UNIVERSITI TEKNOLOGI MARA

# FUNCTIONAL TIME SERIES MODELS IN FORECASTING MONTHLY DIURNAL MAXIMUM AIR POLLUTANT INDEX (API) CURVES

WAN NAJIHA BINTI WAN MAT DIN

MSc

July 2019

UNIVERSITI TEKNOLOGI MARA

# FUNCTIONAL TIME SERIES MODELS IN FORECASTING MONTHLY DIURNAL MAXIMUM AIR POLLUTANT INDEX (API) CURVES

**WAN NAJIHA BINTI WAN MAT DIN**

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

**Faculty of Computer and Mathematical Sciences**

**July 2019**

**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 thesis 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 Postgraduate, Universiti Teknologi MARA, regulating the conduct of my study and research.

| | | |
|---|---|---|
| Name of Student | : | Wan Najiha Binti Wan Mat Din |
| Student I.D. No. | : | 2017595651 |
| Programme | : | Master of Applied Statistics – CS702 |
| Faculty | : | Faculty of Computer and Mathematical Sciences |
| Thesis Title | : | Functional Time Series Models in Forecasting Monthly Diurnal Maximum Air Pollutant Index (API) Curves |
| Signature of Student | : | ..... |
| Date | : | July 2019 |

# ABSTRACT

In Malaysia, the incidence of air pollution often recorded. Thus, due to the harmful consequences of air pollutants, monitoring and investigating air quality status has become an important task of concern to the government. Air Pollutant Index (API) has been used to measure the level of air quality status with respect to health risk. The computation of API involved six major air pollutants including $PM_{10}$, $PM_{2.5}$, $O_3$, CO, $SO_2$ and $NO_2$. The main aim of this study is to apply Functional Data Analysis focusing on Functional Time Series modelling on API data. Two models were compared to forecast the API data at Shah Alam and Pasir Gudang which include multi-step ahead and iterative one-step ahead approach. The results show that multi-step ahead has produced the best performance with the lowest FMSE, FRMSE and FMAPE; 91.17, 9.55 and 11.77 respectively for Shah Alam and 46.78, 6.84 and 6.62 respectively for Pasir Gudang. Based on the multi-step ahead forecast model, the estimated maximum API over a continuum of 24 hours period was illustrated and the confidence interval was also obtained. Functional descriptive analysis has shown that the normal day-to-day API for both Shah Alam and Pasir Gudang are at the healthy level. However, as shown by the functional standard deviation, API level fluctuates within the day period with the highest variation between 9 a.m. to 10 a.m. at both locations. This result indicates the contribution of vehicles emission towards API. Overall, it can be concluded that the application of functional time series gives several advantages in this study. The method provides the ability to visualize, describe, evaluate and predict continuous variation of API over a continuum of time.

# ACKNOWLEDGEMENT

Praise be to Allah the Almighty and the Most Merciful for giving me the opportunity to embark on my Master and for completing this long and challenging journey successfully.

I would like to express my deepest appreciation to all those who provided me the possibility to complete this report. A special gratitude to my supervisor, Dr. Norshahida Shaadan for her continuous encouragement, invaluable guidance and help for completing this dissertation report.

Furthermore, I would also like to acknowledge with much appreciation to the crucial role of Department of Environment Malaysia for providing the information and data.

I also acknowledge with a deep sense of reverence, my gratitude towards my parents and member of family, who has always supported me morally as well as economically.

Last but not least, gratitude goes to all of my friends who directly or indirectly helped me to complete this dissertation report.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# LIST OF ABBREVIATIONS

**Abbreviations**

| | |
|---|---|
| **ANN** | Artificial Neural Network |
| **API** | Air Pollutant Index |
| **ARIMA** | Autoregressive Integrated Moving Average |
| **CO** | Carbon Monoxide |
| **COHb** | Carboxyhaemoglobin |
| **DOE** | Department of Environment |
| **FPCA** | Functional Principal Component Analysis |
| **FPCR** | Functional Principal Component Regression |
| **GLO** | Ground-level Ozone |
| **MAQI** | Malaysian Air Quality Index |
| **$NO_2$** | Nitrogen Dioxide |
| **$O_3$** | Ozone |
| **PAHs** | Polycyclic Aromatic Hydrocarbons |
| **PCA** | Principal Component Analysis |
| **$PM_{10}$** | Particulate matter < 10μm |
| **$PM_{2.5}$** | Particulate matter < 2.5μm |
| **PSI** | Pollutant Standard Index |
| **$SO_2$** | Sulphur Dioxide |
| **VOCs** | Volatile Organic Compounds |

# CHAPTER ONE
# INTRODUCTION

## 1.1 Research Background

Every day the fresh air of the atmosphere is getting polluted due to the mixing of particulates, biological molecules and other harmful substances. Lots of dirty wastes produces by people on daily basis particularly in the big cities polluting the atmospheric to a great extent. Air pollution is a serious issue that needs to be given immediate and serious attention by all relevant authorities around the globe, as it is one of the most important factors that contributes to the quality of life and living (Azid, Juahir, Toriman, Kamarudin, Saudi, Hasnam, Aziz, Azaman, Latif, Zainuddin, Osman, & Yamin, 2014).

In Malaysia, due to the harmful consequences of air pollutants, monitoring and investigating air quality status has become an important task of concern to the government. Air Pollutant Index (API) has been used to measure air quality level and the level of potential health risk because of over exposed to air pollution. Department of Environment, 2018 indicated that the computation of API involved six major air pollutants including particulate matter<10μm ($PM_{10}$), particulate matter<2.5μm ($PM_{2.5}$), ozone ($O_3$), carbon monoxide (CO), sulphur dioxide ($SO_2$) and nitrogen dioxide ($NO_2$).

API is verified to provide the public with easily understandable information about air pollution (Department of Environment, 2000). Predicting future level of API is crucial for the society as they will benefit through the information the information gathered and help them keep aware about air pollution. The most important benefit is that the findings of this study would help the government to provide an alternative solution on how to prevent the severity of API by means of management, mitigation and adaption of worst air quality. Thus, suitable technique to gain more informative prediction results in need.

Several studies have been conducted to gain information on the behaviour including the pattern recognition of Malaysia air quality, possible sources of air pollutants and the spatial patterns (Mutalib, Juahir, Azid, Mohd Sharif, Latif, Aris, Zain, & Dominick, 2013), application of principal component analysis (PCA) and artificial neural network (ANN) to forecast the air pollutant index (Azid et al., 2014), seasonal ARIMA for forecasting API (Lee, Rahman, Suhartono, Latif, Nor, & Kamisan, 2012), etc. Most of the study had focuses the study of API on the prediction in the term of point estimation.

Recent advances in computer recording and storing technology have massively increased the presence of functional data, whose graphical depiction can be infinite-dimensional curve, shape or image. Data are known as functional time series when the same functional object is observed over a period (Kokoszka & Reimherr, 2017). To increase information on the API behaviour, this study is conducted by focusing on the pattern of monthly diurnal maximum API behaviour in two monitoring stations; Shah Alam, Selangor represents urban area and Pasir Gudang, Johor represents industrial area. This study intended to observe and seek the pattern of monthly diurnal maximum API based on multi-step-ahead and iterative one-step-ahead forecast using functional times series analysis. The findings of this study could increase the understanding on the behaviour of API pattern. Moreover, none of the studies on API behaviour in the Malaysian environment using functional time series analysis.

## 1.2 Problem Statement

Air pollution is now fully recognised to be an important public health problem worldwide including Malaysia. Of greater concern, modern type of pollution in today's urban environments is undetectable at ground level but manifests in chronic health effects (Ferrante, Fiorce, Conti, & Ledda, 2012). Air Pollutant Index (API) is used to measure air quality status in Malaysia. Department of Environment (2018) stated that the higher the value of API higher is the air pollution risk to human health. Thus, predicting future air quality is very important to help government to monitor air quality.

Based on the literatures, the researcher had commonly used classical time series technique to predict API values. In the modelling process, the data used to predict API were discrete observations recorded by hourly, daily basis and so on. As a result, the predicted value obtained from the model is in the form of point value which contains less information and static in nature. However, the information is seen insufficient and limited for observing air quality status or the level of risk throughout a day process continuously, at every hour or any time of the day. Thus, an alternative approach to obtain API prediction on the whole fluctuation of API level over a particular period of time, for example within a day period is required.

Functional data is defined as continuous observation that exists within a continuum of time or space (Ramsay & Silverman, 2006). The data can be represented by functions. Physically the data is visualized as curves. Functional time series (FTS) is a set of statistical techniques to analyse curve data. Thus, by applying FTS, daily API can be predicted using FTS methodology. Moreover, due to the effectiveness and advantages of the application of functional data approach in predicting continuous mortality data that have reported in several research, thus in this study, the functional data technique will be adopted for modelling dynamic time series of Malaysia's API.

## 1.3 Research Questions

The research questions for this study are:

i) Which functional time series method is the best in predicting monthly diurnal maximum API curves between the multi-step-ahead forecast and iterative one-step-ahead forecast?

ii) What is the future pattern of monthly diurnal maximum API curves at particular location in Malaysia?

iii) What is the significant range of forecasted monthly diurnal maximum API curves?

## 1.4 Research Objectives

The research objectives for this study are:

i) To compare the performance between multi-step-ahead and iterative one-step-ahead functional time series methods in predicting of monthly diurnal maximum API curves.

ii) To determine the future pattern of monthly diurnal maximum API curves at two hotspot locations in Malaysia using the best model as result in Objective 1.

iii) To determine the confidence interval for the forecasted monthly diurnal maximum API curves.

## 1.5 Significance of Study

Generally, API information is very important because it is the measure of the risk impact level due to air pollution exposure. By conducting research on API fluctuation and prediction, the information could provide on the possible severity of the air pollution towards human's health with respect to the temporal (time) and spatial (place) occurrence.

In addition, the findings of this study will redound to the benefit of society as a basis for preparing an unexpected serious event of air pollution. Consequently, society will benefit from the information gathered and help them to be aware of air pollution in their day-to-day activities.

Ultimately, the government can apply the recommended approach derived from the results of this study as an input for air pollution mitigation as well as increase understanding to manage air pollution. Furthermore, as for private sector, this study might provide a predictive analytic tool in enhancing the current practice to predict a continuous fluctuation of air quality level throughout one day period. So that, the model can be used in computer application system.

## 1.6 Limitation of Study

In order to achieve the objectives of this study, there is an unavoidable limitation during the research process. Due to the time limit and cost, this study only focused on two locations in Malaysia that has been identified as hotspot locations which are Shah Alam and Pasir Gudang. Hence the findings only restricted for the particular monitoring locations. Plus, the study conducted only able to compare the prediction of monthly diurnal maximum API by comparing two (2) models; which are multi-step-ahead and iterative one-step-ahead models. Moreover, since this study used new analysis approach that is not yet used by many, the R codes programming to run the analyses quite hard to be found and limited.

# CHAPTER TWO
# LITERATURE REVIEW

## 2.1 Introduction

Air pollution is a serious issue that needs to be given immediate and serious attention by all relevant authorities around the globe, as it is one of the most important factors that contribute to the quality of life and living (Azid et al., 2014). Due to relentless development in this world, Malaysia particularly, possibility of being affected by pollution is high. Pollution, an unwanted destruction of natural environment by person and naturally prompted insults, is an issue confronting the current world (Mabahwi, Leh, & Omar, 2015).

Air is always impure and contaminated with gases such as CO, $NO_2$, $SO_2$ and others (which are toxic in nature) and finely separated solid and liquid particles and smog (Mabahwi et al., 2015). Air ends up polluted due to the presence of these contaminants. According to Mabahwi et al. (2015), the existence of these contaminants in the air is known as air pollution and the substances which pollute the air are known as air pollutants. "There are many pollutants of suspended materials such as dust, fumes, smokes, mists, gaseous pollutants, hydrocarbons, volatile organic compound (VOCs), polycyclic aromatic hydrocarbons (PAHs) and halogen derivatives in the air which at the high concentrations cause vulnerability to many diseases including different types of cancers" (Ghorani, Riahi, & Balali, 2016). Each pollutant has a different impact on human health. The major important air pollutants and their poisonous effects on different human body organs and related illnesses have been briefly explained below.

### 2.1.1 Ozone and its Negative Impacts

Ozone ($O_3$) with the chemical formula $O_3$ is an unseen gas that is the most vital component of the atmosphere. According to Banan, Latif, Junen and Ahamad (2013), the $O_3$ was identified as the dominant pollution in some suburban and rural areas due to the downwind effect that transport the $O_3$ pollutants. It is found both at the ground level

and in the upper regions of the atmosphere which is called troposphere and stratosphere respectively as shown in Figure 2.1. Stratosphere also known as good ozone since it shields life on Earth from the sun's harmful ultraviolet (UV) rays. Ground-level Ozone (GLO) is produced as a result of chemical reaction between oxides of nitrogen and VOCs released from natural sources and/or due to human activities. VOCs and NOx are released from many different sources such as motor vehicles, chemical solvents and other industrial causes (Sharma, Jain, Khirwadkar, & Kulkarni, 2013). Gorai, Tuluri and Tchounwou (2014) stated that GLO is believed to have a plausible association with increased risk of respiratory diseases, particularly asthma.

Figure 2.1 Type of Ozone

Bad ozone can cause various infections to human health such as shortness of breath, coughing, throat irritation and lung disease. Ozone also triggers asthma and may aggravate other respiratory illnesses such as pneumonia, emphysema and bronchitis (Mabahwi, Leh, & Omar, 2014). Asthma attacks may let chest pain when breathing and also can grew risk of respiratory disorders.

At concentrations that occur in many urban areas, $O_3$ causes a variety of toxic effects in humans and experimental animals. Based on previous studies to evaluate the consequences of ozone on the respiratory tract, ozone exposure has been reported to induce airway inflammation, increase airway responsiveness and reduce pulmonary function in persons and experimental animal models (Fiévez, Kirschvink, Dogné,

Jaspar, Merville, Bours, Lekeux, & Bureau, 2001). From an ecological point of view, $O_3$ can reduce carbon absorption in trees that can lead to deforestation which may influence global food security in the long-term exposure (Fares, Vargas, Detto, Goldstein, Karklik, Paoletti, & Vitale, 2013; Wilkinson, Mills, Illidge, & Davies, 2012).

### 2.1.2 Carbon Monoxide and its Negative Impacts

CO is a toxic gas that cannot be seen or smelled which can cause sudden illness and death. CO usually comes from sources that are not properly maintained or vented in or near home. CO can be found in combustion fumes, such as those produced by trucks and cars, lanterns, stoves, small gasoline engines, burning charcoal and woods and poorly functioning heating systems (Goldstein, 2008). CO poisoning can attack anyone. The elderly, infants, unborn babies and persons with chronic heart disease, respiratory problems or anaemia are commonly more at stake than others.

Carboxyhaemoglobin (COHb) percentage is the most commonly used diagnosis of carbon monoxide exposure. "No human health effects have been showed for carboxyhaemoglobin (COHb) levels lower than 2%, while levels above 40% may be fatal. Hypoxia, apoptosis and ischemia are known to be underlying diseases for CO toxicity" (Akyol, Erdogan, Idiz, Celik, Kaya, Ucar, Dane, & Akyol, 2014). The symptoms of CO poisoning are dizziness, headache, nausea, and vomiting. CO causes loss of consciousness, coma and eventually dead at very high levels. Long-term exposure to moderate and high levels of CO has also been related with an escalated risk of heart disease (Chuang, Yan, Chiu, & Cheng, 2011). Persons who can endure serious level of CO poisoning might experience long-term health difficulties.

### 2.1.3 Sulphur Dioxide and its Negative Impacts

$SO_2$ is a colourless, highly reactive gas, which is considered a s an important air pollutant. It is mostly emitted from fossil fuel consumption, natural volcanic activities and industrial processes. $SO_2$ is extremely harmful for animal, plant life and human health. Person with lung disease, the elderly, kids and those who are most subjected to $SO_2$ are at greater risk of the lung and skin diseases.

Respiratory ache and dysfunction, and also aggravation of existing cardiovascular disease are the main health troubles related with exposure to high concentrations of $SO_2$. $SO_2$ is mostly absorbed in the upper airways which can cause bronchospasm and mucus secretion in humans. Mahajan (2009) stated that residents of industrialised regions encountered with $SO_2$ even at lower concentrations (<1ppm) in the polluted ambient air might experience a high level of bronchitis.

$SO_2$ penetration into the lungs is higher during mouth breathing compared to nose breathing. Rapid breathing increases the gas's concentration into the deeper lung. Those who exercise in the polluted air would therefore inhale more $SO_2$ and are likely to suffer from greater irritation (Johns & Linn, 2011).

According to Environmental Protection Agency (EPA) of the USA, the annual $SO_2$ standard level is 0.03 ppm. $SO_2$ is responsible for acid rain creation and acidification of soils due to its solubility in water. $SO_2$ cuts the amount of oxygen in the water initiating the death of both animals and plants of marine species. According to Ghorani et al. (2016), $SO_2$ exposure may cause eye damage (lacrimation and corneal opacity), respiratory tracts, skin damage (redness, blisters) and mucous membranes. The most common clinical findings related to $SO_2$ exposure are pulmonary edema, bronchospasms, acute airway blockage and pneumonitis (Chen, Kuschner, Gokhale, & Shofer, 2007).

### 2.1.4 Nitrogen Dioxide and its Negative Impacts

"Nitrogen dioxides are important ambient air pollutants that can escalate the risk of respiratory infections" (Chen et al., 2007). The main source of $NO_2$ is air pollutants related to traffic, such as smog emitted from motor engines. They are deep lung irritants that can cause pulmonary edema if been inhaled at high concentrations. They are usually less toxic than $O_3$, but $NO_2$ can present clear toxicological problems.

The usual complication of $NO_x$ toxicity is wheezing and coughing, but there may also be irritations in the eyes, nose or throat, chest pain, dyspnea, fever, headache, pulmonary edema and bronchospasm. In another report by Hesterberg (2009), it is

suggested that the level of nitrogen oxide between 0.2 and 0.6 ppm is harmless for human population.

### 2.1.5 Particle Pollutants and its Negative Impacts

Particle pollutants are major parts of air pollutants. In a simple definition, they are a mixture of particles found in the air. Particle pollutant which is more known as PM is linked with most of pulmonary and cardiac-associated morbidity and mortality (Ghorani et al., 2016; Sahu, Kannan, & Vijayaraghavan, 2014). They have varied in size ranging mostly from 2.5 to 10 μm ($PM_{2.5}$ to $PM_{10}$).

The size of particle pollutants is directly associated with the onset and progression of the lungs and heart diseases. Particles of smaller size reach the lower respiratory tract and thus have greater potential for causing the lungs and heart diseases. In addition, numerous scientific data have shown that fine particle pollutants cause premature death in people with heart and/or lung disease including non-fatal heart attacks, deceased lung functions, aggravated asthma and cardiac dysrhythmias. Particulate pollutants can cause mild and severe disease depends on the exposure level. Cough, wheezing, dry mouth and activities limitation due to breathing problems are the most common clinical signs of respiratory diseases caused by air pollution (Guillam, Pédrono, Bouquin, Huneau, Gaudon, Leborgne, Dewitte, & Ségala, 2013). Residents of industrialised regions with $SO_2$ in polluted atmosphere at even lower concentrations (<1ppm) may experience a high level of bronchitis.

Exposure to current environmental PM concentrations for a long-term could lead to a significant reduction in life expectancy. The main reasons for the reduction in life expectancy are the increase in cardiopulmonary and lung cancer mortality. Reduced lung functions in adults and children that lead to chronic obstructive pulmonary disease (COPD) and asthmatic bronchitis are also serious diseases which prompt lower quality of life spans. Strong evidence of the effect of long-term exposure to PM on cardiopulmonary mortality comes from study conducted by Zhou, Ito, Lall, Lippmann and Thurston (2011).

Table 2.1
Summary of Pollutants in API, Sources and Effects

| Pollutant | Sources | Effects |
|---|---|---|
| Carbon Monoxide ($CO_2$) | Vehicle and engine fuel combustion | Reduces the quantity of oxygen that reaches the organs and tissues of the body; exacerbates heart disease, leading to chest pain and other symptoms. |
| Ground-level Ozone ($O_3$) | Secondary pollutant formed in the presence of sunlight through the chemical reaction of volatile organic compounds (VOCs) and $NO_x$ | Decreases lung function and triggers respiratory symptoms such as coughing and shortness of breath, as well as exacerbates asthma and other lung diseases. |
| Nitrogen Dioxide ($NO_2$) | Combustion of fuel (electric utilities, large industrial boilers, cars) and burning of wood. | Worsens pulmonary illness that lead to breathing symptoms, accelerated susceptibility to breathing infection. |
| Sulphur Dioxide ($SO_2$) | $SO_2$ is derived from combustion of fuel (particularly high-sulphur coal); electrical utilities and industrial processes as well as natural events such as volcanoes. | It exacerbates asthma and makes it hard to breath. It also makes up to the formation of particles with potential health effects. |
| Particulate Matter (PM) | This is created by chemical reactions, combustion of fuel (e.g. coal burning, timber, diesel), industrial processes, agriculture (plowing, field burning), and unpaved highways or road buildings. | Short-term exposures can exacerbate heart or lung disease and trigger breathing issues. Long-term exposures can lead to heart or lung disease and premature fatalities sometimes. |

## 2.2 Air Pollutant Index (API)

By the year 2020, Malaysia will have advanced to become an industrialize nation. According to that determination, air quality in Malaysia is therefore a major concern. Abdullah, Samah and Jun (2012) said factories, power plants, dry cleaner, vehicles, windblown dust and wildfires are examples of various sources contributing to air pollution. Air pollution monitoring duties are disseminated at different levels consisting of international protocols and agreements, and community legislation at the national and regional levels (Rani, Azid, Khalit, Juahir, & Samsudin, 2018).

Air quality monitoring has become part of the initial pollution prevention strategy in Malaysia. Air quality guidelines for air pollutants were formulated by the Malaysian Department of Environment (DOE) in 1989. The recommended Malaysia Air Quality Guidelines (RMG) defined concentration limits of selected air pollutants that might adversely affect the general public's health and welfare. In 1993, the DOE established its first quality index system, known as the Malaysian Air Quality Index (MAQI), and played an important role in informing both decision makers and the general public about ambient air quality status ranging from good to emergency. By applying this index particularly in industrialized countries, the management of air quality and public health protection become effective. In 1996, the DOE Malaysia revised its index system for easy evaluation with countries as well as for regional harmonization where the Air Pollutant Index (API) was adopted, which closely follows the United States system known as pollutant standard index (PSI) (Department of Environment,2000). PSI is one of the first synthetic indices agreed by the United States Environmental Protection Agency (USEPA) as developed by Ott and Hunt (1976). However, in 1999, the EPA changed replaced PSI with air quality index (AQI) while Malaysia stuck with API.

