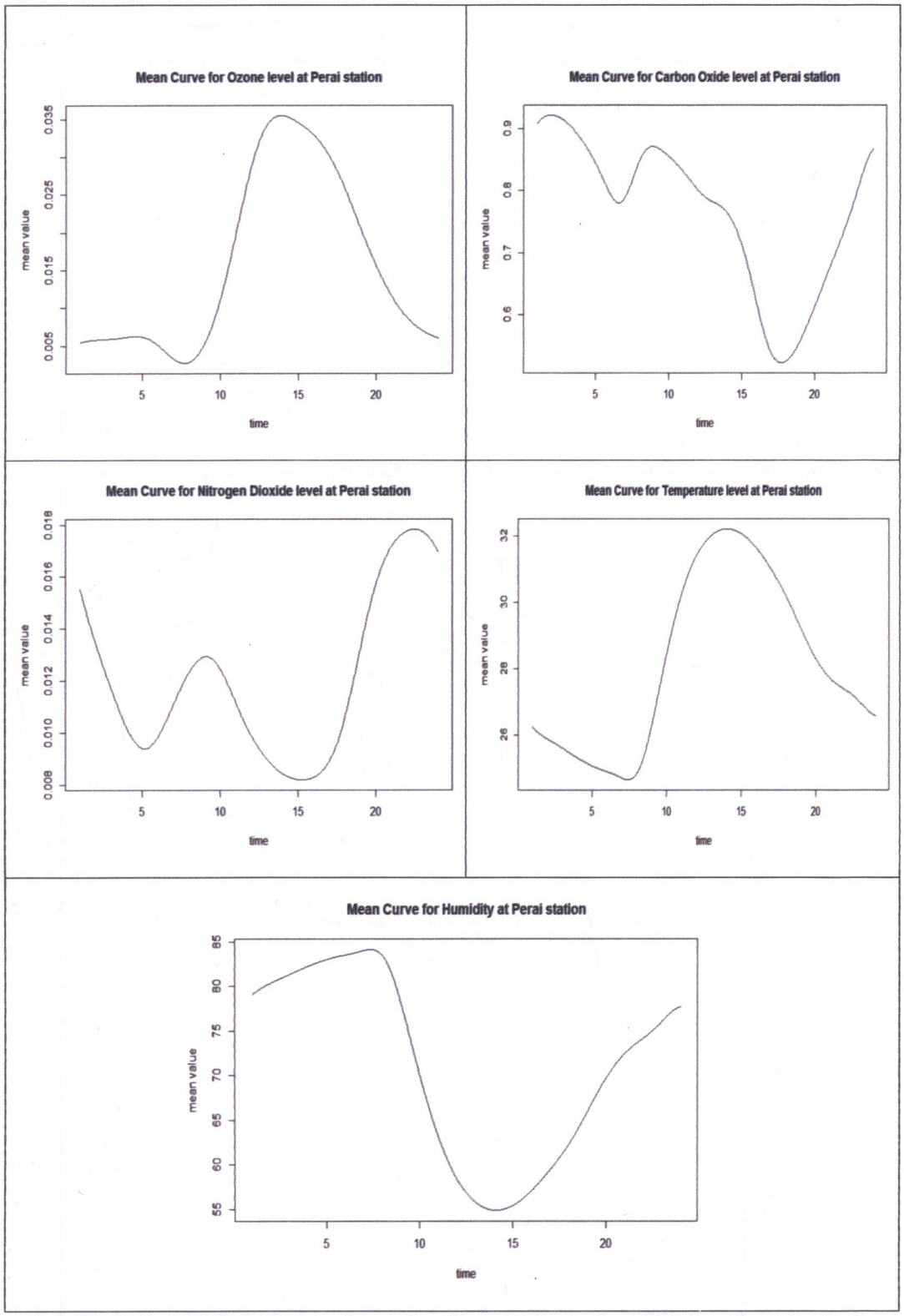


Figure 4.9: Summary of Mean Curve for all variables in Perai Air Quality Monitoring Station

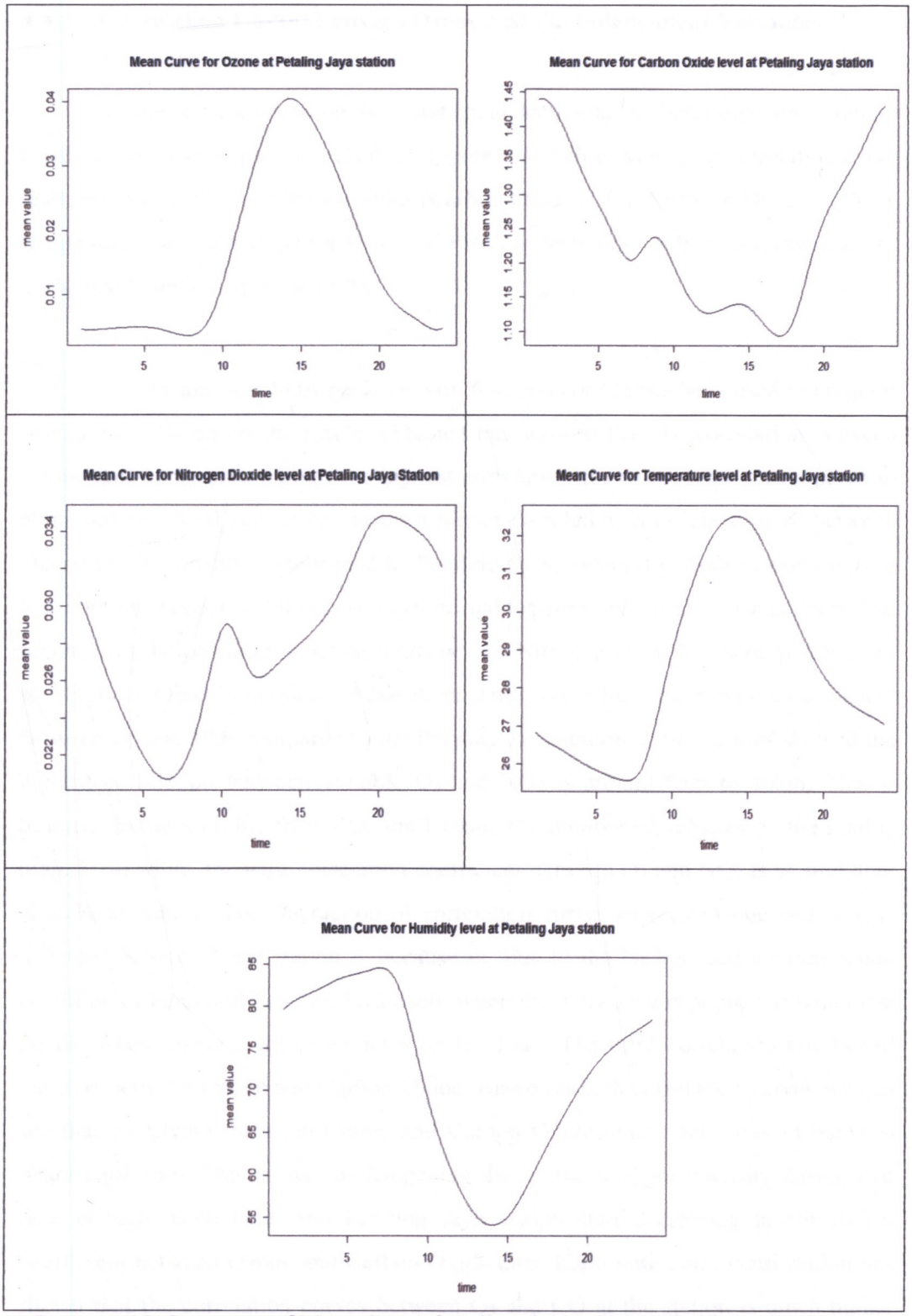


Figure 4.10: Summary of Mean Curve for all variables in Petaling Jaya Air Quality Monitoring Station

4.4.3 Correlation Curves between Ozone and the Independent Variables

In statistics, correlation is a statistical technique to determine the strength relationship between pairs of variables. Figure 4.11 below shows the correlation curve between Ozone (O_3) and the variables (Carbon Oxide (CO), Nitrogen Dioxide (NO_2), temperature, and humidity) for Perai and Petaling Jaya air quality monitoring stations computed based on equation (3.8).

In this analysis, FDA package with function cor.fda has been used to program in Rstudio. Based on the results obtained has showed that the correlation between Ozone and Temperature gives the highest correlation for both stations. Compared to both stations, Perai station has given a higher correlation coefficient curve between Ozone and temperature compared to Petaling Jaya. Generally, both stations have a low correlation coefficient between Ozone and temperature variable around 5am. The second correlation curve is between Ozone and Nitrogen Dioxide where in the graph shows the red line. It has showed that Perai station has a higher correlation coefficient between O_3 and NO_2 compared to the Petaling Jaya station. Both stations showed the highest correlation between variable O_3 and NO_2 is around 8am to 10am. This is because, in the morning from 8am until 10am, the number of vehicles on the road is high. Meanwhile, the least correlation coefficient between O_3 and NO_2 is around 8pm is at Perai station. The fluctuation of correlation curve between Ozone and NO_2 is different between Perai station is because of, due to the background environmental condition of Perai and Petaling Jaya itself, where Petaling Jaya is populated with most people where using a lot of vehicles on the road. The third correlation coefficient curve is between Ozone and Carbon Oxide. Based on both correlation curves we can see that correlation between Ozone and Carbon Oxide reach their peak in between 10am until 3pm. This is may be happening due to the sunlight intensity during that time is high. Both Perai and Petaling Jaya station start decreasing in correlation coefficient between Ozone and Carbon Oxide after 12am until 5am. Perai station has shown that the correlation curves between O_3 and CO at the station is much higher with more than 0.1 for a long hour compared to the Petaling Jaya station before slowly declining the correlation coefficient level. The last pair is the correlation coefficient curve between Ozone and humidity variables. It has shown that the

Petaling Jaya station has the lowest correlation coefficient with lower than -0.4. This mean that, Ozone and humidity as a negative correlation coefficient effect. Increasing of O_3 level will decrease the humidity level.

