# PREDICTION MODEL OF $PM_{10}$ BY USING MACHINE LEARNING

NURUL LATIFFAH ABD. RANI

DOCTOR OF PHILOSOPHY
UNIVERSITI SULTAN ZAINAL ABIDIN

2019

# PREDICTION MODEL OF $PM_{10}$ BY USING MACHINE LEARNING

**NURUL LATIFFAH ABD RANI**

**Thesis Submitted in Fulfilment of the Requirement for the Degree of Doctor of Philosophy in the Faculty of Bioresources and Food Industry Universiti Sultan Zainal Abidin**

**2019**

## MODEL RAMALAN $PM_{10}$ MENGGUNAKAN PEMBELAJARAN MESIN


## ABSTRAK

Pepejal terampai ($PM_{10}$) adalah petunjuk utama indeks pencemaran udara (IPU) di Malaysia. Meskipun kepekatan $PM_{10}$ menunjukkan kepentingan dalam menentukan tahap IPU di Malaysia, terdapat beberapa data kepekatan $PM_{10}$ yang gagal dianalisis telah dikesan pada hari tertentu disebabkan oleh kegagalan mesin pencerap untuk beroperasi. Oleh itu, percubaan bagi membangunkan model ramalan $PM_{10}$ yang sahih dan munasabah telah dihasilkan. Parameter meteorologi (kelajuan angin, arah angin, suhu, kelembapan) dan elemen pencemar ($NO_x$, NO, $SO_2$, $NO_2$, CO, $O_3$) yang digunakan dalam kajian ini diperolehi daripada 52 stesen pengawasan kualiti udara di seluruh Malaysia yang bermula dari Januari 2010 sehingga Disember 2015. Dalam kajian ini, model Rangkaian Neural Buatan (ANN) dan Regresi Linear Berganda (MLR) dengan input yang diperolehi daripada Analisis Komponen Utama (PCA) berkorelasi tinggi (> 0.75) digunakan untuk meramal kepekatan $PM_{10}$ berdasarkan tiga klasifikasi kumpulan oleh kumpulan kelompok hirarki (AHC), iaitu kawasan mempunyai pencemaran tinggi (HPR), pencemaran sederhana (MPR) dan pencemaran rendah (LPR) di Malaysia. Parameter input yang diperolehi apabila data kepekatan $PM_{10}$ pada hari sebelumnya tidak dipilih bagi HPR, MPR, LPR yang diperolehi melalui PCA masing-masing ialah $NO_x$, NO, $NO_2$, CO, $O_3$, suhu; $NO_x$, NO, $O_3$, arah angin dan $NO_x$, NO, $NO_2$, suhu. Nilai $R^2$ dan RMSE yang diperoleh daripada parameter input ini masing-masing ialah 0.4736, 24.14; 0.1851, 23.63 dan 0.0760, 19.66 bagi HPR, MPR dan LPR melalui ANN. Sementara itu, nilai $R^2$ dan RMSE masing-masing ialah 0.3407, 27.02; 0.1332, 23.98 dan 0.0569, 20.05 bagi HPR, MPR dan LPR melalui MLR. Analisis menunjukkan bahawa model ANN memberikan ramalan yang lebih baik berbanding dengan MLR. Walau bagaimanapun, kedua-dua model menunjukkan ketepatan kurang dengan memberikan nilai $R^2$ yang rendah dan nilai RMSE yang tinggi di kalangan kawasan. Percubaan telah dibuat bagi meningkatkan prestasi model dengan memasukkan data kepekatan $PM_{10}$ pada hari sebelumnya. Parameter input yang diperolehi termasuk data kepekatan $PM_{10}$ pada hari sebelumnya bagi HPR, MPR, LPR melalui PCA masing-masing ialah $NO_x$, NO, $NO_2$, suhu, $O_3$, $PM_{10}$; $NO_x$, NO, suhu, arah angin, $PM_{10}$ dan $NO_x$, NO, $NO_2$, suhu, $PM_{10}$. Nilai $R^2$ dan RMSE yang diperoleh daripada parameter input ini masing-masing ialah 0.7718, 16.11; 0.7590, 12.63 dan 0.7721, 9.86 bagi HPR, MPR dan LPR melalui ANN. Sementara itu, nilai $R^2$ dan RMSE yang diperoleh masing-masing ialah 0.7609, 16.27; 0.7442, 13.03 dan 0.7642, 10.02 bagi HPR, MPR dan LPR melalui MLR. Kedua-dua model menunjukkan ketepatan yang meningkat apabila data kepekatan $PM_{10}$ pada hari sebelumnya digunakan. Perbandingan antara keputusan yang diperoleh menunjukkan bahawa model ANN masih memberikan ramalan lebih baik daripada MLR bagi setiap kawasan. Sebagai kesimpulan, ANN memberikan model ramalan $PM_{10}$ yang sahih dan munasabah berbanding dengan MLR apabila data kepekatan $PM_{10}$ pada hari sebelumnya digunakan sebagai parameter input. Walau bagaimanapun, kedua-dua model ramalan (ANN dan MLR) menunjukkan data yang terbaik apabila data kepekatan $PM_{10}$ sebelumnya digunakan sebagai input dengan masing-masing memberikan prestasi tinggi dan rendah bagi $R^2$ dan RMSE.

## PREDICTION MODEL OF $PM_{10}$ BY USING MACHINE LEARNING


### ABSTRACT

Particulate matter ($PM_{10}$) is the key indicator of air pollution index (API) in Malaysia. Regardless of its importance in determining the API level in Malaysia, absence of some $PM_{10}$ concentrations is noticed for certain day possibly because of equipment failure. Therefore, an attempt to develop an accurate $PM_{10}$ prediction model was made. This study used parameters, such as meteorological (wind speed, wind direction, temperature, humidity) and air pollutants ($NO_x$, NO, $SO_2$, $NO_2$, CO, $O_3$) data obtained from 52 continuous air quality monitoring stations in Malaysia, beginning from January 2010 until December 2015. In this study, artificial neural network (ANN), multiple linear regression (MLR) models and input obtained from the principal component analysis (PCA) with high correlation (>0.75) were utilised to predict the $PM_{10}$ concentration based on three classification groups by agglomerative hierarchical cluster (AHC), namely high polluted region (HPR), moderate polluted region (MPR) and low polluted region (LPR) in Malaysia. The input parameters obtained with excluded lagged $PM_{10}$ concentration data for HPR, MPR, LPR through PCA were $NO_x$, NO, $NO_2$, CO, $O_3$, temperature; $NO_x$, NO, $O_3$, wind direction and $NO_x$, NO, $NO_2$, temperature, respectively. The values of $R^2$ and RMSE obtained from ANN by using these input parameters were 0.4736, 24.14; 0.1851, 23.63 and 0.0760, 19.66 for HPR, MPR and LPR, respectively. Meanwhile, the values of $R^2$ and RMSE obtained from MLR by using these input parameters were 0.3407, 27.02; 0.1332, 23.98 and 0.0569, 20.05 for HPR, MPR and LPR, respectively. This analysis showed that the ANN model gave a better prediction as compared to the MLR. However, both models were less accurate by giving low $R^2$ and high RMSE values among the regions. An attempt was made to enhance the model performance by including the lagged $PM_{10}$ concentration data. The input parameters obtained with included lagged $PM_{10}$ concentration data for HPR, MPR, LPR through PCA were $NO_x$, NO, $NO_2$, temperature, $O_3$, $PM_{10}$; $NO_x$, NO, temperature, wind direction, $PM_{10}$ and $NO_x$, NO, $NO_2$, temperature, $PM_{10}$, respectively. The values of $R^2$ and RMSE obtained from ANN by using these input parameters were 0.7718, 16.11; 0.7590, 12.63 and 0.7721, 9.86 for HPR, MPR and LPR, respectively. Meanwhile, the values of $R^2$ and RMSE obtained from MLR by using these input parameters were 0.7609, 16.27; 0.7442, 13.03 and 0.7642, 10.02 for HPR, MPR and LPR, respectively. Both models showed an accuracy enhancement when the lagged $PM_{10}$ concentration data were included as one of the input data. By comparing the results, the ANN model still gave better prediction than the MLR for each region. In conclusion, ANN is a more accurate $PM_{10}$ concentration prediction model as compared to MLR. However, both prediction models gave the best when lagged $PM_{10}$ concentration data were applied as inputs, and gave higher and lower $R^2$ and RMSE performance, respectively.

## ACKNOWLEDGEMENT

First and foremost, I owe my deepest gratitude to Allah S.W.T for His blessings throughout the course of my study. I am sincerely thankful to my supervisor, Dr. Azman Azid and my co-supervisor, Dr. Saiful Iskandar Khalit whose encouragement, guidance and support have enabled me to complete this thesis successfully.

My gratitude also goes to Universiti Sultan Zainal Abidin (UniSZA) and the Ministry of Higher Educational Malaysia (MOHE) for their financial support through MyPHD. A special gratitude goes to the Air Quality Division of the Department of Environment (DOE), Ministry of Natural Resource and Environment of Malaysia for providing useful air quality data set for this study, besides helping in a number of ways.

Last but not least, I would like to express my gratitude to my husband Mohamad Tarmizi bin Abu Bakar, my son Muhammad Syahrul Faiz and also my family members as well as colleagues who gave their never-ending help and support throughout my course of study.

## APPROVAL

I certify that an Examination Committee has met on _____ to conduct the final examination of Nurul Latiffah Binti Abd Rani on her thesis entitled 'Prediction Model of $PM_{10}$ by using Machine Learning' in accordance with the regulations approved by the Senate of Universiti Sultan Zainal Abidin. The Committee recommends that the candidate be awarded the relevant degree, and it has been accepted by the Senate of Universiti Sultan Zainal Abidin as fulfilment of the requirements for the Doctor of Philosophy. The members of the Examination Committee are as follows:

**Tengku Mohammad Ariff Bin Raja Hussin, PhD**
Professor
Faculty of Medicine
Universiti Sultan Zainal Abidin
(Chairperson)

**Khamsah Suryati Mohd, PhD**
Associate Professor
Faculty of Bioresources and Food Industry
Universiti Sultan Zainal Abidin
(Internal Examiner)

**Sharifuddin M. Zain, PhD**
Professor
Department of Chemistry
Faculty of Science
University of Malaya
(External Examiner)

_____
**AHMAD PUAD BIN MAT SOM, PhD**
Professor/Dean of Graduate School
Universiti Sultan Zainal Abidin

Date:

## DECLARATION BY CANDIDATE

I hereby declare that the thesis is based on my original work except for quotations and citations, which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at Universiti Sultan Zainal Abidin or other institutions.

NURUL LATIFFAH ABD RANI

Date: 17/10/2019

## DECLARATION BY THE SUPERVISORS

This is to confirm that:
The research conducted and the writing of this thesis was under our supervision.

Signature:_____
Name of Main Supervisor: Dr Azman Bin Azid
Faculty: Faculty of Bioresources and Food Industry

Signature:_____
Name of Co-Supervisor 1: Dr Saiful Iskandar Bin Khalit
Faculty: Faculty of Bioresources and Food Industry

## TABLE OF CONTENTS

## LIST OF TABLES

## LIST OF FIGURES

## LIST OF ABBREVIATIONS

| | |
|---|---|
| AHC | Agglomerative Hierarchical Cluster |
| ANN | Artificial Neural Network |
| API | Air Pollution Index |
| ASMA | Alam Sekitar Malaysia |
| % | Percentage |
| ° | Degree |
| °C | Degree Celcius |
| CO | Carbon Monoxide |
| $CH_4$ | Methane |
| DA | Discriminant Analysis |
| DOA | Department of Agriculture |
| DOE | Department of Environment |
| DOS | Department of Statistics |
| E | Efficiency |
| EEA | European Environment Agency |
| EMB | Expectation-Maximization Based (EMB) Algorithm |
| HC | Hydrocarbon |
| HCN | Hydrogen cyanide |
| HPR | High Polluted Region |
| IA | Index of Agreement |
| IARC | International Agency for Research on Cancer |
| IT | Interim target |
| km/h | kilometre per hour |
| KMO | Kaiser-Meyer-Olkin |
| L/min | Liter per minute |

| | |
|---|---|
| LPR | Low Polluted Region |
| MAQI | Malaysian Air Quality Index |
| MAQS | Malaysia Ambient Air Quality Standard |
| MAR | Missing at Random |
| MBPJ | Majlis Bandaraya Petaling Jaya |
| MCAR | Missing Completely at Random |
| MNAR | Missing Not at Random |
| MLP-FF-ANN | Multilayer perceptron feed-forward artificial neural network |
| MLR | Multiple Linear Regression |
| MPR | Moderate Polluted Region |
| $\mu g/m^3$ | Microgram per cubic meter |
| $\mu m$ | micrometres |
| NEM | Northeast Monsoon |
| $NH_3$ | Ammonia |
| $NH_4$ | Methane |
| NMHC | Non-methane Hydrocarbon |
| NO | Nitrogen Monoxide |
| $NO_2$ | Nitrogen Dioxide |
| $NO_x$ | Nitrogen Oxide |
| $O_3$ | Ozone |
| PCA | Principal Component Analysis |
| $PM_{10}$ | Particulate matter with a diameter less than or equal to 10 micrometres ($\mu m$) |
| ppb | Part per billion |
| ppm | Part per million |
| PSI | Pollutant Standard Index |
| $R^2$ | Coefficient of Determination |

| | |
|---|---|
| RMAQG | Recommended Malaysian Air Quality Guideline |
| RMSE | Root Mean Square Error |
| $SO_2$ | Sulphur Dioxide |
| SWM | Southwest Monsoon |
| THC | Total Hydrocarbon |
| USEPA | United States Environmental Protection Agency |
| UVB | Ultraviolet-b |
| WD | Wind Direction |
| $W/m^2$ | Watt per square metre |
| WS | Wind Speed |
| WHO | World Health Organisation |

# CHAPTER 1

# INTRODUCTION

## 1.1 Introduction

This chapter discusses the overview of air pollution scenario that happened in the world, especially in Malaysia. A description of the prediction of air quality will also be mentioned. Besides, the study problems, aim of study, research questions, scopes and study significance as well as the study outline with theoretical framework are also presented.

## 1.2 Background of the study

Life on Earth includes humans, plants, and animals as well as a general ecosystem regulation that needs a complex natural gas which is represented by the atmosphere. However, pollutants in air consist of solids, liquids or gaseous materials that are obtained from the stationary or mobile sources which can deteriorate atmospheric air quality. These pollutants may experience chemical or physical transformations before they are deposited into the surface. The transformation involves interactions with other molecules in the atmosphere, such as ozone ($O_3$) photochemical formation from hydrocarbons (HC). It also involves interactions between molecules in the atmosphere and water, vegetation, surfaces and animals, for instance, acid rain formation from sulphur dioxide ($SO_2$), direct effect of $O_3$, staining of buildings by particles and respiratory damage by acidified aerosol.

In general, anthropogenic is a term used to indicate pollution created by human activities, while biogenic is the term used to indicate pollution that is caused by natural

events, such as dust storms and volcanic eruptions. According to Ahmad Isiyaka et al. (2014), anthropogenic is the main contributor to air pollution.

In Malaysia, there are five basic pollutants, namely carbon monoxide (CO), ozone ($O_3$), particulate matter ($PM_{10}$), sulphur dioxide ($SO_2$) and nitrogen dioxide ($NO_2$) that are intended for atmospheric air quality determination (DOE, 2015a). Generally, pollution sources can be divided into stationary, mobile and trans-boundary. Stationary sources comprise dust emission from urban construction and quarries works, open burning and power plants, while mobile sources comprise emissions from the motor vehicles. However, transboundary pollution is air pollution that is transported from the neighbouring countries, for instance, forest fire and volcanic eruptions.

Transboundary pollution sources are faced by Southeast Asia countries, such as southern Thailand, Brunei, Singapore and Malaysia. These countries were affected by uncontrolled burning for land clearing by farmers and plantation owners in Indonesia sometime in 2006. In addition, the El-Niño phenomenon had also generated the transboundary pollutants from Sumatra and Kalimantan in Indonesia, which were indirectly carried by extreme dry weather conditions. The 1997 haze scenario was the worst year experienced by Malaysia (Mutalib et al., 2013). The episode of haze has become a yearly event faced by Asia since 2015. Moreover, another haze scenario that occurred caused around 100, 300 deaths across Indonesia, Malaysia and Singapore (Koplitz et al., 2016).

In general, the air quality in Asia is worsened by rapid urban growth and from local and regional air pollution (Beer et al., 2018). Therefore, Malaysia being geographically situated at the south east Asia centre has indirectly experienced this situation. Azmi et

al. (2010) mentioned that severe air quality problems occur in highly urbanised areas in regard to the dust fall-out, suspended PM and Pb (lead) in ambient air along jammed roads, which are mainly caused by motor vehicle emissions and industrial activities (Zizi, 2018).

In urban areas, transportation and fuel combustion at stationary sources (industrial activities) contribute to urban air pollution (Mohtar et al., 2018). To achieve the industrial country status by 2020, Malaysia has experienced industrial pollution problems and urban environments degradation from the time the country set a target for its rapid economic growth. This is supported by the results obtained by Mutalib et al. (2013) who found that major air pollution sources are from fossil fuel combustion by motor vehicles and industrial activities. In addition, Amran et al. (2015) found that in Malaysia, development is a major contributor of local or regional air pollution.

According to a study by Zhang and Cao (2015), the highly populous urban areas are faced with a great number of pollution episodes. Jin et al. (2016) revealed that long-term pollution events reveal high pollution intensity and extensive geographical coverage. As a consequence, this will worsen the air quality as well as human health. Meanwhile, in many developing countries, air pollution is one of the most serious environmental concerns (Kanada et al., 2013). Chen et al. (2016) mentioned that in urban areas, pollutants, such as $PM_{10}$ (particulate matter with aerodynamic diameter of less than or equal to 10 μm), $O_3$, $SO_2$, $NO_2$, and CO are some of the major pollutants.

In addition, change in global climate complicates the situation, resulting in frequent air pollution episodes driven by weather actions (Mohtar et al., 2018). Climate change and pollutions are mainly major environmental problems, whereby air pollution problems

are defined as scales of multiple spatial and temporal, together with complex chemical and physical mechanisms. Human activities and highly non-linear condition of air may worsen the problem.

Worldwide, many urban areas are miserably facing stalled air quality. Poor dispersal and flow of air allow pollution to accumulate and cause smog in cities. During such periods, susceptible individuals, such as those with asthma or heart disease may face acute exacerbation, which demands hospital admission and increased medication. The 'invisible killer' comprises $NO_x$, $O_3$, and PM. PM is a common air pollutant that gives health effects.

The major sources in urban areas that come from fossil fuel combustion are mainly from road transport, power stations and factories. Meanwhile, sources in rural and semi-urban areas come from burning of biomass fuels through open fires or traditional stoves, which generate indoor PM concentration that surpass outdoor concentration that is safely deliberated (Kelly & Fussell, 2015). Air pollution raises hazard to the human health and natural environment, which can be activated by activities from the factory, power plants, motor vehicles, wildfires and windblown dust.

Sources of air pollution may be varied depending on regions. Therefore, chemometric techniques, such as Hierarchical Agglomerative Cluster Analysis (HACA), Principal Component Analysis (PCA), Discriminant analysis (DA) and $PM_{10}$ prediction models viz. MLR and ANN will be applied to identify the most significant pollutants that contribute to air pollution as well as to develop a $PM_{10}$ prediction model. This can help to reduce possible effects to human health and the environment by the presence of $PM_{10}$ as one of the main air pollutants.

## 1.3 Problem statement

Air pollution, which is mainly due to natural activities, such as volcano eruptions open burning and industrial processes, is the world's most predominant pollution type (Afroz et al., 2003). Urbanisation and industrialisation in most developed and developing countries also increase the air pollution concentration, which mostly results in indirect serious health impact. According to Zheng et al. (2014), the biggest contributor to air pollution is from human activities. Therefore, population density plays an important role as a contributor to environmental conditions. In urban areas, fuel burning from vehicles is the main reason for the increase in air pollution. Besides affecting human health, this situation also affects environmental conditions by increasing global warming due to the greenhouse effect, which in the long term can worsen the dangers to Earth (Rahman et al., 2016).

Pollutants emitted into the atmosphere will result in increasing level of air pollution index (API) (Yunus & Hasan, 2017). In general, in Malaysia, the major air pollutants applied to determine the air quality status for API consist of $NO_2$, CO, $SO_2$, $O_3$ and $PM_{10}$. Basically, API was produced in view of the API presented by the United State Environmental Protection Agency (USEPA). Sub-indexes for the five main pollutants are calculated with the highest value among these chosen sub-indexes to obtain the API value (Rahman et al., 2016). According to DOE (2014), the API value in Malaysia is frequently influenced by the $PM_{10}$ concentration because the concentration of $PM_{10}$ is always higher than other pollutants. High $PM_{10}$ levels in the atmosphere were recorded to significantly consort with haze, which has become an annual problem in the country since the 1980s. In Malaysia, there were many haze episodes that happened in 1994, 1997, 2005, 2006, 2010, 2011, 2012, 2013, 2014 and 2015 (DOE, $2015_b$).

High API was assigned to the high $PM_{10}$ concentrations, which were the main air pollutant and have become the key indicator of API in Malaysia. Trend analysis of $PM_{10}$ for the last six years starting from 2010 to 2015 in Malaysia showed an increase and decrease in $PM_{10}$ pollutant concentration for each region at the high polluted region (HPR), moderate polluted region (MPR) and low polluted region (LPR), with some of the $PM_{10}$ concentrations exceeded the Malaysia Ambient Air Quality Guidelines, which is 40 $\mu g/m^3$ for annual averaging time.

Regardless of their importance in determining the API level in Malaysia, there are some absence of $PM_{10}$ concentrations noticed for certain days possibly owing to equipment failure. Absent values is a common situation in almost all environmental research studies, especially in air pollution studies due to equipment failure. According to the data obtained from the Department of Environment (DOE), each year starting from 2010 to 2015, Malaysia was faced with $PM_{10}$ data absence. The number of data absence obtained for 2010, 2011, 2012, 2013, 2014, 2015 and 2016 were 128, 80, 82, 89, 69 and 78, respectively. These $PM_{10}$ absence values might contribute to the negative impact on humans as well as the environment.

Therefore, prediction of $PM_{10}$ is necessary to counteract the intensifying situation over the long run. Realising the importance of $PM_{10}$ in API, an attempt was made to develop the valid and trustworthy $PM_{10}$ prediction model by using ANN and MLR to ensure that the API can be calculated even with the $PM_{10}$ concentration absence. Accurate $PM_{10}$ predictions give important information to the public with the intention of minimising hazards to human health and also to alert the susceptible public, especially those with special health conditions like asthma patients. Therefore, these predictions will focus on $PM_{10}$ concentrations that have major effects on humans and

the environment. Specific policies in managing air pollution are needed for implementation, especially for the faster growth emission sources of $PM_{10}$, especially at residential megacities. Therefore, air pollution prediction is a critical part of the air pollution management policy.

Air quality prediction systems are required to develop good policies and warning bulletins when air pollutants surpass the maximum limiting values. Nowadays, an ANN is an extraordinary among most well-known prediction models. It has an awesome ability to predict or model the pollutants (Yildirim, 2006). Besides, over the years the MLR has been implemented to predict the $PM_{10}$ concentration in Malaysia, with an attempt to demonstrate the relation between at least two explanatory variables and a response variable by fitting a linear equation to the observed data (Abdullah et al., $2017_a$).

## 1.4 Scope of the study

The study covers all 52 air quality monitoring stations located in Malaysia. However, a station named ILP Miri in Sarawak was excluded from the analysis because there was no complete and continuous data available from 2010 to 2011. The study area was categorised according to the air quality monitoring stations, namely high polluted region (HPR), moderate polluted area (MPR) and low polluted area (LPR). The five main pollutants being focused on were CO, $NO_2$, $SO_2$, $O_3$, $PM_{10}$. Meanwhile, other pollutants ($NO_x$, NO, $CH_4$, NMHC, THC) and meteorological parameters (wind speed, wind direction, temperature, UVB, humidity) were also included. However, to determine the most significant parameters for the $PM_{10}$ prediction model input, only CO, $NO_2$, $SO_2$, $O_3$, $PM_{10}$, $NO_x$, NO, wind speed, wind direction, temperature and humidity were considered in the analysis. Parameters of UVB, $CH_4$, NMHC, and THC

were excluded from the analysis due to the high percentage of missing data. The statistical data for the $PM_{10}$ prediction model was performed by using cluster analysis, discriminant analysis and PCA, while the prediction model used MLR and ANN.

## 1.5 Objectives of the study

The study aims to obtain significant factors that affect $PM_{10}$ presence as the main pollutant as compared to others. Besides, other pollutants and meteorological parameters for the $PM_{10}$ prediction model input also play an important role to develop an accurate prediction model by using a trained learning machine that can be implemented by the Malaysian authorities to improve the air quality status in the study areas. To achieve this aim, the specific research objectives are:

i) To classify the air quality monitoring stations into three regions.
ii) To identify the presence of five main pollutants by the area based on yearly trend analysis.
iii) To determine the effect of monsoon seasons on $PM_{10}$ concentration.
iv) To develop the best $PM_{10}$ prediction model by using a trained learning machine.

## 1.6 Research questions

Research on interpreting and predicting the concentration of $PM_{10}$ in this study was guided by the following research questions based on the problem statement and objectives mentioned above:

1) Which air quality monitoring station can be classified as high polluted region (HPR), moderate polluted region (MPR) and low polluted region (LPR)?

2) How does the presence of five main pollutants, namely CO, $NO_2$, $SO_2$, $O_3$ and $PM_{10}$ affect the area based on yearly trend analysis?
3) Do monsoon seasons give a different $PM_{10}$ concentration pattern?
4) Will the ANN model predicts more accurately than the MLR model by using the most suitable parameters as inputs?

## 1.7 Significance of the study

This research is important to obtain a more comprehensive and accurate $PM_{10}$ prediction model as it is known as the main pollutant present in the atmosphere as compared to other pollutants. The increasing number of industries, development, as well as number of vehicles in Malaysia, has nowadays contributed to the daily increase in $PM_{10}$ concentration and deteriorates the air quality. According to Syafei (2014), daily data may be the best option for better derive policies. Besides transportation, $PM_{10}$ also comes from other complex activities that are high at a certain time frame in the morning, evening or even night; hence, suggesting a unique $PM_{10}$ emission pattern each day. In view of this scenario, it is important to develop a daily prediction model of $PM_{10}$ concentration in Malaysia as a unique daily presence pattern of $PM_{10}$ emission, which may give a negative effect to human health and the environment.

The $PM_{10}$ prediction model developed in this study can be used for developing strategies to control any negative incidents in regard to high $PM_{10}$ concentration, especially about the $PM_{10}$ data absence. The $PM_{10}$ prediction model will notify policy makers as well as residents to stay alert regarding the potential risk of $PM_{10}$ measured to legislate for cleaner air and keep early precautions. Right attitude and public behaviour through monitoring, prediction and reporting of air pollution are some of the implementations headed for a healthier environment, where systems can progressively

achieve information. Moreover, various $PM_{10}$ concentrations will correlate with corresponding available areas at the air quality monitoring stations. Therefore, the region was divided into high polluted region (HPR), moderate polluted region (MPR) and low polluted region (LPR) to ensure that an accurate $PM_{10}$ prediction model can be developed for each region because each region gives different levels and compositions of the pollutant according to the emission sources and backgrounds (Mohtar et al., 2018).

It is important to expand on air quality improvement instead of a controllable challenge. Translating the accurate prediction model unarguably has the possibility to diminish air pollution that as it will not pose a harmful toll on public health. Therefore, it is imperative to enlighten the public about the air pollution levels and associated health risks so that they can take their healthcare action. Accurate prediction is important to ensure early action is taken when pollutant concentration is present with missing data. Therefore, the air pollution concentration risk can be reduced by taking action against the predicted data. From the fundamental and statistical analyses, researchers, stakeholders and residents are clarified by the cause-effect mechanism. Besides, preventive actions can be taken to diminish air pollution influences and eventually control the air quality.

## 1.8 Organisation of the study

Overall, this study is organised in five chapters as follows:

**CHAPTER 1** This chapter focuses on the study background, research problem statement, study objectives followed by a list of research questions to answer the study objectives. Moreover, the study is also explained based on the study scope. The

significance of the study on accurate $PM_{10}$ prediction model towards the government, companies and humans are also discussed in this chapter.

**CHAPTER 2** This chapter discusses related literature review in regard to an overview of air pollution, characteristics of air pollutant parameters, source of air pollution, impacts of air pollution, as well as air quality standards and legislation. Other studies about $PM_{10}$ prediction models are also stated in this chapter.

**CHAPTER 3** This chapter provides a description of the methodology, including study areas, data collection as well as data analysis, which consist of data pretreatment, chemometric analysis (HACA, DA, PCA) and modelling analysis (ANN, MLR) to achieve the study objectives and answer all research questions.

**CHAPTER 4** This chapter presents and discusses the research findings in view of the objectives and research questions. Generally, the findings can be divided into two main sections, namely data interpretation and data prediction. Data are interpreted based on HPR, MPR and LPR obtained from HACA. Besides, changes in air quality from 2010 to 2015 are also discussed in regard to the trend analysis based on area from API annual frequencies, an annual diurnal variation of main pollutants (CO, $NO_2$, $O_3$, $SO_2$, and $PM_{10}$) and monthly $PM_{10}$ variation affected by the monsoon season. Meanwhile, for data prediction, inputs for the $PM_{10}$ prediction model are obtained from the PCA. $PM_{10}$ prediction models are developed by using lagged $PM_{10}$ data as input instead of without lagged $PM_{10}$ data as input.

**CHAPTER 5** This chapter gives a brief conclusion of the study purpose. Besides, the study implication also discusses further study in the future.

## 1.9 Theoretical framework

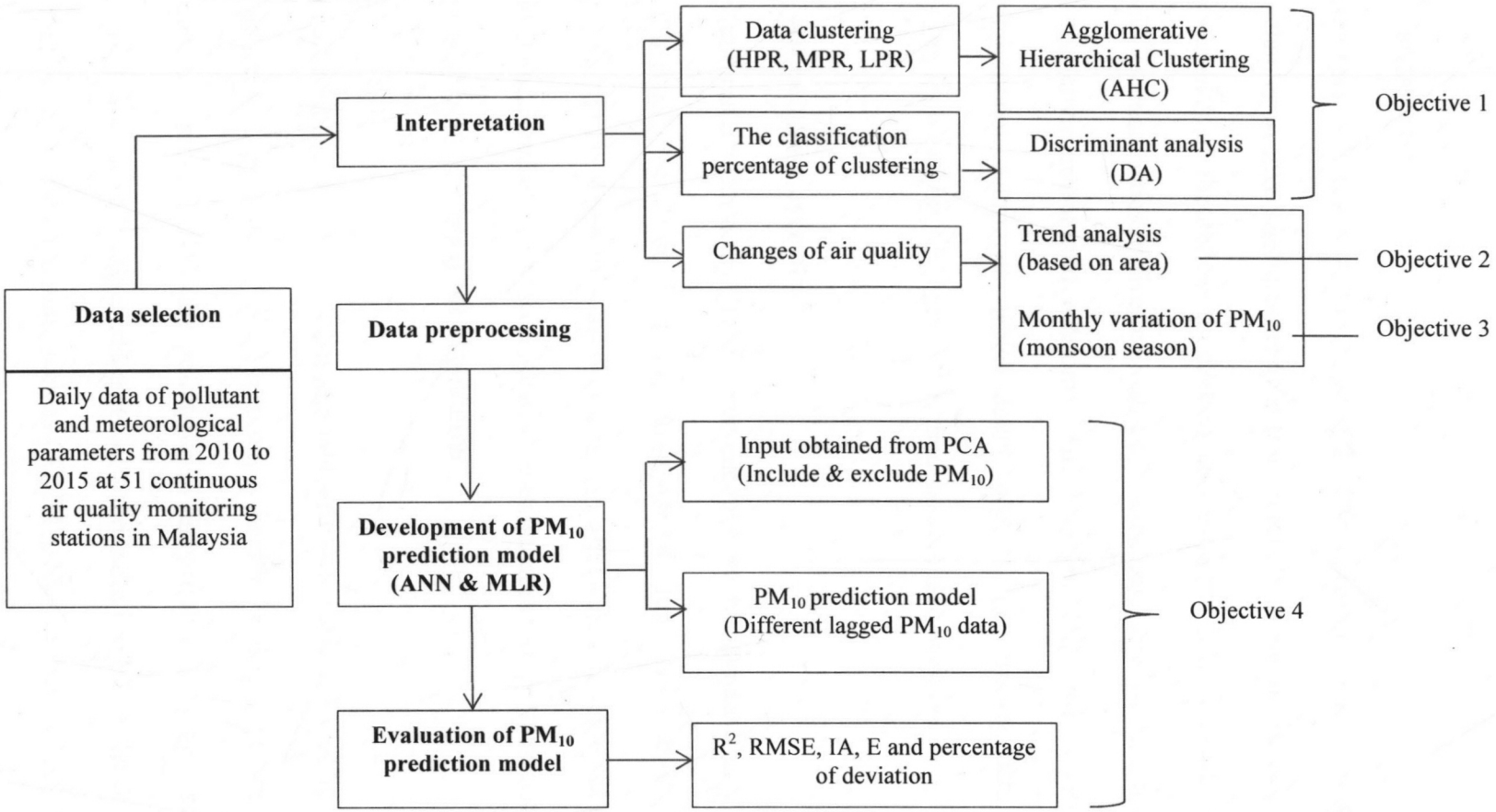



Figure 1.1. Theoretical framework

# CHAPTER 2

# INTRODUCTION

## 2.1 Introduction

Literature review in regard to the study of the $PM_{10}$ prediction model was constructed in this chapter to give useful indicators from relevant literature review. Besides, the review consists of theoretical foundation and concepts related to the air pollution overview, characteristics of five main air pollutants, their sources and impacts towards health, the environment and economy. The Malaysia Air Quality Guidelines and prediction model techniques, which detail more on the missing data pattern and various studies based on ANN and MLR models were also reviewed.

## 2.2 Air pollution overview

Air pollution is a noteworthy global issue encountered by humans and it is in the top ten health risks (Akhtar et al., 2018). Animals with lungs are subjected to oxygen supply from the air. Therefore, air pollution can affect the unintentional consumption of various undesired elements with adverse consequences. Air pollution shows a major risk to health besides other types of pollution.

As a result, reducing levels of air pollution directly reduces health problems. Low level of air pollution contributes to good respiratory health for both long and short terms. According to Darbre (2018), the pollutants in air pollution may be natural or man-made, which may arise as gases, droplets, or solid particles. For instance, it may arise in gaseous form from compounds, such as CO, $NO_2$, $SO_2$, and $O_3$, or they may occur as particulate matter (PM) in droplets or solid particles from compounds, such as dust,

soil, acids, organic molecules, and some metals. Seaspray is one of the examples for the natural PM source in droplet form. According to Ahmat et al. (2015), among other pollutants, $PM_{10}$ is the main pollutant present in the air as compared to other pollutants that are considered for the API calculation, namely CO, $NO_2$, $SO_2$, and $O_3$.

In 2016, there were 4.2 million premature deaths worldwide both in the cities and rural areas. In addition, 91% of the population lived in places that did not meet the air quality level guidelines (WHO, 2018). Many premature deaths were recorded in South East Asia and the Western Pacific region where most places were from low and middle-income countries. The highest and lowest number of deaths were recorded in the Democratic People's Republic of Korea and Brunei Darussalam with 238.4 deaths per 100, 000 population and 0.2 per 100, 000 population, respectively.

This was coherent with the finding by WHO International Agency for Research on Cancer (IARC) in 2013, which found that air pollution was cancer-causing to humans since the increase in cancer occurrence was connected with the component of PM in air pollution. For instance, air pollution had increased urinary tract or bladder cancer. In addition, Soni and Shukla (2012) mentioned that introduction to the air pollution with high concentrations over a short period (frequently seconds) was more significant in destructing human health than a long period of introduction. WHO mentioned that South East Asia is one of the regions that experience a high number of premature deaths due to air quality deterioration. Malaysia is situated at the centre of South East Asia and may experience a high number of premature deaths. Nowadays, the air quality in Malaysia is influenced by rapid growth of development, whereby it is one of the major contributors to local or regional air pollution problems. Besides, concentrated population density also increases the air pollution problems, especially

during peak hours. Among others, the urban and industrial areas also experience concentrated air pollution. However, polluted air quality may be dispersed by seasonal wind pattern and local factors, such as land and sea breeze as well as wind that transfer from rural to surrounding area (Amran et al., 2015).

Air pollution management is quite challenging owing to the source's intricacy as well as inadequate resources for enforcement. Therefore, exposure to air pollution may arise due to these problems. According to Wolff (2014), air pollution can be reduced by decreasing the air pollution emitters from polluted area; hence, to control the measures, it is important to predict the $PM_{10}$ concentration, which is the major pollutant present in air, especially in the polluted areas. According to Kumar and Goyal (2011), air quality prediction is effective in controlling measures and helps to perform regulations.

## 2.3 Sources of air pollution in Malaysia

Malaysia is facing air pollution due to the hasty growth in development, which is the main contributor to upsurge the local and regional air pollution level (Amran et al., 2015). Air pollution is one of the most imperative concerns around the world and is becoming an important pollution problem in Malaysia (Azid et al., 2016). Highly urbanised and industrialised areas usually experience severe air quality problems (Azmi et al., 2010). Therefore, air pollution monitor techniques and strategies need attention, especially to the five main pollutants that are considered in Malaysia for air pollution index (API), namely $SO_2$, $NO_2$, $O_3$, CO and $PM_{10}$.

Instantaneous and serious attention is needed from the authorities in regard to the serious air pollution issue around the world since it is one of the main factors that

cause life and living deterioration. According to Azid et al. (2014) and Mutalib et al. (2013), increasing number in transportation and industrial activities, as well as transboundary pollution from neighbouring countries by mobile, stationary and transboundary sources, contribute to the air pollution problems, which cause the main environmental issues in Malaysia. In Malaysia, 70% to 75%, 20% to 25% and 3% to 5% of total air pollution come from the mobile sources, stationary sources and transboundary sources, respectively (Azid et al., 2013). This shows that major sources of air pollution mainly come from mobile sources.

Besides man-made air pollution, natural sources of air pollution must not be disregarded. Heavy smoke usually produced from wildfires, for instance, forest fire and agricultural burning could be harmful to the human respiratory system. Carbon emissions from peat fires are proportional to 15% emission of anthropogenic, creating worries to the worldwide carbon cycle (Hu et al., 2018). Unfortunately, Indonesia, which is located adjacent to Malaysia, is stated as one of the countries with peatland areas besides Russia, Canada and the USA. Therefore, there is a high probability of transboundary pollution from a neighbouring country to give effect to the presence of pollutants in Malaysia and perhaps can be the main cause of air pollution deterioration, especially during the haze phenomenon. Figure 2.1 visualises biomass burning during a wildfire for the flaming of the grass and the smouldering of organic soil (Rein, 2008).

(a) Flaming | (b) Smouldering

Figure 2.1. Biomass burning during a wildfire (a) flaming of grass (b) smouldering of organic soil (Rein, 2008)

The comparison between flaming and smouldering peatland fire is shown in Table 2.1. There is a strong fire plume during flaming fire, while smouldering fire has a weak fire plume. Flaming fire is a surface occurrence with fast-moving diffusion flames, produces black smoke with abundant soot, short soil heating and minimal soil/ecosystem damage. In contrast, smouldering fire is a volumetric occurrence with creeping flameless reaction, produces whitish or yellowish smoke with abundant organic carbon, larger soil thermal severity with longer residence time and lethal damage to soil properties and biological systems (Rein, 2008).

Table 2.1. Comparison between flaming and smouldering peatland fire (Hu et al., 2018)

| Flaming fire | Smouldering fire |
|---|---|
| Strong fire plume | Weak fire plume |
| Surface occurrence | Volumetric occurrence |
| Fast moving diffusion of flames | Creeping Flameless moving diffusion of flames |
| Black smoke (abundant soot) | Whitish/yellowish smoke (abundant organic carbon) |
| Short soil heating | Longer soil heating |
| Minimal damage of soil/ecosystem | Lethal damage of soil/ecosystem |

Generally, smouldering fires are the most persistent type of combustion phenomena and directly produce significant amount of aerosols that result in severe regional haze episodes (Hu et al., 2018). Roots, seeds and plant stem are affected because of the prolonged heating through the smouldering fires for long-term damage (Rein et al., 2008). In general, peat fires emit a various complex mixture of gases and an aerosol (Stockwell et al., 2016), for instance, $CO_2$, CO, $CH_4$, HCN (hydrogen cyanide) and $NH_3$ (ammonia). Moreover, PM is emitted into the atmosphere in various sizes. However, $PM_{10}$ presence seemed to be the most on 20 October 2015, whereby Palangkaraya, in Indonesia gave a high $PM_{10}$ concentration number, which was 3741 $\mu g/m^3$. The source of pollutants (CO, $SO_2$, $NO_2$, $O_3$, $PM_{10}$) in the atmosphere, which was due to the natural and anthropogenic, is summarised in Table 2.2.

Table 2.2. Natural and anthropogenic sources of pollutants, namely CO, $SO_2$, $NO_2$, $O_3$ and $PM_{10}$ in the atmosphere (Arif et al., 2013)

| Pollutant | Source (natural) | Source (anthropogenic) |
|---|---|---|
| CO | Wildfires | Incomplete fuel combustion (natural gas, coal, wood, petrol) |
| $SO_2$ | Volcanoes | Activities from industrial; cause acid rain from the pollutant of $SO_2$ emitted with atmospheric water |
| $NO_2$ | Thunderstorms | Combustion with high-temperature |
| $O_3$ | None (secondary pollutant) | Fossil fuel combustion |
| $PM_{10}$ | Volcanoes, dust storms, forest fires, sea spray | Combustion of fuel from motor vehicles, marine vessels, aircraft, power plants; industrial processes; cigarette smoking. |

## 2.4 Impacts of air pollution

With rapid industrial development and urbanisation in Malaysia, air pollution has become dominant and harmful, affecting human health, the environment and economy.

### 2.4.1 Health

A huge number of health issues and respiratory diseases are caused by air pollution. Poor air quality gives symptoms, such as irritation of eye, skin, nose and throat,

headache, tiredness, dizziness, and difficulty in breathing experienced in humans. In 2012, WHO reported that outdoor urban and rural sources caused human death of around 3.7 million (WHO, 2014). Diseases, such as ischaemic heart disease, stroke, chronic obstructive pulmonary disease (COPD), lung cancer and acute lower respiratory infections in children contribute to 40%, 40%, 11%, 6% and 3% of the deaths cause, respectively.

Western Pacific and South East Asia with low and middle-income countries showed the largest outdoor air pollution problem with 2.6 million related to deaths (WHO, 2014) caused by heavy industry and air pollution hotspots surrounded by the developing nations of these areas. PM and $O_3$ are the riskiest elements that could harm the human health (Mofijur et al., 2016). This is coherent with the study done by Mahiyuddin et al. (2013), who mentioned that PM and $O_3$ were the vital pollutants related to common mortality.

A few epidemiological investigations revealed that the positive relation between PM and several negative impacts on the human good health, along these lines turning into a major challenge to public health, particularly the respiratory and cardiovascular framework (Shahadin et al., 2018). In general, PM results in lung cancer and cardiopulmonary deaths; $NO_x$ irritates the lungs and leads to oedema, bronchitis and pneumonia; CO influences young child tissue's development besides fetal growth in pregnant women; while $O_3$ causes coughing, lung function, headaches and physical discomfort (Mofijur et al., 2016).

### 2.4.2 Environment

In addition to health, air pollution also gives an impact to the environment. Greenhouse gas ($CO_2$) that is emitted from transportation results in climate change. Besides, emitted $CO_2$ by changes in area is mainly due to deforestation, which is an important factor for global emissions. Acidification through reaction among pollutants also indirectly affects the environment, for instance, acid rain produced from the sulphur emissions in the atmosphere (Mofijur et al., 2016). According to Goudie (2018), the productivity of forest can be either a decrease or an increase through acid rain contingent upon local factors. For instance, acid rain with moderate inputs is probably going to stimulate forest growth if there are abundant nutrient cations and deficient sulphur and nitrogen. However, soil character is generally changed by acid rain, whereby essential nutrients are leached, and thus make the soil less accessible for plants to utilise.

Besides, the trans-boundary haze pollution has adverse influences on the environment because of greenhouse gas emissions in addition to ecosystem and biodiversity, which affect climate changes (Aghamohammadi & Isahak, 2018). Haze generated from uncontrolled forest fire also affects the environment, whereby this air pollution gave the worst haze episode from July to October 2007 at most parts of Malaysia, Indonesia, southern Thailand, Singapore, southern Philippines and Brunei. This incident reduced visibility, because of the scattering of particles. Areas affected by haze experience much warmer situation since there was less rainfall recorded during the haze event (Arif et al., 2013).

High amount of toxic smoke is produced during haze episodes. This indirectly affects the wildlife by changing their surroundings. A study done by Erb et al. (2018) showed

that ecological activity of orangutans in Central Kalimantan decreased during the haze episodes, in which their rest time increased during and after the smoke period, while their distance and travel time decreased and fat catabolism increased after smoke. Results indicated that the wildfires smoke adversely affected the orangutan state, which was most likely related to immune response. The largest orangutan population populates the Central Kalimantan peat swamp forests. Therefore, burning more than 20, 000 $km^2$ of the forest, killed hundreds of orangutans in 2015 alone (Ancrenaz et al., 2018).

### 2.4.3 Economy

The air pollution phenomenon also indirectly affects Malaysia's economy. A review study done by Manan et al. (2018) showed that there was increasing prematurity of mortality as well as respiratory problems during the Asian Haze that occurred in 1997. Therefore, indirect increase in number of hospitalisations with the cost of hospital admission was MYR21 million. Unfortunately, the cost was increased after that. The range of cost on hospital admission from 2005 to 2013 was noted as MYR1.8 million to MYR118.9 million, respectively. Besides, reducing crop productivity also affects Malaysia's economy .

This happened since the haze caused a decrease in light that reached the Earth's surface. This indirectly lowers the rate of photosynthesis, and eventually reduces plant production. A study was done by Aziz et al. (2018) during the haze scenario on 2014. It showed that the haze scenario was reduced by 12.9% to 53.2% for the net photosynthetic rate and 26.1% to 73.8% for stomatal conductance of rice throughout the day. Due to this scenario, there was 10% to 19% reduction in rice yield as compared to the potential yield.

## 2.5 Characteristics of five main pollutants in the air

Air pollutant is a substance in the form of solid particles, droplets or gases in the air that can harm humans and the environment, which could happen naturally or from man-made, and is classified as either primary or secondary. This subchapter discusses the characteristics of the five main air pollutants ($PM_{10}$, $SO_2$, $NO_2$, CO, $O_3$).

### 2.5.1 Particulate matter ($PM_{10}$)

$PM_{10}$ alludes to particles with a diameter of less than or equal to 10 micrometres (μm), which comprise an assortment of discrete particles, either droplets (aerosols) or solids suspended in the atmosphere (Chen et al., 2016). They are also called as thoracic particles with an aerodynamic diameter of less than or equal to 10 μm and greater than 2.5 μm. These small inhalable particles can penetrate deeply into the lungs, indirectly causing health effects to humans. According to USEPA (2015), thoracic particles are mostly produced from the mechanical processes and uncontrolled burning.

In general, respiration processes through the nose remove particles larger than 15 μm. Therefore, small inhalable particles of thoracic particles may be deposited in the respiratory tract, such as nasal cavity, pharynx and larynx that eventually could create some irritation effects like inflammation and nose or throat dryness (Famoso et al., 2015). However, chemical and physical properties of particulates differ prominently with time, region, meteorology as well as the source of emissions (USEPA, 2015).

According to Yong and Awang (2017), the components of PM vary and fluctuately produced from a wide variety of sources, including natural and anthropogenic. Natural sources comprise of particle emission from the volcano eruptions, forest fires and particles of salt-forming from the sea spray evaporation (Ahmed et al., 2016).

Meanwhile, emitted $PM_{10}$ from anthropogenic source comprises of particle emissions mobile and stationary sources, such as industries, transportation, construction sites and open burning. According to DOE ($2015_a$), local and transboundary haze are the main sources of $PM_{10}$.

Besides being emitted directly from the source, this pollutant is also emitted through chemical reactions among precursor gases, for instance, $NO_x$ and $SO_2$ (Yong & Awang, 2017). In addition, Shukla and Sharma (2008) also mentioned that secondary $PM_{10}$ is emitted into the atmosphere by primary precursor gases, for example, $SO_2$, CO, $NO_x$ and ammonia ($NH_3$), which indirectly produce secondary $PM_{10}$, namely sulphates, nitrates, complex carbon compounds and condensed organic compounds. This shows that air pollutants, for instance, CO, $O_3$, $NO_2$ and $SO_2$ can affect to the $PM_{10}$ concentration.

$PM_{10}$ is a well-known air pollutant that is mainly related with harmful health effects. Studies have shown that $PM_{10}$ worsen respiratory and cardiopulmonary roles besides increasing the mortality level of related diseases. Additionally, exposure to $PM_{10}$ during pregnancy may result in birth abnormalities. Meanwhile, particular populations, such as children, elders, and those with asthma and cardiovascular problems have high probability to be affected by the $PM_{10}$ pollution (Ng & Awang, 2018).

It also can cause mild to severe illnesses reliant upon the exposure level with the most predominant clinical symptoms of respiratory disease caused by air pollution; wheezing, cough, dry mouth and restraint in activities owing to breathing problems. This also may reduce life expectancy due to the increase in cardiopulmonary and lung cancer (Ghorani-Azam et al., 2016).

### 2.5.2 Sulphur dioxide ($SO_2$)

According to Jeong and Park (2013), $SO_2$ is one of the air quality indicators besides other pollutants. In the urban areas, 73% of $SO_2$ comes from fossil fuel burning, for instance, coal and heavy fuel oil. Meanwhile 20% comes from other industrial facilities and less than 7% of $SO_2$ source comes from the natural sources (Nazari et al., 2012). This is coherent with a study done by Kim and Lee (2018) who mentioned that $SO_2$ is present at high level in urban area because of high sulphur content usage, for example, coal and heavy oils.

However, the $SO_2$ concentration is also influenced by the diesel engines emissions from motor vehicles (Mutalib et al., 2013). This is supported by a study done by Afzali et al. (2017), whereby $SO_2$ concentrations is more from on-road vehicle emission sources rather than the industrial emission sources. In general, industrial raw materials, such as crude oil and coal are usually related to the $SO_2$ gas, whereby when fuel or coal burns, the $SO_2$ gas is released. Industries and power generation plants usually cause the presence of $SO_2$ in the atmosphere (Hosseiniebalam & Ghaffarpasand, 2015). A similar finding also found that higher $SO_2$ levels in Venice sites could likely be associated with industrial emissions. For instance, thermal power plant, oil refinery and municipal solid waste incinerator (Masiol et al., 2017).

Meanwhile, in the metropolitan area, $SO_2$ presence represents the effects of industrialisation on the environment (Luvsan et al., 2012). Besides industrialisation, previous study showed that many factors contributed to the presence of $SO_2$ concentration in a region (Iqbal et al., 2014). According to the research done by Ray and Kim (2014), meteorological factors viz. climate conditions also affect the $SO_2$ concentration. For example, certain $SO_2$ reactions in water might produce an acidic

solution (Akabueze et al., 2012) and the production of acid rain in which $SO_2$ is the major component can destroy the environment.

The reaction that causes acid rain processes is shown below:

$$SO_2 + \frac{1}{2}O_2 \rightarrow SO_3$$

$$SO_3 + H_2O \rightarrow H_2SO_4 \quad (2.1)$$

Therefore, the pollutant of $SO_2$ is the main element of acid rain that indirectly harms the environment. According to Jeong and Park (2013), air quality indicator includes $SO_2$, which gives adverse effect on human health. The $SO_2$ concentration can also be naturally emitted from volcanic eruptions. However, since Malaysia is located far from the volcano, other sources mentioned earlier, such as urban, transportation and industrial might be the cause for the presence of $SO_2$concentration. This type of gas emitted from various sources contributes to health problems, such as respiratory system and eyes irritation owing to the presence of diluted sulphuric acid around the eyes. One can also experience permanent damage to the respiratory system when exposed to high $SO_2$ concentration.

Besides, the reaction between $SO_2$ and other chemicals form tiny sulphate particles in the air can contribute to difficulty in breathing as these tiny particles may be inhaled and gathered in the lungs and cause difficulty in breathing as well as premature death (Soni & Shukla, 2012; USEPA, 2011). In addition, according to USEPA (2015), people with asthma are mostly at risk whereby short-term exposures of $SO_2$ cause elevation of $SO_2$ level. Besides, they might experience difficulties in breathing and shortness of breath at the moderate level exposure of $SO_2$.

### 2.5.3 Nitrogen dioxide ($NO_2$)

Nitrogen dioxide ($NO_2$) is categorised in the nitrogen oxides ($NO_x$) family, whereby this highly reactive gas consists of nitrogen and oxygen. In general, nitrogen oxides ($NO_x$) consist of NO, $NO_2$ and $NO_3$ (Soni & Shukla, 2012). The $NO_2$ present in the urban outdoor air is predominant as a result of the traffic and power plants due to high temperature of combustion processes (Famoso et al., 2015). According to the DOE ($2015_a$) in Malaysia, 65%, 27%, 6% and 2% emissions of $NO_2$ are designated from industries, motor vehicles, power stations and other sources, respectively. Increase in vehicles number, as well as processes of combustion, cause $NO_2$ concentrations to be continuously high in urban and industrial areas. Meanwhile, indoor $NO_2$ pollution is released from the gas stoves and unvented heaters (Mishra & Goyal, 2015).

$NO_2$ and other $NO_x$ react as a precursor to harmful secondary air pollutants, namely $HNO_3$, $NO_3^-$ in secondary inorganic aerosols, and $O_3$. However, the reactions take some time, depending on the atmospheric composition and meteorological parameters. Besides, before secondary pollutants are produced, air can travel to some distance. Therefore, health risks due to $NO_x$ may possibly come from $O_3$, a secondary particle from the products of reaction, as well as from the $NO_2$ itself.

On top of that, human resistance to respiratory infections becomes lower and aggravates existing chronic respiratory diseases during $NO_2$ long-term exposures. People, especially with asthma, may face more frequent and intense attack due to the increase in $NO_2$ levels. It is riskier to children with asthma and older people with heart disease (USEPA, 2015). Road vehicles, industry and households contribute to the production of $NO_2$ toxic gas. One may experience eye, nose, throat and lung irritations when facing short-term exposure from 30 minutes to 24 hours. While long-term

exposure may permanently distress lung function. Additionally, it is the foremost precursor for the $O_3$ formation at ground level that is risky to the human health (Catalano et al., 2016). Besides, $O_3$ formed from the $NO_2$ contributor also has the potential to give adverse effects on terrestrial and aquatic ecosystems.

Equations below show the reaction formation of $NO_2$ in the atmosphere:

$$N_2 + O_2 \rightarrow 2NO$$
$$2NO + O_2 \rightarrow 2NO_2 \quad (2.2)$$

Where, combustion processes allow the reaction between $N_2$ and $O_2$ in the atmosphere with temperature higher than 1200°C to produce NO. While during the soot cooling processes, NO reacts with $O_2$ to form $NO_2$.

### 2.5.4 Carbon monoxide (CO)

Inappropriate combustion from the burning of coal and wood emits colourless and odourless CO gas (Ghorani-Azam et al., 2016). According to a study done by Dominick et al. (2012), CO was the main air pollutant that influenced the high $PM_{10}$ concentration, owing to the combustion processes predominantly originating from transportation, besides meteorological factors, for instance, ambient temperature, wind speed and humidity.

According to Guerreiro et al. (2016), an urban area especially during the peak hour at traffic locations, experienced the highest CO concentration, besides during the downwind emissions from the large industry. This is supported by Mutalib et al. (2013), who mentioned that CO was mostly emitted from motor vehicle emissions

during high traffic. Incomplete fuel burning from the malfunction vehicle engine caused higher CO concentration presence. According to the statistical report published by DOE ($2015_a$), motor vehicles contributed to 95% emission of CO in urban areas.

Combustion process of hydrocarbons produces CO. Equations below show the reaction product of CO in the atmosphere:

$$2C + O_2 \rightarrow 2CO$$
$$2CO + O_2 \rightarrow 2CO_2$$
$$CO_2 \rightarrow CO + O$$
$$CO_2 + C \rightarrow 2CO \quad (2.3)$$

The presence of high CO concentrations is harmful to the cardiovascular system, such as hypertension and heart attacks as well as the nervous system and fetuses. Besides, it can combine with haemoglobin, which functions as an oxygen carrier in lieu of oxygen (Famoso et al., 2015). This causes a decrease in oxygen supply to tissues and organs, for example, the heart. Therefore, people, especially with heart disease, are recognised as at the highest risk when exposed to high CO concentration.

Besides, people with chronic obstructive pulmonary disease, anaemia, diabetes, prenatal and elderly life stages are also potentially at risk in regard to the CO concentration present in air (USEPA, 2015). According to Ghorani-Azam et al. (2016), the poisoning of CO symptoms are very alike to other illnesses, for example, food poisoning or viral infections, such as a headache, dizziness, weakness, nausea, vomiting, and finally loss of consciousness depending on the CO concentration and length of exposure.

### 2.5.5 Ozone ($O_3$)

Colourless gas of $O_3$ is one of the major components of atmosphere, which can be found both at the ground and upper levels of the atmosphere called the troposphere. The reaction between VOCs that are emitted from natural sources and human activities with $NO_x$ will produce ground-level $O_3$. According to Famoso et al. (2015), $O_3$ is a strongly unstable compound formed in relation to the ultraviolet light absorption by $NO_2$, which is separated in NO and O atoms. Free oxygen (O) is produced from the process and then reacts with molecular oxygen ($O_2$) to form $O_3$.

