PREDICTION OF PM_{10} CONCENTRATION USING MULTIPLE LINEAR REGRESSION AND BAYESIAN MODEL AVERAGING

HAFIZAHIZZATI BINTI ISMAIL

SCHOOL OF CIVIL ENGINEERING
UNIVERSITI SAINS MALAYSIA
2017

PREDICTION OF PM_{10} CONCENTRATION USING MULTIPLE LINEAR REGRESSION AND BAYESIAN MODEL AVERAGING

By

HAFIZAHIZZATI BINTI ISMAIL

This dissertation is submitted to

UNIVERSITI SAINS MALAYSIA

As partial fulfilment of requirement for the degree of

BACHELOR OF ENGINEERING (HONS.)
(CIVIL ENGINEERING)

School of Civil Engineering,
Universiti Sains Malaysia

June 2017

UNIVERSITI SAINS MALAYSIA

SCHOOL OF CIVIL ENGINEERING
ACADEMIC SESSION 2014/2015

FINAL YEAR PROJECT EAA492/6
DISSERTATION ENDORSEMENT FORM

Title: Prediction of PM_{10} Concentration Using Multiple Linear Regression and Bayesian Model Averaging

Name of Student: Hafizahizzati Binti Ismail

I hereby declare that all corrections and comments made by the supervisor(s)and examiner have been taken into consideration and rectified accordingly.

Signature: Approved by:

_____ _____

(Signature of Supervisor)

Date : 23/6/2017 Name of Supervisor :

Date :

Approved by:

_____

(Signature of Examiner)

Name of Examiner :

Date :

ACKNOWLEDGEMENT

Alhamdulillah, thanks to Allah the creator to all being up in the sky and on the earth because of His grace, I am still alive and have to do responsibilities given to me as a Muslim. First of all, I want to thank Professor Ahmad Shukri bin Yahaya for guiding me to do this research from start until finish, which is required of me as a student of the School of Civil Engineering to meet part of the requirement for final year project that relate to the field in environmental and evaluation in final semester and finally graduate in this course. I believe without his support and excellent supervision, this project never have been successful.

Besides that, I would like to thank all lecturers in Universiti Sains Malaysia (USM) Engineering Campus especially lectures from the School of Civil Engineering for the encouragement, advices and ideas that were given to me during this project. I will use all the knowledge and experiences that I have gain during the way to finish this project for work and future. Without their advice and guidance, it would be difficult for me to complete my project. Furthermore, a special thank is also given to the Department of Environment Malaysia who has given me a permission and providing the data needed to carry out this project.

To my family especially my beloved parents and siblings because always provided me with full support and encouragement me along the journey of completing this final year project. Not forget a special to thank all my friends from school Civil Engineering for their support and advice. Finally, nobody is perfect and as human I cannot run away from mistakes, I want to put ten fingers and apologies if there are mistakes.

ABSTRAK

PM_{10} ambien (zarah terampai dengan diameter aerodynamic kurang daripada 10μm) adalah salah satu pencemar yang mempunyai kesan negatif ke atas kesihatan manusia dan alam sekitar. Ia dipengaruhi oleh parameter cuaca dan gas. Kajian ini adalah untuk meramal kepekatan zarah terampai (PM_{10}) dengan menggunakan model linear regresi beganda dan purata model Bayesian. Empat stesen telah dipilih untuk tiga tahun (2013 hingga 2015) yang terletak di Jerantut, Nilai, Seberang Jaya dan Shah Alam. Sebelum memulakan analisis, data dibahagikan kepada dua kategori iaitu data latihan dan data pengesahan. Data latihan adalah 70% daripada data yang diperhatikan (bermula pada hari 1 hingga hari 255) telah digunakan untuk mendapatkan model. Satu lagi 30% daripada data yang diperhatikan (bermula pada hari 256 hingga hari 365) telah digunakan untuk tujuan pengesahan. Analisis deskriptif menunjukkan bahawa pada tahun 2015, Nilai mencatatkan purata kepekatan PM_{10} yang tertinggi berbanding stesen lain. Nilai maksimum kepekatan PM_{10} yang paling tinggi dicatatkan di stesen Seberang Jaya yang berlaku pada tahun 2015 disebabkan oleh musim peralihan monsun yang menunjukkan tahap PM_{10} melebihi tahap keselamatan berdasarkan garis panduan kualiti udara di Malaysia. Untuk mendapatkan parameter yang menyumbang kepada pencemaran udara bagi ramalan zarah terampai untuk hari keesoknya ($PM_{10,D1}$), data latihan dianalisis dengan menggunakan perisian SPSS bagi model regresi linear berganda dan perisian R bagi purata model Bayesian. Keputusan menunjukkan bahawa stesen Shah Alam menyumbang parameter utama kepada ramalan zarah terampai untuk hari esok ($PM_{10,D1}$) yang mempunyai nilai $\overline{R}^2$ tertinggi dengan menggunakan model linear regresi beganda. Penilaian prestasi model menunjukkan bahawa purata model Bayesian adalah model yang terbaik untuk meramalkan kepekatan PM_{10} untuk hari esok ($PM_{10,D1}$) dengan menggunakan data pengesahan.