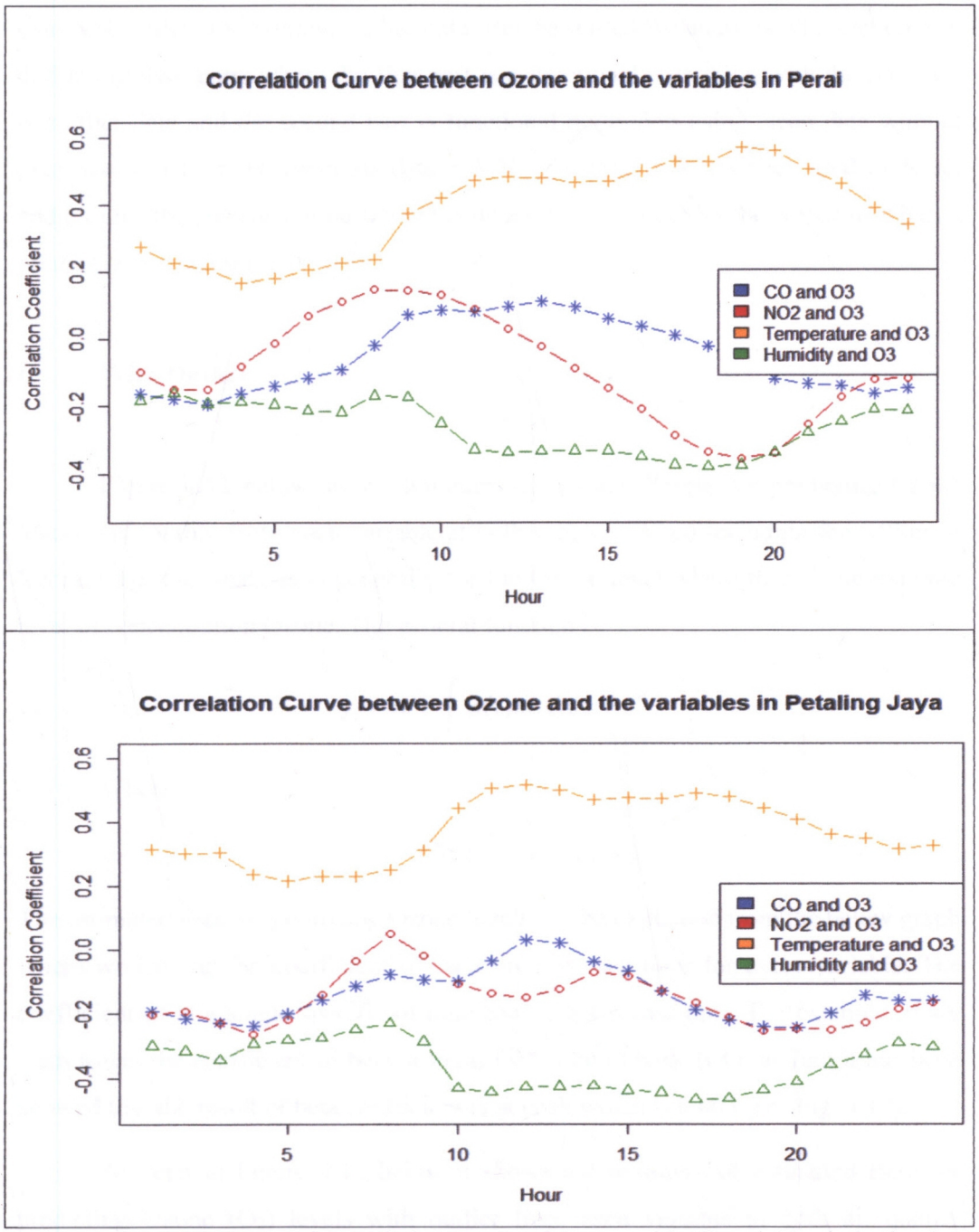


Figure 4.11: Summary of Correlation Curve of Ozone with all variables in both Air Quality Monitoring Station

4.4.4 Functional Regression Analysis

In this section, the functional regression analysis result will be shown based on data with outlier and without outlier data. Before started the analysis, this section will divide into two parts where the first part is a functional regression with the presence of outlier data and the second part is functional regression using clean data without presence of outlier. To clean the data, a MVN package in R has been used to detect and remove the presence of outlier in the data set. The outlier has been removed based on the Ozone data set.

I. With Outlier

Figure 4.12 below shows summary of estimated Beta for predicting Ozone levels with outlier from each variable at both stations. When including the outlier of Ozone (O_3), this analysis is generally for the Ozone level where there is an extreme level of concentration Ozone. The general function is

$$y_i = \propto + \int \beta(t)x_i(t)dt + \varepsilon_i$$

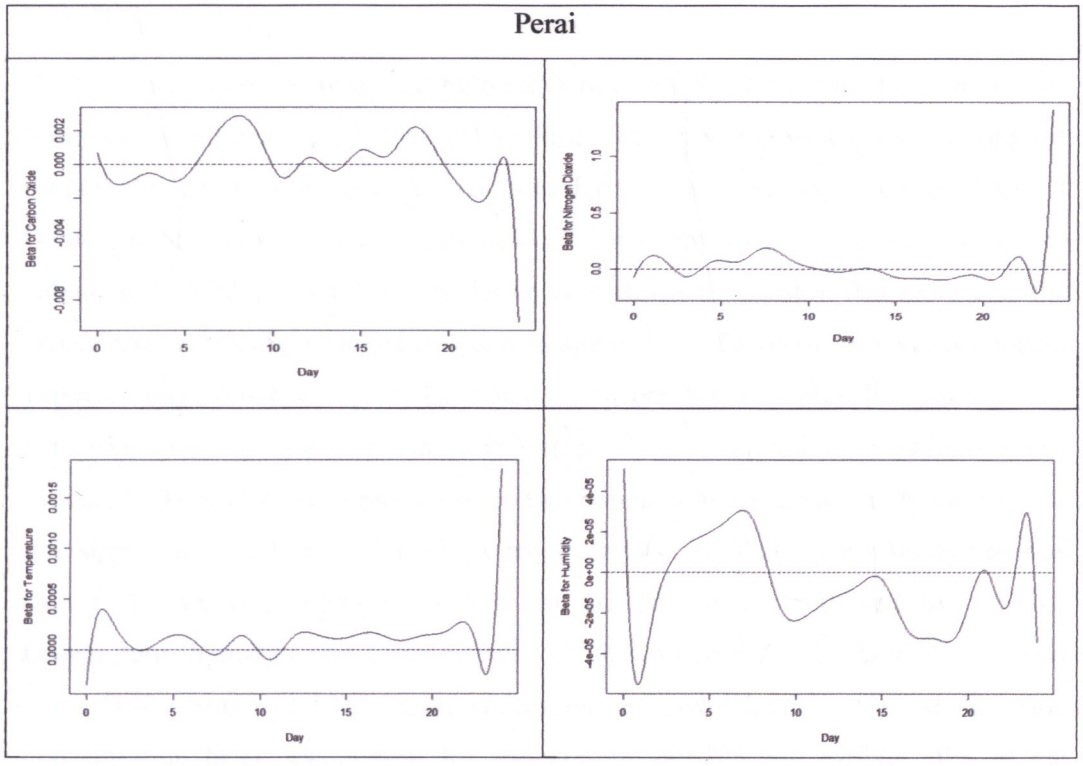
Where,

$$t = t_1, t_2, t_3, t_4 \dots t_n,$$

The estimated beta for predicting Ozone levels can be explained from the below graph where we can get the coefficient of each time or each hour for every variable. The coefficient of at a single specific of time can be get across time. For example, if we want to get the coefficient of beta at Perai for Carbon Oxide (CO) at 7am it has been showed that the result of beta coefficient is at peak which is 0.003 (see Fig. 4.12).

As seen in Figure 4.12 below it shows the summary of estimated Beta for predicting Ozone (O_3) levels with outlier from each variable in both air quality monitoring stations. This is mean this analysis include with extreme reading of Ozone (O_3). From the first association between Ozone level and Carbon Oxide (CO). The association of Ozone (O_3) and Carbon Oxide (CO) is high at 8am for both Perai and

Petaling Jaya station. It was clearly shown the strong negative relationship 3pm to 8pm in the Petaling Jaya station contradict with the Perai station where during that time it shows the high relationship between Ozone and Carbon Oxide. The next pair is the relationship between Ozone (O_3) and Nitrogen Dioxide (NO_2). Both station shows similar fluctuation where around 3am, 5am, 8am and 10pm it shows there are strong positive association for Nitrogen Dioxide (NO_2) and Ozone (O_3) level. The average out with considering the extreme value of Ozone level and temperature has shown that it is clearly highly associated at 1pm to 4pm for the Petaling Jaya station. Meanwhile, at Perai station, fluctuate more rapidly and has a higher association between Ozone and temperature. Petaling Jaya stations have shown the highest association in between 3pm while in Perai station did not show obvious peak, which means the high relationship is between the same range for each peak. The last relationship is between Ozone (O_3) level and humidity level. It has shown that the high relationship for Ozone (O_3) and humidity level for both stations appeared at around the same hour which are at the midnight. Generally, it can be said that all the variables have an indirect association between Ozone (O_3) across the day since at certain times their association is positive and in certain have a negative association.