The interaction between ultraviolet light absorption by $NO_2$ and $O_3$ formation is shown below:

$$NO_2 + hv \rightarrow NO + O$$
$$O + O_2 \rightarrow O_3 \qquad (2.4)$$

$O_3$ acts as a greenhouse gas (GHG) that occurs all the way through the lower part of the Earth's atmosphere (tropospheric), warming the Earth's surface by trapping heat from the sun.

Pollutant reaction emitted from industries and electric utilities can produce ground-level $O_3$, besides other $O_3$ precursors emitted from chemicals as well as from natural sources, such as trees and other plants. Besides, formation, transportation and dispersion of $O_3$ are influenced by meteorological conditions. Predominantly, formation and accumulation of surface $O_3$ at the surface atmosphere result from intense solar irradiation, low wind speeds and high temperature. On the whole, according to Mutalib et al. (2013), temperature tends to give high levels of $O_3$ in the

atmosphere, frequently in warm sunny atmosphere through the reaction between $NO_x$ and VOCs.

According to Sari et al. (2016), characteristics and $O_3$ levels in urban and rural areas have major dissimilarities, whereby urban areas have lower $O_3$ as compared to the rural area, owing to high levels of $NO_x$ and low VOCs in urban areas. Higher $NO_x$ concentrations in urban areas cause $O_3$ deterioration. At low $NO_x$, the production rate of $O_3$ increases until it reaches the maximum, while at high $NO_x$, the production rate of $O_3$ decreases. However, exact sources of $O_3$ occurrences are difficult to identify as it is involved in the complexity of photochemical reactions. Besides, surface $O_3$ and its precursors involve transportation and dispersion by advection, vertical intrusion, and mixing in the planetary boundary layer. These increase difficulty to figure out the exact source of $O_3$ (Ahamad et al., 2014).

Exposure to $O_3$ can cause susceptibility to respiratory illnesses, namely asthma and bronchitis (USEPA, 2015). Irritation on mucous membranes and respiratory tract is the most unsafe among other effects on health due to the concentrated $O_3$ located between bronchioles and alveoli known as the terminal part of a respiratory system where it exercises its intense oxidising action. Besides, mortality among prematures is also related to the short-term $O_3$ exposure. Meanwhile, long-term exposure of $O_3$ is related to permanent lung tissue damage.

## 2.6 Malaysia Air Quality Guidelines

The effects of air pollution have become a critical part in the thought of the Malaysian government to control the pollution stages that could pose threats to public health. Researchers showed that in some cities, air pollutants were increased with time and

beyond the acceptable levels according to the air quality guidelines. To control issues in regard to air pollution in Malaysia, the New Malaysia Ambient Air Quality Standard and air pollution index (API) were enacted. To recover and maintain air quality and protect public health, the Malaysian Department of Environment (DOE) has set up the API and established the national air-quality standards through the Malaysia Ambient Air Quality Standard (MAQS) for each of these pollutants.

### 2.6.1 Malaysia Ambient Air Quality Standard (MAQS)

In order to measure the air quality status in Malaysia, the Department of Environment (DOE) has issued the Recommended Malaysia Ambient Air Quality Guidelines (RMAQG) since 1989. The older RMAQG that was used since 1989 was replaced by the new MAQS, whereby the concentration limit of air pollutants will be reinforced gradually until 2020. Table 2.3 shows the new MAQS implemented by the government onto pollutants of $PM_{10}$, $PM_{2.5}$, $SO_2$, $NO_2$, $O_3$ and CO with three interim targets, which consist of interim Target 1 (IT-1), interim Target 2 (IT-2) and the full implementation of the standard in 2015, 2018 and 2020, respectively.

Table 2.3. New Malaysia Ambient Air Quality Standard (DOE, 2018)

| Pollutants | Averaging time | Ambient Air Quality Standards | | |
|---|---|---|---|---|
| | | IT-1 (2015) | IT-2 (2018) | Standard (2020) |
| $PM_{10}$ ($\mu g/m^3$) | 1 Year | 50 | 45 | 40 |
| | 24 Hour | 150 | 120 | 100 |
| $PM_{2.5}$ ($\mu g/m^3$) | 1 Year | 35 | 25 | 15 |
| | 24 Hour | 75 | 50 | 35 |
| $SO_2$ ($\mu g/m^3$) | 1 Year | 350 | 300 | 250 |
| | 24 Hour | 105 | 90 | 80 |
| $NO_2$ ($\mu g/m^3$) | 1 Hour | 320 | 300 | 280 |
| | 24 Hour | 75 | 75 | 70 |
| $O_3$ ($\mu g/m^3$) | 1 Hour | 200 | 200 | 180 |
| | 8 Hour | 120 | 120 | 100 |
| CO ($mg/m^3$) | 1 Hour | 35 | 35 | 30 |
| | 8 Hour | 10 | 10 | 10 |

### 2.6.2 Air Pollution Index (API)

The amount of air pollutant concentration limit that may adversely influence the general public health and welfare are defined through the RMAQG, which was formulated by the Department of Environment (DOE), Malaysia in 1989. The Malaysian Air Quality Index (MAQI) was developed in 1993 to expose the public about the ambient air quality status, ranging from good to emergency. Air pollution management, as well as public health protection, have become more effective, especially in industrialised countries.

Malaysia's API, which closely follows the United States system known as Pollutant Standard Index (PSI), was adopted in 1996 with the aim of easy evaluation with countries in addition for regional harmonisation. Nevertheless, in 1999, the PSI was replaced with the Air Quality Index (AQI), but for the meantime, Malaysia remains with API. Five main air pollutants, namely $PM_{10}$, $SO_2$, $NO_2$, $O_3$ and CO were used in calculating and reporting the API. Range values from good to emergency are reported for the API, which reflect its result on the human health. The scales of status indicator used to describe the air quality levels are in Table 2.4.

Table 2.4. Air Pollution Index (API) status indicator (DOE, 2000)

| API | DESCRIPTOR |
|---|---|
| 0-50 | Good |
| 51-100 | Moderate |
| 101-200 | Unhealthy |
| 201-300 | Very unhealthy |
| >300 | Hazardous |
| >500 | Emergency |

The air quality is considered good, moderate, unhealthy, very unhealthy, hazardous and emergency if the API values are ranging between 0 to 50, 51 to 100, 101 to 200, 201 to 300, greater than 300 and greater than 500, respectively. There is severe

atmospheric pollution if the value of API becomes higher and greater than the health concerns. Through API, the public can easily know the air quality status for defining health precautions as well as being up-to-date about atmosphere cleanliness.

### 2.6.2.1 API calculation and health measures

To determine the status indicator of API for a given time period, the sub-index value for each of the main pollutants was calculated. The API value was selected from the maximum sub-index among all pollutants. Then, actions are needed based on the reported health effects category. The flowchart process of API selection value at a given time is detailed in Figure 2.2.

The general public can easily understand a simple, comprehensive API to define the air quality status. Each pollutant has different averaging values. For instance, $PM_{10}$ and $SO_2$ are averaged over a 24-hour running period, CO is averaged over an eight-hour period, and $O_3$ and $NO_2$ have averaged over a one-hour running period before calculating the index with the use of sub-index functions for each pollutant according to the standpoint of human health implications. According to WHO (2006), pollutants are measured at a different averaging time based on its different implications on human health. Individual indices are calculated based on individual pollutants.

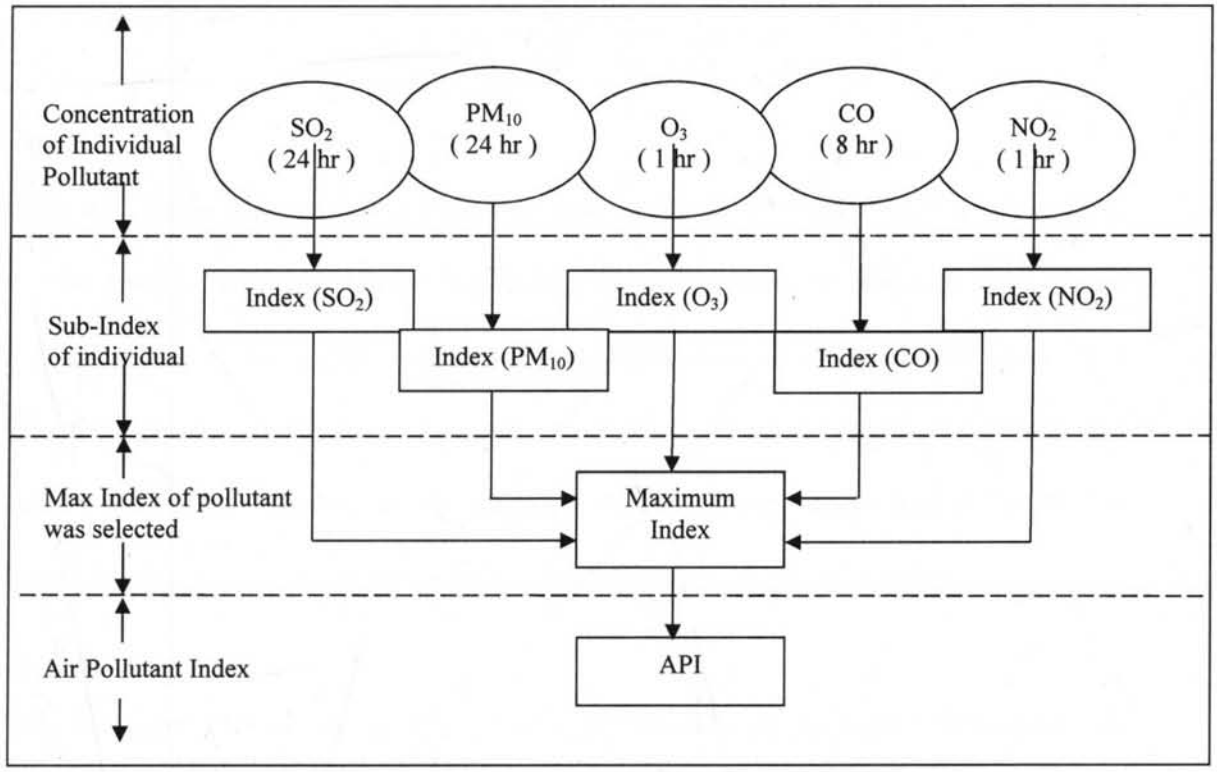



Figure 2.2. Flowchart of the API process (DOE, 2000)

The maximum index among the pollutants was selected and then considered responsible for the API value. All the pollutants sub-indices can be calculated by using the equation as shown in Table 2.5.

Table 2.5. API equation for each of pollutants (DOE, 2000)

| **Pollutant** | | **Equation for API calculation** |
|---|---|---|
| CO | conc. < 9 ppm | API = conc. × ll.11lll |
| (Based on 8-hour | 9 < conc. <15 | API = 100 + {[conc. - 9] × 16.66667} |
| average concentration) | 15 < conc. < 30 | API = 200 + {[conc. - 15] × 6.66667} |
| | conc. > 30 ppm | API = 300 + {[conc. - 30] × 10} |
| $O_3$ | conc. < 0.2 ppm | API = conc. x 1000 |
| (Based on 1-hour | 0.2 < conc. < 0.4 | API = 200 + {[conc. - 0.2] × 500} |
| average concentration) | conc. > 0.4 ppm | API = 300 + {[conc. - 0.4] × 1000} |
| $NO_2$ | conc. < 0.17 ppm | API = conc. × 588.23529 |
| (Based on 1-hour | 0.17<conc.<0.6 | API = 100 + {[conc. - 0.17] × 232.56} |
| average concentration) | 0.6<conc.<1.2 | API = 200 + {[conc. - 0.6] × 166.667} |
| | conc. > 1.2 ppm | API = 300 + {[conc. - 1.2] × 250} |
| $SO_2$ | conc. < 0.04 ppm | API= conc. × 2500 |
| (Based on 24-hour | 0.04<conc.<0.3 | API= 100 + {[conc. - 0.04] × 384.61} |
| average concentration) | 0.3<conc.<0.6 | API= 200 + {[conc. - 0.3] × 333.333} |
| | conc. > 0.6 ppm | API= 300 + {[conc. - 0.6] × 500} |
| $PM_{10}$ | conc. < 50μg/m$^3$ | API = conc. |
| (Based on 24-hour | 50 < conc. < 150 | API = 50 + {[conc. - 50] × 0.5} |
| average concentration) | 150 < conc.< 350 | API = 100 + {[conc. - 150] × 0.5} |
| | 350< conc. <420 | API = 200 + {[conc. - 350] × 1.4286} |

| | |
|---|---|
| $420 < \text{conc.} < 500$ | $API = 300 + \{[\text{conc.} - 420] \times 1.25\}$ |
| conc. > 500 $\mu g/m^3$ | $API = 400 + [\text{conc.} - 500]$ |

Each API status has a different impact on human health (DOE, 2000). For instance, at a good level of API, there is no implication of health since the pollution level is low and the quality of air is significantly cleaner. Meanwhile, at moderate API level, very sensitive groups on $O_3$ and $PM_{10}$ may experience moderate health. However, the quality of air is still acceptable. At unhealthy API level, individuals with difficulties in breathing should lessen outdoor exercise since slight irritations may happen.

This API level gives mild aggravation of symptoms among the highest risk group. At very unhealthy API level, people with breathing or heart problems as well as elders should stay indoors and limit their activities, while healthy people will be noticeably affected. At hazardous API level, elders and the sick should stay indoors and keep off activities. At this level, healthy individuals also should keep away from outdoor activities. There may be strong irritations and symptoms that may cause other illnesses. At emergency API level, people and public health are at high risk.

In Malaysia, $PM_{10}$ is one of the major pollutants that usually determine the API level. It is recognised as a pollutant that affects human health, especially during haze episodes. During the normal days, this pollutant might be influenced by the traffic volume and amount of industrial activities (Mutalib et al., 2013). Comparing to other pollutants, $PM_{10}$ gives the highest concentration and is considered as the main pollutant (Amran et al., 2015). According to Azid et al. (2014), $PM_{10}$ pollutant is the most noteworthy in determining the API value contrasted with other air pollutants based on some past research reports.

## 2.7 Machine learning

Machine learning (ML) is the most important sub-field of computational intelligence, which is also called artificial intelligence (AI). This computer science branch makes computers capable of performing a task without being explicitly programmed. Its main objective is to extract information from data by using computational methods. This method is widely used in environmental sciences, such as data processing, forecasting of air quality and hydrology as well as weather and climate predictions (Kang et al., 2018; Peng, 2015).

For the last two decades, ML techniques were used in air quality modelling (Gardner & Dorling, 1998). Many studies used numerous ML methods in air quality modelling to search more practical models. ML methods, for example Artificial Neural Networks (ANN), Multiple Linear Regression (MLR), Decision Tree, Multivariate Adaptive Regression Splines (MARS) and Support Vector Machine (SVM) can build comparable and better accuracy air quality models. Besides, machine learning algorithms can handle complex and non-linear relations that exist between air quality parameters (Esplin, 1995) and produce models that implement well in predicting unseen data.

### 2.7.1 Artificial Neural Network (ANN)

Development of ANN is the consequence of simplified mathematical formulation models in view of the brain physiology. There is a highly complex system of information processing in the human brain with the main element as the capability of parallel and non-linear external information-stimuli processing. The architectural structure of ANN comprises a great parallel number, adaptive processing units interconnected set and organisation of hierarchical to interrelate with the environment.

ANN and the network of biological share the similarities by providing information to the networks through the process of training (Alimissis et al., 2018). It is an advanced model used for predicting the pollutant concentration (Moustris et al., 2013). In this study, a multilayer perceptron feed-forward artificial neural network (MLP-FF-ANN), which is widely used in the applications of atmospheric science, was implemented to predict the concentration of $PM_{10}$. Figure 2.3 shows the network structure of the MLP-FF-ANN model, whereby the structure comprises of layers, namely, input layer (independent parameter), hidden layer and output layer (dependent parameter) arranged together with multiple neurons or nodes.

Information from the input system, also known as the independent parameters, flows from the input layer to the hidden layer to process the signal. Finally, the signal comes to the output layer, which is also known as the dependent parameter. ANN pre-modelling stages are considered to have a better prediction model. The data preparation procedures for development of ANN modelling consist of training, testing and validation datasets.

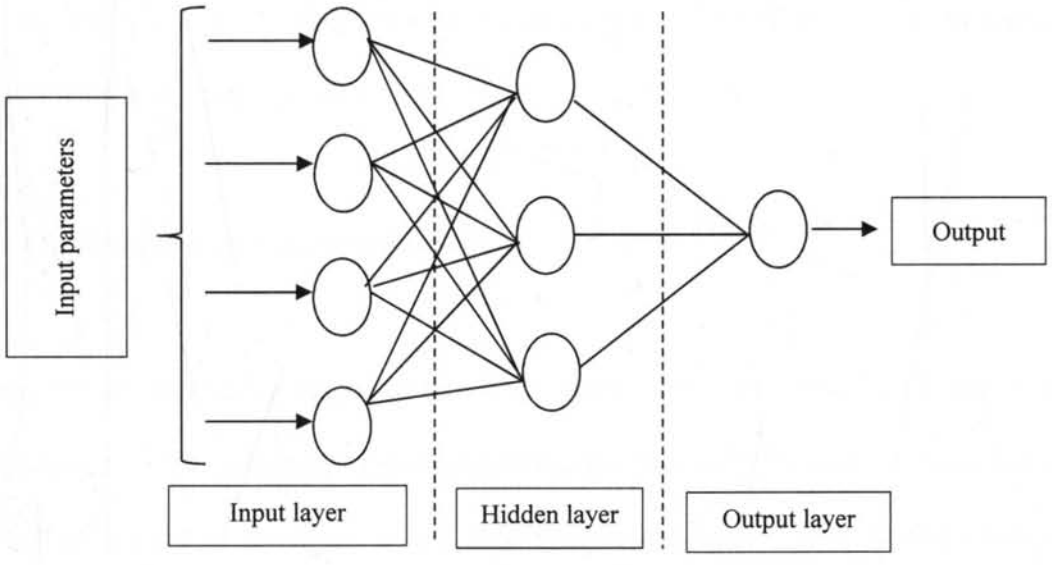



Figure 2.3. Network structure of MLP-FF-ANN model

According to Saptoro (2010), the way the dataset was distributed into the training, testing and validation sets were said to have a great effect on the predictive ability of the system. Unfortunately, there is no specific guidance on how these datasets ought to be split. However, Chaloulakou et al. (2003) suggested that the training set should involve two-thirds of the data. The large dataset was chosen for the prediction model to ensure that the data was representative. To attain a high level of generalisation, large representative datasets are necessary (Alimissis et al., 2018). Therefore, in this study, the datasets were divided into unequal subsets, namely training (67%), testing (16%) and validation (17%) to run within the ANN model.

The dividing was randomly done to certify that each parameter of the subsets represents the whole dataset (Cabaneros et al., 2017). Large and representative datasets in the training subset cause the network to be effectively developed due to its possibility to simplify the new data present. Meanwhile, the validation set is important in the training procedures since it is implemented as the technique of early-stopping to elude the overfitting of the network in the data. In contrast, the purposed testing data is only for the assessment of the network's generalisation ability. Therefore, in general, only the training data are considered for the assessment of the prediction model instead of the testing and validation data.

Data transformation and activation function

During the pre-modelling stages of the ANN model, the transformation of data is an important issue. The selection of data transformation procedure depends exclusively on the sort of activation functions in the hidden layer (Saptoro, 2010). In this study, a hyperbolic tangent function, which was bounded in the range of -1 to 1 was chosen

since this transfer function resulted in smaller errors of prediction than those of the more frequently used, for example, the sigmoid (logistic) transfer function. The equation of the hyperbolic tangent function used in this study is as shown below:

$$Hyperbolic\ tangent\ function = \frac{e^{2x}-1}{e^{2x}+1} \tag{2.5}$$

Training algorithm

In this study, the Levenberg-Marquardt algorithm was used for the network training. This training algorithm showed faster intersection and could discover better minima of error. Besides, this standard technique was adopted in numerous fields or disciplines (Abdullah et al., 2016).

Number of hidden nodes and hidden layers

A number of hidden nodes and layers are important since it will affect the model complexity, besides its predictive capability. It is important to determine the number of nodes. According to He et al. (2015), a high number of hidden nodes contributed to overfitting, while less number of hidden nodes contributed to the inadequate information capture. Therefore, the proposed equation was applied in this analysis to have a better model without overfitting and insufficient information. The number of hidden nodes considered for this analysis was based on the proposed equation by Fletcher and Goss (1993) to determine the appropriate number of nodes ranges as shown below:

$$Number\ of\ nodes\ ranges = 2S^{1/2} - 0\ \ to\ \ 2S - 1 \tag{2.6}$$

Where,

S = nodes input number

O = nodes output number

Meanwhile, for a number of hidden layers, the results do not show much better with the multi hidden layer. In contrast, one hidden layer show is sufficient as the function approximator (Saptoro, 2010). Besides, theoretical studies by Hornik et al. (1989) showed that ANN is able to estimate a non-linear function by using only one hidden layer. This was proven by a research done by Chaloulakou et al. (2003), which found that there was no improvement when two hidden layers were applied in the networks. In addition, Abderrahim et al. (2016) mentioned that the hidden layer number is an important factor, while the related hidden layer number is not so powerful.

### 2.7.2 Multiple Linear Regression (MLR)

MLR is broadly utilised in atmospheric modelling. This technique has been used for examining the relation among numerous independent and dependent parameters by fitting a linear equation for data observation. In this study, it was utilised to justify the relation between the parameters of air quality and $PM_{10}$ data. The linear regression was simplified by the model, whereby each independent parameter value was associated with a dependent parameter value. The equation of the model is as shown below:

$$y_i = \beta_O + \beta_{1i}X_{1i} + \cdots + \beta_k X_{ki} + \varepsilon_i \quad (2.7)$$

where,

i = 1,…,n

$\beta_1$ = Regression coefficient

$X_1$= Independent variable and

$\varepsilon$ = Error associated with the regression

### 2.7.3 Decision Tree (DT)

The DT is produced through a recursive partitioning algorithm, which decides the rules related with each binary split in the tree. At first, all the data are used to decide the rule, which divide the data into the two most dissimilar sets. By minimising the node impurity measure (Gini index), the best splitting criteria for each node is found. Following the initial rule, each subset is recursively divided until a large tree is developed, with each terminal node containing only a few observations.

The oversized tree is then pruned until the generalisation performance, described by either evaluating the error rate on a test set or by cross-validation, which indicate no overfitting. Decision trees readily lend themselves to be displayed graphically, making them easier to interpret. Furthermore, a DT is often unable to identify the interactive effects of multiple inputs (Zickus et al., 2002).

### 2.7.4 Multivariate Adaptive Regression Splines (MARS)

The MARS is a multivariate nonparametric regression procedure first proposed by Friedman (1991), which approximates complex relations by a series of linear regressions (splines) on different intervals of the independent parameters. MARS uses the so-called basis functions, which re-express independent parameters by mapping them to new parameters. It is possible to approximate any functional shape by mixing the different types of basic functions and providing adequate values for knots. MARS is a very flexible model building and data analysis tool as it can adapt any functional form. It is suitable for predicting outcomes and for exploratory data analysis.

### 2.7.5 Support Vector Machine (SVM)

The SVM technique is based on the structured risk minimisation principle, which aims to minimise the upper bound of the generalisation error (Pai et al., 2010). This governs the SVM with a greater potential to regress the input-output relation during its training phase, and to obtain good performance for new input data (Chen, 2011). For regression, the SVM maps the input data x into a high-dimensional feature space F by non-linear mapping, to yield and solve a linear regression problem in this feature space.

## 2.8 Prediction models

In recent years, humans, buildings, crops and ecosystems were affected by the major issue of air pollution. To ensure the presence of air pollution does not pose a threat to public health, it becomes a concern for the government to control the air pollution level. Air pollution in some air monitoring stations shows high-level pollutant concentration that sometimes exceeds the Air Quality Guidelines. The mortality will decrease with the highly concern of air pollution. Therefore, to overcome the problem, a prediction model of $PM_{10}$ as a major pollutant in the atmosphere was implemented to overcome respiratory health problems and loss of welfare.

### 2.8.1 Pattern of missing data

In the real world, missing data is unavoidable. After implementing data mining techniques to the real-world data, there are frequent presences of missing data for some parameters. Almost each real-world dataset consists of missing data as it is omnipresent. There are a variety of reasons that contribute to missing data. Commonly, due to human and equipment and environmental condition errors that consequence in noise and propagation distortion. Eventually, caused missed intercepted signals and

incomplete data (Jordanov et al., 2018). Chen et al. (2016) and Noor and Zainudin (2008) also mentioned that equipment failure is one of the reasons for missing data presences. For certain dataset applications, missing data can be particularly unfavourable mainly when the missing data are not uniformly distributed and there is an unidentified potential mechanism that could enlighten the missing values (Garciarena & Santana, 2017). A lot of voids contribute to instability in the imputations (Gómez-Carracedo et al., 2014). Therefore, it is important to deal with the missing data in order to improve accuracy in the task of machine learning (Jordanov et al., 2018). Imputation methods are the most used to manage missing data.

Influences of the imputation method on the performance are very reliant on the type of missing data. Inappropriate selection of imputation method can create a low classification quality on the data due to the biases (Garciarena & Santana, 2017). Owing to this situation, a missing value can be calculated by using suitable imputation methods related to it. The best imputation methods can be applied based on the appropriate type of missing data that implements the imputed data. Generally, missing data mechanism can be divided into three types, namely, Missing Completely at Random (MCAR), Missing at Random (MAR) and Missing Not at Random (MNAR) (Garciarena & Santana, 2017; Jordanov et al., 2018).

For MCAR, there is no pattern recognised as failures for the database measurement. Commonly, there is a variety of reasons that lead to the MCAR. One of them is probably due to the loss of data during the transfer of information (Garciarena & Santana, 2017; Jordanov et al., 2018). This type of missing data gives impact to the classification algorithm as it is influenced by the distribution of missing data over the data. It has fewer biases when presented in the database if the distribution of the

missing data is more uniform. According to the Jordanov et al. (2018), the missingness has no relation between observed and unobserved parameters.

Meanwhile, for MAR there is a particular pattern recognised, whereby observations with missing values present in the prevalent parameter can be discovered. For MAR, it is easier to conclude the cause of missing data by learning other dataset parameters (Garciarena & Santana, 2017). Therefore, it is easier to assume its origin through the dataset variables and missingness has a relation between observed and unobserved parameters, whereby missingness can be fully considered by the complete parameters (Jordanov et al., 2018).

For MNAR type of missing data, this type of missing data is alike MAR. However, the values that affect others to be missing are not recognised. This type of missing data might due to the parameters that take a value out of its representation range, and probably the parameters were not observed (Garciarena & Santana, 2017). According to Jordanov et al. (2018), MNAR gives a much more difficult condition and biased estimates as compared to MCAR and MAR.

Recognising any pattern in the missing data is a key feature when considering techniques to manage missing observations. Besides, prediction quality is directly impacted by the type of missing data from the classification methods applied to the data (Garciarena & Santana, 2017). According to Arroyo et al. (2018), in air quality data, the mechanism of missing data is generally random, which is known as missing at random (MAR). Besides, Ramli et al. (2013) mentioned that MAR is the ultimate method for multiple imputations. Obviously, the performance of imputation is not only influenced by the amount of missing data but also on the features of missing data

patterns. It is important to differentiate the pattern of missing data, which defines the observed and missing values, mechanism of missing data as well as the relationship between the value of parameters and missingness in the dataset.

### 2.8.2 Various studies on $PM_{10}$ prediction model based on the models used for this thesis

This part depicts in detail the studies that were done on $PM_{10}$ prediction model based on the models utilised in this thesis. Table 2.6 involves the various studies with a short description of the results obtained in regard to the study done in Malaysia as well as in other countries. This review demonstrates the usefulness of the methods used in this work to determine the appropriate methods for the $PM_{10}$ prediction model.

Table 2.6: Reviews on $PM_{10}$ prediction model study in (a) Malaysia and (b) other countries

(a)

| Study | Site | Method | Inputs | Results |
|---|---|---|---|---|
| Ng & Awang (2018) | Seberang Jaya, Petaling Jaya, Sungai Petani, Seremban, Batu Muda, (Malaysia) | MLR | CO, $O_3$, humidity, temperature, wind speed, wind direction | Reveals that including lagged $PM_{10}$ concentration increased the precision of prediction. Prediction model of MLR with and without lagged $PM_{10}$ concentration gsve RMSE values of 7.754 and 8.711, respectively. Low RMSE value showed a high accuracy of the model. Thus, including the lagged concentration of $PM_{10}$ gave high accuracy as compared to model that excluded lagged concentration of $PM_{10}$. |
| Nazif et al. (2018) | Kuching, Balok Baru, (Malaysia) | MLR<br><br>PCR (PCA and MLR) | MLR<br>Humidity, temperature, wind speed, wind direction and previous-day concentration of $PM_{10}$ (2006 -2010)<br>PCR<br>Kuching.<br>PC 1 (temperature, humidity);<br>PC 2 (previous concentration of $PM_{10}$);<br>PC 3 (wind speed).<br>Balok Baru<br>PC 1 (temperature, humidity);<br>PC 2 (wind direction) | They concluded that a hybrid model, which was a combination between the PCA and MLR improved predictability models and gave better results.<br><br>The $R^2$ values obtained for MLR analysis for Kuching and Balok Baru were 0.84 and 0.64, respectively. While the $R^2$ values obtained for PCR (combination between PCA and MLR) analysis for Kuching and Balok Baru were 0.93 and 0.90, respectively.<br><br>Thus, PCR showed better result compared to MLR alone. However, both models showed significant values with $R^2$ values greater than 0.5. |
| Ul-saufie et al. (2013) | Nilai (Malaysia) | MLR<br><br>ANN<br><br>PCA-MLR<br><br>PCA-ANN | MLR & ANN (FFBP)<br>$SO_2$, $NO_2$, CO, the previous-day concentration of $PM_{10}$, humidity, temperature, wind speed (2003-2010)<br>PCA-MLR & PCA-ANN<br>PC 1 ($PM_{10}$, CO, $NO_2$);<br>PC 2 (humidity, wind speed, | They did improvement by applying PCA onto both prediction models for predicting next day, next two-day and next three-day prediction model of $PM_{10}$ concentration.<br>The $R^2$ values obtained for MLR, PCA-MLR, ANN and PCA-ANN for the next day were 0.6246, 0.7595, 0.6358, 0.7812; for the next two days were 0.4232, 0.6358, 0.4451, 0.5904; for the next three days were 0.3508, 0.5998, 0.3578 and 0.4551 respectively. |

| | | | | |
|---|---|---|---|---|
| | | | temperature); PC 3 ($SO_2$) | This study revealed that PCA-ANN is the best to be applied among others, for a next day prediction model of $PM_{10}$ while PCA-MLR is the best to be applied among others for next two days and next three-day prediction model of $PM_{10}$. |
| Afzali et al. (2014) | Pasir Gudang, Johor, (Malaysia) | MLR<br>ANN<br>(FFNN) | Wind speed,<br>Temperature,<br>Solar radiation,<br>Humidity rate | The value of $R^2$ obtained when excluded lagged $PM_{10}$ as input was 0.18, while the accuracy increased when the lagged concentration of $PM_{10}$ was being used as the input with $R^2$ obtained at 0.47 for MLR. While ANN gives high accuracy with $R^2$ value at 0.69. |
| Azid et al. (2014) | Pasir Gudang, Kuching, Bukit Rambai, Tasek, Nilai, Klang, Balok Baru, Pengkalan Chepa, Paka, Labuan, (Malaysia) | ANN<br>(MLP-FF) | All parameters<br>CO, $O_3$, $PM_{10}$, $SO_2$ $NO_2$, $CH_4$, NMHC THC<br>Input after PCA<br>$CH_4$, NMHC, THC, $O_3$, and $PM_{10}$. | Five parameters namely $CH_4$, NMHC, THC, $O_3$, and $PM_{10}$ were obtained from the rotated PCA in the study area<br><br>Model A (used all parameters) gave $R^2 = 0.615$,<br>Model B (obtained from rotated PCA) gave $R^2 = 0.618$<br><br>Reduction of predictor parameters obtained from factor scores through varimax rotation was more efficient and effective owing to the no losing important information.<br><br>The combination of ANN and factor scores after varimax rotation has been proven as useful tools in modelling of air pollution. |
| AhmadIsiyaka et al. (2014) | Kemaman, Paka-Kertih, Kuala Terengganu, Putra Jaya, Kuala Lumpur (Malaysia) | ANN<br>(MLP-FF)<br><br>MLR | CO, $O_3$, $PM_{10}$, $SO_2$, $NO_2$, ambient temperature and wind speed | ANN gave a better prediction compared with MLR<br><br>ANN ($R^2 = 0.93$, RMSE = 4.87)<br>MLR ($R^2 = 0.70$, RMSE = 9.57) |
| Hua (2018) | Klang, Petaling Jaya, Shah Alam, Kuala Selangor, Putrajaya, Cheras, Batu Muda, Banting (Malaysia) | MLR | $NO_2$, $O_3$, $SO_2$, CO, $PM_{10}$<br>PCA-MLR<br>LPS (CO, $PM_{10}$, $SO_2$, $NO_2$)<br>MPS (CO, $NO_2$, $SO_2$, $O_3$)<br>SHPS ($O_3$, CO, $NO_2$, $PM_{10}$) | $R^2$ values obtained for LPS, MPS and SHPS were 0.837, 0.878, 0.894, while RMSE values obtained for LPS, MPS and SHPS were 3.324, 2.829, 1.808, respectively. |
| Abdullah et al. | SK Pusat Chabang Tiga | MLR | Previous day $PM_{10}$, | The MLR models for NEM, Inter-Monsoon 1, SWM and Inter- |

| | | | |
|---|---|---|---|
| ($2017_a$) | (Malaysia) | temperature, humidity, wind speed, Mean Sea Level Pressure and rainfall amount | Monsoon 2 disclosed $R^2$ of 0.68, 0.58, 0.57, and 0.63, respectively.<br><br>In conclusion, the developed MLR models are appropriate for forecasting $PM_{10}$ concentrations at the local level for each monsoon |

(b)

Table 2.6: Continued

| Study | Site | Method | Inputs | Results |
|---|---|---|---|---|
| Sarwat & El-Shanshoury (2018) | Ain Sokhna city, Egypt | MLR<br><br>ANN<br><br>PCA-MLR<br><br>PCA-ANN | <u>MLR & ANN</u><br>CO, NO, $NO_2$, $NO_X$, $PM_{10}$, $SO_2$, wind direction, wind speed, temperature, sigma theta<br><br><u>PCA-MLR & PCA-ANN</u><br>PC 1 ($NO_x$, $NO_2$, NO, $SO_2$)<br>PC 2 (sigma theta, wind direction, temperature)<br>PC 3 (CO)<br>PC 4 ($PM_{10}$, wind speed) | They found out that ANN using original parameters as input performed better than MLR.<br><br>Input parameters obtained from PCA-ANN with varimax rotation showed a higher accuracy compared to the PCA-MLR.<br><br>In contrast, the accuracy was reduced when PCA by using equamax and quartimax was used for both prediction models.<br><br>This shows that PCA with varimax rotation is the best when applied together with ANN.<br><br>This indicates that the ANN method can be applied successfully as a tool for decision-making and problem-solving for better atmospheric management. |
| Nidzgorska-Lencewicz (2018) | Poland | ANN | Temperature, humidity, air pressure, wind speed, $PM_{10}$<br><br>Winter season (December, January, February) 2002/2003 until 2016/2017 | In their study, ANN was used to predict levels 1 to 6 hours ahead concentrations of $PM_{10}$.<br><br>This study shows the satisfactory result of ANN with $R^2$, index of agreement (IA) and RMSE obtained were from 0.452 to 0.848, 0.693 to 0.957 and 8.80 to 23.56, respectively.<br><br>This study also shows that ANN is capable to predict 1 h ahead concentration of $PM_{10}$ rather than 6 h ahead. |
| Pawul & Śliwka (2016) | Poland | ANN (MLP) | Maximum, minimum and average temperature, | Three-layer perception with back-propagation algorithm gave the best results |

| | | | | |
|---|---|---|---|---|
| | | | average wind speed, the average temperature of the previous day, average daily concentrations of particulate matter $PM_{10}$ of the previous day. (2014 and 2015 at three measuring stations) | Krasinski Avenue Nowa ($R^2 = 0.908$)<br>Nowa Huta ($R^2 = 0.921$)<br>Kraków-Kurdwanów ($R^2 = 0.933$)<br><br>The differences in distributions between the expected value and the predicted value were similar for each measurement station. The performances of the MLP networks were satisfactory. |
| Baawain & Al-Serihi (2014) | Oman | ANN (MLP-BP) | WS, WD, T, H | It is best performed using a hidden layer with 47 hidden neuron numbers ($R^2 = 0.82$) |
| Özdemir & Taner (2014) | Turkey | MLR<br>ANN (BP) | Temperature, humidity, air pressure, wind speed, wind direction | Urban ($R^2 = 0.74$), Industrial ($R^2 = 0.36$)<br><br>Backpropagation feedforward network type with two hidden layers (showed the best)<br><br>ANN results had higher $R^2$ values than MLR.<br>ANN – Urban ($R^2 = 0.87$), Industrial ($R^2 = 0.49$)<br>ANN – Urban ($R^2 = 0.74$), Industrial ($R^2 = 0.36$)<br><br>In the ANN model, backpropagation network with two hidden layers achieved the best prediction efficiency determination. |
| Cortina–Januchs et al. (2015) | Mexico | MLR<br>ANN (MLP) | $PM_{10}$, wind speed, wind direction, temperature, relative humidity | $R^2$ values obtained for Cruz Raja, Nativitas and Dif were 0.53, 0.65, 0.43 for MLR, while $R^2$ values obtained for Cruz Raja, Nativitas and Dif were 0.64, 0.72, 0.51, respectively, for ANN. |

Although Table 2.6 infers that MLR was broadly utilised for $PM_{10}$ prediction, the MLR results showed weakness as compared to the ANN. The ability of ANN to simulate complex problems with non-linear behaviour, besides the incorporation of many heterogeneous parameters, implementation speed is one of the advantages for the ANN to perform well as compared to MLR, which is more effective to be applied onto linear behaviour.

For the selection of input parameters, PCA is often used. Some study applied the hybrid method by combining PCA with the prediction models (MLR and PCA). Most of them found the best accuracy when applying the hybrid method. Besides, utilisation of previous or lagged $PM_{10}$ concentration as an input parameter can significantly improve the prediction results. In general, ANN performs better than MLR in predicting the $PM_{10}$ concentration. Different studies showed different types of input parameters used as well as the number of hidden numbers. Therefore, no specified structure can be applied to the models since according to Ding et al. (2016), the mechanism of the model is not limited by the data structure. However, it is robust because of its non-parametric nature. Most recently, pollutants prediction of Multi-Layer Perceptron (MLP) has been wide. Good results for different pollutants were obtained when this method was implemented. Several studies implemented this method and it performed better than MLR (Esfandani & Nematzadeh, 2016).

## 2.9 Summary

Among these methods, ANN and MLR are widely used for the $PM_{10}$ prediction model. However, ANN has the most complex mathematical structure and can simulate human learning and pattern recognition, which extracts information from imprecise and non-linear data. Overall, this chapter has covered the overview of air pollution in the

atmosphere and characteristics of the five main air pollutants ($PM_{10}$, $SO_2$, $NO_2$, CO, $O_3$). This chapter also discussed the sources and impacts of air pollution. To reduce the air pollution problem and up-to-date status of air pollution, Malaysia has implemented the Malaysia Ambient Air Quality Standard (MAQS) and air pollution index (API). This chapter also discussed the missing data pattern that is usually a problem for the prediction model due to incomplete data. In the end, this chapter deliberated the various studies conducted on $PM_{10}$ prediction model based on the models used in this study.

# CHAPTER 3

# METHODOLOGY

## 3.1 Introduction

This chapter presents a comprehensive description of the research methodology implemented to achieve the objectives and answer the research questions. Section 3.2 discusses the study areas, while Section 3.3 gives an explanation about research data, comprising data collection and data pretreatment. This section discusses details about the instruments used and also the imputation method suitability to be applied to the missing data. The transformation of data to be applied for the next prediction model is also discussed. Section 3.4 describes chemometric analysis that consists of AHC, DA and PCA applied in the study. Meanwhile, Section 3.5 gives an explanation about the evaluation of each performance model. Section 3.6 discusses steps analysis of modelling applied for $PM_{10}$ prediction using ANN and MLR. Lastly, Section 3.7 discusses about statistical analysis tools implemented in the study.

## 3.2 Study area

Malaysia is a coastal nation encompassed by Thailand and Indonesia in the northwestern and southern parts, respectively, while Brunei and the South China Sea are on the eastern part. It comprises of west and east Malaysia with one-third of the Borneo Island. These two regions are divided by 640 miles of the South China Sea and cover a total area of 330, 252 square kilometers ($km^2$). Each year, Malaysia encounters a moderately uniform temperature that range from 26 °C to 28°C. Besides, Malaysia also experience monsoon seasons, such as Northeast monsoon and Southwest monsoon, which occur from November to March and from May to September,

respectively. However, natural disasters, for example, volcanic eruption and typhoon do not occur in Malaysia.

There are 52 continuous air monitoring stations in Malaysia. The coordinates, station ID, stations, state and classification of the stations based on area are shown in Table 3.1. However, only 51 stations were considered in the analysis since one station, namely ILP Miri, has a high percentage of missing data that will affect the analysis of the $PM_{10}$ prediction model. Figure 3.1 visualises the locality of continuous air quality monitoring stations located in Malaysia.

Table 3.1. Air quality monitoring stations located in Malaysia

| Station No. | Lat. (N) | Long. (E) | Station ID | Stations | State | Classification |
|---|---|---|---|---|---|---|
| 1 | N04°16.260 | E103°25.826 | CA0002 | Sek.Ren.Bukit Kuang, Teluk Kalung, Kemaman | Terengganu | Industrial |
| 2 | N04°35.880 | E103°26.096 | CA0024 | Kuarters TNB, Paka-Kerteh | Terengganu | Industrial |
| 3 | N05°18.455 | E103°07.213 | CA0034 | Sek.Keb.Chabang Tiga, Kuala Terengganu | Terengganu | Urban |
| 4 | N03°00.602 | E101°24.484 | CA0011 | Sek.Men.Perempuan Raja Zarina, Klang | Selangor | Urban |
| 5 | N03°06.612 | E101°42.274 | CA0016 | SK Bandar Utama Damansara, Petaling Jaya | Selangor | Industrial |
| 6 | N03°06.286 | E101°33.367 | CA0025 | Sek.Ren. Keb. TTDI Jaya, Shah Alam | Selangor | Urban |
| 7 | N03°19.592 | E101°15.532 | CA0048 | Sek.Men.Sains Kuala Selangor | Selangor | Sub Urban |
| 8 | N02°49.001 | E101°37.381 | CA0060 | Kolej Mara Banting, Selangor | Selangor | Urban |
| 9 | N04°45.529 | E115°00.813 | CA0031 | Dewan Suarah, Limbang | Sarawak | Sub Urban |
| 10 | N01°27.308 | E110°29.498 | CA0035 | Pejabat Perumahan Kota Samarahan | Sarawak | Rural/Background |
| 11 | N01°14.425 | E111°27.629 | CA0036 | Pejabat Perumahan Sri Aman | Sarawak | Sub Urban |
| 12 | N02°00.875 | E112°55.640 | CA0055 | Stadium Tertutup Kapit | Sarawak | Rural/Background |
| 13 | N04°08.447 | E114°01.247 | CA0045 | ILP Miri | Sarawak | Rural/Background |
| 14 | N01°33.734 | E110°23.329 | CA0004 | Medical Store, Kuching | Sarawak | Industrial |
| 15 | N02°18.856 | E111°49.906 | CA0026 | Ibu Pejabat Polis Sibu | Sarawak | Sub Urban |
| 16 | N03°10.587 | E113°02.433 | CA0027 | Balai Polis Pusat Bintulu | Sarawak | Sub Urban |
| 17 | N04°25.456 | E114°00.731 | CA0028 | Sek.Men.Dato Permaisuri, Miri | Sarawak | Sub Urban |
| 18 | N02°07.992 | E111°31.351 | CA0029 | Balai Polis Pusat Sarikei | Sarawak | Sub Urban |
| 19 | N05°53.623 | E116°02.596 | CA0030 | Sek.Men.Keb.Putatan,Tg.Aru, Kota Kinabalu | Sabah | Urban |
| 20 | N04°15.016 | E117°56.166 | CA0039 | Pejabat JKR, Tawau | Sabah | Urban |
| 21 | N05°20.313 | E116°09.769 | CA0049 | Sek.Men.Keb.Gunsanad, Keningau | Sabah | Sub Urban |
| 22 | N05°51.865 | E118°05.479 | CA0050 | Pejabat JKR, Sandakan | Sabah | Sub Urban |
| 23 | N04°37.781 | E101°06.964 | CA0008 | Sek.Men.Jalan Tasik Ipoh | Perak | Industrial |
| 24 | N04°53.940 | E100°40.782 | CA0020 | Sek.Ren.Keb. Air Puteh, Taiping | Perak | Industrial |
| 25 | N04°12.038 | E100°39.841 | CA0041 | Pejabat Pentadbiran Daerah Manjung | Perak | Sub Urban |
| 26 | N03°41.267 | E101°31.466 | CA0045 | UPSI, Tanjung Malim | Perak | Sub Urban |
| 27 | N04°33.155 | E101°04.856 | CA0046 | Sek.Men.Pegoh, Ipoh | Perak | Urban |

Table 3.1: Continued

| Station No. | Lat. (N) | Long. (E) | Station ID | Stations | State | Classification |
|---|---|---|---|---|---|---|
| 28 | N03°58.238 | E102°20.863 | CA0007 | Pejabat Kajicuaca, Batu Embun, Jerantut | Pahang | Rural/Background |
| 29 | N03°49.138 | E103°17.817 | CA0014 | Sek.Keb.Indera Mahkota, Kuantan | Pahang | Sub Urban |
| 30 | N03°57.726 | E103°22.955 | CA0015 | Sek.Keb.Balok Baru, Kuantan | Pahang | Industrial |
| 31 | N05°23.470 | E100°23.213 | CA0003 | Sek. Keb.Cenderawasih, Tmn Inderawasih, Prai | Pulau Pinang | Industrial |
| 32 | N05°23.890 | E100°24.194 | CA0009 | Sek.Keb.Seberang Jaya 2, Perai | Pulau Pinang | Sub Urban |
| 33 | N05°21.528 | E100°17.864 | CA0038 | Universiti Sains Malaysia | Pulau Pinang | Sub Urban |
| 34 | N02°49.246 | E101°48.877 | CA0010 | Taman Semarak (Phase 2), Nilai | Negeri Sembilan | Industrial |
| 35 | N02°43.418 | E101°58.105 | CA0047 | Sek.Men.Teknik Tuanku Jaafar, Seremban | Negeri Sembilan | Urban |
| 36 | N02°26.458 | E101°51.956 | CA0056 | Pusat Sumber Pendidikan N.S. Port Dickson | Negeri Sembilan | Urban |
| 37 | N02°15.510 | E102°10.364 | CA0006 | Sek.Men.Kebangsaan Bukit Rambai, Melaka | Melaka | Industrial |
| 38 | N02°12.789 | E102°14.055 | CA0043 | Sekolah Tinggi Melaka, Melaka | Melaka | Urban |
| 39 | N06°09.520 | E102°15.059 | CA0022 | Sek. Men. Tanjung Chat, Kota Bahru | Kelantan | Urban |
| 40 | N05°48.671 | E102°08.000 | CA0059 | Sek. Men. Keb. Tanah Merah | Kelantan | Industrial |
| 41 | N05°37.886 | E100°28.189 | CA0017 | Sek. Keb. Bakar Arang, Sungai Petani | Kedah | Sub Urban |
| 42 | N06°19.903 | E099°51.517 | CA0032 | Kompleks Sukan Langkawi | Kedah | Sub Urban |
| 43 | N06°08.218 | E100°20.880 | CA0040 | Sek.Men.Agama Mergong, Alor Setar | Kedah | Urban |
| 44 | N06°25.424 | E100°11.046 | CA0033 | Institut Latihan Perindustrian (ILP) Kangar, Perlis | Perlis | Sub Urban |
| 45 | N01°28.225 | E103°53.637 | CA0001 | Sek.Men.Pasir Gudang 2 | Johor | Industrial |
| 46 | N01°29.815 | E103°43.617 | CA0019 | Inst. Perguruan Malaysia Temenggong Ibrahim, Larkin, Johor | Johor | Industrial |
| 47 | N02°03.715 | E102°35.587 | CA0044 | Sek. Men. Teknik, Muar | Johor | Sub Urban |
| 48 | N01°33.500 | E104°13.310 | CA0057 | Sek. Men. Agama Bandar Penawar, Kota Tinggi | Johor | Urban |
| 49 | N05°19.980 | E115°14.315 | CA0042 | Taman Perumahan MPL | Labuan | Sub Urban |
| 50 | N02°55.915 | E101°40.909 | CA0053 | Sek. Keb. Putrajaya 8(2), Jln P8/E2, Presint 8, Putrajaya | Kuala Lumpur | Urban |
| 51 | N03°06.376 | E101°43.072 | CA0054 | Sek. Men. Keb. Sri Permasuri, Cheras | Kuala Lumpur | Urban |
| 52 | N03°12.748 | E101°40.929 | CA0058 | Sek. Keb. Batu Muda, Kuala Lumpur | Kuala Lumpur | Urban |

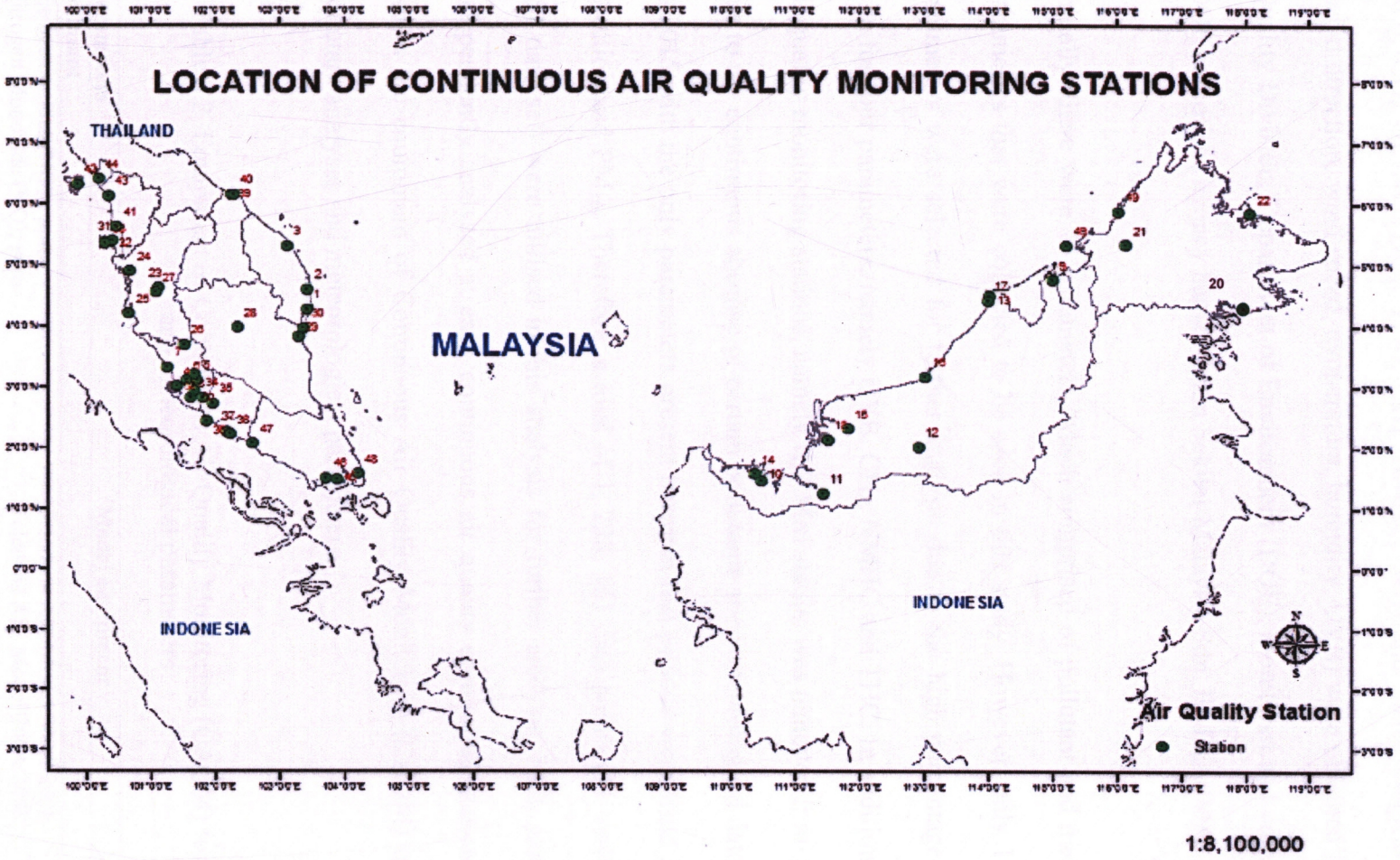



Figure 3.1. Continuous air quality monitoring stations located in Malaysia (DOE, $2015_a$)

### 3.3 Data collection

Six years of data from January 1, 2010 to December 31, 2015 for air pollutants ($PM_{10}$, $NO_2$, CO, $O_3$, NO, $NO_x$, THC, $CH_4$, NMHC, $SO_2$) and meteorological parameters (wind direction, wind speed, temperature, humidity, UVB) were obtained from the Air Quality Division, Department of Environment (DOE), monitored and collected by the DOE authorised agency named Alam Sekitar Malaysia Sdn. Bhd. (ASMA).

Initially, there were 15 parameters, which comprised of pollutant and meteorological parameters that were collected to be used in this study. However, only 11 out of 15 parameters were selected for further analysis due to the high percentage of missing data for four parameters, namely UVB, $CH_4$, NMHC and THC. In addition, one of the air quality monitoring stations, namely ILP Miri station was removed from the analysis due to the continuous absence of certain pollutants and meteorological data from 2012 to 2015 with the only parameters present between that periods were wind speed, wind direction and $PM_{10}$. Therefore, a total of 1, 228, 161 data points (11 variables × 111 651 data set) were utilised in this analysis for further analysis. Each parameter was independently analysed at each continuous air quality monitoring station. Table 3.2 shows the equipment of Continuous Air Quality Monitoring (CAQM) used for the pollutant analyses and meteorological parameters.

Table 3.2. Equipment of Continuous Air Quality Monitoring (CAQM) for pollutant and meteorological parameters

| Parameters | Model equipment |
|---|---|
| **Pollutant** | |
| Nitrogen oxides ($NO_x$), ppm | Teledyne API Model 200A/200E |
| Nitrogen Monoxide (NO), ppm | Teledyne API Model 200A/200E |
| Sulphur Dioxide ($SO_2$), ppm | Teledyne API Model 100A/100E |
| Nitrogen Dioxide ($NO_2$), ppm | Teledyne API Model 200A/200E |
| Ozone ($O_3$), ppm | Teledyne API Model 400/400E |
| Carbon Monoxide (CO), ppm | Teledyne API Model 300/300E |
| Particulate matter ($PM_{10}$), $\mu g/m^3$ | BAM-1020 Beta Attenuation Mas Monitor |
| Methane ($CH_4$), ppm | Teledyne API M4020 |

| | |
|---|---|
| Non-methane Hydrocarbon (NMHC), ppm | Teledyne API M4020 |
| Total hydrocarbon (THC), ppm | Teledyne API M4020 |
| **Meteorological** | |
| Wind Speed, km/hr | Met One 010C |
| Wind Direction, ° | Met One 010C |
| Temperature, °C | Met One 062 |
| Humidity, % | Met One 083D |
| UVB, $J/m^2h$ | UV radiometer |

According to ASMA (2007), pollutant parameters, namely $SO_2$, CO, $O_3$ and $NO_x$, were measured by using the instruments from the Teledyne Technologies Inc., USA, namely the Teledyne API Model 100A/100E, Teledyne API Model 300/300E, Teledyne API Model 400/400E and Teledyne API Model 200A/200E, respectively. Besides, $NO_x$, pollutants of NO and $NO_2$ were also measured by Teledyne API Model 200A/200E. Meanwhile, the pollutant of $PM_{10}$ was measured by using BAM-1020 Beta Attenuation Mass Monitor from Met One Instrument Inc. USA.

Meanwhile, pollutants, such as THC, $CH_4$, and NMHC were measured by using Teledyne API M4020 from Teledyne Technologies Inc.. Besides pollutants, meteorological parameters were also measured with instruments from Met One Instrument, Inc. For example, wind speed, wind direction, ambient temperature and relative humidity were measured by using Met One 010C sensor, Met One 020C sensor, Met One 062 and Met One 083D sensor, while UV radiation was measured by using UV radiometer.

Measurement of $SO_2$ was based on the UV fluorescence method with a detection range of 0 ppm –20 ppm, meanwhile pollutant of CO was measured by using non-dispersive, infrared absorption (Beer-Lambert) method with 0.5% precision with the lowest detection of 0.04 ppm. Meanwhile, the pollutant of $O_3$ was measured through the UV absorption (Beer-Lambert) method, which is a microprocessor controlled device with a detection limit of 0.4 ppb (Mohammed et al., 2013). In order to detect concentrations

of $NO_x$, NO and $NO_2$ in the ambient air, the chemiluminescent detection principle was applied into the analyser. Meanwhile, for measurement of $PM_{10}$ pollutant, high resolution of 0.1 $\mu g\ m^{-3}$ at a 16.7 L $min^{-1}$ flow rate was used. Lower detection limit for this instrument (BAM-1020 Beta Attenuation Mass Monitor) was as low as 4.8 $\mu g/m^3$ and 1.0 $\mu g/m^3$ for one hour and 24 hours, respectively. Meanwhile, measurement of THC, $CH_4$, and NMHC was equipped with a flame-ionisation detector (FID) with 1% measurement accuracy.

According to Teledyne (2008), pollutant concentration monitoring ability was enhanced through the microprocessor technologies in the analyser, which was used to offer sensibility, stability and simplicity of use for ambient continuous monitoring requirement (DOE, 2010). The concentration of air pollutant and meteorological parameters were recorded automatically based on the value fixed by the United States Environmental Protection Agency (USEPA) standard (Muhamad et al., 2015).