ABSTRACT

Ambient PM_{10} (particulate matter with an aerodynamic diameter less than 10μm) is one of the pollutant that has negative impacts on human health and environment. It is influenced by weather and gaseous parameters. This study is to predict particulate matter (PM_{10}) concentration by using multiple linear regression and Bayesian model averaging. Four stations were selected for three years (2013 until 2015) which are located in Jerantut , Nilai, Seberang Jaya and Shah Alam. Before the analysis, the data was divided into two categories which are training data and validation data. The training data is 70% of observed data (beginning on day 1 until day 255) used to obtain the model. Another 30% of observed data (beginning on day 256 until day 365) were used for validation purpose. The descriptive analysis showed that in 2015, Nilai recorded the highest mean value of PM_{10} concentration compared to other stations while the highest maximum value of PM_{10} concentration was recorded at Seberang Jaya station that happened in 2015 due to inter-monsoon season that indicate PM_{10} level is above threshold value following Malaysia Ambient Air Quality Guideline (MAAQG). To obtain the parameters that contribute to air pollutant for the prediction of particulate matter for the next day ($PM_{10,D1}$), the training data was analysed using SPSS software for multiple linear regression model and R Software for Bayesian model averaging. The results showed that Shah Alam station is contributing the main parameters which have highest value of adjusted $\overline{R}^2$ by using multiple linear regression models. Assessment of model performance indicated that Bayesian model averaging (BMA) is the better model to predict PM_{10} concentration for the next day ($PM_{10,D1}$) by using the validation data.

TABLE OF CONTENTS

ACKNOWLEDGEMENT ..... **I**

ABSTRAK ..... **II**

ABSTRACT ..... **III**

TABLE OF CONTENTS ..... **IV**

LIST OF FIGURES ..... **VI**

LIST OF TABLES ..... **VII**

LIST OF ABBREVIATIONS ..... **VIII**

CHAPTER 1 ..... **1**

1.1 Background 1

1.2 Problem Statement 4

1.3 Objectives 6

1.4 Scope of Work 7

1.5 Outline of Research 7

CHAPTER 2 ..... **9**

2.1 Introduction 9

2.2 Particulate Matter (PM), Weather Parameters and Gaseous Parameters 9

2.3 Application of Statistical Modelling For Prediction of Particulate Matter 11

2.4 Bayesian Approach 14

2.4.1 Bayesian Model Averaging Approach 15

2.5 Summary of Literature Review 20

CHAPTER 3 ..... **21**

3.1 Introduction 21

3.2 Data Collection 23

3.3 Site Description 25

3.4 Instrument Specification 26

3.5 Descriptive Analysis 27

3.5.1 Minimum and Maximum Value 27

3.5.2 Mean 28

3.5.3 Median 28
3.5.4 Mode 29
3.5.5 Standard Deviation 29
3.5.6 Coefficient of Variation (CV) 30
3.5.7 Skewness and Kurtosis 30
3.5.8 Box and Whisker Plot 31
3.6 Statistical Analysis 32
3.6.1 Multiple Linear Regression Model 33
3.6.2 Bayesian Model Averaging 34
3.6.3 Performance Indicator 36
CHAPTER 4 38
4.1 Introduction 38
4.2 Descriptive Statistics 38
4.2.1 Descriptive Statistics at Jerantut Station 38
4.2.2 Descriptive Statistics at Nilai Station 41
4.2.3 Descriptive Statistics at Seberang Jaya Station 43
4.2.4 Descriptive Statistics at Shah Alam Station 45
4.2.5 Box Plot of PM_{10} Concentration in 2013, 2014 and 2015 47
4.2.6 Descriptive Statistics of PM_{10} Concentration with Gaseous and Weather Parameters in 2013, 2014 and 2015 49
4.2.7 Box Plot of PM_{10} Concentration for All Years 52
4.3 Multiple Linear Regression Model (MLR) and Bayesian Model Averaging (BMA) 53
4.3.1 Multiple Linear Regression (MLR) and Bayesian Model Averaging (BMA) for All stations 57
4.4 Performance Indicator 58
4.4.1 Performance indicators for PM_{10} concentration at Jerantut, Nilai, Seberang Jaya and Shah Alam Station 59
4.4.2 Performance Indicator for PM_{10} concentration for All Stations 62
CHAPTER 5 64
5.1 Introduction 64
5.2 Conclusions 64
5.3 Recommendation 66
REFERENCES 67