The status indicator of API was divided into a few categories. For instances, good, moderate, unhealthy, very unhealthy, hazardous and emergency as mentioned in Table 2.2, which can be of air quality management level or decision making for data interpretation processes.

Table 2.2
API in Malaysia

| API | Status | Level of Pollution | Health Measure |
|---|---|---|---|
| 0 – 50 | Good | Low pollution without any bad effect on health | No activity restrictions for all groups of individuals. |
| 51 – 100 | Moderate | Moderate pollution that does not pose any bad effect on health | No activity restrictions for all groups of individuals. |
| 101 – 200 | Unhealthy | Worsen the health condition of high-risk people who is the people heart with heart and lung complications | Limiting outdoor activities for individuals at high risk.<br>Outdoor activities should be reduced by the public. |
| 201 – 300 | Very unhealthy | Worsen the health condition and low tolerance of physical exercises to people with heart and lung complications.<br>Affect public health. | Individuals suffering from heart or lung disease should remain indoors and prevent physical activities. |
| 301 – 500 | Hazardous | Hazardous to high-risk people and public health. | Individuals suffering from heart or lung disease should remain indoors and prevent physical activities.<br>Outdoor activities should be avoided by the public. |
| Above 500 | Emergency | Severe aggravation and health hazards | Public recommended to follow the National Security Council's order and always follow the announcement via mass media |

Source: Department of Environment, 2018; Rahman, 2016

The index system known as API is a comprehensive approach for defining air quality status that can be understood easily by the general public. It is categorized based on the highest values from five main air pollutant index values: $O_3$, $PM_{10}$, $NO_2$, $SO_2$ and CO for a particular time period, and where $SO_2$ and $PM_{10}$ hourly value averaged over a 24-hour running period, CO is averaged over an eight-hour period, and $NO_2$ and $O_3$ are read hourly before an hourly index is computed with the use of sub-index functions for each pollutants according to the standpoint of human health implications. All the sub-indices of pollutants can be calculated as shown in Table 2.3. Individual indices were calculated based on individual pollutants. The maximum index among the pollutants was selected. That index is then considered as API.

Table 2.3
API Equation for Each Pollutant

| Pollutant | | API Calculation Equation |
|---|---|---|
| CO | conc. < 9 ppm | API = conc. x 11.111111 |
| (Based on eight-hour average concentrations) | 9 < conc. < 15 ppm | API = 100 + {[conc. – 9] x 16.666667} |
| | 15 < conc. < 30 ppm | API = 200 + {[conc. – 15] x 6.666667} |
| | conc. > 30 ppm | API = 300 + {[conc. – 30] x 10} |
| $O_3$ | conc. < 0.2 ppm | API = conc. x 1000 |
| (Based on one-hour average concentrations) | 0.2 < conc. < 0.4 ppm | API = 200 + {[conc. – 0.2] x 500} |
| | conc. > 0.4 ppm | API = 300 + {[conc. – 0.4] x 1000} |
| $NO_2$ | conc. < 0.17 ppm | API = conc. x 588.23529 |
| (Based on one-hour average concentrations) | 0.17 < conc. < 0.6 ppm | API = 100 + {[conc. – 0.17] x 232.56} |
| | 0.6 < conc. < 1.2 ppm | API = 200 + {[conc. – 0.6] x 16 6.667} |
| | conc. > 1.2 ppm | API = 300 + {[conc. – 1.2] x 250} |
| $SO_2$ | conc. < 0.17 ppm | API = conc. x 588.23529 |
| (Based on 24-hour average concentrations) | 0.17 < conc. < 0.6 ppm | API = 100 + {[conc. – 0.17] x 232.56} |
| | 0.6 < conc. < 1.2 ppm | API = 200 + {[conc. – 0.6] x 16 6.667} |
| | conc. > 1.2 ppm | API = 300 + {[conc. – 1.2] x 250} |
| $PM_{10}$ | conc. < 50 pg/$m^3$ | API = conc. |
| (Based on 24-hour average concentrations) | 50 < conc. < 150 | API = 50 + {[conc. – 50] x 0.5} |
| | 150 < conc. < 350 | API = 100 + {[conc. – 150] x 0.5} |
| | 350 < conc. < 420 | API = 200 + {[conc. – 350] x 14286} |
| | 420 < conc. < 500 | API = 300 + {[conc. – 420] x 1.25} |
| | conc. > 500 pg/$m^3$ | API = 400 + [conc. – 500] |

Source: Department of Environment, 2000

## 2.3 Study of Air Pollution in Malaysia

Many studies have been carried out by Malaysian researchers on air pollution. The studies conducted were essentially aimed at determining the air pollution level in the Malaysia environment.

### 2.3.1 Issues and Methods

Rani et al. (2018) implemented XLSTAT add-in 2014 to determine the Malaysia's 2010 – 2015 Air Pollutant Index (API). The study analyses indicate that air monitoring station at Sekolah Menengah Teknik Muar in Johor illustrated the highest API reading value with 663 on June 23, 2013 (emergency level), where Malaysia experienced its worst air quality on that day linked to haze episodes.

Isiyaka and Azid (2015) has studied air quality pattern assessment in Malaysia using multivariate techniques such as Hierarchical Agglomerative Cluster Analysis (HACA), Discriminant Analysis (DA), Principal Component Analysis (PCA) and Artificial Neural Network (ANN).

Another study was conducted by Azid et al. (2014) have utilised eight air quality parameters in ten monitoring stations in Malaysia for years 2005 - 2011 to study on the pattern recognition of Malaysia's air quality based on the data attained from DOE. The study applied two methods; principal component analysis (PCA) and artificial neural networks (ANN) to determine the sources predictive ability for the API. The study concluded that CH4, NmHC, THC, O3, and PM10 are the most significant parameters and the The PCA-ANN showed better predictive ability in the determination of API with fewer variables.

Abdullah et al. (2012) reviewed the API for the period from 2001 to 2009 in Klang Valley, Malaysia. The general air quality in the Klang Valley was moderate for 66% of the days throughout 2009, whereas at the unhealthy level only 5% of the days were classified. The major contributors to the deterioration of air quality in the Klang Valley were urbanisation, vehicles, industry and forest fires (Abdullah et al., 2012).

Therefore, the study suggested that the government should introduce a more sustainable strategy to address the problem of air pollution.

A study by Dominick, Juahir, Latif, Zain and Aris (2012) based on a two-year database (2008 – 2009) shows possible sources of air pollutants and the spatial patterns in the eight Malaysian air monitoring stations selected. The multivariate analysis was used on the dataset. It featured Multiple Linear Regression (MLR) to assess the contribution percentage of each air pollutant, PCA to determine the main sources of the air pollutions and Hierarchical Agglomerative Cluster Analysis (HACA) to access the spatial patterns. Based on the characteristics of the air pollutants and meteorological parameters, the HACA outcomes grouped the eight monitoring stations into three distinct clusters. The PCA analysis showed that the main sources of air pollution were releases from industries, aircraft, motor vehicles and areas of high population density. The MLR analysis revealed that the main pollutant contributing to inconsistency in the API at all stations was $PM_{10}$.

Table 2.4
Summary of Air Pollution Study in Malaysia

| Citations | Research Title | Objective | Method Used |
|---|---|---|---|
| Rani et al. (2018) | Air Pollution Index Trend Analysis in Malaysia, 2010 – 2015. | To determine the API trend in Malaysia from 2010 to 2015. | XLSTAT add-in 2014 |
| Isiyaka & Azid (2015) | Air Quality Pattern Assessment in Malaysia using Multivariate Techniques. | To investigate the spatial characteristics in the pattern of air quality monitoring sites | Hierarchical Agglomerative Cluster Analysis (HACA) |
| | | To identify the most discriminating parameters contributing to air pollution | Discriminant Analysis (DA) |
| | | To predict the level of Air | Principal Component Analysis (PCA) |

| | | | |
|---|---|---|---|
| | | Pollution Index. | Artificial Neural Network (ANN) |
| Azid et al. (2014) | Prediction of the level of air pollution using principal component analysis and artificial neural network techniques: A case study in Malaysia. | To identify the sources of pollution in the study locations.<br>To determine the predictive ability for Air Pollutant Index (API). | Principal Component Analysis (PCA)<br>Artificial Neural Network (ANN) |
| Abdullah et al. (2012) | An Overview of the Air Pollution Trend in Klang Valley, Malaysia. | To review the air pollution trend as well as the factors that contribute to the air quality | Descriptive Analysis |
| Dominick et al. (2012) | Spatial assessment of air quality patterns in Malaysia using multivariate analysis. | To investigate possible sources of air pollutants and the spatial patterns within the eight selected Malaysian air monitoring stations based on a two-year database (2008–2009). | Multiple Linear Regression (MLR)<br>Hierarchical Agglomerative Cluster Analysis (HACA)<br>Principal Component Analysis (PCA) |

### 2.3.2 Time Series Analysis of Air Pollution and its Limitation

The emergence of air pollution in urban areas has received a great deal of attention globally in recent years. The air pollutants have arisen harmful effects on living conditions and health. Time series forecasting is the important method nowadays with the ability to forecast the future events. In the study carried out by Rahman, Lee

and Latif (2013), the forecast is based on 10 years of API monthly data in residential and industrial monitoring station areas in Malaysia. ANN, fuzzy time series and the Autoregressive Integrated Moving Average (ARIMA) were used as the methods to predict the API values.

Lee et al. (2012) also has carried out a study to examine the monthly and seasonal fluctuations of API at all monitoring stations in Johor. In this study, time series models are used to evaluate future air quality, modelling and forecasting monthly future air quality in Malaysia. A Box-Jenkins ARIMA method was used to analyse the API values in Johor. High API values recorded at Sekolah Menengah Pasir Gudang Dua among all these three stations. This situation shows that Pasir Gudang is the most polluted area in Johor.

Another study has been conducted by Siew, Chin and Wee (2008) with the objective to fit and illustrate the use of time series models in API forecasting in Shah Alam, Selangor. The data employed in this study comprises of 70 monthly API observations (from March 1998 to December 2003) published in the Department of Environment's annual report. The time series models considered were the ARIMA and the Integrated Long Memory Model (ARFIMA) models. The smallest values of MAE, MAPE and RMSE were used as the criteria for model selection. The embedded ARFIMA model claims to be the better model compared to ARIMA since it has the smallest MAPE value.

Based on above literatures, the researcher had commonly used classical time series technique to predict API values. In the modelling process, the data used to predict were values measured discretely by hourly, daily, monthly basis, etc. However, the predicted value obtained from the model is a point value which contains less information. This is because, this point value is seen as a static statistic. Thus, an alternative approach to obtain a dynamic predictive result is important and is required to increase understanding as well as to predict the whole fluctuation of API level over a particular period of time, for example within a day period.

Table 2.5
Summary of Time Series Analysis in Malaysia: Current Methods

| Citations | Research Title | Objective | Time Series Methods | Results |
|---|---|---|---|---|
| Rahman et al. (2013) | Forecasting of Air Pollution Index with Artificial Neural Network. | To construct and develop the accurate statistical forecasting models to predict the monthly API and to evaluate such models in order to monitor the air quality status. | The autoregressive integrated moving average (ARIMA), fuzzy time series (FTS) and artificial neural network (ANNs) were used as the methods to forecast the API values.<br>The performance of each method is compare using the root mean square error (RMSE).<br>**NOTE:** The study used point-wise approach. | The result shows that the ANNs give the smallest forecasting error to forecast API compared to FTS and ARIMA. |
| Lee et al. (2012) | Seasonal ARIMA for forecasting air pollution index: A case study. | To determine the monthly and seasonal variations of Air Pollution Index (API) at all monitoring stations in Johor. | A Box-Jenkins ARIMA approach was applied in order to analyse the API values in Johor.<br>**NOTE:** The study used point-wise approach. | The most polluted area in Johor located in Pasir Gudang. |

| | | | | |
|---|---|---|---|---|
| Siew et al. (2008) | ARIMA and integrated ARFIMA models for forecasting air pollution index in Shah Alam, Selangor. | To fit and illustrate the use of time series models in forecasting the API in Shah Alam, Selangor. | The time series models considered were the ARIMA and the Integrated Long Memory Model (ARFIMA) models.<br>The smallest values of MAE, MAPE and RMSE were used as the criteria for model selection.<br>**NOTE:** The study used point-wise approach. | The embedded ARFIMA model claims to be the better model compared to ARIMA since it has the smallest MAPE value. |

## 2.4 Functional Data Analysis

Functional data analysis (FDA) is increasingly being used to better data analysis, modelling and time series forecasting. Key features of FDA involve selection of smoothing techniques, data reduction, clustering adjustment, functional linear modelling and forecasting methods.

A study conducted by Ullah and Finch (2013) shows that a total of 84 articles on FDA application have been identified; 75.0% of the articles submitted have been published since 2005. The use of FDA appears in a large number of publications in different science fields; the majority is related to applications and biomedicine (21.4%). In total, 72 studies (85.7%) reported the type of smoothing techniques used, with B-spline smoothing (29.8%) being the most popular. Functional principal component analysis (FPCA) for the extraction of information from functional data was reported in 51 (60.7%) studies. One quarter (25.0%) of the published studies used functional linear models to describe relationships between explanatory and outcome variables and only 8.3% used FDA for time series data forecasting.

Despite its obvious advantages for analysing time series data, full understanding of the key components and value of FDA have been limited to date, though the applications show its significance to many public health and biomedical problems. Broader application of the FDA to all studies involving correlated measurements should allow for better modelling and predictions of such data in the future, especially as FDA makes no a priori age and time effects assumptions.

Commonly, time series data are considered as multivariate data because they are given as a finite, discrete time series. This common multivariate method totally disregards crucial information about the smooth functional behaviour of the generating process that underpins the data (Green & Silverman, 1994). It also experiences problems related to highly correlated measurements within each functional object. The general idea behind FDA is to convey discrete observations resulting from time series in the form of a function (to create functional data) which represents the entire measured function as a single observation. And then, to draw modelling and/or forecast

information from a collection of functional data by using statistical concepts from multivariate data analysis. In this way, it has the advantage of generating models that can be portrayed by continuous smooth dynamics, enabling accurate estimation of parameters for use in the analysis stages, effective data noise reduction through curve smoothing and applicability to data with irregular time sampling schedules.

There are several practical reasons for considering functional data as explained by Ramsay and Dalzell (1991):

i) interpolation and smoothing methods can generate functional representations of a finite set of observations

ii) it is more natural to think in a functional way through modelling problems; and

iii) the objectives of an analysis can be functional in nature, as would be the case if finite data were used to estimate the entire function, its derivatives, or other functional values.

Table 2.6
Summary of Functional Data Analysis Application

| Citations | Fields of Study | Outcome of Interest | FDA Features | | | |
|---|---|---|---|---|---|---|
| | | | Smoothing | Data Reduction | FLM | Forecasting |
| Crane et al. (2010) | Biomechanics | Kinematic gait data | Polynomial Spline | FPCA | - | - |
| Hyndman & Shang (2010) | Demography | Age-specific mortality rates | Kernel | FPCA | | |
| Erbas et al. (2010) | Medicine | Age-specific breast cancer mortality | Penalized regression spline | FPCA | - | State space model |
| Lee, Meyer, & Bradlow (2009) | Meteorology | Clickstream web data (Hurricane Katrina) | B-spline | - | FANOVA | - |
| Hyndman & Booth (2008) | Demography | Mortality, fertility and migration rates | Weighted penalized regression spline | FPCA | - | State space model |
| Laukaitis (2008) | Finance | Cash flow and transactions | Wavelet | FPCA | - | FAR |

Source: Ullah and Finch, 2013

## 2.5 Functional Time Series Application

Functional time series often consists of random functions observed at regular time intervals. Depending on whether or not the continuum is also a time variable, functional time series can be grouped into two categories.

Moreover, functional time series can occur by dividing a nearly continuous time record into natural consecutive intervals such as days, months or years. Examples of these include daily financial stock price curves (Kokoszka Miao, & Zhang, 2012), and monthly sea surface temperature in climatology (Shang & Hyndman, 2011). Figure 2.2 shows monthly sea surface temperatures (in °C) from January 1950 to December 2006. The sea surface temperatures were smoothed using a smoothing spline with the smoothing parameters determined by generalized cross validation. Each curve represents smoothed sea surface temperatures in year.

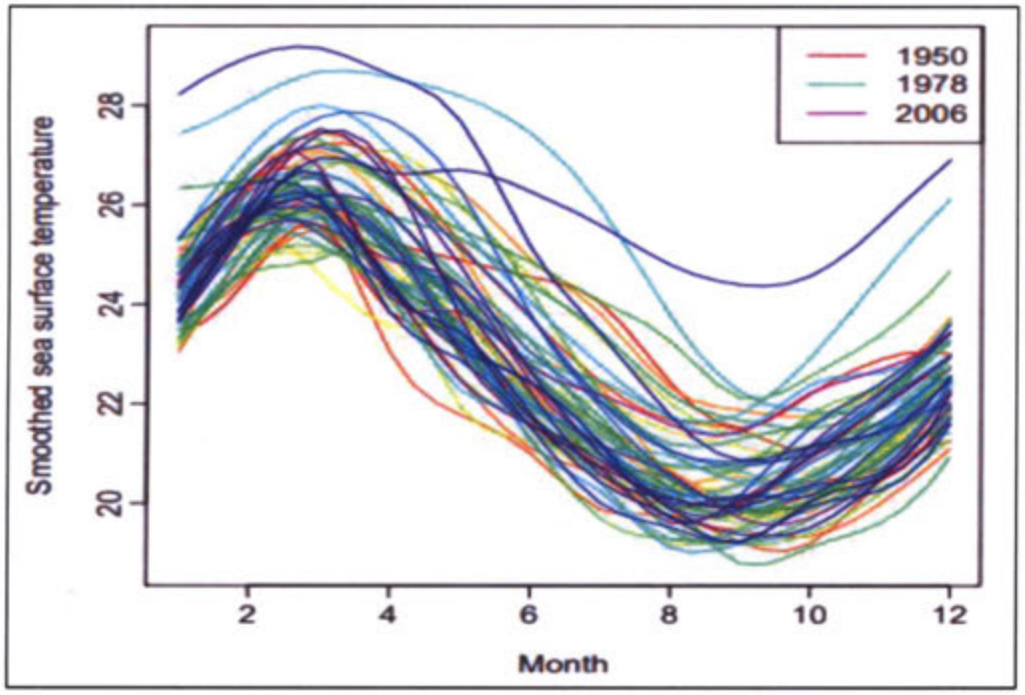


Figure 2.2 Smoothed Monthly Sea Surface Temperatures (in °C) from 1950 to 2006

On the other hand, functional time series can also arise when observations in a time period can be considered together as finite realizations of an underlying continuous function; for example, annual age-specific mortality rates in demography (Gao, Shang, & Yang, 2018). The data set covers yearly age-specific mortality rates over a span of

41 years from 1975 to 2015. The observations are the yearly mortality curves from ages 0 to 110 years, where age is treated as the continuum in the rate function. The curves represented as shown in Figure 2.3.

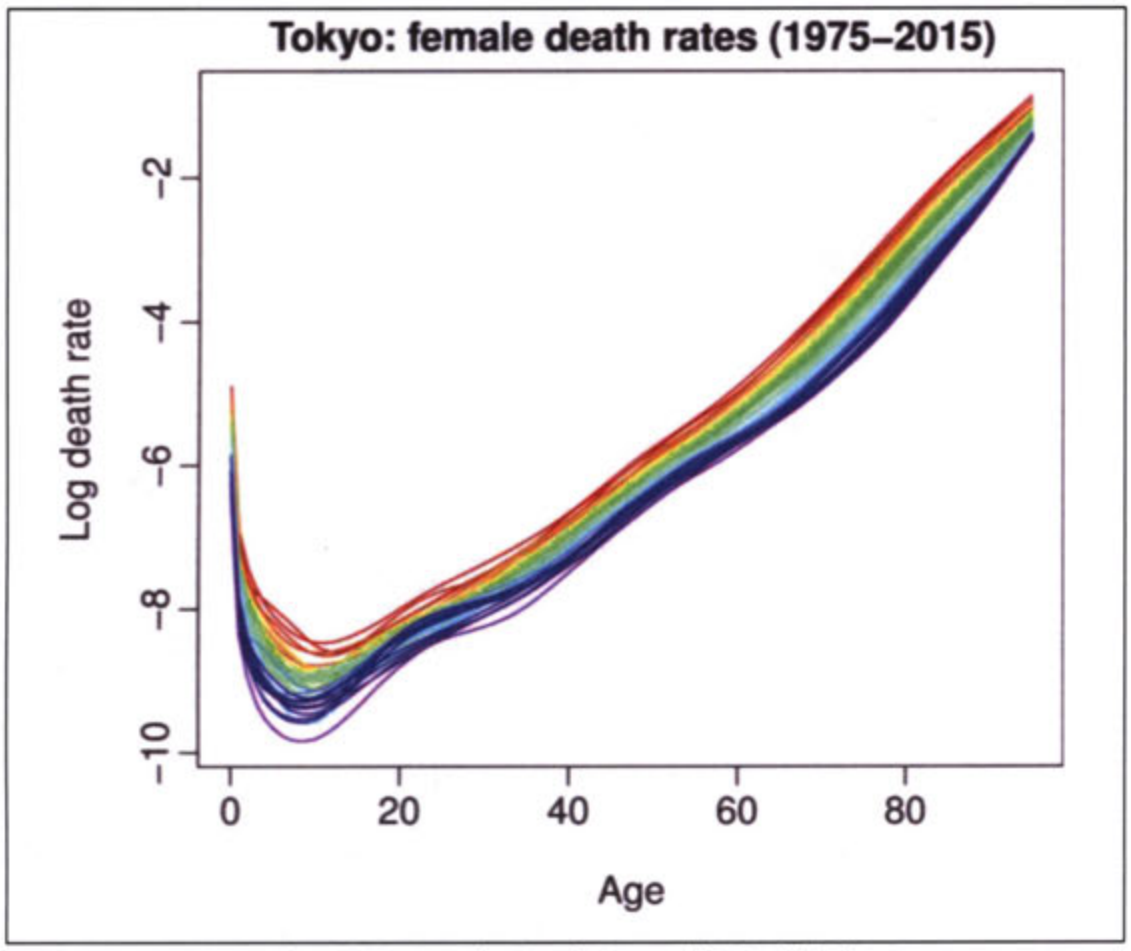


Figure 2.3 Female Death Rates (1975 – 2015)

Hyndman and Shang (2009) also previously have conducted study to forecast Australian fertility rates. A rainbow plot is illustrated in Figure 2.4, where the mortality rates are smoothed by weighted penalized regression without monotonic constraint (see Hyndman and Ullah, 2009).