### 3.4 Data pretreatment

To obtain reliable data for the prediction model, the raw data first underwent pretreatment before further analysis. In this study, the raw data underwent missing data treatment since some of the parameters noticed the presence of a high missing data percentage. Besides, data transformation was also implemented to ensure that all the parameters are in the same range since according to Zheng et al. (2018), normalisation was done to eliminate the difference in magnitude. Therefore, there were no dominated parameters in network learning.

### 3.4.1 Outliers

Some parameters showed a high variation between the lowest and highest concentrations. This indicated the complexity of the atmospheric chemical reactions that rely on many aspects comprising their origins, including chemical and physical behaviours in the atmosphere (Behera & Balasubramanian, 2014). According to Burke (1999), outliers might give true measurements and they should not be removed. Besides, the parameters with outlier are common yearly phenomena , especially for the $PM_{10}$ pollutant.

This is coherent with a study done by Abdullah et al. ($2017_a$) who found that there was a high value of outliers for the concentration of $PM_{10}$ due to the biomass burning in Sumatera and Kalimantan, Indonesia during the Southwest monsoon. It is evident from the study done by Campa et al. (2018) who found out that the highest mean of $PM_{10}$ concentration was due to biomass burning as compared to other four sources, namely traffic, dust of mineral and marine aerosol. Besides, Saptoro (2010) mentioned that vital information inside the datasets could be eliminated when the outliers are deleted.

### 3.4.2 Missing data imputation methods

Some of the air quality monitoring stations noted incomplete concentration of pollutants. To overcome this problem, various types of imputation methods (mean, nearest neighbour, EMB) were analysed in this study in order to obtain the best imputation method applied to the data. The procedure to select the best imputation methods on missing data is as shown below:

1) Presence of missing data and patterns of missing data is investigated by modelling missingness map

2) Parameters with missing values are defined from the modelling missingness map to be intended for the imputation model
3) Simulation study was done on 5%, 10%, 15%, 25%, 40% of $PM_{10}$ missing data to determine the appropriate imputation method applied
4) Value of missing data was imputed by giving complete data sets through the selected imputation method (mean, nearest neighbour, EMB)
5) The efficiency of the imputation methods for every percentage of simulated data is described from the performance of the coefficient of determination ($R^2$) and root mean square error (RMSE)
6) Imputed datasets were run by using the best imputation methods (high value of $R^2$ and low value of RMSE)

#### 3.4.2.1 Mean

This imputation method used the mean value of the observed parameter to replace the missing values. It is a fast and simple imputation method. However, it frequently caused biased on the evaluation of parameters and with this obstacles, Dohoo (2015) did not recommend this type of imputation method. On top of that, this imputation method underestimates the dataset with variance. Consequently, may modify some other determined chemometric study (Gómez-Carracedo et al., 2014). Conversely, MCAR pattern of missing data did not give bias results when applying this imputation method (Liu & Gopalakrishnan, 2017). The Equation 3.1 was used for the mean imputation method as shown below:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{i=n} y_i \qquad (3.1)$$

where,

n = number of available data

$y_i$ = data points

### 3.4.2.2 Nearest neighbour

Imputation method of nearest neighbour was implemented by utilising known values at a neighbouring location with endpoints of the gaps used in this imputation method to estimate values of missing data. This imputation method studies the distance between each point and the nearest point to it. According to Junninen et al., (2004) this imputation method is appropriate to be applied onto 3% of missing data. The equation used for the nearest neighbor imputation method is as shown below:

$$y = y_1 \quad \text{if} \quad x \leq x_1 + [(x_2 - x_1) / 2]$$
$$y = y_1 \quad \text{if} \quad x \geq x_1 + [(x_2 - x_1) / 2] \qquad (3.2)$$

where,

y = represents the interpolate

x = time point of the interpolate

$y_1$ and $x_1$ = coordinates of the starting point of the gap

$y_2$ and $x_2$ = the end points of the gaps

### 3.4.2.3 Expectation-Maximisation Based (EMB) Algorithm

Expectation-Maximisation Based (EMB) Algorithm was implemented based on the R platform by using Amelia package (Amelia II). This imputation method implements multiple imputations that generate multiple complete forms of the incomplete dataset. According to Honaker et al. (2011), multiple imputations appeared to increase efficiency and reduce bias, while the mean imputation method can prompt serious

biases. Figure 3.2 shows a schematic diagram of multiple imputations with EMB algorithm. The diagram illustrates that multiple bootstrapped samples of the original incomplete data use the familiar EM (expectation-maximisation) algorithm to draw values of the complete-data parameters. Imputed values from each set of bootstrapped drew from this algorithm will replace the missing values.

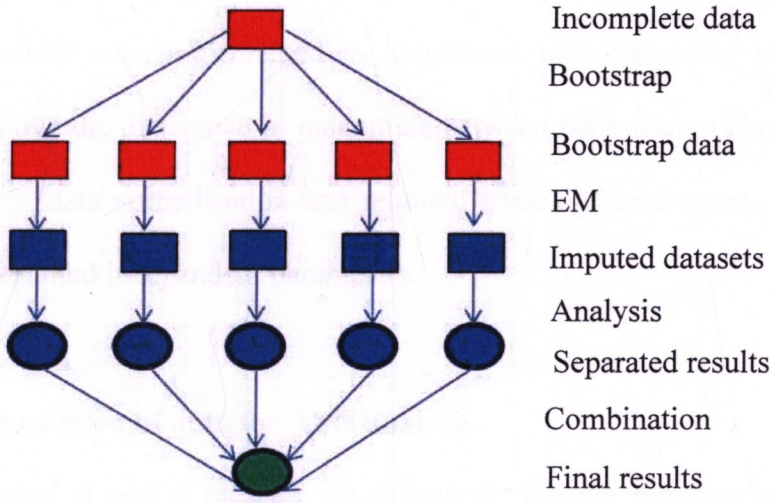



Figure 3.2. Schematic diagram of multiple imputations with EMB algorithm

The basic idea for procedure of multiple imputations as imputation methods on missing data is explained below:

Imputation: During this first step, the uncertainty of this model is reflected by the sets of plausible values, which can be used M times to generate M 'completed' data sets. At this stage, bootstrap data stimulate estimation uncertainty from the incomplete data approaching with bootstrap.

Analysis: Each M dataset was performed by the desired analysis by using standard complete data methods, whereby EM algorithm was run to learn the posterior mode for

the bootstrapped data, which gave necessary uncertainty of the EMB algorithm for details.

Combination: The results were combined during this final step allowed the uncertainty about the imputation to be considered.

### 3.4.3 Data transformation

The normalisation method was implemented to preprocess each parameter before further analysis to remove the difference in magnitude between parameters (Zheng et al., 2018). In this study, data normalisation was required since different ranges were noticed for the dependent and independent parameters.

#### 3.4.3.1 Transformation of data for ANN analysis

The data scales changeto -1 and 1 through the hyperbolic tangent function of the transfer function in the hidden layer. According to Asadollahfardi et al. (2016), the transformation of data is important to avoid network saturation. The equation of hyperbolic tangent function mentioned in this study is as follows:

$$\text{The hyperbolic tangent function} = \frac{e^{2x}-1}{e^{2x}+1} \tag{3.3}$$

where,

x is the linear combination of the X variables

#### 3.4.3.2 Transformation of data for MLR analysis

The scaled data within the range of -1 to 1 is produced from the normalisation procedure. This does not introduce bias since it maintains exactly all relations in the data. This scaling is appropriate for enhancing the precision of numeric calculation by

the MLR models for the better outputs by using the min-max system. The data normalisation is obtained by the following transformation (Abdullah et al., 2017a):

$$z_i = \frac{x_i - x_{min}}{x_{max} - x_{min}} \quad (3.4)$$

where,

$x = (x_1, \ldots, x_n)$

$z_i$ = normalised data

## 3.5 Chemometric analysis

Chemometric techniques, otherwise known as analysis of multivariate techniques, are outstanding tools that are frequently used in the field of environment to recognise the pollution sources. Besides, they are also useful for the analysis of databases for better understanding of the air quality status condition in a particular region (Hua, 2018). Chemometric techniques can evade the misinterpretation of large and complex environmental monitoring data (Simeonov et al., 2002). Chemometric techniques are widely applied used in many scientific studies, for instance, Agglomerative Hierarchical Cluster (AHC), Discriminant Analysis (DA) and Principal Component Analysis (PCA).

### 3.5.1 Agglomerative Hierarchical Cluster (AHC)

AHC is used to structure homogeneous groups of classes based on the parameter description. In this study, the dissimilarity coefficient used was the Euclidean distance. Based on the minimum distance or the nearest neighbour rule, the Euclidean distance was calculated on the basis of the single linkage method. Distances between two clusters were defined from this measure as the minimum distance was found between

the first cluster and second cluster (Yim & Ramdeen, 2015). The equation used for the Euclidean distance is as shown below:

$$(D_{link} / D_{max}) \times 100 \quad (3.5)$$

where,

$D_{link}$ = linkage distance

$D_{max}$ = maximal distance

The quotient, which is symbolised by the y-axis is multiplied by 100 to standardise the linkage distance (Shrestha & Kazama, 2007). In clustering analysis, the dissimilarity between groups of $PM_{10}$ concentration was calculated and summarised by using Ward's method. The Ward's method agglomerate groups of two. Therefore, to remain as homogeneous clusters, inertia within group increases as little as possible. According to Hamid et al. (2018), it has a high probability to merge the smaller cluster if the centres for a pair of clusters are far apart from Ward's method.

To form small group consisting of 51 continuous air quality monitoring stations located in different locations into three clusters, namely HPR, MPR and LPR with AHC based on the concentration of $PM_{10}$ was implemented. The results of AHC visualised in dendrogram revealed the dissimilarity between the clusters involved. The highest, modest and lowest average $PM_{10}$ concentrations among these regions were named as HPR, MPR and LPR, respectively.

### 3.5.2 Discriminant Analysis (DA)

Specific classification of data that shares common properties can be differentiated by using one of the multivariate methods, namely discriminant analysis (DA) (Ali et al.,

2014). The total corrected percentage or efficiency is necessary to be defined after clusterisation to find out the reliability of the result (Ismail et al., 2017). There is a discriminant function (DF) for each cluster, whereby DF can be determined by the equation below (Mohamad et al., 2018):

$$f(G_i) = k_i + \sum_{j-1}^{n} W_{ij} P_{ij} \quad (3.6)$$

where,

i = groups number (G)

$k_i$ = the constant innate to each group

n = parameters number used to categorise data set into a given group

$w_j$ = weight coefficient allotted by DF analysis to a given parameter ($P_j$)

In this analysis, raw data were used to develop DF for each of the group, whereby this function characterises a method to divide data into regions. Effective data set gave a high correct percentage of classification (Mutalib et al., 2013).

### 3.5.3 Principal Component Analysis (PCA)

Kaiser Meyer Olkin (KMO) and Bartlett's tests were performed at the beginning of the PCA to test the suitability of data before applying the selected parameters for the next analysis. Generally, KMO test can be used to determine whether data of interest are a well factor or not, while Bartlett's test can be used to confirm correlated parameters in the analysis. Table 3.3 shows the guiding rules for interpreting the results of KMO test. Adequacy of samples was verified by applying the KMO measure of sampling adequacy (MSA) before extracting the factors in the PCA. According to Gazzaz et al. (2012), MSA is suitable if the KMO value were ranged from 0.60 to 1.00.

Table 3.3. Rules of guidance for interpreting results of KMO test

| Value of KMO | Interpretation |
|---|---|
| 0.90-1.00 | Marvellous |
| 0.80-0.89 | Meritorious |
| 0.70-0.79 | Middling |
| 0.60-0.69 | Mediocre |
| 0.50-0.59 | Miserable |
| 0.00-0.49 | Unacceptable |

Meanwhile, the Bartlett test was able to evaluate the correlation probability in a matrix. In this test, the null hypothesis ($H_0$) represents no correlation significantly different from 0 between the parameters, while the alternative hypothesis ($H_a$) represents at least one of the correlations between the parameters is significantly different from 0. $H_0$ and $H_a$ would be rejected and accepted, respectively, if the p-value computed is lower than the significant alpha level, which is 0.05. Therefore, the parameters used in the PCA are considered correlated if the $H_0$ is rejected.

With the aim to find the principal parameters as inputs for the prediction model, PCA was applied onto 11 parameters, namely wind speed, wind direction, temperature, humidity, $NO_x$, NO, $NO_2$, $O_3$, $SO_2$ and CO for further analysis. Besides, interrelation between sets of parameters can also be analysed and interpreted by using this analysis since the analysis is also capable of being utilised to identify the source of emission (Azid et al., $2015_b$). Principal components (PCs) are the new parameters created from the PCA. According to Juahir et al. (2011), the maximum new number of parameters is equal to the original parameter numbers.

However, there is some drawbacks that affected the analyses, such as the PCs generated by PCA at the time were not readily interpreted. Therefore, the varimax rotation needs to be applied onto the PCA in order to rotate the PCs generated by the PCA to ensure that they are readily interpreted. Component complexity can be reduced

by producing large loadings larger and the small loadings smaller within each component. On top of that, the interpretation became easier since the method of a factor had been simplified.

However, according to Kim and Mueller (1987), to attain the new parameters, only the PCs with eigenvalues greater than one are considered significant and implemented for the next analysis. For varimax rotation method, the VF (varimax factor) is reflected as strong, moderate and weak factor loadings if the values are greater than 0.75, ranging between 0.50 and 0.75, and range between 0.30 and 0.49, respectively. In this study, greater than 0.75 absolute values of VF were chosen as the selection threshold beside representing the most significant parameters and strong factor loadings. These selected inputs were then applied for the $PM_{10}$ prediction model by the ANN and MLR models.

In this study, the AHC and prediction models (ANN and MLR) were combined with PCA with the intention of generating the most powerful $PM_{10}$ prediction model. By applying this, only the most significant parameters obtained from PCA were implemented as input parameters for the specific type of region that was classified through AHC, namely HPR, MPR and LPR. The equation used for assessing the PC is as shown below (Azid et al., $2015_b$):

$$z_{ij} = a_{i1}x_{1j} + a_{i2}x_{2j} + \cdots + a_{im}x_{mj} \tag{3.7}$$

where,

z = score of the component

a = factor loading

x = measured value of parameters

i = number of component

j = number of parameter

m = parameters total number

### 3.6 Performance evaluation

The performance of both prediction models was examined according to the value of statistical analysis, namely coefficient of determination ($R^2$), root means square error (RMSE), index of agreement (IA), the percentage of deviation and coefficient of efficiency (E) obtained. Accuracy for both prediction models was based on the comparison between the observed and predicted concentrations. The optimal values that gave better performance and highest accuracy between actual and predicted values were 0 for RMSE and 1 for $R^2$, IA and E, while for percentage of deviation, the prediction is correctly predicted if the deviation of sample is less than 20% while it is less correctly predicted if the deviation of sample is greater than 70%. The performance of each statistical analysis was examined according to the following equations:

Table 3.4. Performance measures used in model evaluations

| Measure | Equation | Description |
|---|---|---|
| The coefficient of determination ($R^2$) (Taspinar & Bozkurt, 2014) | $\left(\frac{\sum_{i=1}^{n}(o_i - \bar{o})^2.(P_i - P)}{n.\sigma_o.\sigma_p}\right)^2$ | Close to 1.0 indicates the greater explained variance |
| Root mean square error (RMSE) (Dotse et al., 2018) | $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(O_i - P_i)^2}$ | The smaller values of RMSE denote the better model performance |
| Index of agreement (IA) (Dotse et al., 2018) | $1-\left(\frac{\sum_{i=1}^{n}(P_i - O_i)^2}{\sum_{i=1}^{n}(\lvert P_i - \bar{O}\rvert + \lvert O_i - \bar{O}\rvert)^2}\right)$ | Closer to 1.0 indicates the better agreement with the selected model |

| | | |
|---|---|---|
| Percentage of deviation (Yang et al., 2016) | $\frac{\lvert P_i - O_i \rvert}{O_i} \; X \; 100\%$ | A sample is correctly predicted if the deviation of a sample is less than 20% or the predicted air quality level matches the observed level |
| The coefficient of efficiency (E) (Nash & Sutcliffe, 1970) | $\frac{\sum_{i=1}^{n}(O_i - \bar{O})^2 - \sum_{i=1}^{n}(O_i - P_i)^2}{\sum_{i=1}^{n}(O_i - \bar{O})^2}$ | If the measured value is the same as all predictions, E is 1. If the E is between 0 and 1, it indicates deviations between measured and predicted values. If E is negative, predictions are very poor |

Notes: n = number of data; $P_i$ = predicted values; $O_i$ = observed values; $\bar{O}$ = mean of the observed values; P = mean of the predicted values; $\sigma_P$ = standard deviation of the predicted values; $\sigma_O$ = standard deviation of the observed values

### 3.7 Steps of the $PM_{10}$ prediction model by using ANN and MLR

The steps in the study of the $PM_{10}$ prediction model by using ANN are as follows:

Step 1: Six years of data from January 1, 2010 until December 31, 2015 for air pollutants ($PM_{10}$, $NO_2$, CO, $O_3$, NO, $NO_x$, THC, $CH_4$, NMHC, $SO_2$) and meteorological parameters (wind direction, wind speed, temperature, humidity, UVB) were obtained from the Air Quality Division, Department of Environment (DOE), Malaysia.

Step 2: In order to develop a $PM_{10}$ prediction model based on HPR, MPR and LPR, the data obtained were classified in this region by using AHC. The reliability of the results was then confirmed by DA.

Step 3: Missing data imputation was applied to the data in order to utilise all the data obtained. In this study, parameters of UVB, $CH_4$, NMHC and THC were removed from the analysis due to their high percentage of missing data. Therefore, only 11

parameters were taken into account with a total of 1, 228, 161 data points (11 parameters × 111 651 data set) for the next analysis.

Step 4: PCA was applied to the data in order to obtain the most contributed parameters towards the atmosphere. Before applying PCA, KMO and Bartlett's tests were also applied to the data during this stage to ensure that the data are suitable to be applied for the next analysis. Then, the selected parameters were used as input parameters for the prediction model. This study used two groups of data as input parameters, which were parameters without $PM_{10}$ ($NO_2$, CO, $O_3$, NO, $NO_x$, $SO_2$, wind direction, wind speed, temperature, humidity, UVB) and parameters with $PM_{10}$ ($PM_{10}$, $NO_2$, CO, $O_3$, NO, $NO_x$, $SO_2$, wind direction, wind speed, temperature, humidity) for each of the regions.

Step 5: Data transformation was then applied to the data in the range of -1 to 1 before applying the data for the prediction model since there was a dissimilar range among the parameters used. The input parameters obtained from PCA were applied for the ANN and MLR prediction models.

Steps 6: For ANN, there were some specific criteria needed to be fulfiled to obtain the best prediction model. In this study, 67%, 16% and 17% of the data were used for the training, testing and validation data sets, respectively. The activation function of the hyperbolic tangent function, which bounded in the range of -1 to 1 was used. Meanwhile, the Levenberg-Marquardt algorithm was used for the network training. A hidden layer was used in this study while a number of hidden nodes were varied for each region based on the number of input and output parameters. For MLR, there were no specific criteria as the ANN prediction model. However, the best input parameters used will give a better prediction model.

Steps 7: The prediction model performance used $PM_{10}$ as input parameters and without $PM_{10}$ as input parameters for both models was evaluated by the statistical values of $R^2$, RMSE, IA, and percentage of deviation.

## 3.8 Statistical analysis tools

In this study, AHC, DA, Pearson correlation, PCA and MLR were performed by using XLSTAT 2014 add-in software developed by Addinsoft. Meanwhile, ANN was computed by the JMP13 software developed by SAS Institute. For the missing data imputation method, mean and nearest neighbour imputation method were analysed by using XLSTAT 2014 add-in software, while EMB algorithm imputation method was analysed by using Amelia II, which exists as a package for the R statistical software package developed by James Honaker, Gary King and Matthew Blackwell. By choosing AmeliaView in the Amelia II programmed, one can directly run Amelia II without having high knowledge about the R programming language.

## 3.9 Summary

This chapter discussed the methods being applied for the $PM_{10}$ prediction model. The raw data underwent pretreatment, such as outliers, missing data imputation and transformation of data before being analysed by using ANN and MLR. The chemometric analysis, namely AHC, DA and PCA were implemented with the aim of classifying the region into HPR, MPR and LPR as well as to determine the input parameters for the prediction model by selecting the most contributing parameters that deteriorated air quality. The processes experienced by the machine learning prediction models of the ANN and MLR were also discussed with their specific criteria to ensure the best prediction model is obtained. The performance evaluation and statistical analysis tools for both models were also discussed in this chapter.

# CHAPTER 4

# RESULTS AND DISCUSSION

## 4.1 Introduction

This chapter discusses the development of $PM_{10}$ prediction model for HPR, MPR and LPR. Section 4.2 discusses the spatial classification of cluster analysis. Section 4.3 focuses on the discriminant analysis (DA) based on spatial classification regions to ensure that all these three regions are well classified. Section 4.4 discusses an overview of data on air quality based on spatial classification visualised by the box-whisker plot. Section 4.5 gives an explanation about the air pollutants trend which consists of the annual air quality status for each region, annual variation of the five main pollutants, namely CO, $NO_2$, $SO_2$, $O_3$, and $PM_{10}$ as well as the monthly variation of $PM_{10}$ in Malaysia from 2010 to 2015 for HPR, MPR and LPR. Section 4.6 describes the correlation between air pollutants and meteorological factors. Meanwhile, Section 4.7 discusses data preprocessing before developing the $PM_{10}$ prediction model. This section also discusses the Kaiser-Meyer-Olkin (KMO) and Bartlett's tests in order to determine the suitability of data to be implemented in the models, the appropriate imputation methods to be applied onto the missing data as well as the normality test. Section 4.8, describes the designing of $PM_{10}$ prediction model starting with determining the main sources of pollutants obtained from PCA with or without lagged $PM_{10}$ and eventually discussing the evaluation on of the models.

## 4.2 Spatial classification of cluster analysis

Figure 4.1 shows three significant clusters that share the characteristics of homogeneity obtained from the AHC analysis using the raw data collected from the 51

sampling stations. In order to ensure that the regions are not influenced by the missing data, a station named CA 0045 (ILP Miri) was excluded from the AHC analysis since it consisted of missing data greater than 25%. Table 4.1 shows the list of three significant clusters, namely Cluster 1, Cluster 2 and Cluster 3 represented by HPR, MPR and LPR, respectively.

Cluster 1 consisted of stations from CA0001, CA0057, CA0019, CA0015, CA0002, CA0011, CA0006, CA0010, CA0044, CA0043, CA0056, CA0047, CA0053, CA0060, CA0054, CA0058, CA0025 and CA0016 (Pasir Gudang, Kota Tinggi, Larkin, Balok Baru, Kemaman, Klang, SMK Bukit Rambai, Nilai, Muar, SM Tinggi, Port Dickson, Seremban, Putrajaya, Banting, Cheras, Batu Muda, Shah Alam and Petaling Jaya).

Cluster 2 consisted of stations from CA0048, CA0041, CA0009, CA0017, CA0020, CA0008, CA0046, CA0022, CA0059, CA0034, CA0032, CA0033, CA0040, CA0038 and CA0003 (Kuala Selangor, Manjung, SK Seberang Jaya, Sg Petani, Taiping, SM Jln Tasek, SM Pagoh, Kota Bharu, Tanah Merah, Kuala Terengganu, Langkawi, ILP Kangar, Alor Setar, USM and SK Cenderawasih Prai).

Cluster 3 consisted of stations from CA0036, CA0004, CA0035, CA0027, CA0029, CA0026, CA0039, CA0042, CA0050, CA0030, CA0049, CA0031, CA0055, CA0028, CA0024, CA0045, CA0007 and CA0014 (Sri Aman, Kuching, Kota Samarahan, Bintulu, Sarikei, Sibu, Tawau, Labuan, Sandakan, Kota Kinabalu, Keningau, Limbang, Kapit, Dato' Permaisuri, Kerteh, UPSI, Jerantut and Indera Mahkota). Cluster 1 was classified as HPR with the average $PM_{10}$ value of 58.78 $\mu g/m^3$, while Cluster 2 and Cluster 3 were classified as MPR and LPR with the average $PM_{10}$ value of 50.50 $\mu g/m^3$ and 41.03 $\mu g/m^3$, respectively.

This finding implied that a number of air quality monitoring stations can be reduced according to the classification regions (HPR, MPR and LPR). Moreover, AHC was more efficient by reducing the time consuming as well as cost saving in order to determine the stations which have the same classification, especially for the $PM_{10}$ prediction model. Therefore, rapid evaluation of the $PM_{10}$ prediction model can be implemented by assessing one station for each cluster since it represents the whole network of the cluster classification.

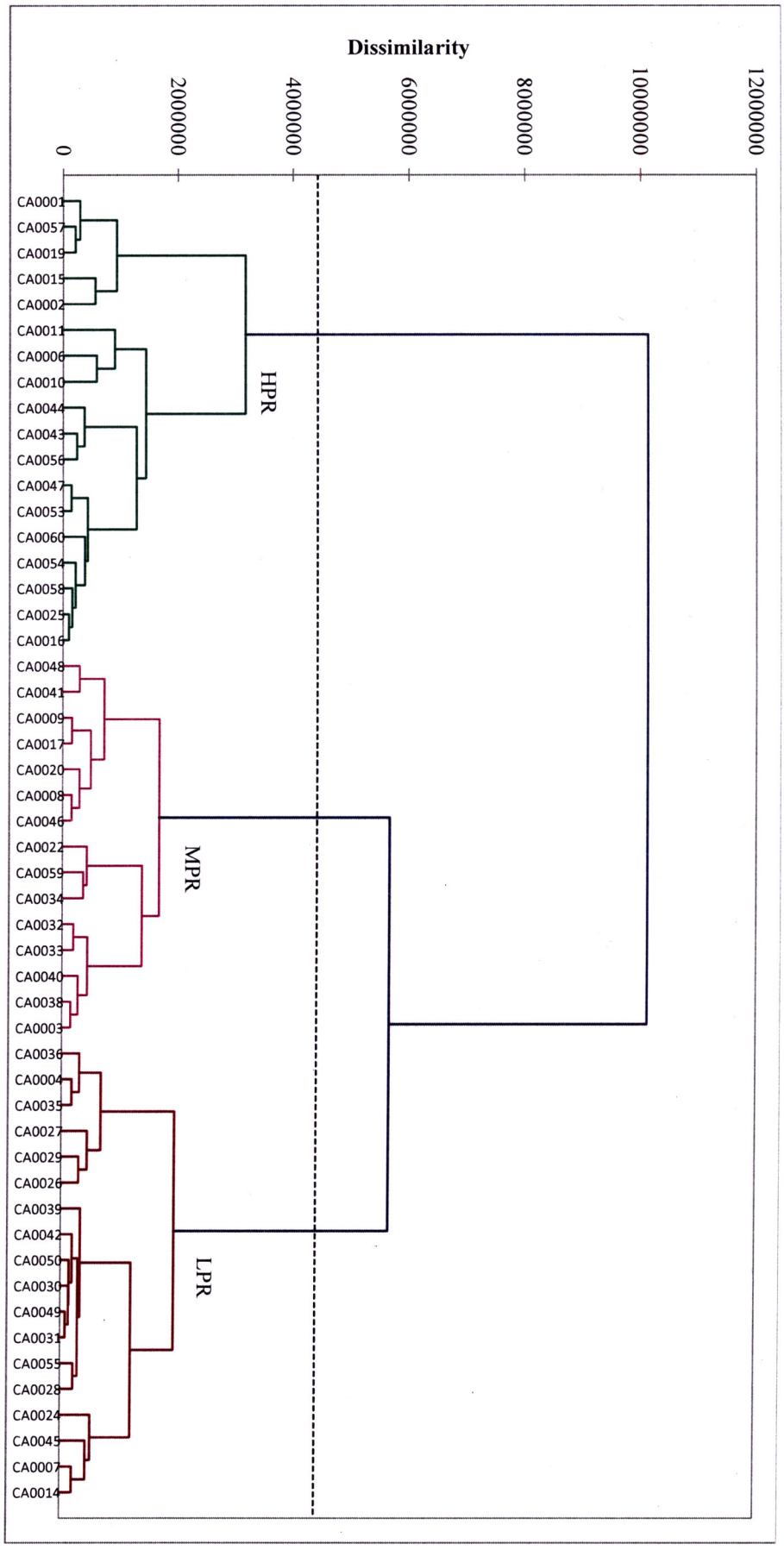



Figure 4.1. Generated dendrogram by using Ward's linkage method for AHC

Table 4.2 shows the number of stations located in the industrial, background, urban and suburban areas based on the regional classification. Each cluster represents the stations located in the industrial, urban and suburban areas. However, only LPR represents the stations located in the background area. HPR consisted the highest number of stations located in the industrial and urban areas and the lowest number of stations located in the suburban areas as compared to MPR and LPR with a number of stations located in the industrial and urban were seven (CA0016, CA0006, CA0010, CA0019, CA0001, CA0002, CA0015) and nine (CA0011, CA0025, CA0058, CA0054, CA0053, CA0047, CA0056, CA0043, CA0057), respectively.

In contrast there were four stations, namely CA0003, CA0020, CA0008, CA0059, located in the industrial areas and four stations, namely CA0040, CA0046, CA0022, CA0034, located in the urban areas for MPR, while for LPR, the number of stations located in the industrial and urban areas were two stations (CA0004, CA0024) and two stations (CA0030, CA0039), respectively. LPR showed the highest number of the stations located in the suburban and background areas and the lowest number of stations located in the urban and industrial areas.

Table 4.1. List of HPR, MPR and LPR obtained from the AHC

| Cluster 1 (HPR) | | Cluster 2 (MPR) | | Cluster 3 (LPR) | |
|---|---|---|---|---|---|
| CA0001 | Pasir Gudang | CA0048 | Kuala Selangor | CA0036 | Sri Aman |
| CA0057 | Kota Tinggi | CA0041 | Manjung | CA0004 | Kuching |
| CA0019 | Larkin | CA0009 | SK Seberang Jaya | CA0035 | Kota Samarahan |
| CA0015 | Balok Baru | CA0017 | Sg Petani | CA0027 | Bintulu |
| CA0002 | Kemaman | CA0020 | Taiping | CA0029 | Sarikei |
| CA0011 | Klang | CA0008 | SM Jln Tasek, Ipoh | CA0026 | Sibu |
| CA0006 | SMK Bukit Rambai | CA0046 | SM Pagoh, Ipoh | CA0039 | Tawau |
| CA0010 | Nilai | CA0022 | Kota Bharu | CA0042 | Labuan |
| CA0044 | Muar | CA0059 | Tanah Merah | CA0050 | Sandakan |
| CA0043 | SM Tinggi Melaka | CA0034 | Kuala Terengganu | CA0030 | Kota Kinabalu |
| CA0056 | Port Dickson | CA0032 | Langkawi | CA0049 | Keningau |
| CA0047 | Seremban | CA0033 | ILP Kangar | CA0031 | Limbang |
| CA0053 | Putrajaya | CA0040 | Alor Setar | CA0055 | Kapit |
| CA0060 | Banting | CA0038 | USM | CA0028 | Dato Permaisuri, Miri |
| CA0054 | Cheras | CA0003 | SK Cenderawasih, Prai | CA0024 | Kerteh |
| CA0058 | Batu Muda | | | CA0045 | UPSI |
| CA0025 | Shah Alam | | | CA0007 | Jerantut |
| CA0016 | Petaling Jaya | | | CA0014 | Indera Mahkota |

Table 4.2. Number of stations located in industrial, background, urban and suburban areas based on regional classification

| Region | Number of stations | | | | Total number of stations |
|---|---|---|---|---|---|
| | Industries | Background | Urban | Suburban | |
| HPR | 7 | 0 | 9 | 2 | 18 |
| MPR | 4 | 0 | 4 | 7 | 15 |
| LPR | 2 | 3 | 2 | 11 | 18 |

Figure 4.2 visualises the classification of regions (HPR, MPR, LPR) as a result of air quality by AHC based on $PM_{10}$ pollutant in Malaysia with most of the HPR located at the Southern region, Central region and some of the East region of Peninsular Malaysia, while most of the MPR located at the Northern region of Peninsular Malaysia. The whole Borneo region was classified as LPR with some of the stations located in the East region and Northern region of Peninsular Malaysia. A high number of stations located in the industrial and urban present in the southern region, central region and some east regions of Peninsular Malaysia caused them were classified as HPR. According to Amran et al. (2015), the higher reading of air quality was recorded at areas with heavy industries and transportation areas that release $PM_{10}$ into the atmosphere.

For instance, Kemaman station is one of the stations classified as HPR. Geographically, Kemaman station is located in the industrial area, which contributes to the high concentration of $PM_{10}$. It is also known for the petroleum discovery site near its town. In Kemaman, there is Kerteh Petrochemical Industrial Area, which is the major industrial site in Terengganu. Besides that, there is also Gebeng Industrial Area that is located near Kemaman, which may also be one of the contributors to the high $PM_{10}$ concentration in Kemaman.

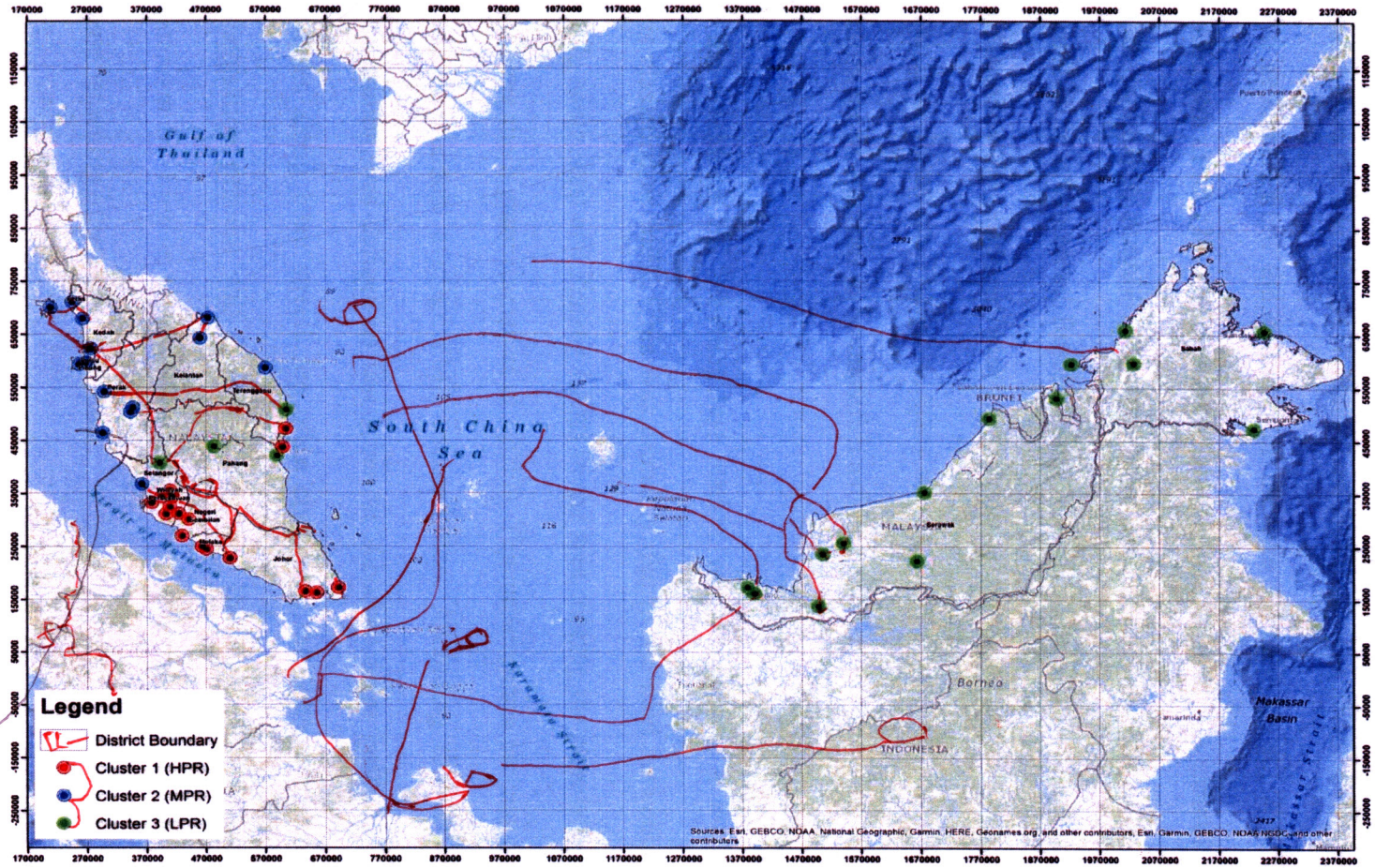



Figure 4.2. Classification of regions as a result of air quality by AHC in Malaysia

Meanwhile, stations located at urban areas experienced high $PM_{10}$ concentration probably due to the emissions from motor vehicles. This is supported by the study done by Mohamad et al. (2015), who mentioned that this type of pollutant usually comes from the use of motor vehicle besides industrial activities. On top of that, it was also produced from the abrasion of the tyre (Mutalib et al., 2013). According to Mahapatra et al. (2018), in ambient air, particulate matter is identified as one of the main pollutants, mainly in urban areas. Therefore, high $PM_{10}$ concentrations were identified in urban areas. Besides, Leh et al. (2012) mentioned that in urban areas, air pollutants were produced with higher rate as compared to less developed areas and natural environment. Therefore, many stations located in the urban area was one of the factors for the stations to be classified as HPR.

In the meantime, SK Seberang Jaya station, which is located at the suburban station was classified as MPR. This was probably due to the road networks connecting the north, south, east and Penang Island in the area making the population of the resident is among the highest (Ismail et al., 2017). This urbanisation process contributes to the increasing annual average $PM_{10}$ concentration. According to Masiol et al. (2017), urbanisation was defined as fast urban or suburban areas build-up while some city features along busy streets are often congested throughout rush hour times. This urbanisation process will indirectly decrease the air quality. This is supported a statement by Ling et al. (2010) which mentioned that process of urbanisation will cause lessening of good days number.

Meanwhile, most of the stations classified under LPR experienced were less polluted rather than stations located in industrial and urban areas due to the lowest number of air quality monitoring stations located at the industrial and urban areas as well as the

presence of air quality monitoring stations located at the background areas. However, although located in the background area, it has the potential to experience high $PM_{10}$ concentration. According to a study by Latif et al. (2014), they concluded that wind direction was a factor that transports the $PM_{10}$ pollutant to the rural areas from more urbanised and industrial areas.

Therefore, the high $PM_{10}$ concentration present in LPR was probably due to the $PM_{10}$ transportation from Sumatra during the forest fire and other urbanised and industrial areas during the Northeast Monsoon season. Forest fire in Sumatra is one of the events that promote the long range transportation of haze to Malaysia. According to Yusof et al. (2008), it was expected that $PM_{10}$ will reach the peak as the weather during this season becomes dry and southwesterly wind brings the pollutant due to burning to the other regions during the Southwest Monsoon.

### 4.3 Discriminant analysis (DA) based on spatial classification regions

Further analysis by using DA, which exposed dissimilarities within the study sites was implemented in view of the clustering obtained from the AHC for HPR, MPR and LPR. Here, the $PM_{10}$ concentrations were considered as the independent parameters, while the study areas were considered as the grouping parameters.

Table 4.3. Classification matrix of DA for HPR, MPR and LPR

| Regions | HPR | MPR | LPR | Total | % correct |
|---|---|---|---|---|---|
| HPR | 33521 | 4342 | 1485 | 39348 | 85.19 |
| MPR | 5020 | 25750 | 2095 | 32865 | 78.35 |
| LPR | 2018 | 2241 | 35179 | 39438 | 89.20 |
| Average | | | | | 84.25 |

Table 4.3 shows the classification matrix of DA, which involved three types of clusters, namely HPR, MPR and LPR. Based on the table, there were 33,521 numbers

of data from the HPR with the percentage accuracy of 85.19%, 25,750 numbers of data from the MPR with the percentage accuracy of 78.35% and 35,179 numbers of data from the LPR with the percentage accuracy of 89.20%. In general, the regions discriminated well with an average of 84.25% correct classification.

Besides, Figure 4.3 also shows that these three classification regions were discriminating well based on the Receiver Operating Characteristics (ROC), which supported the accuracy of each cluster. According to Hidalgo et al. (2018), the performance of a classifier can be assessed by the curve of ROC, whereby the accuracy of the classifier can be measured through the area under the curve. It is a perfect test if an area that represents 1 while it is a worthless test if an area represents 0.5. The ROC curve for HPR, MPR and LPR was 0.8109, 0.8809 and 0.7547, respectively. From the results obtained, LPR showed the best classification followed by HPR and MPR perhaps due to a slight misclassification of stations in HPR and MPR as compared to LPR. However, it showed that the accuracy for each classification region was acceptable in area represented above the worthless test value.

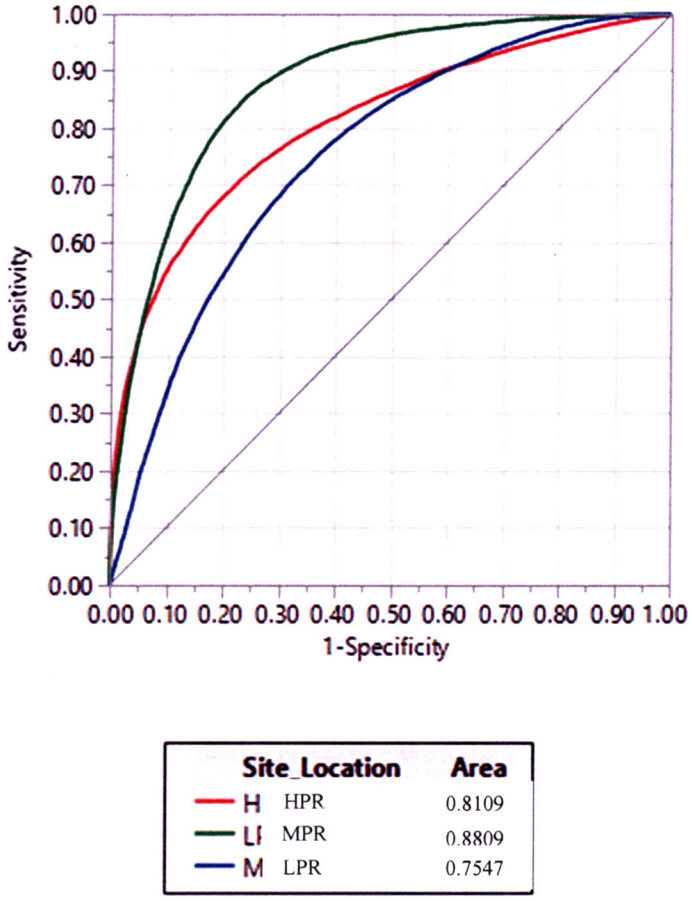



Figure 4.3. Receiver Operating Characteristics (ROC) for the classification of regions

Besides, the test will become more accurate if the curve becomes closer to the left-hand border as well as to the top border space of ROC. In contrast, it becomes less accurate if the curve gets closer to the 45-degree diagonal space of ROC. Figure 4.3 shows that all of the classification regions followed the left-hand border closely with the LPR classification region shows closer to the top border of the ROC space followed by HPR and MPR, respectively.

On top of that, the Kruskal–Wallis test was also conducted and the $PM_{10}$ concentration showed a significant difference if the *p*-value was less than 0.05. The null hypothesis ($H_0$) states that the mean vectors of the three classes come from the same population

(equal). The alternative hypothesis ($H_a$) states that the mean vectors of the three classes did not come from the same population (different). The computed *p*-value, which was less than 0.05 ($p < 0.0001$) obtained from the Kruskal-Wallis test proved that the three clusters were indeed different from one another.

## 4.4 Overview data of air quality based on spatial classification (Box-whisker plot)

The boxplots in Figure 4.4 until Figure 4.6 show the daily average data of meteorological (wind speed, wind direction, temperature, UVB, humidity) and air pollutant ($NO_x$, NO, $CH_4$, NMHC, THC, $SO_2$, $NO_2$, $O_3$, CO, $PM_{10}$) parameters from 2010 to 2015 for HPR, MPR and LPR all around Malaysia. The boxplot was used to visualise the minimum, maximum, 1$^{st}$ quartile, median, 3$^{rd}$ quartile, mean values, as well as outliers, which were not included between the whiskers (the endpoints of the lines attached to the box), and the data is displayed in Table 4.4 until Table 4.6.

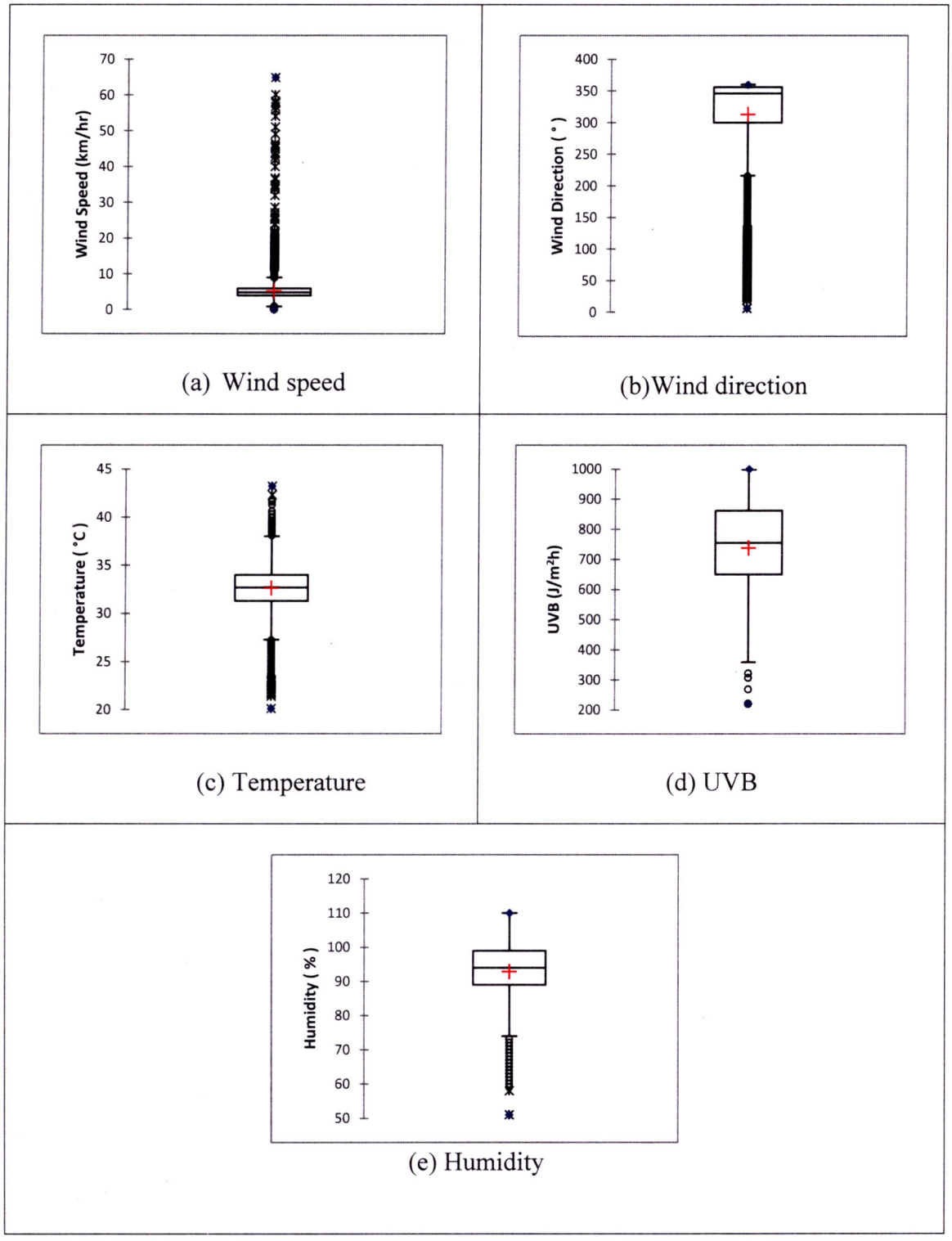



Figure 4.4. Boxplots of (a) wind speed, (b) wind direction, (c) temperature, (d) UVB, (e) humidity) (f) $NO_x$, (g) NO, (h) $CH_4$, (i) NMHC, (j) THC, (k) $SO_2$, (l) $NO_2$, (m) $O_3$, (n) CO, and (o) $PM_{10}$ for HPR

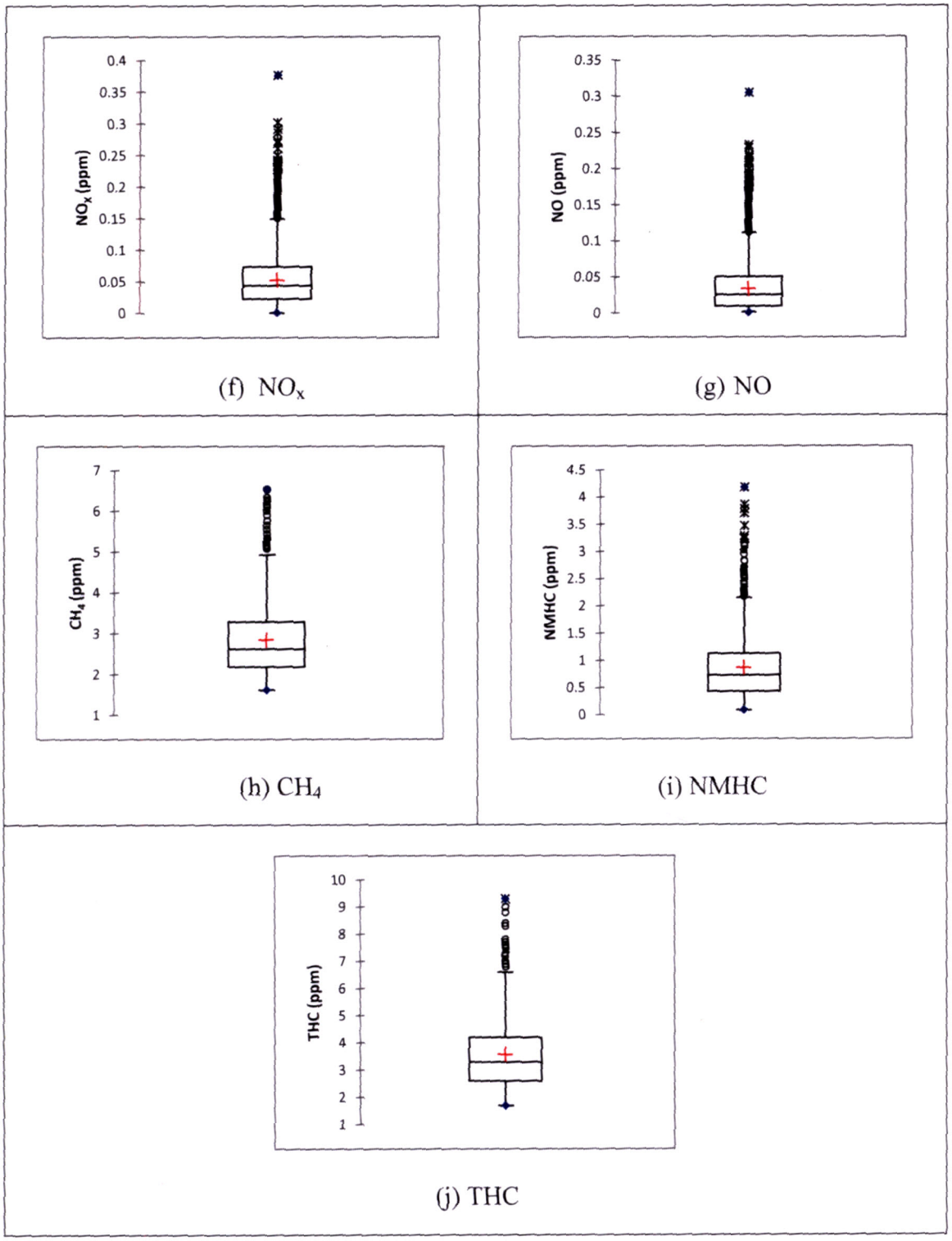



Figure 4.4. Continued

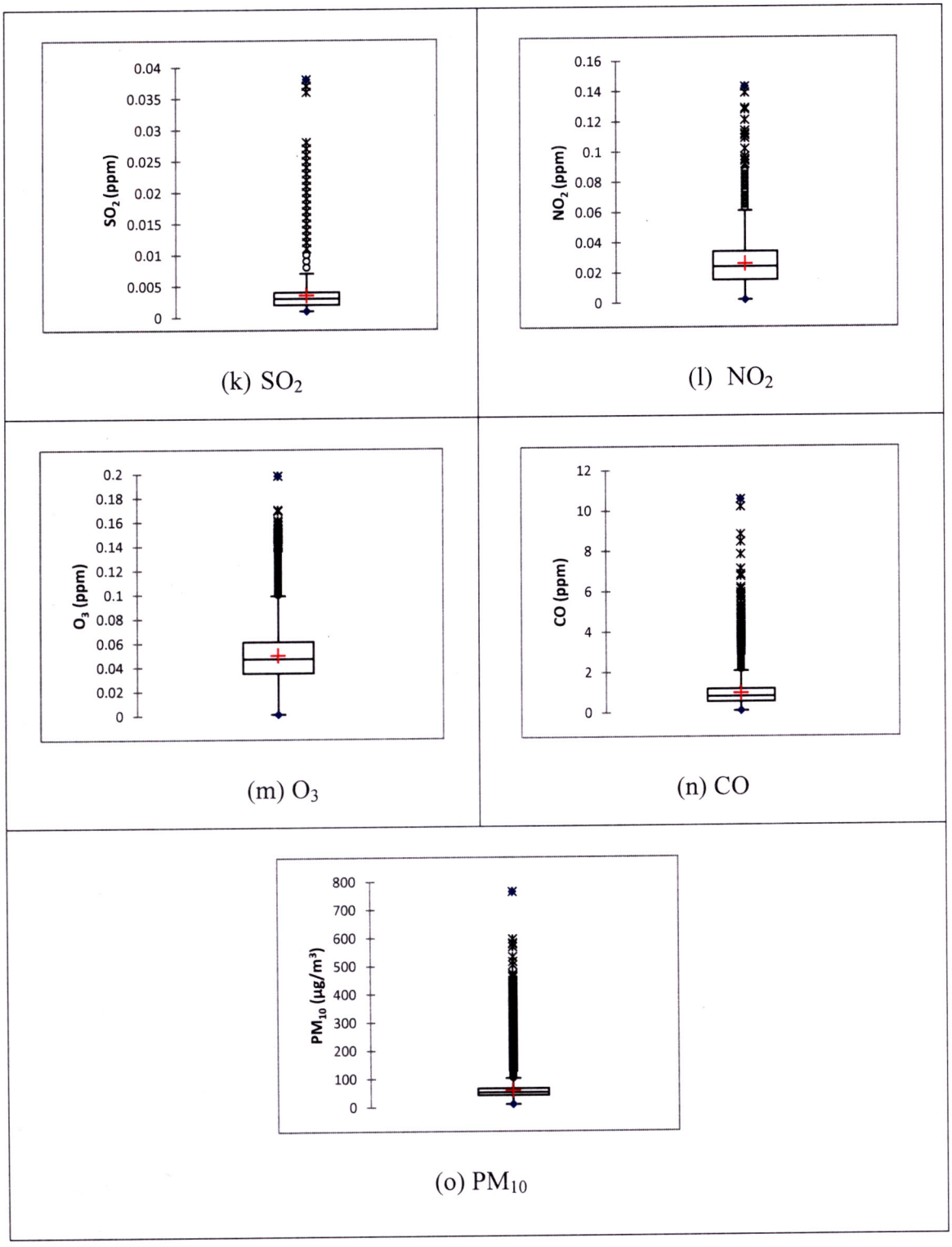



Figure 4.4. Continued

Table 4.4. Value of minimum, maximum, 1st quartile, median, 3rd quartile, mean, skewness and kurtosis for (a) meteorological parameters and (b) pollutant parameters for HPR

| Statistic | WS (km/hr) | WD (°) | Temp (°C) | UVB ($J/m^2h$) | Humidity (%) |
|---|---|---|---|---|---|
| Minimum | 0.09 | 6 | 20.10 | 220 | 51 |
| Maximum | 64.90 | 360 | 43.20 | 999 | 110 |
| 1st Quartile | 3.88 | 300 | 31.30 | 650 | 89 |
| Median | 4.79 | 346 | 32.70 | 755 | 94 |
| 3rd Quartile | 5.90 | 356 | 34.00 | 862 | 99 |
| Mean | 5.13 | 313 | 32.63 | 737 | 93 |
| Skewness (Pearson) | 6.4881 | -2.0283 | -0.2653 | -0.6362 | -0.8929 |
| Kurtosis (Pearson) | 113.6592 | 3.8067 | 0.8533 | 0.1819 | 0.8186 |

Table 4.4. Continued

| Statistic | $NO_X$ (ppm) | NO (ppm) | $CH_4$ (ppm) | NMHC (ppm) | THC (ppm) | $SO_2$ (ppm) | $NO_2$ (ppm) | $O_3$ (ppm) | CO (ppm) | $PM_{10}$ ($\mu g/m^3$) |
|---|---|---|---|---|---|---|---|---|---|---|
| Minimum | 0.001 | 0.001 | 1.62 | 0.09 | 1.70 | 0.001 | 0.002 | 0.001 | 0.09 | 10 |
| Maximum | 0.377 | 0.305 | 6.53 | 4.18 | 9.30 | 0.038 | 0.143 | 0.198 | 10.56 | 763 |
| 1st Quartile | 0.023 | 0.009 | 2.19 | 0.43 | 2.60 | 0.002 | 0.015 | 0.035 | 0.54 | 42 |
| Median | 0.044 | 0.025 | 2.63 | 0.73 | 3.30 | 0.003 | 0.024 | 0.047 | 0.81 | 52 |
| 3rd Quartile | 0.074 | 0.050 | 3.29 | 1.13 | 4.20 | 0.004 | 0.034 | 0.061 | 1.17 | 66 |
| Mean | 0.053 | 0.034 | 2.85 | 0.87 | 3.57 | 0.004 | 0.026 | 0.050 | 0.95 | 59 |
| Skewness (Pearson) | 1.1390 | 1.3283 | 1.2847 | 1.6014 | 1.3196 | 2.5114 | 0.8439 | 0.8686 | 2.2804 | 4.8765 |
| Kurtosis (Pearson) | 1.4309 | 1.9925 | 1.7445 | 3.4717 | 2.2698 | 12.0010 | 0.9019 | 1.1432 | 14.3918 | 41.6048 |

Table 4.4 shows the statistical value of the minimum, maximum, 1st quartile, median, 3rd quartile and mean for meteorological and pollutant parameters, respectively, during the entire study period from 2010 to 2015 for HPR. The ranges for the meteorological parameters obtained for HPR were: wind speed (0.09-64.90) km/hr, wind direction (6–360)°, temperature (20.10–43.20)°C, UVB (220–999) J/m$^2$h and humidity (51–110)% with mean values of wind speed, wind direction, temperature, UVB and humidity were 5.13 km/hr, 313°, 32.63°C, 737 J/m$^2$hr and 93%, respectively. Meanwhile, the median values of wind speed, wind direction, temperature, UVB and humidity were 4.79 km/hr, 346°, 32.70°C, 755 J/m$^2$hr and 94%, respectively.