LIST OF FIGURES

Figure 1.1	The Particulate matter (PM) emission load by sources (in percentage), started 2004 until 2012	6
Figure 3.1	Flow chart for study procedure	22
Figure 3.2	Continuous Air Quality Monitoring Station (CAQMS) in Malaysia	24
Figure 3.3	Beta Attenuation Monitor (BAM -1020)	27
Figure 3.4	Element of Box plot and Whisker plot	32
Figure 3.5	Occam Window of interpreting the Posterior odds for Nested Models	36
Figure 4.1 (a-d)	Box Plot for PM_{10} concentration in (a) Jerantut, (b) Nilai, (c) Seberang Jaya and (d) Shah Alam	48
Figure 4.2	Box Plot for PM_{10} concentration at Jerantut, Nilai, Seberang Jaya and Shah Alam	52

LIST OF TABLES

Table 1.1	Malaysia Air Pollution Index (API)	2
Table 3.1	Detailed Location of The Monitoring Sites	23
Table 3.2	Formulation of performance indicator	37
Table 4.1	Descriptive statistics at Jerantut station	40
Table 4.2	Descriptive statistics at Nilai station	42
Table 4.3	Descriptive statistics at Seberang Jaya station	44
Table 4.4	Descriptive statistics at Shah Alam station	46
Table 4.5	Descriptive statistics of PM_{10} oncentration with gaseous and weather parameters in 2013, 2014 and 2015	51
Table 4.6	Multiple Linear Regression model (MLR) and Bayesian Model Averaging (BMA) of $PM_{10,D1}$ at Jerantut and Nilai,	55
Table 4.7	Multiple Linear Regression model (MLR) and Bayesian Model Averaging (BMA) of $PM_{10,D1}$ at Seberang Jaya and Shah Alam station	56
Table 4.8	Multiple Linear Regression model (MLR) and Bayesian Model Averaging (BMA) of $PM_{10,D1}$ for all stations	58
Table 4.9	Performance indicators of PM_{10} concentration at Jerantut and Nilai stations	60
Table 4.10	Performance indicators for PM_{10} concentration at Seberang Jaya and Shah Alam stations	61
Table 4.11	Performance indicators of PM_{10} for all stations in 2013, 2014 and 2015	63

LIST OF ABBREVIATIONS

API	Air Pollution Indec
BAM	Beta Attenuation Monitor
BMA	Bayesian Model Averaging
CAQMS	Continuous Air Quality Monitoring Station
CO	Carbon Monoxide
DOE	Department of Environment
FACT2	Factor of two
IA	Index of Agreement
MAAQG	Malaysia Ambient Air Quality Guidelines
MLR	Multiple Linear Regression
MBE	Mean Bias Error
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
NO_2	Nitrogen Dioxide
O_3	Ozone
PM_{10}	Particulate Matter with an aerodynamic diameter less than 10μm
$PM_{10,Do}$	Particulate Matter with an aerodynamic diameter less than 10μm for the previous day

$PM_{10,D1}$	Particulate Matter with an aerodynamic diameter less than 10μm for the next day
PI	Performance Indicator
PA	Prediction Accuracy
RH	Relative Humidity
SO_2	Sulphur Dioxide
SPSS	Statistical Package and Services Solution
Temp	Temperature
USEPA	United States Environmental Protection Agency

CHAPTER 1

INTRODUCTION

1.1 Background

Air pollution poses enormous health risks to billions of people around the world every day, contributing to heart disease, lung failure along with many other fatal human illnesses and diseases. Managing air pollution is critical to human, economic and environmental growth in communities throughout the world today (Ernesto, 2013). Air pollution is a process that introduces diverse pollutant into atmosphere that cause harm to human, other living organism, and the natural environment (Kinney, 2008; Brauer et al., 2012; Kim et al., 2013). Air pollutants, such as carbon monoxide (CO), sulphur dioxide (SO_2), nitrogen oxides (NO_2), volatile organic compounds (VOCs), ozone (O_3), heavy metals, and respirable particulate matter ($PM_{2.5}$ and PM_{10}), differ in their chemical composition, reaction properties, emission, time of disintegration and ability to diffuse in long or short distances (Kampa et al., 2008).