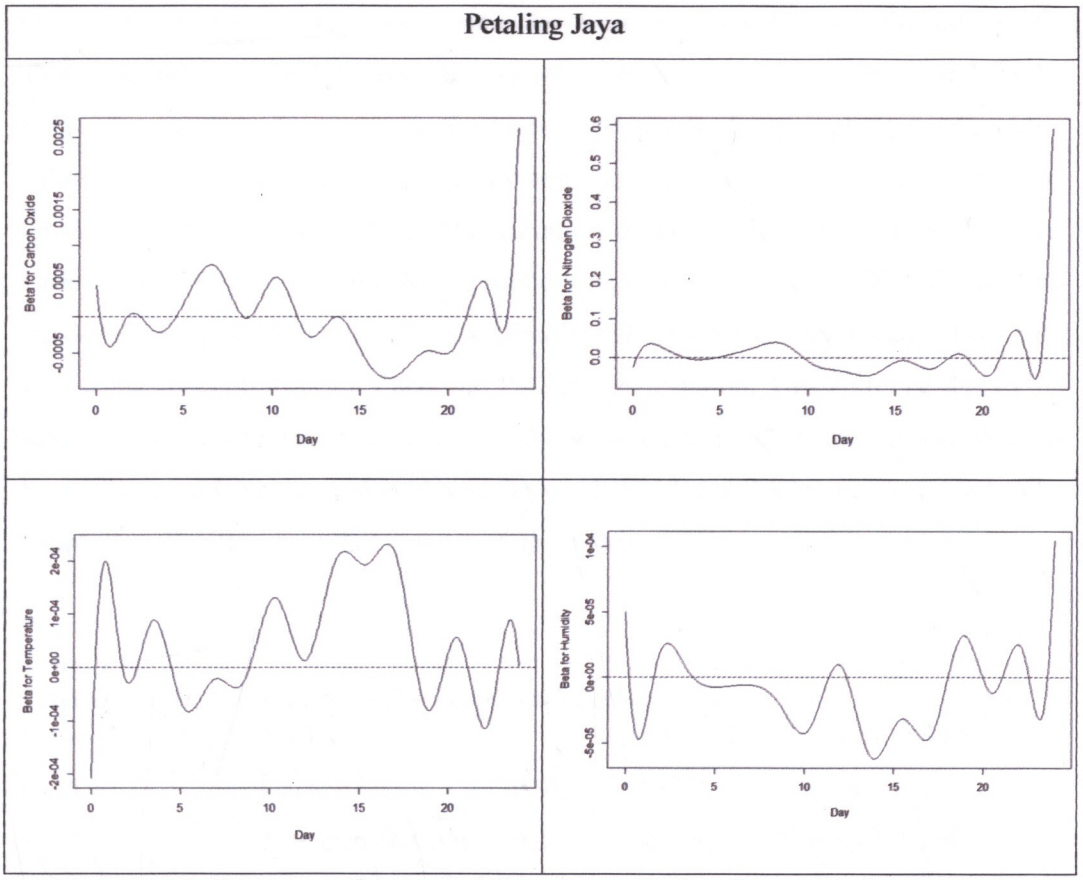


Figure 4.12: Summary of estimated Beta for predicting Ozone levels with outlier from each variable in both Air Quality Monitoring Station

A summary of functional regression result of R^2 and F-ratio with outlier for both stations is given in Table 4.6 below. Since the analysis considers of including the outlier, this means that this analysis considers the extreme concentration level of Ozone. In this analysis, squared correlation R^2 and F-ratio test will be used. R^2 or also called as the multiple squared correlation is a statistical measure that represents the proportion and can be changed to a percentage of the total variance in the dependent variables that influenced by the existence of independent variables. It is useful since the correlation can determine the relationships between variables. F-ratio is a test statistic to determine the means between two populations are significantly different or not significantly different. Based on the Perai station, the highest R-squared has shown for variable Nitrogen Dioxide in the first place, followed by variable temperature, humidity and Carbon Oxide (CO) with R^2 value 0.32, 0.29, 0.09 and 0.08. This means that 33% Ozone concentration is explained by Nitrogen Dioxide concentration level. Meanwhile, for temperature variable will explain 29% of the

Ozone concentration level. The F-ratio for Nitrogen Dioxide and temperature also shows higher compared to the other variables with F-ratio value 88.78 and 74.22. While for Petaling Jaya station, the highest R^2 result shows highest for humidity variable which is 0.23 and followed by temperature, Nitrogen Dioxide and Carbon Oxide. This means that 23% of the variation of Ozone is explained by humidity variable. The F-ratio for humidity is higher compared to other variable followed by temperature, Nitrogen Dioxide and Carbon Oxide. Different factor that influencing the Ozone concentration for Perai and Petaling Jaya station. This could be due to the different environmental background condition and geographical area where Perai is located near to beach and affect from the wind speed and the temperature tend to be higher than Petaling Jaya that is far from the beach.

Table 4.6:
Summary of functional regression result with outlier data

Station	Variable	R^2	F-ratio	F critical values
Perai	Carbon Oxide	0.08	16.79	$F_{(18,3268)} \approx 1.5$
	Nitrogen Dioxide	0.33	88.78	$F_{(18,3268)} \approx 1.5$
	Temperature	0.29	74.22	$F_{(18,3268)} \approx 1.5$
	Humidity	0.09	18.51	$F_{(18,3268)} \approx 1.5$
Petaling Jaya	Carbon Oxide	0.04	8.83	$F_{(16,3270)} \approx 1.5$
	Nitrogen Dioxide	0.16	38.82	$F_{(16,3270)} \approx 1.5$
	Temperature	0.21	54.05	$F_{(16,3270)} \approx 1.5$
	Humidity	0.23	59.84	$F_{(16,3270)} \approx 1.5$

Since the F-statistic value for all model is more than the F-table (critical value), thus, it is shown that the model are significant.

II. Without Outlier

For this analysis, a dataset without influencing outlier has been used. Sometimes, influencing outlier tend to be bias that will influence the goodness of fit results. A multivariate model approach that is called as cook's distance is used in order to remove influencing outlier in this analysis. Table 4.7 shows a summary of outlier percentages in Perai and Petaling Jaya station. Extreme values or influencing outliers is showing the high and extreme Ozone (O_3) concentration that will harm people. Removing the influencing outliers will give the norm Ozone (O_3) level, which mean the Ozone level is at normal level. A function in the base package called cooks.distance is used to detect the influential outlier. The summary of the outlier percentages is shown in the Table 4.7 below.

Figure 4.13 below shows the influential observations by cook's distance for Perai and Petaling Jaya station.The results are given in Table 4.7 for the summary of outlier percentages and shown graphically in Figure 4.14 below. In the beginning all the data set consists of 3287 days, which is 9 years (2009-2017). The outlier is based on the Ozone (O_3) level concentration. Perai has shown the highest number of influencing outlier with 105 out of 3287 which is 3.19%. The second station is Petaling Jaya where it shows the number of influencing outliers is 68 out of 3287 which is equivalent to 2.09%. For Figure 4.13 below, the red dots present the outliers that appeared in the dataset and the number of observation is shown beside the red dots. Initially, the total number of data set for both stations are 3287 days and after removing the outlier the total number of days for Perai station is 3182 days while for the Petaling Jaya station is 3219 days. The red dots appear depart far from the black dots is generally known as the influencing outlier.

Table 4.7:
Summary of Outlier Percentages

Stations	Before removing outlier (data)	Influencing outlier	After removing outlier (data)	Percentage of removing (%)
Perai	3287	105	3182	3.19
Petaling Jaya	3287	68	3219	2.09