The ranges for the pollutant parameters obtained for HPR were: $NO_x$ (0.001–0.377) ppm, NO (0.001–0.305) ppm, $CH_4$ (1.62–6.53) ppm, NMHC (0.09–4.18) ppm, THC (1.70–9.30) ppm, $SO_2$ (0.001–0.038) ppm, $NO_2$ (0.002–0.143) ppm, $O_3$ (0.001–0.198) ppm, CO (0.09–10.56) ppm and $PM_{10}$ (10–763) μg/m$^3$ with mean concentrations of $NO_x$, NO, $CH_4$, NMHC, THC, $SO_2$, $NO_2$, $O_3$, CO and $PM_{10}$ were 0.053 ppm, 0.034 ppm, 2.85 ppm, 0.87 ppm, 3.57 ppm, 0.004 ppm, 0.026 ppm, 0.05 ppm, 0.95 ppm and 59 μg/m$^3$, respectively. Meanwhile, the median values of $NO_x$, NO, $CH_4$, NMHC, THC, $SO_2$, $NO_2$, $O_3$, CO and $PM_{10}$ were 0.044 ppm, 0.025 ppm, 2.63 ppm, 0.73 ppm, 3.30 ppm, 0.003 ppm, 0.024 ppm, 0.047 ppm, 0.81 ppm and 52 μg/m$^3$, respectively.

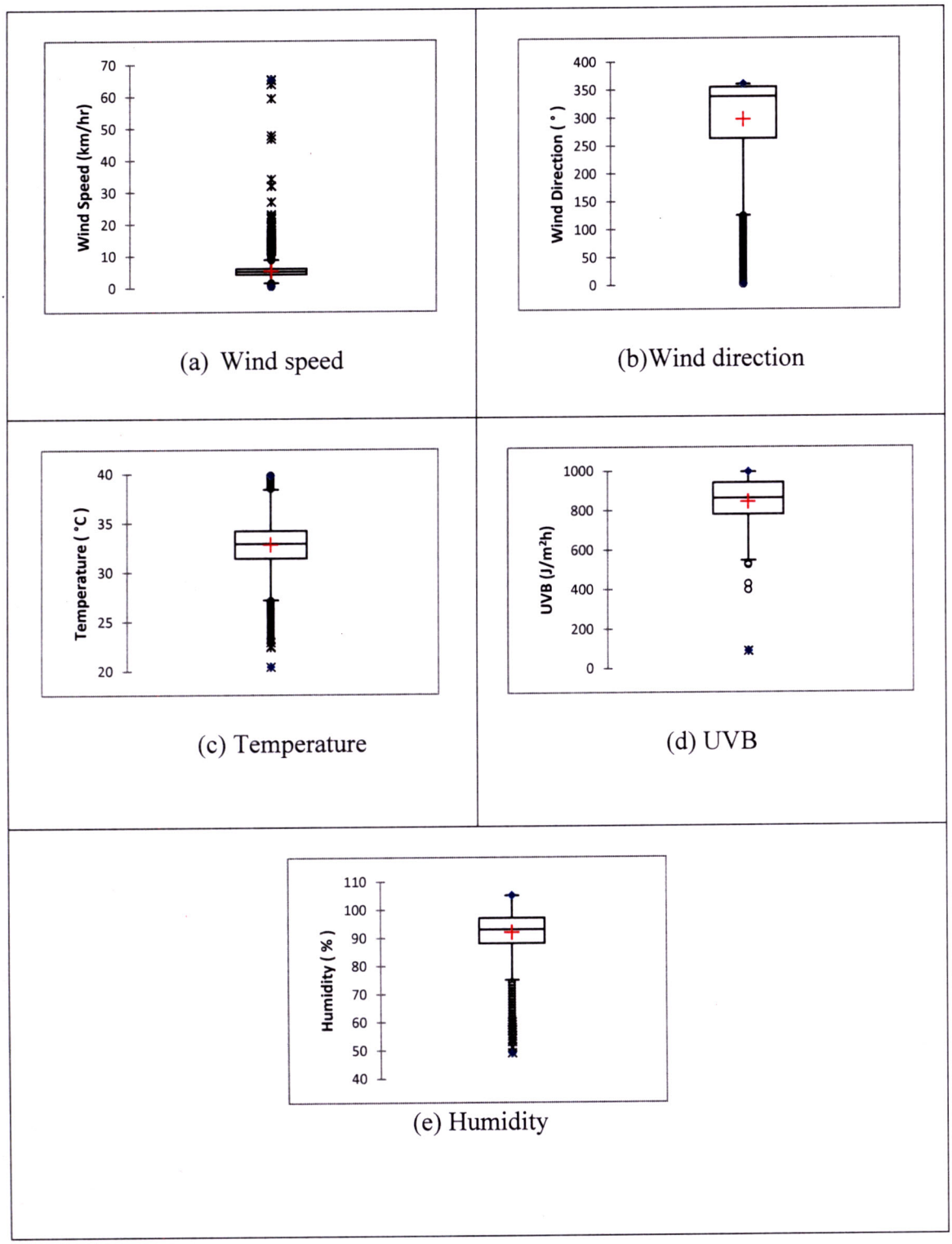



Figure 4.5. Boxplots of (a) wind speed, (b) wind direction, (c) temperature, (d) UVB, (e) humidity (f) $NO_x$, (g) NO, (h) $CH_4$, (i) NMHC, (j) THC, (k) $SO_2$, (l) $NO_2$, (m) $O_3$, (n) CO, and (o) $PM_{10}$ for MPR

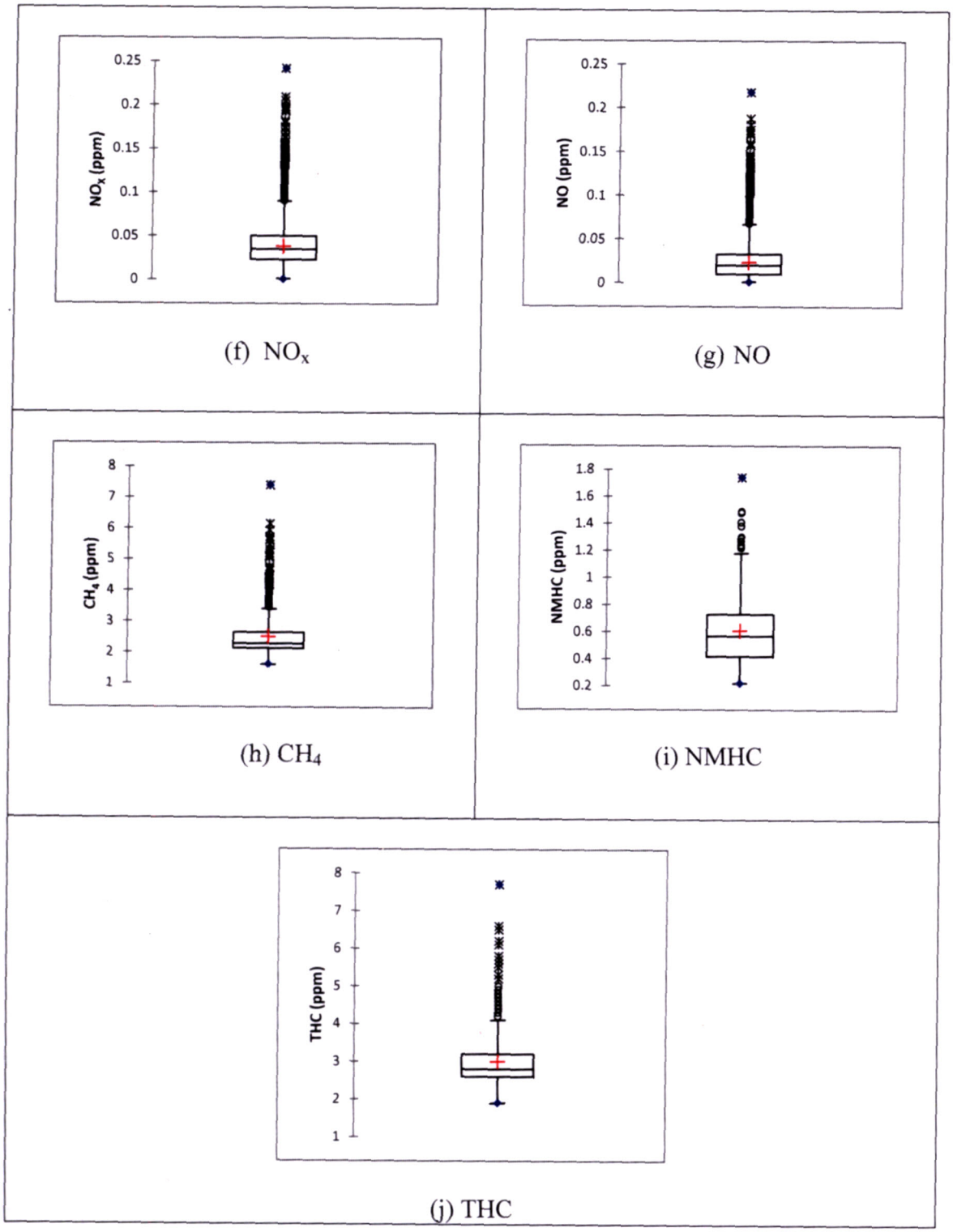



Figure 4.5. Continued

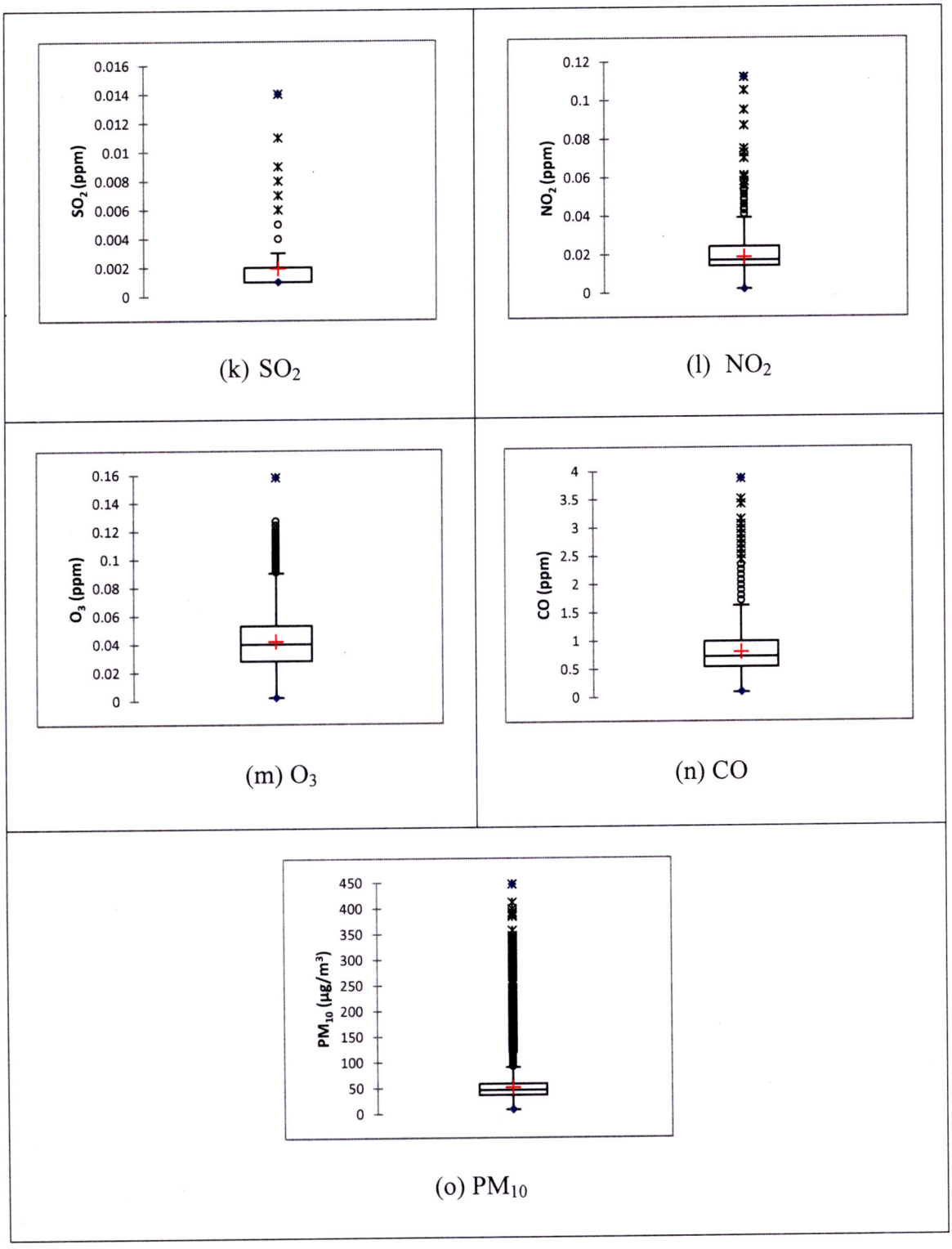



Figure 4.5. Continued

Table 4.5. Value of minimum, maximum, 1st quartile, median, 3rd quartile, mean, skewness and kurtosis for (a) meteorological parameters and (b) pollutants parameters for MPR

| Statistic | WS (km/hr) | WD (°) | Temp (°C) | UVB ($J/m^2h$) | Humidity (%) |
|---|---|---|---|---|---|
| Minimum | 0.40 | 2 | 20.40 | 88 | 49 |
| Maximum | 65.33 | 360 | 39.80 | 995 | 105 |
| 1st Quartile | 4.23 | 263 | 31.40 | 780 | 88 |
| Median | 5.11 | 338 | 32.90 | 863 | 93 |
| 3rd Quartile | 6.06 | 355 | 34.20 | 941 | 97 |
| Mean | 5.28 | 297 | 32.77 | 842 | 92 |
| Skewness (Pearson) | 7.6135 | -1.5042 | -0.3576 | -1.8167 | -1.1016 |
| Kurtosis (Pearson) | 191.3960 | 1.4053 | 0.8108 | 6.1348 | 1.6861 |

Table 4.5. Continued

| Statistic | $NO_X$ (ppm) | NO (ppm) | $CH_4$ (ppm) | NMHC (ppm) | THC (ppm) | $SO_2$ (ppm) | $NO_2$ (ppm) | $O_3$ (ppm) | CO (ppm) | $PM_{10}$ ($\mu g/m^3$) |
|---|---|---|---|---|---|---|---|---|---|---|
| Minimum | 0.001 | 0.001 | 1.62 | 0.22 | 1.90 | 0.001 | 0.002 | 0.002 | 0.09 | 8 |
| Maximum | 0.242 | 0.218 | 7.40 | 1.74 | 7.70 | 0.014 | 0.112 | 0.158 | 3.87 | 446 |
| 1st Quartile | 0.023 | 0.010 | 2.14 | 0.42 | 2.60 | 0.001 | 0.014 | 0.028 | 0.54 | 36 |
| Median | 0.035 | 0.020 | 2.29 | 0.57 | 2.80 | 0.001 | 0.017 | 0.040 | 0.72 | 46 |
| 3rd Quartile | 0.050 | 0.033 | 2.65 | 0.73 | 3.20 | 0.002 | 0.024 | 0.053 | 0.99 | 58 |
| Mean | 0.038 | 0.024 | 2.50 | 0.61 | 2.99 | 0.002 | 0.019 | 0.042 | 0.80 | 51 |
| Skewness (Pearson) | 1.2384 | 1.5068 | 2.7647 | 0.9263 | 2.3126 | 1.7479 | 0.9167 | 0.5992 | 1.4023 | 4.1463 |
| Kurtosis (Pearson) | 3.3738 | 4.4088 | 10.4991 | 0.9388 | 8.2817 | 5.4292 | 2.3698 | 0.2268 | 5.0916 | 33.6158 |

Table 4.5 shows the statistical value of minimum, maximum, 1st quartile, median, 3rd quartile and mean for meteorological and pollutant parameters, respectively, during the entire study period from 2010 to 2015 for MPR in Malaysia. The ranges for the meteorological parameters obtained for MPR were: wind speed (0.40–65.33) km/hr, wind direction (2–360)°, temperature (20.40–39.80)°C, UVB (88–995) J/m$^2$h and humidity (49–105) % with mean values of wind speed, wind direction, temperature, UVB and humidity were 5.28 km/hr, 297°, 32.77°C, 842 J/m$^2$hr and 92%, respectively. Meanwhile, the median values of wind speed, wind direction, temperature, UVB and humidity were 5.11 km/hr, 338°, 32.90 °C, 863 J/m$^2$hr and 93%, respectively.

The ranges for the pollutant parameters obtained for MPR were: $NO_x$ (0.001–0.242) ppm, NO (0.001–0.218) ppm, $CH_4$ (1.62–7.40) ppm, NMHC (0.22–1.74) ppm, THC (1.90–7.70) ppm, $SO_2$ (0.001–0.014) ppm, $NO_2$ (0.002–0.112) ppm, $O_3$ (0.002–0.158) ppm, CO (0.09–3.87) ppm and $PM_{10}$ (8–446) μg/m$^3$ with mean concentrations of $NO_x$, NO, $CH_4$, NMHC, THC, $SO_2$, $NO_2$, $O_3$, CO and $PM_{10}$ were 0.038 ppm, 0.024 ppm, 2.50 ppm, 0.61 ppm, 2.99 ppm, 0.002 ppm, 0.019 ppm, 0.042 ppm, 0.80 ppm and 51 μg/m$^3$, respectively. Meanwhile, the median values of $NO_x$, NO, $CH_4$, NMHC, THC, $SO_2$, $NO_2$, $O_3$, CO and $PM_{10}$ were 0.035 ppm, 0.02 ppm, 2.29 ppm, 0.57 ppm, 2.80 ppm, 0.002 ppm, 0.017 ppm, 0.04 ppm, 0.72 ppm and 46 μg/m$^3$,respectively.

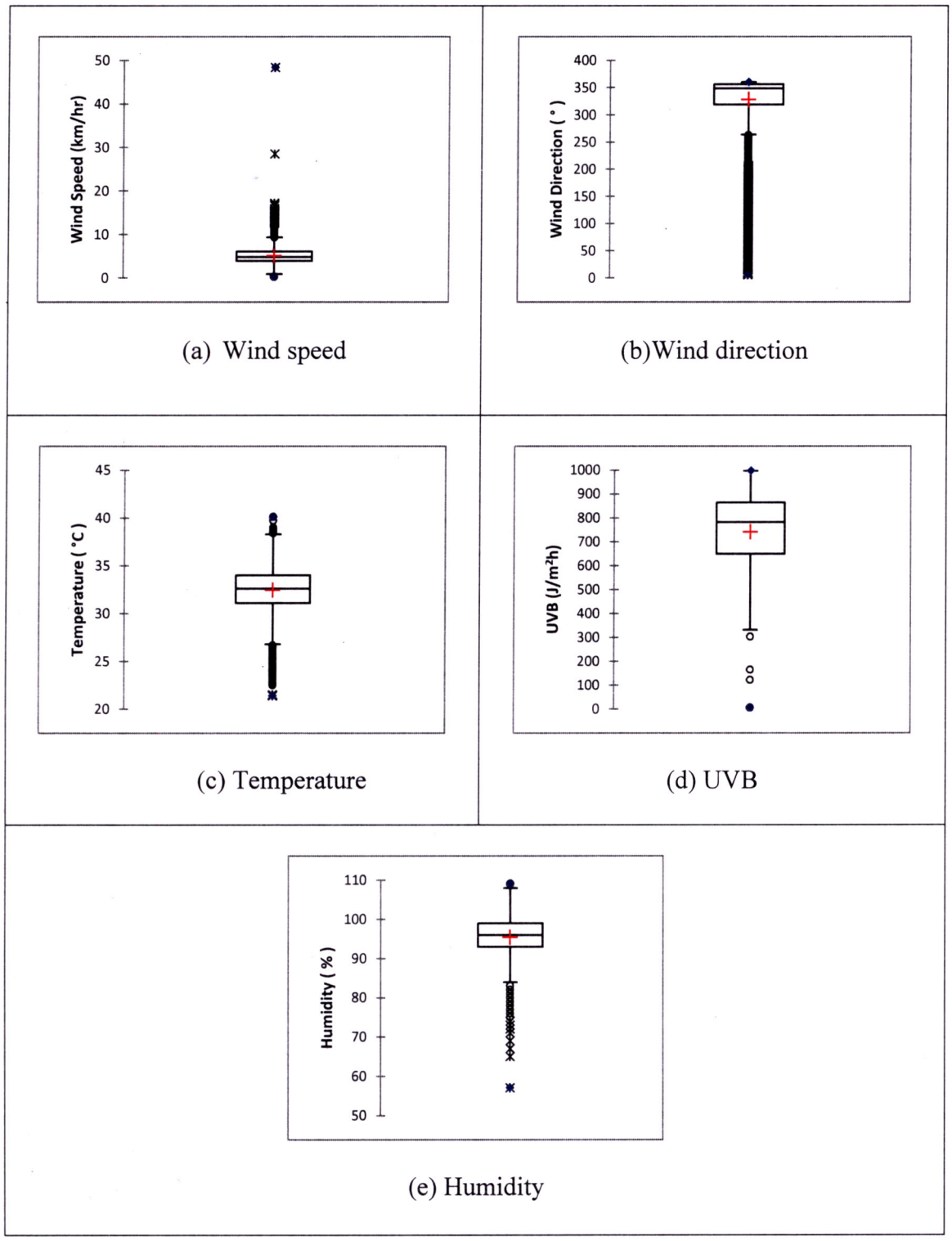



Figure 4.6. Boxplots of (a) wind speed, (b) wind direction, (c) temperature, (d) UVB, (e) humidity (f) $NO_x$, (g) NO, (h) $CH_4$, (i) NMHC, (j) THC, (k) $SO_2$, (l) $NO_2$, (m) $O_3$, (n) CO, and (o) $PM_{10}$ for LPR

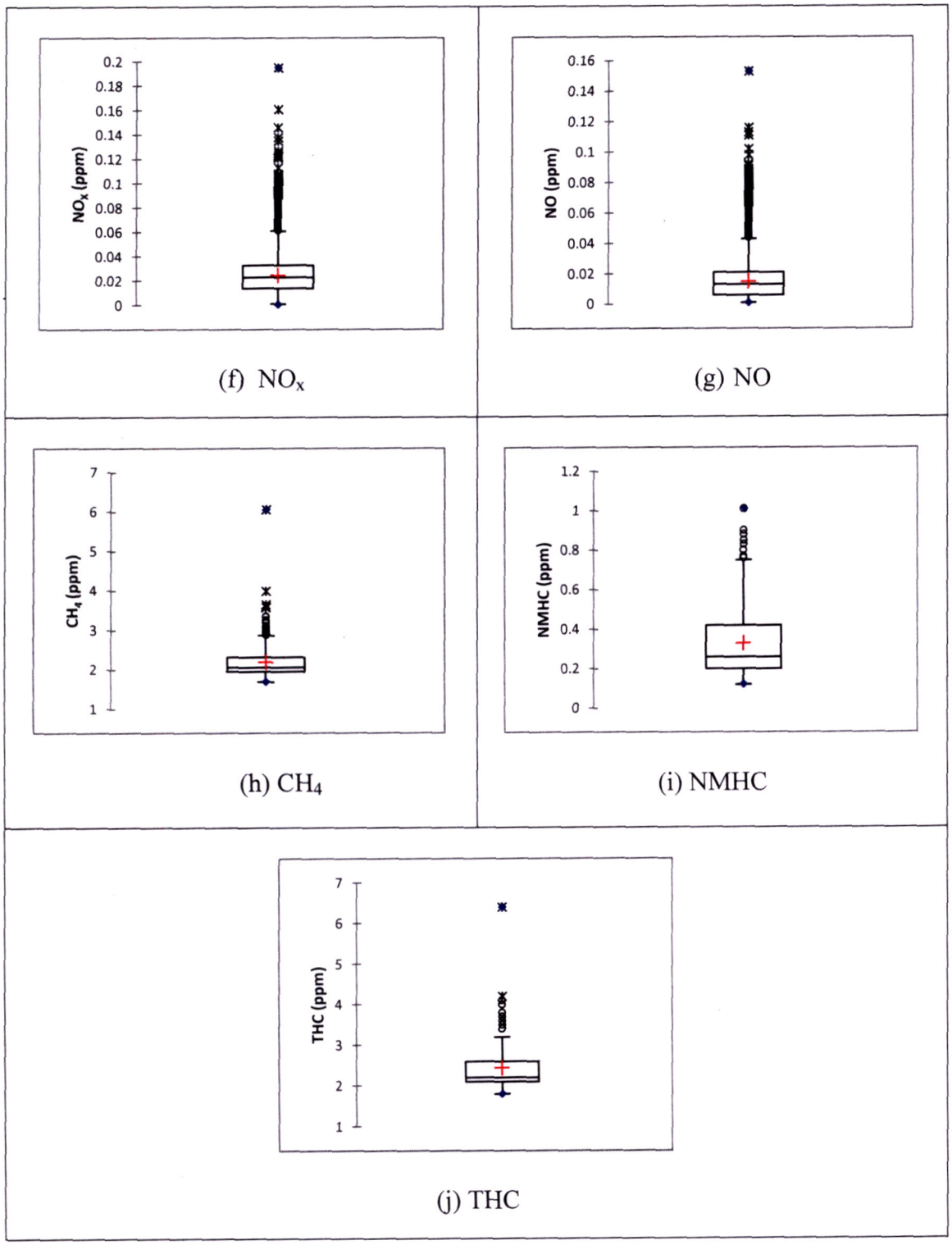



Figure 4.6. Continued

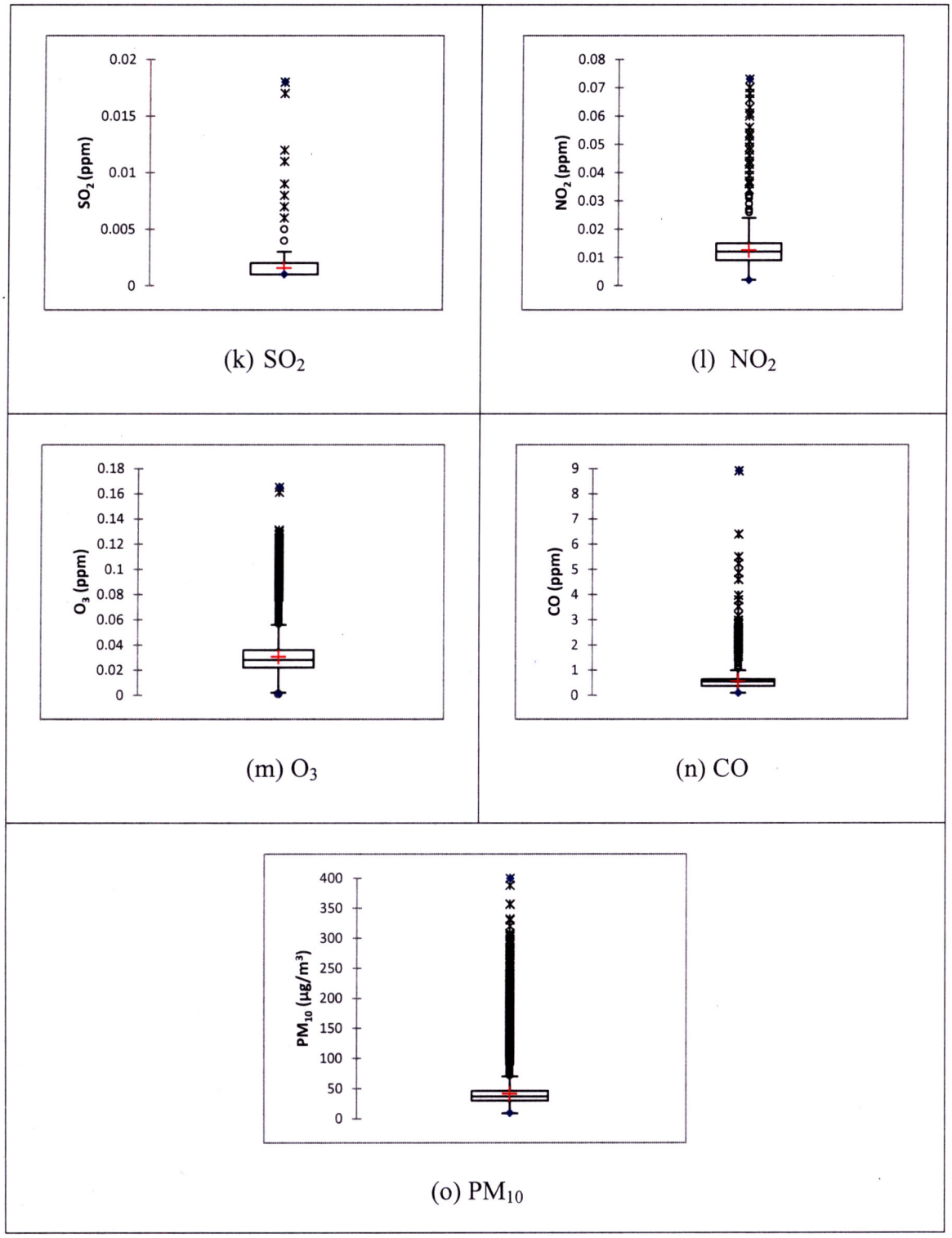



Figure 4.6. Continued

Table 4.6. Value of minimum, maximum, 1st quartile, median, 3rd quartile, mean, skewness and kurtosis for (a) meteorological parameters and (b) pollutants parameters for LPR

| Statistic | WS (km/hr) | WD (°) | Temp (°C) | UVB ($J/m^2h$) | Humidity (%) |
|---|---|---|---|---|---|
| Minimum | 0.28 | 6 | 21.40 | 5 | 57 |
| Maximum | 48.40 | 360 | 40.10 | 997 | 109 |
| 1st Quartile | 3.92 | 319 | 31.10 | 649 | 93 |
| Median | 4.84 | 348 | 32.60 | 782 | 96 |
| 3rd Quartile | 6.08 | 356 | 34.00 | 864 | 99 |
| Mean | 5.06 | 328 | 32.44 | 740 | 95 |
| Skewness (Pearson) | 1.5133 | -3.0099 | -0.4893 | -1.2447 | -1.0601 |
| Kurtosis (Pearson) | 19.5206 | 11.5498 | 0.5257 | 2.0412 | 1.8255 |

Table 4.6. Continued

| Statistic | $NO_X$ (ppm) | NO (ppm) | $CH_4$ (ppm) | NMHC (ppm) | THC (ppm) | $SO_2$ (ppm) | $NO_2$ (ppm) | $O_3$ (ppm) | CO (ppm) | $PM_{10}$ ($\mu g/m^3$) |
|---|---|---|---|---|---|---|---|---|---|---|
| Minimum | 0.001 | 0.001 | 1.69 | 0.12 | 1.80 | 0.001 | 0.002 | 0.001 | 0.090 | 9 |
| Maximum | 0.195 | 0.153 | 6.05 | 1.01 | 6.40 | 0.018 | 0.073 | 0.165 | 8.910 | 399 |
| 1st Quartile | 0.014 | 0.006 | 1.95 | 0.20 | 2.10 | 0.001 | 0.009 | 0.022 | 0.360 | 30 |
| Median | 0.023 | 0.013 | 2.06 | 0.26 | 2.20 | 0.001 | 0.012 | 0.028 | 0.540 | 37 |
| 3rd Quartile | 0.033 | 0.021 | 2.32 | 0.42 | 2.60 | 0.002 | 0.015 | 0.036 | 0.630 | 46 |
| Mean | 0.025 | 0.015 | 2.19 | 0.33 | 2.44 | 0.002 | 0.013 | 0.030 | 0.551 | 41 |
| Skewness (Pearson) | 1.0381 | 1.4241 | 3.6978 | 1.4411 | 2.6063 | 2.9964 | 0.9714 | 1.7061 | 4.5108 | 4.7581 |
| Kurtosis (Pearson) | 3.1368 | 4.2944 | 24.5966 | 1.6736 | 12.7391 | 27.1919 | 3.6200 | 6.2201 | 76.9946 | 41.0222 |

Table 4.6 shows the statistical value of minimum, maximum, 1st quartile, median, 3rd quartile and mean for meteorological and pollutant parameters, respectively, during the entire study period from 2010 to 2015 for LPR in Malaysia. The ranges for the meteorological parameters obtained for LPR were: wind speed (0.28–48.40) km/hr, wind direction (6–360)°, temperature (21.40–40.10)°C, UVB (5–997) J/m$^2$h and humidity (57–109) % with mean values of wind speed, wind direction, temperature, UVB and humidity were 5.06 km/hr, 328°, 32.44 °C, 740 J/m$^2$hr and 95%, respectively. Meanwhile, the median values of wind speed, wind direction, temperature, UVB and humidity were 4.84 km/hr, 348°, 32.60 °C, 782 J/m$^2$hr and 96%, respectively.

The ranges for the pollutant parameters obtained for LPR were: $NO_x$ (0.001-0.195) ppm, NO (0.001-0.153) ppm, $CH_4$ (1.69-6.05) ppm, NMHC (0.12-1.01) ppm, THC (1.80-6.40) ppm, $SO_2$ (0.001-0.018) ppm, $NO_2$ (0.002-0.073) ppm, $O_3$ (0.001-0.165) ppm, CO (0.09-8.91) ppm and $PM_{10}$ (9-399) μg/m$^3$ with mean concentrations of $NO_x$, NO, $CH_4$, NMHC, THC, $SO_2$, $NO_2$, $O_3$, CO and $PM_{10}$ were 0.025 ppm, 0.015 ppm, 2.19 ppm, 0.33 ppm, 2.44 ppm, 0.002 ppm, 0.013 ppm, 0.03 ppm, 0.551 ppm and 41 μg/m$^3$, respectively. Meanwhile, the median values of $NO_x$, NO, $CH_4$, NMHC, THC, $SO_2$, $NO_2$, $O_3$, CO and $PM_{10}$ were 0.023 ppm, 0.013 ppm, 2.06 ppm, 0.26 ppm, 2.20 ppm, 0.001 ppm, 0.012 ppm, 0.028 ppm, 0.54 ppm and 37 μg/m$^3$, respectively.

Among the concentrations of pollutants, $PM_{10}$ pollutant gave the highest mean and median concentration value as compared to other pollutants. All concentrations of air pollutants showed higher value of mean than median. It was also shown that the distributions of pollutant were skewed to the right and the extreme events happened (Wahab et al., 2016; Sansuddin et al., 2011; Zakaria & Noor, 2018). This coincided

with the positive number of the skewness for the pollutant parameters in each of the region, as shown in Table 4.4 until Table 4.6. According to Yusof et al. (2009), positive numbers of the skewness reveal the existences of high pollution levels. Therefore, in this study, the descriptive analysis revealed that there were extreme events that occurred during the study period that affected the concentration of pollutants in the atmosphere, especially pollutant of $PM_{10}$ that gave the highest concentration among others.

## 4.5 Trend of air pollutants

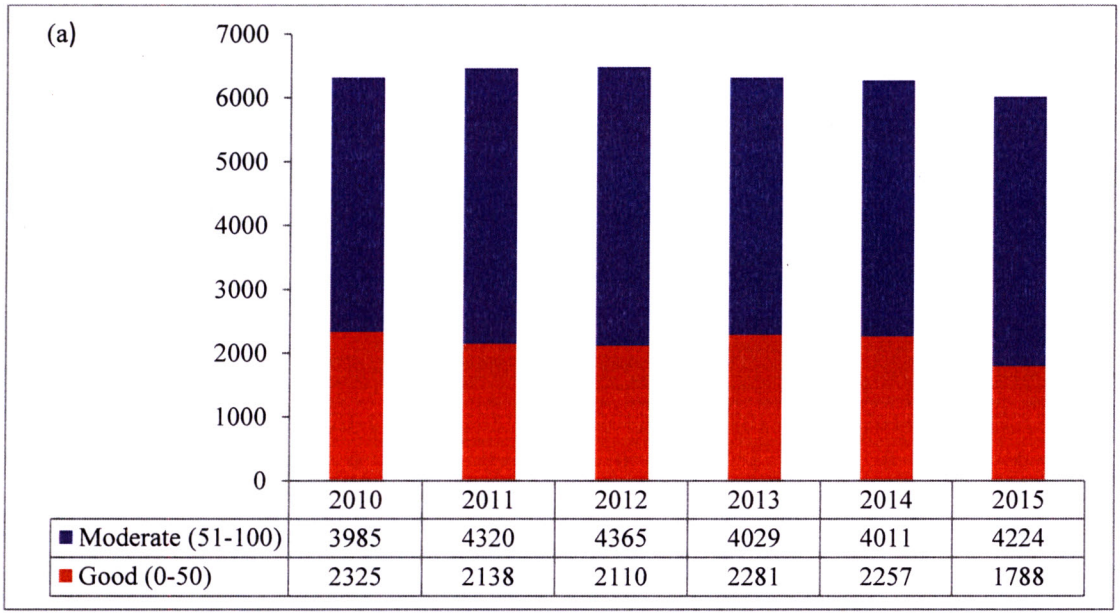
Annual air quality status, annual variation trend of main pollutants (CO, $NO_2$, $SO_2$, $O_3$ and $PM_{10}$) and monthly variation trend of $PM_{10}$ were analysed to determine the variation of pollutants in the Malaysian atmosphere from 2010 to 2015.

### 4.5.1 Air quality status in HPR, MPR and LPR

Figure 4.7 until Figure 4.9 show the annual frequencies of air quality status for HPR, MPR and LPR based on the API that was categorised into good, moderate, unhealthy, very unhealthy, hazardous and emergency.

#### 4.5.1.1 Air quality status in HPR

Figure 4.7 shows the annual frequencies of air quality status for HPR that consisted of good, moderate, unhealthy, very unhealthy, hazardous and emergency, which vary over the years. It is shown that moderate levels of API has the highest frequencies followed by a good level of API from 2010 to 2015 with the API annual frequencies are 3985, 4320, 4365, 4029, 4011 and 4224, respectively.

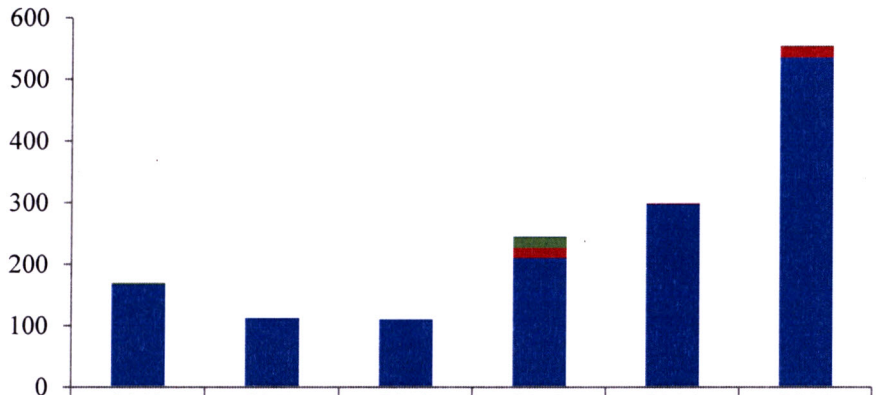



| | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
|---|---|---|---|---|---|---|
| Moderate (51-100) | 3985 | 4320 | 4365 | 4029 | 4011 | 4224 |
| Good (0-50) | 2325 | 2138 | 2110 | 2281 | 2257 | 1788 |

Figure 4.7. Annual frequencies of air quality status (a) Good, Moderate (b) Unhealthy, Very unhealthy, Hazardous, Emergency in Malaysia from 2010 to 2015 for HPR

The unhealthy level decreased in 2011 and 2012, and then continuously increased until 2015. It was obviously observed that in 2015, the unhealthy level of API became domain as compared to other levels of API (very unhealthy, hazardous, and emergency) with frequencies of unhealthy API level of 535, which was drastically increased from 2014 with frequency of 297. The highest API frequency of very unhealthy level was also visualised in 2015 with the frequency at 18, and followed by 2013 and 2014 with the API frequency at 16 and 1, respectively. Meanwhile, the highest frequencies for hazardous and emergency API levels were observed during 2013 and that year was the year with the worst air quality experienced by Malaysia, which was recorded as the highest frequency reading of the API.

(b)



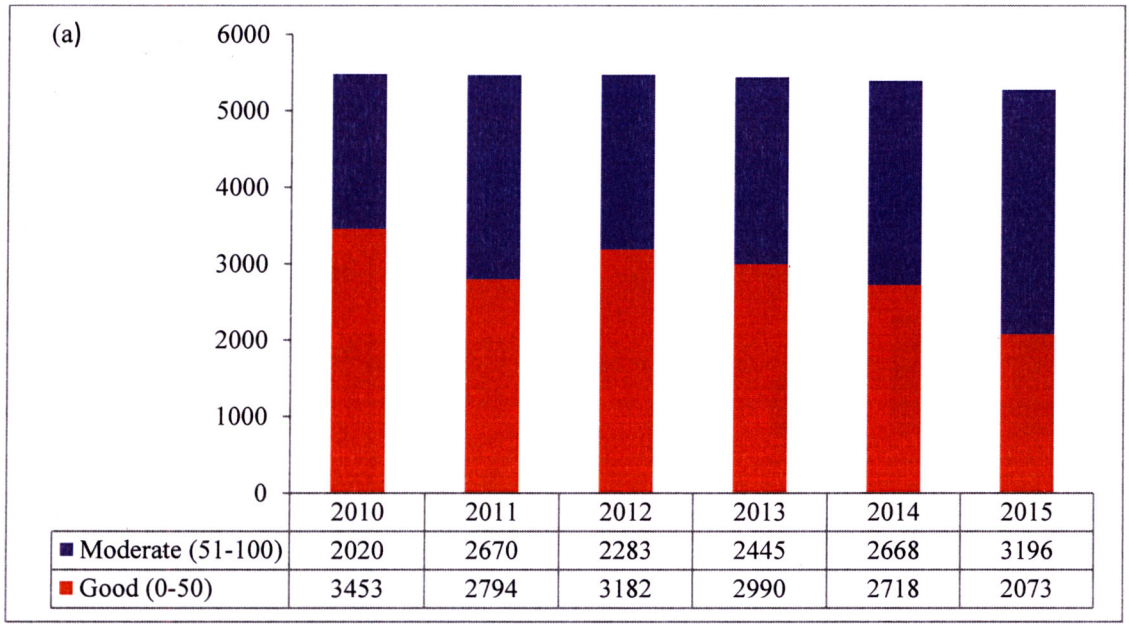

| | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
|---|---|---|---|---|---|---|
| Emergency (>500) | 0 | 0 | 0 | 1 | 0 | 0 |
| Hazardous | 2 | 0 | 0 | 17 | 0 | 1 |
| Very unhealthy | 0 | 0 | 0 | 16 | 1 | 18 |
| Unhealthy | 167 | 112 | 110 | 210 | 297 | 535 |

Figure 4.7. Continued

Generally, most of the time, air quality in the HPR was between good to moderate. However, Muar recorded one-day emergency API due to the transboundary haze pollution in 2013. Other stations, namely, Bukit Rambai (4 days), Klang (2 days), SM Tinggi Melaka (2 days), Port Dickson (2 days), Pasir Gudang (2 days), Muar (2 days), Kota Tinggi (2 days) and Shah Alam (1 day) experienced hazardous days in 2013. Besides, the Shah Alam station also recorded one day of hazardous day in 2015 and two days of hazardous day in 2010 at Muar station.

#### 4.5.1.2 Air quality status in MPR

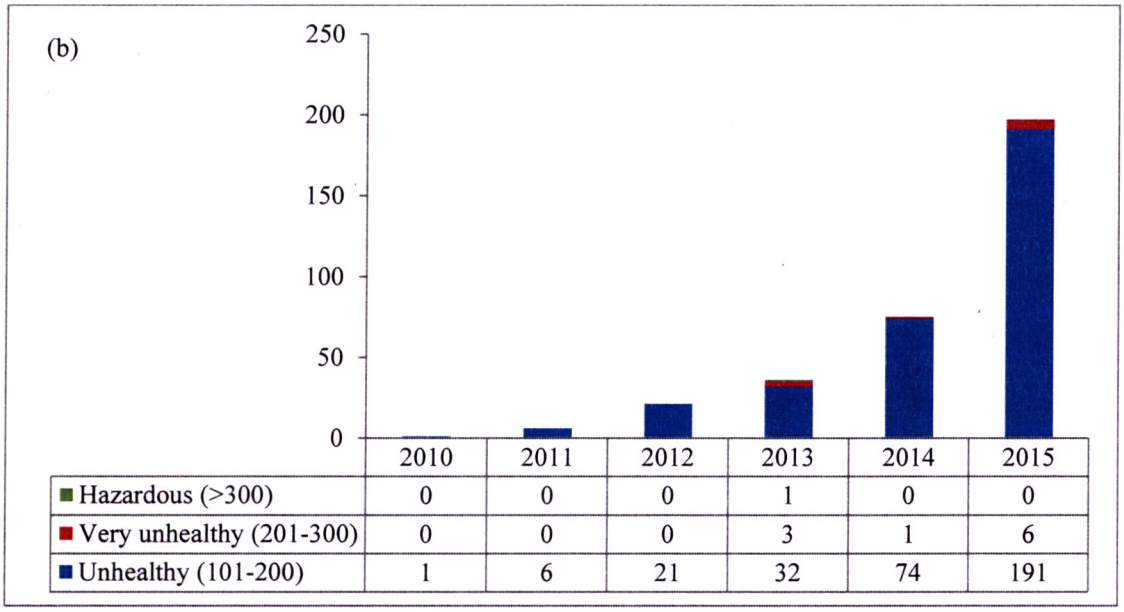
Figure 4.8 shows the annual frequencies of air quality status for MPR which consisted of good, moderate, unhealthy, very unhealthy and hazardous, which vary over the years. It is shown that a good level of API shows the highest frequencies followed by a moderate level of API from 2010 to 2014 with the API annual frequencies at 3453, 2794, 3182, 2990 and 2718, respectively, and vice versa during 2015, which shows that the moderate API level is higher than good API level with frequency of 3196.




| | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
|---|---|---|---|---|---|---|
| Moderate (51-100) | 2020 | 2670 | 2283 | 2445 | 2668 | 3196 |
| Good (0-50) | 3453 | 2794 | 3182 | 2990 | 2718 | 2073 |

Figure 4.8. Annual frequencies of air quality status (a) Good, Moderate (b) Unhealthy, Very unhealthy, Hazardous in Malaysia from 2010 to 2015 for MPR

The unhealthy level increased from 2010 to 2015. Like HPR, it was obviously observed that in 2015, the unhealthy level of API became domain as compared to other levels of API (very unhealthy and hazardous) with the frequencies of API unhealthy level at 191. The year of 2015 also showed the highest frequency of very unhealthy level as compared to other years with frequencies at 6 followed by the year 2013 and 2014 with frequencies 3 and 1, respectively. Malaysia experienced a hazardous level of the API only in 2013 for MPR.



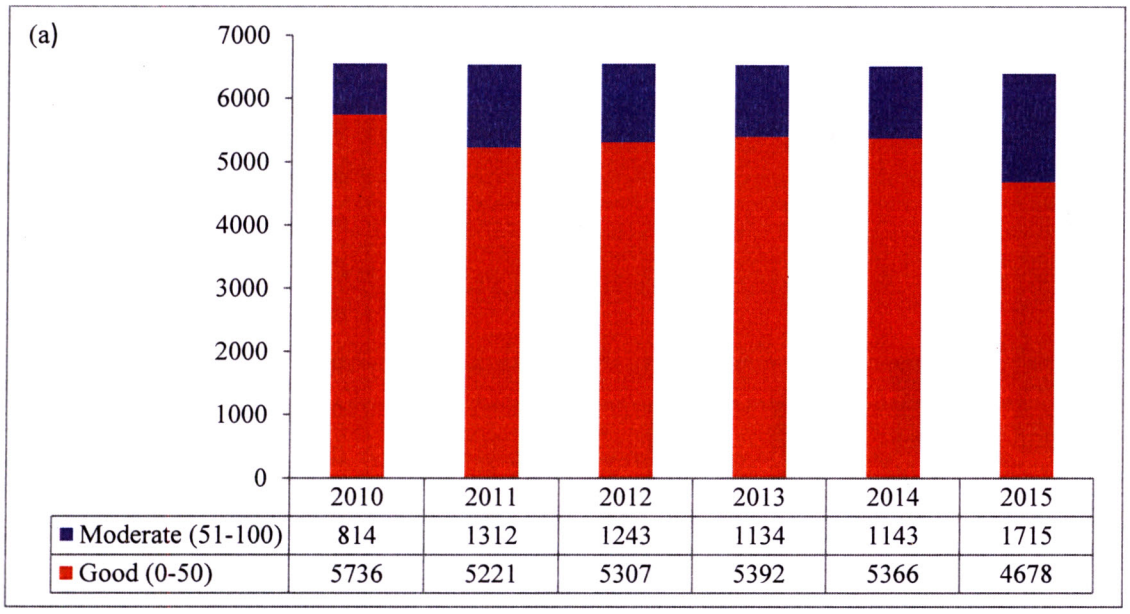

| | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
|---|---|---|---|---|---|---|
| Hazardous (>300) | 0 | 0 | 0 | 1 | 0 | 0 |
| Very unhealthy (201-300) | 0 | 0 | 0 | 3 | 1 | 6 |
| Unhealthy (101-200) | 1 | 6 | 21 | 32 | 74 | 191 |

Figure 4.8. Continued

The overall air quality for MPR was between good to moderate most of the time. However, Manjung recorded one hazardous day. Very unhealthy days were also recorded in other stations viz. Kuala Selangor (2 days), Pegoh (1 day) in 2013; Manjung (1 day) in 2014; Perai, ILP Kangar, USM, Sg Petani, Langkawi and Alor Setar (1 day for each station).

#### 4.5.1.3 Air quality status in LPR

Figure 4.9 shows the annual frequencies of air quality status for LPR which consisted of good, moderate, unhealthy, very unhealthy and hazardous, which vary over the years. It is shown that a good level of API showed the highest frequencies followed by moderate, unhealthy and very unhealthy with the total frequencies of good API level from 2010 to 2015 were 5736, 5221, 5307, 5392, 5366 and 4678, respectively.

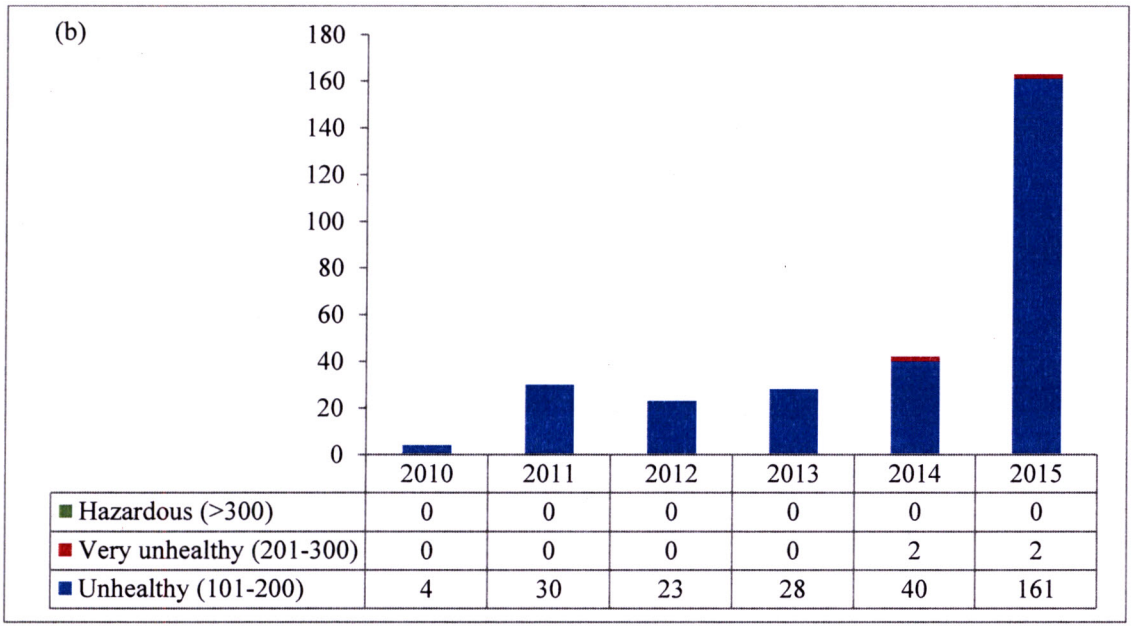



| | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
|---|---|---|---|---|---|---|
| Moderate (51-100) | 814 | 1312 | 1243 | 1134 | 1143 | 1715 |
| Good (0-50) | 5736 | 5221 | 5307 | 5392 | 5366 | 4678 |

Figure 4.9. Annual frequencies of air quality status (a) Good, Moderate (b) Unhealthy, Very unhealthy, Hazardous in Malaysia from 2010 to 2015 for LPR

The unhealthy level increased from 2011 to 2010. However, it was decreased in 2012 and continuously increased until 2015. It is obviously observed that in 2015, the unhealthy level of API became dominant as compared to other levels of API (very unhealthy and hazardous) with the unhealthy level of API at 161. The year 2014 and 2015 experienced two days of very unhealthy API level. None of the hazardous API level was observed from 2010 to 2015 for LPR.

The overall air quality for LPR was between good to moderate most of the time with the best API level. However, very unhealthy days were also recorded at other stations viz. Sibu (2 days) in 2014; Kuching (1 day) and Kota Samarahan (1 day) in 2015.




| | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
|---|---|---|---|---|---|---|
| Hazardous (>300) | 0 | 0 | 0 | 0 | 0 | 0 |
| Very unhealthy (201-300) | 0 | 0 | 0 | 0 | 2 | 2 |
| Unhealthy (101-200) | 4 | 30 | 23 | 28 | 40 | 161 |

Figure 4.9. Continued

Each region experienced a very unhealthy level of air quality status. However, HPR showed the most unhealthy air quality status followed by MPR and LPR. According to Awang et al. (2000), $PM_{10}$ is the main contributor to the API in the atmosphere. The main factor which contributed to the higher API value present was probably due to presence of $PM_{10}$ in the air quality, which was mainly due to the massive forest fire in Indonesia (Ahmat et al., 2015), sea spray originated from the ocean (Ismail et al., 2015) and due to the emission from industrial activities and vehicles (Razak et al., 2014).

### 4.5.2 Annual variation of CO, $NO_2$, $SO_2$, $O_3$ and $PM_{10}$ from 2010 to 2015 by areas

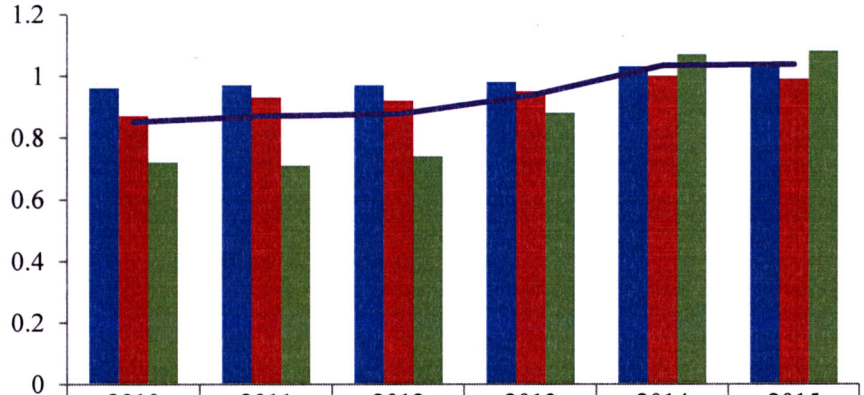
Figure 4.10 until Figure 4.14 show the annual average concentrations of CO, $NO_2$, $SO_2$, $O_3$ and $PM_{10}$ from the year 2010 to 2015 for HPR, MPR and LPR in Malaysia. MAQS stipulated for daily 24-hour $NO_2$ and $SO_2$ were 0.04 ppm and 0.03 ppm, respectively. Meanwhile, MAQS stipulated for the annual average $PM_{10}$ concentration

of 40 $\mu g/m^3$. Since the concentration of CO and $O_3$ was only limited to 8-hour monitoring basis, there was no limit comparison for both of these pollutants.