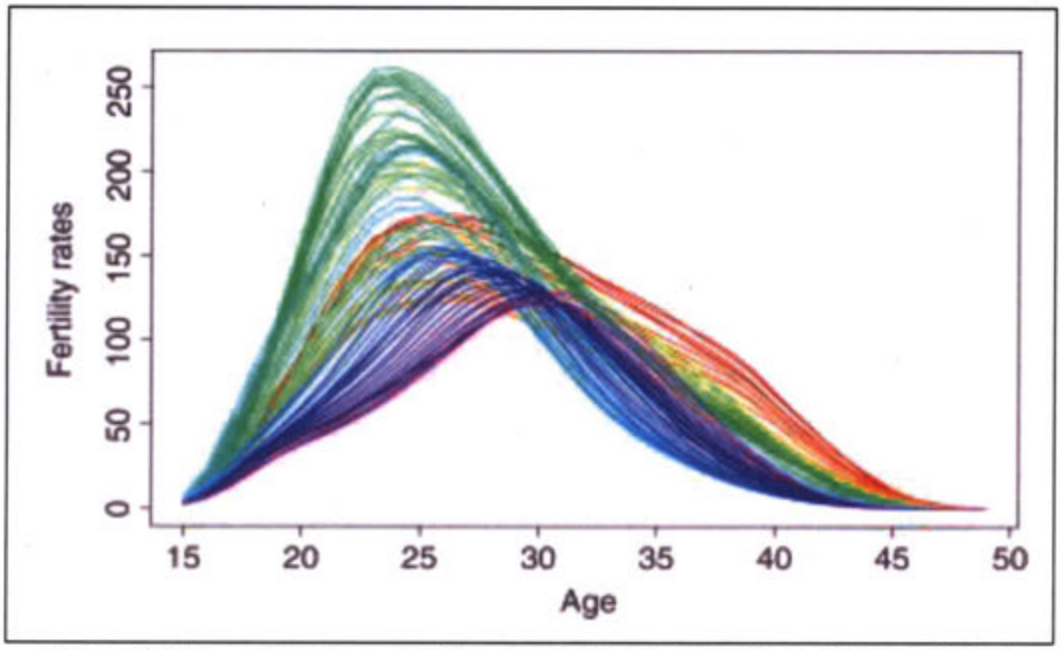

Figure 2.4 Smoothed Age-specific Australian Fertility Rates (1921 – 2006)

Table 2.7
Summary of Functional Time Series Application

| Citations | Research Title | Application on |
|---|---|---|
| Gao et al. (2018) | High-dimensional functional time series forecasting:<br>An application to age-specific mortality rates. | Japanese age-specific mortality rates (1975 – 2015) |
| Kokoszka et al. (2012) | Functional prediction of intraday cumulative returns | Daily financial stock price curves |
| Shang & Hyndman (2011) | Nonparametric time series forecasting with dynamic updating. | Monthly sea surface temperatures (in degree Celcius) from 1950 to 2006 |
| Hyndman & Ullah (2009) | Forecasting functional time series. | Age-specific Australian fertility rates from 1921 to 2006 |

# CHAPTER THREE
# RESEARCH METHODOLOGY

## 3.1 Introduction

This chapter provides a detail explanation on the methods applied in this research study. Secondary data of hourly API observed for 8 years (2011 – 2018) will be used. These data will be obtained from Air Quality Division, Department of Environment Malaysia (DOE). This research study will be conducted following the seven (7) phases of analyses which are data collection, data preparation, data conversion and smoothing, functional time series model development, model comparison, model validation and lastly obtaining confidence interval for the forecasted curves. The analyses will be conducted using Microsoft Excel and R software.

## 3.2 Research Framework

The study will be performed by following the research framework as shown in Figure 3.1 to accomplish the objectives of this study.

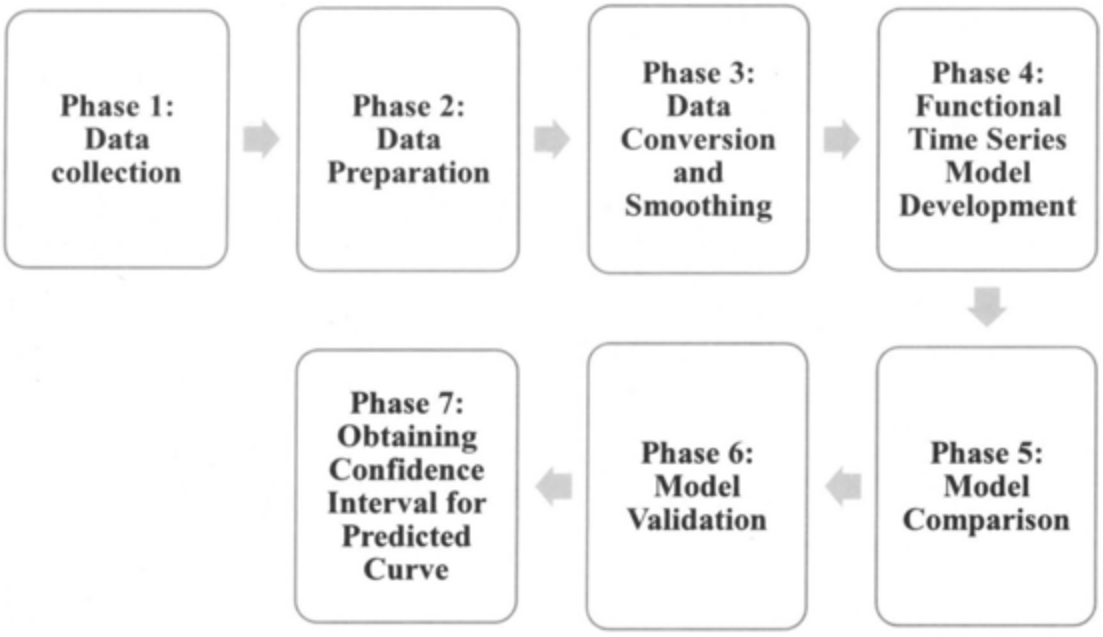


Figure 3.1 Research Famework

Firstly, data collection of hourly API will be obtained from Air Quality Division, Department of Environment Malaysia (DOE). Then, the data will go through the data preparation processes which are data arrangement and data cleaning. The third phase of this study is data conversion and smoothing which the analysis will use the basis expansion function. The phases then continued with functional time series model development, model comparison and model validation. Phase seven (7) which is the last phase of this study was to obtain the confidence interval for the forecasted curve. All the seven phases details will be explained in the following sections.

## 3.3 Data Collection

This study used data hourly recorded Air Pollutant Index (API) from year 2011 to 2018 which the data was obtained from Air Quality Division of Department of Environment (DOE), Malaysia. Permission was first been asked from the head of DOE due to the historical API data unavailable to the public and confidential. It took about two (2) weeks to get the data from the DOE. The original data consist of 24-hourly measured API with the total observations of 70,128 hourly API readings for each station covered 8 years period. A total of 140,256 data sets (70,128 observations × 2 stations) were used for the analysis. Two (2) air monitoring stations; Shah Alam which was located within residential or urban area and Pasir Gudang which was located within industrial area were used and the air quality monitoring station details were shown in in Table 3.1.

Table 3.1
Continuous Air Quality Monitoring Station Studied

| Station ID | Air Monitoring Station | Representative Name | Background |
|---|---|---|---|
| CA0001 | Sek. Men. Pasir Gudang 2, Pasir Gudang | Pasir Gudang | Industrial |
| CA0025 | Sek. Keb. TTDI, Shah Alam | Shah Alam | Urban |

Most of previous studies (Azid et al., 2014; Rahman et al., 2013; Siew et al., 2008) used Shah Alam monitoring stations as it has high API records. Shah Alam is

Malaysia's most develop region and is more susceptible to air pollution due to its geographical location, the development of large-scale industrial and commercial activities, densely populated areas and also high vehicular traffic.

Johor City, like any other city in the globe, is constantly evolving. Industrial development in the southern part of Johor could deteriorate the quality of life, especially the air quality in the urban area, and therefore should be greatly taken into consideration.

Figure 3.2 Continuous Air Quality Monitoring Station Studied

## 3.4 Data Preparation

After data has been collected, data preparation for the study analysis will be conducted which include data arrangement and data cleaning.

### 3.4.1 Data Arrangement

The observed data is arranged into a matrix or a data frame form. The purpose of this data arrangement is to prepare for data analysis such that days are the cases; the variables are hours, ie: there will be 24 variables representing 24 hours recorded variables. Figure 3.3 can be used to visualise the format of data representation.

$$\begin{array}{cc} Day & \begin{matrix} Hour1 & Hour2 & \dots & Hour24 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ \vdots \\ 2{,}922 \end{matrix} & \begin{bmatrix} y_{1,1} & y_{1,2} & \cdots & y_{1,p} \\ y_{2,1} & y_{1,2} & \cdots & y_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n,1} & y_{1,2} & \cdots & y_{n,p} \end{bmatrix} \end{array}$$

Figure 3.3 API Data Arrangement

Table 3.2
Actual API Data from DOE

| Site ID | Site_Location | Date | Time | API |
|---|---|---|---|---|
| CA0025 | Sek. Keb. TTDI Jaya, Shah Alam | 20110101 | 0100 | 36 |
| CA0025 | Sek. Keb. TTDI Jaya, Shah Alam | 20110101 | 0200 | 36 |
| CA0025 | Sek. Keb. TTDI Jaya, Shah Alam | 20110101 | 0300 | 36 |
| CA0025 | Sek. Keb. TTDI Jaya, Shah Alam | 20110101 | 0400 | 36 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| CA0025 | Sek. Keb. TTDI Jaya, Shah Alam | 20181231 | 2300 | 28 |
| CA0025 | Sek. Keb. TTDI Jaya, Shah Alam | 20181231 | 0000 | 28 |
| CA0001 | Sek. Men. Pasir Gudang 2, Pasir Gudang | 20110101 | 0100 | 31 |
| CA0001 | Sek. Men. Pasir Gudang 2, Pasir Gudang | 20110101 | 0200 | 31 |
| CA0001 | Sek. Men. Pasir Gudang 2, Pasir Gudang | 20110101 | 0300 | 31 |
| CA0001 | Sek. Men. Pasir Gudang 2, Pasir Gudang | 20110101 | 0400 | 31 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| CA0001 | Sek. Men. Pasir Gudang 2, Pasir Gudang | 20181231 | 2300 | 11 |
| CA0001 | Sek. Men. Pasir Gudang 2, Pasir Gudang | 20181231 | 0000 | 11 |

Table 3.2 shows some of the actual data given by DOE. In order to fit the analysis, the data has been transformed into the column of Year, Month, Day and Hours (from Hour 1 to Hour 24) for each location. The transformed data as shown in Table 3.3. Then, the data has been imported to R console to proceed with further analyses. Same process has been applied to the API data of two monitoring stations.

Table 3.3
Data Arrange into 24-Hour Columns

| Year | Month | Day | Hour | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | 1 | 2 | 3 | 5 | | 6 | ... | 24 |
| 2011 | January | 1 | 31 | 31 | 31 | 31 | 31 | 31 | ... | 40 |
| 2011 | January | 2 | 40 | 41 | 42 | 42 | 42 | 43 | ... | 41 |
| 2011 | January | 3 | 40 | 40 | 39 | 39 | 38 | 38 | ... | 32 |
| 2011 | January | 4 | 33 | 33 | 33 | 33 | 33 | 33 | ... | 35 |
| 2011 | January | 5 | 34 | 34 | 33 | 33 | 34 | 33 | ... | 40 |
| 2011 | January | 6 | 40 | 41 | 42 | 42 | 43 | 43 | ... | 40 |
| ⋮ | | | | | | | | | | |
| 2018 | December | 30 | 26 | 26 | 26 | 27 | 27 | 30 | ... | 18 |
| 2018 | December | 31 | 17 | 17 | 16 | 16 | 15 | 31 | ... | NA |

### 3.4.2 Data Cleaning

The accessibility of a complete data set is crucial in various statistical analyses. However, the issue of missing data often happens in the context of air quality information due to numerous factors such as machinery malfunction, human error and calibration process (Shaadan, Deni, & Jemain, 2015). When using incomplete records as an input in the analysis, the results of any statistical model and analysis might deteriorate. Because of this, the estimation of missing values in the process of data preparation becomes the first priority.

Two general approaches to solve this issue are case deletion and imputation, which are the common techniques used (Shaadan et al., 2015). Although deletion of cases is the most prevalent and easiest technique, it has the disadvantage of losing information due to data reduction. Thus, this study applied an imputation technique to handle the missing values.

Hundreds of missing records have been generated from the API data set has been recorded daily at hourly basis at the Shah Alam and Pasir Gudang air quality monitoring stations located in the west and south parts of Peninsular Malaysia respectively. A matrix data set with 2922 days (rows) by 24 hours (column) which was recorded in the year 2011 to 2018 is considered for each station. The data contains small percentage of missing values; 2.96% missing value for Shah Alam and 2% missing value for Pasir Gudang (as shown in Table 3.4). The missing values have been imputed with the Principal Components Analysis (PCA) approach.

Table 3.4
Percentage of Missing Values

| Station ID | Location | Days Missing | Percent Missing |
|---|---|---|---|
| CA0001 | Pasir Gudang | 2,073 | 2.96% |
| CA0025 | Shah Alam | 1,457 | 2.00% |

Given that a few data points are missing within the 24 hours period of a day curve, the aimed was to replace the missing value $y_{miss}$ at time $t$ with a value on the estimated curve $x(t)$ at the same time $t$ being missing, where $y_{miss} = x(t)$. Examples for the possible condition of the missing values are illustrated in Figure 3.4.

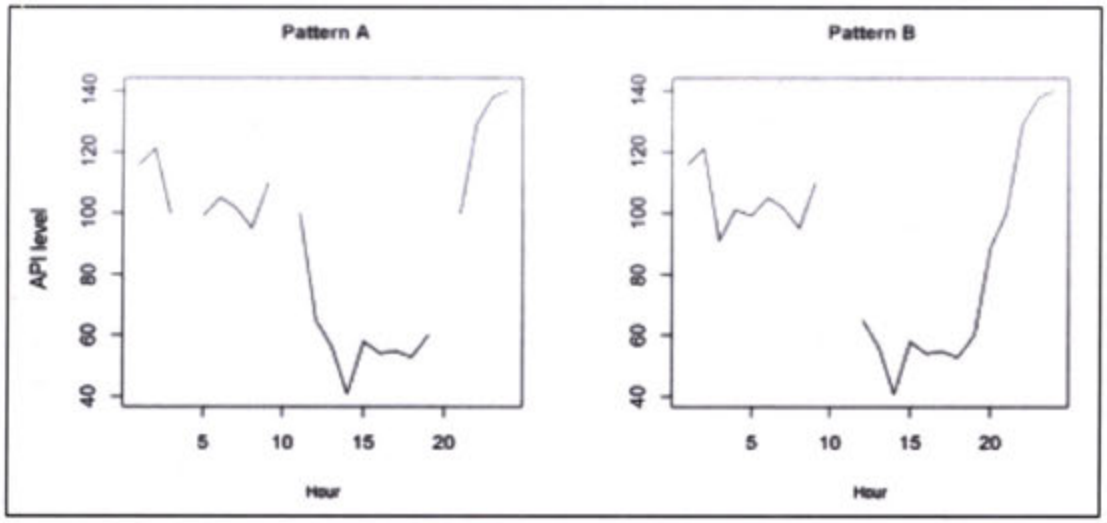

Figure 3.4 Possible Patterns of Missing Values Within A Day Curve

## 3.5 Data Conversion and Smoothing

### 3.5.1 Data Conversion

Before the data been smoothed and conversed to the functional time series, the actual data were daily data in hourly basis covered period of 2011 – 2018. Since this study aimed to forecast the monthly diurnal maximum API, thus the maximum for each month were obtained using commands in R (as attach in Appendix).

In the study analyses, the monthly and hourly recorded maximum API data are treated as functional data or curves instead of discrete monthly maximum values. Therefore, the original monthly by hourly recorded data must be first converted into functional data or curves. Figure 3.5 shows the example of five months API curves after data conversion process was conducted. There are n = 5 months API curves and $x$ is variable time within the period of $T$ that lies between 1 to 24 such that is $x \in T$ where $T = [1,24]$ where $i = 1,2, \dots ,24$. Data conversion is defined as converting discrete data into functional data or curves.

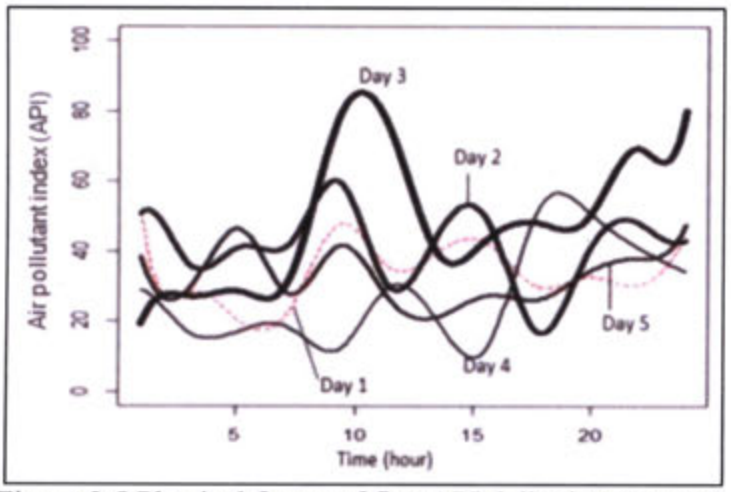


Figure 3.5 Physical forms of five API daily data (curves)

Let's consider a set of discrete observations $y = [y_{1,1}, y_{1,2} \cdots, y_{t,24}]$ recorded at time point $t = [1,2, \dots, n]$. Data conversion is defined as transforming the discrete data $y$ into a continuous function $y_t(x_i)$ that can be computed at any time $t$ by the following mathematical model:

$$y = y_t(x_i) + \varepsilon \tag{3.1}$$

where notation $\varepsilon$ is known as random noise in the data.

### 3.5.2 Data Smoothing Using Basis Expansion Method

The term $y_t(x_i)$is assumed as a smooth function and will be estimated using the basis function expansion. For this study, any day $y_t$, the API data $(y)$ recorded on month $t = [1,2,\dots,n]$ and at hourly time $i = [1,2,\dots,24]$, is converted into a continuous function $y_t(x_i)$ using basis function:

$$y_t(x_i) = \sum_{k=1}^{K} C_k \varphi_k(x_i) \tag{3.2}$$

The term $\varphi_k(x_i)$ is the chosen basis system consisting of $k$ number of basis functions. There are various systems of basis function including Fourier, spline, polynomial, quadratic, constant and etc that can be used as the basis. However, which basis is to be chosen depends on the underlying pattern of the data. The Fourier basis is appropriate for periodic data, whereas the spline is suitable for non-periodic data and was reported as a more flexible method (Ramsay & Silverman, 2006). The term $C_k$ is the corresponding basis coefficient.

## 3.6 Functional Time Series Model Development

The development of functional time series model involves the combination of Functional principal Component Regression (FPCR) and Functional Principal Component Analysis (FPCA).

### 3.6.1 Functional Principal Component Regression (FPCR)

Following the research by Hyndman and Shang (2009), this study employs the

application of FPCR to model and predict functional time series. First, the researcher defines the problem more precisely before reviewing the FPCR. Let $y_t(x_i)$ represent a function of API for the continuous hour variable $x_i$ in day $t$. Functional time series techniques able to pick up the underlying dynamic of the multiple API level at multiple time in the data.

The researcher assumes that there is an underlying smooth function $f_t(x_i)$, which observes with error at discretized grid points of $x$. In this study, the researcher observes $\{x_i, y_t(x_i)\}$ for $t = 1,2,\dots,n$ and $i = 1,2,\dots,p$, from which we obtain a smooth function $f_t(x_i)$, presented by

$$y_t(x_i) = f_t(x_i) + \sigma_t(x_i)\varepsilon_{t,i} \tag{3.3}$$

where $\varepsilon_{t,i}$ is an independent and identically distributed (iid) standard normal random variable, $\sigma_t(x_i)$ allows the amount of noise vary with $x_i$, and $\{x_1, x_2, \dots, x_p\}$ is a set of discrete data points. Given a set of functional data $f(x) = [f_1(x), f_2(x), \dots, f_n(x)]^{\mathsf{T}}$, the researcher is interested in discovering underlying patterns using the FPCR, from which he/she acquire forecasts of $y_{n+h}(x)$, where h represents the forecast horizon.

The aim of this study is to forecast the next day curves based on series of previous day curves with the equation:

$$y_{n+1}(x_i) = f_{n+1}(x_i) + \sigma_t(x_i)\varepsilon_{n+1,i} \tag{3.4}$$

Now let's discuss on the term $f_t(x_i)$. At a population level, a stochastic process indicated by $f$ can be composed into mean function and the sum of the products of orthogonal functional principal components and uncorrelated principal component scores. It can be expressed as

$$f = \mu + \sum_{k=1}^{\infty} \beta_k \emptyset_k \tag{3.5}$$

where μ is the unobservable population mean function, $\beta_k$ is the $k^{\text{th}}$ principal component scores, and $\emptyset_k$ is the $k^{\text{th}}$ population functional principal component.

In this study, following Hyndman and Ullah (2007), the researcher used nonparametric smoothing on each curve $y_t(x_i)$ separately to get estimates of the smooth function $\{f_t(x_i)\}$. Then a functional principal component method proposed to decompose the time series of functional data into a number of principal components and their scores. The model can be written as below:

$$f_t(x_i) = \mu(x_i) + \sum_{k=1}^{K} \beta_{t,k} \emptyset_k(x_i) + \varepsilon_t\,(x_i) \tag{3.6}$$

where $\emptyset_k(x_i)$ is the $k^{\text{th}}$ principal component function, $\{\beta_{1,k}, \ldots, \beta_{n,k}\}$ are the corresponding scores, $\varepsilon_t(x_i)$ denote independent and identically distributed random functions with zero mean, and $K < n$. Because principal component scores are uncorrelated, Hyndman and Ullah (2007) recommended that each univariate time series $\{\beta_{t,k}\}$, $k = 1, \ldots, K$, can be forecasted separately using univariate time series model. The estimated future curves are obtained by multiplying the forecasted principal component scores with the principal components. The mean curve is represented as in Figure 3.6.