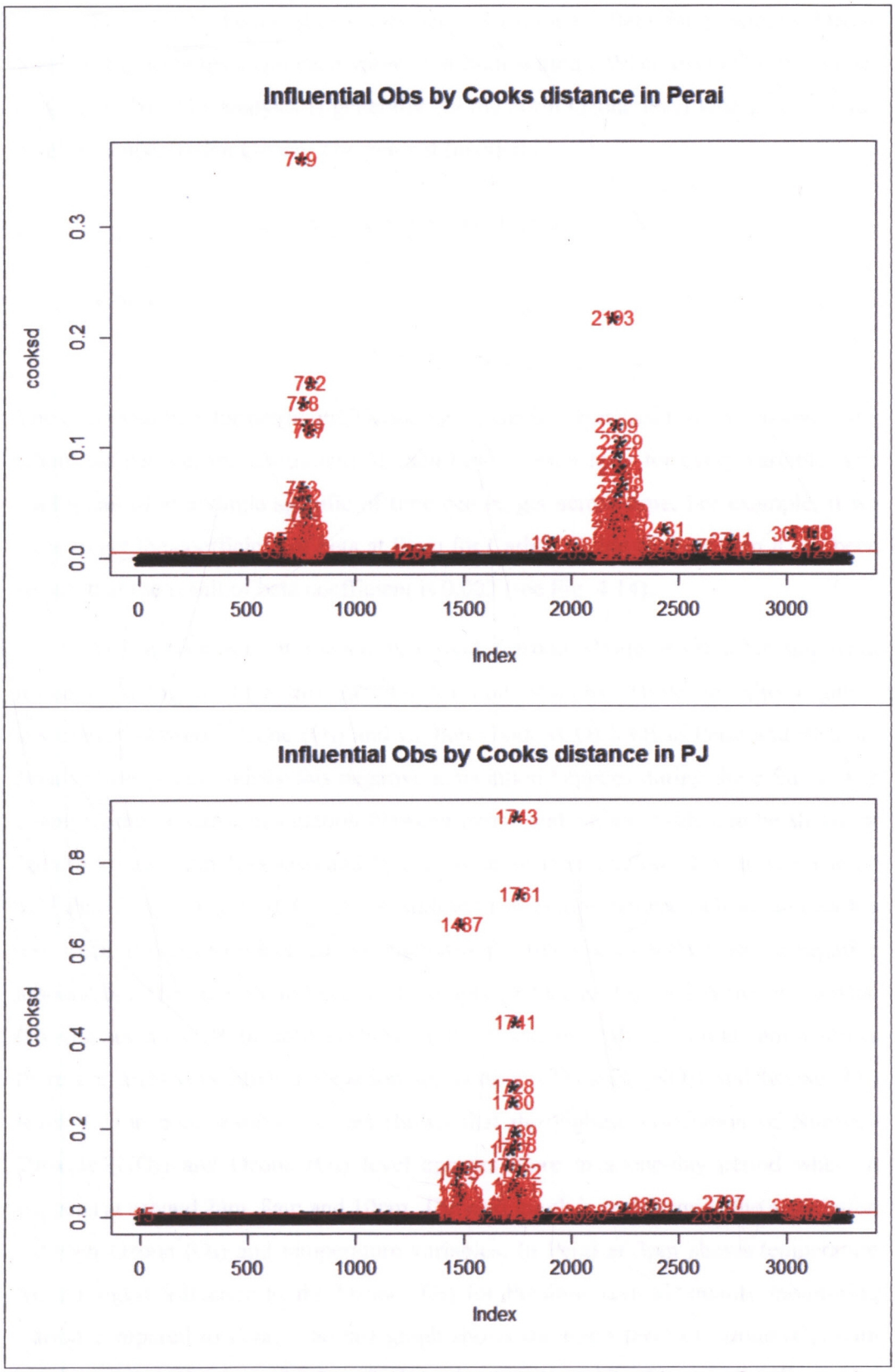


Figure 4.13: Summary of Influential Observations by Cooks Distance for Every Air Quality Monitoring Stations

Figure 4.14 below shows summary of estimated Beta for predicting Ozone levels without outlier from each variable at both stations. When excluding the outlier of Ozone (O_3), this analysis is generally for the norm Ozone level where no extreme level of concentration Ozone. The general function is

$$y_i = \propto + \int \beta(t)x_i(t)dt + \varepsilon_i$$

Where,

$$t = t_1, t_2, t_3, t_4 \dots t_n,$$

The estimated beta for predicting Ozone levels can be explained from the below graph where we can get the coefficient of each time or each hour for every variable. The coefficient of at a single specific of time can be get across time. For example, if we want to get the coefficient of beta at Perai for Carbon Oxide (CO) at 7am it has been shown that the result of beta coefficient is 0.003 (see Fig. 4.14).

The association of Ozone (O_3) and Carbon Oxide (CO) after removing influence outlier is high around 7am for both stations. There are also negative association between Ozone (O_3) and Carbon Oxide (CO) level in Perai and Petaling Jaya station, where mostly this negative association happens during the evening. For example, the negative association between ozone and carbon oxide can be shown at 3pm to 8pm for Petaling Jaya and 5pm to 10pm at Perai station. Generally, it can be said that Carbon Oxide (CO) has an indirect association between Ozone across the day since at certain times their association is positive and in certain have a negative association. This also same goes to the graph for Ozone (O_3) and Nitrogen Dioxide (NO_2) association. Both stations show similar fluctuation where around 7am it shows there are highest positive association for Nitrogen Dioxide (NO_2) and Ozone (O_3) level. In the both stations, it was shown that the highest association of Nitrogen Dioxide (NO_2) and Ozone (O_3) level appears more in a one-day period where it happens at around 2am, 8am and 10pm. The next graph is to determine the association between Ozone (O_3) and temperature variables. In Perai at 3pm shows temperature has strongest influence to the Ozone (O_3) for Petaling Jaya air quality monitoring station compared to Perai. The last graph shows the norm level of Ozone (O_3) with humidity level. On the average in both stations, the humidity has strong influence to the Ozone (O_3) level at early in the morning.

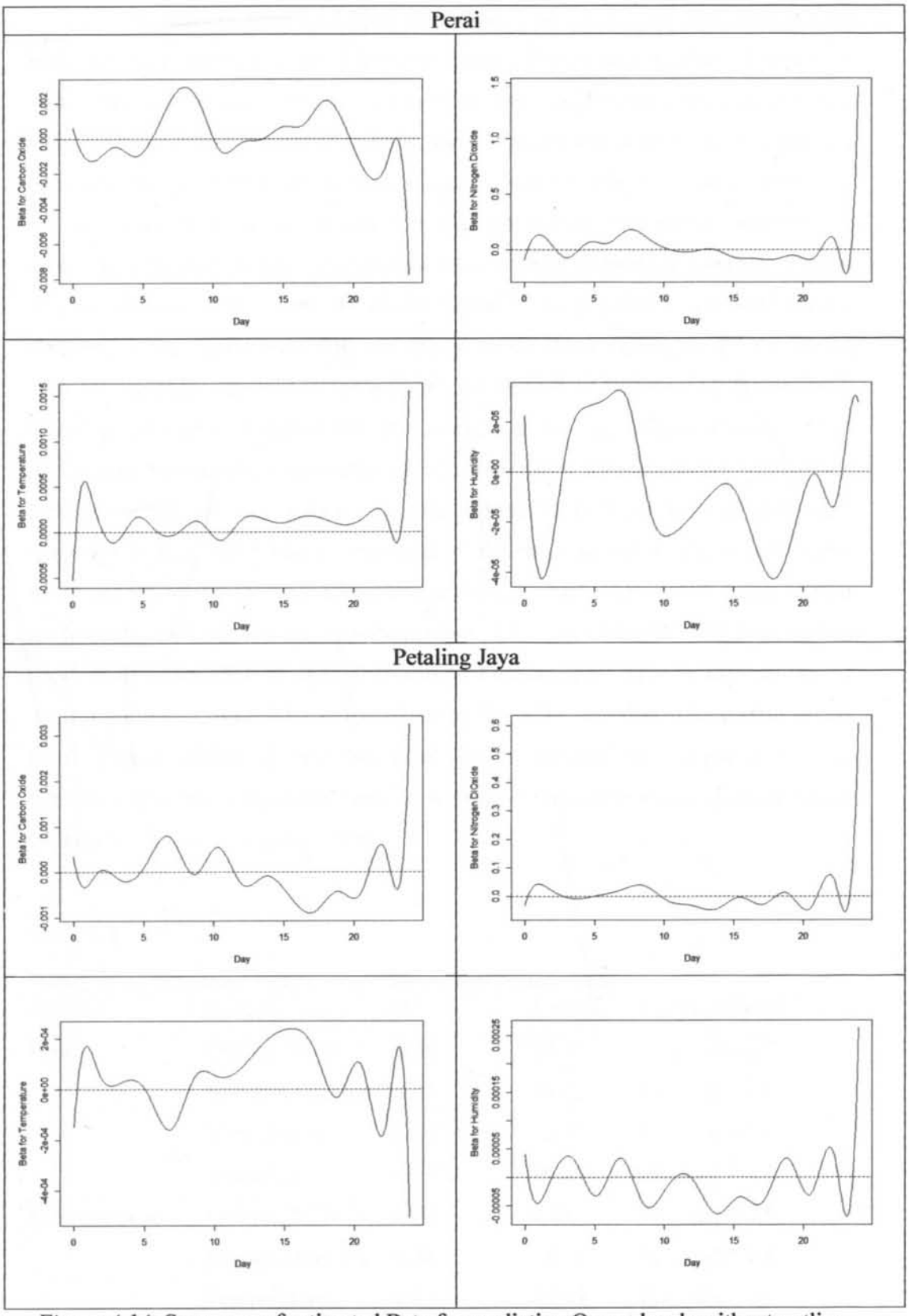


Figure 4.14: Summary of estimated Beta for predicting Ozone levels without outlier from each variable at both stations

A summary of functional regression result of R^2 and F-ratio without outlier for both stations is given in Table 4.8 below. Since the analysis considers of removing influencing outlier, this mean that this analysis only considers the norm concentration level of Ozone. R^2 or also called as the squared correlation is a statistical measure that represents the proportion and can be changed to a percentage of the total variance in the dependent variables that influenced by the existence of independent variables. It is useful since the correlation can determine the relationships between variables. F-ratio is a test statistic to determine the means between two populations are significantly different or not significantly different. Based on the Perai station, the R^2 has shown high for variable Nitrogen Dioxide (NO_2) which is 0.33 followed by temperature, humidity and Carbon Oxide (CO). This means that 33% of Nitrogen Dioxide (NO_2) will explain Ozone (O_3) concentration level. Meanwhile, for temperature will explain 29% of the Ozone concentration level. The F-ratio for Nitrogen Dioxide (NO_2) and temperature also shows higher compared to the other variables with F-ratio value 86.22 and 73.46. While for Petaling Jaya station, the highest R^2 result shows highest for humidity variable followed by temperature, Nitrogen Dioxide (NO2) and Carbon Oxide (C0) which are 0.24, 0.23, 0.16 and 0.04 values of R^2. This means that 24% of the variation of Ozone (O_3) is explained by humidity variable. The F-ratio in the Petaling Jaya station of humidity level also shows highest compared to other variables. This means humidity variable is more significant in influencing the Ozone concentration level in Petaling Jaya.

Table 4.8:

Summary of functional regression result without outlier data

Station	Variable	R^2	F-ratio	F critical value
Perai	Carbon Oxide	0.09	17.02	$F_{(18,3163)} \approx 1.5$
	Nitrogen Dioxide	0.33	86.22	$F_{(18,3163)} \approx 1.5$
	Temperature	0.29	73.47	$F_{(18,3163)} \approx 1.5$
	Humidity	0.11	20.65	$F_{(18,3163)} \approx 1.5$
Petaling Jaya	Carbon Oxide	0.04	8.84	$F_{(16,3202)} \approx 1.5$
	Nitrogen Dioxide	0.16	38.54	$F_{(16,3202)} \approx 1.5$
	Temperature	0.23	59.53	$F_{(16,3202)} \approx 1.5$
	Humidity	0.24	63.56	$F_{(16,3202)} \approx 1.5$

Since the F-statistic value for all model is more than the F-table (critical value), thus, it is shown that the model is significant.