### 4.5.2.1 Annual variation of CO from 2010 to 2015 for HPR, MPR and LPR by areas

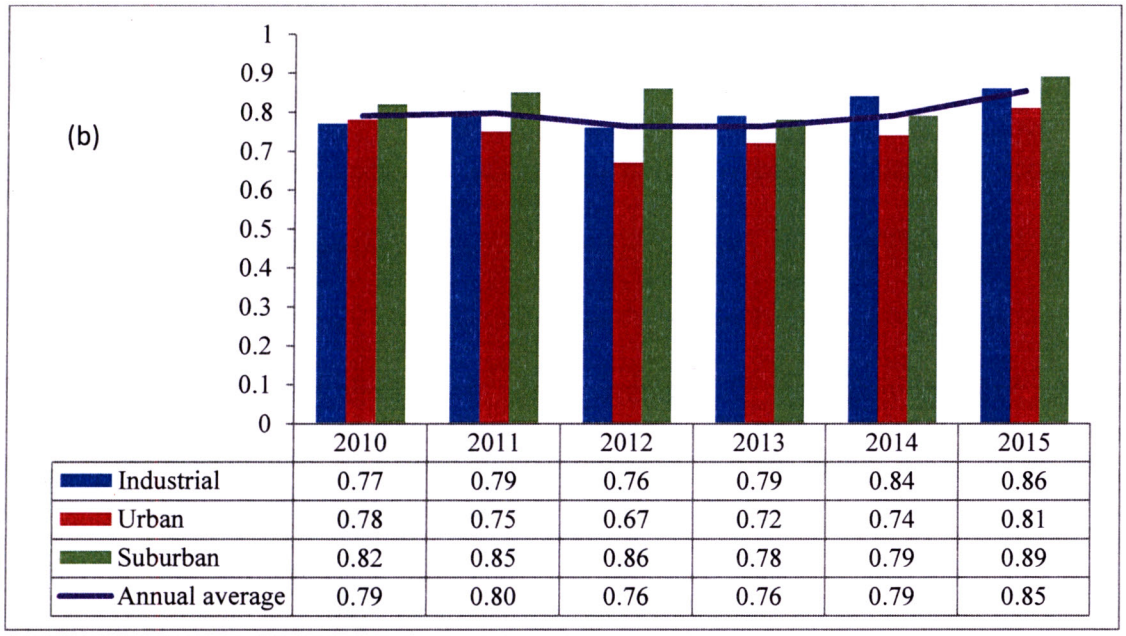
Figure 4.10 (a) to Figure 4.10 (c) show the annual average CO concentration by year, from 2010 to 2015 for HPR MPR and LPR in Malaysia, respectively. For the annual average CO concentration in HPR, it was visualised as increased annual average concentration of CO from 2010 until 2015. While for MPR, the annual average CO concentration was found to be increased in 2011 from 2010. However, it decreased in 2012 and 2013 with a constant concentration of CO annual average (0.76 ppm). The annual average CO concentration continuously increased in a year from 2014 to 2015. For LPR, the annual average CO concentration increased in 2011 and 2012 with a constant concentration of CO annual average (0.52 ppm) and decreased in 2013. It is became increasingly increased in 2014 and continued until 2015.

(a)



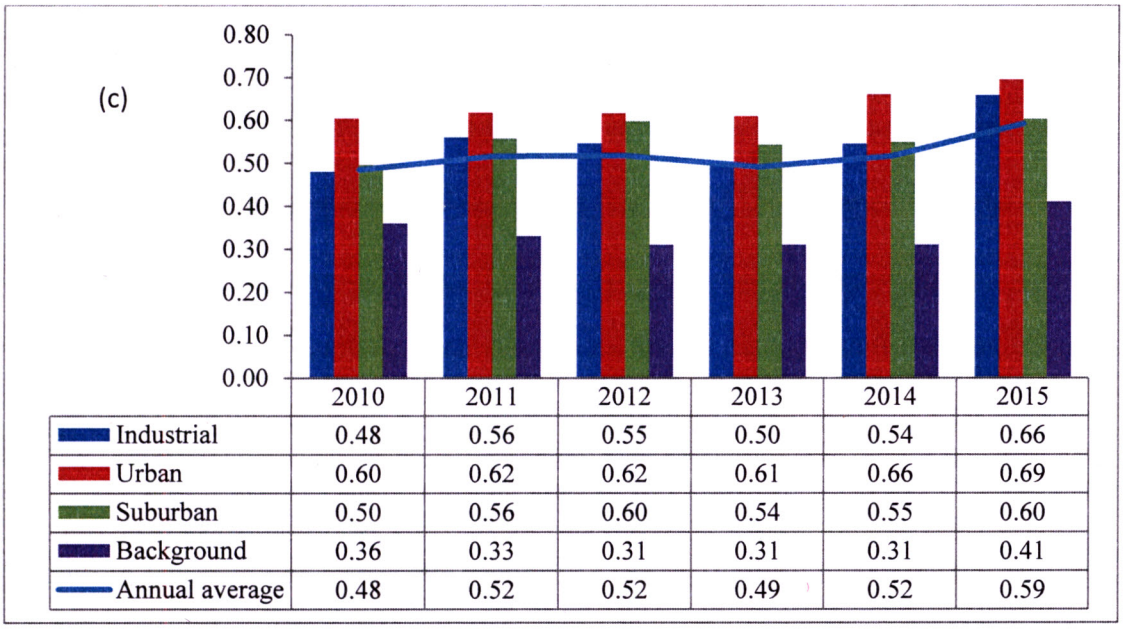

| | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
|---|---|---|---|---|---|---|
| Industrial | 0.96 | 0.97 | 0.97 | 0.98 | 1.03 | 1.04 |
| Urban | 0.87 | 0.93 | 0.92 | 0.95 | 1.00 | 0.99 |
| Suburban | 0.72 | 0.71 | 0.74 | 0.88 | 1.07 | 1.08 |
| Annual average | 0.85 | 0.87 | 0.88 | 0.94 | 1.03 | 1.04 |

Figure 4.10. Annual average concentration of CO by year (2010-2015) for (a) HPR (b) MPR and (c) LPR in Malaysia




| | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
|---|---|---|---|---|---|---|
| Industrial | 0.77 | 0.79 | 0.76 | 0.79 | 0.84 | 0.86 |
| Urban | 0.78 | 0.75 | 0.67 | 0.72 | 0.74 | 0.81 |
| Suburban | 0.82 | 0.85 | 0.86 | 0.78 | 0.79 | 0.89 |
| Annual average | 0.79 | 0.80 | 0.76 | 0.76 | 0.79 | 0.85 |

Figure 4.10. Continued




| | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
|---|---|---|---|---|---|---|
| Industrial | 0.48 | 0.56 | 0.55 | 0.50 | 0.54 | 0.66 |
| Urban | 0.60 | 0.62 | 0.62 | 0.61 | 0.66 | 0.69 |
| Suburban | 0.50 | 0.56 | 0.60 | 0.54 | 0.55 | 0.60 |
| Background | 0.36 | 0.33 | 0.31 | 0.31 | 0.31 | 0.41 |
| Annual average | 0.48 | 0.52 | 0.52 | 0.49 | 0.52 | 0.59 |

Figure 4.10. Continued

In general, HPR recorded the highest annual average CO concentration as compared to MPR and LPR. The higher CO concentration present in the industrial area at HPR was due to the location of air quality monitoring stations being in the industrial areas. Guerreiro et al. (2016) mentioned that industrial areas gave the highest levels of CO, which was usually found during emissions downwind from industries. For instance,

the air quality monitoring station located at Petaling Jaya in one of the stations located at industrial area. In Petaling Jaya, there were about 2780 industrial projects, comprising numerous sectors, for instance, the electrical industry, chemical production, machine manufacturing and fabricated metal products (Majlis Bandaraya Petaling Jaya (MBPJ), 2011). Therefore, perhaps emission downwind from these industries contributed to the highest annual average CO concentration present at the HPR as mentioned by a study done by Guerreiro et al. (2016) before.

Besides, according to Razak et al. (2014), the Petaling Jaya station is also encircled with other industrial and residential areas, for example, Subang, Shah Alam and Kuala Lumpur. Instead of being affected by the highest annual average CO concentration presence due to its location in the industrial area, and was near to the national capital of Malaysia of Kuala Lumpur, which also caused the highest annual average CO concentration presence among the regions in Malaysia. The distance between Petaling Jaya and Kuala Lumpur is about 15 km, which is highly populated. In addition, Mutalib et al. (2013) mentioned that this station is also surrounded by very busy streets which include mostly industrial, residential and commercial areas.

Therefore, there will be a high emissions level of the diesel engines from buses, trucks, motorcars and motorcycles, whereby incomplete combustion inside the exhaust of motor vehicles produces CO emissions that contribute to the presence of CO in the atmosphere. Besides, in Selangor, Petaling Jaya is the largest and most developed district. Perhaps the industrial factors and other factors, such as transportation and urbanisation, caused the Petaling Jaya station to face the highest annual average of CO concentration. Besides industrial areas, Guerreiro et al. (2016) mentioned that urban areas give the highest levels of CO during peak hour in urban areas. According to the

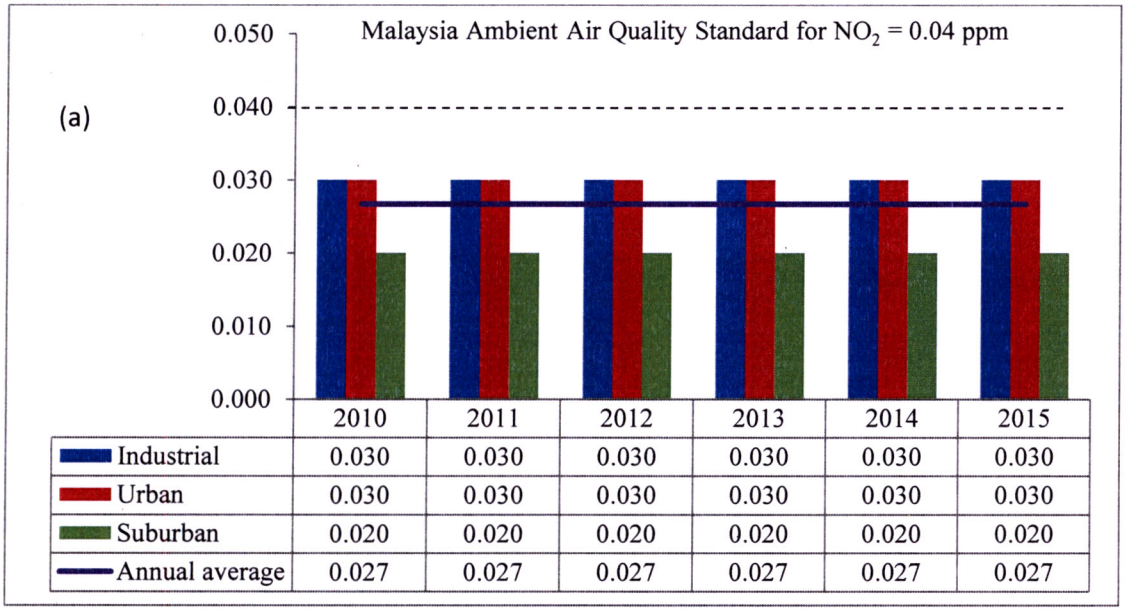
report done by DOE ($2015_a$), there was 95% emission of CO in 2015 with the main emission source from the motor vehicles in urban areas. Therefore, some of the air quality monitoring stations located at the urban area categorised in HPR, namely Klang, Kota Tinggi, SM Tinggi Melaka, Port Dickson, Seremban, Putrajaya, Cheras, Batu Muda and Shah Alam might be influenced by this scenario that contributed to the high annual average CO concentration in 2015.

In addition, the high CO concentration present in suburban area at HPR was probably due to the locality of air quality monitoring stations. For instance, Muar is one of the air quality monitor stations located in the suburban area. It is located adjacent to the tropical peatland at the Riau province coastal region, which has served as a major source of transboundary haze. The peatland fires emitted more intense carbon emission than vegetation fires. This indirectly affected the air quality at Muar station as a consequent of aerosol particles transported eastward from the wildfire during the Southwest monsoon. According to the study done by Kuwata et al. (2018), instead of $PM_{10}$, the CO pollutant was also present in high concentration during the wildfire haze episodes. Besides, the station is located near to one of the busiest shipping lanes, namely, the Straits of Melaka. Therefore, there is CO emission from the boats along with the other emission from the transportation at the land area. According to Mutalib et al. (2013), incomplete fuel burning due to vehicle engine malfunction which will emit higher CO emission. It was proven by a study done by Desai (2018), whereby CO is a major pollutant emitted due to the incomplete combustion of vehicle engines.

#### 4.5.2.2 Annual variation of $NO_2$ from 2010 to 2015 for HPR, MPR and LPR by areas

Figure 4.11 (a) to Figure 4.11 (c) show the annual average $NO_2$ concentration by year, from 2010 to 2015, for HPR MPR and LPR in Malaysia, respectively. For the annual average $NO_2$ concentration in HPR, it visualised no significant change from 2010 to 2015 (0.027 ppm). While for MPR, there was no significant change from 2010 to 2013 (0.020 ppm) and slightly decreased in 2014. The annual average $NO_2$ concentration remained constant from 2014 to 2015 (0.017 ppm). For LPR, the annual average of $NO_2$ concentration increased from 2010 to 2012. However, the annual average $NO_2$ concentration slightly decreased in 2013 and remained constant until 2015 (0.011 ppm).

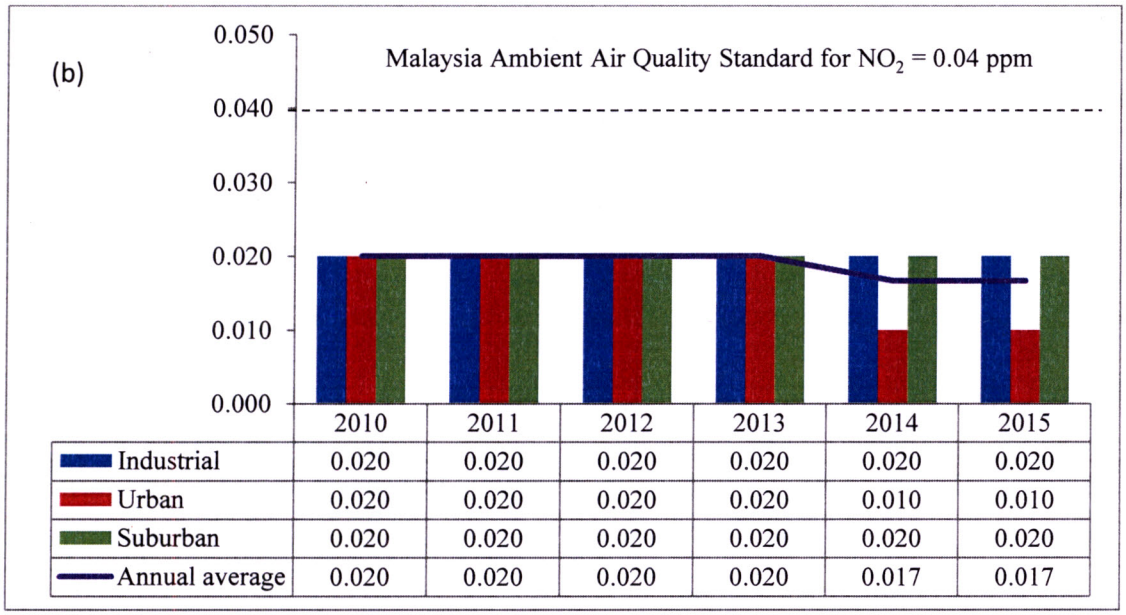
The overall trend on the annual average daily 24-hour in ambient air from 2010 to 2015 was well below the limit of 0.04 ppm. Comparing all areas at LPR, the urban area showed the highest annual average of $NO_2$ concentration among others with concentration of 0.016 ppm in 2015. However, that value was only slightly different from the highest annual average of $NO_2$ concentration recorded by the industrial and suburban areas. In contrast, the background area showed the lowest annual average of $NO_2$ concentration as compared to other areas, which normally recorded low concentration of pollutants. The overall trend on the annual average daily 24-hour in ambient air from 2010 to 2015 was well below the limit of 0.04 ppm.




| | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
|---|---|---|---|---|---|---|
| Industrial | 0.030 | 0.030 | 0.030 | 0.030 | 0.030 | 0.030 |
| Urban | 0.030 | 0.030 | 0.030 | 0.030 | 0.030 | 0.030 |
| Suburban | 0.020 | 0.020 | 0.020 | 0.020 | 0.020 | 0.020 |
| Annual average | 0.027 | 0.027 | 0.027 | 0.027 | 0.027 | 0.027 |

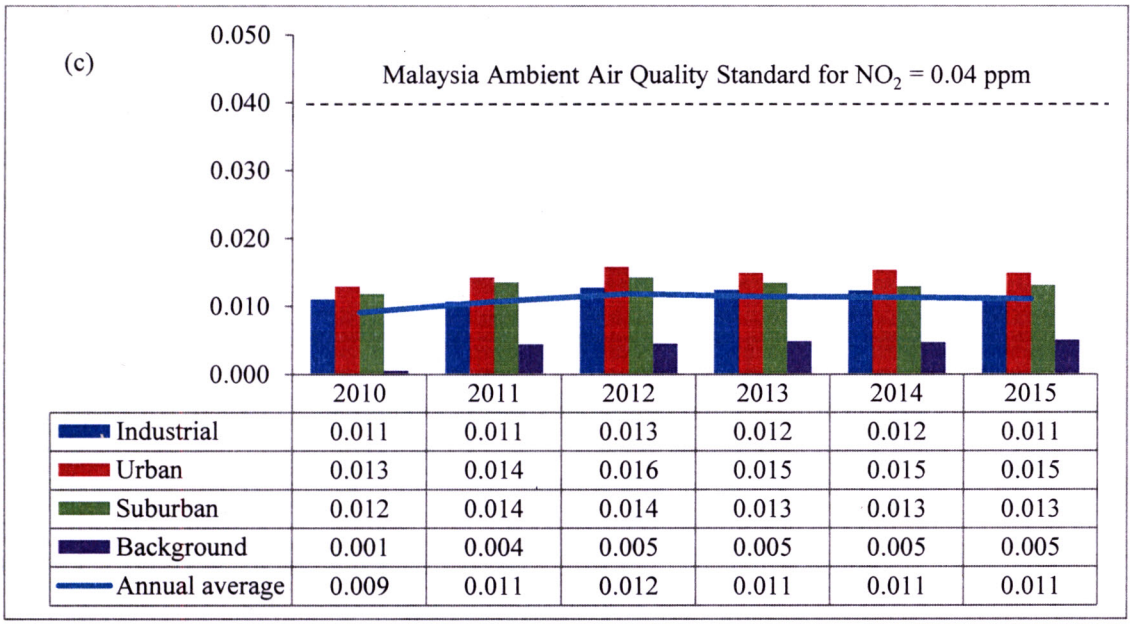
Figure 4.11 Annual average concentration of $NO_2$ by year (2010-2015) for (a) HPR (b) MPR and (c) LPR in Malaysia

In general, HPR recorded the highest annual average of $NO_2$ concentration as compared to MPR and LPR with none of them above the limit stipulated by MAQS (0.04 ppm). According to the DOE ($2015_a$), increasing the processes of combustion caused the $NO_2$ concentration to remain in the air especially in the industrial areas. Besides, DOE also reported that 65% were contributed through industrial, 27% from motor vehicles, 6% from the power plant and 2% from other sources which indicated emission load of $NO_2$. For instance, Muar air quality monitoring station at the suburban area, which was categorised as HPR indirectly contributed to the presence of $NO_2$ since it was surrounded by a small number of industries.




| | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
|---|---|---|---|---|---|---|
| Industrial | 0.020 | 0.020 | 0.020 | 0.020 | 0.020 | 0.020 |
| Urban | 0.020 | 0.020 | 0.020 | 0.020 | 0.010 | 0.010 |
| Suburban | 0.020 | 0.020 | 0.020 | 0.020 | 0.020 | 0.020 |
| Annual average | 0.020 | 0.020 | 0.020 | 0.020 | 0.017 | 0.017 |

Figure 4.11. Continued




| | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
|---|---|---|---|---|---|---|
| Industrial | 0.011 | 0.011 | 0.013 | 0.012 | 0.012 | 0.011 |
| Urban | 0.013 | 0.014 | 0.016 | 0.015 | 0.015 | 0.015 |
| Suburban | 0.012 | 0.014 | 0.014 | 0.013 | 0.013 | 0.013 |
| Background | 0.001 | 0.004 | 0.005 | 0.005 | 0.005 | 0.005 |
| Annual average | 0.009 | 0.011 | 0.012 | 0.011 | 0.011 | 0.011 |

Figure 4.11. Continued

Besides, the Petaling Jaya air quality monitoring station, which is located at an industrial area contributed to the high annual average of $NO_2$ concentration. This is correlated with the report issued by the DOE, whereby industrial is the dominant contributor, followed by motor vehicles, power plant and other sources which indicated an estimation emission load of $NO_2$. As discussed before, besides it is location, which was located at the industrial area, it was also surrounded by busy

roads, which made the Petaling Jaya station experienced the highest annual average $NO_2$ concentration than another regions. In addition, the Nilai station located in the industrial area also contributes to the high annual average of $NO_2$ concentration. The growth of industry in Nilai increased the population density due to jobs availability (Ul-Saufie et al., 2013). Besides, the Nilai station is located near to the air and rail network of transportations, such as the Kuala Lumpur International Airport (KLIA) and Kereta Api Tanah Melayu Railway (KTM) Network, which brought the urbanisation process in Nilai.

Besides, the high annual average of $NO_2$ concentration in the urban area is probably due to transportation. According to the DOE ($2015_a$), the increase in the number of motor vehicles caused the $NO_2$ concentration to remain in the air, especially in urban areas. This correlated with a study done by Dominick et al. (2012), whereby, the urban area were categorised as strong loading of $NO_2$, which implied to be the source of air pollution due to fuel combustion in automobile and aircraft. Besides, burning of nitrogen in fuel also contributed in producing a large proportion of $NO_2$ (Rößler et al., 2017).

Meanwhile SK Seberang Jaya air quality monitoring station, located at sub-urban area categorised as MPR, contributed to the high annual average of $NO_2$ concentration was probably due to the vehicle exhaust promoted in that area. Ismail et al. (2017) mentioned that there were road networks in Seberang Jaya that connect the north, south, east and Penang Island. Therefore, increasing the vehicle exhaust in that area will indirectly increase the annual average of $NO_2$ concentration present which was supported by Mohamad et al. (2015), whereby $NO_2$ was generally produced from vehicle exhaust. Meanwhile, Sibu air quality monitoring station, located at suburban

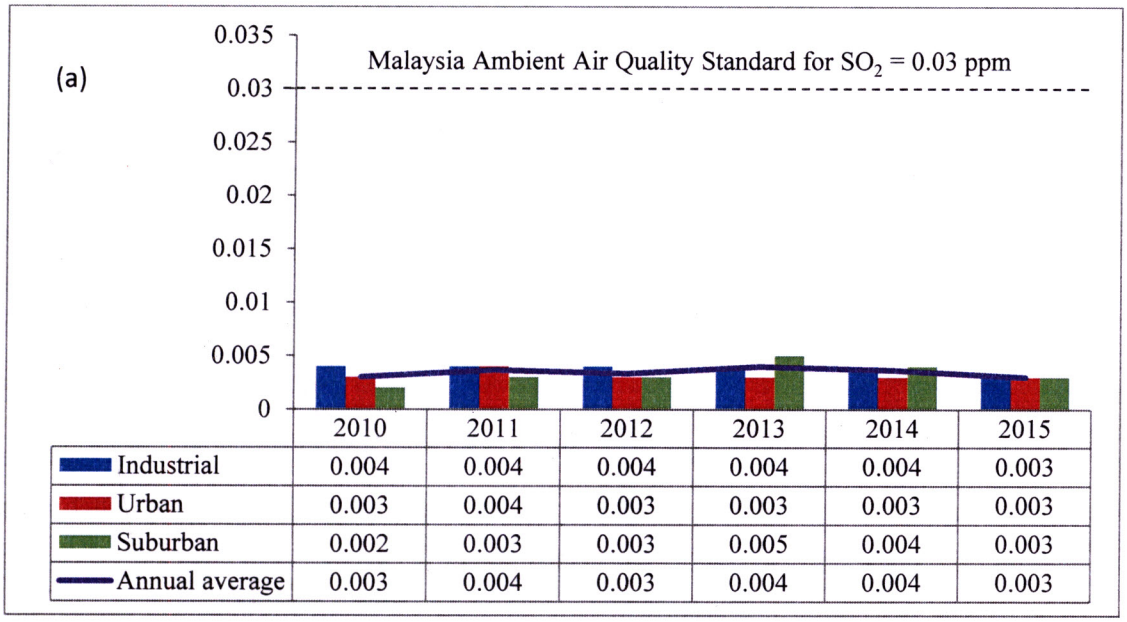
area that was categorised as LPR, contributed to the high annual average of $NO_2$ concentration was probably due to the location of the station which was near to the Sibu domestic airport. The presence high annual average concentration in Sibu station was perhaps due to the $NO_x$ aircraft emissions from the Sibu domestic airport (Mutalib et al., 2013).

#### 4.5.2.3 Annual variation of $SO_2$ from 2010 to 2015 for HPR, MPR and LPR by areas

Figure 4.12 (a) to Figure 4.12 (c) show the annual average $SO_2$ concentration by year from 2010 to 2015 for HPR MPR and LPR in Malaysia by area, respectively. For the annual average $SO_2$ concentration in HPR, it was visualised a slight increase in the annual average of $SO_2$ concentration from 2010 (0.003 ppm) to 2011 (0.004 ppm). However, it decreased in 2012 (0.003 ppm) and continued to increase a year after that until 2014 (0.004 ppm). In 2015, the annual average of $SO_2$ concentration decreased (0.003 ppm). While for MPR, the annual average of$SO_2$ concentration remained constant from 2010 to 2014 with the annual average concentration of 0.0020 ppm. However, it decreased in 2015 with an annual average concentration of 0.0017 ppm. For LPR, the annual average $SO_2$ concentration slightly decreased from 2010 to 2014 and increased in 2015. The overall trend on the annual average daily 24-hour in ambient air from 2010 to 2015 by area was well below the limit of 0.03 ppm.

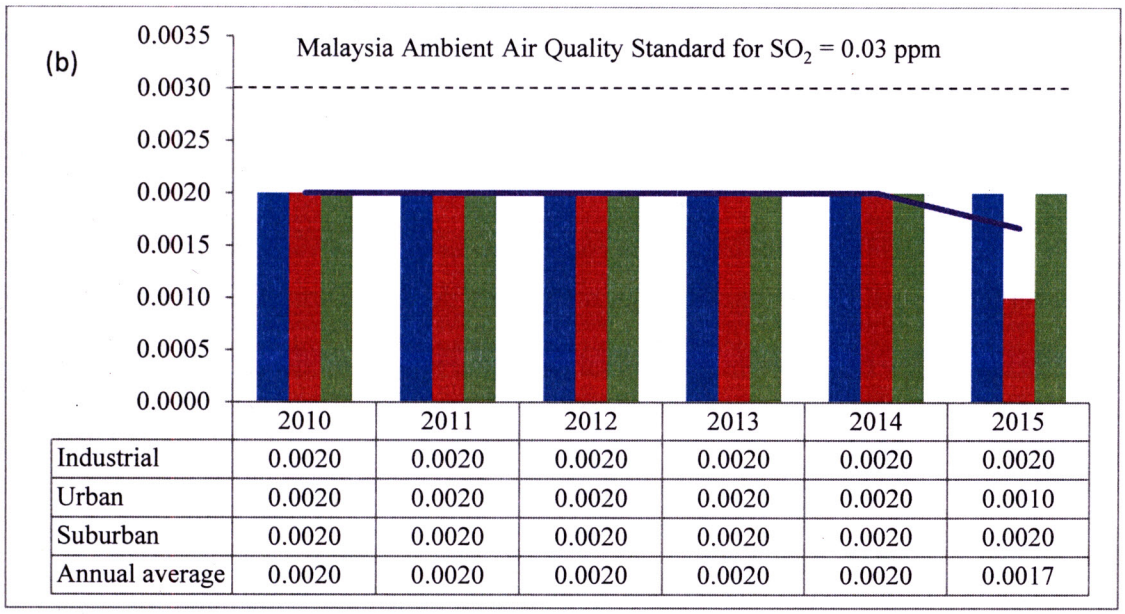
In general, HPR recorded the highest annual average $SO_2$ concentration as compared to MPR and LPR with none of them being above the limit stipulated by MAQS (0.03 ppm). According to Masiol et al. (2017) and Mutalib et al. (2013), the presence of $SO_2$ in the atmosphere was usually the result of industrial activities. The Pasir Gudang air

quality monitoring station located in HPR experienced a great amount of air pollution due to its locality in the rapid growth of industrial areas. Due to its geographical location in industrial areas, it was recorded as the highest polluted area in Johor (Lee et al., 2012).



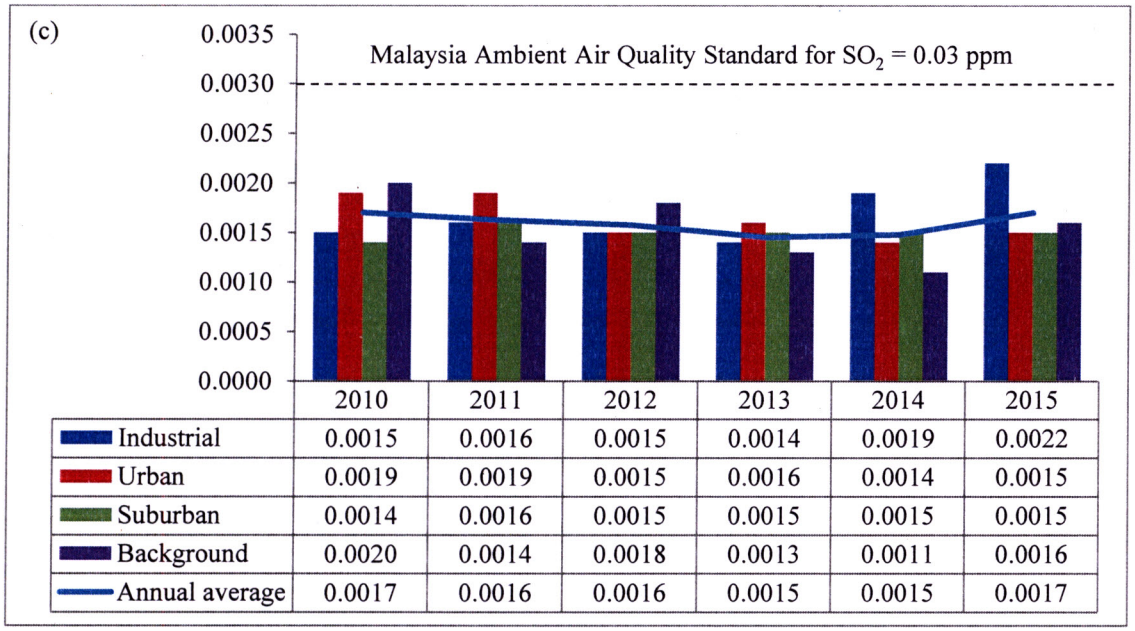

| | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
|---|---|---|---|---|---|---|
| Industrial | 0.004 | 0.004 | 0.004 | 0.004 | 0.004 | 0.003 |
| Urban | 0.003 | 0.004 | 0.003 | 0.003 | 0.003 | 0.003 |
| Suburban | 0.002 | 0.003 | 0.003 | 0.005 | 0.004 | 0.003 |
| Annual average | 0.003 | 0.004 | 0.003 | 0.004 | 0.004 | 0.003 |

Figure 4.12. Annual average concentration of $SO_2$ by year (2010-2015) for (a) HPR (b) MPR and (c) LPR in Malaysia




| | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
|---|---|---|---|---|---|---|
| Industrial | 0.0020 | 0.0020 | 0.0020 | 0.0020 | 0.0020 | 0.0020 |
| Urban | 0.0020 | 0.0020 | 0.0020 | 0.0020 | 0.0020 | 0.0010 |
| Suburban | 0.0020 | 0.0020 | 0.0020 | 0.0020 | 0.0020 | 0.0020 |
| Annual average | 0.0020 | 0.0020 | 0.0020 | 0.0020 | 0.0020 | 0.0017 |

Figure 4.12. Continued




| | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
|---|---|---|---|---|---|---|
| Industrial | 0.0015 | 0.0016 | 0.0015 | 0.0014 | 0.0019 | 0.0022 |
| Urban | 0.0019 | 0.0019 | 0.0015 | 0.0016 | 0.0014 | 0.0015 |
| Suburban | 0.0014 | 0.0016 | 0.0015 | 0.0015 | 0.0015 | 0.0015 |
| Background | 0.0020 | 0.0014 | 0.0018 | 0.0013 | 0.0011 | 0.0016 |
| Annual average | 0.0017 | 0.0016 | 0.0016 | 0.0015 | 0.0015 | 0.0017 |

Figure 4.12. Continued

This study correlated with a study done by Rahman et al. (2016), which found that the high concentration of $SO_2$ in Pasir Gudang station as compared to Muar was due to the location of Pasir Gudang station at industrial and residential besides congested roads. This proves that pollutants concentration was influenced by the local surroundings. According to Retnam et al. (2013), Pasir Gudang is an important industrial zone in Malaysia, whereby there was an industrial area with transportation and logistics, shipbuilding, petrochemicals and other heavy industries, as well as palm oil storage and distribution (Amin et al., 2015). Meanwhile, SK Cenderawasih, Prai air quality monitoring station perhaps one of the contributors to the high annual average concentration of $SO_2$ present in MPR. According to Ismail et al. (2017), Prai areas are the most affected by the industrial activities with 1,066 out of 4,838 industries have the potential to create air pollution in Penang.

In Prai, there are wood industries, electronics, metals, chemical and rubber industries that contribute to the major pollution in Prai with anpercentage of 2.75%, 27.52 %, 39.45 %, 24.77% and 5.50%, respectively. Besides, air quality monitoring at Kerteh

categorized in LPR perhaps one of the contributors to the high annual average concentration of pollutants present in LPR since it is located in the industrial area and known for it petroleum discovery site. There is Kerteh Petrochemical Industrial area which is the major industrial sites in Terengganu. Besides, Gebeng Industrial area which located near Kerteh perhaps one of the contributor to the high annual average $SO_2$concentration present in Kerteh.

The Klang station, which was categorised in HPR, is located in the urban area. Present of the high annual average concentration of $SO_2$ in Klang station probably due to the presence of Port Klang. Port Klang is the largest port in the country. Besides, one of the oldest gas turbine and coal-fired power stations activated in Malaysia was also occupied in the area. According to Mohamad et al. (2015), coal combustion activities emit various air pollutants, namely $NO_x$, $SO_2$, $PM_{10}$, $CO_2$ and other substances. Therefore, the activities contributed to the high annual average concentration of $SO_2$ present in the Klang station. Besides, this district is developing in terms of urbanisation, industrialisation and infrastructure that trigger the air quality in the atmosphere. Usually, volcanoes, oceans, biological decay and forest fire contribute to the presence of $SO_2$. However, due to the absence of volcanic actions in Malaysia, $SO_2$ is substantially emitted from the combustion of coal and petroleum from vehicles and power stations.

Besides, emissions from the diesel engines of motor vehicles in the urban area also contribute to the high annual average $SO_2$ concentration (Mutalib et al., 2013). According to Afzali et al. (2017), they found out that the contribution of $SO_2$ concentrations on-road vehicle emission was more than the industrial emission sources. This showed that besides the industrial areas, emission from the

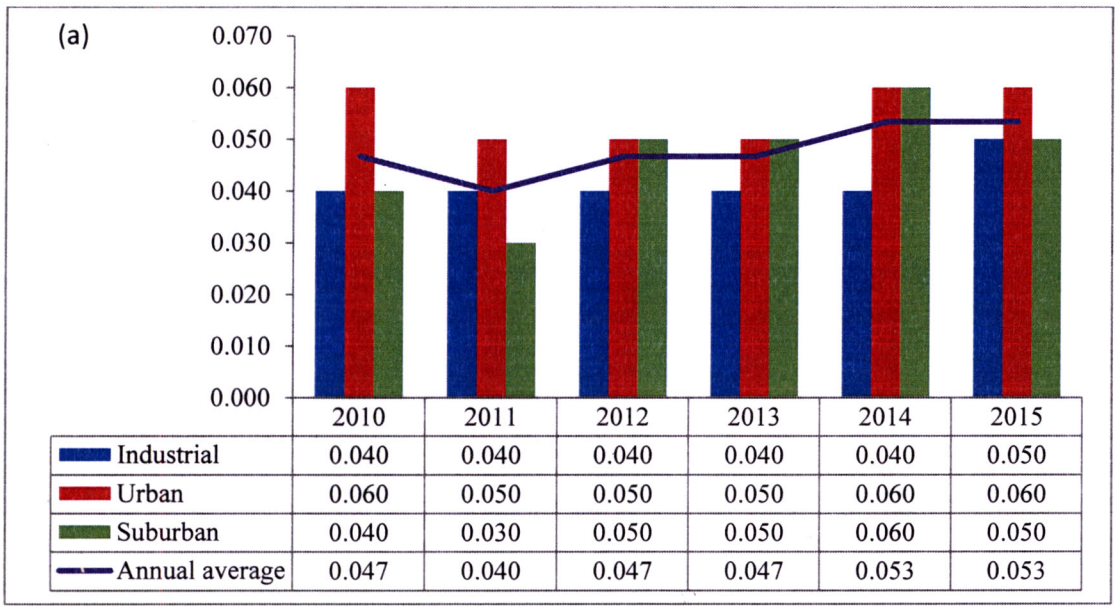
transportation also contributes to the presence of $SO_2$ in the atmosphere. Muar station, which is located at the suburban area is surrounded by a small number of industries. This caused the Muar station, which was categorised as HPR to experience a high annual average $SO_2$ concentration. Emissions of $SO_2$ are primarily related to the combustion of sulphur-containing fuels. It was noticed that the annual average concentration of $SO_2$slightly decreased from 2014 to 2015 for HPR and MPR. This was attributed by the use of better fuel quality petrol EURO-4M RON 97 in Malaysia starting from 2015, as well as the implementation of EURO- 5 Diesel with the sulphur content of less than 10 mg/L. Even though the number of vehicles which used both types of fuels is low, they still have an impact on the $SO_2$ trend (DOE, $2015_a$).

#### 4.5.2.4 Annual variation of $O_3$ from 2010 to 2015 for HPR, MPR and LPR by areas

Figure 4.13 (a) to Figure 4.13 (c) show the annual average $O_3$ concentration from 2010 to 2015 for HPR MPR and LPR in Malaysia by area, respectively. For the annual average concentration of $O_3$ in HPR, it visualised a decrease annual average concentration of $O_3$ in 2011. However, the annual average concentration of $O_3$ increased after that until 2015.

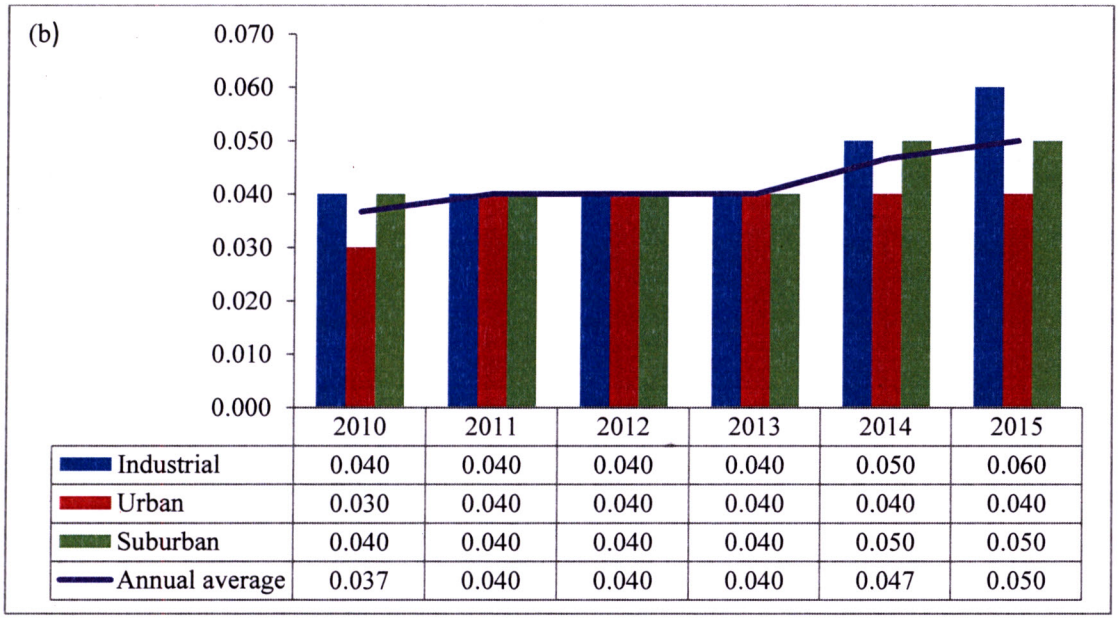
While for MPR, the annual average concentration of $O_3$ was found to increase from 2010 to 2015, with an annual average concentration of $O_3$ that remained constant from 2011 to 2013, with an annual average concentration of $O_3$ at 0.040 ppm. For LPR, the annual average concentration of $O_3$ increased from 2010 to 2012 with a constant annual average concentration of at 0.029 ppm in 2011 and 2012. In 2013, the annual average concentration of $O_3$ decreased and increased in 2014 and 2015, respectively.

In general, HPR recorded the highest annual average concentration of $O_3$ as compared to MPR and LPR. The present high concentration of $O_3$ in industrial area at HPR was probably due to the locality of air quality monitoring stations at the industrial areas. For instance, the Kemaman air quality monitoring station that was categorised as HPR probably contributed to the high annual average concentration of $O_3$ present in that area. According to AhmadIsiyaka and Azid (2015), the Kemaman station houses large activities of industrial and commercial.



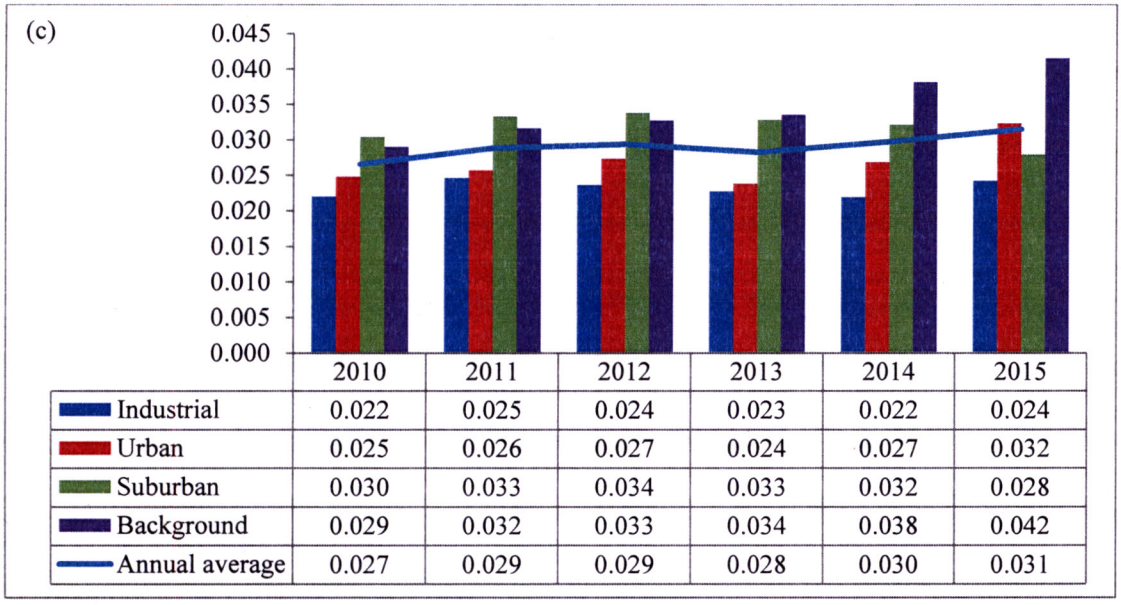

| | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
|---|---|---|---|---|---|---|
| Industrial | 0.040 | 0.040 | 0.040 | 0.040 | 0.040 | 0.050 |
| Urban | 0.060 | 0.050 | 0.050 | 0.050 | 0.060 | 0.060 |
| Suburban | 0.040 | 0.030 | 0.050 | 0.050 | 0.060 | 0.050 |
| Annual average | 0.047 | 0.040 | 0.047 | 0.047 | 0.053 | 0.053 |

Figure 4.13. Annual average concentration of $O_3$ by year (2010-2015) for (a) HPR (b) MPR and (c) LPR in Malaysia




| | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
|---|---|---|---|---|---|---|
| Industrial | 0.040 | 0.040 | 0.040 | 0.040 | 0.050 | 0.060 |
| Urban | 0.030 | 0.040 | 0.040 | 0.040 | 0.040 | 0.040 |
| Suburban | 0.040 | 0.040 | 0.040 | 0.040 | 0.050 | 0.050 |
| Annual average | 0.037 | 0.040 | 0.040 | 0.040 | 0.047 | 0.050 |

Figure 4.13. Continued




| | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
|---|---|---|---|---|---|---|
| Industrial | 0.022 | 0.025 | 0.024 | 0.023 | 0.022 | 0.024 |
| Urban | 0.025 | 0.026 | 0.027 | 0.024 | 0.027 | 0.032 |
| Suburban | 0.030 | 0.033 | 0.034 | 0.033 | 0.032 | 0.028 |
| Background | 0.029 | 0.032 | 0.033 | 0.034 | 0.038 | 0.042 |
| Annual average | 0.027 | 0.029 | 0.029 | 0.028 | 0.030 | 0.031 |

Figure 4.13. Continued

Besides, petrochemical and steel plants arelocated along the shoreline (Awang et al., $2015_b$). According to Finardi et al. (2018), concentration of $O_3$ also increased due to the sea breeze since the cycle of sea breeze can cause recirculation of pollutants for prolonged periods, favouring the formation of $O_3$ because both local and regional sources. Since the Kemaman station is located near the coastal region, and thus it experienced a high annual average concentration of $O_3$ (Mutalib et al., 2013). Besides,

Tasek, Ipoh air quality monitoring station that was categorised as MPR probably contributed to the high annual average concentration of $O_3$ present in that area. There is the biggest industry, mostly focusing on metal wood, plastic and production of paper with 17.86% and 82. 14% of rubber mill and oil palm mills near in Perak, respectively (Ismail et al., 2017). On top of that, Tasek, Ipoh station, which is also located within the area of the airport, will indirectly increase the annual average concentration of $O_3$ present due to the presence of $NO_x$ from the aircraft emissions (Mutalib et al., 2013), which was one of the $O_3$ precursors.

Besides, the concentration of $O_3$ especially in urban and suburban areas was perhaps contributed by the $NO_x$ precursors from the industrial activities (Azid et al., $2015_a$). High annual average concentration of $O_3$ in stations located in urban areas, such as Cheras, Kota Tinggi, Klang, SM Tinggi Melaka, Port Dickson, Seremban, Putrajaya, Batu Muda and Shah Alam and were categorised as HPR,was probably due to the $O_3$ precursors, namely, CO, $NO_x$ and VOC which predominantly came from motor vehicles and sources of natural emission (Che et al., 2011). In general, temperature tends to cause the high $O_3$ concentration besides other meteorological parameters, such as solar radiation, wind speed and pressure (Mittal et al., 2007). In the warm sunny urban atmosphere, $NO_x$ goes through some chemical reactions with VOC and forms $O_3$ (Mutalib et al., 2013), whereby these pollutants ($NO_x$ and VOC) usually come from industries and exhaust of motor vehicles that are mainly in the urban areas.

Higher traffic capacity and conducive condition of atmospheric resulting the formation of $O_3$ in urban areas (DOE, $2015_a$). It was evident from the study done by Kasparoglu et al. (2018), who found out that 57% and 60% of $O_3$ concentration which exceeded the rural and urban stations, respectively, showed that urban contribute more $O_3$ than

rural area. Besides, the location of Cheras station, which is located at SM Kebangsaan Seri Permasuri, also experienced a high annual average concentration of $O_3$ since it is located close to the major road, which is the Maju Expressway (Mohamad et al., 2015). Therefore, $O_3$ precursor emanated from motor vehicles contributed to the presence of $O_3$ in that area.

In suburban areas, the Bintulu air quality monitoring station that was categorised as LPR contributed to the high annual average $O_3$ concentration owing to the effect of downwind, bringing precursors of $O_3$ viz. $NO_x$ and VOC produced from motor vehicles and industries. In addition, a study done by Zong et al. (2018) showed that downwind plume of the $O_3$ precursor from urban areas was one of the contributors to the high annual average concentration of $O_3$ present in the suburban areas. Besides being produced by the reaction of VOC and $NO_x$ that emanate from motor vehicles and industries, the Bintulu station experienced the highest annual average concentration of $O_3$ in Sarawak region was probably due to the outsized forest and agricultural areas in Borneo. According to Duncan et al. (2003), outsized forest and agricultural areas also contributed to the VOC, which was one of the $O_3$ precursors in the atmosphere. This was also one of the reasons for background area had experienced the highest annual average concentration of $O_3$.

#### 4.5.2.5 Annual variation of $PM_{10}$ from 2010 to 2015 for HPR, MPR and LPR by areas

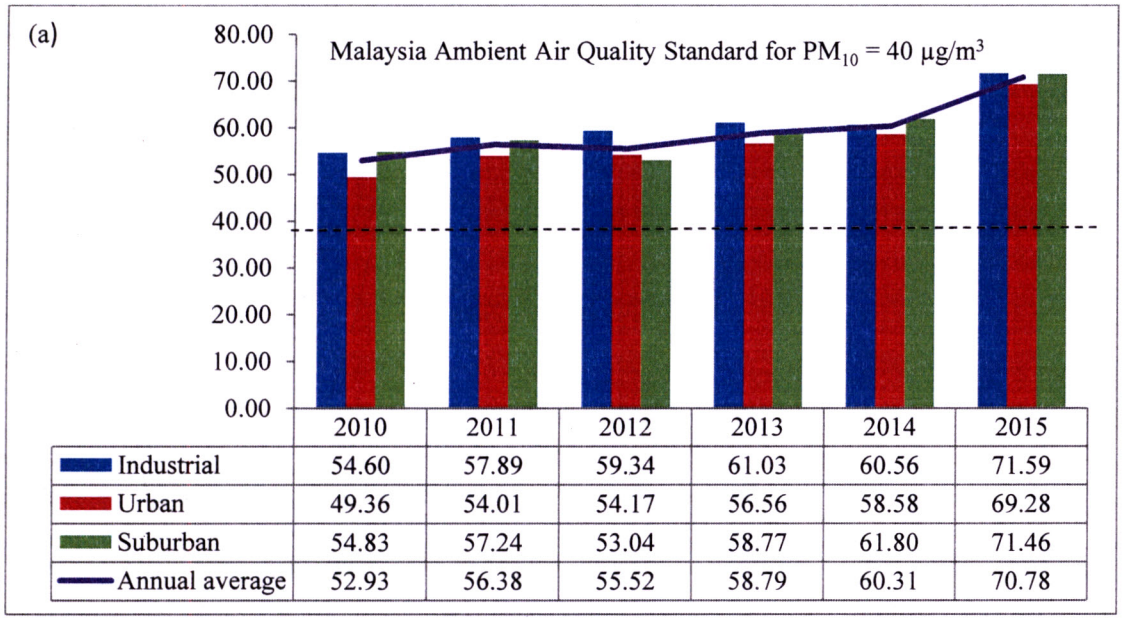
Figure 4.14 (a) to Figure 4.14 (c) show annual average concentration of$PM_{10}$ by year from 2010 to 2015 for HPR, MPR, and LPR in Malaysia by area, respectively. For the annual average concentration of $PM_{10}$ in HPR, it visualised an increase in the annual average concentration of $PM_{10}$ in 2011, decreased in 2012 and continuously increased

from 2013 to 2015. Meanwhile for MPR, the annual average concentration of $PM_{10}$ was found to increase in 2011, slightly decreased in 2012 and 2013, slightly increased in 2014 and continuously increased in 2015. For LPR, the annual average concentration of $PM_{10}$ increased in 2011. Then it decreased until 2013, and continuously increased in 2014 and 2015. The overall trend on the annual average $PM_{10}$ concentration in ambient air for all the areas from 2010 to 2015 exceeded the limit of MAQS, which was 40 $\mu g/m^3$ for HPR and MPR, while some of them exceeded the LPR limit.

In general, HPR recorded the highest annual average $PM_{10}$ concentration as compared to MPR and LPR. Geographically, Bukit Rambai station, which is one of the air quality monitoring stations categorised as HPR, is located in the rapid growth of the industrial areas, causing a great amount of air pollution (Yap & Hashim, 2013). Furthermore, the Bukit Rambai station that is located in the southern part of Peninsular Malaysia was predisposed to the transboundary due to the smoke produced from the Sumatera region via forest fire (Ahmat et al., 2015).

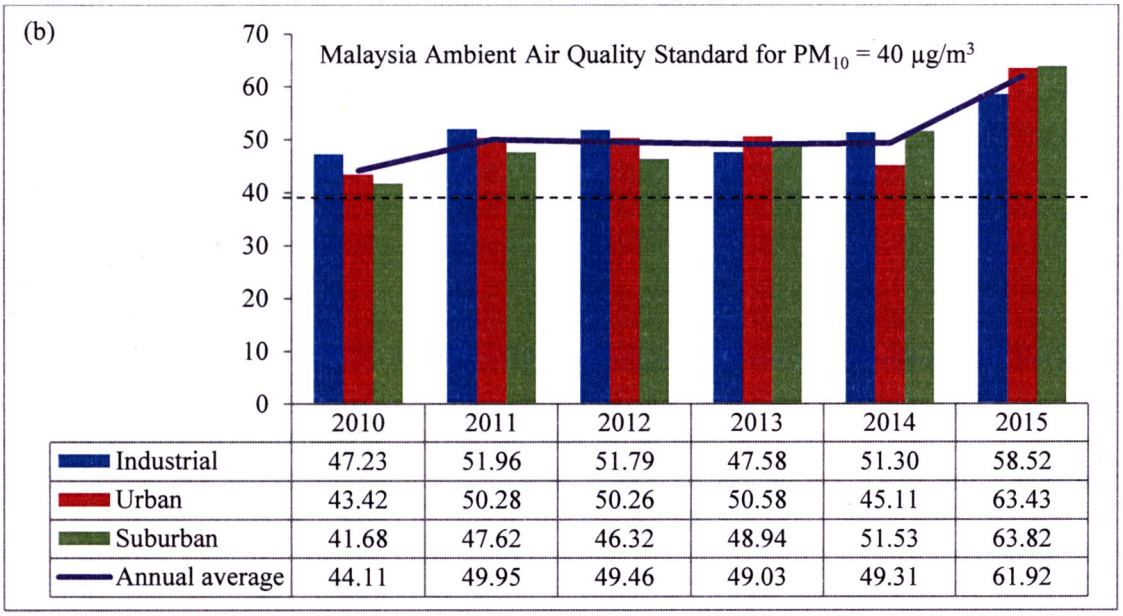
The Sg Petani station located at the suburban area, which was categorised as MPR, contributed the highest annual average concentration of $PM_{10}$ probably due to the presence of the Bakar Arang industrial area and Tikam Batu industrial area located within the Sg Petani station (Ismail et al., 2017). The presence of these industrial activities contributed to the high annual average concentration of $PM_{10}$ present in the Sg Petani station. Meanwhile, Labuan station, which is located at suburban areas categorised as LPR and facing the South China Sea, caused Labuan station to contribute high annual average concentration of $PM_{10}$. According to the European Environment Agency (EEA, 2012), the presence of $PM_{10}$ is probably due to the sea

spray emanate from the ocean, which is formed through the combination of inorganic sea salt and organic matter.. It consisted of sodium chloride (NaCl), magnesium (Mg) and sulphate ($SO_4^{2-}$) and was formed from the droplet through the bubble bursting mechanism.



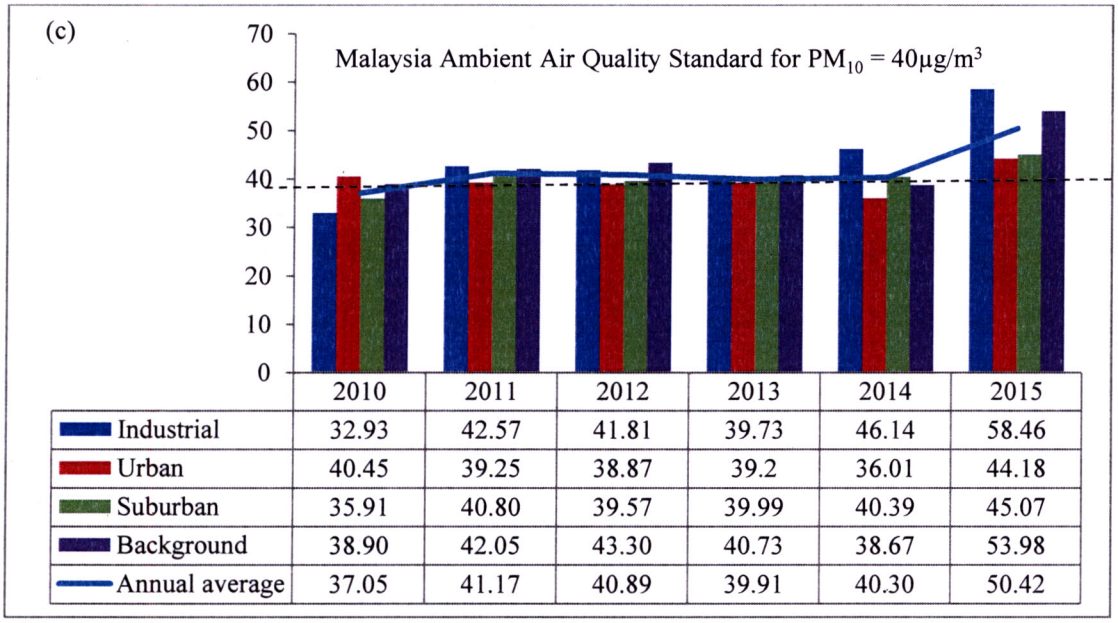

| | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
|---|---|---|---|---|---|---|
| Industrial | 54.60 | 57.89 | 59.34 | 61.03 | 60.56 | 71.59 |
| Urban | 49.36 | 54.01 | 54.17 | 56.56 | 58.58 | 69.28 |
| Suburban | 54.83 | 57.24 | 53.04 | 58.77 | 61.80 | 71.46 |
| Annual average | 52.93 | 56.38 | 55.52 | 58.79 | 60.31 | 70.78 |

Figure 4.14. Annual average concentration of $PM_{10}$ by year (2010-2015) for (a) HPR (b) MPR and (c) LPR in Malaysia




| | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
|---|---|---|---|---|---|---|
| Industrial | 47.23 | 51.96 | 51.79 | 47.58 | 51.30 | 58.52 |
| Urban | 43.42 | 50.28 | 50.26 | 50.58 | 45.11 | 63.43 |
| Suburban | 41.68 | 47.62 | 46.32 | 48.94 | 51.53 | 63.82 |
| Annual average | 44.11 | 49.95 | 49.46 | 49.03 | 49.31 | 61.92 |

Figure 4.14. Continued




| | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
|---|---|---|---|---|---|---|
| Industrial | 32.93 | 42.57 | 41.81 | 39.73 | 46.14 | 58.46 |
| Urban | 40.45 | 39.25 | 38.87 | 39.2 | 36.01 | 44.18 |
| Suburban | 35.91 | 40.80 | 39.57 | 39.99 | 40.39 | 45.07 |
| Background | 38.90 | 42.05 | 43.30 | 40.73 | 38.67 | 53.98 |
| Annual average | 37.05 | 41.17 | 40.89 | 39.91 | 40.30 | 50.42 |

Figure 4.14. Continued

Therefore, the presence of sea spray which originated from the ocean contributed to the presence of $PM_{10}$. This is coherent with a study done by Almeida et al. (2005), whereby they found out that at coastal areas, sea spray is the major source of coarse mode aerosol. Therefore, the anthropogenic (industries) and environmental (sea spray) factors contributed to the presence of high $PM_{10}$ concentration. The highest annual average concentration of $PM_{10}$ was noticed in the background area. Although located in the background area, which normally experienced a low concentration of pollutants, it has the potential to experience the highest annual average concentration of $PM_{10}$ due to the $PM_{10}$ transport through wind direction (Latif et al., 2014). Moreover, the background area is widely known as an area that depends mostly on agriculture. Therefore, $PM_{10}$ presence in the atmosphere might be due to the use of pesticides in agriculture practices (Abdullah et al., 2012).