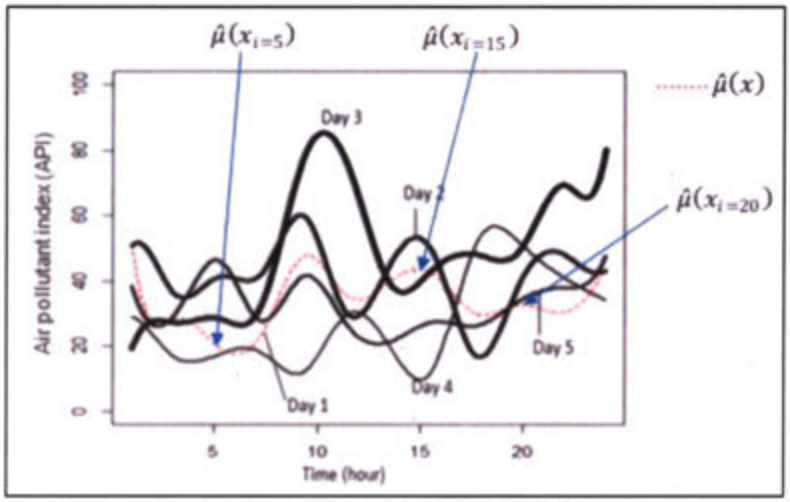

Figure 3.6 Mean Curve $\hat{\mu}(x)$

The mean function $\mu(x_i)$ is estimated using weighted average

$$\hat{\mu}(x_i) = \sum_{t=1}^{n} \omega_t \hat{f}_t(x_i) \tag{3.7}$$

where $\hat{f}_t(x_i)$ is the smoothed curve estimated from $y_t$, and $\omega_t = K(1-K)^{n-t}$ is a geometrically decreasing weight with $0 < K < 1$. If one requests for a robust estimator, then L median of the estimated smoothed curves can be used instead (Hyndman & Ullah, 2007). The mean or median-adjusted functional data are represented as $f_t^*(x_i) = \hat{f}_t(x_i) - \hat{\mu}(x_i)$.

### 3.6.2 Functional Principal Component Analysis (FPCA)

FPCA is one of the most important exploratory tools in FDA. The analysis is conducted using mean centred curves because the interest is primarily in characterizing the main deviations of each curve from the average curves (Ramsay & Silverman, 2006). In comparison to FPCA, PCA is known as a classical approach to the exploration of variation in multivariate data, while FPCA is used for functional data or curves. Both methods are generally aimed and used for data reduction, but the key difference is that, PCA uses discrete point data, while FPCA uses functional data. The methods use an eigenvalue decomposition of the covariance matrix to find direction in the observation space along which the data have the highest variability. The direction of variation in

PCA is represented by loading vectors, while in the functional context; each principal component is specified by a principal component weight function $\xi(t)$.

In this study, FPCA is conducted using the mean centered curves $z_i(t)=x_i(t)-\bar{x}(t)$, $i=1,\ldots,24$ for 24 average diurnal curves from 24 hourly API. A degree three b-spline is chosen as the basis, and $k = 15$ equal number of basis functions were used in the data conversion process. The number of basis $k$ was determined by applying the Bayesian Information Criteria (BIC) to the functional mean curves, $x(t)$ of the 24 hours (Huang & Shen, 2004). The main aim of FPCA is to search for several important (principal) components that can describe the major variation in the API curves. Before FPCA is conducted, it is a need to determine how many numbers of the principal components (PC) to be retained so that the components are enough to convey important information in the data. In this case, following approach by Ramsay, Hooker and Graves (2009), the technique of principal component ranking with respect to eigenvalues using scree-plot is used.

The first step in FPCA is to find the first eigen-weight function $\xi(t)$ that maximizes the mean square of the component score, that is $n^{-1}\sum_i S_{i1}$ for which $S_{i1} = \int \xi_1(t)z_i(t)dt$, $i = 1,\ldots,24$ subject to the normality constraint $\int \xi_1(t)^2\,dt = 1$. Repeat the subsequent step until the desired number of PC for example $m$. The weight function for $m^{\text{th}}$ PC, the $\xi_m(t)$ is also required to satisfy $\int \xi_m(t)^2 dt = 1$ and $\int \xi_m(t)\xi_k(t)dt = 0$, $k < m$. Given the variance-covariance function such as:

$$v(s,t) = n^{-1}\sum_{i=1}^{n} z_i(s)z_i(t) \tag{3.8}$$

We can obtain a set of eigen-weight functions $\xi(t) = [\xi_1, \ldots, \xi_m]$ and a set of eigenvalues $\lambda = [\lambda_1, \ldots, \lambda_m]$ by solving the eigen-equation problem given by:

$$\int v(s,t)\xi(t)dt = \lambda\xi(s) \tag{3.9}$$

Each principal component accounts for a different proportion of the variability in the curves which is given by an eigenvalue. The first capture the greatest amount of the variation, the second captures the second greatest amount of the variation, and so on, and are independently indicating different information. In terms of the computational aspect, the eigen-analysis problem is transforming into matrix eigen-analysis task, either by discretizing the functions or using basis function expansion method was used by adopting the approach by Ramsay and Silverman (2006).

## 3.7 Model Comparison

By conditioning on the set of smoothed functions $f(x) = [f_1(x), f_2(x), \ldots, f_n(x)]^{\mathsf{T}}$ and the fixed functional principal components $B = [\widehat{\emptyset}_K(x), \widehat{\emptyset}_K(x), \ldots, \widehat{\emptyset}_K(x)]^{\mathsf{T}}$, the $h$-step-ahead forecasts can be obtained using $y_{n+h}(x)$ as in (3.10).

$$\hat{y}_{n+h|n}(x) = E[y_{n+h}(x)|f(x), B] = \bar{f}(x) + \sum_{K=1}^{K} \hat{\beta}_{n+h|n,k} \widehat{\emptyset}_K(x) \tag{3.10}$$

where $\hat{\beta}_{n+h|n,k}$ denotes the $h$-step-ahead forecasts of $\beta_{n+h,k}$ using a univariate time series.

### 3.7.1 Multi-Step-Ahead Forecast

"Direct" forecasts or multi-step ahead times series forecasts are created using a horizon-specific forecasted model, where the dependent variable is the multi-period ahead value being predicted. For example, assumed $n = 365$, forecast of 3-step ahead forecast can be obtained as

$$\hat{y}_{368} = \hat{y}_{365+3|365}(x) = E[y_{365+3}(x)|f(x), B] = \bar{f}(x) + \sum_{K=1}^{K} \hat{\beta}_{365+3|365,k} \widehat{\emptyset}_K(x)$$

### 3.7.2 Iterative One-Step-Ahead Forecast

Iterative one-step ahead time series forecasts are created using a one-period ahead model, iterated forward for the number of periods desired. For example, as comparison for above forecast, the three (3) iterative one-step ahead forecast can be obtained as

Iteration 1:

$$\hat{y}_{366} = \hat{y}_{365+1|n}(x) = E[y_{365+1}(x)|f(x), B] = \bar{f}(x) + \sum_{K=1}^{K} \hat{\beta}_{365+1|365,k}\widehat{\emptyset}_K(x)$$

Iteration 2:

$$\hat{y}_{367} = \hat{y}_{366+1|366}(x) = E[y_{366+1}(x)|f(x), B] = \bar{f}(x) + \sum_{K=1}^{K} \hat{\beta}_{366+1|366,k}\widehat{\emptyset}_K(x)$$

Iteration 3:

$$\hat{y}_{368} = \hat{y}_{367+1|367}(x) = E[y_{367+1}(x)|f(x), B] = \bar{f}(x) + \sum_{K=1}^{K} \hat{\beta}_{367+1|367,k}\widehat{\emptyset}_K(x)$$

### 3.7.3 Performance Indicator for Model Comparison

This section briefly describes accuracy measures that can be calculated for the residuals of a prediction model in FDA. In general, there is not a standard criterion for functional errors and it usually depends on the application. Nevertheless, some functional accuracy measures can be computed based on the $L^p$ norm:

The Functional Mean Average Error (FMAE), also called the Mean Integrated Average Error (MIAE), is defined as

$$FMAE = N^{-1}\sum_{t=1}^{T}\|Y_t - \hat{Y}_t\|_{L_1} = N^{-1}\sum_{t=1}^{T}\int|Y_t - \hat{Y}_t(u)|\,du \qquad (3.11)$$

The Functional Mean Square Error (FMSE), also called the Mean Integrated Square Error (MISE), is defined as

$$FRMSE = \sqrt{N^{-1}\sum_{t=1}^{T}\| Y_t - \hat{Y}_t \|_{L_2}^2} = \sqrt{N^{-1}\sum_{t=1}^{T}\int \left(Y_t - \hat{Y}_t(u)\right)^2} \tag{3.12}$$

These error measurements can provide some results that are hard to interpret because the integration is done in the domain range of the functions $u \in [a, b]$. A way to normalize the results is by dividing each measure by $b - a$.

Finally, a functional version of MAPE could be defined as the norm of the residual normalized by the norm of the real functional observation

$$FMAPE = N^{-1}\sum_{t=1}^{T}\frac{\| Y_t - \hat{Y}_t \|_{L^2}}{\| Y_t \|_{L^2}} = N^{-1}\sum_{t=1}^{T}\frac{\int \left| Y_t(u) - \hat{Y}_t(u) \right| du}{\int |Y_t(u)|\, du} \tag{3.13}$$

### 3.7.4 Model Validation

By comparing the predicted future curves using the best model obtained with benchmark curves.

## 3.8 Prediction Interval

Prediction intervals are an important tool for evaluating the probabilistic doubt related with point forecasts. As was stressed by Chatfield (2000), it is crucial to provide both interval forecasts and point forecasts in order to

i) evaluate future uncertainty level;

ii) allow for the planning of different strategies for the range of possible outcomes indicated by the interval forecasts;

iii) compare forecasts from different methods more thoroughly; and

iv) explore different situations based on different assumptions.

The researcher calculates the forecast variance that follows from (3.3) and (3.6) to construct prediction interval. Because of orthogonality, the forecast variance can be approximated by the sum of component variances

$$\begin{aligned}\xi_{n+h}(x) &= Var[y_{n+h}(x)|f(x),\beta] \\ &= \hat{\sigma}_{\mu}^{2}(x) + \sum_{k=1}^{K} u_{n+h,k}\emptyset_{k}^{2}(x) + v(x) + \sigma_{n+h}^{2}(x)\end{aligned} \tag{3.14}$$

where $u_{n+h,k} = Var(\beta_{n+h,k}|\beta_{1,k},\beta_{2,k},\ldots,\beta_{n,k})$ can be gained from the time series model, and the model error variance $v(x)$ is estimated by averaging $\{\hat{\varepsilon}_1^2(x),\hat{\varepsilon}_2^2(x),\ldots,\hat{\varepsilon}_n^2(x)\}$ for each $x$, and $\hat{\sigma}_{\mu}^{2}(x)$ can be gained from the nonparametric smoothing methods used.

Based on the normality assumption, the 95% prediction interval for $y_{n+h}(x)$ is composed as $\hat{y}_{n+h|n}(x) \pm z_{0.05}\sqrt{\xi_{n+h}(x)}$ where $z_{0.05}$ is the $(1-\frac{0.05}{2})$ standard normal quantile. A 95% confidence interval is a range of values that we can be 95% certain contains the true mean of the population. This study used the 95% confidence level which only allow 5% error due to in forecasting, the error cannot be too large, otherwise it will give wrong information to the people.

# CHAPTER FOUR
# ANALYSIS AND DISCUSSIONS OF RESULTS

## 4.1 Introduction

This chapter discusses in detail regarding the results and findings of this research study. The monthly diurnal maximum API data for two (2) chosen air monitoring stations; Shah Alam and Pasir Gudang was analysed using Microsoft Excel and R programming. The data first went through the preliminary investigation where the data has been describe using the graph, mean and standard deviation. Then, the functional time series analyses were run in order to achieve all the three (3) objectives of this study. The results and discussions will be further explained in the following sections.

## 4.2 Functional Descriptions of Monthly Diurnal Maximum API Data

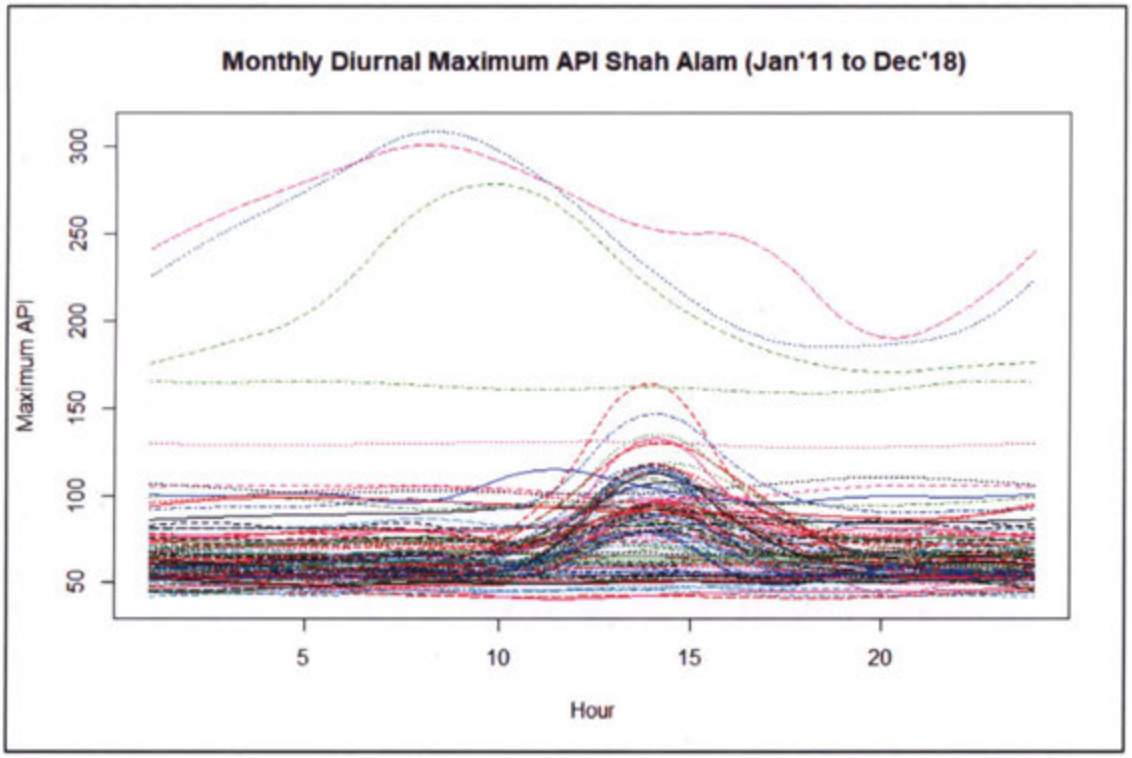


Figure 4.1 Monthly Diurnal Maximum API Curves for Shah Alam (Jan'11 to Dec'18)

Figure 4.1 shows that for Shah Alam air monitoring station, the monthly maximum API curves over a 24-hour period are about the same with the highest peak at 3 p.m. A few curves, however, seem to have the peak level at 10 a.m. before declining until night. These few curves are the curve of maximum API for June 2013, September and October 2015, where deterioration of air quality has occurred due to transboundary pollution, massive land and forest fire in Kalimantan and Sumatra.

Furthermore, from Figure 4.1 seem there are two significant groups of curve patterns across the hours. A group of curves with consistent behaviour was observed at the lower level of maximum API. On the other hand, a sparser curve at the higher level indicated extreme behaviour. Thus, the obvious difference in the behaviour of the diurnal monthly curves might be attributed to the different emission source. The changes in the meteorological condition and the air pollution emission from the transportation activities caused the occurrence of the maximum peak at around 15:00 in the afternoon for most of the months, whereby the highest API of 300 was recorded at around 9 a.m. in June 2013.

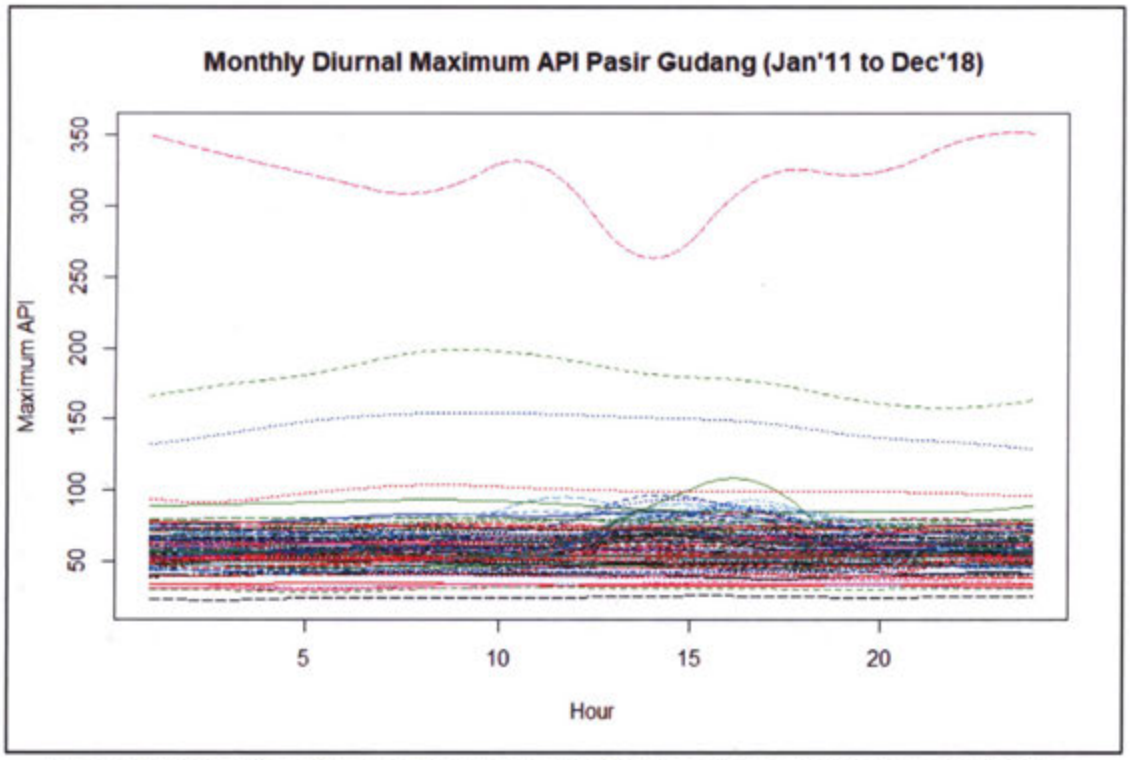


Figure 4.2 Monthly Diurnal Maximum API Curves for Pasir Gudang (Jan'11 to Dec'18)

Figure 4.2 reveals that, except for 3 months, the maximum API curves for Pasir Gudang station over a 24 -hour period reported moderate API readings (API level below 100). The peak level with 350 API reading occurred at midnight indicates hazardous air quality status. It is apparently seen that there are two significant groups of curve patterns across the hours. A group of curves with consistent behaviour was observed at the lower level of maximum API. On the other hand, a sparser curve at the higher level indicated extreme behaviour. Thus, the obvious difference in the behaviour of the diurnal monthly curves might be attributed to the different emission source or due to particular events/incidents such as haze, transboundary pollution, etc.

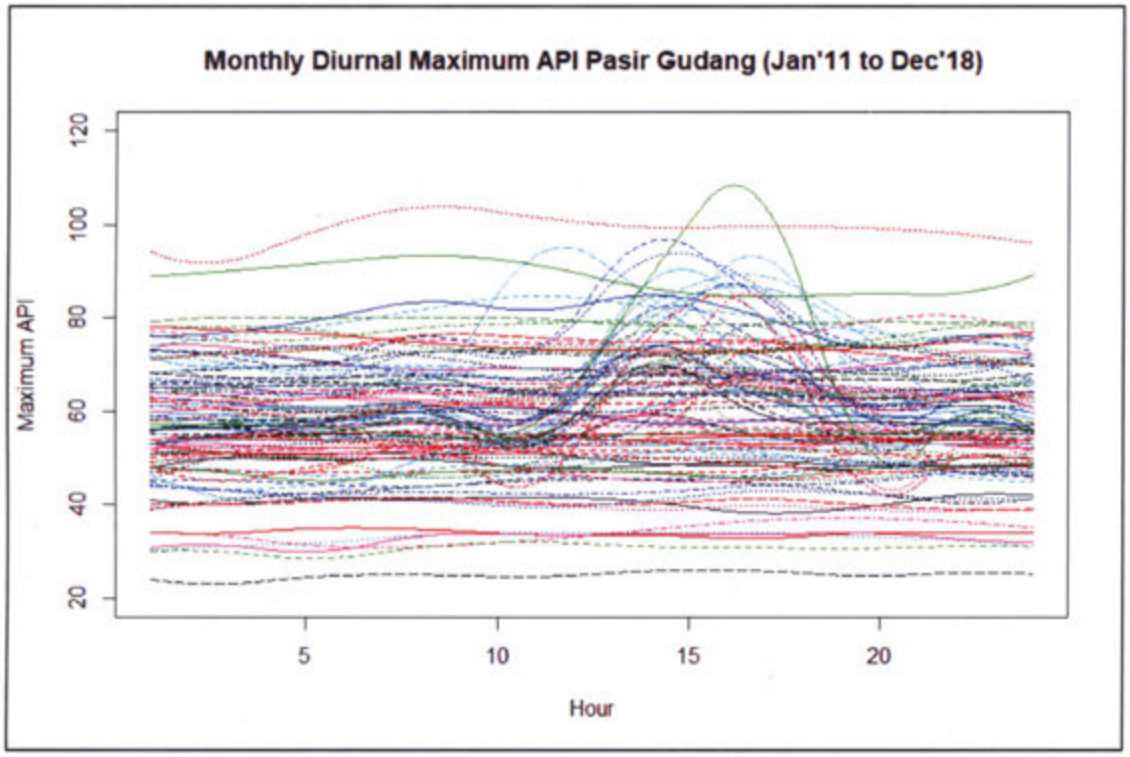


Figure 4.3 Monthly Diurnal Maximum API Curves for Pasir Gudang (focus on maximum value of 20 till 120)

Figure 4.3 was generated to further investigate Pasir Gudang's pattern of maximum API curves concentrating the reading at the lower level of maximum API; API reading below 120. Commonly, at 15:00 in the afternoon, Pasir Gudang station recorded the highest peak of maximum API. Maximum API readings recorded lower in the morning and night, same as Shah Alam. However, in comparison with Shah Alam, Pasir Gudang reported lower maximum API readings. The changes in the meteorological condition and the air pollution emission from the transportation

activities caused the occurrence of the maximum peak at around 15:00 in the afternoon for most of the months.