4.5 Comparison Result data with Outlier and Without Outlier

I. R^2

Table 4.9 summarizes the comparison of the R-squared for data with and without the outlier. Overall, the result shows the R-squared is slightly different whenever the analysis is analyzing with or without influencing extreme Ozone concentration level (with or without influencing outlier). Based on Perai station, it has shown that the value of R^2 of Nitrogen Dioxide is same in both with and without outlier which is 0.33. There is only a slightly differences for the correlation coefficient result between the data with influencing outlier and without influencing outlier for the overall result. Same goes to temperature variable at Perai station where without outlier the R^2 value is 0.2948 for both datasets (with and without outlier). It is shown for both dataset, 29% of variation of ozone is explained by temperature. Meanwhile, for Carbon Oxide shows a slightly difference where when using with outlier data is 0.08 while using without outlier dataset is 0.09. Humidity variable also experiences the same situation where only a slight different when using both dataset which is 0.09 for with outlier and 0.11 for without outlier dataset. Same situation goes to the Petaling Jaya station where the R^2 is slightly higher in without influence outlier data. The highest give influence to Ozone in Petaling Jaya station is the temperature variable followed by humidity, Nitrogen Dioxide (NOx) and Carbon Oxide (CO). Meanwhile, at Perai station, the highest R^2 value is goes to Nitrogen Dioxide (NOx) followed by temperature, humidity and lastly Carbon Oxide (CO). This shows that, the functional regression model is actually a robust model to Malaysia dataset as well as there is only slight difference in the result whenever using with or without the outlier. R-squared value of this study is to determine the association of Ozone and the variable. However, the R-squared value cannot do the prediction since the R-squared value is small that is not suitable for it.

Table 4.9
Comparison R-squared with outlier and without influencing outlier

Station	Variable	With outlier	Without outlier
Perai	Carbon Oxide	0.08	0.09
	Nitrogen Dioxide	0.33	0.33
	Temperature	0.29	0.29
	Humidity	0.09	0.11
Petaling Jaya	Carbon Oxide	0.04	0.04
	Nitrogen Dioxide	0.16	0.16
	Temperature	0.21	0.23
	Humidity	0.23	0.24

II. F-Ratio

Table 4.10 summarizes the comparison of F-ratio for data with and without the influencing outlier. Overall, all the variables in both stations give high F-ratio value when using the without influencing outlier except for the Nitrogen Dioxide (NO_2) variable. Both stations have shown a high value of F-ratio of Nitrogen Dioxide (NO_2) variable when including extreme Ozone level of concentration. The result contradicts with the other variables where when using without outlier data is more F-ratio value result. At Perai station, the F-ratio for Nitrogen Dioxide is the highest when using data with outlier with F-Ratio value 88.78 followed by temperature, humidity and Carbon Oxide (CO). Same goes to for Perai station when using a clean without influence outlier data where the highest F-Ratio goes to Nitrogen Dioxide with 86.22 followed by temperature, humidity and Carbon Oxide (CO). While in Petaling Jaya station, when using data with outlier, the humidity tends to be the highest F-Ratio which is 59.84 followed by temperature, Nitrogen Dioxide (NO_2) and Carbon Oxide (CO). Same result rank for the F-ratio value for without outlier data in Petaling Jaya, it shows different where the highest F-Ratio value goes to humidity with 63.56 followed by temperature, Nitrogen Dioxide (NO_2) and Carbon Oxide (CO). Overall, as shown in table 4.10, Carbon Oxide and humidity gives a more significant result with using without the outlier data in Perai meanwhile in Petaling Jaya more significant result whenever use without outlier data for variable Carbon Oxide, temperature and humidity. There is only slightly different from the result using with or without the

outlier. This analysis support that functional model is a robust model. Therefore, it shows that the functional regression model is actually a robust model to Malaysia dataset as well as there is only a slight difference to the result whenever using with or without influencing outlier. The result does not give big changes or impact whenever using with or without outlier dataset.

Table 4.10

Comparison F-ratio with outlier and without influencing outlier

Station	Variable	With outlier	Without outlier
Perai	Carbon Oxide	16.80	17.01
	Nitrogen Dioxide	88.78	86.22
	Temperature	74.22	73.47
	Humidity	18.51	20.65
Petaling Jaya	Carbon Oxide	8.83	8.84
	Nitrogen Dioxide	38.82	38.54
	Temperature	54.05	59.53
	Humidity	59.84	63.56

CHAPTER 5

CONCLUSION AND RECOMMENDATIONS

5.1 Introduction

This chapter will be discussed in the summary of all the findings that has been obtained based on the analysis that have been conducted according to the research objectives of this study. Other than that, some recommendation or suggestions are made to be used for a better future research study for the future researcher.

5.2 Conclusion

There are three objectives of this study that need to be concluded in this section where the first one is to describe the diurnal and spatial behavior of Ozone (O3), the precursors (CO, NO2) and meteorological variables (temperature and humidity) at Petaling Jaya and Perai. The second objective is to investigate the diurnal inter - association pattern between O3 and the precursors as well as meteorological variables at Petaling Jaya and Perai. The third objective is model the diurnal relationship between Ozone (O3) and the precursors (CO and NO2) as well as the meteorological variables (temperature and humidity) using Bivariate Functional Linear Regression at Petaling Jaya and Perai. This study was conducted using hourly data for 3287 days which is equivalent to nine years that is from 2009 to 2017. In order to achieve the objectives of this study, the main initial step is to get the functional of curves data by converting the observed data. A Regression Smoothing methodology is used to obtain the curves by means of basis expansion approach with K number of cubic B-spline basis function method. The appropriate number of K for the study is different for each location and each variable. A set of curve data was obtained to visualize the physical fluctuation for nine years daily, hourly data which is from the year 2009 to the year 2017. The special of curves data is that the concentration variable for each variable can be obtained at any time within the day

period. Furthermore, the fluctuation of the whole data can be seen through the entire days. Based on the all set curves obtained, the figures show a slight difference in concentration level of each variable but quite similar pattern for both stations. The same peak tends to appear at around 3 pm in the evening. Petaling Jaya station was observed to have higher peak for Ozone (O_3), Carbon Oxide (CO), Nitrogen Dioxide (NO_2) and temperature variable compared to Perai station. Meanwhile the humidity level tends to have valley at the curves with the lowest humidity concentration level at 3 pm for both stations.

The functional median curve was used to visualize the pattern of the day to day variation. Based on the result, both Perai and Petaling Jaya station depicted same pattern for each same variable where a unimodal pattern was shown for variable Ozone (O_3) and temperature, bimodal pattern was shown for variable Carbon Oxide and Nitrogen Dioxide (NO_2), while for humidity variable showed U-shape pattern for the median curve as well as mean curve. Besides median curve, the mean curve was used to visualize the average pattern of variables in day to day. The functional median curve has shown that Perai station has slightly higher concentration level in Ozone (O_3) variable compared to the Petaling Jaya station. Meanwhile, for Carbon Oxide (CO) variable, the functional median curve appeared to be higher in Petaling Jaya station air quality monitoring station with 1.4 ppbv compared to Perai station which is 0.9 ppbv. Same to Nitrogen Dioxide (NO_2) variable where Petaling Jaya to be highest in the level of Nitrogen Dioxide (NO_2) concentrations. For temperature and humidity variables, both tend to have the same range for peak and valley of the concentration in Perai and Petaling Jaya station. This result has the same pattern for the functional mean curve. In conclusion, for functional median and mean curves, air pollution factors tend to be high in Petaling Jaya station since Ozone (O_3), Carbon Oxide (CO) and Nitrogen Dioxide (NO_2) showed higher level concentration compared to Perai station. This may happen due to its location as a background station, where Petaling Jaya is a high polluted station that is believed has a higher number of people and a higher number of emissions comes from vehicles. It also believed that the level of concentration of pollutant in Perai is much lower since it is affected by the wind speed as Perai is located near to the sea. Functional standard deviation curves are used to describe the variation of Ozone dispersion through 24 hours, one-day period. Based on the functional standard deviation curves, results shown that Petaling Jaya station

experience the highest hourly dispersion of all variable levels. The result also shown that both stations appeared similar dispersion occurred for Ozone (O_3), temperature and humidity variables have a peak level at 3pm.

The next objective is to investigate the diurnal inter - association pattern between O_3 and the precursors as well as meteorological variables at Petaling Jaya and Perai. Function corr.fd in fda package was used to determine the correlation of Ozone with the variables. Temperature variable showed as the most influential factor with the Ozone (O_3) variable since the correlation coefficient of the temperature with Ozone is the highest compared to other variables in both Perai and Petaling Jaya station. During around 5 am, the correlation coefficient curves of temperature with Ozone (O_3) tends to be the lowest around 0.2. Meanwhile the least related to Ozone variable is a humidity variable since the result of correlation coefficient curves based on the graph showed that they have a negative relationship. This means that, if the Ozone (O_3) is at a high level, the humidity will be at low concentration. The highest peak for correlation curves between humidity and Ozone (O_3) can be found around 8 am for Perai and Petaling Jaya station.