As shown in Figure 4.10 to Figure 4.14, the annual average concentration of $PM_{10}$ showed the highest concentration for each of the regions as compared to other pollutants. This showed that $PM_{10}$ was the major pollutant in 2010 to 2015 and its

presence might give negative impact to human health as well as the environment. It was also shown that there was inequitable exposure of pollutants distribution for pollutants, such as CO, $NO_2$, $SO_2$, $O_3$ and $PM_{10}$ since they gave different distributions pattern owing to the increase and decrease in the annual average concentration of pollutant. The line chart of the annual average $PM_{10}$ concentration by year as shown in figure visualised that each region experienced different patterns of pollutant concentration, whereby the pollutant was inequitably exposed through the yearly pattern. In contrast, some of the stations showed the same pattern by year. However, according to Ismail et al. (2017), the same pattern of pollutant concentration did not mean pollutants remain in the region since most of the pollutant concentration patterns rely upon source availability, either internal or external.

According to Rosofsky et al. (2018), inequality of pollutants exposure might be due to the shifts of spatiotemporal in air pollution. This situation is supported with the different sources of pollution in terms of spatial that affect the concentration of pollutants present. This variety of station locations will give different air pollutant data readings for each region (Amran et al., 2015). Therefore, various sources of pollutants obtained from the urban, suburban, industrial and background areas promote inequitable exposure of pollutants distribution.

The highest annual average concentration of CO, $NO_2$, $SO_2$, $O_3$ and $PM_{10}$ experienced by the stations located in the urban, suburban, industrial as well as rural areas supported the inequitable exposure of pollutants distribution due to the spatiotemporal shifts in the air pollution since the background areas also experienced a high concentration of pollutants, which is usually experienced by the urban, suburban and industrial areas.

In terms of spatial, generally, the urban, suburban, industrial and background areas promote the presence of air pollutants into the atmosphere. Comparing the whole atmosphere, the urban ambient air is more polluted (Leh et al., 2012). This is due to the higher rate of pollutants produced in the urban areas as compared to the less developed areas owing to the high human population density and the activities done. According to the Department of Statistics (DOS, 2016), population increased from 2010 to 2015 with statistics of population at 28588, 29062, 29510, 30213, 30708, 31186 in the year of 2010, 2011, 2012, 2013, 2014 and 2015, respectively. These high population densities indirectly contribute to the high pollutants rate in the atmosphere.

Besides, burning activities in the urban area also contribute to the presence of the pollutants (Leh et al., 2012). In urban areas, motor vehicles are the main air pollution sources, which consist of dust fall-out, suspended particular matter and lead (Awang et al., 2000). This is supported by Ghazali et al. (2009) who mentioned that the major sources of pollutants, namely, $PM_{10}$, $NO_2$ and CO are produced from the combustion of fuel in automobiles._Beside motor vehicles, industrial waste incinerators, power plant and dust emission from construction works and quarries at urban areas also contribute to air pollution (Rahman et al., 2015). A high level of pollutants in each region is due to the rapid urbanisation process. This is supported by the study done by Azmi et al. (2010) who mentioned that severe air quality occurred in highly urbanised areas. However, in this study, industrial also showed a great annual average concentration of pollutants besides the urban area itself, especially for HPR.

## 4.6 Correlation between air pollutants and meteorological factors

The association strength between two variables can be measured by using correlation analysis. In this study, the correlations between at least two continuous parameters

were measured by using the Pearson correlation ($r$). The general formula of the $r$ is shown below:

$$r = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\left(\sqrt{\sum X^2 - \frac{(\sum X)^2}{N}}\right)\left(\sqrt{\sum Y^2 - \frac{(\sum XY)^2}{N}}\right)}$$

Where,

N = sample size

X = independent parameter value

Y = dependent parameter value

The sign of coefficient loading indicates a direct or inverse relation between the coefficient loading and the parameters. While the correlation strength between the parameters is designated by the numeric value (Yuce et al., 2014).

Table 4.7 presents the relation between meteorological and pollutant parameters in Malaysia for HPR, MPR and LPR. For HPR, the pollutants of $NO_x$, NO, $NO_2$ and CO were strongly correlated with each other with the correlation value between the pollutants as shown in the parenthesis; $NO_x$ and NO (0.9555), $NO_x$ and $NO_2$ (0.7916), $NO_x$ and CO (0.6757), NO and $NO_2$ (0.6550), NO and CO (0.6228), $NO_2$ and CO (0.61002). Meanwhile for MPR, the pollutants of $NO_x$, NO, $NO_2$ and CO were strongly correlated with each other with the correlation value between the pollutants as shown in the parenthesis; $NO_x$ and NO (0.9260), $NO_x$ and $NO_2$ (0.6227), $NO_x$ and CO (0.5231). For LPR, the pollutants of $NO_x$, NO, $NO_2$ and CO werestrongly correlated with each other with the correlation value between the pollutants as shown in the

parenthesis; $NO_x$ and NO (0.9265), $NO_x$ and $NO_2$ (0.7410), $NO_x$ and CO (0.5530), NO and $NO_2$ (0.5720), $NO_2$ and CO (0.5198).

The relation between the pollutants ($NO_x$, NO, $NO_2$) was explained in a study done by Famoso et al. (2015), whereby they mentioned that in urban areas, combustion processes from engine cars contributed to the formation of $NO_x$. Meanwhile, NO is formed when there is a relation between gaseous nitrogen ($N_2$) existing in the air with the atmospheric oxygen during the processes of combustion at temperatures greater than 1200 °C. Meanwhile, during the process of soot cooling, NO interacts with the $O_2$ and transforms partially into $NO_2$ and a mix of two oxides called $NO_x$. In the meantime, CO gives the highest correlation with $PM_{10}$ with the correlation values between CO and $PM_{10}$ were 0.4869, 0.5142 and 0.4828 for HPR, MPR and LPR, respectively. According to Dominick et al. (2012), CO is the main pollutant that impelled the high concentration of $PM_{10}$ due to the combustion process initiating from motor vehicles.

Meanwhile, poor correlation was observed between the pollutant and meteorological parameters, including wind speed, wind direction, temperature and humidity for each region. However, they remained important and indirectly affected the concentration of pollutant present. According to Akhtar et al. (2018), wind speed is the main meteorological parameter in diluting pollutants. Strong winds scatter the pollutants while light winds contribute stagnant pollutants that cause accumulation of pollutants in certain area. Meanwhile, the increase in temperature would contribute to the chemical reactions that cause particulate matter to be finely divided (Afzali et al., 2014). Therefore, indirectly increased the concentration of $PM_{10}$ with increased temperature.

Table 4.7. Correlation between air pollutants and meteorological factors for each region

| Stations | Parameters | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Ws | Wd | Temp | Humidity | $NO_x$ | NO | $SO_2$ | $NO_2$ | $O_3$ | CO | $PM_{10}$ |
| HPR | Wd | -0.2636 | 1 | | | | | | | | | |
| | Temp | -0.0165 | 0.0301 | 1 | | | | | | | | |
| | Humidity | -0.2654 | 0.1625 | -0.0855 | 1 | | | | | | | |
| | $NO_x$ | -0.1813 | 0.1210 | 0.2012 | -0.0801 | 1 | | | | | | |
| | NO | -0.1869 | 0.1210 | 0.1857 | -0.0183 | **0.9555** | 1 | | | | | |
| | $SO_2$ | -0.0075 | 0.0498 | 0.0784 | -0.0913 | 0.3372 | 0.3165 | 1 | | | | |
| | $NO_2$ | -0.1715 | 0.0986 | 0.2235 | -0.1249 | **0.7916** | **0.6550** | 0.3052 | 1 | | | |
| | $O_3$ | -0.0780 | 0.0541 | 0.3365 | 0.0451 | 0.1470 | 0.0859 | 0.0838 | 0.3233 | 1 | | |
| | CO | -0.1402 | 0.0918 | 0.1717 | -0.1223 | **0.6757** | **0.6228** | 0.2591 | **0.6102** | 0.1918 | 1 | |
| | $PM_{10}$ | 0.0334 | -0.0181 | 0.0591 | -0.1363 | 0.1183 | 0.0740 | 0.1872 | 0.1815 | 0.2173 | 0.4869 | 1 |
| MPR | Wd | -0.2863 | 1 | | | | | | | | | |
| | Temp | 0.0707 | 0.0149 | 1 | | | | | | | | |
| | Humidity | -0.1227 | 0.0723 | -0.1832 | 1 | | | | | | | |
| | $NO_x$ | -0.2000 | 0.1563 | 0.0638 | -0.1046 | 1 | | | | | | |
| | NO | -0.1941 | 0.0992 | 0.0649 | -0.0275 | **0.9260** | 1 | | | | | |
| | $SO_2$ | -0.0369 | 0.0741 | 0.0046 | -0.0927 | 0.2708 | 0.2249 | 1 | | | | |
| | $NO_2$ | -0.1320 | 0.2414 | 0.0879 | -0.2252 | **0.6227** | 0.4063 | 0.2871 | 1 | | | |
| | $O_3$ | 0.0268 | 0.1077 | 0.2713 | -0.1944 | 0.1507 | 0.0241 | 0.1685 | 0.4323 | 1 | | |
| | CO | -0.1554 | 0.1244 | 0.0128 | -0.0838 | **0.5231** | 0.4474 | 0.2094 | 0.4555 | 0.2097 | 1 | |
| | $PM_{10}$ | -0.0086 | 0.0699 | 0.0675 | -0.0797 | 0.1238 | 0.0403 | 0.1527 | 0.2576 | 0.3430 | **0.5142** | 1 |
| LPR | Wd | -0.1562 | 1 | | | | | | | | | |
| | Temp | 0.0122 | -0.0037 | 1 | | | | | | | | |
| | Humidity | -0.1763 | 0.1155 | -0.1251 | 1 | | | | | | | |
| | $NO_x$ | -0.0508 | 0.0255 | 0.0851 | -0.1273 | 1 | | | | | | |
| | NO | -0.0615 | 0.0298 | 0.1002 | -0.0998 | **0.9265** | 1 | | | | | |
| | $SO_2$ | -0.0113 | 0.0133 | 0.0207 | -0.0207 | 0.2987 | 0.2696 | 1 | | | | |
| | $NO_2$ | -0.0397 | 0.0109 | 0.1041 | -0.1085 | **0.7410** | **0.5720** | 0.2751 | 1 | | | |
| | $O_3$ | -0.0626 | 0.0251 | 0.2139 | -0.0160 | 0.1984 | 0.1392 | 0.1660 | 0.3391 | 1 | | |
| | CO | -0.0423 | -0.0095 | 0.1054 | -0.0923 | **0.5530** | 0.4903 | 0.2096 | **0.5198** | 0.2306 | 1 | |
| | $PM_{10}$ | 0.0308 | -0.0505 | 0.1190 | -0.1008 | 0.1167 | 0.0590 | 0.1410 | 0.1999 | 0.2691 | 0.4828 | 1 |

Meanwhile, the concentration of pollutants in air decreased when rain wash-out the atmospheric aerosols (Afzali et al., 2014; Abderrahim et al., 2016). Franceschi et al. (2018) hold the same opinion by performing negative correlation between $PM_{10}$ and humidity. Meanwhile, according to Wang and Ogawa (2015), pollutants can be affected by one of the important parameters, which is wind direction since different directions of winds transport different pollutant amounts. Inconsistency in direction of wind causes poorly negative and positive correlation between wind direction and the pollutants. Despite poor correlation with $PM_{10}$, possible emission sources can be identified and estimated by referring to the direction of wind whichdominantly flew (Syafei, 2014).

Studies by Jaafar et al. (2018) and Noor et al. (2015) showed that winds most frequently came from the southwesterly wind flow from Sumatra in Indonesia and northeasterly wind during the Southwest and Northeast Monsoons with the southwesterly wind flow contributing high concentration of $PM_{10}$ by bringing the particles from forest fires in several provinces. Based on Table 4.7, besides $PM_{10}$, there is also a weak correlation between wind direction and other parameters. In this study, these parameters were not expected to have such relations, but rather a seasonal fluctuation, since according to Jaafar et al. (2018), the patterns of wind flow changed with various monsoons.

Among the pollutants, $O_3$ gave the highest correlation with temperature as compared to other pollutants for each of the regions. The highest temperatures recorded in HPR, MPR and LPR were 40.10 °C, 43.20 °C and 39.80 °C, respectively. This was agreeable with the finding by Mutalib et al. (2013), who mentioned that temperature tends to frequently give high levels of $O_3$ in the atmosphere in warm sunny

atmosphere. This study is also coherent with Sari et al. (2016) who found that high temperature influenced the concentration of $O_3$ present.

Meanwhile, the positive correlation between $PM_{10}$ and other gaseous pollutants, such as $SO_2$ and $NO_2$ recommends that these gases also cause the formation of $PM_{10}$ as water-soluble ion, such as sulphate ($SO_4^{2-}$) and nitrate ($NO_3^-$), which are produced from these gases are part of the $PM_{10}$ concentration (Jaafar et al., 2018). A similar result was also showed through a study done by Custodio et al. (2018) and Li et al. (2018), whereby they found that $SO_4^{2-}$ and nitrate $NO_3^-$ were dominant water-soluble ions in air pollution samples. Therefore, it can be concluded that other air pollutants also influence the concentration of $PM_{10}$.

Although this study, gave low correlation loadings between the pollutant and meteorological parameters, the atmospheric dynamics and the conditions of meteorological were assume as the essential part in controlling the air pollutants. According to Özdemir and Taner (2014), occasionally, the air quality fluctuates at any place, despite the fact that the emission sources are constant. This is due to the influenced patterns of regional weather that transport and disperse air pollutants in the atmosphere. Therefore, the concentration of pollutants present is indirectly affected by the meteorological parameters and it plays an important part in the dispersion of pollutants.

### 4.7 Monthly variation of $PM_{10}$ from 2010 to 2015 by regions

Figure 4.15 shows the monthly average concentration of $PM_{10}$ from 2010 to 2015 based on the region at continuous air quality monitoring stations in Malaysia. Based on the graph plotted, it is observed that HPR shows the highest monthly concentration of

$PM_{10}$ as compared to MPR and LPR, except for certain months which are in January whereby the monthly concentration of $PM_{10}$ is same with MPR and in February, whereby the monthly concentration of $PM_{10}$ is slightly lower than MPR.

The highest monthly concentration of $PM_{10}$ is recorded in September with monthly concentration of $PM_{10}$at 78.09 $\mu g/m^3$, followed by October, June, July, March, August, May, February, April, January, November and December at 73.14 $\mu g/m^3$, 71.95 $\mu g/m^3$, 63.87 $\mu g/m^3$, 61.62 $\mu g/m^3$, 59.58 $\mu g/m^3$, 53.01 $\mu g/m^3$, 52.28 $\mu g/m^3$, 52.12 $\mu g/m^3$, 48.84 $\mu g/m^3$, 44.93 $\mu g/m^3$ and 44.42 $\mu g/m^3$. respectively. Meanwhile, the lowest monthly $PM_{10}$ concentration was recorded in November at 32.49 $\mu g/m^3$ followed by December, January, May, April, February, March, July, June, October, August and September at 32.82 $\mu g/m^3$, 35.29 $\mu g/m^3$, 37.06 $\mu g/m^3$, 37.29 $\mu g/m^3$, 38.05 $\mu g/m^3$, 40.96 $\mu g/m^3$, 45.15 $\mu g/m^3$, 45.32 $\mu g/m^3$, 45.83 $\mu g/m^3$, 45.85 $\mu g/m^3$ and 55.04 $\mu g/m^3$, respectively.

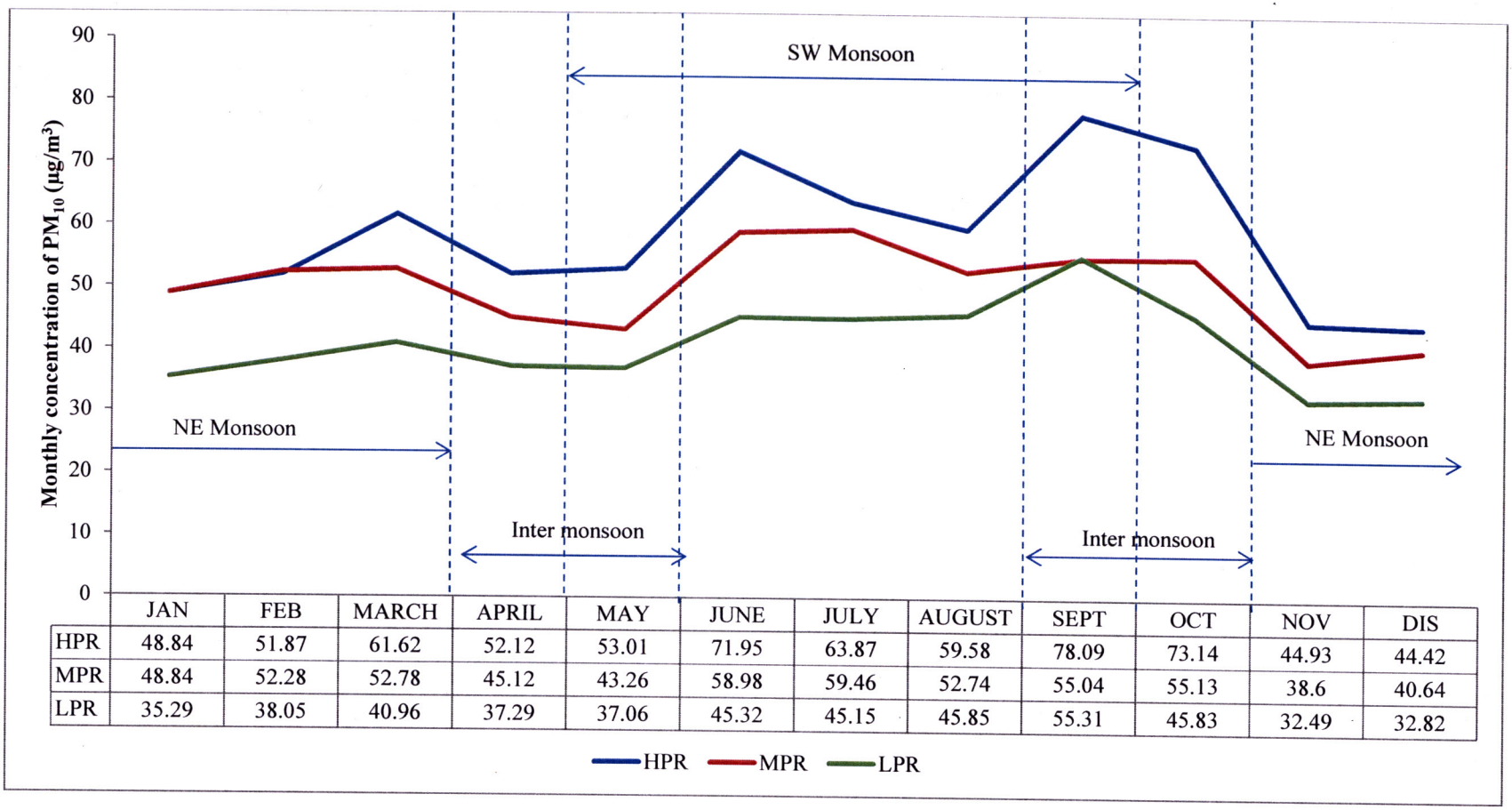
It was noticed that the highest and lowest monthly $PM_{10}$ concentration occurred in September and November, respectively. The Southwest Monsoon usually occurr from May until September while the Northeast Monsoon occur in November until March. Therefore, the highest and lowest monthly concentration of $PM_{10}$ that occurred in September and November was probably due to the monsoon phenomena. In general, the concentration of $PM_{10}$ recorded during the Southwest Monsoon (May until September) was higher than that of the Northeast Monsoon (November until March) (Abdullah et al., $2017_a$).

The higher value of $PM_{10}$ concentration during this period is generally attributable to the drier weather condition, stable atmosphere, local effects and also air pollutants

transboundary transport from the biomass burning from the neighbouring countries (Abdullah et al., 2011). Low rainfall and stable atmospheric condition motivated the high $PM_{10}$ concentration since according to Asif et al. (2018), pollutants are scavenged from the air by rain or snow, or fog in wet deposition. Meanwhile, a study done by Rahman et al. (2015) showed that the $PM_{10}$ levels remained highly concentrated during the Southwest Monsoon (hot and dry season).

The inter-monsoon alsooccur in September, whereby during inter-monsoon, the wind is frequently inconsistent and might coincidently bring the $PM_{10}$ through the transboundary air pollution from Sumatra and Kalimantan, Indonesia. Meanwhile, in October, the inter-monsoon period continue to prevail before starting the Northeast Monsoon in November. During this period, the wind is frequently inconstant and gives strong convective clouds in the late mornings and early afternoons (Fakaruddin et al., 2015).

Perhaps the transboundary of air pollution from the land and forest fires in Sumatra and Kalimantan, Indonesia contributed to the high concentration of $PM_{10}$ presentin Sabah since in 2015 there was massive land and forest fires in Sumatra and Kalimantan, Indonesia. However, the high concentration of $PM_{10}$ in other months, except during the Southwest Monsoon, was probably due to other factors, such as urbanisation process and traffic emissions.




| | JAN | FEB | MARCH | APRIL | MAY | JUNE | JULY | AUGUST | SEPT | OCT | NOV | DIS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| HPR | 48.84 | 51.87 | 61.62 | 52.12 | 53.01 | 71.95 | 63.87 | 59.58 | 78.09 | 73.14 | 44.93 | 44.42 |
| MPR | 48.84 | 52.28 | 52.78 | 45.12 | 43.26 | 58.98 | 59.46 | 52.74 | 55.04 | 55.13 | 38.6 | 40.64 |
| LPR | 35.29 | 38.05 | 40.96 | 37.29 | 37.06 | 45.32 | 45.15 | 45.85 | 55.31 | 45.83 | 32.49 | 32.82 |

Figure 4.15. Monthly average concentration of $PM_{10}$ from 2010 to 2015 based on region

In addition, throughout the monthly $PM_{10}$ concentration trend, some abruptness was observed. According to Gurjar et al. (2016), this abrupt behaviour was probably due to the local meteorological and environmental factors. The abrupt trend of $PM_{10}$ concentration might also be due to the declined air quality with regard to the the transboundary haze pollution from massive land and forest fires in Sumatra and Kalimantan, Indonesia besides the monsoon phenomena. The unevenness of $PM_{10}$ concentration was probably due to the stability and trapping of particles. This is supported by a study done by Soleiman et al. (2003), who found that stability and trapping particles were the main factors which affect the pollution during haze episodes. On top of that, during the summer monsoon (May to September), there was an abrupt peak in the trend of $PM_{10}$ concentration. Perhaps this was due to the stronger transport energy in the atmosphere during the summer monsoon season, whereby during this time the pollutants were lifted, suspended and transported away with a greater amount of energy from the fire emission sources in Sumatra (Show & Chang, 2016).

## 4.8 Development of $PM_{10}$ prediction model

### 4.8.1 Missing data in air quality data

Missingness map is a useful tool for discovering the missingness in a dataset as the data can be summarised quickly and easily. This missingness map visualises the colour grid based on the missingness status, whether the data are missing or observed. Meanwhile the grid columns represent parameters and grid rows represent observations. In any analysis, missing data can represent a severe problem for the network quality, especially for the prediction model. Therefore, this missingness map can point out the ways to improve the missingness by applying proper imputation method through the mechanism of missingness.

Figure 4.16 shows the missingness map for 51 continuous air monitoring stations in Malaysia from 2010 to 2015 for the meteorological (wind speed, wind direction, temperature, humidity, UVB) and pollutant (THC, NMHC, $CH_4$, $SO_2$, $O_3$, CO, NO, $NO_2$, $NO_x$, $PM_{10}$) parameters. It was clear that UVB, THC, NMHC and $CH_4$ variables showed a great number of missing data. In this missingness map, a number of missing data are arranged in decreasing order of missingness from left to right. Therefore, UVB shows a great number of missing data followed by THC, NMHC, $CH_4$, $SO_2$, wind speed, $O_3$, CO, NO, $NO_2$, $NO_x$, temperature, humidity, wind direction and $PM_{10}$.

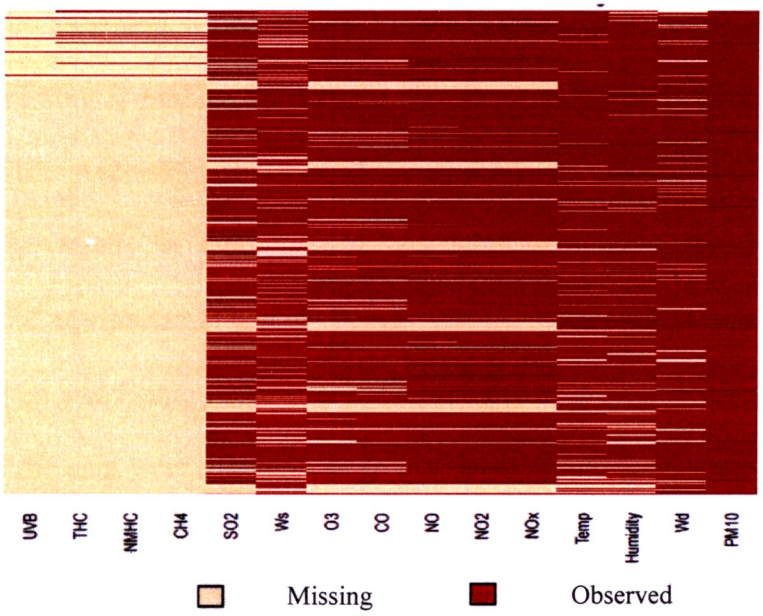



Figure 4.16. Missingness map for 51 continuous air monitoring stations in Malaysia from 2010 to 2015 for meteorological (wind speed, wind direction, temperature, humidity, UVB) and pollutant (THC, NMHC, $CH_4$, $SO_2$, $O_3$, CO, NO, $NO_2$, $NO_x$, $PM_{10}$) parameters

Based on the missingness map (Figure 4.16) UVB, THC, NMHC and $CH_4$ parameters show high proportion of missing data, where the data have more than 90% of missing data. According to Widaman (2006), the missing data will be considered excessive if the data experience more than 50% of missing data. Besides, Munir et al. (2013) mentioned that any parameters with less than 75% of data capture should be removed

from the analysis. This correlated with the study done by Clements et al. (2016), which mentioned that the missing data are removed if there is greater than 25% of missing data present. Meanwhile, according to Ramli et al. (2013), there was no imputation method to implement if there is more than 60% of missing data. Therefore, parameters with a high proportion of missing data, namely UVB, THC, NMHC and $CH_4$ were removed from the analysis since they had more than 60% of missing data. Meanwhile, other parameters with less proportion of missing data still remained and were fully used for the analysis and air quality modelling.

From the missingness map, it was shown that the mechanism of missingness was non-structured or missing at random (MAR) for all the parameters except for UVB, THC, NMHC and $CH_4$. According to Noor et al. (2008), the mechanism for air quality missing data is usually in random. This mechanism of missingness owes to the shutdown of the equipment (Plaia & Bondì, 2006). Gómez-Carracedo et al. (2014) agreed with Plaia and Bondì (2006) by mentioning that there is MAR mechanism of missing data if the data were missing due to the shutdown of system and power supply . Eventually, proper imputation was applied to the remaining parameters for accurate results of air quality modelling based on this mechanism of missingness. In addition, according to Ramli et al. (2013), MAR mechanism of missingness was suitable to be applied by using EMB imputation method.

#### 4.8.1.1 Simulation of missing data

It is vital to assess and study how imputation affects data. To attain this aim, simulation of missing data was done by applying a complete daily $PM_{10}$ concentration obtained from continuous air monitoring stations in Malaysia from 2010 to 2015. Table 4.8 shows the summary of $PM_{10}$ characteristics data. There were 526 missing

data points of $PM_{10}$ from 113,112 total data points. Observed data points were 112,586, and were used for missing data simulation. The lowest and highest concentrations of $PM_{10}$ used for the simulation were 8 $\mu g/m^3$ and 763 $\mu g/m^3$, respectively, with a mean value of 49.91 $\mu g/m^3$. $PM_{10}$ was chosen since it was the most dominant pollutant recorded in many areas in Malaysia (DOE, $2015_a$).

From the complete $PM_{10}$ data set, a simulation study on imputation method viz. nearest neighbour, mean and EMB were conducted based on the various proportions of missing data (5%, 10%, 15%, 25%, 40%). Missing data were simulated to evaluate the performance of the imputation method on the percentage value of missing data. This proportion was based on the missing data categorised as followed; small (5%, 10%), medium (15%, 25%) and large (40%) (Noor et al., 2014).

Table 4.8. Characteristics of $PM_{10}$ data

| | |
|---|---|
| No. of observed data points | 112, 586 |
| No. of missing data points | 526 |
| No. of total data points | 113, 112 |
| Mean | 49.91 |
| Minimum | 8 |
| Maximum | 763 |

Table 4.9 shows the descriptive statistics of simulated missing data with the total number of data points at 112,586 and number of missing data at 5630, 11260, 16888, 28148, 45035 based on the proportions of missing data (5%, 10%, 15%, 25% and 40%).

Table 4.9. Descriptive statistics of simulated missing data

| % of missing data | 5% | 10% | 15% | 25% | 40% |
|---|---|---|---|---|---|
| No. of observed data points | 106956 | 101326 | 95698 | 84438 | 67551 |
| No. of missing data points | 5,630 | 11,260 | 16,888 | 28,148 | 45,035 |
| No. of total data points | 112,586 | 112,586 | 112,586 | 112,586 | 112,586 |
| Mean | 50.09 | 49.31 | 48.26 | 44 | 44 |
| Min | 8 | 8 | 8 | 8 | 10 |
| Max | 763 | 763 | 763 | 763 | 763 |

Table 4.10 shows the performance of three imputation methods (mean, nearest neighbour and EMB) for various proportions of missing values with imputation methods of mean and nearest neighbour is great to be applied onto small percentage of missing data which is 5% as compared to EMB imputation method with $R^2$ and RMSE value obtained for mean, nearest neighbour and EMB imputation methods are 0.9274, 0.9318, 0.9084 and 7.47, 7.47 and 8.58, respectively.

However, it became less efficient when the percentage of missing data was increased since the $R^2$ values obtained from the performance of mean (0.9274, 0.8117, 0.6484, 0.5401, 0.3911) and nearest neighbour (0.9318, 0.8126, 0.6546, 0.5459, 0.3946) imputation method decreased with increase in percentage of missing data (5%, 10%, 15%, 25%, 40%). Besides that, the RMSE values increased for mean (7.47, 12.36, 16.90, 19.13, 22.07) and nearest neighbour (7.47, 12.27, 16.68, 19.13, 21.76) imputation methods. EMB imputation method also decreased the value of $R^2$ (0.9084, 0.8468, 0.7530, 0.5791, 0.5004) and increased values of RMSE (8.58, 11.18, 14.20, 18.53, 20.48) when the proportion of missing data was increased. This showed that when the missing values increase, the estimates of validity decrease (Zakaria & Noor, 2018). Conversely, when there are excessive numbers of missing values in the data, the estimations have a tendency to deviate from the true value (Razak et al., 2014).

Table 4.10 Performance of nearest neighbor, mean and EMB imputation methods for various proportions of missing values

| | Proportion of missing data | Method | $R^2$ | RMSE |
|---|---|---|---|---|
| Small | 5% | Nearest neighbor | 0.9318 | 7.47 |
| | | Mean | 0.9274 | 7.47 |
| | | EMB | 0.9084 | 8.58 |
| | 10% | Nearest neighbor | 0.8126 | 12.27 |
| | | Mean | 0.8117 | 12.36 |
| | | EMB | 0.8468 | 11.18 |
| Medium | 15% | Nearest neighbor | 0.6546 | 16.68 |
| | | Mean | 0.6484 | 16.90 |
| | | EMB | 0.7530 | 14.20 |
| | 25% | Nearest neighbor | 0.5458 | 19.13 |
| | | Mean | 0.5400 | 19.13 |
| | | EMB | 0.5791 | 18.53 |
| Large | 40% | Nearest neighbor | 0.3946 | 21.76 |
| | | Mean | 0.3910 | 22.07 |
| | | EMB | 0.5004 | 20.48 |

Besides, the simulation study showed that nearest neighbor and means imputation method was best applied with a small degree of complexities (5%) and that they become less suitable for the medium (15-25%) and large (40%) degree of complexities. In contrast, the EMB imputation method was less suitable to be applied to the small proportion of missing data (5%) as compared to the mean and nearest neighbour imputation methods. However, it gave a high value of $R^2$ and low value of RMSE for the medium and large proportions of missing data as compared to the mean and nearest neighbour imputation methods.

#### 4.8.1.2 Implementation of imputation methods on data studied

Table 4.11 shows the $R^2$ and RMSE values for the overall missing data in this study, with the proportion of missing data was at16.64% based on the parameters of wind speed, wind direction, temperature, humidity, $O_3$, CO, NO, $NO_2$, $NO_x$, $SO_2$ and $PM_{10}$ and the proportion of missing data of the overall study was categorised in medium degree of complexities.

Table 4.11. $R^2$ and RMSE for overall missing data in the study

| | Before imputation | After imputation | | |
|---|---|---|---|---|
| | | Nearest neighbour | Mean | EMB |
| $R^2$ | 0.3853 | 0.4128 | 0.4078 | **0.5168** |
| RMSE | 22.94 | 21.83 | 21.78 | **19.79** |

After applying the imputation methods of missing data, the three imputation methods, namely, nearest neighbour, mean and EMB showed good results of $R^2$ (0.4128, 0.4078, 0.5168) by comparing with the $R^2$ value before imputation (0.3853), respectively. It is crucial to decide which imputation methods give the best missing values substitution. According to Azid et al. (2016), accuracy can be determined by the lowest RMSE value and highest $R^2$ value with RMSE and $R^2$ values equal to 0 and 1, respectively. Based on the results obtained, the EMB imputation method showed the highest $R^2$ (0.5168) and lowest RMSE (19.79) values when compared with the nearest neighbour (0.4128, 21.83) and mean imputation (0.4078, 21.78) methods, respectively. Thus, the accuracy of the prediction by using EMB imputation was greater than the nearest neighbor and mean imputation method.

Therefore, it is best to have EMB imputation method rather than the nearest neighbour and mean imputation methods in the case of medium proportion of missing data. Besides, according to Zhang (2016), EMB imputation method through Amelia package is powerful that as it allows multiple imputations, especially for time series data. Besides, the imputation process not only did well in modelling with varied parameters and continuous data, yet in addition makes full utilisation of the whole dataset to accomplish increased efficiency (Hu, 2017). In addition, Ramli et al. (2013) mentioned that multivariate imputation (EMB) is the ultimate method for MAR type of missing data. A study done by Garciarena and Santana (2017) showed that the same imputation method with the difference in pattern of missing data gave difference in performance

of classification. Therefore, it is important to determine the pattern of missing data when applying the method of imputation.

In addition, according to Junger and De Leon (2015), when the proportion of missing data exceeded 10%, the validity of the estimates degenerated. Therefore, it is suggested to apply multivariate imputation method so that the efficiency loss of information can be lessened. These two imputation methods viz. nearest neighbour and mean showed very small difference in number of $R^2$ and RMSE values for the various proportion of missing values. This concluded that there is no difference between nearest neighbour and mean imputation method in the case of the imputation method. Eventually, complete data together with imputation data were applied for further PCA analysis and model construction.

### 4.8.2 Kaiser-Meyer-Olkin (KMO) and Barlett's tests

Kaiser-Meyer-Olkin (KMO) and Barlett's tests were implemented at the beginning of the analysis before extracting the data in PCA in order to determine the suitability of data for the factor analysis as well as to evaluate the adequacy of the data (Azid et al., $2015_a$). In this study, PCA was run with all of the excepted parameters with a high proportion of missing data, which were UVB, $CH_4$, NMHC and THC. According to Giussani et al. (2016), PCA can effectively manage a low percentage of missing data (7%) and a great percentage of missing data will give unreliable data for the PCA analysis. On top of that, PCA is sensitive to missing data as a result of scarce data given by the parameters and prediction models would be unreliable due to the large amount of missing data experienced by the parameters (Franceschi et al., 2018).

KMO was applied to test the adequacy of samples, which probably might be due to underlying causes (Shrestha & Kazama, 2007). Table 4.12 shows the KMO measure of sampling adequacy. The value of reference point for KMO test was consented to be greater than 0.5 (Ul-Saufie et al., 2013). From the results obtained, the measure of sampling adequacy (MSA) was acceptable as the values of KMO achieved from the test were greater than 0.5 and sufficient for PCA extraction, which were 0.6674, 0.6302 and 0.6591 for HPR, MPR and LPR, respectively.

Table 4.12. Kaiser-Meyer-Olkin measure of sampling adequacy

| | Communality 11 parameters | | |
|---|---|---|---|
| | HPR | MPR | LPR |
| Ws | 0.6900 | 0.7175 | 0.5744 |
| Wd | 0.7146 | 0.6914 | 0.6017 |
| Temp | 0.7110 | 0.5462 | 0.6232 |
| Humidity | 0.5381 | 0.7079 | 0.6552 |
| $NO_x$ | 0.6188 | 0.5755 | 0.6031 |
| NO | 0.6206 | 0.5391 | 0.6014 |
| $SO_2$ | 0.9202 | 0.9299 | 0.9429 |
| $NO_2$ | 0.6770 | 0.6216 | 0.6749 |
| $O_3$ | 0.6213 | 0.7112 | 0.7421 |
| CO | 0.8288 | 0.7868 | 0.8136 |
| $PM_{10}$ | 0.5225 | 0.6081 | 0.5861 |
| KMO | 0.6674 | 0.6302 | 0.6591 |

According to Li and Weng (2016), besides the KMO values, the parameters should be removed from the analysis if they give small values of communality, which indicated that the parameters were unsuitable for the factor solution in PCA extraction. In this analysis, all the parameters were accepted for the analysis since none of the parameters showed small communality values. Therefore, all parameters were considered adequate for further analysis. In addition, the values of KMO obtained from the analysis designated that there was no significant issue of multicollinearity for the data and appropriate to construct PCA since KMO value obtained also showed the identification of similarity in the values of correlation that existed between the parameters (Azid et al., 2018). Therefore, the similarity in correlation between the parameters will be

measured in a similar feature. This confirms that the data would factor well and agreeable to PCA extraction.

Besides KMO, the Barlett's sphericity test was used in the identification of the correlation between parameters and adequacy of the data in order to construct the PCA with the point of reference's value less than 0.05 (Abdullah et al., 2016; Li & Weng, 2016; Ul-Saufie et al., 2013). The result of the Barlett's test obtained showed that the *p*-value was less than 0.0001 with alpha value 0.05 for each region. Therefore, there was adequate data to construct PCA and the result indicated a high degree of relation between the parameters, and thus the data were appropriate for factor analysis.

Since the *p*-value was lower than a significant level of alpha value, which was 0.05, the null hypothesis ($H_o$) was rejected and the alternative hypothesis ($H_a$) was accepted, whereby $H_o$ represents no correlation significantly different from zero between the parameters, while $H_a$ represents at least one of the correlations between the parameters was significantly different from 0. Therefore, parameters of air quality were correlated and not orthogonal, and thus, PCA will allow for variability of interpretation in the data (Azid et al., $2015_a$).

This correlated with Gazzaz et al. (2012), who mentioned that rejecting $H_o$ showed that there was a correlation between the parameters. There were satisfied requirements preceding the PCA subjected to the data, which the KMO of sampling adequacy values obtained were 0.6674, 0.6302 and 0.6591 (>0.50) for HPR, MPR and LPR, respectively, while the Bartlett's test value obtained was <0.0001 (<0.05) for all regions.

### 4.8.3 Normality test

For large samples, the normality test can be done by applying the Jarque-Bera test (Sansuddin et al., 2011), whereby this statistical test can be used to determine the normal distribution of the series. From the analysis done, it was shown that the data were not normally distributed since the *p*-value (< 0.0001) was lower than the significance level alpha (0.05). This result showed that one should reject the null hypothesis ($H_0$) and accept the alternative hypothesis ($H_a$), whereby the $H_0$ represented the extracted variable that followed a normal distribution, while the $H_a$ represents the extracted variable that did not follow a normal distribution.

As stated by Box (1976), in nature, there was never a normal distribution, there never was a straight line, yet with normal and linear assumptions, known to be false. On the other hand, they were still useful approximations to the statisticians. Usually, data transformation was done to ensure that the data is normally distributed. Moreover, transformations can avoid the performance being conquered by any variables (Zhang & Ding, 2017).

This correlated with a study done by He et al. (2015), which mentioned that it was inappropriate to analyse not normalised data directly for prediction as there are different ranges among the parameters used, and they indicated different influenced factors. This coincides with Bing et al. (2015), who stated that the best capability of prediction model is influenced by the ranges of input variables. Since the data applied for this study were not normally distributed, it was best to transform the data first before proceeding with the next analysis to avoid the influence of parameters with different ranges.

## 4.9 Designing of the $PM_{10}$ prediction model

This subsection presents the results from the analysis by using ANN and MLR model. ANN was acquired by a different range of hidden neurons number obtained from the number of input and output parameters. Besides, input parameters obtained from the Pearson coefficient with high correlation, which was greater than 0.75 were applied for each region (HPR, MPR, LPR) in this prediction model of $PM_{10}$. The same procedures were also applied for the MLR model, excluding the experimental analysis of model with hidden neurons number.

There are numerous challenges to predict the air quality. The complexity of the air quality data makes it incompetent to capture the non-linear relation between the dissimilar parameters. Therefore, widely used prediction models of ANN and MLR were applied to predict the $PM_{10}$ concentration. These models were done based on the historical data since the air quality data were established in time series. The models were developed to resolve the public health concerned with the prediction of $PM_{10}$ in air quality.

With the intention of having an insightful assessment between the proposed models, statistical performances of $R^2$, RMSE, IA, E and percentage of deviation were calculated for the ANN and MLR models. Statistical performancse of $R^2$, RMSE, IA and E are presented in Table 4.21 while percentages of deviation for both models are presented in Figure 4.18. Measured statistical performance as shown herein disclosed that the different prediction models, namely ANN and MLR, appeared to have the important influence to predict the concentration of $PM_{10}$. In this study, whole year data were applied, starting from 2010 to 2015 for each region. According to Arhami et al. (2013), it was vital to apply whole year data by bearing in mind the full coverage

variation of the seasonal pollutant levels and meteorological parameters for the air pollution modelling. The prediction model of $PM_{10}$ by using whole year data was crucial to ensure the prediction model for concentrations of $PM_{10}$ can fully describe the quality of air, especially at the specific regions. This can indirectly help people to take early precautions, either by wearing a mask or choosing another route to avoid the location with an unhealthy $PM_{10}$ presence, especially regions that exceed the permissible level since this situation can give negative impacts to human health and the ecosystem.

### 4.9.1 Main sources of pollution for HPR, MPR and LPR

Table 4.13 shows the loading factors of meteorological and pollutant parameters for each region (HPR, MPR, LPR) obtained from the PCA. Lesser factor numbers with the most significant factor was determined by the highest eigenvalues. According to Ahmad Isiyaka and Azid (2015), the eigenvalues were considered significant with a value of one or greater. While the components with eigenvalue of less than 1 were rejected (Bishoi et al., 2009). According to Yong and Pearce (2013), the eigenvalues were used to decide what number of factors to hold. An eigenvalue of less than 1 implied that the factor explained less variance and along these lines ought not to be thought as a significant factor (Pai et al., 2016).

Therefore, eigenvalues of greater than 1 were chosen for input parameters selection by using PCA based on the Pearson coefficient with high loadings. Pearson coefficient with high loading (>0.75) is represented with the black bold font as shown in Table 4.13. Small loading, whether with the negative or positive sign were removed for the interpretation since it was poorly counted for the factor and only large loading, which was greater than 0.75, were chosen for interpretation. Since factor loading is axes

oriented, the loading can be of negative or positive value. Negative loading showed that all station points with positive scores were lower than average value. Meanwhile, positive loading in factor showed that all station points with negative scores were higher than average values for the $PM_{10}$ parameters.

For the HPR, there are three factors that represented 62.40% ofvariability after varimax rotation. Selection of input parameters based on Pearson coefficient with high loading (> 0.75) consisted of $NO_x$, NO, $NO_2$, CO, $O_3$, and temperature. For MPR, there were three factors which represented 58.08% of variability after varimax rotation. Selection of input parameters was based on the Pearson coefficient with high loading (> 0.75), which consisted of $NO_x$, NO, $O_3$ and wind direction. For LPR, there were three factors which represented 56.70% of variability after varimax rotation. Selection of input parameters was based on the Pearson coefficient with high loading (> 0.75), which consisted of $NO_x$, NO, $NO_2$ and temperature.

Each region shows different input parameters. Therefore, it is important to study the source of air pollutants based on the region since the levels and composition of the pollutant might be different from one region to another, relying upon the sources of emission and background. Hence, it is vital to implement monitoring and modelling regionally. In addition, Pawul and Sliwka (2016) suggested to create each model of the neural network, which will work for each of the regions since the selection of its design, including the number of hidden neurons and the selection of input parameters for ANN, can considerably have an impact on its performance. Moreover, each model is trained for every particular region, facing different conditions of weather, emissions of pollutant and traffic conditions (Cortina–Januchs et al., 2015) that will affect the prediction model upon different existing data.

In general, each region experienced high loading of $NO_x$ and NO. However, only MPR experienced moderate loading of $NO_2$ (0.6505), while HPR and LPR experienced high loading of $NO_2$ (0.8164 and 0.8084, respectively). Although $NO_2$ appearred to be in moderate loading in MPR, it could indirectly become one of the important pollutants that deteriorated the air quality. Furthermore, these three pollutants were in the same group of factor loading which was Factor 1 (F1) for each of the regions, which indicated that they were coming from the same sources. Besides, high loading of pollutants in F1 contributed more, followed by F2 and F3 since F1 gave a high percentage of variability as compared to others. In this study, parameters with high loading were chosen because it was more solid to explain the most significant contributor to air deterioration.

Based on the correlation between the pollutant and meteorological parameters (Table 4.7), the CO pollutant showed the highest correlation with $PM_{10}$ as compared to other pollutants with coefficient of correlation ($R^2$) values of 0.4869, 0.5142 and 0.483 for HPR, MPR and LPR, respectively. This was coherent with the study done by Dominick et al. (2012), which found out that CO demonstrated concentration of $PM_{10}$ in the atmosphere. Among the regions, HPR showed the highest loading of CO with a loading value of 0.7837. Although MPR showed the highest correlation between CO and $PM_{10}$, HPR showed the highest value of CO loading, indicating the highest presence of $PM_{10}$ as compared to other regions.

Table 4.13. Loading factors of meteorological and pollutant parameters for (a) HPR (b) MPR and (c) LPR

| | F1 | F2 | F3 |
|---|---|---|---|
| WS (km/hr) | -0.1531 | -0.7418 | -0.0248 |
| WD (°) | 0.0965 | 0.6290 | 0.0319 |
| Temp (°C) | 0.1347 | -0.0728 | **0.7777** |
| Humidity (%) | -0.1982 | 0.7045 | -0.0280 |
| $NO_x$ (ppm) | **0.9520** | 0.1003 | 0.0762 |
| NO (ppm) | **0.9052** | 0.1364 | 0.0143 |
| $SO_2$ (ppm) | 0.4832 | -0.1070 | 0.0026 |
| $NO_2$ (ppm) | **0.8164** | 0.0584 | 0.2744 |
| $O_3$ (ppm) | 0.0829 | 0.1052 | **0.8324** |
| CO (ppm) | **0.7837** | 0.0342 | 0.1373 |
| Eigenvalue | 3.5502 | 1.4632 | 1.2271 |
| Variability (%) | 35.502 | 14.632 | 12.271 |
| Cumulative (%) | 35.502 | 50.133 | 62.404 |

HPR

| | F1 | F2 | F3 |
|---|---|---|---|
| WS (km/hr) | -0.1812 | 0.2057 | -0.6977 |
| WD (°) | 0.0632 | 0.1346 | **0.8123** |
| Temp (°C) | 0.0068 | 0.6003 | -0.0618 |
| Humidity (%) | -0.1423 | -0.5543 | 0.3276 |
| $NO_x$ (ppm) | **0.9518** | 0.0059 | 0.0716 |
| NO (ppm) | **0.9010** | -0.1485 | 0.0070 |
| $SO_2$ (ppm) | 0.3991 | 0.2002 | 0.0229 |
| $NO_2$ (ppm) | 0.6505 | 0.4429 | 0.2328 |
| $O_3$ (ppm) | 0.1299 | **0.7809** | 0.1758 |
| CO (ppm) | 0.6781 | 0.1119 | 0.1271 |
| Eigenvalue | 3.1079 | 1.5244 | 1.1752 |
| Variability (%) | 31.079 | 15.244 | 11.752 |
| Cumulative (%) | 31.079 | 46.323 | 58.075 |

MPR

| | F1 | F2 | F3 |
|---|---|---|---|
| WS (km/hr) | -0.0756 | -0.6980 | -0.0672 |
| WD (°) | 0.0251 | 0.6043 | 0.0639 |
| Temp (°C) | 0.0093 | -0.1296 | **0.7940** |
| Humidity (%) | -0.1205 | 0.6498 | -0.1775 |
| $NO_x$ (ppm) | **0.9519** | -0.0313 | 0.0083 |
| NO (ppm) | **0.8895** | -0.0179 | -0.0328 |
| $SO_2$ (ppm) | 0.4309 | 0.0510 | 0.0676 |
| $NO_2$ (ppm) | **0.8089** | -0.0117 | 0.1942 |
| $O_3$ (ppm) | 0.2344 | 0.1384 | 0.7181 |
| CO (ppm) | 0.6990 | -0.0313 | 0.1519 |
| Eigenvalue | 3.2176 | 1.3189 | 1.1336 |
| Variability (%) | 32.176 | 13.189 | 11.336 |
| Cumulative (%) | 32.176 | 45.365 | 56.701 |

LPR

Pearson coefficient with high loading (>0.75) represented with black bold font

This showed that HPR indirectly contributed to the high presence of $PM_{10}$ concentration as compared to other regions due to the high correlation of CO presence with $PM_{10}$ in HPR. Besides, $NO_x$, NO and $NO_2$ also showed a strong correlation with CO with the correlation between $NO_x$ and CO (0.6757), NO and CO (0.6228), $NO_2$ and CO (0.6102) in HPR. This indirectly showed that HPR was one of the most contributor to the presence of $PM_{10}$, which was the main pollutant in air due to the high loadings of pollutants, namely $NO_x$, NO and $NO_2$ since these types of pollutants has high correlation with CO, whereby CO was the only pollutant that has high correlation with the presence of $PM_{10}$. According to Masiol et al. (2017), $NO_x$, CO and $PM_{10}$ were significantly higher in urban areas due to the higher anthropogenic pressure.

Besides, a study done by Azid et al. ($2015_b$) found that the existence of $NO_x$, NO, $NO_2$ and CO were produced from activities related to transportation and agriculture. $NO_x$, NO and $NO_2$ in the atmosphere came from the anthropogenic sources, whichowe to the traffic or combustion processes from engine cars. Besides transportation, Mukhopadhyay and Forssell (2005) mentioned that the existence of $NO_x$, NO, $NO_2$ and CO was associated with the fossil fuel combustion of the agricultural systems. This envisages that agriculture perhaps was one of the main factors that contributed to the presence of these pollutants instead of transportation. For example, deposition of nitrogen due to the agricultural practices that caused sequestered of additional carbon (De Vries et al., 2006).

In spite of the fact that industrialisation and urbanisation are constantly featured in HPR, HPR were also mostly surrounded by the districts of agriculture. Agriculture in Selangor is similarly important for the states along with the country as a whole. Agricultural lands in Selangor can be found in almost all districts, namely, Gombak,

Hulu Langat, Hulu Selangor, Klang, Kuala Langat, Kuala Selangor, Petaling, Sabak Bernam and Sepang but are concentrated in the district of Hulu Selangor and Kuala Langat for the production of main vegetables and cash crops, respectively (DOA, 2015). Since most of the air quality stations in Selangor were categorised as HPR, agriculture in Selangor has become one of the contributors to the presence of $NO_x$, NO, $NO_2$ and CO.

On top of that, high loading of $NO_x$, NO, $NO_2$ and $O_3$ was probably due to the transboundary haze pollution through forest fire in Indonesia, which is one of the most severe regional air pollution issues that occur every year that influenced the air quality in Southeast Asia (SEA),which consista of Malaysia, Singapore, Brunei, Indonesia and Southern Thailand. According to Rahman et al. (2016), haze that came from the biomass burning in Indonesia was the key reason for the hazardous catastrophe. Peat fires are reflected as one of the main causes of smoke haze in the midst of the other types of biomass burned in Indonesia (Ahmed et al., 2016).

According to Ahmed et al. (2016), volatile organic compound (VOC) was one of the elements that is emitted during biomass burning and it was also one of the $O_3$ precursors (Ahamad et al., 2014). This situation will indirectly contribute to the presence of $O_3$. Likewise, air masses regarded as dry and warm environments frequently related to the haze occurrence in Malaysia also contributed to the highest $O_3$ concentration (Abdullah et al., $2017_b$). Burning substances containing nitrogen during the forest fires would emit NO and $NO_2$ (Behera & Balasubramanian, 2014). Besides, the emission of $NO_x$ also relies on the composition of biomass, whereby $NO_x$ will increase with the increased content of nitrogen (Mladenović et al., 2016).

Besides pollutant parameters, meteorological parameters, namely, wind direction and temperature give high loading values. A meteorological parameter is usually related to the monsoon seasons (Masseran & Razali, 2016). This is supported through the study done by Bertazzon and Shahid (2017), where areas in the Northeast and South in Canada give poor air quality due to the sources of seasonal pollution as well as prevailing winds. According to Wen et al. (2016), wind direction has a directproportional impact to the haze problem that occurred in Malaysia instead of forest fire severity.

Furthermore, wind direction is one of the pollution sources present in a certain area. This was approved by a study done by Ahmed et al. (2016), which found that almost all the paths of wind originate from Indonesia (Sumatra and Kalimantan) to their study area at Kampar in the north of Peninsular Malaysia. Therefore, these wind paths that originate from Indonesia to Peninsular Malaysia will indirectly increase the concentration of $PM_{10}$, especially during the local peat fire in Sumatra and Kalimantan that caused haze in Malaysia. These showed that meteorological factors, especially wind direction play an important role in the formation of $PM_{10}$.

Besides, wind direction is one of the important agents that control the transport and dilution of $O_3$ (Toh et al, 2013). This was supported by Gong et al. (2018) who mentioned that wind direction has a great influence on $O_3$ concentration. Besides, it also influences the distribution of $NO_2$ (Gorai et al., 2015) as well as playing an important role in the high concentration of $PM_{10}$ presence (Dotse et al., 2016). According to the study done by Charabi et al. (2018), in Oman during the monsoon season, blowing of wind from the south-eastern direction were major factors that affected the vehicle dispersion pollutants of CO and $NO_X$.

Meanwhile, temperature affects the air quality due to the temperate inversion, with the warm air above cooler air acts as a lid, thereby trapping the cooler air at the surface and suppressing vertical mixing (Akhtar et al., 2018). This was supported by Trinh et al. (2018) who found that the temperature inversion values were 22.3% higher than the normal values in Hanoi, Vietnam and recorded a slightly higher hospital admission during temperature inversion as compared to the normal days.

According to Latif et al. (2014), higher temperature in tropical countries like Malaysia cause high evaporation and particles resuspension in ambient air. Particles amount increase in ambient air due to the hot weather from the high temperature. Weather shifting also influence the air pollution since transport, dispersion, and transformation of air pollutants at multiple scales change due to changes in temperature as well as wind (Dias et al., 2012).