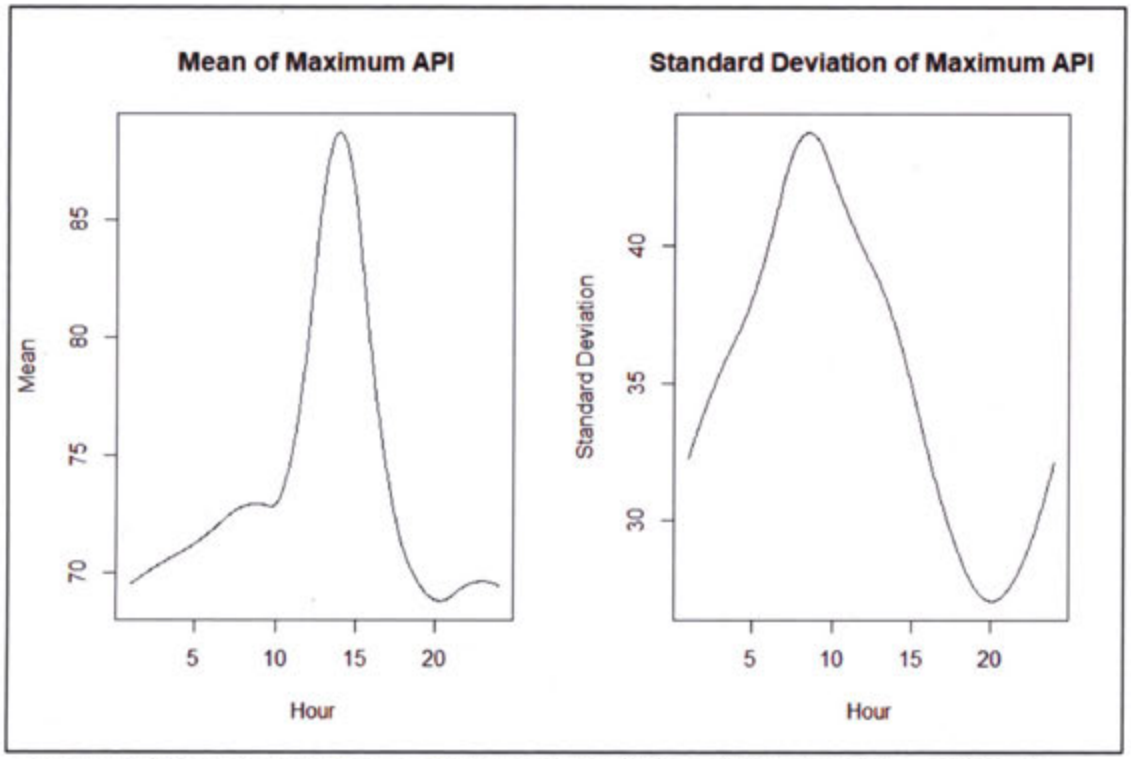


Figure 4.4 Functional Mean and Standard Deviation of Monthly Diurnal Maximum API Shah Alam

Figure 4.4 shows the mean and standard deviation curves of monthly maximum API over 24 hours period for Shah Alam station. The mean curve level is between API readings of 60 to 90 and the curve's highest level was at 3 p.m. in the evening with the mean approximately 90. Based on the mean curve, it seems that air pollution of day-to-day basis does not look harmful or serious; the API readings still below 100 indicates moderate air quality.

However, by looking at the standard deviation curve (as shown in Figure 4.4), it seems that the variation quite high which indicating that some of the days, API reading might shoot high. The standard deviation curve's highest level was at 10 a.m. in the morning indicates that dispersion is the highest at that time. This might due to several factors that need to be further studied. It is suggested that statistical process control investigation and analysis need to be done. So that, we can detect the sources that cause the high variation in the API values at some particular hours.

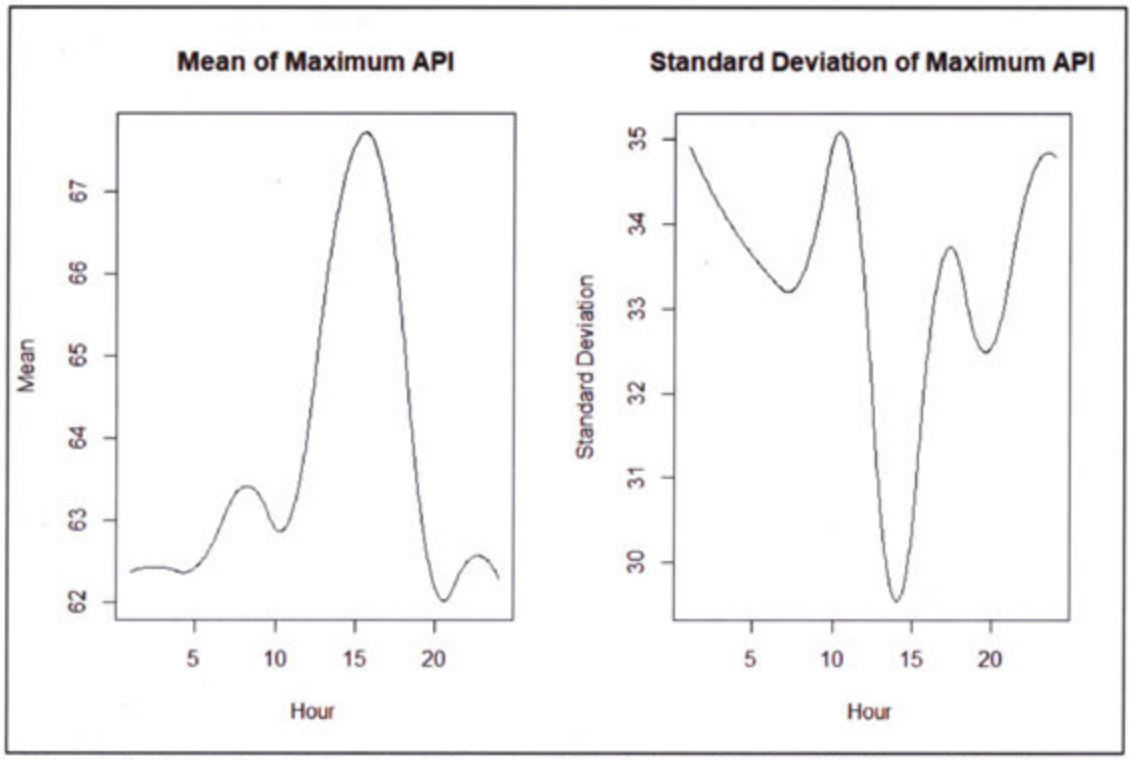


Figure 4.5 Functional Mean and Standard Deviation of Monthly Diurnal Maximum API Pasir Gudang

Figure 4.5 shows the mean and standard deviation curves of monthly maximum API over 24 hours period for Pasir Gudang station. The mean curve level is between API readings of 62 to 68 and the curve's highest level was at 3 p.m. in the evening with the mean approximately 68. Based on the mean curve, day-to-day air pollution does not appear harmful or serious with the API readings still below 100 throughout day indicates moderate air quality.

However, by looking at the standard deviation curve (as shown in Figure 4.5), it seems that the variation quite high which indicating that some of the days, API reading might shoot high. The curve's highest level was at 10 a.m. in the morning indicates that the dispersion is very high at that time. This might due to several factors that need to be further studied. It is suggested that statistical process control investigation and analysis need to be done. Thus, we can detect the sources that cause the high variation in the API values. In addition, the least dispersion happened at 2 p.m. that the data did not vary greatly in comparison with other hours.

From both Figure 4.4 and Figure 4.5, it can be summarised that Shah Alam experienced higher air pollution compared to Pasir Gudang. Pasir Gudang, however, reported greater dispersion of API readings compared to Shah Alam.

## 4.3 To Compare the Performance Between Multi-Step-Ahead and Iterative One-Step-Ahead Forecast in Forecasting Monthly Maximum API

### 4.3.1 Multi-Step-Ahead Forecast

By conditioning on the set of smoothed functions $f(x)$ and the fixed functional principal components $B$, the $h$-step-ahead forecasts obtained using $y_{n+h}(x)$ and resulting to the values as in Table 4.1 and Table 4.2, as well as the results depicted as in Figure 4.7 and Figure 4.9.

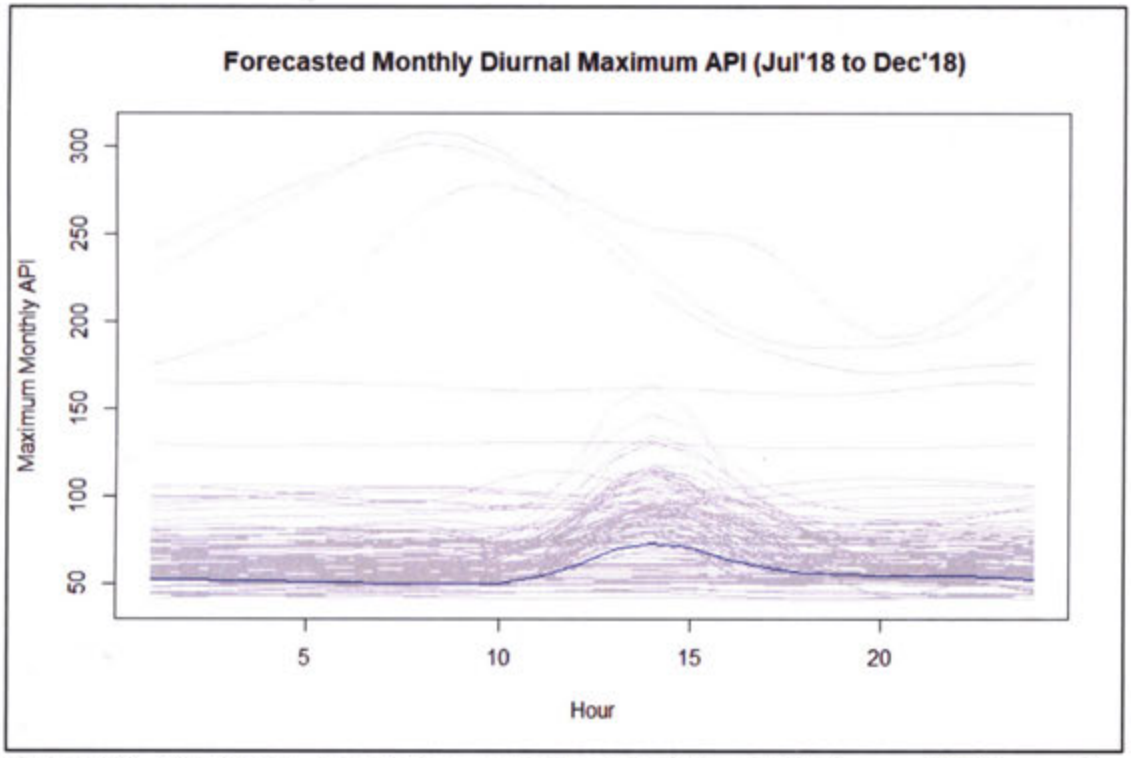


Figure 4.6 Multi-Step-Ahead Forecasted Monthly Diurnal Maximum API Curves for Shah Alam (Jul'18 to Dec'18)

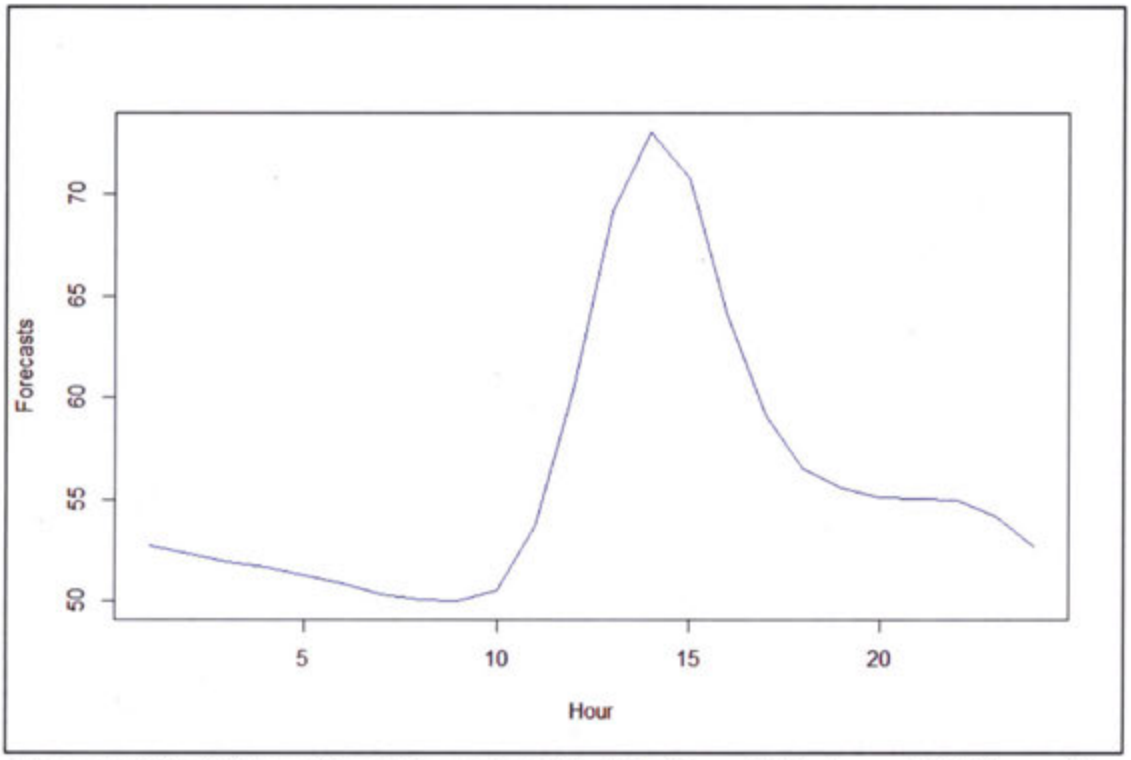

Figure 4.7 Multi-Step-Ahead Forecasted Monthly Diurnal Maximum API Curves for Shah Alam (Jul'18 to Dec'18)

Figure 4.6 shows the forecasts of maximum API data of Shah Alam from July 2018 to December 2018 highlighted in blue colour, while the data used for estimation are greyed out. The forecasted curves should exhibit rainbow colour with 6 different curves indicating a curve to a month. This multi-step-ahead forecast method, however, cannot exhibit multiple curves for maximum API data, resulting same curves exhibit for every month forecasted.

Figure 4.7 which shows clear visualisation (enlarged form of curve depicted in Figure 4.6) of the forecasted curves for July 2018 to December 2018 indicates that the maximum API achieved the highest peak level of about 80 at 15:00 in the evening. The curves are below 100 throughout the 24 hours of the forecasted months indicating that Shah Alam's air quality will be in a healthy state for the second half of year 2018.

Table 4.1 below shows the details of the forecasted values, which it can be seen that the forecasted diurnal maximum API were the same for every forecasted month. This insight shows that this model can be further improvise in future.

Table 4.1
Forecasted Values of Monthly Diurnal Maximum API for Shah Alam using Multi-Step-Ahead Forecasts

| Hour | Month | | | | | |
|---|---|---|---|---|---|---|
| | **Jul 2018** | **Aug 2018** | **Sep 2018** | **Oct 2018** | **Nov 2018** | **Dec 2018** |
| 1 | 52.6989 | 52.6989 | 52.6989 | 52.6989 | 52.6989 | 52.6989 |
| 2 | 52.3119 | 52.3119 | 52.3119 | 52.3119 | 52.3119 | 52.3119 |
| 3 | 51.9607 | 51.9607 | 51.9607 | 51.9607 | 51.9607 | 51.9607 |
| 4 | 51.6443 | 51.6443 | 51.6443 | 51.6443 | 51.6443 | 51.6443 |
| 5 | 51.2934 | 51.2934 | 51.2934 | 51.2934 | 51.2934 | 51.2934 |
| 6 | 50.8644 | 50.8644 | 50.8644 | 50.8644 | 50.8644 | 50.8644 |
| 7 | 50.3686 | 50.3686 | 50.3686 | 50.3686 | 50.3686 | 50.3686 |
| 8 | 50.0560 | 50.0560 | 50.0560 | 50.0560 | 50.0560 | 50.0560 |
| 9 | 50.0287 | 50.0287 | 50.0287 | 50.0287 | 50.0287 | 50.0287 |
| 10 | 50.5049 | 50.5049 | 50.5049 | 50.5049 | 50.5049 | 50.5049 |
| 11 | 53.7495 | 53.7495 | 53.7495 | 53.7495 | 53.7495 | 53.7495 |
| 12 | 60.3684 | 60.3684 | 60.3684 | 60.3684 | 60.3684 | 60.3684 |
| 13 | 69.1714 | 69.1714 | 69.1714 | 69.1714 | 69.1714 | 69.1714 |
| 14 | 73.0558 | 73.0558 | 73.0558 | 73.0558 | 73.0558 | 73.0558 |
| 15 | 70.8299 | 70.8299 | 70.8299 | 70.8299 | 70.8299 | 70.8299 |
| 16 | 64.1425 | 64.1425 | 64.1425 | 64.1425 | 64.1425 | 64.1425 |
| 17 | 59.1659 | 59.1659 | 59.1659 | 59.1659 | 59.1659 | 59.1659 |
| 18 | 56.4953 | 56.4953 | 56.4953 | 56.4953 | 56.4953 | 56.4953 |
| 19 | 55.5893 | 55.5893 | 55.5893 | 55.5893 | 55.5893 | 55.5893 |
| 20 | 55.1284 | 55.1284 | 55.1284 | 55.1284 | 55.1284 | 55.1284 |
| 21 | 55.0518 | 55.0518 | 55.0518 | 55.0518 | 55.0518 | 55.0518 |
| 22 | 54.9495 | 54.9495 | 54.9495 | 54.9495 | 54.9495 | 54.9495 |
| 23 | 54.1692 | 54.1692 | 54.1692 | 54.1692 | 54.1692 | 54.1692 |
| 24 | 52.7023 | 52.7023 | 52.7023 | 52.7023 | 52.7023 | 52.7023 |

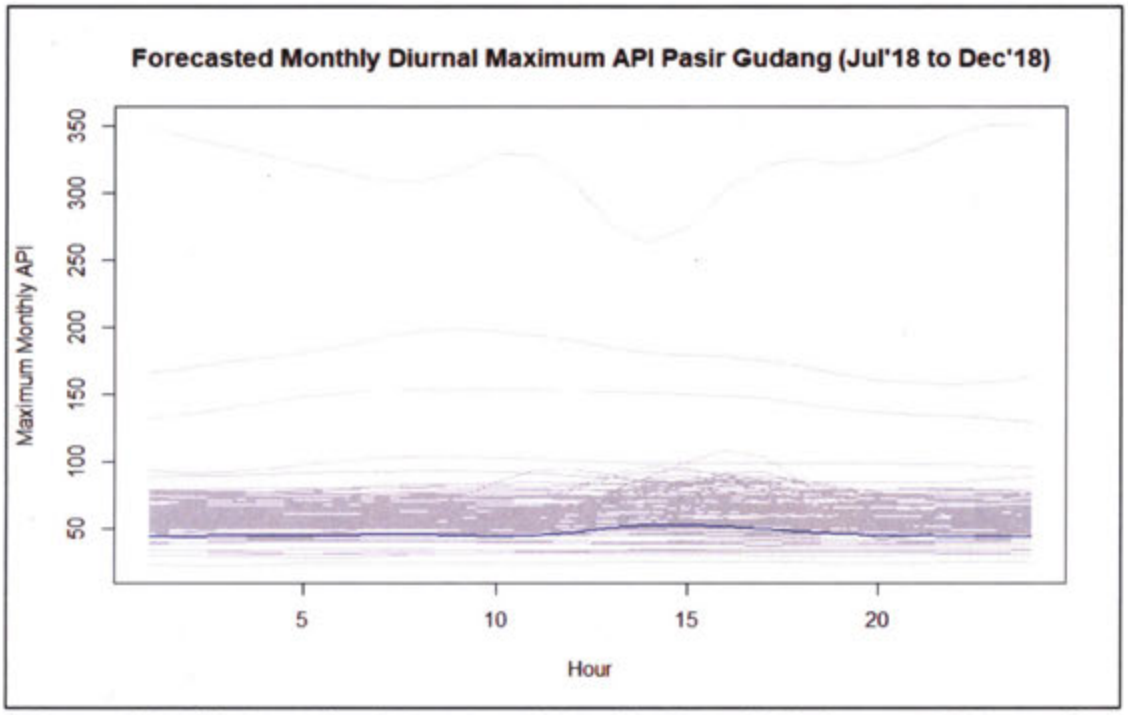


Figure 4.8 Multi-Step-Ahead Forecasted Monthly Diurnal Maximum API Curves for Pasir Gudang (Jul'18 to Dec'18)

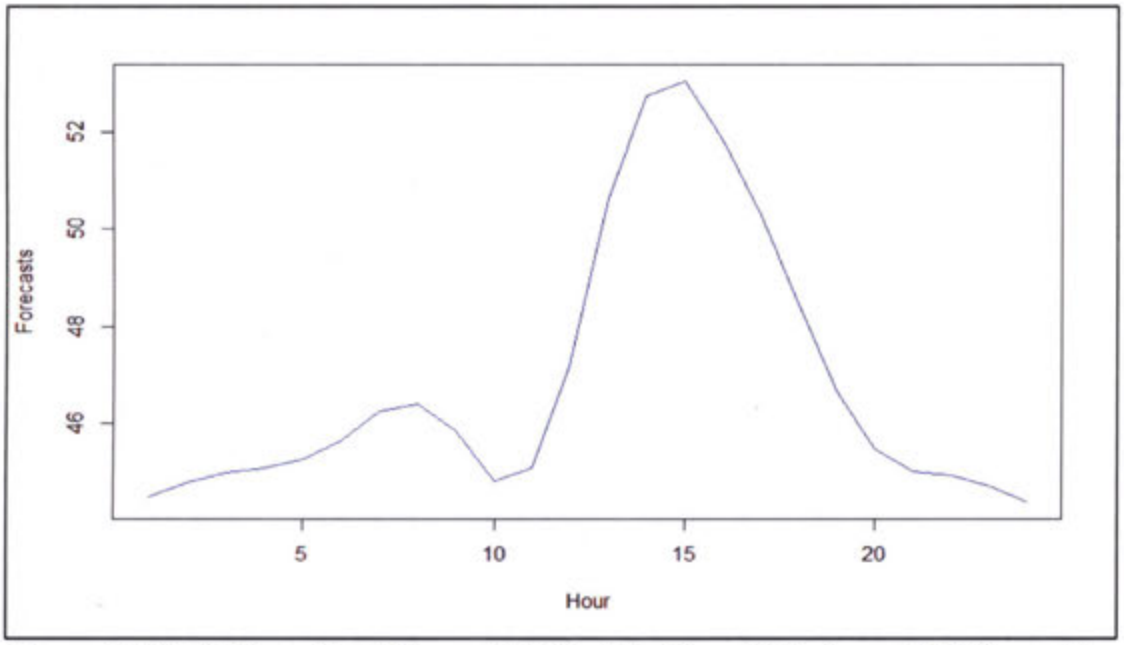


Figure 4.9 Multi-Step-Ahead Forecasted Monthly Diurnal Maximum API Curves for Pasir Gudang (Jul'18 to Dec'18)

Figure 4.8 shows the forecasts of maximum API data of Pasir Gudang from July 2018 to December 2018 highlighted in blue colour, while the data used for estimation are greyed out. The forecasted curves should exhibit rainbow colour with 6 different curves indicating a curve to a month. However, this forecast method cannot exhibit multiple curves for maximum API data, this method resulting to same curves for every

month forecasted.