The third objective to model the diurnal relationship between Ozone (O3) and the precursors (CO and NO2) as well as the meteorological variables (temperature and humidity) using Bivariate Functional Linear Regression at Petaling Jaya and Perai. In this analysis, R^2 and F-ratio test was used. The functional regression is tested with two data that are with outlier and without outlier data. With outlier data means that the analysis will include the extreme values of Ozone (O_3) whereas without outlier data means that the analysis will be conducted at a norm Ozone level. R-squared value of this study is to determine the amount of variation in the predictive variable. Overall the R^2 value of data without outlier showed slightly higher values than testing with the norm level of Ozone (O_3) that is with outlier data but not much different. In R-squared values, the variables Nitrogen Dioxide and temperature in Perai station showed the same R-squared value either using with or without outlier. While in Petaling Jaya, Carbon Oxide and Nitrogen Dioxide has showed the same value either using with or without outlier dataset. It also can be showed that Perai gives a higher result for all variables except the humidity variable when compared to the Petaling Jaya station. Overall, when using Ozone (O_3) data with outlier gives slightly similar result to both station and for each the variable but higher in without influencing the data. This mean,

both datasets does not give big influence to the model performance. However, the model cannot be used for the prediction since the R-squared value is small. Next, for the F-ratio test showed that all variables are shown significant at 5% level. In conclusion, it can be said that for this study there is only slightly different for the result for the models either with or without outlier.

5.3 Recommendations

As recommendations, several interesting aspects may be explored further for a better future research. For example, increasing the number of air quality monitoring station in the industrial area to get a broader and better result across one Malaysia. The stations that are included in industrial areas are Pasir Gudang, Shah Alam, Nilai, Paka, Tanjung Malim and Bukit Rambai. More stations across the Malaysia can give a conclusion for Ozone (O_3) pattern in an industrial area in Malaysia. Other than that, multiple functional regression can be used for further research. Since in this current situation, only bivariate linear functional regression is common and readily being used the method as the theory of functional multiple regression is being developed. Other than that, the number of sample size can be increased to get a wide time range such as for 20 years or 30 years. Greater sample size could give more accurate and reliable estimation with greater precision and power.

REFERENCES

(MAA), M. A. A. (2019). Market review for 2018 and outlook for 2019.

Abdullah, A. M., Samah, M. A. A., & Jun, T. Y. (2012). An Overview of the Air Pollution Trend in Klang Valley, Malaysia. *Open Environmental Sciences, 6*, 13-19.

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

Ahamad, F., Latifa, M. T., Dominick, D., Juahir, H., Tang, R., & Juneng, L. (2014). Variation of surface ozone exceedance around Klang Valley, Malaysia. *Atmospheric Research, 139*, 116-127.

Ashaha, N. R. N. (2019). Jerebu: Rompin, Shah Alam catat IPU tak sihat. *Sinar Harian.*

Awang, Elbayoumi, M., Ramli, N. A., & Yahaya, A. S. (2015). Diurnal variations of ground-level ozone in three port cities in Malaysia. *Air Qual Atmos Health.* doi:10.1007/s11869-015-0334-7

Awang, M. B., Jaafar, A. B., Abdullah, A. M., Ismail, M. B., Hassan, M. N., Abdullah, R., . . . Noor, H. (2001). Air quality in Malaysia: Impacts, management issues and future challenges. *ResearchGate, 5*, 183-196.

Azam, A. G., Zanjani, B. R., & Mood, M. B. (2016). Effects of air pollution on human health and practical measures for prevention in Iran. *J Res Med Sci, 21*, 65. doi:10.4103/1735-1995.189646

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 Qual Atmos Health, 3*, 53-64. doi:10.1007/s11869-009-0051-1

Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple Imputation by Chained Equations: What is it and how does it work? *NIH Public Access*, 40-49. doi:10.1002/mpr.329.

Banan, N., & Latif, M. T. (2011). An Investigation into Ozone Concentration at Urban and Rural Monitoring Stations in Malaysia. *World Academy of Science, Engineering and Technology, 77.*

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

Brugha, R., Edmondson, C., & Davies, J. C. (2018). Outdoor air pollution and cystic fibrosis. *Paediatr Respir Rev, 28*, 80-86. doi:10.1016/j.prrv.2018.03.005

DOSM. (2018a, 31 July 2018). Current Population Estimates, Malaysia, 2017-2018. Retrieved from https://www.dosm.gov.my/v1/index.php?r=column/cthemeByCat&cat=155&bul_id=c1pqTnFjb29HSnNYNUpiTmNWZHArdz09&menu_id=L0pheU43NWJwRWVSZklWdzQ4TlhUUT09

DOSM. (2018b, 25 June 2019). Selangor. *Malaysia at a glance.* Retrieved from https://www.dosm.gov.my/v1/index.php?r=column/cone&menu_id=eGUyTm9RcEVZSllmYW45dmpnZHh4dz09

Du, Y., Ma, C., Wu, C., Xu, X., Guo, Y., Zhou, Y., & Li, J. (2017). A Visual Analytics Approach for Station-Based Air Quality Data. *Sensors, 17*(30), 1-17. doi:10.3390/s17010030

Febrero-Bande, M., & Fuente, M. O. d. l. (2012). Statistical Computing in Functional Data Analysis: The R Package fda.usc. *JournalofStatisticalSoftware, 51*(4).

Fuente, M. O. d. l., Febrero-Bande, M., Muñoz, M. a. P., & Domı´nguez, À. (2018). Predicting seasonal influenza transmission using functional regression models with temporal dependence. *PLOS One*, 18. doi:https://doi.org/10.1371/journal.pone.0194250

Gervini, D. (2012). Functional robust regression for longitudinal data. *ResearchGate*, 21.

Hooker, G. (2017). Functional Data Analysis A Short Course.

Huang, J. Z., & Shen, H. (2004). Functional Coefficient Regression Models for Nonlinear Time Series: A Polynomial Spline Approach. *Scandinavian Journal of Statistics 31, 31*(4). doi:https://doi.org/10.1111/j.1467-9469.2004.00404.x

Hullait, H., Leslie, D. S., Pavlidis, N. G., & King, S. (2019). Robust Function-on-Function Regression. *ResearchGate*, 34.

Kalogridis, I., & Aelst, S. V. (2018). Robust functional regression based on principal components. *ResearchGate*, 33.

Kang, H. (2013). The prevention and handling of the missing data.

Kjellstrom, T., Lodh, M., McMichael, T., Ranmuthugala, G., Shrestha, R., & Kingsland, S. Air and Water Pollution: Burden and Strategies for Control. In *Disease Control Priorities in Developing Countries* (pp. 817-832).

Korkmaz, S., Goksuluk, D., & Zararsiz, G. (2014). MVN: An R Package for Assessing Multivariate Normality. *R Journal, Vol. 6/2*, 151-162.

Lancet, T. (2016). Air pollution: consequences and actions for the UK, and beyond. *The Lancet, 387*(10021). doi:10.1016/s0140-6736(16)00551-1

Latif, M. T., Dominick, D., Ahamad, F., Khan, M. F., Juneng, L., Hamzah, F. M., & Nadzir, M. S. M. (2014). Long term assessment of air quality from a background station on the Malaysian Peninsula. *Science of the Total Environment*, 336–348. doi:http://dx.doi.org/10.1016/j.scitotenv.2014.02.132

Li, Ho, S. S. H., Gong, S., Ni, J., Li, H., Han, L., . . . Zhao, D. (2019). Characterization of VOCs and their related atmospheric processes in a central Chinese city during severe ozone pollution periods. *Atmosheric Chemistry and Physics, 19*, 617-638. doi:https://doi.org/10.5194/acp-19-617-2019

Li, H., Fan, H., & Mao, F. (2016). A Visualization Approach to Air Pollution Data Exploration—A Case Study of Air Quality Index (PM2.5) in Beijing, China. *Atmosphere, 7*(35), 1-20. doi:10.3390/atmos7030035

Liu, J., Li, W., Wu, J., & Liu, Y. (2018). Visualizing the intercity correlation ofPM2.5 time series in the Beijing-Tianjin-Hebei region using ground-based air quality monitoring data. *PLOS One, 13*(2), 1-14. doi:https://doi.org/10.1371/journal.pone.0192614

M¨uller, H.-G., & Stadtm¨uller, U. (2005). Generalized Functional Linear Models. *The Annuals of Statistics, 33*(2), 774-805. doi:10.1214/009053604000001156

Mackenzie, J. (2016). Air Pollution: Everything You Need to Know. Retrieved from https://www.nrdc.org/stories/air-pollution-everything-you-need-know#sec1

Madhoun, W. A. A., Ramli, N. A., & Yahaya, A. S. (2012). Monitoring the Total Volatile Organic Compounds (TVOCs) and Benzene Emitted at Different Locations in Malaysia. *Journal of Engineering Science, Vol. 8*, 59–67.