### 4.9.2 Prediction model of $PM_{10}$ for HPR, MPR and LPR by ANN and MLR models

Table 4.14 to Table 4.16 show the values of $R^2$ and RMSE obtained from the ANN and MLR models for prediction model of $PM_{10,}$ whereby the ANN model was analysed based on the different ranges of hidden nodes with a hidden layer according to the number of input parameters obtained from the Pearson coefficient with high loading ($> 0.75$) for each region (HPR, MPR and LPR) as discussed before in the main source of pollution for HPR, MPR and LPR. According to Ababneh et al. (2014), in assessing the performance of air quality models, the correlation coefficient ($R^2$) and the root mean square error (RMSE) gave sufficient information.

Since the hidden nodes reflected to be of great importance for ANN, the performance of different hidden nodes obtained based on the numbers of input and output were implemented. Eventually, the hidden nodes with maximum $R^2$ and minimum RMSE were selected. It was shown that the network with 10, 5 and 4 hidden nodes in the hidden layer was the best performance of the ANN model for HPR, MPR and LPR with $R^2$ (RMSE) of 0.4736 (24.14), 0.1851 (23.63) and 0.0760 (19.66), respectively. This showed that network analysis did not perform better as the hidden nodes increased but the performance depended on the selected optimal hidden nodes based on the measured performance results viz. $R^2$ and RMSE. This proved that hidden node plays an important role in obtaining the accurate model.

The analyses for the prediction model of $PM_{10}$ were continued by applying them onto the MLR model. Table 4.14 until Table 4.16 show the values of $R^2$ and RMSE obtained from MLR based on the input parameters obtained from the Pearson coefficient with high loading (>0.75) for each region (HPR, MPR and LPR). The performance of $R^2$ (RMSE) for HPR, MPR and LPR was 0.3407 (27.02), 0.1332 (23.98) and 0.0569 (20.05), respectively. PCA basically decreased the series of extents, whereby it changed several correlated parameters into a small number of parameters that covered great amounts of the original data (Bai et al., 2018). Intrinsically, the multivariate statistical analysis utilisation, for example, PCA among the ANN and MLR modellings will facilitate us to extract the important statistical findings (AhmadIsiyaka et al., 2014). Eliminating the correlation between parameters from PCA is believed to improve the prediction performance. Moreover, pollutants and meteorological are believed to contribute to the presence of $PM_{10}$ in their own ways. Therefore, only the main sources of pollution at that region were chosen as input parameters.

According to Jayamurugan et al. (2013), along with sources of emission, meteorological factors, such as wind speed, wind direction, temperature and humidity can influence the variability of atmospheric pollutants. A study done by Munir et al. (2013) showed that $PM_{10}$ level was not reliant on the sources of emission but also on meteorological parameters, for instance, wind, temperature and humidity. Furthermore, the fate and composition of $PM_{10}$ during the transport processes will be determined by meteorological factors, such as temperature, rainfall, humidity, and atmospheric stability besides physical and chemical interactions (Dotse et al., 2016).

From the results in Table 4.14 to Table 4.16, it seemed that the $PM_{10}$ prediction models by ANN and MLR were less accurate with $R^2$ value of below 0.5 for each region. Therefore, an attempt was made by using lagged 1-day data as the input parameter for HPR, MPR and LPR to obtain more precise $PM_{10}$ prediction model. To find out the effectiveness of the lagged day data being used as the input parameter, lagged 2-day and 3-day data of $PM_{10}$ were also implemented as input parameters besides the lagged 1-day data.

Table 4.14. Results of the ANN and MLR for HPR

| | | | $R^2$ | |
|---|---|---|---|---|
| | Input variables | HN | ANN | MLR |
| No lagged $PM_{10}$ data used for prediction model of $PM_{10}$ ($PM_{10}(t+1)$) | $NO_x$, NO, $NO_2$, CO, Temp, $O_3$ | 5 | 0.4606 (24.43) | 0.3407 (27.02) |
| | | 6 | 0.4450 (24.78) | |
| | | 7 | 0.4545 (24.57) | |
| | | 8 | 0.4634 (24.37) | |
| | | 9 | 0.4682 (24.26) | |
| | | **10** | **0.4736 (24.14)** | |
| | | 11 | 0.4726 (24.16) | |

Note: values in parentheses are the RMSE values

Table 4.15. Results of the ANN and MLR for MPR

| | | | $R^2$ | |
|---|---|---|---|---|
| | Input variables | HN | ANN | MLR |
| No lagged $PM_{10}$ data used for prediction model of $PM_{10}$ ($PM_{10}(t+1)$) | $NO_x$, NO, $O_3$, Wind direction | 3 | 0.1435 (24.23) | 0.1332 (23.98) |
| | | 4 | 0.1840 (23.64) | |
| | | **5** | **0.1851 (23.63)** | |
| | | 6 | 0.1730 (23.81) | |
| | | 7 | 0.1848 (23.63) | |

Note: values in parentheses are the RMSE values

Table 4.16. Results of the ANN and MLR for LPR

| | | | $R^2$ | |
|---|---|---|---|---|
| | Input variables | HN | ANN | MLR |
| No lagged $PM_{10}$ data used for prediction model of $PM_{10}$ ($PM_{10}(t+1)$) | $NO_x$, NO, $NO_2$, Temperature | 3 | 0.0684 (19.74) | 0.0569 (20.05) |
| | | **4** | **0.0760 (19.66)** | |
| | | 5 | 0.0618 (19.81) | |
| | | 6 | 0.0630 (19.79) | |
| | | 7 | 0.0646 (19.78) | |

Note: values in parentheses are the RMSE values

### 4.9.3 Prediction model of $PM_{10}$ using lagged 1-day, 2-day and 3-day of $PM_{10}$ data for HPR, MPR and LPR by ANN and MLR models

Table 4.17 shows the loading factors of meteorological and pollutant parameters, including $PM_{10}$ pollutant for each region (HPR, MPR, LPR) obtained from PCA. For the HPR, there were four factors that represented 67.90% percentage of variability after varimax rotation. Selection of input parameters based on the Pearson coefficient with high loading (> 0.75) consisted of $NO_x$, NO, $NO_2$, $PM_{10}$, $O_3$, temperature.

For MPR, there were four factors that represented 65.09% percentage of variability after varimax rotation. Selection of input parameters based on Pearson coefficient with high loading (> 0.75) consisted of $NO_x$, NO, $PM_{10}$, temperature and wind direction. For LPR, there were four factors that represented 63.26% percentage of variability after varimax rotation. Selection of input parameters based on the Pearson coefficient with high loading (> 0.75) consisted of $NO_x$, NO, $NO_2$, $PM_{10}$ and temperature. The results obtained from the PCA showed that $PM_{10}$ was one of the main sources of pollution and perhaps can enhance the prediction model of $PM_{10}$ by including the lagged concentration of $PM_{10}$.

Table 4.18 until Table 4.20 shows the values of $R^2$ and RMSE obtained from the ANN and MLR models for the $PM_{10}$ prediction model using lagged 1-day, 2-day and 3-day data for HPR, MPR and LPR respectively. The ANN model was analysed based on the different ranges of hidden nodes with a hidden layer according to the number of input parameters obtained from the Pearson coefficient with high loading (> 0.75) for each region (HPR, MPR and LPR), including the $PM_{10}$ parameter. It is shown that the network with 7, 6 and 6 hidden nodes in the hidden layer gave the best performance of the ANN model for HPR, MPR and LPR 1-day lagged $PM_{10}$ data were used as input

for $PM_{10}$ prediction model as compared to 2-day and 3-day lagged $PM_{10}$ data as input with $R^2$ (RMSE) of 0.7718 (16.11), 0.7590 (12.63) and 0.7721 (9.86), respectively.

The analyses for the prediction model of $PM_{10}$ using lagged 1-day, 2-day and 3-day $PM_{10}$ data were continued by applying to the MLR model. Table 4.18 also shows the values of $R^2$ and RMSE obtained from MLR based on the input parameters obtained from the Pearson coefficient with high loading (>0.75) for each region (HPR, MPR and LPR), including a parameter of $PM_{10}$. The best performance values of $R^2$ (RMSE) for HPR, MPR and LPR were 0.7609 (16.27), 0.7442 (13.03) and 0.7642 (10.02), respectively, when 1-day lagged $PM_{10}$ data was used as input for the $PM_{10}$ prediction model as compared to the 2-day and 3-day lagged $PM_{10}$ data.

From the results in Table 4.18 until Table 4.20, it seemed that the prediction model of $PM_{10}$ by both models showed the best performance when using 1-day lagged $PM_{10}$ data were used as input rather than the 2-day and 3-day lagged $PM_{10}$ data as input for HPR, MPR and LPR. However, the 2-day and 3-day lagged $PM_{10}$ data performed better than the prediction model of $PM_{10}$ with no lagged $PM_{10}$ data. Sayegh et al. (2014) mentioned that in the atmosphere, particles can stay longer and indirectly contribute to the concentration of particle present days later.

This showed that the concentration of particles present in the atmosphere today could affect to the concentration of particles present in the atmosphere for the next day. Conversely, a day later gave the best performance of $PM_{10}$ prediction model than concentration of $PM_{10}$ present two and three days later. The best ANN models obtained for the $PM_{10}$ prediction model in Malaysia based on HPR, MPR and LPR are shown in Figure 4.17.

Table 4.17. Loading factors of meteorological and pollutant parameters (including $PM_{10}$ pollutant) for (a) HPR (b) MPR and (c) LPR

| | F1 | F2 | F3 | F4 |
|---|---|---|---|---|
| WS (km/hr) | -0.1674 | -0.7363 | -0.0270 | 0.0199 |
| WD (°) | 0.0935 | 0.6427 | 0.0110 | 0.0600 |
| Temp (°C) | 0.1665 | -0.0998 | **0.8105** | -0.1279 |
| Humidity (%) | -0.1700 | 0.6944 | -0.0125 | -0.1297 |
| $NO_x$ (ppm) | **0.9697** | 0.0725 | 0.0911 | -0.0192 |
| NO (ppm) | **0.9323** | 0.1030 | 0.0370 | -0.0764 |
| $SO_2$ (ppm) | 0.4396 | -0.0758 | -0.0481 | 0.2743 |
| $NO_2$ (ppm) | **0.8115** | 0.0498 | 0.2684 | 0.1044 |
| $O_3$ (ppm) | 0.0446 | 0.1356 | **0.7938** | 0.2692 |
| CO (ppm) | 0.7185 | 0.0493 | 0.0975 | 0.4815 |
| $PM_{10}$ ($\mu g/m^3$) | 0.1019 | -0.0646 | 0.0878 | **0.9309** |
| Eigenvalue | 3.6369 | 1.5267 | 1.2581 | 1.0477 |
| Variability (%) | 33.063 | 13.879 | 11.437 | 9.525 |
| Cumulative (%) | 33.063 | 46.942 | 58.379 | 67.904 |

HPR

| | F1 | F2 | F3 | F4 |
|---|---|---|---|---|
| WS (km/hr) | -0.1797 | 0.0287 | 0.1862 | -0.7156 |
| WD (°) | 0.0304 | 0.0861 | 0.1001 | **0.8186** |
| Temp (°C) | 0.0112 | -0.0931 | **0.7558** | 0.0442 |
| Humidity (%) | -0.1401 | -0.0618 | -0.5961 | 0.2888 |
| $NO_x$ (ppm) | **0.9566** | 0.0913 | 0.0783 | 0.1247 |
| NO (ppm) | **0.9397** | 0.0476 | -0.0201 | 0.0768 |
| $SO_2$ (ppm) | 0.3462 | 0.2980 | 0.0732 | -0.0289 |
| $NO_2$ (ppm) | 0.5702 | 0.3895 | 0.3459 | 0.2246 |
| $O_3$ (ppm) | 0.0006 | 0.5180 | 0.6032 | 0.1392 |
| CO (ppm) | 0.5268 | 0.6483 | -0.0855 | 0.0646 |
| $PM_{10}$ ($\mu g/m^3$) | 0.0031 | **0.8875** | 0.0143 | -0.0066 |
| Eigenvalue | 3.2432 | 1.6049 | 1.2301 | 1.0816 |
| Variability (%) | 29.484 | 14.590 | 11.183 | 9.833 |
| Cumulative (%) | 29.484 | 44.073 | 55.256 | 65.089 |

MPR

| | F1 | F2 | F3 | F4 |
|---|---|---|---|---|
| WS (km/hr) | -0.0618 | -0.0335 | -0.7037 | 0.0032 |
| WD (°) | 0.0410 | -0.1219 | 0.6308 | 0.1610 |
| Temp (°C) | 0.0233 | 0.0863 | -0.0067 | **0.8687** |
| Humidity (%) | -0.1504 | 0.0713 | 0.5999 | -0.3881 |
| $NO_x$ (ppm) | **0.9692** | 0.0294 | 0.0004 | 0.0677 |
| NO (ppm) | **0.9234** | -0.0592 | 0.0159 | 0.0669 |
| $SO_2$ (ppm) | 0.3862 | 0.2889 | 0.0355 | -0.1504 |
| $NO_2$ (ppm) | **0.7829** | 0.2383 | 0.0189 | 0.1254 |
| $O_3$ (ppm) | 0.1465 | 0.5497 | 0.1936 | 0.4155 |
| CO (ppm) | 0.5909 | 0.5571 | -0.0506 | 0.0025 |
| $PM_{10}$ ($\mu g/m^3$) | 0.0473 | **0.8669** | -0.1370 | 0.0328 |
| Eigenvalue | 3.3232 | 1.3805 | 1.2489 | 1.0064 |
| Variability (%) | 30.211 | 12.550 | 11.353 | 9.149 |
| Cumulative (%) | 30.211 | 42.761 | 54.114 | 63.263 |

LPR

Pearson coefficient with high loading (>0.75) represent with black bold font

Table 4.18. Results of the ANN using lagged 1-day, 2-day and 3-day $PM_{10}$ data for HPR

| | Input variables | HN | $R^2$ | |
|---|---|---|---|---|
| | | | ANN | MLR |
| 1-day lagged $PM_{10}$ data used for prediction model of $PM_{10}(t+1)$ | $NO_x$, NO, $NO_2$, Temp, $O_3$, $PM_{10}$ | 5 | 0.7656 (16.33) | 0.7609 (16.27) |
| | | 6 | 0.7647 (16.36) | |
| | | **7** | **0.7718 (16.11)** | |
| | | 8 | 0.7621 (16.08) | |
| | | 9 | 0.7713 (16.12) | |
| | | 10 | 0.7676 (16.25) | |
| | | 11 | 0.7619 (16.08) | |
| 2-day lagged $PM_{10}$ data used for next 2 day prediction of $PM_{10}(t+2)$ | $NO_x$, NO, $NO_2$, Temp, $O_3$, $PM_{10}$ | 5 | 0.5332 (22.62) | 0.5129 (23.23) |
| | | 6 | 0.5311 (22.68) | |
| | | 7 | 0.5339 (22.61) | |
| | | 8 | 0.5330 (22.63) | |
| | | 9 | 0.5339 (22.61) | |
| | | **10** | **0.5363 (22.55)** | |
| | | 11 | 0.5334 (22.62) | |
| 3-day lagged $PM_{10}$ data used for next 3 day prediction of $PM_{10}(t+3)$ | $NO_x$, NO, $NO_2$, Temp, $O_3$, $PM_{10}$ | 5 | 0.4256 (25.23) | 0.3856 (26.08) |
| | | 6 | 0.4252 (25.24) | |
| | | 7 | 0.4251 (25.25) | |
| | | **8** | **0.4272 (25.20)** | |
| | | 9 | 0.4253 (25.24) | |
| | | 10 | 0.4231 (25.29) | |
| | | 11 | 0.4271 (25.19) | |

Note: values in parentheses are the RMSE values

Table 4.19. Results of the ANN using lagged 1-day, 2-day and 3-day $PM_{10}$ data for MPR

| | Input variables | HN | $R^2$ | |
|---|---|---|---|---|
| | | | ANN | MLR |
| 1-day lagged $PM_{10}$ data used for prediction model of $PM_{10}(t+1)$ | $NO_x$, NO, Temp, WD, $PM_{10}$ | 4 | 0.7505 (12.85) | 0.7442 (13.03) |
| | | 5 | 0.7523 (12.80) | |
| | | **6** | **0.7590 (12.63)** | |
| | | 7 | 0.7527 (12.79) | |
| | | 8 | 0.7523 (12.80) | |
| | | 9 | 0.7552 (12.73) | |
| 2-day lagged $PM_{10}$ data used for next 2 day prediction of $PM_{10}(t+2)$ | $NO_x$, NO, Temp, WD, $PM_{10}$ | 4 | 0.4794 (18.63) | 0.4594 (18.94) |
| | | 5 | 0.4804 (18.61) | |
| | | 6 | 0.4811 (18.59) | |
| | | 7 | 0.4810 (18.60) | |
| | | 8 | 0.4826 (18.57) | |
| | | 9 | 0.4803 (18.61) | |
| 3-day lagged $PM_{10}$ data used for next 3 day prediction of $PM_{10}(t+3)$ | $NO_x$, NO, Temp, WD, $PM_{10}$ | 4 | 0.3328 (20.99) | 0.3139 (21.34) |
| | | 5 | 0.3357 (20.94) | |
| | | 6 | 0.3321 (21.00) | |
| | | 7 | 0.3322 (21.00) | |
| | | 8 | 0.3294 (21.04) | |
| | | 9 | 0.3320 (21.00) | |

Note: values in parentheses are the RMSE values

Table 4.20. Results of the ANN using lagged 1-day, 2-day and 3-day $PM_{10}$ data for LPR

| | Input variables | HN | $R^2$ | |
|---|---|---|---|---|
| | | | ANN | MLR |
| 1-day lagged $PM_{10}$ data used for prediction model of $PM_{10}(t+1)$ | $NO_x$, NO, $NO_2$, Temp, $PM_{10}$ | 4 | 0.7713 (9.88) | 0.7642 (10.02) |
| | | 5 | 0.7715 (9.88) | |
| | | **6** | **0.7721 (9.86)** | |
| | | 7 | 0.7712 (9.89) | |
| | | 8 | 0.7706 (9.90) | |
| | | 9 | 0.7719 (9.87) | |
| 2-day lagged $PM_{10}$ data used for next 2-day prediction of $PM_{10}(t+2)$ | $NO_x$, NO, $NO_2$, Temp, $PM_{10}$ | 4 | 0.5387 (13.97) | 0.5297 (14.16) |
| | | 5 | 0.5398 (13.95) | |
| | | 6 | 0.5370 (14.00) | |
| | | 7 | 0.5363 (14.01) | |
| | | 8 | 0.5396 (13.96) | |
| | | 9 | 0.5432 (13.90) | |
| 3-day lagged $PM_{10}$ data used for next 3-day prediction of $PM_{10}(t+3)$ | $NO_x$, NO, $NO_2$, Temp, $PM_{10}$ | 4 | 0.4135 (15.89) | 0.4043 (15.94) |
| | | 5 | 0.4172 (15.84) | |
| | | 6 | 0.4181 (15.83) | |
| | | 7 | 0.4186 (15.82) | |
| | | 8 | 0.4201 (15.80) | |
| | | 9 | 0.4214 (15.78) | |

Note: values in parentheses are the RMSE values

(a)

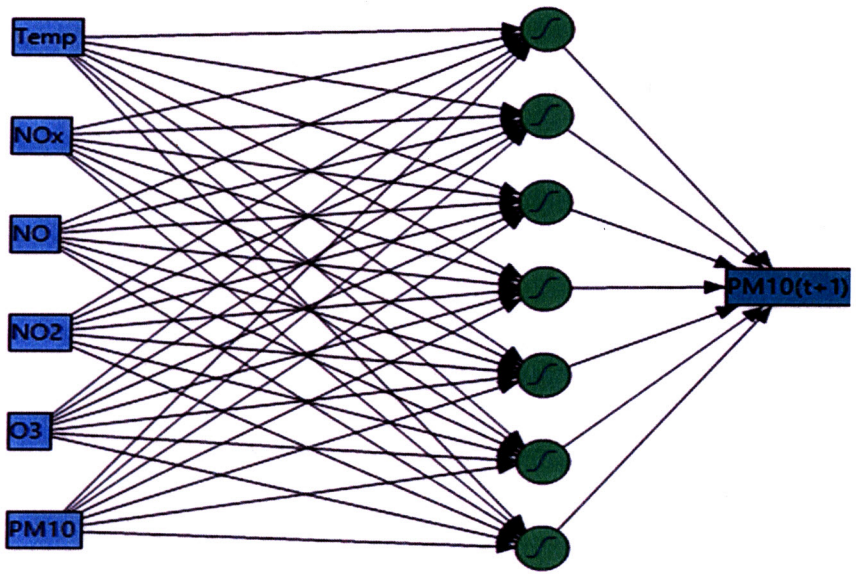



Figure 4.17. The best ANN model obtained after reducing the inputs through PCA for (a) HPR (b) MPR (c) LPR

(b)

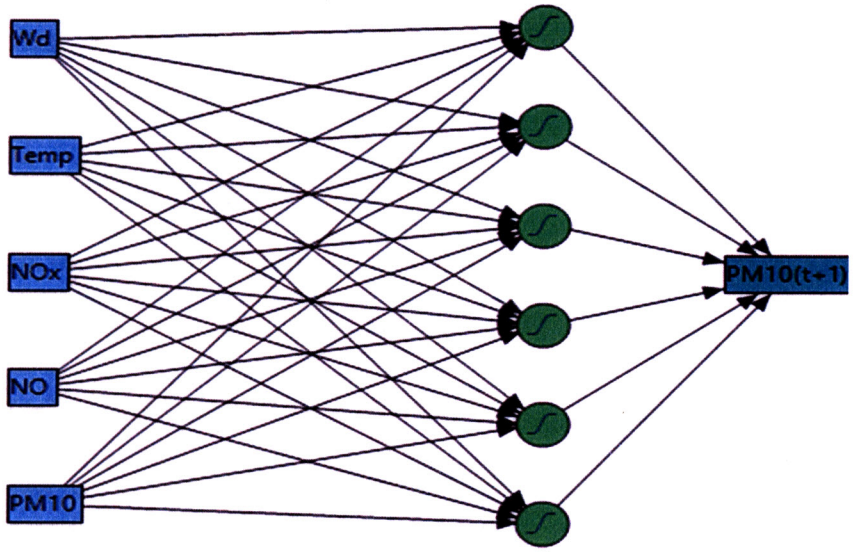



Figure 4.17. Continued

(C)

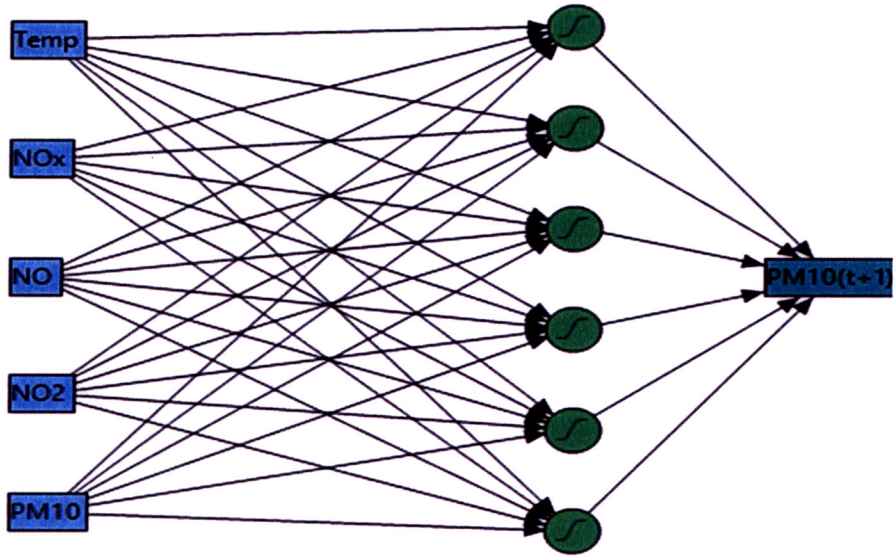



Figure 4.17. Continued

### 4.9.4 Model evaluation for prediction model of $PM_{10}$ by ANN and MLR models

There are numerous performance measurements applied to evaluate the ability of ANN and MLR for the prediction. In this research, the performance and accuracy of the $PM_{10}$ prediction model were evaluated by using $R^2$, RMSE, Index of agreement (IA), Efficiency (E) as well as the percentage of deviation between predicted $PM_{10}$ and observed $PM_{10}$. Table 4.21 shows the performance statistics of $R^2$, RMSE, IA and E obtained from the prediction model of ANN and MLR for each of the regions. The optimal values that gave better performance and highest accuracy between actual and predicted values were near to 0 for RMSE and 1 for $R^2$, IA and E.

According to Özdemir and Taner (2014), the model is more efficient if the $R^2$ is high and RMSE is low. Relations between predicted and actual $PM_{10}$ were clarified by the values of $R^2$ for ANN and MLR statistical models with the best $R^2$ values shown by the ANN model as compared to the MLR model for each region. $R^2$ values obtained through the ANN model for HPR, MPR and LPR were 0.7718, 0.7590 and 0.7721, respectively.

Meanwhile, $R^2$ values obtained through the MLR model for HPR, MPR and LPR were 0.7609, 0.7442 and 0.7642 respectively. Besides, the ANN model also showed a low value of RMSE than the MLR model for each region with RMSE values obtained through the ANN model for HPR, MPR and LPR were 16.11, 12.63 and 9.80, respectively. Meanwhile, the RMSE values obtained through the MLR model for HPR, MPR and LPR were 16.27, 13.03 and 10.02, respectively. Based on the results acquired from the two models, it is obvious that ANN gives better prediction ability and accuracy in predicting $PM_{10}$ concentration contrast with MLR. In general, it is proven that the accuracy of ANN prediction is better than traditional statistical prediction (MLR).

According to Nash and Sutcliffe (1970), E is equal to 1 if all the observed values are the same as all predictions, while E value is between 0 and 1 if there are deviations between observed and predicted values. Meanwhile, the predictions are very poor if the value of E is in a negative value. Each of the models for each region showed a deviation between the predicted and actual $PM_{10}$ concentration since the E value obtained was between 0 to 1 for both models, with the values of E were 0.7663, 0.7510 and 0.7657 for HPR, MPR and LPR, respectively, by the ANN model, while the values of E were 0.7609, 0.7442 and 0.7642 for HPR, MPR and LPR, respectively, by the MLR model.

Although both models experienced deviation between observed and predicted $PM_{10}$, both models showed good agreement between the predicted and actual concentration of $PM_{10}$ with the values of IA obtained were 0.9307, 0.9230, 0.9302 by ANN and 0.9277, 0.9217, 0.9292 by MLR for HPR, MPR and LPR, respectively. Generally, the overall performance in predicting the $PM_{10}$ concentration based on the results

presented so far is satisfactory for ANN and MLR models. However, the accuracy of the MLR model was slightly lower as compared to the ANN model.

Table 4.21. Value of $R^2$, RMSE, IA and E obtained from the prediction model by using ANN and MLR for each region

| Region | Prediction model | $R^2$ (1) | RMSE (0) | IA (1) | E (1) |
|---|---|---|---|---|---|
| HPR | ANN | **0.7718** | **16.11** | **0.9307** | **0.7663** |
| | MLR | 0.7609 | 16.27 | 0.9277 | 0.7609 |
| MPR | ANN | **0.7590** | **12.63** | **0.9230** | **0.7510** |
| | MLR | 0.7442 | 13.03 | 0.9217 | 0.7442 |
| LPR | ANN | **0.7721** | **9.80** | **0.9302** | **0.7657** |
| | MLR | 0.7642 | 10.02 | 0.9292 | 0.7642 |

The pie chart in Figure 4.18 shows the distribution percentage of deviation for predicted and observed $PM_{10}$ obtained from ANN and MLR models for HPR, MPR and LPR. According to Yang et al. (2016), the data is correctly predicted if the deviation of the sample is less than 20% while it is less correctly predicted if the deviation of the sample is greater than 70%. Both models showed high distribution percentage at 74.25%, 75.43% and 77.98% for HPR, MPR and LPR, respectively by ANN.

Meanwhile, the values of distribution percentage of deviation obtained from the MLR model were 74.07%, 75.16% and 77.98%, respectively. In addition, both models also showed low distribution percentage of deviation for HPR, MPR and LPR. This strongly agreed that both models correctly predicted $PM_{10}$ pollutant by giving high and low percentage of distribution percentage of deviation. Both models showed that the predicted $PM_{10}$ data matched the observed values very well on the whole sequence of six years. However, in general, the ANN model was chosen for $PM_{10}$ prediction model since accuracy obtained from this model was higher as compared to MLR.

By comparing all the results obtained from ANN and MLR, the ANN can model the intricate and non-linear relationship between meteorological and pollutant parameters due to its flexibility (AhmadIsiyaka et al., 2014). According to Bai et al. (2018), ANN has good capability in fitting the non-linear character of $PM_{10}$ that was probably due to the emitted $PM_{10}$ subjected by a few meteorological parameters, for instance, wind speed, temperature and humidity, which varied it .

Therefore, it is important to apply the non-linear model instead of the linear model for the nonlinearity $PM_{10}$ pollutant. ANN is a well-known method that is used in statistical modelling in regard to the predicting of non-linear air pollution. Transmission, deposition, and scattering of $PM_{10}$ concentration in the atmosphere were controlled by the meteorological parameters. In addition, there are many parameters that influence pollution and the relation is difficult for enlightening and enhancing the ANN prediction accuracy (Bai et al., 2018). Therefore, it is imperative to consider the meteorological parameters in the development of $PM_{10}$ prediction model besides the pollutants.

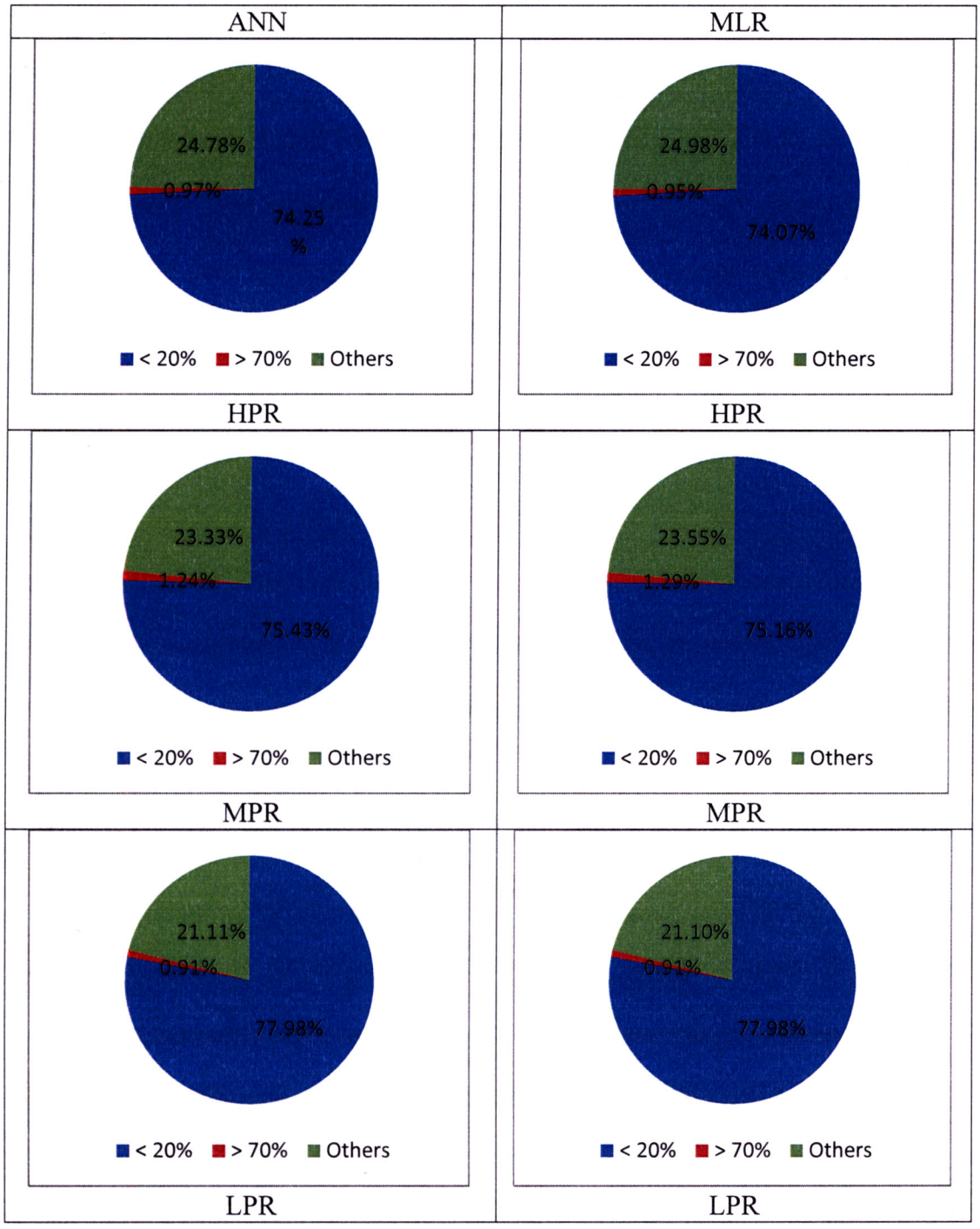



Figure 4.18. The percentage deviation of predicted and observed $PM_{10}$ for HPR, MPR and LPR obtained from ANN and MLR models

Less accurate predictions were attained for the MLR linear model that assumed a linear relationship between the meteorological and pollutant parameters as compared to the non-linear ANN model. Therefore, the results acquired with the linear regression model were less accurate than the ones acquired with the ANN models (Cortina–

Januchs et al., 2015). Besides, ANN can generalise by considering the designs utilised for the duration of the training process (Elangasinghe et al., 2014). Generally, the ANN model offered the potential for prediction model of $PM_{10}$ with the ANN model being clearly superior to the nonlinear method.

Figure 4.19 shows the scatter plot of observed $PM_{10}$ in comparison with predicted $PM_{10}$ for each region (HPR, MPR, LPR). According to Alimissis et al. (2018), the error magnitude for low, medium and high concentration levels can be examined through the scatter diagram. From the graph plotted, some of the data points visualised the fluctuations of daily $PM_{10}$ levels for each of the region that could interrupted the result of the prediction. According to Syafei (2014), the fluctuation and the range of training dataset caused the errors that indirectly influenced the prediction.

The error occurred especially at the high season since the external factors, such as wind speed, wind direction, humidity and temperature influence directly the level of $PM_{10}$ (Wongsathan & Seedadan, 2016). The ANN model demonstrates its capability if the training data set used is not highly fluctuated. However, in order to consider all seasonal and meteorological pollutant parameters among the years from 2010 to 2015 in the prediction model, all the training data were used for the prediction model. A perfect line created by the points would yield perfect predictions. A common approach of error can be used to measure how far away the predicted value is from the actual value.

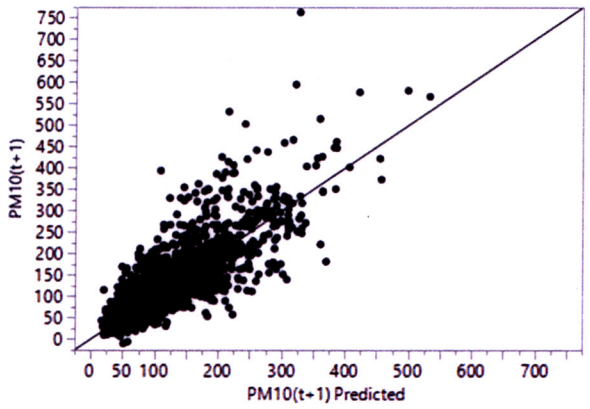
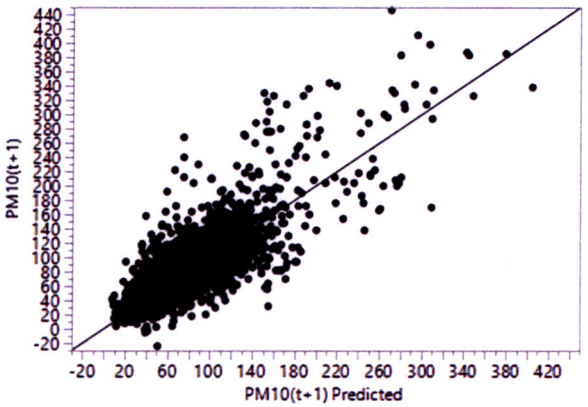
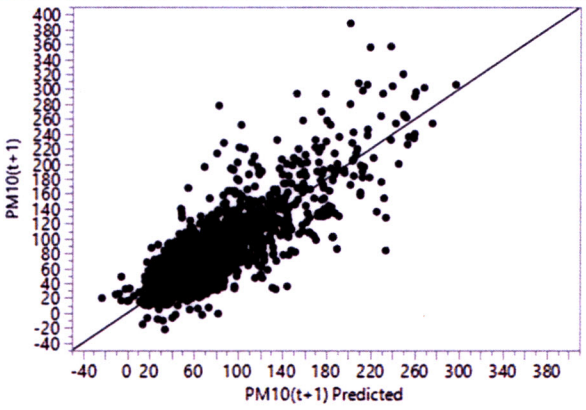
| Regions | Scatter plot of observed $PM_{10}$ vs. predicted $PM_{10}$ |
|---|---|
| HPR |  |
| MPR |  |
| LPR |  |



Figure 4.19. Scatter plot of observed $PM_{10}$ in comparison with predicted $PM_{10}$ for each region (HPR, MPR, LPR)

Nevertheless, the statistical performance showed that the model accurately predicted the concentration of $PM_{10}$ since the fluctuated data contributed less percentage on the prediction model (Figure 4.18). Table 4.22 shows the summary of equations derived from ANN for the $PM_{10}$ prediction model for each region. The equations were derived from the input parameters obtained from PCA. These equations can be applied to predict the concentration of $PM_{10}$ at each region.

Table 4.22. Equations derived from ANN for the $PM_{10}$ prediction model

| Region | Models (ANN) |
|---|---|
| HPR | $PM_{10}(t+1) = 8.18 + 30.55(NO_x) - 34.84(NO) - 19.62(NO_2) + 0.056(Temp) - 58.90(O_3) + 0.88(PM_{10})$ |
| MPR | $PM_{10}(t+1) = 9.34 - 30.78(NO_x) + 20.86(NO) + 0.86(PM_{10}) - 0.11(Temp) + 0.0065(Wd)$ |
| LPR | $PM_{10}(t+1) = 7.15 + 19.47(NO_x) - 33.05(NO) - 20.74(NO_2) + 0.88(PM_{10}) - 0.053(Temp)$ |

## 4.10 Summary

Within this chapter, continuous air quality monitoring stations in Malaysia were classified into HPR, MPR and LPR with the reliability of the results confirmed by DA. The trend analysis showed that $PM_{10}$ was the most dominant parameter as compared to others in the year 2010 to 2015. For the prediction model, ANN performed better than MLR to predict the concentration of $PM_{10}$ with lagged $PM_{10}$ being used as one of the input parameters. Lagged 1-day $PM_{10}$ concentration gave a better prediction model as compared to lagged 2-day and lagged 3-day $PM_{10}$ concentration as the input parameters.

# CHAPTER 5

# CONCLUSIONS AND RECOMMENDATIONS

## 5.1 Introduction

This chapter is structured into a few sections. In Section 5.2 and Section 5.3, the conclusions on the region's classification and annual variation pattern of CO, $NO_2$, $SO_2$, $O_3$ and $PM_{10}$ by the area were highlighted, respectively, while in Section 5.4 and Section 5.5, the conclusions on the monthly $PM_{10}$ concentration pattern due to the monsoon season and $PM_{10}$ prediction models by using ANN and MLR were highlighted, respectively. Overall conclusions were briefly mentioned in Section 5.6. Lastly, in Section 5.7 the recommendations regarding this study were proposed with the aim to enhance the effectiveness of the $PM_{10}$ prediction model to be implemented in Malaysia.

## 5.2 Region's classification results

In this study, multivariate statistics and environmental modelling techniques, namely AHC, DA, PCA as well as MLR, ANN, were used to evaluate the variation in atmospheric air quality as well as to develop a $PM_{10}$ prediction model throughout continuous air quality monitoring stations in Malaysia. Based on the degree of homogeneity, AHC grouped the stations into three clusters. From the analysis, 18 out of 52 air quality monitoring stations with average $PM_{10}$ value of 58.78 $\mu g/m^3$ were classified into the HPR, namely CA0001, CA0057, CA0019, CA0015, CA0002, CA0011, CA0006, CA0010, CA0044, CA0043, CA0056, CA0047, CA0053, CA0060, CA0054, CA0058, CA0025 and CA0016 (Pasir Gudang, Kota Tinggi, Larkin, Balok Baru, Kemaman, Klang, SMK Bukit Rambai, Nilai, Muar, SM Tinggi, Port Dickson,

Seremban, Putrajaya, Banting, Cheras, Batu Muda, Shah Alam and Petaling Jaya, respectively). Meanwhile, 15 out of 52 air quality monitoring stations with average $PM_{10}$ value of 50.50 $\mu g/m^3$ were classified into the MPR, namely CA0048, CA0041, CA0009, CA0017, CA0020, CA0008, CA0046, CA0022, CA0059, CA0034, CA0032, CA0033, CA0040, CA0038 and CA0003 (Kuala Selangor, Manjung, SK Seberang Jaya, Sg Petani, Taiping, SM Jln Tasek, SM Pagoh, Kota Bharu, Tanah Merah, Kuala Terengganu, Langkawi, ILP Kangar, Alor Setar, USM and SK Cenderawasih Prai, respectively). Besides that, 18 out of 51 air quality monitoring stations with average $PM_{10}$ value of 41.03 $\mu g/m^3$ were classified into LPR, namely CA0036, CA0004, CA0035, CA0027, CA0029, CA0026, CA0039, CA0042, CA0050, CA0030, CA0049, CA0031, CA0055, CA0028, CA0024, CA0045, CA0007 and CA0014 (Sri Aman, Kuching, Kota Samarahan, Bintulu, Sarikei, Sibu, Tawau, Labuan, Sandakan, Kota Kinabalu, Keningau, Limbang, Kapit, Dato' Permaisuri, Kerteh, UPSI, Jerantut and Indera Mahkota, respectively).

By figuring out the classification of regions (HPR, MPR and LPR) between stations, only the selected stations from each region were taken into account for the upcoming study in order to implement the $PM_{10}$ concentration prediction, which indirectly reduced the time and cost by removing unnecessary sampling sites for each of the stations in each cluster, since it represented the whole network for cluster classification. In this study, regions which were well discriminated with the percentage of correct classification were 85.19%, 78.35% and 89.20% for the regions of HPR, MPR and LPR, respectively. The average percentage of overall correct classification was 84.25%. Besides, the Kruskal-Wallis tests with $p < 0.05$ demonstrated that the three clusters were divergent from one another. This implied that all the stations were spatially discriminated within the study area.

## 5.3 Annual variation pattern of CO, $NO_2$, $SO_2$, $O_3$ and $PM_{10}$ by the area results

The presence of five main pollutants, namely CO, $NO_2$, $SO_2$, $O_3$ and $PM_{10}$ by area based on yearly data concluded that there was difference in pattern recorded for each pollutant by the area along 2010 to 2015, which was perhaps due to the internal and external factors as well as spatiotemporal shifts in the different sources of air pollution. Through the homogeneity results obtained by AHC, HPR consists of the highest number of the CAQMS located in the industrial and urban areas and lowest number of stations located in the suburban area with none of the stations located at the rural area. Conversely, LPR consisted of the lowest number of CAQMS located in the industrial and urban areas. However, the LPR showed the highest number of stations located in suburban and rural areas. By comparing all regions, MPR showed a moderate number of CAQMS located in the industrial, urban and suburban areas with none of the stations located in the rural area.

Overall, a high number of industrial and urban areas present in HPR gave a high concentration of pollutants in the region as compared to other regions. Besides, urbanisation process and industrial activities contributed to a high level of pollutants present in the atmosphere. The atmosphere in the urban ambient air was more polluted as compared to others due to the high population rate and human activities, such as transportation. This concludes that the location of CAQMS based on the type of area indirectly contributed to the concentration of pollutant present. In addition, as compared to the pollutants at each region, $PM_{10}$ was the major pollutant along 2010 to 2015. This showed that $PM_{10}$ was the major pollutant and its presence gave a huge impact towards human health, the environment along with the economy of Malaysia as compared to other pollutants.

## 5.4 Monthly $PM_{10}$ concentration pattern due to monsoon season results

Meteorological factor, especially wind directions also influence the concentration of most pollutants. From this study, it was observed that in September and November, $PM_{10}$ presence was at the highest and lowest concentrations, respectively. This is due to the Southwest and Northeast wind direction during the Southwest Monsoon season and the Northeast Monsoon season, which usually occur from May to September and November to March, respectively. Dry weather condition in addition to air pollutant transboundary transport from the neighbouring country biomass burning caused the presence of high $PM_{10}$ concentration during the Southwest Monsoon season. Abrupt trend of $PM_{10}$ concentration was observed for certain months from 2010 until 2015 due to the particle stability during haze occurrences.

## 5.5 $PM_{10}$ prediction models using ANN and MLR

PCA was used to determine the main pollutants that contributed to the occurrence of air pollution in the HPR, MPR and LPR. For HPR, the main pollutants that contributed to the occurrence of air pollution were $NO_x$, NO, $NO_2$, temperature, $O_3$ and $PM_{10}$, while for MPR the main pollutants that contributed to the occurrence of air pollution were $NO_x$, NO, $NO_2$, temperature, $O_3$ and $PM_{10}$. For LPR, the main pollutants that contributed to the occurrence of air pollution were $NO_x$, NO, $NO_2$, temperature, $O_3$ and $PM_{10}$. The main pollutants obtained from PCA for each of the regions were selected as the input parameters for the $PM_{10}$ prediction model.

Input parameters play an important role to develop an accurate prediction model. It can be concluded that the $PM_{10}$ prediction models, which used a lagged concentration of $PM_{10}$ data as one of the input parameters performed better prediction than excluded lagged $PM_{10}$ as an input parameter. The values of $R^2$ and RMSE obtained for HPR,

MPR and LPR through ANN, including lagged $PM_{10}$ were 0.7718, 16.11; 0.7590, 12.63; and 0.7721, 9.86, respectively. Meanwhile, the values of $R^2$ and RMSE obtained for HPR, MPR and LPR through ANN, without using lagged $PM_{10}$, were 0.4736, 24.14; 0.1851, 23.63; and 0.0760, 19.66, respectively. Meanwhile for MLR, the values of $R^2$ and RMSE obtained for HPR, MPR and LPR, including lagged $PM_{10}$, were 0.7609, 16.27; 0.7442, 13.03; and 0.7642, 10.02, respectively. The values of $R^2$ and RMSE obtained for HPR, MPR and LPR through MLR, without using lagged $PM_{10}$, were 0.3407, 27.02; 0.1332, 23.98; and 0.0569, 20.05, respectively. From these comparisons, ANN with lagged $PM_{10}$ as one of the input parameters was shown to be the best.

To determine the best number of days for $PM_{10}$ lagged data, 2-day and 3-day lagged $PM_{10}$ data as input parameters were also applied for both prediction models. The results showed that 1-day lagged concentration of $PM_{10}$ showed a better prediction as compared to 2-day and 3-day lagged $PM_{10}$ data as input parameters for both prediction models. However, the ANN model once again showed higher accuracy than MLR. It can be concluded that the $PM_{10}$ prediction model of ANN gave better accuracy as compared to MLR with a 1-day lagged concentration of $PM_{10}$ as one of the input parameters with the best performance of the ANN model by using 7, 6 and 6 hidden nodes in the hidden layer for HPR, MPR and LPR, respectively. The values of $R^2$, RMSE, IA and E obtained for HPR, MPR and LPR through ANN were 0.7718, 16.11, 0.9307, 0.7663; 0.7590, 12.63, 0.9230, 0.7510; and 0.7721, 9.80, 0.9302, 0.7657, respectively. While, the values of $R^2$, RMSE, IA and E obtained for HPR, MPR and LPR through MLR were 0.7609, 16.27, 0.9277, 0.7609; 0.7442, 13.03, 0.9217, 0.7442; and 0.7642, 10.02, 0.9292, 0.7642, respectively.

This model requires improvement by adding a 1-day lagged concentration of $PM_{10}$ in order to provide a better result of the $PM_{10}$ prediction. This concluded that 1-day lagged concentration of $PM_{10}$ plays an important role with the aim to obtain accurate $PM_{10}$ prediction model since particles can stay longer and indirectly contribute to the concentration of particles present a day later.

## 5.6 Overall conclusions

In brief, this study concluded that all CAQMS discriminated spatially into HPR, MPR and LPR by using AHC. It was noticed that each main pollutant, namely, CO, $NO_2$, $SO_2$, $O_3$ and $PM_{10}$ showed a different annual pattern along 2010 to 2015 by area due to the internal and external factors as well as spatiotemporal shifts in the different sources of air pollution with $PM_{10}$ pollutant as the major pollutant among others. Therefore, its presence might give a huge impact on human health, environment as well as to the economy. Meteorological parameters, especially wind direction could also give impact to the presence of $PM_{10}$ pollutant. Southwest and northeast wind direction during the Southwest Monsoon season and Northeast Monsoon season, respectively, will affect the concentration of $PM_{10}$ present in the atmosphere. Dry weather condition during the Southwest Monsoon season as well as the transboundary air pollutant from the neighbouring country's biomass burning caused a high concentration of $PM_{10}$ presence during the Southwest Monsoon season. This also indirectly caused an abrupt trend in $PM_{10}$ concentration for certain months due to the particles stability and trapping, especially during haze occurrences. For the $PM_{10}$ prediction model, adding 1-day lagged concentration of $PM_{10}$ as one of the input parameters through ANN by using the PCA gave an accurate result as compared to without adding it. In general, ANN gave a better prediction model as compared to the MLR and can provide an accurate prediction of $PM_{10}$.

## 5.7 Recommendations

The developed $PM_{10}$ model prediction was based on limited air quality history. However, the ANN prediction model can be enhanced by incorporating the following aspects:

1) Update the current ANN model routinely owing to the expansion of AQMS located in Malaysia (Starting from April 2017, there are 64 CAQMS with one of the previous CAQMS located in Muar been removed). Details on the new CAQMS can be referred in Appendix A.

2) The periodic maintenance of monitoring stations is compulsory in order to get more consistent data with less percentage of missing data as well as to obtain more accurate prediction model.

3) This study can be extended by taking into account the risk assessment of human health as a result exposure of air pollutants that affect human health.

4) Besides input parameters from the pollutant ($PM_{10}$, $NO_2$, CO, $O_3$, NO, $NO_x$, THC, $CH_4$, NMHC, $SO_2$) and meteorological parameters (wind direction, wind speed, temperature, humidity, UVB), other parameters, such as emission data can also be used as another input parameters since Malaysia is experiencing great development and urbanisation process.

## REFERENCES


Ababneh, M. F., AL-Manaseer, O., & Hjouj Btoush, M. (2014). $PM_{10}$ Forecasting Using Soft Computing Techniques. *Research Journal of Applied Sciences, Engineering and Technology*, 7(16): 3253–3265.

Abderrahim, H., Chellali, M. R., & Hamou, A. (2016). Forecasting $PM_{10}$ in Algiers: efficacy of multilayer perceptron networks. *Environmental Science and Pollution Research*, 23(2): 1634-1641.

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

Abdullah, N. A., Shuhaimi, S. H., Ying, T. Y., & Shapee, A. H. (2011). The study of seasonal variation of $PM_{10}$ concentration in Peninsular, Sabah and Sarawak. *Malaysian Meteorological Department*, 9: 1–28.

Abdullah, S., Ismail, M., & Fong, S. Y. ($2017_a$). Multiple Linear Regression (MLR) models for long term $PM_{10}$ concentration forecasting during different monsoon seasons. *Journal of Sustainability Science and Management*, 12(1): 60–69.

Abdullah, S., Ismail, M., Fong, S. Y., & Ahmed, A. N. (2016). Neural Network Fitting using Levenberg-Marquardt Training Algorithm for $PM_{10}$ Concentration Forecasting in Kuala Terengganu. *Journal of Telecommunication, Electronic and Computer Engineering (JTEC),* 8(12): 27-31.

Abdullah, S., Ismail, M., & Fong, S. Y. ($2017_b$). The Relationship between Daily Maximum Temperature and Daily Maximum Ground Level Ozone Concentration. *Polish Journal of Environmetal Studies,* 26(2): 517-523

Abidin, Z. (2008). *Penyuaian taburan kebarangkalian bagi data halaju angin.* Dissertasi Sarjana Sains (Statistik). Fakulti Sains dan Teknologi.Universiti Kebangsaan Malaysia.

Afroz, R., Hassan, M. N., & Ibrahim, N. A. (2003). Review of air pollution and health impacts in Malaysia. *Environmental Research*, 92(2): 71–77.

Afzali, A., Rashid, M., Afzali, M., & Younesi, V. (2017). Prediction of air pollutants concentrations from multiple sources using AERMOD coupled with WRF prognostic model. *Journal of Cleaner Production*, 166: 1216–1225.

Afzali, A., Rashid, M., Sabariah, B., & Ramli, M. (2014). $PM_{10}$ pollution: Its prediction and meteorological influence in Pasir Gudang, Johor. *IOP Conference Series: Earth and Environmental Science*.

Aghamohammadi, N., & Isahak, M. (2018). Climate Change and Air Pollution in Malaysia. *Climate Change and Air Pollution*, pp. 241-254.

Ahamad, F., Latif, M. T., Tang, R., Juneng, L., Dominick, D., & Juahir, H. (2014). Variation of surface ozone exceedance around Klang Valley, Malaysia.

*Atmospheric Research*, 139: 116–127.

AhmadIsiyaka, H. A., & Azid, A. (2015). Air Quality Pattern Assessment in Malaysia using Multivariate Techniques. *Malaysian Journal of Analytical Science*, *19*(5), 966–978.

AhmadIsiyaka, H. A., Juahir, H., Toriman, M. E., Gasim, B. M., Azid, A., Amri, M. K., Ibrahim, A., & Usman, U. N. (2014). Spatial Assessment of Air Pollution Index Using Environmetric Modeling Techniques. *Advances in Environmental Biology*, 8(24): 244–256.

Ahmat, H., Yahaya, A. S., & Ramli, N. A. (2015). $PM_{10}$ Analysis for Three Industrialized Areas using Extreme Value. *Sains Malaysiana*, 44(2): 175–185.

Ahmed, M., Guo, X., & Zhao, X. M. (2016). Determination and analysis of trace metals and surfactant in air particulate matter during biomass burning haze episode in Malaysia. *Atmospheric Environment*, *141*, pp. 219–229.

Akabueze, C.I., Tsafe, A.I., Itodo, A.U., & Uba, A. (2012). Influence of climate and height on the levels of sulfur dioxide ($SO_2$) in Sokoto high traffic density and near atmospheric region. *World Environment*, 2: 51-55.

Akhtar, A., Masood, S., Gupta, C., & Masood, A. (2018). Prediction and analysis of pollution levels in Delhi using multilayer perceptron. *Data Engineering and Intelligent Computing* (pp. 563-572). Springer, Singapore.

Alimissis, A., Philippopoulos, K., Tzanis, C. G., & Deligiorgi, D. (2018). Spatial estimation of urban air pollution with the use of artificial neural network models. *Atmospheric Environment*, 191, 205-213.

Ali, Z. M., Ibrahim, N. A., Mengersen, K., Shitan, M., & Juahir, H. (2014). Discriminant Analysis of Water Quality Data in Langat River. *From Sources to Solution* (pp. 597-601). Springer, Singapore.

Almeida, S.M., Pio, C.A., Freitas, M.C., Reis, M.A., Trancoso, M.A. (2005). Source apportionment of fine and coarse particulate matter in a sub-urban area at the Western European Coast. *Atmospheric Environment*, 39: 3127-3138.

Amin, N. A. M., Adam, M. B., & Aris, A. Z. (2015). Bayesian Extreme for Modeling High $PM_{10}$ Concentration in Johor. *Procedia Environmental Sciences*, 30: 309–314.

Amran, M. A., Azid, A., Juahir, H., Toriman, M. E., Mustafa, A. D., Hasnam, C. N. C., Azaman, F., Kamarudin, M. K. A., Saudi, A. S. M., & Yunus, K. (2015). Spatial Analysis of The certain Air Pollutants Using Environmetric Techniques. *Jurnal Teknologi*, 75(1): 241–249.

Ancrenaz, M., Gumal, M., Marshall, A.J., Meijaard, E., Wich , S.A. & Husson, S. (2018). Pongo pygmaeus. *IUCN Red List of Threatened Species 2016.*

Arhami, M., Kamali, N., & Rajabi, M. M. (2013). Predicting hourly air pollutant levels using artificial neural networks coupled with uncertainty analysis by Monte

Carlo simulations. *Environmental Science and Pollution Research*, 20(7): 4777–4789.

Arif, N. L., Abdullah, A. M., Juahir, H., & Chng, L. K. (2013). Evaluation of Meteorological Condition During 2005 Haze Episode In Klang Valley Using Mesoscale Model MM5. *International Journal of Basic & Applied Sciences*, 13(2).

Arroyo, Á., Herrero, Á., Tricio, V., Corchado, E., & Woźniak, M. (2018). Neural models for imputation of missing ozone data in air-quality datasets. *Complexity*.

Asadollahfardi, G., Tayebi Jebeli, M., Mehdinejad, M., & Rajabipour, M. J. (2016). Short-term prediction of atmospheric concentrations of ground- level ozone in Karaj using artificial neural network. *Pollution*, 2(4): 475–488.

Asif, Z., Chen, Z., & Han, Y. (2018). Air quality modeling for effective environmental management in the mining region. *Journal of the Air & Waste Management Association*, 68(9): 1001-1014.

ASMA (2007). Standard Operating Procedure for Continuous Air Quality Monitoring. Shah Alam, Selangor Malaysia

Awang, M., Jaafar, A.B., Abdullah A.M., Ismail, M., Hassan, M. N., Abdullah, R., Johan, S., & Noor, H. (2000). Air quality in Malaysia: impacts, management issues and future challenges. *Respirology*, 5(2): 183-196.

Awang, N. R., Ramli, N. A., Yahaya, A. S., & Elbayoumi, M. ($2015_a$). Multivariate methods to predict ground level ozone during daytime, nighttime, and critical conversion time in urban areas. *Atmospheric Pollution Research*, 6(5): 726–734.