Figure 4.9 which shows clear visualisation of the forecasted curves July 2018 to December 2018 indicates that the maximum API achieved it peak level of about 53 at 3 p.m. in the evening then remained at the lower level during night. Thus, the result gives the information to the public that Pasir Gudang's air quality status for the second half of 2018 will be in a healthy state and people can schedule their day-to-day activities freely. Table 4.2 below shows the details of the forecasted values, which the forecasted diurnal maximum API were the same for every forecasted month.

Table 4.2
Forecasted Values of Monthly Diurnal Maximum API for Pasir Gudang using Multi-Step-Ahead Forecasts

| Hour | Month | | | | | |
|---|---|---|---|---|---|---|
| | **Jul 2018** | **Aug 2018** | **Sep 2018** | **Oct 2018** | **Nov 2018** | **Dec 2018** |
| 1 | 44.4712 | 44.4712 | 44.4712 | 44.4712 | 44.4712 | 44.4712 |
| 2 | 44.7798 | 44.7798 | 44.7798 | 44.7798 | 44.7798 | 44.7798 |
| 3 | 44.9785 | 44.9785 | 44.9785 | 44.9785 | 44.9785 | 44.9785 |
| 4 | 45.0698 | 45.0698 | 45.0698 | 45.0698 | 45.0698 | 45.0698 |
| 5 | 45.2520 | 45.2520 | 45.2520 | 45.2520 | 45.2520 | 45.2520 |
| 6 | 45.6497 | 45.6497 | 45.6497 | 45.6497 | 45.6497 | 45.6497 |
| 7 | 46.2343 | 46.2343 | 46.2343 | 46.2343 | 46.2343 | 46.2343 |
| 8 | 46.3837 | 46.3837 | 46.3837 | 46.3837 | 46.3837 | 46.3837 |
| 9 | 45.8429 | 45.8429 | 45.8429 | 45.8429 | 45.8429 | 45.8429 |
| 10 | 44.7897 | 44.7897 | 44.7897 | 44.7897 | 44.7897 | 44.7897 |
| 11 | 45.0676 | 45.0676 | 45.0676 | 45.0676 | 45.0676 | 45.0676 |
| 12 | 47.1691 | 47.1691 | 47.1691 | 47.1691 | 47.1691 | 47.1691 |
| 13 | 50.6375 | 50.6375 | 50.6375 | 50.6375 | 50.6375 | 50.6375 |
| 14 | 52.7500 | 52.7500 | 52.7500 | 52.7500 | 52.7500 | 52.7500 |
| 15 | 53.0500 | 53.0500 | 53.0500 | 53.0500 | 53.0500 | 53.0500 |
| 16 | 51.8433 | 51.8433 | 51.8433 | 51.8433 | 51.8433 | 51.8433 |
| 17 | 50.2735 | 50.2735 | 50.2735 | 50.2735 | 50.2735 | 50.2735 |
| 18 | 48.4506 | 48.4506 | 48.4506 | 48.4506 | 48.4506 | 48.4506 |
| 19 | 46.6383 | 46.6383 | 46.6383 | 46.6383 | 46.6383 | 46.6383 |
| 20 | 45.4802 | 45.4802 | 45.4802 | 45.4802 | 45.4802 | 45.4802 |
| 21 | 45.0058 | 45.0058 | 45.0058 | 45.0058 | 45.0058 | 45.0058 |
| 22 | 44.9091 | 44.9091 | 44.9091 | 44.9091 | 44.9091 | 44.9091 |
| 23 | 44.7026 | 44.7026 | 44.7026 | 44.7026 | 44.7026 | 44.7026 |
| 24 | 44.3800 | 44.3800 | 44.3800 | 44.3800 | 44.3800 | 44.3800 |

### 4.3.2 Iterative One-Step-Ahead Forecast

By conditioning on the set of smoothed functions $f(x)$ and the fixed functional principal components $B$, the one-step-ahead forecasts obtained using $y_{n+1}(x)$ and resulting to the values of July 2018 as in Table 4.3 and Table 4.4 and the process iterated forward for the number of periods desired; for example this study iterated forward 6 periods to get the forecasted (curves) values of month July 2018 – December 2018. And the plot of the forecasted curves as in Figure 4.11 and Figure 4.13.

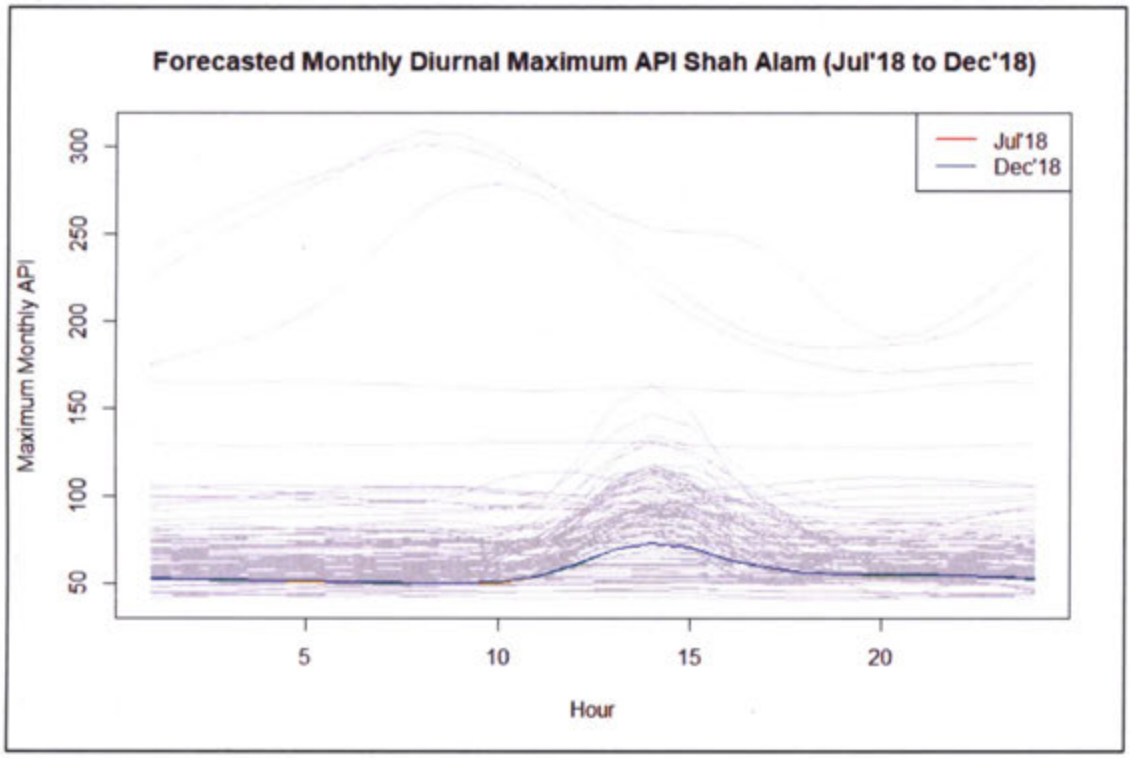


Figure 4.10 Iterative One-Step-Ahead Forecasted Monthly Diurnal Maximum API Curves for Shah Alam (Jul'18 to Dec'18)

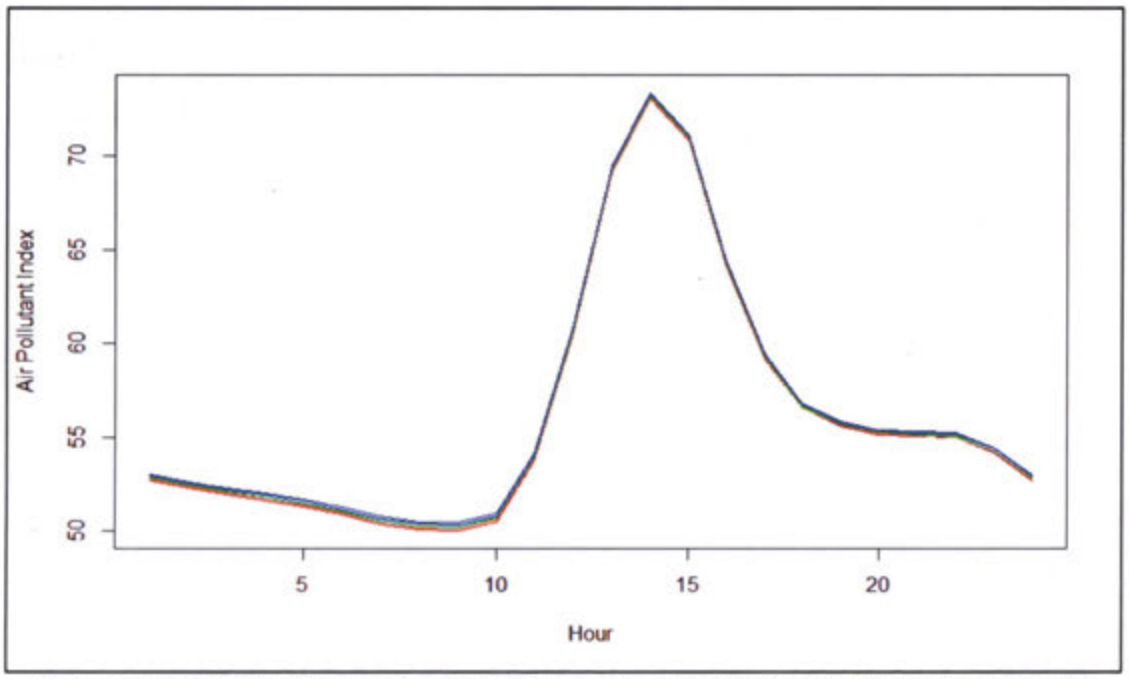

Figure 4.11 Iterative One-Step-Ahead Forecasted Monthly Diurnal Maximum API Curves for Shah Alam (Jul'18 to Dec'18)

Figure 4.10 shows the forecasts of maximum API data of Shah Alam from July 2018 to December 2018 highlighted in rainbow colour, while the data used for estimation are greyed out. The forecasted curves should exhibit rainbow colour with 6 different curves indicating a curve to a month. Fortunately, this forecast method able to exhibit multiple curves for maximum API data.

Figure 4.11 of the enlarged form of curve depicted in Figure 4.10 which shows clear visualisation of the forecasted curves July 2018 to December 2018 indicates that the maximum API achieved it peak level of about 80 at 3 p.m. in the evening then continued at the lower level till midnight. Furthermore, the result gives the information to the public that Shah Alam's air quality status for the second half of 2018 will be in a healthy state and people can freely plan their outdoor daily activities.

Table 4.3 below shows the details of the forecasted values of monthly diurnal maximum API for Shah Alam for July to December 2018 using iterative one-step ahead forecasts. These values have been depicted as in Figure 4.10 and Figure 4.11.

Table 4.3
Forecasted Values of Monthly Diurnal Maximum API for Shah Alam using Iterative One-Step-Ahead Forecasts

| Hour | Month | | | | | |
|---|---|---|---|---|---|---|
| | **Jul 2018** | **Aug 2018** | **Sep 2018** | **Oct 2018** | **Nov 2018** | **Dec 2018** |
| 1 | 52.6989 | 52.7657 | 52.8048 | 52.8870 | 52.9465 | 53.0143 |
| 2 | 52.3119 | 52.3821 | 52.4232 | 52.5096 | 52.5721 | 52.6433 |
| 3 | 51.9607 | 52.0340 | 52.0769 | 52.1671 | 52.2323 | 52.3067 |
| 4 | 51.6443 | 51.7205 | 51.7650 | 51.8586 | 51.9264 | 52.0036 |
| 5 | 51.2934 | 51.3727 | 51.4190 | 51.5164 | 51.5870 | 51.6674 |
| 6 | 50.8644 | 50.9474 | 50.9959 | 51.0980 | 51.1719 | 51.2561 |
| 7 | 50.3686 | 50.4560 | 50.5070 | 50.6145 | 50.6922 | 50.7809 |
| 8 | 50.0560 | 50.1462 | 50.1990 | 50.3099 | 50.3902 | 50.4818 |
| 9 | 50.0287 | 50.1195 | 50.1725 | 50.2841 | 50.3649 | 50.4570 |
| 10 | 50.5049 | 50.5939 | 50.6459 | 50.7553 | 50.8345 | 50.9248 |
| 11 | 53.7495 | 53.8347 | 53.8845 | 53.9893 | 54.0652 | 54.1516 |
| 12 | 60.3684 | 60.4479 | 60.4945 | 60.5923 | 60.6631 | 60.7439 |
| 13 | 69.1714 | 69.2439 | 69.2863 | 69.3755 | 69.4400 | 69.5136 |
| 14 | 73.0558 | 73.1230 | 73.1624 | 73.2451 | 73.3050 | 73.3732 |
| 15 | 70.8299 | 70.8942 | 70.9318 | 71.0109 | 71.0681 | 71.1333 |
| 16 | 64.1425 | 64.2056 | 64.2424 | 64.3200 | 64.3762 | 64.4402 |
| 17 | 59.1659 | 59.2272 | 59.2630 | 59.3383 | 59.3929 | 59.4551 |
| 18 | 56.4953 | 56.5540 | 56.5883 | 56.6605 | 56.7128 | 56.7723 |
| 19 | 55.5893 | 55.6454 | 55.6782 | 55.7471 | 55.7970 | 55.8539 |
| 20 | 55.1284 | 55.1835 | 55.2157 | 55.2834 | 55.3324 | 55.3883 |
| 21 | 55.0518 | 55.1076 | 55.1402 | 55.2089 | 55.2586 | 55.3152 |
| 22 | 54.9495 | 55.0076 | 55.0416 | 55.1131 | 55.1648 | 55.2238 |
| 23 | 54.1692 | 54.2309 | 54.2669 | 54.3428 | 54.3977 | 54.4603 |
| 24 | 52.7023 | 52.7688 | 52.8076 | 52.8894 | 52.9485 | 53.0160 |

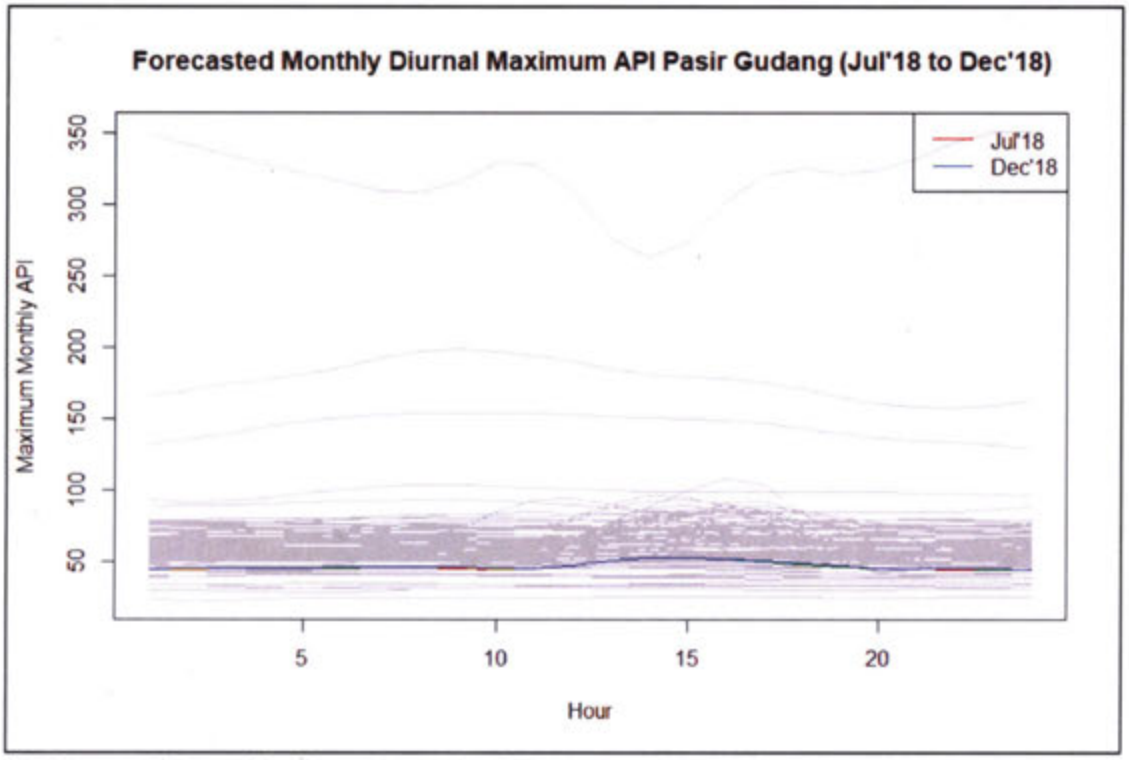


Figure 4.12 Iterative One-Step-Ahead Forecasted Monthly Diurnal Maximum API Curves for Pasir Gudang (Jul'18 to Dec'18)

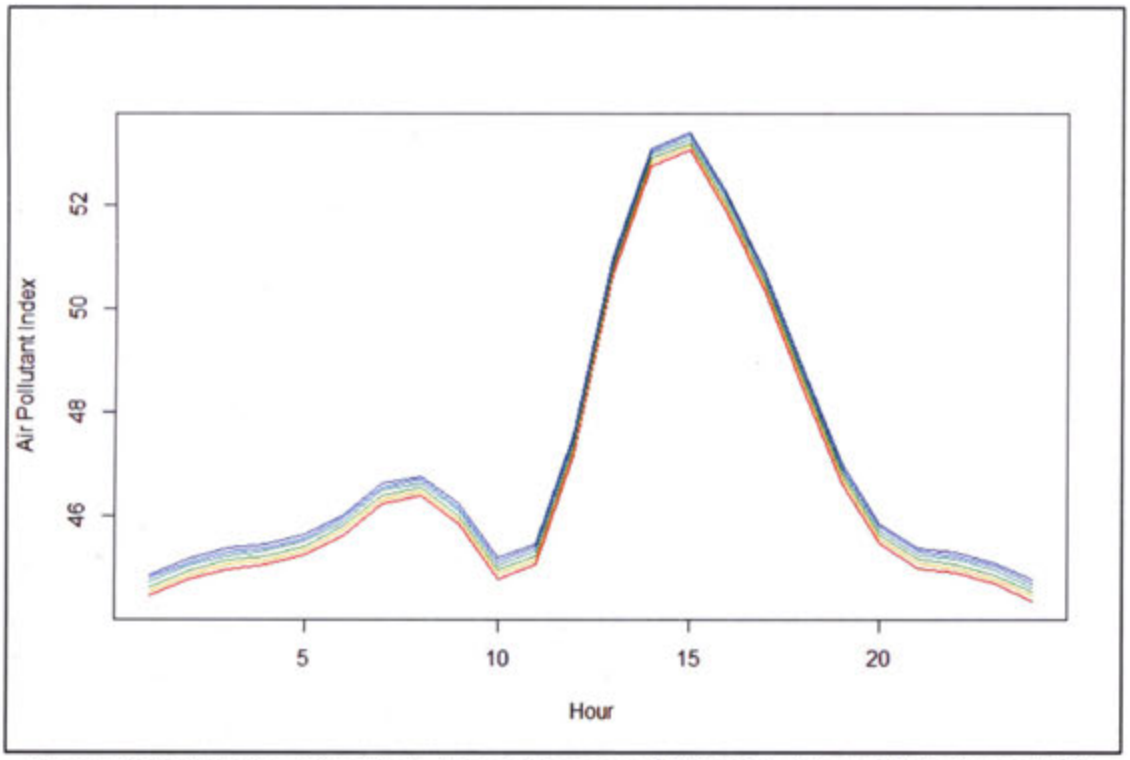


Figure 4.13 Iterative One-Step-Ahead Forecasted Monthly Diurnal Maximum API Curves for Pasir Gudang (Jul'18 to Dec'18)

Figure 4.12 shows the forecasts of maximum API data of Pasir Gudang from July 2018 to December 2018 highlighted in rainbow colour, while the data used for

estimation are greyed out. The forecasted curves should exhibit rainbow colour with 6 different curves indicating a curve to a month. Fortunately, this forecast method able to exhibit multiple curves for maximum API data. Figure 4.13 which shows clear visualisation of the forecasted curves July 2018 to December 2018 indicates that the maximum API achieved it peak level of about 53 at 3 p.m. in the evening then continued at the lower level till midnight. In addition, the result provides the public with the information that the air quality status of Pasir Gudang will be in a good state for the second half of 2018, thus people can schedule their daily activities especially outdoor daily activities freely.