Martínez, J., Saavedra, Á., García-Nieto, P. J., Piñeiro, J. I., Iglesias, C., Taboada, J., . . . a, J. P. (2014). Air quality parameters outliers detection using functional data analysis in the Langreo urban area (Northern Spain). *Applied Mathematics and Computation, 241*, 1-10. doi:http://dx.doi.org/10.1016/j.amc.2014.05.004

Masselot, P., Chebana, F., Ouarda, T. B. M. J., Bélanger, D., St-Hilaire, A., & Gosselin, P. (2018). A new look at weather-related health impacts through functional regression. *Scienctific reports, 8*(1), 9. doi:10.1038/s41598-018-33626-1

Mofijur, M., Rasul, M. G., Hyde, J., Azad, A. K., Mamat, R., & Bhuiya, M. M. K. (2016). Role of biofuel and their binary (diesel–biodiesel) and ternary (ethanol–biodiesel–diesel) blends on internal combustion engines emission reduction. *Renewable and Sustainable Energy Reviews, 53*, 265-278. doi:http://dx.doi.org/10.1016/j.rser.2015.08.046

Mohammed, N. I., Ramli, N. A., & Yahya, A. S. (2013). Ozone phytotoxicity evaluation and prediction of crops production in tropical regions. *Atmospheric Environment, 68*, 343-349. doi:http://dx.doi.org/10.1016/j.atmosenv.2012.09.010

Monteiro, A., Strunk, A., Carvalho, A., Tchepel, O., Miranda, A. I., Borrego, C., . . . Elbern, H. (2012). Investigating a high ozone episode in a rural mountain site. *Environmental Pollution, 162*, 176-189. doi:10.1016/j.envpol.2011.11.008

Morand, C. P., & Maesano, I. A. (2004). Air pollution: from sources of emissions to health effects*s1*(2), 109-119. Retrieved from

Pereira, I., Pablik, E., Schwarzhans, F., Resch, H., Fischer, G., Vass, C., & Frommlet, F. (2018). A Functional Regression Model of the Retinal Nerve Fiber Layer Thickness in Healthy Subjects. *translational vision science and technology (tvst), 7*(1), 1-12. doi:https://doi.org/10.1167/tvst.7.1.9

Ramsay, J. O., Hooker, G., & Graves, S. (2009). Functional Data Analysis with R and MATLAB. *1-203*. doi:10.1007/978-0-387-98185-7

Ramsay, J. O., & Silverman, B. W. (2006). *Functional Data Analysis*. United States of America: Springer Science+Business Media.

Rani, N. L. A., Azid, A., Juahir, H., Khalit, S. I., & Samsudin, M. S. (2018). Air Pollution Index Trend Analysis in Malaysia, 2010-15. *Polish Journal of Environmental Studies, 27*(2), 801-807. doi:10.15244/pjoes/75964

Ritchie, H., & Roser, M. (2017). Air Pollution. Retrieved from https://ourworldindata.org/air-pollution#empirical-view

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Canada: John Wiley & Sons.

Sadanaga, Y., Sengen, M., Takenaka, N., & Bandow, H. (2012). Analyses of the Ozone Weekend Effect in Tokyo, Japan: Regime of Oxidant (O3 + NO2) Production. *Aerosol and Air Quality Research, 12*, 161-168. doi:10.4209/aaqr.2011.07.0102

Schraufnagel, D. E., Balmes, J. R., Cowl, C. T., De Matteis, S., Jung, S. H., Mortimer, K., . . . Wuebbles, D. J. (2019). Air Pollution and Noncommunicable Diseases: A Review by the Forum of International Respiratory Societies' Environmental Committee, Part 2: Air Pollution and Organ Systems. *Chest, 155*(2), 417-426. doi:10.1016/j.chest.2018.10.041

Scoullos, M. (2013). *Education for sustainable development in biosphere reserves and other designated areas: a resource book for educators in South-Eastern Europe and the Mediterranean* (R. M. M. Anastasia Roniotes Ed.): United Nations Educational, Scientific and Cultural Organization.

Shaadan, N., Deni, S. M., & Jemain, A. A. (2015). Application of Functional Data Analysis for the Treatment of Missing Air Quality Data. *Sains Malaysiana, 44*(10), 1531–1540.

Shaadan, N., Jemain, A. A., & Deni, S. M. (2014). Data Preparation for Functional Data Analysis of PM10 in Peninsular Malaysia. doi:10.1063/1.4887701

Shaadan, N., Nazeri, L. N., Jalani, M. F. M., Rahman, N. F. A. A., & Roslan, R. R. R. R. (2017). Data Visualization Of Temporal Ozone Pollution Between Urban And Sub-Urban Locations In Selangor Malaysia. *Journal of Fundamental and Applied Sciences, 9*(6s), 490-507. doi:http://dx.doi.org/10.4314/jfas.v9i6s.37

Shah, M. F., & Nordin, R. (2019, 14 March 2019). Numbers rise: 2,775 people affected by Pasir Gudang chemical pollution. *The Star Online*. Retrieved from https://www.thestar.com.my/news/nation/2019/03/14/numbers-rise-2775-people-affected-by-pasir-gudang-chemical-pollution/

Sinharay, S., & Russell, D. W. (2001). The Use of Multiple Imputation for the Analysis of Missing Data. *Psychological Method, 6*(4), 317-329. doi:DOI: 10.1037/1082-989X.6.4.317

Smaga, Ł., & Matsui, H. (2018). A note on variable selection in functional regression via random subspace method. *Stat Methods Appl, 27*, 455–477. doi:https://doi.org/10.1007/s10260-018-0421-7

Torresa, J. M., Nietob, P. J. G., Alejanoc, L., & Reyesc, A. N. (2011). Detection of outliers in gas emissions from urban areas using functional data analysis.

Journal of Hazardous Materials, 186, 144-149. doi:10.1016/j.jhazmat.2010.10.091

Ullah, S., & Finch, C. F. (2013). Applications of functional data analysis: A systematic review. *BMC Medical Research Methodology, 13*(43), 1-13.

Vallero, D. A. (2008). *Fundamentals of Air Pollution FOURTH EDITION* (4 ed.): Elsevier Inc.

Wang, J.-L., Chiou, J.-M., & M¨uller, H.-G. (2015). Review of functional data analysis. 1-41. doi:10.1146

WHO. (2005). WHO Air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide, 22. Retrieved from

WHO. (2011) Public health, environmental and social determinants of health (PHE). In, *https://www.who.int/phe/health_topics/outdoorair/databases/cities-2011/en*: WHO.

WHO. (2016). *Ambient air pollution: A global assessment of exposure and burden disease*. In I. communication (Ed.), (pp. 121).

WHO. (2018, 2 May 2018). 9 out of 10 people worldwide breathe polluted air, but more countries are taking action. Retrieved from https://www.who.int/news-room/detail/02-05-2018-9-out-of-10-people-worldwide-breathe-polluted-air-but-more-countries-are-taking-action

Williams, M. (2016). What Causes Air Pollution? Retrieved from https://www.universetoday.com/81977/causes-of-air-pollution/

Yassen, M. E., Jahi, J. M., & Ahmad, S. (2005). Evaluation of Long Term Trends in Oxide of Nitrogen Concentrations in the Klang Valley Region, Malaysia. *Malaysian Journal of Environmental Management, 6*, 59-72.

Zhang, F., Xie, D., Liang, M., & Xiong, M. (2016). Functional Regression Models for Epistasis Analysis of Multiple Quantitative Traits. *PLOS Genetics*, 26. doi:10.1371/journal.pgen.1005965