Awang, N. R., Ramli, N. A., Yahaya, A. S., & Elbayoumi, M. ($2015_b$). High nighttime ground-level ozone concentrations in Kemaman: NO and $NO_2$ concentrations attributions. *Aerosol and Air Quality Research*, 15(4): 1357–1366.

Azid, A., Amran, M. A., Samsudin, M. S., Latiffah, N., & Rani, A. (2018). Assessing Indoor Air Quality Using Chemometric Models. *Polish Journal Environmental Studies,* 27(6): 1-8.

Azid, A., Juahir, H., Ezani, E., Toriman, M. E., Endut, A., Rahman, M. N. A., Yunus, K., Kamarudin, M. K. A., Hasnam, C. N. C., Saudi, S. M., & Umar, R. ($2015_b$). Identification source of variation on regional impact of air quality pattern using chemometric. *Aerosol and Air Quality Research*, 15(4), 1545–1558.

Azid, A., Juahir, H., Latif, M. T., & Zain, S. M. (2013). Feed-Forward Artificial Neural Network Model for Air Pollutant Index Prediction in the Southern Region of Peninsular Malaysia. *Journal of Environmental Protection*, 4: 1-10.

Azid, A., Juahir, H., Toriman, M., Endut, A., Abdul Rahman, M., Amri Kamarudin, M., Latif, M., Mohd Saudi, A., Che Hasnam, C., & Yunus, K.. (2016). Selection of the Most Significant Variables of Air Pollutants Using Sensitivity

Analysis. *Journal of Testing and Evaluation*, 44(1): 376-384.

Azid, A., Juahir, H., Toriman, M. E., Endut, A., Kamarudin, M. K. A., Rahman, M. N. A., Hasnam, C. N. C., Saudi, A. S. M., & Yunus, K. ($2015_a$). Source apportionment of air pollution: A case study in Malaysia. *Jurnal Teknologi*, 72(1): 83–88.

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., Zainudin, 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 Soil Pollution*. 225(8): 2-14.

Aziz, R. M., Zabawi, A. M., Azdawiyah, A. S., & Fazlyzan, A. (2018). Effects of haze on net photosynthetic rate, stomatal conductance and yield of Malaysian rice (Oryza sativa L.) varieties. *Journal of Tropical Agriculture and Food Science*, 46(2): 157-169.

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

Bai, L., Wang, J., Ma, X., & Lu, H. (2018). Air Pollution Forecasts: An Overview. *International Journal of environmental research and public health*, 15(4): 780.

Battista, G., & Vollaro, R. D. L. (2017). Correlation between air pollution and weather data in urban areas: Assessment of the city of Rome (Italy) as spatially and temporally independent regarding pollutants. *Atmospheric Environment,* 165: 240-247.

Beer, T., Li, J., & Alverson, K. (2018). *Global change and future earth: The geoscience perspective.* Cambridge university press. First edition.

Behera, S. N., & Balasubramanian, R. (2014). Influence of biomass burning on temporal and diurnal variations of acidic gases, particulate nitrate, and sulfate in a tropical urban atmosphere. *Advances in Meteorology*.

Bertazzon, S., & Shahid, R. (2017). Schools, air pollution, and active transportation: An exploratory spatial analysis of calgary, Canada. *International Journal of Environmental Research and Public Health*, 14(8): 834.

Bing, G., Ordieres-Meré, J., & Cabrera, C. B. (2015). Prediction models for ozone in metropolitan area of Mexico City based on artificial intelligence techniques. *International Journal of Information and Decision Sciences*, 7(2): 115–139.

Bishoi, B., Prakash, A., & Jain, V. K. (2009). A Comparative Study of Air Quality Index Based on Factor Analysis and US-EPA Methods for an Urban Environment. *Aerosol and Air Quality Research*, 9(1): 1–17.

Box, G. E. P. (1976). Science and Statistics. *Journal of the American Statistical Association*. 71(356): 791-799.

Burke, S. (1999). Missing values, outliers, robust statistics and non-parametric methods. *LC*GC Europe Onlie Suplement.* 19-24.

Cabaneros, S. M. S., Calautit, J. K. S., & Hughes, B. R. (2017). Hybrid Artificial Neural Network Models for Effective Prediction and Mitigation of Urban Roadside $NO_2$ Pollution. *Energy Procedia,* 142: 3524-3530.

Campa, S. D. L. A. M., Salvador, P., Fernández-Camacho, R., Artiñano, B., Coz, E., Márquez, G., Sánchez-Rodas, D., & de la Rosa, J. (2018). Characterization of biomass burning from olive grove areas: A major source of organic aerosol in $PM_{10}$ of Southwest Europe. *Atmospheric Research*, 199: 1–13.

Catalano, M., Galatioto, F., Bell, M., Namdeo, A., & Bergantino, A. S. (2016). Improving the prediction of air pollution peak episodes generated by urban transport networks. *Environmental Science & Policy*, 60: 69-83.

Charabi, Y., Abdul-wahab, S., Al-rawas, G., & Al-wardy, M. (2018). Investigating the impact of monsoon season on the dispersion of pollutants emitted from vehicles : A case study of Salalah City , Sultanate of Oman. *Transportation Research Part D*, 59: 108–120.

Che, W., Zheng, J., Wang, S., Zhong, L., & Lau A., (2011). Assessment of motor vehicle emission control policies using model-3/ CMAQ model for the Pearl River Delta Region, China. *Atmosperic Environment*, 45: 1740–1751.

Chen, W., Tang, H., & Zhao, H. (2016). Urban air quality evaluations under two versions of the national ambient air quality standards of China. *Atmospheric Pollution Research*, 7(1): 49–57.

Chen, K. Y. (2011). Combining linear and nonlinear model in forecasting tourism demand. *Expert Syst Appl*, 38(8):10368–76.

Chaloulakou, A., Grivas, G., & Spyrellis, N. (2003). Neural network and multiple regression models for $PM_{10}$ prediction in Athens: a comparative assessment. *Journal of the Air & Waste Management Association,* 53: 1183–1190.

Clements, N., Hannigan, M. P., Miller, S. L., Peel, J. L., & Milford, J. B. (2016). Comparisons of urban and rural $PM_{10\text{-}2.5}$ and $PM_{2.5}$ mass concentrations and semi-volatile fractions in northeastern Colorado. *Atmospheric Chemistry and Physics*, 16(11): 7469–7484.

Cortina–Januchs, M. G., Quintanilla–Dominguez, J., Vega–Corona, A., & Andina, D. (2015). Development of a model for forecasting of $PM_{10}$ concentrations in Salamanca, Mexico. *Atmospheric Pollution Research*, 6(4): 626–634.

Custodio, D., Alves, C., Jomolca, Y., & Castro, P. De. (2018). Carbonaceous components and major ions in $PM_{10}$ from the Amazonian Basin. *Atmospheric Research*, 215: 75–84.

Darbre, P. D. (2018). Overview of air pollution and endocrine disorders, *International Journal of General Medicine,* 11:191–207.

Desai, A. A. (2018). A review on Assessment of Air Pollution due to Vehicular Emission in Traffic Area. *International Journal of Current Engineering and Technology*, 8(2): 356–360.

De Vries, W., Butterbach B.K., Denier V.D.G.H. & Oenema, O. (2006). The Impact of Atmospheric Nitrogen Deposition on the Exchange of Carbon Dioxide, Nitrous Oxide and Methane from European Forests. *Global Change Biology,* 12: 1151–1173.

Dias, D., Tchepel, O., Carvalho, A., Miranda, A.I., & Borrego, C. (2012). Particulate matter and health risk under a changing climate: Assessment for Portugal. *Scientific World Journal*.

DOA (2015). *Vegetables and Cash Crops Statistics Malaysia.* Department of Agriculture, Malaysia.

DOE (2000). *A guide to Air Pollutant Index (API) in Malaysia*. Department of Environment Ministry of Natural Resources and Environment Malaysia.

DOE ($2015_a$). *Malaysia Environmental Quality Report.* Department of Environment, Ministry of Natural Resources and Environment Malaysia.

DOE (2018). *Air Quality Standard.* Retrieved November 2, 2018, from http://www.doe.gov.my/portalv1/wp-content/uploads/2013/01/Air-Quality-Standard-BI.pdf

DOE ($2015_b$). *Chronology of haze episodes in Malaysia*. Retrieved March 15, 2016, fr om https://www.doe.gov.my/portalv1/wp-content/uploads/2015/09/Chronology-of-Haze-Episodes-in-Malaysia.pdf

DOE (2014). *Malaysia Environmental Quality Report.* Department of Environment, Ministry of Natural Resources and Environment Malaysia.

DOE (2010). *Malaysia Environmental Quality Report*. Department of Environment, Ministry of Sciences, Technology and the Environments, Malaysia.

DOE (2011). *Malaysia Environmental Quality Report*. Department of Environment, Ministry of Sciences, Technology and the Environments, Malaysia.

Dohoo, I. R. (2015). Dealing with deficient and missing data. *Preventive Veterinary Medicine*, 122(1–2): 221–228.

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.

DOS (2016) *Population projections, Malaysia 2010-2040*. Retrieved October 22, 2018, from https://www.dosm.gov.my

Dotse, S. Q., Dagar, L., Petra, M. I., & De Silva, L. C. (2016). Influence of Southeast Asian Haze episodes on high $PM_{10}$ concentrations across Brunei Darussalam. *Environmental Pollution*, 219: 337–352.

Dotse, S. Q., Petra, M. I., Dagar, L., & De Silva, L. C. (2018). Application of computational intelligence techniques to forecast daily $PM_{10}$ exceedances in Brunei Darussalam. *Atmospheric Pollution Research*, 9(2): 358–368.

Duncan, B. N., Martin, R. V., Staudt, A. C., Yevich, R. & Logan., J. A. (2003). Interannual and seasonal variability of biomass burning emissions constrained by satellite observation. *Journal of Geophysical Research: Atmospheres,* 108(D2).

EEA (2012). *Particulate matter from natural sources and related reporting under the EU Air Quality Directive in 2008 and 2009*. European Environment Agency, Copenhagen.

Elangasinghe, M. A., Singhal, N., Dirks, K. N., Salmond, J. A., & Samarasinghe, S. (2014). Complex time series analysis of $PM_{10}$ and $PM_{2.5}$ for a coastal site using artificial neural network modelling and k-means clustering. *Atmospheric Environment*, 94:106–116.

Erb, W. M., Barrow, E. J., Hofner, A. N., Utami-Atmoko, S. S., & Vogel, E. R. (2018). Wildfire smoke impacts activity and energetics of wild Bornean orangutans. *Scientific reports*, 8(1), 7606.

Esplin, G.J. (1995). Approximate explicit solution to the general line source problem. *Atmospheric Environment*, 29: 1459–1463.

Fakaruddin, F. J., Saleh, F. Z., Yik, D. J., Adam, M. K. M., Sang, Y. W., Chang, N. K., Yunus, F.,& Abdullah, M. H. (2015). *Weather analysis from July until October 2015*. Technical note no. 2015. Malaysian Meteorological Department, Ministry of Science, Technology and Innovation.

Famoso, F., Lanzafame, R., Monforte, P., Oliveri, C., & Scandura, P. F. (2015). Air quality data for Catania: Analysis and investigation casestudy 2012-2013. *Energy Procedia*, 81(2): 644–654.

Farsani, M. H., Shirmardi, M., Alavi, N., Maleki, H., Sorooshian, A., Babaei, A., Asgharnia, H., Marzouni, M. B., & Goudarzia, G. (2018). Evaluation of the relationship between $PM_{10}$ concentrations and heavy metals during normal and dusty days in Ahvaz, Iran. *Aeolian Research*, 33: 12-22.

Finardi, S., Agrillo, G., Baraldi, R., Calori, G., Carlucci, P., Ciccioli, P., D'Allura, A., Gasbarra, D., Gioli, B., Magliulo, V., Radice, P., Toscano, P., & Zaldei, A. (2018). Atmospheric Dynamics and Ozone Cycle during Sea Breeze in a Mediterranean Complex Urbanized Coastal Site. *Journal of Applied Meteorology and Climatology*, 57(5): 1083–1099.

Fletcher, D., & Goss, E. (1993). Forecasting with neural networks: an application using bankruptcy data. *Information & Management,* 24(3): 159-167.

Franceschi, F., Cobo, M., & Figueredo, M. (2018). Discovering relationships and forecasting $PM_{10}$ and $PM_{2.5}$ concentrations in Bogotá, Colombia, using Artificial Neural Networks, Principal Component Analysis, and k-means clustering. *Atmospheric Pollution Research*, 9(5): 912-922.

Friedman, J. H. (1991). Multivariate adaptive regression splines (with discussion). *Ann. Statis*. 19: 1–141.

Garciarena, U., & Santana, R. (2017). An extensive analysis of the interaction between missing data types, imputation methods, and supervised classifiers. *Expert Systems with Applications*, 89: 52–65.

Gardner, M.W., Dorling, S.R. (1998). Artificial neural networks (the multilayer perceptron)— a review of applications in the atmospheric sciences. *Atmospheric Environment*, 32: 2627–2636.

Gazzaz, N. M, Yusoff, M. K, Ramli, M. F, Aris, A. Z., & Juahir, H. (2012) Characterization of spatial patterns in river water quality using chemometric pattern recognition techniques. *Marine Pollution Bulletin*, 64(4): 688-698.

Ghazali, N. A., Ramli, N. A., Yahaya, A. S. (2009). A study to investigate and model the transformation of nitrogen dioxide into ozone using time series plot. *European Journal of Scientific Research*, 37(2): 192-205.

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*, 21(5).

Giussani, B., Roncoroni, S., Recchia, S., & Pozzi, A. (2016). Bidimensional and multidimensional principal component analysis in long term atmospheric monitoring. *Atmosphere*, 7(12).

Gómez-Carracedo, M. P., Andrade, J. M., López-Mahía, P., Muniategui, S., & Prada, D. (2014). A practical comparison of single and multiple imputation methods to handle complex missing data in air quality datasets. *Chemometrics and Intelligent Laboratory Systems*, 134: 23-33.

Gong, X., Hong, S., & Jaffe, D. A. (2018). Ozone in China: Spatial distribution and leading meteorological factors controlling $O_3$ in 16 Chinese cities. *Aerosol Air Quality Research*, 18: 2287-2300.

Goudie, A. S. (2018). *Human impact on the natural environment*. John Wiley & Sons.

Gorai, A. K., Tuluri, F., Tchounwou, P. B., Ambinakudige, S. (2015). Influence of local meteorology and $NO_2$ conditions on ground-level ozone concentrations in the eastern part of Texas, USA. *Air Quality, Atmosphere & Health*. 8(1): 81-96.

Guerreiro, C., Gonzalez Ortiz, A., de Leeuw, F., Viana, M., & Horalek, J. (2016). *Air quality in Europe — 2016 report*.

Gurjar, B. R., Ravindra, K., & Nagpure, A. S. (2016). Air pollution trends over Indian megacities and their local-to-global implications. *Atmospheric Environment*, 142: 475–495.

Hamid, H. A., Rahmat, M. H., & Sapani, S. A. (2018). The classification of $PM_{10}$ concentrations in Johor Based on Seasonal Monsoons. *Earth and Environmental Science*, 140(1): 12-28.

He, H. D., Lu, W. Z., & Xue, Y. (2015). Prediction of particulate matters at urban intersection by using multilayer perceptron model based on principal components. *Stochastic Environmental Research and Risk Assessment*, 29(8): 2107-2114.

Hidalgo, M. J., Pozzi, M. T., Furlong, O. J., Marchesky, E. J., & Pellerano, R. G. (2018). Classification of organic olives based on chemometric analysis of elemental data. *Microchemical Journal,* 142:30-35.

Honaker, J., King, G., & Blackwell, M. (2011). Amelia II : A Program for Missing Data. *Journal of Statistical Software*, 45(7): 1–54.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feed-forward networks are universal approximators. *Neural Networks*, 2: 359–366.

Hosseiniebalam, F., & Ghaffarpasand, O. (2015). The effects of emission sources and meteorological factors on sulphur dioxide concentration of Great Isfahan, Iran. *Atmospheric Environment*, 100: 94–101.

Hua, A. K. (2018). Applied chemometric approach in identification sources of air quality pattern in Selangor, Malaysia. *Sains Malaysiana*, 47(3): 471–479.

Hu, Y., Fernandez-Anez, N., Smith, T. E., & Rein, G. (2018). Review of emissions from smouldering peat fires and their contribution to regional haze episodes. *International Journal of Wildland Fire*, 27(5): 293-312.

Iqbal, M. A., Kim, K. H., Shon, Z. H., Sohn, J. R., Jeon, E. C., Kim, Y. S., & Oh, J. M. (2014). Comparison of ozone pollution levels at various sites in Seoul, a megacity in Northeast Asia. *Atmospheric Research*, 138: 330-345.

Ismail, A. S., Abdullah, A. M., & Samah, M. A. A. (2017). Environmetric study on air quality pattern for assessment in Northern region of Peninsular Malaysia. *Journal of Environmental Science and Technology,* 10: 186-196.

Ismail, M., Yuen, F. S., & Abdullah, S. (2015). Trend and status of particulate matter ($PM_{10}$) concentration at three major cities in east coast of Peninsular Malaysia. *Research Journal of Chemical and Environmental Sciences*, 3(5):25-31.

Jaafar, S. A., Latif, M. T., Razak, I. S., Wahid, N. B. A., Khan, M. F., & Srithawirat, T. (2018). Composition of carbohydrates, surfactants, major elements and anions in $PM_{2.5}$during the 2013 Southeast Asia high pollution episode in Malaysia. *Particuology*, 37: 119–126.

Jayamurugan, R., Kumaravel, B., Palanivelraja, S., & Chockalingam, M. P. (2013). Influence of temperature, relative humidity and seasonal variability on ambient air quality in a coastal urban area. *International Journal of Atmospheric Sciences.*

Jeong, J. I., & Park, R. J. (2013). Effects of the meteorological variability on regional air quality in East Asia. *Atmospheric Environment*, 69: 46-55.

Jin, L., Luo, X., Fu, P., & Li, X. (2016). Airborne particulate matter pollution in urban China: A chemical mixture perspective from sources to impacts. *National*

*Science Review*, 4(4): 593-610.

Jordanov, I., Petrov, N., & Petrozziello, A. (2018). Classifiers Accuracy Improvement Based on Missing Data Imputation. *Journal of Artificial Intelligence and Soft Computing Research*, 8(1): 31–48.

Juahir, H., Zain, S. M., Yusoff, M. K., Hanidza, T. I. T., Armi, A. S. M., Toriman, M. E. & Mokhtar, M. (2011). Spatial Water Quality Assessment of Langat River Basin (Malaysia) Using Chemometric Techniques. *Environmental Monitoring and Assess*ment, 173: 625–641.

Junger, W. L., & De Leon, A. P. (2015). Imputation of missing data in time series for air pollutants. *Atmospheric Environment*, 102: 96–104.

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

Kanada, M., Dong, L., Fujita, T., Fujii, M., Inoue, T., Hirano, Y., Togawa, T., & Geng, Y. (2013). Regional disparity and cost-effective $SO_2$ pollution control in China: A case study in 5 mega-cities. *Energy Policy*, 61: 1322–1331.

Kang, G. K., Gao, J. Z., Chiao, S., Lu, S., & Xie, G. (2018). Air quality prediction: Big data and machine learning approaches. *International Journal of Environmental Science and Development*, 9(1): 8-16.

Kasparoglu, S., Incecik, S., & Topcu, S. (2018). Spatial and temporal variation of $O_3$, NO and $NO_2$ concentrations at rural and urban sites in Marmara Region of Turkey. *Atmospheric Pollution Research*, 9(6): 1009-1020.

Kelly, F. J., & Fussell, J. C. (2015). Air pollution and public health: emerging hazards and improved understanding of risk. *Environmental Geochemistry and Health*, 37(4): 631–649.

Kim, J. O., & Mueller, C. W. (1987). *Introduction to factor analysis: what it is and how to do it. Quantitative applications in the social science series*. Newbury Park: Sage University Press.

Kim, Y. P., & Lee, G. (2018). Trend of Air Quality in Seoul: Policy and Science. *Aerosol and Air Quality Research*, 18: 2141–2156.

Koplitz, S. N., Mickley, L. J., Marlier, M. E., Buonocore, J. J., Kim, P. S., Liu, T., Sulprizio, M. P., De Fries, R. S., Jacob, D. J., Schwartz, J., Pongsiri, M., & Myers, S. S. (2016). Public health impacts of the severe haze in Equatorial Asia in September–October 2015: demonstration of a new framework for informing fire management strategies to reduce downwind smoke exposure. *Environmental Research Letters.*

Kumar, A., & Goyal, P. (2011). Forecasting of daily air quality index in Delhi. *Science of the Total Environment*, 409(24): 5517-5523.

Kuwata, M., Neelam-Naganathan, G.-G., Miyakawa, T., Khan, M. F., Kozan, O., Kawasaki, M., Sumin, S.,& Latif, M. T. (2018). Constraining the Emission of

Particulate Matter from Indonesian Peatland Burning Using Continuous Observation Data. *Journal of Geophysical Research: Atmospheres*, 123(17): 9828-9842.

Latif, M. T., Dominick, D., Ahamad, F., Khan, M. F., Juneng, L., Hamzah, F. M., & Nadzir, M. S. M. (2014). Long term assessment of air quality from a background station on the Malaysian Peninsula. *Science of the Total Environment*, 482–483(1): 336–348.

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.

Leh, O. L. H., Ahmad, S., Aiyub, K., Jani, Y. M., & Hwa, T. K. (2012). Urban air environmental health indicators for Kuala Lumpur city. *Sains Malaysiana*, 41(2): 179-191.

Li, G., & Weng, Q. (2016). Measuring the quality of life in city of Indianapolis by integration of remote sensing and census data, *International Journal of Remote Sensing*, 28(2): 249-267.

Li, J., Chen, B., de la Campa, A. M. S., Alastuey, A., Querol, X., & Jesus, D. (2018). 2005–2014 trends of $PM_{10}$ source contributions in an industrialized area of southern Spain. *Environmental Pollution*, 236: 570-579.

Ling, O. H. L., Ting, K. H., Shaharuddin, A., Kadaruddin, A., & Yaakob, M. J. (2010). Urban growth and air quality in Kuala Lumpur city, Malaysia. *Environment Asia*, 3(2): 123-8.

Liu, Y., & Gopalakrishnan, V. (2017). An Overview and Evaluation of Recent Machine Learning Imputation Methods Using Cardiac Imaging Data. *Data*, 2(1): 8.

Luvsan, M. E., Shie, R. H., Purevdori, T., Badarch, L., Baldorj, B., & Chan, C. C. (2012). The influence of emission sources and meteorological conditions on $SO_2$ pollution in Mongolia. *Atmospheric Environment*, 61: 542-549.

Mahapatra, P. S., Sinha, P. R., Boopathy, R., Das, T., Mohanty, S., Sahu, S. C., & Gurjar, B. R. (2018). Seasonal progression of atmospheric particulate matter over an urban coastal region in peninsular India: Role of local meteorology and long-range transport. *Atmospheric Research*, 199: 145–158.

Mahiyuddin, W. R. W., Sahani, M., Aripin, R., Latif, M. T., Thach, T. Q., & Wong, C. M. (2013). Short-term effects of daily air pollution on mortality. *Atmospheric environment*, 65: 69-79.

MBPJ (2011) *Laporan tahunan 2011*. Majlis Bandaraya Petaling Jaya.

Manan, N. A., Abdul Manaf, M. R., & Hod R. (2018). The Malaysia Haze and its health economy impact: A literature review. *Malaysian Journal of Public Health Medicine*, 18(1): 38-45.

Masiol, M., Squizzato, S., Formenton, G., Harrison, R. M., & Agostinelli, C. (2017).

Air quality across a European hotspot: Spatial gradients, seasonality, diurnal cycles and trends in the Veneto region, NE Italy. *Science of the Total Environment*, 576: 210–224.

Masseran, N., & Razali, A. M. (2016). Modeling the wind direction behaviors during the monsoon seasons in Peninsular Malaysia. *Renewable and Sustainable Energy Reviews,* 56: 1419-1430.

Mishra, D., & Goyal, P. (2015). Development of artificial intelligence based $NO_2$ forecasting models at Taj Mahal, Agra. *Atmospheric Pollution Research*, 6(1): 99–106.

Mittal, M. L., Hess, P. G., Jain, S. L., Arya B. C and Sharma, C. (2007). Surface ozone in the Indian region, *Atmospheric Environment*, 41: 6572–6584.

Mladenović, M. R., Dakić, D. V, Nemoda, S. Đ., Paprika, M. J., & Komatina, M. S. (2016). The combustion of biomass – The impact of its types and combustion technologies on the emission of nitrogen oxide. *Hemijska industrija,* 70(3): 287–298.

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

Mohamad, M., Kamarudin, M. K. A., Juahir, H., Mat Ali, N. A., Karim, F., Badarilah, N., Muhammad, N., & Mohd Ridzuan, M. S. (2018). Development of spatial distribution model using GIS to identify Social Support Index among drug-abuse inmates. *International Journal of Engineering and Technology (UAE)*, 7(2): 0–7.

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

Mohammed, N. I., Ramli, N. A., Yahya, A. S. (2013). Ozone phytotoxicity evaluation and prediction of crops production in tropical regions. *Atmospheric Environment,* 68: 343–349.

Mohtar, A., Asma, A., Latif, M. T., Baharudin, N. H., Ahamad, F., Chung, J. X., Othman, M., & Juneng, L. (2018). Variation of major air pollutants in different seasonal conditions in an urban environment in Malaysia. *Geoscience Letters,* 5(1): 21

Moustris, K. P., Larissi, I. K., Nastos, P. T., Koukouletsos, K. V. & Paliatsos, A. G. (2013). Development and Application of Artificial Neural Network Modeling in Forecasting $PM_{10}$ Levels in a Mediterranean City. *Water, Air, & Soil Pollution*, 224(8): 1-11.

Muhamad, M., Ul-Saufie, A. Z., & Deni, S. M. (2015). Three Days Ahead Prediction of Daily 12 Hour Ozone ($O_3$) concentrations for Urban Area in Malaysia. *Journal of Environmental Science and Technology*, 8(3): 102.

Mukhopadhyay, K., & Forssell, O. (2005). An Empirical Investigation of Air Pollution from Fossil Fuel Combustion and Its Impact on Health in India during 1973–1974 to 1996–1997. *Ecological Economics*, 55(2): 235–250.

Munir, S., Habeebullah, T. M., Seroji, A. R., Morsy, E. A., Mohammed, A. M. F., Saud, W. A., Abdou, A. E. A., & Awad, A. H. (2013). Modeling particulate matter concentrations in Makkah, applying a statistical modeling approach. *Aerosol and Air Quality Research*, 13(3): 901–910.

Mutalib, S. A., Juahir, H., Azid, A., Mohd Sharif, S., Latif, M. T., Aris, A. Z., Zain, S. M., & Dominick, D. (2013). Spatial and temporal air quality pattern recognition using environmetric techniques: a case study in Malaysia. *Environmental Science: Processes & Impacts*, 15(9): 1717.

Nash, J. E., & Sutcliffe, J. V. (1970). River Flow Forecasting through Conceptual Models Part I-A Discussion of Principles. *Journal of hydrology*, 10(3): 282-290.

Nazari, S., Shahhosseini, O., Sohrabi-Kashani, A., Davari, S., Sahabi, H., & Rezaeian, A. (2012). $SO_2$ pollution of heavy oil-fired steam power plants in Iran. *Energy Policy*, 43: 456-465.

Nazif, A., Mohammed, N. I., Malakahmad, A., & Abualqumboz, M. S. (2018). Multivariate analysis of monsoon seasonal variation and prediction of particulate matter episode using regression and hybrid models. *International Journal of Environmental Science and Technology*, 1-14.

Ng, K. Y., & Awang, N. (2018). Multiple linear regression and regression with time series error models in forecasting $PM_{10}$ concentrations in Peninsular. *Environmental monitoring and assessment*, 190(2): 63.

Nidzgorska-Lencewicz, J. (2018). Application of Artificial Neural Networks in the Prediction of $PM_{10}$ Levels in the Winter Months: A Case Study in the Tricity Agglomeration, Poland. *Atmosphere*, 9(6): 203.

Noor, N. M., Shukri, Y. A., Azam, R. N., & Al Bakri, A. M. M. (2008). Estimation of missing values in air pollution data using single imputation techniques. *Science Asia*, 34(3): 341–345.

Noor, N. M., Al Bakri Abdullah, M. M., Yahaya, A. S., & Ramli, N. A. (2014). Comparison of Linear Interpolation Method and Mean Method to Replace the Missing Values in Environmental Data Set. *Materials Science Forum*, 803: 278–281.

Noor, N. M., Yahaya, A. S., Ramli, N. A., Luca, F. A., Abdullah, M. M. A. B., & Sandu, A. V. (2015). Variation of Air Pollutant (Particulate Matter-PM10) in Peninsular Malaysia Study in the southwest coast of peninsular Malaysia. *Revista de Chimie*, 66(9): 1443-1447.

Noor, N. M., & Zainudin, M. L. (2008). A review: Missing values in environmental data sets. *In Proceeding of International Conference on Environment.*

Oliveira, P. L. D., Figueiredo, B. R. D., & Cardoso, A. A. (2011). Atmospheric pollutants in São Paulo state, Brazil and effects on human health – A review. *Geochimica Brasiliensis, Ouro Preto*, 25(1): 17 - 24.

Özdemir, U., & Taner, S. (2014). Impacts of meteorological factors on $PM_{10}$: Artificial neural networks (ANN) and multiple linear regression (MLR) approaches. *Environmental Forensics*, 15(4): 329-336.

Pai, M. L. (2016). *ANN based Data Mining Technique to Achieve Improved Accuracy to Predict ISMR from Ocean–Atmosphere State Variables* (Doctoral dissertation). Cochin University of Science and Technology.

Pai, P. F, Lin, K. P, Lin, C. S, Chang, P. T. (2010). Time series forecasting by a seasonal support vector regression model. *Expert Syst Appl*, 37(6):4261–5.

Pawul. M., & Śliwka, M. (2016). Application of Artificial Neural Networks for Prediction of Air Pollution Levels in Environmental Monitoring. *Journal of Ecological Engineering*. 17(4): 190–196.

Peng, H. (2015). *Air Quality Prediction by Machine Learning Methods* (Master dissertation). The University of British Columbia.

Plaia, A., & Bondì, A. L. (2006). Imputation of missing values in air quality data sets. *Riunione Scientifica Della Società Italiana Di Statistica*, 667-670.

Rahman, N. H. A. M., Lee, H. M., & Latif, T. M. (2016). Evaluation Performance of Time Series Approach for Forecasting Air Pollution Index in Johor, Malaysia. *Sains Malaysiana*, 45(11): 1625–1633.

Rahman, S. R. A., Ismail, S. N. S., Ramli, M. F., Latif, M. T., Abidin, E. Z., & Praveena, S. M. (2015). The Assessment of Ambient Air Pollution Trend in Klang Valley, Malaysia. *World Environment*, 5(1): 1–11.

Ramli, M. N., Yahaya, A. S., Ramli, N. A., Yusof, N. F. F. M., & Abdullah, M. M. A. (2013). Roles of imputation methods for filling the missing values: A review. *Advances in Environmental Biology*, 7: 3861–3869.

Razak, N. A., Zubairi, Y. Z., & Yunus, R. M. (2014). Imputing missing values in modelling the PM concentrations. *Sains Malaysiana*, 43(10): 1599–1607.

Ray, S., & Kim, K. H. (2014). The pollution status of sulfur dioxide in major urban areas of Korea between 1989 and 2010. *Atmospheric research*, 147: 101-110.

Rein, G., Cleaver, N., Ashton, C., Pironi, P., & Torero, J. L. (2008). The severity of smouldering peat fires and damage to the forest soil. *Catena*, 74(3): 304-309.

Retnam, A., Pauzi, M., Juahir, H., Zaharin, A., Abdul, M., & Fadhil, M. (2013). Chemometric techniques in distribution, characterisation and source apportionment of polycyclic aromatic hydrocarbons (PAHS) in aquaculture sediments in Malaysia. *Marine Pollution Bulletin*, 69(1–2): 55–66.

Rößler, M., Koch, T., Janzer, C., & Olzmann, M. (2017). Mechanisms of the $NO_2$ Formation in Diesel Engines. *MTZ Worldwide,* 78(7–8): 70–75.

Rosofsky, A., Levy, J. I., Zanobetti, A., Janulewicz, P., & Fabian, M. P. (2018). Temporal trends in air pollution exposure inequality in Massachusetts. *Environmental Research*, 161: 76–86.

Sansuddin, N., Ramli, N. A., Yahaya, A. S., Yusof, N. F. F. M., Ghazali, N. A., & Madhoun, W. A. Al. (2011). Statistical analysis of $PM_{10}$ concentrations at different locations in Malaysia. *Environmental Monitoring and Assessment*, 180(1–4): 573–588.

Saptoro, A. (2010). *An integrated approach to artificial neural network based process modelling* (Doctoral dissertation) . Curtin University.

Sari, D., İncecik, S., & Ozkurt, N. (2016). Surface ozone levels in the forest and vegetation areas of the Biga Peninsula, Turkey. *Science of the Total Environment*, 571(2): 1284–1297.

Sarwat, E., & El-Shanshoury, G. I. (2018). Estimation of Air Quality Index by Merging Neural Network with Principal Component Analysis. *International Journal of Computer Application,* 8(1): 2250-1797.

Sayegh, A. S., Munir, S., & Habeebullah, T. M. (2014). Comparing the performance of statistical models for predicting $PM_{10}$ concentrations. *Aerosol and Air Quality Research*, 14(3): 653–665.

Shahadin, M. S., Ab Mutalib, N. S., Latif, M. T., Catherine, G. M., & Tidi, H. (2018). Challenges and future direction of molecular research in air pollution-related lung cancers. *Lung Cancer,* 118: 69-75.

Show, D. L., & Chang, S.-C. (2016). Atmospheric impacts of Indonesian fire emissions: Assessing Remote Sensing Data and Air Quality During 2013 Malaysian Haze. *Procedia Environmental Sciences*, 36: 176–179.

Shrestha S., & Kazama F. (2007). Assessment of surface water quality using multivariate statistical techniques: A case study of the Fuji river basin, Japan. *Environmental Modelling & Software,* 22: 464–75.

Shukla, S.P., & Sharma, M. (2008). Source apportionment of atmospheric $PM_{10}$ in Kanpur, India. *Environmental Engineering Science*, 25(6): 849-862.

Simeonov, V., Einax, J. W., Stanimirova, I., & Kraft, J. (2002). Environmetric modeling and interpretation of river water monitoring data. Analytical and bioanalytical chemistry, 374(5): 898-905.

Soleiman, A., Othman, M., Samah, A. A., Sulaiman, N. M., & Radojevic, M. (2003). The Occurrence of Haze in Malaysia: A Case Study in an Urban Industrial Area. *Pure and Applied Geophysics*, 160(1–2): 221–238.

Soni, A., & Shukla, S. (2012). Application of neuro-fuzzy in prediction of air pollution in urban areas. *IOSR Journal of Engineering*, 2(5), 1182-1187.

Stockwell, C. E., Jayarathne, T., Cochrane, M. A., Ryan, K. C., Putra, E. I., Saharjo, B. H., Nurhayati, A. D., Albar, I., Blake, D. R., Simpson, I.J., Stone, E. A., &

Yokelson, R. J. (2016). Field measurements of trace gases and aerosols emitted by peat fires in Central Kalimantan, Indonesia, during the 2015 El Niño. *Atmospheric Chemistry and Physics*, 16(18): 11711-11732.

Wen, S. Y., Fauzan bin Mohd Nor, A., Bt. Fazilan, N. N., & Bt. Sulaiman, Z. (2016). Transboundary Air Pollution in Malaysia: Impact and Perspective on Haze. *Nova Journal of Engineering and Applied Sciences*, 5(1): 1–11.

Syafei, A. D. (2014). Analyzing and Interpreting Air Quality Monitoring Data in Surabaya Analyzing and Interpreting Air Quality Monitoring Data in Surabaya, (September).

Syafei, A. D. (2014). *Analyzing and Interpreting Air Quality Monitoring Data in Surabaya* (Doctoral dissertation). Hiroshima University.

Taspinar, F., & Bozkurt, Z. (2014). Application of Artificial Neural Networks and Regression Models in the Prediction of Daily Maximum $PM_{10}$ Concentration in Duzce, Tukey. *Fresenius Environmental Bulletin*, 23(10): 2450–2459.

Toh, Y.Y., Fook, L.S., & Von Glasow, R. (2013). The Influence of Meteorological Factors and Biomass Burning on Surface Ozone Concentrations at Tanah Rata, Malaysia. *Atmospheric Environment,* 70: 435–446.

Trinh, T. T., Trinh, T. T., Le, T. T., & Tu, B. M. (2018). Temperature inversion and air pollution relationship, and its effects on human health in Hanoi City, Vietnam. *Environmental geochemistry and health*, 1-9.

Ul-Saufie, A. Z., Yahaya, A. S., Ramli, N. A., Rosaida, N., & Hamid, H. A. (2013). Future daily $PM_{10}$ concentrations prediction by combining regression models and feedforward backpropagation models with principle component analysis (PCA). A*tmospheric Environment,* 77: 621-630.

USEPA (2011). *Six Common Air Pollutants, Sulphur Dioxide*. Retrieved October 5, 2017, from http://www.epa.gov/oaqps001/sulphur dioxide/health.html

USEPA (2015). *Criteria Air Pollutants.* Retrieved March 10, 2017, from https://www.epa.gov

Wahab, N. A., Kamarudin, M. K. A., & Rahim, K. A. (2016). Prediction of Damage Cost of Bronchitis Due to Haze in Malaysia. *Malaysian Journal of Applied Sciences,* 1(2): 1-8.

Wang, J., & Ogawa, S. (2015). Effects of Meteorological Conditions on $PM_{2.5}$ Concentrations in Nagasaki, Japan. *International Journal of Environmental Research and Public Health*, 12(8): 9089-9101.

WHO (2006). *Air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide.* Retrieved May 12, 2015, from http://apps.who.int/iris/bitstream/handle/10665/69477/WHO_SDE_PHE_OEH_06.02_eng.pdf;jsessionid=09F909A9D6238C8F0BE1A91DA0C76AB4?sequence=1

WHO (2014). *7 million premature deaths annually linked to air pollution*. Retrieved September 27, 2018 from https://www.who.int/mediacentre/news/releases/2014/air-pollution/en/

WHO (2018). *Ambient (outdoor) air quality and health.* Retrieved September 27, 2018, from http://www.who.int/en/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health

Widaman, K. (2006). Missing data: What to do with or without them. *Monographs of the Society for Research in Child Development*, 71(3): 42–64.

Wolff, H. (2014). Keep your clunker in the suburb: low-emission zones and adoption of green vehicles. *The Economic Journal,* 124: 481-512.

Wongsathan, R., & Seedadan, I. (2016). A hybrid ARIMA and Neural Networks model for $PM_{10}$ pollution estimation : The case of Chiang Mai city moat area. *Procedia Computer Science*, 86: 273–276.

Yaacob, K. K. K., Ali, A., & Isa, M. M. (2007). *Keadaan laut perairan Semenanjung Malaysia untuk panduan nelayan.* Departemen Penyelidikan dan Pengurusan Sumber Perikanan Marin Jabatan Perikanan Malaysia.

Yang, X., Zhang, Z., Zhang, Z., Sun, L., Xu, C., & Yu, L. (2016). A long-term prediction model of Beijing haze episodes using time series analysis. *Computational intelligence and neuroscience*.

Yap, X. Q., & Hashim, M. (2013). A robust calibration approach for $PM_{10}$ prediction from MODIS aerosol optical depth. *Atmospheric Chemistry and Physics*, 13(6): 3517–3526.

Yim, O., & Ramdeen, K. T. (2015). Hierarchical Cluster Analysis: Comparison of Three Linkage Measures and Application to Psychological Data. *The Quantitative Methods for Psychology*, 11(1): 8–21.

Yong, A. G., & Pearce, S. (2013). A Beginner's Guide to Factor Analysis: Focusing on Exploratory Factor Analysis. *Tutorials in Quantitative Methods for Psychology*, 9(2): 79-94.

Yong, N. K., & Awang, N. (2017). Quantile regression for analysing $PM_{10}$ concentrations in Petaling Jaya, *Malaysian Journal of Fundamental and Applied Sciences,* 13(2): 86–90.

Yuce, B., Mastrocinque, E., Packianather, M. S., Pham, D., Lambiase, A., & Fruggiero, F. (2014). Neural network design and feature selection using principal component analysis and Taguchi method for identifying wood veneer defects. P*roduction & Manufacturing Research,* 2(1): 291-308.

Yunus, R. M., & Hasan, M. M. (2017). *Predicting Hourly $PM_{10}$ Concentration in Seberang Perai and Petaling Jaya Using Log-Normal Linear Data.* Proceedings of IASTEM International Conference, Wellington, New Zealand

Yusof, N. F. F. M., Ramli, N. A., Yahaya, A. S., Sansuddin, N., Ghazali, N. A. & Madhoun, W. (2008). Monsoonal differences and probability distribution of

PM10 concentration. Environ Monit Assess. DOI 10.1007/s10661-009-0866-0

Zakaria, N. A., & Noor, N. M. (2018). Imputation Methods for Filling Missing Data in Urban Air Pollution Data for malaysia. *Urbanism. Arhitectura. Constructii*, 9(2): 159.

Zhang, J., & Ding, W. (2017). Prediction of air pollutants concentration based on an extreme learning machine: The case of Hong Kong. *International Journal of Environmental Research and Public Health*, 14(2): 1–19.

Zhang, Y. L., & Cao, F. (2015). Fine particulate matter ($PM_{2.5}$) in China at a city level. *Scientific Reports*, *5*: 14884.

Zhang, Z. (2016). Multiple imputation for time series data with Amelia package. *Annals of Translational Medicine*, 4(3): 56.

Zheng, D., Qian, Z. Dong, Liu, Y., & Liu, C. Bo. (2018). Prediction and sensitivity analysis of long-term skid resistance of epoxy asphalt mixture based on GA-BP neural network. *Construction and Building Materials*, 158: 614–623.

Zheng, S., Cao, C. X., & Singh, R. P. (2014). Comparison of ground based indices (API and AQI) with satellite based aerosol products. *Science of the Total Environment*, 488-489(1): 398–412.

Zickus, M., Greig, A. J., & Niranjan, M. (2002). Comparison of four machine learning methods for predicting $PM_{10}$ concentrations in Helsinki, Findland. *Water, Air, and Soil Pollution: Focus*, 2: 717–729.

Zizi, M. N., Noor, M. N., Hashim, I. M. N., & Yusuf, S. Y. (2018). Spatial and Temporal Characteristics of Air Pollutants Concentrations in Industrial Area in Malaysia. *IOP Conference Series: Materials Science and Engineering*, 374(1): 12-94.

Zong, R., Yang, X., Wen, L., Xu, C., Zhu, Y., Chen, T., Yao, L., Wang, L., Zhang, J., Yang, L., Wang, X., Shao, M., Zhu, T., Xue, L., & Wang, W. (2018). Strong ozone production at a rural site in the North China Plain: Mixed effects of urban plumes and biogenic emissions. *Journal of Environmental Sciences,* 71: 261-270.

# APPENDICES

## Appendix 1: Locality of Continuous Air Quality Monitoring Stations starting from April 2017

| No. | Site State | Location | Latitude | Longitude | Zone | Classification |
|---|---|---|---|---|---|---|
| 1 | Perlis | Institut Latihan Perindustrian (ILP) Kangar | 06° 25' 47.71" N | 100° 12' 39.84" E | North | Sub Urban |
| 2 | Kedah | Kompleks Sukan Langkawi | 06° 19' 53.54" N | 099° 51' 30.45" E | North | Sub Urban |
| 3 | Kedah | Sek. Men. Agama Kedah, Alor Setar | 06° 08' 13.49" N | 100° 20' 48.71" E | North | Sub Urban |
| 4 | Kedah | Sek.Men. Keb. Tunku Ismail, Sungai Petani | 05° 37' 46.63" N | 100° 28' 03.83" E | North | Sub Urban |
| 5 | Kedah | Kulim Hitech Park, Kulim | 05° 24' 05.82" N | 100° 35' 22.70" E | North | Industrial |
| 6 | P.Pinang | Sek.Keb. Seberang Jaya II, Seberang Jaya | 05° 23' 53.41" N | 100° 24' 14.20" E | North | Urban |
| 7 | P.Pinang | Kolej Vokasional Seberang Perai, Perai | 05° 19' 45.68" N | 100° 26' 36.51" E | North | Sub Urban |
| 8 | P.Pinang | Universiti Sains Malaysia (USM), Minden | 05° 21' 22.35" N | 100° 18' 28.51" E | North | Urban |
| 9 | P.Pinang | Kolej Vakasional Balik Pulau, Balik Pulau | 05° 20' 13.61" N | 100° 12' 59.21" E | North | Sub Urban |
| 10 | Perak | Sek. Keb. Ayer Puteh, Taiping | 04° 53' 55.86" N | 100° 40' 44.78" E | North | Sub Urban |
| 11 | Perak | Sek. Men. Keb. Jalan Tasek, Ipoh | 04° 37' 45.99" N | 101° 06' 59.94" E | North | Urban |
| 12 | Perak | Sek. Keb. Jalan Pegoh, Pegoh | 04° 33' 12.00" N | 101° 04' 48.84" E | North | Sub Urban |
| 13 | Perak | Pentadbiran Daerah Manjung, Seri Manjung | 04° 12' 01.23" N | 100° 39' 48.08" E | North | Rural |
| 14 | Perak | Universiti Perguruan Sultan Idris, Tanjung Malim | 03° 41' 15.92" N | 101° 31' 28.17" E | North | Sub Urban |
| 15 | K.Lumpur | Sek. Keb. Batu Muda, Batu Muda | 03° 12' 44.78" N | 101° 40' 56.02" E | Central | Sub Urban |
| 16 | K.Lumpur | Sek. Men. Keb. Seri Permaisuri, Cheras | 03° 06' 22.44" N | 101° 43' 04.50" E | Central | Urban |
| 17 | Putrajaya | Sek. Keb. Presint 18, Putrajaya | 02° 54' 53.33" N | 101° 41' 24.17" E | Central | Sub Urban |
| 18 | Selangor | Sek. Men. Sains Selangor, Kuala Selangor | 03° 19' 16.70" N | 101° 15' 22.47" E | Central | Rural |

| | | | | | | |
|---|---|---|---|---|---|---|
| 19 | Selangor | Sek. Keb. Bandar Utama, Petaling Jaya | 03° 07' 59.40" N | 101° 36' 28.83" E | Central | Sub Urban |
| 20 | Selangor | Sek. Keb. TTDI Jaya, Shah Alam | 03° 06' 16.98" N | 101° 33' 22.39" E | Central | Urban |
| 21 | Selangor | Klinik Kesihatan Pandamaran, Klang | 03° 00' 53.60" N | 101° 24' 47.19" E | Central | Sub Urban |
| 22 | Selangor | Kolej Mara Banting, Banting | 02° 49' 00.08" N | 101° 37' 23.36" E | Central | Sub Urban |
| 23 | N.Sembilan | Sek. Keb. Taman Semarak (Fasa 2), Nilai | 02° 49' 18.09" N | 101° 48' 41.34" E | Central | Sub Urban |
| 24 | N.Sembilan | Sek. Men. Teknik Tuanku Jaafar, Seremban | 02° 43' 24.17" N | 101° 58' 06.58" E | Central | Urban |
| 25 | N.Sembilan | Pusat Sumber Pendidikan N. Sembilan, Port Dickson | 02° 26' 28.97" N | 101° 52' 00.68" E | Central | Sub Urban |
| 26 | Melaka | Sek. Men. Keb. Seri Pengkalan, Alor Gajah | 02° 22' 15.33" N | 102° 13' 28.53" E | South | Rural |
| 27 | Melaka | Sek. Keb. Tanjung Minyak 2, Bukit Rambai | 02° 16' 06.57" N | 102° 11' 37.19" E | South | Sub Urban |
| 28 | Melaka | Sekolah Tinggi Melaka, Bandaraya Melaka | 02° 11' 27.36" N | 102° 15' 25.40" E | South | Urban |
| 29 | Johor | Klinik Kesihatan Bandar IOI, Segamat | 02° 29' 38.09" N | 102° 51' 45.69" E | South | Sub Urban |
| 30 | Johor | MRSM Batu Pahat, Batu Pahat | 01° 55' 09.56" N | 102° 51' 59.82" E | South | Sub Urban |
| 31 | Johor | Kolej Kejururawatan Kluang, Kluang | 02° 02' 16.37" N | 103° 18' 43.42" E | South | Rural |
| 32 | Johor | Institut Perguruan Temenggong Ibrahim, Larkin | 01° 29' 40.65" N | 103° 44' 09.50" E | South | Urban |
| 33 | Johor | Sek. Men. Pasir Gudang 2, Pasir Gudang | 01° 28' 12.43" N | 103° 53' 36.44" E | South | Sub Urban |
| 34 | Johor | Sek. Keb. Lepau, Pengerang | 01° 23' 22.16" N | 104° 08' 58.50" E | South | Industrial |
| 35 | Johor | Sek. Men. Agama Bandar Penawar, Kota Tinggi | 01° 33' 50.60" N | 104° 13' 31.10" E | South | Sub Urban |
| 36 | Pahang | Sek. Men. Keb. Seri Rompin, Rompin | 02° 55' 35.92" N | 103° 25' 09.11" E | East | Rural |
| 37 | Pahang | Meteorologi Temerloh, Temerloh | 03° 28' 17.77" N | 102° 22' 35.06" E | East | Sub Urban |
| 38 | Pahang | SMK. Jerantut, Jerantut | 03° 56' 54.09" N | 102° 21' 59.87" E | East | Sub Urban |
| 39 | Pahang | Sek. Keb. Indera Mahkota, Indera Mahkota | 03° 49' 09.18" N | 103° 17' 47.57" E | East | Sub Urban |
| 40 | Pahang | Sek. Keb. Balok Baru, Balok Baru | 03° 57' 38.31" N | 103° 22' 55.76" E | East | Industrial |

| | | | | | | |
|---|---|---|---|---|---|---|
| 41 | Terengganu | SMK Bukit Kuang, Kemaman | 04° 15' 43.46" N | 103° 25' 32.90" E | East | Industrial |
| 42 | Terengganu | Kuarters TNB Paka, Paka | 04° 35' 53.03" N | 103° 26' 05.34" E | East | Industrial |
| 43 | Terengganu | Sek. Keb. Chabang Tiga, Kuala Terengganu | 05° 18' 29.13" N | 103° 07' 13.41" E | East | Urban |
| 44 | Terengganu | Sek. Keb. Nyiur Tujuh, Besut | 05° 44' 54.41" N | 102° 30' 56.27" E | East | Sub Urban |
| 45 | Kelantan | Sek. Men. Tanah Merah, Tanah Merah | 05° 48' 40.21" N | 102° 08' 04.20" E | East | Sub Urban |
| 46 | Kelantan | Sek. Men. Keb. Tanjong Chat, Kota Bharu | 06° 08' 50.75" N | 102° 14' 57.24" E | East | Sub Urban |
| 47 | Sabah | JKR Tawau, Tawau | 04° 14' 59.22" N | 117° 56' 09.11" E | Sabah | Sub Urban |
| 48 | Sabah | JKR Sandakan, Sandakan | 05° 51' 52.08" N | 118° 05' 27.92" E | Sabah | Sub Urban |
| 49 | Sabah | Sek. Men. Keb. Tansau, Kota Kinabalu | 05° 52' 46.87" N | 116° 03' 23.13" E | Sabah | Sub Urban |
| 50 | Sabah | Sek. Men. Keb. Bongawan II, Kimanis | 05° 32' 17.60" N | 115° 51' 02.00" E | Sabah | Industrial |
| 51 | Sabah | Sek. Men. Keb. Gunsanad, Keningau | 05° 20' 21.54" N | 116° 09' 49.16" E | Sabah | Background |
| 52 | Labuan | Kolej Vokasional, Labuan | 05° 19' 57.81" N | 115° 14' 17.62" E | Labuan | Sub Urban |
| 53 | Sarawak | Dewan Suarah Limbang, Limbang | 04° 45' 32.00" N | 115° 00' 49.20" E | Sarawak | Rural |
| 54 | Sarawak | Institut Latihan Perindustrian Miri, Permyjaya | 04° 29' 41.24" N | 114° 02' 36.29" E | Sarawak | Rural |
| 55 | Sarawak | Sek. Men. Dato Permaisuri, Miri | 04° 25' 28.84" N | 114° 00' 44.73" E | Sarawak | Sub Urban |
| 56 | Sarawak | Samalaju Industrial Estate, Samalaju | 03° 32' 13.41" N | 113° 17' 42.60" E | Sarawak | Industrial |
| 57 | Sarawak | Balai Polis Pusat Bintulu, Bintulu | 03° 10' 37.50" N | 113° 02' 27.92" E | Sarawak | Sub Urban |
| 58 | Sarawak | Politeknik Mukah, Mukah | 02° 52' 59.65" N | 112° 01' 11.07" E | Sarawak | Rural |
| 59 | Sarawak | Stadium Tertutup Kapit, Kapit | 02° 00' 52.19" N | 112° 55' 38.49" E | Sarawak | Rural |
| 60 | Sarawak | Ibu Polis Sibu, Sibu | 02° 18' 51.86" N | 111° 49' 54.89" E | Sarawak | Sub Urban |
| 61 | Sarawak | Balai Polis Pusat Sarikei, Sarikei | 02° 07' 58.11" N | 111° 31' 22.33" E | Sarawak | Rural |
| 62 | Sarawak | Kompleks Sukan Sri Aman, Sri Aman | 01° 13' 10.76" N | 111° 27' 53.25" E | Sarawak | Rural |
| 63 | Sarawak | Daerah Perumahan Samarahan, Samarahan | 01° 27' 17.47" N | 110° 29' 29.41" E | Sarawak | Rural |

| | | | | | | |
|---|---|---|---|---|---|---|
| 64 | Sarawak | Depot Ubat Kementerian Kesihatan Malaysia, Kuching | 01° 33' 44.02" N | 110° 23' 20.24" E | Sarawak | Urban |
| 65 | Johor | Tapak Semaian Majlis Daerah Tangkak | 02°17' 41.25" N | 102° 34'17.74" E | Johor | Sub urban |

## LIST OF PUBLICATIONS

**Rani, N. L. A.**, Azid, A., Khalit, S. I., & Juahir, H. (2018). Prediction model of missing data: a case study of $PM_{10}$ across Malaysia Region. *Journal of Fundamental and Applied Sciences*, 10(1S): 182-203. **(ISI)**

**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): 1-8. **(Q4, ISI/SCOPUS)**

**Rani, N. L. A.**, Azid, A., Khalit, S. I., Gasim, M. B., & Juahir, H. (2017). Selected Malaysia air quality pollutants assessment using chemometrics techniques. *Journal of Fundamental and Applied Sciences*, 9(2S): 335-351. **(ISI)**

Azid, A., Amin, S. N. S. M., Khalit, S. I., Ismail, S., Samsudin, M. S., Yusof, K. M. K. K., **Rani, N. L. A.**, Amran, M. A., Yunus, K., & Saudi, A. S. M. (2018). Determination of selected heavy metals in airborne particles in industrial area: A baseline study. *Malaysian Journal of Fundamental and Applied Sciences*, 14(2): 251-256. **(ISI)**

Azid, A., **Rani, N. L. A.**, Samsudin, M. S., Khalit, S. I., Gasim, M. B., Kamarudin, M. K. A., Yunus, K., Saudi, A. S. M., & Yusof, K. M. K. K. (2017). Air quality modelling using chemometric techniques. *Journal of Fundamental and Applied Sciences*, 9(2S): 443-466. **(ISI)**

Azid, A., Amran, M. A., Samsudin, M. S., **Rani, N. L. A.**, Khalit, S. I., Gasim, M. B., Yunus, K., Saudi, A. S. M., Amin, S. N. S. M., & Yusof, K. M. K. K. Assessing Indoor Air Quality Using Chemometric Models. *Polish Journal of Environmental Studies,* 27(6): 1-8. **(Q4, ISI/SCOPUS)**

**Accepted papers:**

**Rani, N. L. A.**, & Azid, A. Trend And Missing Data Prediction Model of $PM_{10}$ in Central Region using ANN and MLR. *Malaysia Journal of Analytical Sciences.International Conference of Analytical Sciences 2018.* **(SCOPUS)**

**Rani, N. L. A.**, Azid, A., Sani, M. S. A., Samsudin, M. S., Yusof, K. M. K. K., & Amin, S. N. S. M. Development of Missing Data Prediction Model for Carbon Monoxide. *Malaysian Journal of Fundamental and Applied Sciences* **(ISI)**

**Rani, N. L. A.**, & Azid, A. Imputation Methods of Mean, Nearest Neighbour and Expectation Maximation Based Algorithm (EMB) on Daily $PM_{10}$ Data (2010-2015). *International Conference on Agriculture, Animal Sciences & Food Technology (ICAFT) 2018* **(ISI)**

## CANDIDATE BIODATA

Nurul Latiffah Abd Rani was born in Kuala Terengganu, Terengganu in 1989. She has obtained a Degree in Environmental Technology at Universiti Teknologi MARA, Shah Alam in 2011. Upon completion of her degree, she has continued her Masters Degree in Science (by Research) in Environmental Radiochemistry from Universiti Teknologi MARA, Shah Alam in 2014. Due to the immerse interest in the field of data analysis, especially chemometrics in the environment, she decided to enhance her knowledge towards the analysis of data besides her experienced towards analysis of laboratoty work during her Masters Degree level by pursuing her PhD study in Applied Sciences (Air Quality-Chemometrics, Artificial Intelligence) at Universiti Sultan Zainal Abidin, Terengganu. During her study, she was awarded MyPHD scholarship from MYBRAIN15 by Kementerian Pendidikan Malaysia (KPM). She currently a Graduate Student of the Faculty of Bioresources and Food Industry, Universiti Sultan Zainal Abidin, Besut Campus, Terengganu.