Table 4.4
Forecasted Values of Monthly Diurnal Maximum API for Pasir Gudang using Iterative One-Step-Ahead Forecasts

| Hour | Month | | | | | |
|---|---|---|---|---|---|---|
| | **Jul 2018** | **Aug 2018** | **Sep 2018** | **Oct 2018** | **Nov 2018** | **Dec 2018** |
| 1 | 44.4712 | 44.5586 | 44.6429 | 44.7344 | 44.8049 | 44.8730 |
| 2 | 44.7798 | 44.8662 | 44.9496 | 45.0401 | 45.1098 | 45.1771 |
| 3 | 44.9785 | 45.0640 | 45.1466 | 45.2363 | 45.3054 | 45.3720 |
| 4 | 45.0698 | 45.1547 | 45.2367 | 45.3257 | 45.3942 | 45.4603 |
| 5 | 45.2520 | 45.3363 | 45.4176 | 45.5060 | 45.5740 | 45.6397 |
| 6 | 45.6497 | 45.7333 | 45.8140 | 45.9017 | 45.9692 | 46.0344 |
| 7 | 46.2343 | 46.3172 | 46.3973 | 46.4843 | 46.5513 | 46.6160 |
| 8 | 46.3837 | 46.4670 | 46.5475 | 46.6349 | 46.7022 | 46.7671 |
| 9 | 45.8429 | 45.9281 | 46.0102 | 46.0995 | 46.1682 | 46.2345 |
| 10 | 44.7897 | 44.8776 | 44.9625 | 45.0547 | 45.1256 | 45.1941 |
| 11 | 45.0676 | 45.1551 | 45.2395 | 45.3312 | 45.4019 | 45.4700 |
| 12 | 47.1691 | 47.2518 | 47.3316 | 47.4184 | 47.4852 | 47.5496 |
| 13 | 50.6375 | 50.7124 | 50.7846 | 50.8631 | 50.9236 | 50.9819 |
| 14 | 52.7500 | 52.8213 | 52.8901 | 52.9648 | 53.0223 | 53.0779 |
| 15 | 53.0500 | 53.1232 | 53.1938 | 53.2705 | 53.3296 | 53.3866 |
| 16 | 51.8433 | 51.9222 | 51.9984 | 52.0811 | 52.1448 | 52.2063 |
| 17 | 50.2735 | 50.3558 | 50.4353 | 50.5216 | 50.5880 | 50.6521 |
| 18 | 48.4506 | 48.5334 | 48.6133 | 48.7000 | 48.7669 | 48.8314 |
| 19 | 46.6383 | 46.7198 | 46.7985 | 46.8840 | 46.9499 | 47.0134 |
| 20 | 45.4802 | 45.5618 | 45.6407 | 45.7263 | 45.7923 | 45.8559 |
| 21 | 45.0058 | 45.0892 | 45.1696 | 45.2570 | 45.3243 | 45.3892 |
| 22 | 44.9091 | 44.9946 | 45.0772 | 45.1669 | 45.2360 | 45.3027 |
| 23 | 44.7026 | 44.7894 | 44.8733 | 44.9643 | 45.0344 | 45.1021 |
| 24 | 44.3800 | 44.4671 | 44.5513 | 44.6426 | 44.7130 | 44.7809 |

Table 4.4 above provides detail information of the forecasted values for Pasir Gudang's monthly diurnal maximum API from July to December 2018 using iterative one-step ahead forecasts. These values have been depicted as in Figure 4.12 and Figure 4.13.

In conclusion, in forecasting monthly diurnal maximum API data set, the two (2) models of multi-step-ahead and iterative one-step-ahead models have recognized that Malaysia's air pollution has varied for each location but still not in such worrying situation. However, most of the site has the common peak time around 3 p.m. at late noon. Through these results, we therefore suggest that the authorities, in particular DOE, perform further tests to determine whether the prevalent peak level has occurred due to the certain possible factors. The possible factors might be due to the hot weather around that time, as well as the air pollution emission from the transportation activities.

Furthermore, from this study we know that both models of the functional time series analysis can be used for the API data set.

### 4.3.3 Model Comparison

Table 4.5
Comparison Model Performance Shah Alam

| Errors | Multi-Step Ahead Forecast | Iterative One-Step Ahead Forecast |
|---|---|---|
| FMSE | 91.1719 | 92.7854 |
| FRMSE | 9.5484 | 9.6325 |
| FMAPE | 11.7742 | 12.1009 |

Table 4.6
Comparison of Model Performance for Pasir Gudang

| Errors | Multi-Step Ahead Forecast | Iterative One-Step Ahead Forecast |
|---|---|---|
| FMSE | 46.7837 | 48.5807 |
| FRMSE | 6.8399 | 6.9700 |
| FMAPE | 6.6192 | 6.3674 |

Two statistical models that estimate daily air quality in Peninsular Malaysia are specified using measured Air Pollutant Index (API) from two monitoring stations; Shah

Alam and Pasir Gudang during the time period 2011 – 2018. The two models are (1) multi-step ahead forecast model and (2) iterative one-step ahead forecast model.

The two models are tested by comparison with API readings observed during July – December 2018; it is concluded that the multi-step ahead model is superior in predicting the maximum API data because it has the lower errors of FMSE, FRMSE and FMAPE; 91.17, 9.55 and 11.77 respectively for Shah Alam and 46.78, 6.84 and 6.62 respectively for Pasir Gudang. The model performance comparison tables are as shown in Table 4.5 and Table 4.6 which the lower errors are highlighted in yellow.

Even though the results indicate that the multi-step-ahead is a better model compared to the iterative one-step-ahead, however the error measures are not much differed. Their extremely similar performance is due to the modelling procedures for both methods do not really differ. Both are using the same number of ten (10) basis in producing smooth curves. The only differ between those two (2); multi-step-ahead forecast and iterative one-step-ahead forecast is the multi-step-ahead times series forecasts are created using a horizon-specific forecasted model while iterative one-step-ahead forecasts are created using one-period ahead model, iterated forward for the number of period desires. Previous study by McElroy (2015) also stated the same that if the model is fixed, and the same parameters are used for the direct (multi-step-ahead forecast) and iterative forecasting formulas, then the forecasts are virtually indistinguishable.

### 4.3.4 Model Validation

During the process of model building, the researcher must be constantly concerned with how closely the model reflects the actual data. The process of determining the degree to which the model corresponds to the real data set, or at least accurately represents the model specification benchmark, is referred to as model validation. In simple settings validation could be accomplished by directly comparing model results to physical or actual measurements for the API and computing a confidence interval. Model validation is therefore vital to ensure accuracy and reliability.

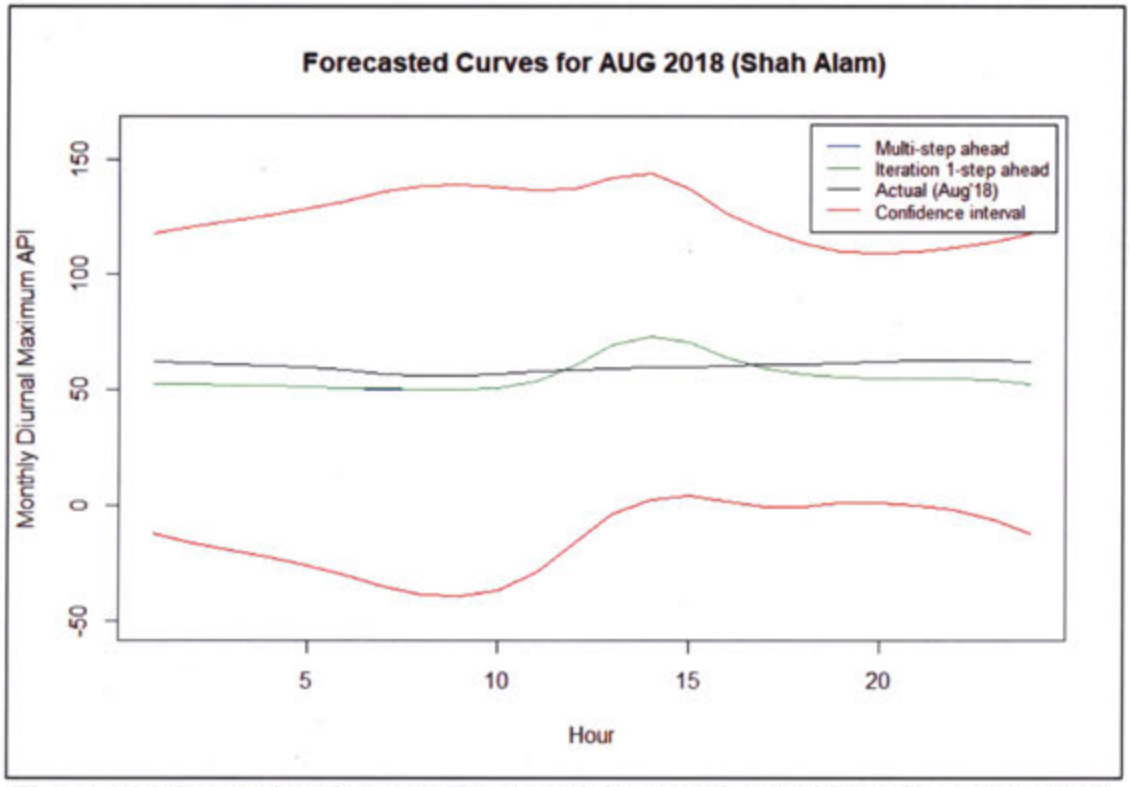


Figure 4.14 Comparison between Forecasted Curve and Actual Data for August 2018 (Shah Alam)

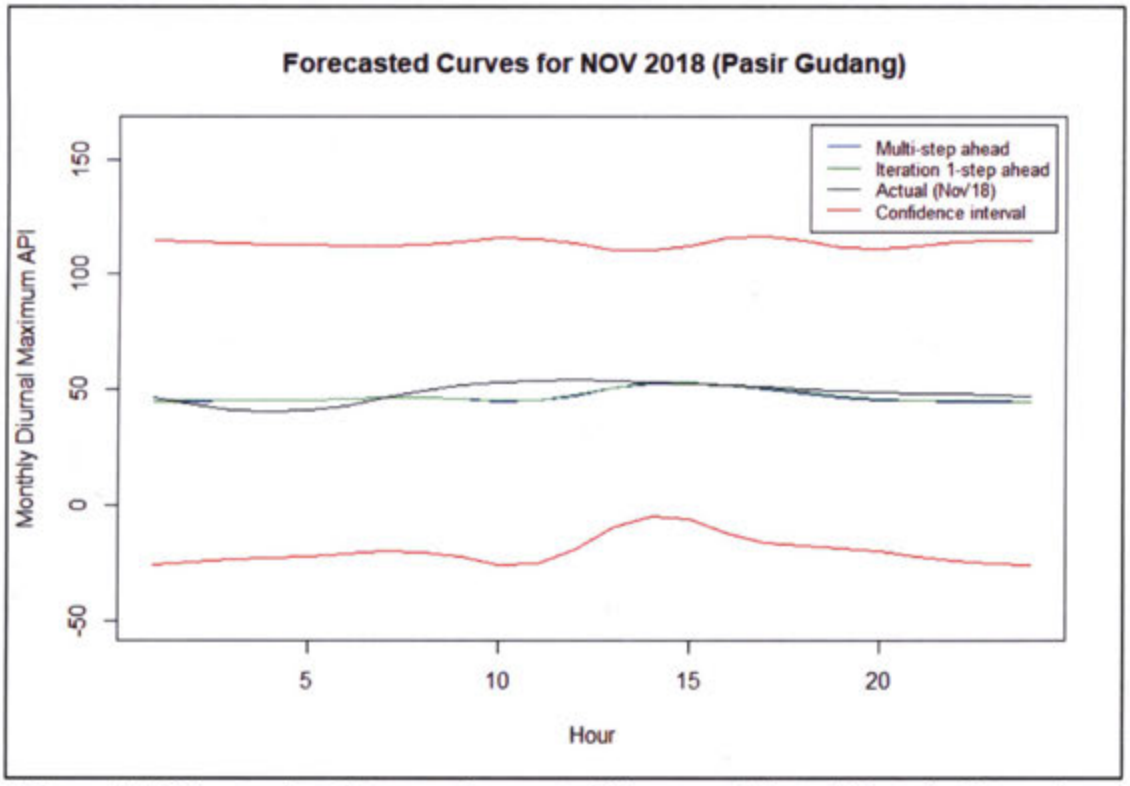


Figure 4.15 Comparison between Forecasted Curves and Actual Data for November 2018 (Pasir Gudang)

Figure 4.14 and Figure 4.15 illustrate comparison between the actual maximum API curves with the forecasted maximum API curves of multi-step-ahead forecast and iterative one-step-ahead forecast for Shah Alam and Pasir Gudang air monitoring stations respectively. Based on both figures, it can be concluded that the forecasted models for both multi-step-ahead and iterative one-step-ahead is acceptable due to the actual curves quite similar with the forecasted curves and lie between the confidence interval curves.

However, since we aim to get the best model between these two (2) models of multi-step-ahead and iterative one-step-ahead forecast, we can conclude that multi-step ahead is a better model since the model validation requirements is meet and it has lower error measurements compared to iterative one-step-ahead forecast model.

## 4.4 To Determine the Future Pattern of Monthly Diurnal Maximum API Curves at Two Hotspot Locations in Malaysia Using the Best Model

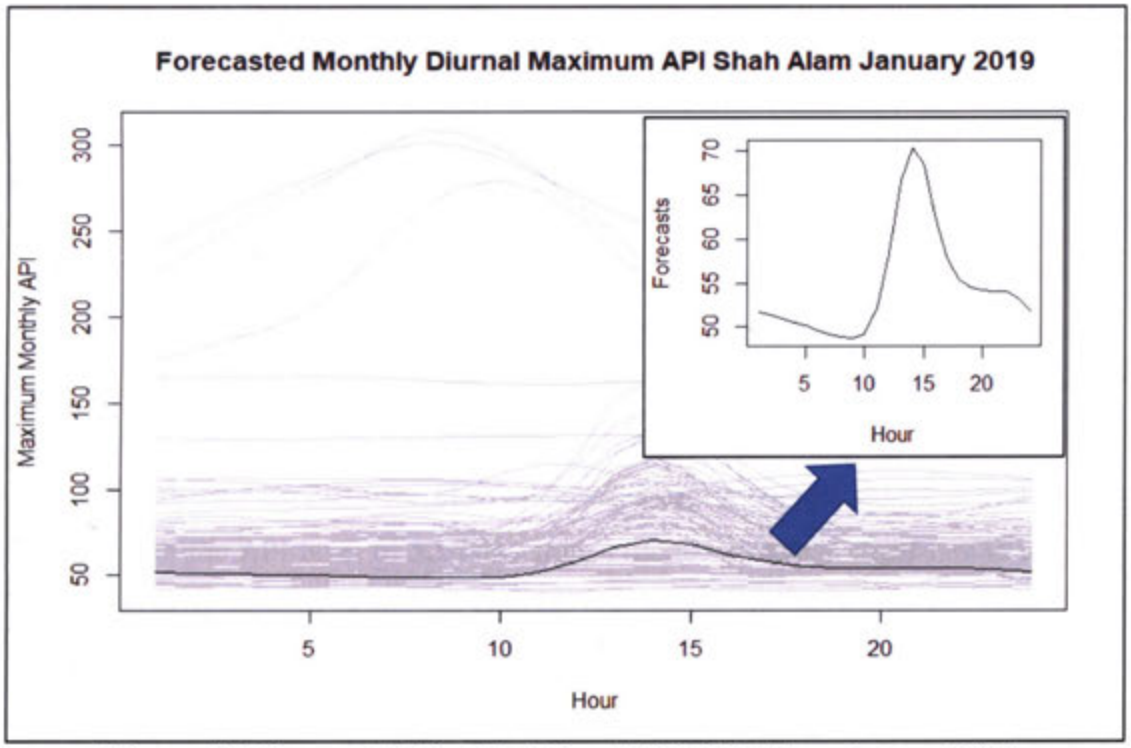


Figure 4.16 Forecasted Monthly Diurnal API Shah Alam January 2019

Figure 4.16 illustrates the forecasted diurnal maximum API curve for January 2019 for Shah Alam air monitoring station. Shah Alam is anticipated to have a moderate air quality on January 2019 because of the maximum API level expected to be below

70 throughout the entire month's 24-hour period. The highest level of API in January 2019 for Shah Alam expected to occur around 3 p.m. in the evening with the API level of 70; indicates moderate air quality status which does not pose any bad effect on public health.

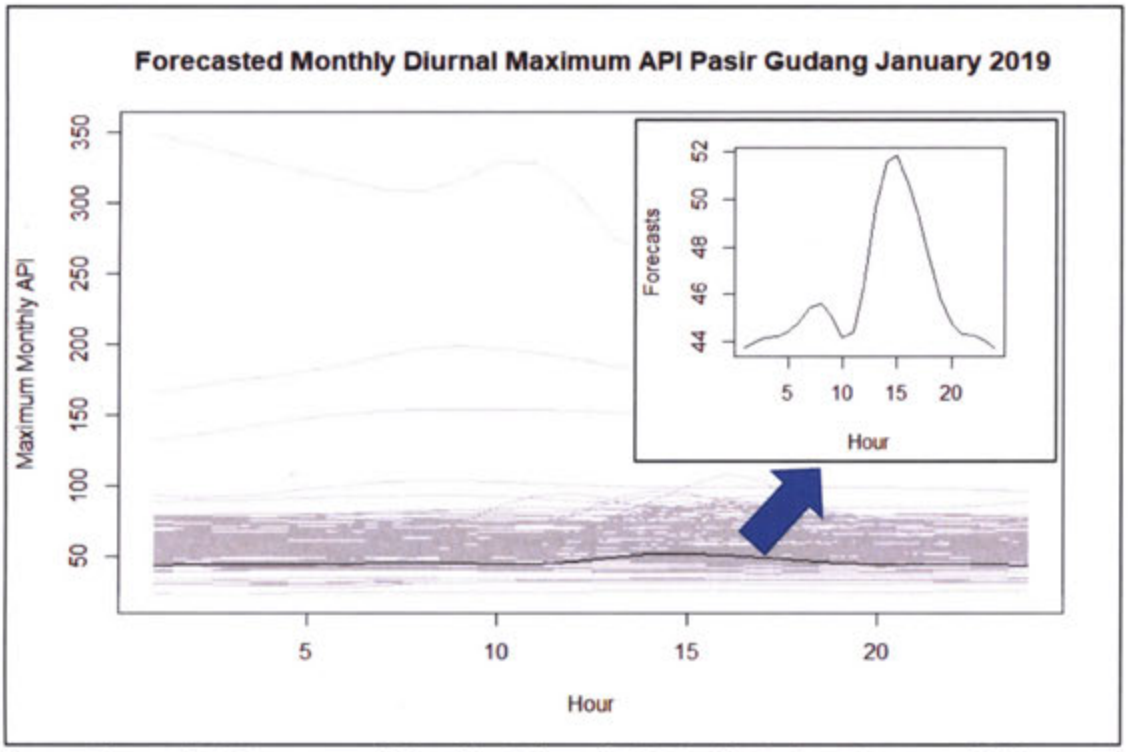


Figure 4.17 Forecasted Monthly Diurnal API Pasir Gudang January 2019

Figure 4.17 illustrates the forecasted diurnal maximum API curve for January 2019 for Pasir Gudang air monitoring station. Pasir Gudang is predicted to experience healthy or good air quality on January 2019 due to the maximum API level expected to be below 70 throughout the entire month's 24-hour period. The highest level of API during month January 2019 for Shah Alam expected to happen around 3 p.m. in the evening with the API level of 52. Note from Figure 4.17 that a slight rise in the API level around 8 a.m. to 9 a.m. could be due to the air pollution emission from transportation as this interval hour is recognized as busy vehicle motion hours as most individuals go to school, college and workplaces.

In conclusion, although the API level is not at worrisome state at both locations, the sudden rise in API readings at 3 p.m. should have a valuable information if we understand the cause of the sudden elevated readings level. The changes in the

meteorological condition and the air pollution emission from the transportation activities may be the causes of the occurrence of the sudden peak at around 15:00. It is therefore suggested that the sources of the high air pollution at certain observed time should be studied in future research study.

## 4.5 To Determine the Confidence Interval for the Forecasted Monthly Diurnal Maximum API Curves

Prediction intervals provide an upper and lower expectation for the real observation. These can be useful for assessing the range of real possible outcomes for a prediction and for better understanding the skill of the model.

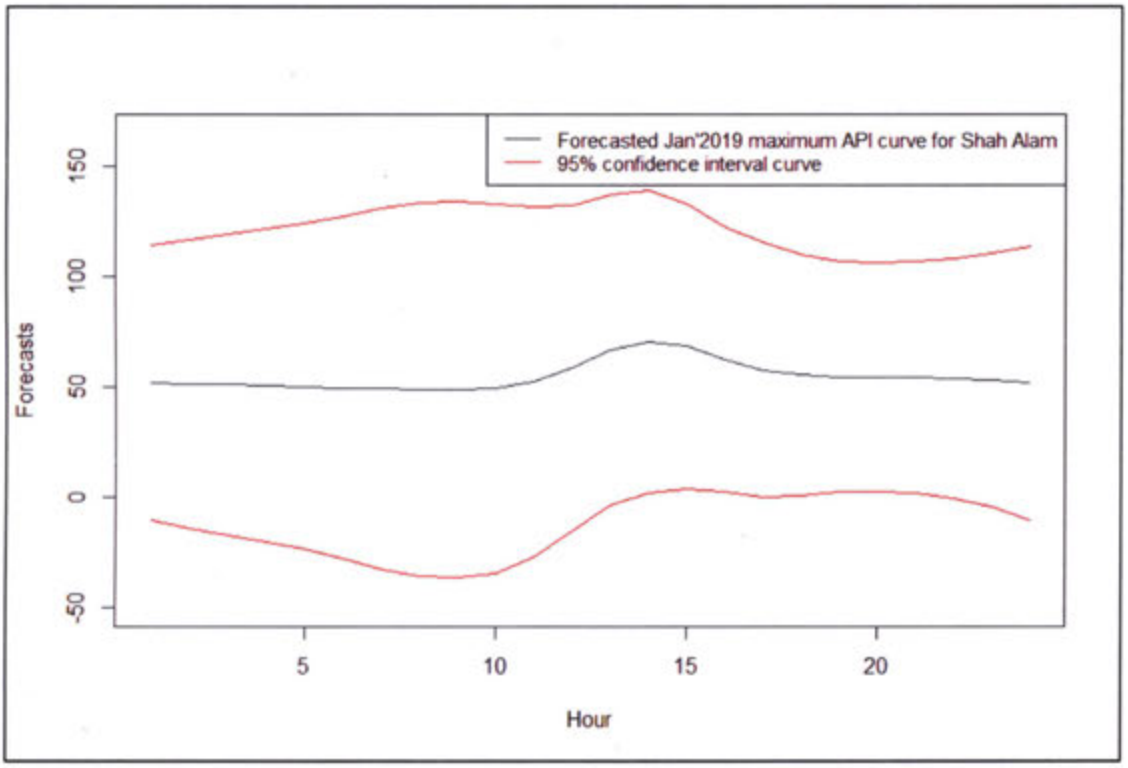


Figure 4.18 Forecasted Curve and Corresponding Confidence Interval Curves for January 2019 Shah Alam

Figure 4.18 illustrates the 95% lower and upper confidence interval curves for forecasted diurnal maximum API curve for January 2019 of Shah Alam air monitoring station. We are 95% confident that the Shah Alam's API curve for January 2019 will lie between API level of 0 to 150. Furthermore, the upper level curve shows that the maximum API is above 100 indicates unhealthy air quality status that might worsen the

health condition of high-risk people. It is advisable for these individuals to reduce outdoor operations in order to avoid the impacts of air pollution.