APPENDICES

APPENDIX A

CODING

```
#####################################################################
                        Call data set
#####################################################################
o3_PJ=read.delim("clipboard",header=TRUE)
co_PJ=read.delim("clipboard",header=TRUE)
no2_PJ=read.delim("clipboard",header=TRUE)
temp_PJ=read.delim("clipboard",header=TRUE)
humidity_PJ=read.delim("clipboard",header=TRUE)

#####################################################################
                        # PETALING JAYA #
#####################################################################
                        Data Imputation
#####################################################################
PJ=read.delim("clipboard",header=TRUE)
str(PJ)
newPJ=PJ[,10:18]
str(newPJ)
imp=mice(newPJ)
newData=complete(imp,1)
summary(newData)

#####################################################################
                        Export data to Excell.csv file
#####################################################################
write.csv(c,"PJ_latest_mice.csv") #the .csv file will be saved in Document directory#

#####################################################################
                        Find RSS
#####################################################################
```

```
# DETERMINING THE OPTIMAL K-basis number #
#########################################################################
o3_PJ=read.delim("clipboard",header=TRUE)
agraval<-1:24
rssval<-function(rangval,agraval,k,o3_PJ)
{
 mean.day<-apply(o3_PJ,2,mean)
 x<-as.vector(mean.day)
 nk<-length(k)
 answer=numeric(nk)
 for(i in 1:nk)
 {basis6<-create.bspline.basis(rangval,k[i])
 SSE<-smooth.basis(agraval,x,basis6)$SSE
 answer[i]<-SSE}
 answer
}
rssval(rangval,agraval,20,o3_PJ)

#########################################################################
                              Find BIC
#########################################################################
#p=5:20
multiBIC=function(n,RSS,p)
{
 jaw=numeric()
 m=length(RSS)
 for (i in 1:m)
 {BIC=log(RSS[i]/n)+log(24)*(p[i]/n)
 jaw[i]=BIC}
 jaw
 ans=cbind(p,jaw)
 ans
}
multiBIC(24,6.778507e-08,20)s
```

```
#####################################################################
    PLOT GRAPH TO DETERMINE APPROPRIATE NUMBER OF BASIS, K
#####################################################################
rangval=c(1,24)
agraval=1:24
RSS.o3_PJ=rssval(rangval,agraval,5:20,o3_PJ)
p=5:20
x=p
bic.o3_PJ=multiBIC(24,RSS.o3_PJ,p)
plot(5:20,bic.o3_PJ[,2],xlim=c(5,20),main="Petaling Jaya",xlab="Number of
Basis(k)",ylab="BIC value",cex.lab=0.8,cex.main=1.0,cex.axis=0.8)
lines(x,bic.o3_PJ[,2])

#####################################################################
    CREATING FUNCTIONAL OBJECTS METHOD:LEAST SQUARE-
                REGRESSION SPLINE
#####################################################################
o3_PJ=read.delim("clipboard",header=TRUE)
rangeval=c(1,24)
argvals=1:24
nbasis=17
o3_PJ.basis=create.bspline.basis(rangeval,nbasis,norder=4,breaks=NULL,dropind=N
ULL,quadvals=NULL,values=NULL,basisvalues=NULL,names="bspl")
plot(o3_PJ.basis,main="Petaling Jaya",xlab="time(hour)")
o3_PJ_fd=smooth.basis(argvals,t(o3_PJ),o3_PJ.basis)$fd
plot(o3_PJ_fd,main="Petaling Jaya",xlab="time(hour)",ylab="Ozone level")
#####################################################################
                DESCRIPTIVE ANALYSIS
#####################################################################
library(matrixStats)
o3_PJ_med.col=colMedians(o3_PJ.matrix)
o3_PJ_med.fd=smooth.basis(argvals,o3_PJ_med.col,o3_PJ_basis)$fd
plot(o3_PJ_med.fd,main="Median Curve for Ozone level at Petaling Jaya station")
o3_PJ_sd.fd=sd.fd(o3_PJ_fd)
```

```
plot(o3_PJ_sd.fd,main="Standard Deviation Curve for Ozone level at PJ station")
o3_PJ_mean.fd=mean.fd(o3_PJ_fd)
plot(o3_PJ_mean.fd,main="Mean Curve for Ozone at Petaling Jaya station")
#########################################################################
                        FUNCTIONAL CORRELATION
#########################################################################
cor.PJ.CO_1<-cor.fd(1:6,co_PJ_fd,1:6,o3_PJ_fd)
cor.PJ.CO_2<-cor.fd(7:12,co_PJ_fd,7:12,o3_PJ_fd)
cor.PJ.CO_3<-cor.fd(13:18,co_PJ_fd,13:18,o3_PJ_fd)
cor.PJ.CO_4<-cor.fd(19:24,co_PJ_fd,19:24,o3_PJ_fd)
plot(cor.PJ.CO)
cor.PJ.CO
plot(cor_o3nco,pch=8,col="blue",type="b",main="Correlation Curve between Ozone
and the variables in Petaling Jaya",ylim=c(-0.5,0.6),ylab="Correlation Coefficient")
#########################################################################
                    FUNCTIONAL REGRESSION WITH OUTLIER
#########################################################################
#Temperature and Ozone
conbasis=create.constant.basis(c(0,24))
betabasis=create.bspline.basis(c(0,24),17)
betalist=vector("list",2)
betalist[[1]]=conbasis
betalist[[2]]=betabasis
fRegressList=fRegress(annualprec,templist,betalist)
plot(tempbetafd,xlab="Day",ylab="Beta for Temperature")
coef(betaestlist[[1]])
annualprechat1=fRegressList$yhatfdobj
annualprecres1=annualprec-annualprechat1
SSE1.1=sum(annualprecres1^2)
SSE0=sum((annualprec-mean(annualprec))^2)
RSQ1=(SSE0-SSE1.1)/SSE0
Fratio=((SSE0-SSE1.1)/16)/(SSE1.1/3270)
```

```
#########################################################################
                    FUNCTIONAL REGRESSION WITHOUT OUTLIER
#########################################################################
#cooks distance without influence outlier in PJ
mod=lm(Hour1~.,data=o3_PJ)
cooksd=cooks.distance(mod)

#detect outlier
plot(cooksd,pch="*",cex=2,main="Influential Obs by Cooks distance in PJ") # plot
cook's distance
abline(h = 4*mean(cooksd, na.rm=T), col="red") # add cutoff line
text(x=1:length(cooksd)+1, y=cooksd, labels=ifelse(cooksd>4*mean(cooksd,
na.rm=T),names(cooksd),""), col="red") # add labels
influential <- as.numeric(names(cooksd)[(cooksd > 4*mean(cooksd, na.rm=T))])#
influential row numbers

#Regression
o3_PJ_new=t(newdata_o3PJ)
conbasis=create.constant.basis(c(0,24))
betalist=vector("list",2)
betalist[[1]]=conbasis
betalist[[2]]=betabasis
fRegressList=fRegress(annualprec,templist,betalist)
betaestlist=fRegressList$betaestlist
tempbetafd=betaestlist[[2]]$fd
plot(tempbetafd,xlab="Day",ylab="Beta for Temperature")
annualprechat1=fRegressList$yhatfdobj
annualprecres1=annualprec-annualprechat1
SSE1.1=sum(annualprecres1^2)
SSE0=sum((annualprec-mean(annualprec))^2)
RSQ1=(SSE0-SSE1.1)/SSE0
Fratio=((SSE0-SSE1.1)/16)/(SSE1.1/1563)
```

APPENDIX B

Summary of Research Findings

Researchers	Title	Variables	Method	Findings
Shaadan et al. (2017)	Data Visualization Of Temporal Ozone Pollution Between Urban And Sub-Urban Locations In Selangor Malaysia	Ozone (O_3) exceedances in urban and sub-urban	Principal Component Analysis (PCA)	An increasing pattern of O_3 pollution occurrence with several modes of distinct diurnal variations at the locations. Banting experience a higher potential for O_3 pollution severity compared to Shah Alam
Banan et al. (2013)	Characteristics of Surface Ozone Concentrations at Stations with Different Backgrounds in the Malaysian Peninsula	Ozone concentration, Nitrogen Oxides (NO and NO_2), non-methane hydrocarbon (NMHC)	Statistical analysis: Normal P-P plot, normal Q-Q plot, One-sample Kolmogorov-Smirnov test, Analysis of variance, Bonferroni Trajectory Analysis: Backward trajectory analyses	The suburban areas had higher levels of O_3 concentration than urban and rural areas.
Li et al. (2019)	Characterization of VOCs and their related atmospheric processes in a central Chinese city during severe Ozone pollution	Ozone (O_3), Volatile organic compounds (VOCs), Nitrogen Oxide (NOx)	Chemical analysis, Positive matrix factorization (PMF), Potential source contribution function (PSCF)	O_3 formation was more sensitive to VOCs than NOx formation in Zhengzhou. Vehicle exhaust, coal and biomass

	periods			burning and solvent usage were the major sources for ambient VOCs at all four sites. The strong emissions from coal and biomass burning and solvent usage were concentrated in the southwest of Shanxi and Henan province
Ahamad et al. (2014)	Variation of surface Ozone exceedance around Klang Valley, Malaysia	O_3, NO, and NO_2, nonmethane hydrocarbon (NMHC)	Hierarchical Agglomerative Cluster Analysis (HACA), Clustered trajectories	O_3 exceedance pattern among the stations. nitrogen oxide (NO), and Nitrogen Dioxide (NO_2) concentrations indicated a strong localised influence on the. O_3 exceedance pattern in the Klang Valley area is strong
Awang et al. (2015)	Diurnal variations of ground-level Ozone in three port cities in Malaysia	Diurnal variations of O_3 concentrations, its precursors, and meteorological parameters	Descriptive statistics, correlation analysis, Correspondence analysis (CA), Principal Component Analysis (PCA)	The concentration of Ozone in the three ports was still below the maximum permissible values prescribed by the MAAQG. Klang exhibited the highest average concentration of O_3. The diurnal

				cycle of Ozone. Concentration has a midday peak (1:00 p.m. to 3:00 p.m.). The diurnal pattern of surface Ozone concentration is strongly influenced by meteorological conditions and prevailing levels of precursors (NOx) and CO. The concentrations of NOx, CO, and O_3 were relatively high during January–May.
Azmi, Latif, Ismail, Juneng, and Jemain (2010)	Trend and status of air quality at three different monitoring stations in the Klang Valley, Malaysia	Ozone, PM_{10}, carbon monoxide, nitrogen oxide, Sulphur dioxide	Statistical analysis: Normal P–P Plot, the Normal Q–Q Plot and the One-Sample Kolmogorov–Smirnov test. Analysis of variance with the additional use of the Bonferroni, The Pearson correlation Trajectory analysis: The backward trajectories	The averaged concentration of all atmospheric pollutants recorded at Petaling Jaya, Shah Alam and Gombak are under the permissible value. The pollution levels are distinctively higher when compared with the background station (Jerantut) used in the study. While the levels of PM_{10} and O3 correlate more closely, the averaged concentration of

				other parameters is influenced more by the number of motor vehicles near the monitoring stations. The high concentration of Ozone recorded at Gombak and Shah Alam is intriguing and warrants further investigation.
Pereira et al. (2018)	A Functional Regression Model of the Retinal Nerve Fiber Layer Thickness in Healthy Subjects	intersubject variability of the circumpapillary retinal nerve fiber layer (RNFL) thickness in healthy subjects.	functional regression approach	The new functional regression approach improves on the multivariate linear regression model and allows an even larger reduction of the amount of intersubject variability, while at the same time using a substantially smaller number of parameters to be estimated.
Masselot et al. (2018)	A new look at weather-related health impacts through functional regression	daily mortality counts and the previous day's temperature, annual mortality – annual temperatures (with FFLM).	Functional data analysis, Smoothing functional data, Functional linear model for scalar response, Fully functional linear model	The predictive performance of functional regression is not vastly superior to those of GAMs and DLNMs. Indeed, the differences in RMSE are small compared to the

				magnitude of daily cardiovascular mortality in Montreal which presents a mean of 17 deaths and a standard deviation of 5 deaths. However, functional regression models present several conceptual and practical advantages over the classical models, as discussed
Zhang, Xie, Liang, and Xiong (2016)	Functional Regression Models for Epistasis Analysis of Multiple Quantitative Traits	epistasis in multiple phenotypes	multiple functional regression (MFRG)	The new functional regression approach improves on the multivariate linear regression model and allows an even larger reduction of the amount of intersubject variability, while at the same time using a substantially smaller number of parameters to be estimated.
Fuente, Febrero-Bande, Muñoz, and Domı´nguez	Predicting seasonal influenza transmission using functional regression models with temporal	Temperature, Influenza rate	Functional generalized least squares regression, Simulation, Variable selection using distance	Influenza may increase due to a cold wave with daily temperature around 7oC for two weeks which is

(2018)	dependence		correlation measure, Prediction using temporal dependence structure	consistent with much of the literature on influenza. Also, the models show that the estimated temporal dependence of the influenza virus is strong and sTable over time.