Table 4.7 shows the forecasted values of the lower and upper bound as per depicts in Figure 4.18. API level ranging from 0 – 50 as heathy air quality status and above 500 as emergency air quality status. Since the lower bound has negative values, as lowest API level is 0, then the negative values are regarded to be 0. Thus, although the curve is below 0 in the Figure 4.18, the lower bound values in Table 4.7 has been changed to the API range.

Table 4.7
Confidence Interval of the Forecasted Maximum API Curve for January 2019 (Shah Alam)

| Hour | Lower Bound | Forecasted Values for API Jan'19 | Upper Bound |
|---|---|---|---|
| 1 | 0.0000 | 51.7599 | 114.2107 |
| 2 | 0.0000 | 51.3425 | 116.9390 |
| 3 | 0.0000 | 50.9491 | 119.4373 |
| 4 | 0.0000 | 50.5791 | 121.6904 |
| 5 | 0.0000 | 50.1755 | 124.2081 |
| 6 | 0.0000 | 49.7025 | 127.3357 |
| 7 | 0.0000 | 49.1715 | 131.1413 |
| 8 | 0.0000 | 48.8302 | 133.7573 |
| 9 | 0.0000 | 48.7799 | 134.2906 |
| 10 | 0.0000 | 49.2218 | 133.1235 |
| 11 | 0.0000 | 52.2473 | 131.9541 |
| 12 | 0.0000 | 58.4155 | 132.3612 |
| 13 | 0.0000 | 66.6245 | 137.0752 |
| 14 | 1.6511 | 70.2968 | 138.9425 |
| 15 | 3.9254 | 68.3291 | 132.7327 |
| 16 | 2.1827 | 62.2417 | 122.3007 |
| 17 | 0.1481 | 57.7267 | 115.3053 |
| 18 | 0.5439 | 55.3328 | 110.1218 |
| 19 | 2.1739 | 54.5553 | 106.9367 |
| 20 | 2.2031 | 54.1634 | 106.1237 |
| 21 | 1.4698 | 54.1006 | 106.7313 |
| 22 | 0.0000 | 53.9864 | 108.4572 |
| 23 | 0.0000 | 53.2155 | 110.8077 |
| 24 | 0.0000 | 51.7800 | 113.9675 |

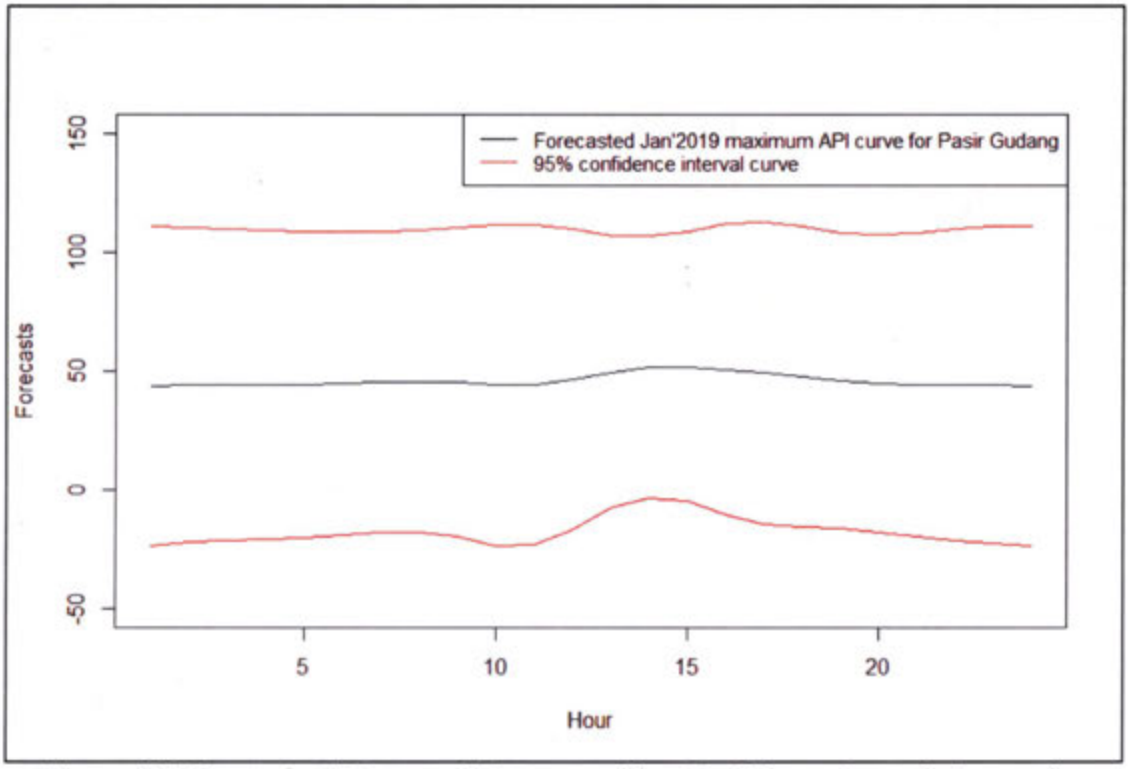


Figure 4.19 Forecasted Curve and Corresponding Confidence Interval Curves for January 2019 Pasir Gudang

Figure 4.19 illustrates the 95% lower and upper confidence interval curves for forecasted diurnal maximum API curve for January 2019 of Pasir Gudang air monitoring station. We are 95% confident that the Pasir Gudang's API curve for January 2019 will lie between API level of 0 to 120. In addition, the upper level curve shows that the maximum API for the month of January 2019 would be slightly above 100 suggests a mildly unhealthy air quality status that could worsen the individuals at high-risk.

Table 4.8 shows the forecasted values of the lower and upper bound as per depicts in Figure 4.19. API level ranging from 0 – 50 as heathy air quality status and above 500 as emergency air quality status. Since the lower bound has negative values, as lowest API level is 0, then the negative values are regarded to be 0. Thus, although the curve is below 0 in the Figure 4.19, the lower bound values in Table 4.8 has been changed to the API range.

Table 4.8
Confidence Interval of the Forecasted Maximum API Curve for January 2019 (Pasir Gudang)

| Hour | Lower Bound | Forecasted Values for API Jan'19 | Upper Bound |
|---|---|---|---|
| 1 | 0.0000 | 43.7394 | 110.9303 |
| 2 | 0.0000 | 43.9835 | 110.2119 |
| 3 | 0.0000 | 44.1442 | 109.6178 |
| 4 | 0.0000 | 44.2240 | 109.1309 |
| 5 | 0.0000 | 44.4085 | 108.7937 |
| 6 | 0.0000 | 44.8145 | 108.6150 |
| 7 | 0.0000 | 45.4141 | 108.6540 |
| 8 | 0.0000 | 45.6011 | 109.1746 |
| 9 | 0.0000 | 45.1269 | 110.2184 |
| 10 | 0.0000 | 44.1585 | 111.7709 |
| 11 | 0.0000 | 44.4262 | 111.7314 |
| 12 | 0.0000 | 46.3924 | 109.6049 |
| 13 | 0.0000 | 49.6294 | 106.9814 |
| 14 | 0.0000 | 51.5869 | 106.8525 |
| 15 | 0.0000 | 51.8371 | 108.6279 |
| 16 | 0.0000 | 50.6729 | 111.8214 |
| 17 | 0.0000 | 49.1884 | 112.6942 |
| 18 | 0.0000 | 47.4892 | 110.6579 |
| 19 | 0.0000 | 45.8149 | 107.9077 |
| 20 | 0.0000 | 44.7510 | 107.3374 |
| 21 | 0.0000 | 44.3245 | 108.3061 |
| 22 | 0.0000 | 44.2432 | 110.0112 |
| 23 | 0.0000 | 44.0420 | 110.9043 |
| 24 | 0.0000 | 43.7148 | 110.9141 |

Based on Figure 4.18 and Figure 4.19 above, it can be concluded that we are 95% confident that the API level for Shah Alam and Pasir Gudang would be between 0 to 150 and 0 to 120 respectively. It shows that both locations might have air quality status ranging from good to unhealthy. It is therefore advisable for the public to always conscious of air pollution in their day-today activities due to the possible unhealthy rate of air pollution predicted for January 2019.

# CHAPTER FIVE
# CONCLUSION AND RECOMMENDATIONS

## 5.1 Conclusion

This study highlights the application of Functional Data methods to analyse monthly diurnal API maximum with the aims to visualize, evaluate and predict the monthly day-to-day (diurnal) API level. The application of the methods is inspired by the nature of the continuous nature of air pollutants concentration in the air. In specific, this study was carried out with three main objectives; to compare the performance between multi-step-ahead and iterative one-step-ahead functional time series models in predicting monthly diurnal maximum API curves, to visualize the future pattern of monthly diurnal maximum API curves at two hot spot locations in Malaysia which include Shah Alam and Pasir Gudang using the best model and finally to determine the confidence interval for the future monthly diurnal maximum API curves.

Based on the results of the functional descriptive analysis in the preliminary investigation, the results have shown that the monthly day-to-day maximum API level within a day period, on the average is considered under healthy level (below 100) with API readings between 60 to 90 for Shah Alam and between 62 to 68 for Pasir Gudang respectively. However, the standards deviations results indicate that API variation is found high within the day period at both locations with the peak occurred between 9 to 10 am. The results have provided the insight that air pollution emission from the transportations activities can be the reason since this interval hour is identified as busy hours of vehicles movement at the two locations.

Between the two functional time series model discussed, multi-step ahead forecast model has produced the best performance with the lowest FRMSE and FMAPE; 9.55 and 11.77 respectively compared to iterative one-step ahead forecast model. The capability of multi-step model is validated by comparing the forecasted future month diurnal curves with the actual observed data. The pattern of the behaviour is quite similar. Forecasted API level also is shown can be estimated using confidence

interval for functional data that enable the predicted API level to be determined continuously over the 24-hour period of time.

In general, the application of Functional Data Analysis in this study gives several advantages. The method provides the ability to visualize, describe, evaluate and predict continuous variation of air pollutants over a continuum of time.

## 5.2 Recommendations

Throughout the years, the Malaysian Environmental Quality Report published by the Department of Environment Ministry of Natural Resources and Environment, Malaysia reported that the API level continues to be the main concern in a few hotspot locations in Malaysia such as Shah Alam and Pasir Gudang. This study able to review the monthly diurnal maximum API curve for period of 2011 – 2018 and forecasted the diurnal maximum API level for January 2019. Even though, most of the time, maximum API level recorded at a non-risk level, but some of the times the level of API varies high. It is suggested that the statistical process control investigation and analysis need to be done in order to detect the sources that cause high variation in API values.

Besides that, this study only covered for two areas which are Shah Alam and Pasir Gudang. Due to the effectiveness and advantages of the application of functional data approach in this study, thus it is suggested that other researchers extend the study using other locations and also might use other continuum of time such as daily basis (this study covered on monthly diurnal maximum API). In addition, we also suggested to add more methods pointwise to be compared with the functional methods studied. Other than that, this current functional method might be improvise by considering hybrid approach or other suitable methods for air quality prediction improvement.

Due to current pollution issues in Johor, it is suggested to add more components into API calculation. This is due to despite the worst pollution in Johor, API level still at the moderate level which does not show the actual health affected by that pollution. It is suggested that the DOE study any other components that highly contribute to air pollution to be add into the calculation of API.

## REFERENCES


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*(1).

Akyol, S., Erdogan, S., Idiz, N., Celik, S., Kaya, M., Ucar, F., Dane, S., & Akyol, O. (2014). The role of reactive oxygen species and oxidative stress in carbon monoxide toxicity: An in-depth analysis. *Redox Report*, *19*(5), 180–189. https://doi.org/10.1179/1351000214y.0000000094

Azid, A., Juahir, H., Toriman, M. E., Kamarudin, M. K. A., Saudi, A. S. M., Hasnam, C. N. C., Aziz, N. A. A., Azaman, F., Latif, M. T., Zainuddin, S. F. M., Osman, M. R., & Yamin, M. (2014). Prediction of the level of air pollution using principal component analysis and artificial neural network techniques: A case study in Malaysia. *Water, Air, and Soil Pollution*, *225*(8). https://doi.org/10.1007/s11270-014-2063-1

Banan, N., Latif, M. T., Juneng, L., & Ahamad, F. (2013). Characteristics of surface ozone concentrations at stations with different backgrounds in the Malaysian Peninsula. *Aerosol and Air Quality Research*, *13*(3), 1090–1106. https://doi.org/10.4209/aaqr.2012.09.0259

Chatfield, C. (2000). *Time-series forecasting*. Chapman and Hall/CRC.

Chen, T. M., Kuschner, W. G., Gokhale, J., & Shofer, S. (2007). Outdoor air pollution: nitrogen dioxide, sulfur dioxide, and carbon monoxide health effects. *The American journal of the medical sciences*, *333*(4), 249-256.

Chuang, K. J., Yan, Y. H., Chiu, S. Y., & Cheng, T. J. (2011). Long-term air pollution exposure and risk factors for cardiovascular diseases among the elderly in Taiwan. *Occupational and environmental medicine*, *68*(1), 64-68.

Crane E. A., Cassidy R. B., Rothman E. D., Gerstner G. E. (2010). Effect of registration on cyclical kinematic data. *J Biomech*, 43:2444–2447.

Department of Environment. (2000). *A Guide to Air Pollutant Index (API) in Malaysia*. Department of Environment: Kuala Lumpur, Malaysia.

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

Department of Environment. (2018). Introduction of PM2.5. Retrieved from http://apims.doe.gov.my/public_v2/pdf/Introduction_of_PM25.pdf

Dominick, D., Juahir, H., Latif, M. T., Zain, S. M., & Aris, A. Z. (2012). Spatial assessment of air quality patterns in Malaysia using multivariate analysis. *Atmospheric environment*, *60*, 172-181.

Erbas, B., Akram, M., Gertig, D. M., English, D., Hopper, J. L., Kavanagh, A. M., & Hyndman, R. (2010). Using Functional Data Analysis Models to Estimate Future Time Trends in Age-Specific Breast Cancer Mortality for the United States and England–Wales. *Journal of Epidemiology*, *20*(2), 159–165. https://doi.org/10.2188/jea.je20090072

Fares, S., Vargas, R., Detto, M., Goldstein, A. H., Karlik, J., Paoletti, E., & Vitale, M. (2013). Tropospheric ozone reduces carbon assimilation in trees: Estimates from

analysis of continuous flux measurements. *Global Change Biology*, *19*(8), 2427–2443. https://doi.org/10.1111/gcb.12222

Ferrante, M., Fiore, M., Conti, G. O., & Ledda, C. (2012). Air Pollution - A Comprehensive Perspective. *Air Pollution - A Comprehensive Perspective*, (August). https://doi.org/10.5772/2591

Fiévez, L., Kirschvink, N., Dogné, S., Jaspar, F., Merville, M. P., Bours, V., Lekeux, P., & Bureau, F. (2001). Impaired accumulation of granulocytes in the lung during ozone adaptation. *Free Radical Biology and Medicine*, *31*(5), 633-641.

Gao, Y., Shang, H. L., & Yang, Y. (2018). High-dimensional functional time series forecasting: An application to age-specific mortality rates. *Journal of Multivariate Analysis*.

Ghorani-Azam, A., Riahi-Zanjani, B., & Balali-Mood, M. (2016). Effects of air pollution on human health and practical measures for prevention in Iran. *Journal of research in medical sciences: the official journal of Isfahan University of Medical Sciences*, *21*.

Goldstein, M. (2008). Carbon monoxide poisoning. *Journal of Emergency Nursing*, *34*(6), 538-542.

Gorai, A. K., Tuluri, F., & Tchounwou, P. B. (2014). A GIS based approach for assessing the association between air pollution and asthma in New York State, USA. *International Journal of Environmental Research and Public Health*, *11*(5), 4845–4869. https://doi.org/10.3390/ijerph110504845

Green P. J., & Silverman B. W. (1994). *Nonparametric regression and generalized linear models. A roughness penalty approach*. London: Chapman and Hall.

Guillam, M., Pédrono, G., Bouquin, S. Le, Huneau, A., Gaudon, J., Leborgne, R., Dewitte, J., & Ségala, C. (2013). *Chronic respiratory symptoms of poultry farmers and model-based estimates of long-term dust exposure*. *20*(2), 307–311.

Hesterberg, T. W., Bunn, W. B., McClellan, R. O., Hamade, A. K., Long, C. M., & Valberg, P. A. (2009). Critical review of the human data on short-term nitrogen dioxide (NO2) exposures: evidence for NO2 no-effect levels. *Critical reviews in toxicology*, *39*(9), 743-781.

Huang, J. Z., & Shen, H. (2004). Functional coefficient regression models for non-linear time series: a polynomial spline approach. *Scandinavian journal of statistics*, *31*(4), 515-534.

Hyndman R. J., & Booth H. (2008). Stochastic population forecasts using functional data models for mortality, fertility and migration. *Int J Forecast*, 24:323–342

Hyndman R. J., & Shang H. L. (2010). Rainbow plots, bagplots, and boxplots for functional Data. *J Comput Graph Stat*, 19:29–45.

Hyndman, R. J., & Shang, H. L. (2009). Forecasting functional time series. *Journal of the Korean Statistical Society*, *38*(3), 199-211.

Hyndman, R. J., & Ullah, M. S. (2007). Robust forecasting of mortality and fertility rates: a functional data approach. *Computational Statistics & Data Analysis*, *51*(10), 4942-4956.

Isiyaka, H. A., & Azid, A. (2015). Air quality pattern assessment in Malaysia using multivariate techniques. *Malaysian Journal of Analytical Sciences*, *19*(5), 966-978.

Johns, D. O., & Linn, W. S. (2011). A review of controlled human SO2 exposure studies contributing to the US EPA integrated science assessment for sulfur oxides. *Inhalation Toxicology*, *23*(1), 33-43.

Kokoszka, P., Miao, H., & Zhang, X. (2012). Functional dynamic factor model for Intraday price curves. *Journal of Financial Econometrics*, *13*(2), 456-477.

Kokoszka, P., & Reimherr, M. (2017). *Introduction to Functional Data Analysis*. CRC Press.

Laukaitis A. (2008). Functional data analysis for cash flow and transactions intensity continuous-time prediction using Hilbert-valued autoregressive processes. *Eur J Oper Res*, 185:1607–1614.

Lee K. L., Meyer R. J., & Bradlow E. T. (2009). Analyzing risk response dynamics on the web: the case of Hurricane Katrina. *Risk Anal*, *29*:1779–1792.

Lee, M. H., Rahman, N. H. A., Suhartono, Latif, M. T., Nor, M. E., & Kamisan, N. A. B. (2012). Seasonal ARIMA for forecasting air pollution index: A case study. *American Journal of Applied Sciences*, *9*(4), 570-578. https://doi.org/10.3844/ajassp.2012.570.578

Mabahwi, N. A. B., Leh, O. L. H., & Omar, D. (2014). Human Health and Wellbeing: Human Health Effect of Air Pollution. *Procedia - Social and Behavioral Sciences*, *153*, 221–229. https://doi.org/10.1016/j.sbspro.2014.10.056

Mabahwi, N. A., Leh, O. L. H., & Omar, D. (2015). Urban air quality and human health effects in Selangor, Malaysia. *Procedia-Social and Behavioral Sciences*, 170, 282-291.

Mahajan, S. P. (2009). *Air pollution control*. The Energy and Resources Institute (TERI).

McElroy, T. (2015). When are direct multi-step and iterative forecasts identical? *Journal of Forecasting*, *34*(4), 315–336. https://doi.org/10.1002/for.2321

Mutalib, S. N. S. A., Juahir, H., Azid, A., Mohd Sharif, S., Latif, M. T., Aris, A. Z., … Dominick, D. (2013). Spatial and temporal air quality pattern recognition using environmetric techniques: A case study in Malaysia. *Environmental Sciences: Processes and Impacts*, *15*(9), 1717–1728. https://doi.org/10.1039/c3em00161j

Ott W. R., & Hunt W. F. A. (1976). Quantitative Evaluation of the Pollutant Standards Index. *Journal of the Air Pollution Control Association*, *26* (11), 1050.

Rahman, N. H. A., Lee, M. H., & Latif, M. T. (2013). Forecasting of air pollution index with artificial neural network. *Jurnal Teknologi (Sciences and Engineering)*, *63*(2), 59-64.

Rahman, H. A. (2016). Air Pollution in Urban Areas and Health Effects. *International Journal of the Malay World and Civilisation*, 25 - 33.

Ramsay, J. O., & Dalzell, C. J. (1991). Some tools for functional data analysis. *Journal*

*of the Royal Statistical Society. Series B (Methodological)*, 539-572.

Ramsay, J., Hooker, G., & Graves, S. (2009). *Functional data analysis with R and MATLAB*. Springer Science & Business Media.

Ramsay, J. O., & Silverman, B. W. (2006). *Functional data analysis (second ed.)*, New York, Springer.

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

Sahu, D., Kannan, G. M., & Vijayaraghavan, R. (2014). Carbon black particle exhibits size dependent toxicity in human monocytes. *International journal of inflammation*, *2014*.

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.

Shang, H. L., & Hyndman, R. J. (2011). Nonparametric time series forecasting with dynamic updating. *Mathematics and Computers in Simulation*, 81(7):1310–1324.

Sharma, S. B., Jain, S., Khirwadkar, P., & Kulkarni, S. (2013). The effects of air pollution on the environment and human health. *Indian Journal of Research in Pharmacy and Biotechnology*, *1*(3), 2320–3471. Retrieved from https://pdfs.semanticscholar.org/a2ab/90fda60b29ef2478dc3b6633c06ae79fb3d2.pdf

Siew, L. Y., Chin, L. Y., & Wee, P. M. J. (2008). ARIMA and integrated ARFIMA models for forecasting air pollution index in Shah Alam, Selangor. *Malaysian Journal of Analytical Sciences*, *12*(1), 257-263.

Ullah, S., & Finch, C. F. (2013). Applications of functional data analysis: A systematic review. *BMC medical research methodology*, *13*(1), 43.

Wilkinson, S., Mills, G., Illidge, R., & Davies, W. J. (2012). How is ozone pollution reducing our food supply? *Journal of Experimental Botany*, *63*(2), 527–536. https://doi.org/10.1093/jxb/err317

Zhou, J., Ito, K., Lall, R., Lippmann, M., & Thurston, G. (2011). Time-series analysis of mortality effects of fine particulate matter components in Detroit and seattle. *Environmental Health Perspectives*, *119*(4), 461–466. https://doi.org/10.1289/ehp.1002613