Besides that, air pollution is contamination of the indoor or outdoor environment by any chemical, physical or biological agent that modifies the natural characteristics of the atmosphere (World Health Organization, 2014). According to Kampa et al., (2008) air pollution has both acute and chronic effects on human health, affecting a number of different systems and organs. It ranges from minor upper respiratory irritation to chronic respiratory and heart disease, lung cancer, acute respiratory infections in children and chronic bronchitis in adults, aggravating pre-existing heart and lung disease, or asthmatic attacks.

Bishoi et al., (2009) define the Air Pollutant Index (API) as number that is used for reporting the quality of air with respect to its effects on human health. The API is an important tool to inform the public about how clean or polluted the atmosphere and what associated health effects might be of concern to us. Based on Department of Environment (2017), Malaysia Air Pollution Index (API) values have been divided into five ranges which are good, moderate, unhealthy, very unhealthy and hazardous based on possible health effects. The API for a given time period is calculated based on the sub-index values (sub-API) for all the five air pollutants included in Malaysian API system such as sulphur dioxide or SO_2, nitrogen dioxide or NO_2, ozone or O_3, carbon monoxide or CO, particulate matter below 10 micron in size or PM_{10}. The range of the API value is shown in Table 1.1 where the higher the level of the API, the greater the level of air pollution and the greater is the health concern.

Table 1.1 : Malaysia Air Pollution Index (Source : Department of Environment,2017)

Air Pollution Index (API)	**Air Status Quality**
0 – 50	Good
51 – 100	Moderate
101 – 200	Unhealthy
210 – 300	Very Unhealthy
>300	Hazardous

In Malaysia, air pollution are caused by industrial activities (manufacturing processes and oil refiners), construction works (building construction), transportation (driving vehicles), power plants stations (coal – fired power plant or fossil fuels) and also by open burning (Hanapi, 2012). According to Economic Planning Unit, n. d., the main sources of air pollution in Malaysia are motor vehicles, power stations, industrial

fuel burning and processes, domestic fuel burning, burning of municipal and industrial waste.

Yahaya et al., (2011) has described that Malaysian Ambient Air Quality Guidelines (MAAQG) were issued and target values for annual and daily mean mass concentrations for various air pollutant were established to control and reduce air pollutant levels in the atmosphere. Stated by Department of Environment (2017), Malaysian Ambient Air Quality Standard (MAAQS), there are three interim targets set which include interim target 1 (IT-1) in 2015, interim target 2 (IT-2) in 2018 and the full implementation of the standard in 2020. For interim target 1 (IT-1), PM_{10} concentration in ambient air in Malaysia are monitored as recommended by Malaysian guideline at a threshold value of 150 $\mu g/m^3$ for 24 hours average and an annual means of 50 $\mu g/m^3$.

In addition, ambient respirable particles or particulate matter with a aerodynamic diameter of less than 10 μm (PM_{10}) have attracted special legislative and scientific attention due to their effects on human health and the particles with a aerodynamic diameter of less than 10 μm constitute the so-called inhalable fraction of particles, which are able to reach the bronchi-tracheal area. PM_{10} is made up of a variety of solid and liquid substances derived from natural sources (e.g., volcanoes, dust storms, forest and grassland fires, living vegetation, and marine salts) and human activities (e.g., central heating, industry, construction works, vehicular traffic, domestic heating, and incinerators) and from a chemical point of view, a complex mixture of organic and inorganic carbon, metals (lead, arsenic, mercury, cadmium, chrome, nickel, and vanadium), nitrates, sulphates and phosphates are present in the particulates (Taheri & Sodoudi, 2016).

1.2 Problem Statement

Particulate Matter (PM) is one of the six criteria pollutants, and the most important in terms of adverse effects on human health and particulate matter (PM) is the term used for a mixture of solid particles and liquid droplets suspended in the air. These particles originate from a variety of sources, such as power plants, industrial processes, and diesel trucks, and they are formed in the atmosphere by transformation of gaseous emissions (Fierro, 2000).

Exposure to PM_{10} has been consistently associated with serious health outcomes, resulting in an increase in mortality and hospital admissions, predominantly related to cardiovascular and respiratory diseases (Pascal et al., 2014). United Stated Environmental Protection Agency (2017) also reported PM_{10} can also coughing, difficulty in breathing, aggravated asthma, chronic bronchitis, irregular heartbeat, non-fatal heart attacks and some cancer.

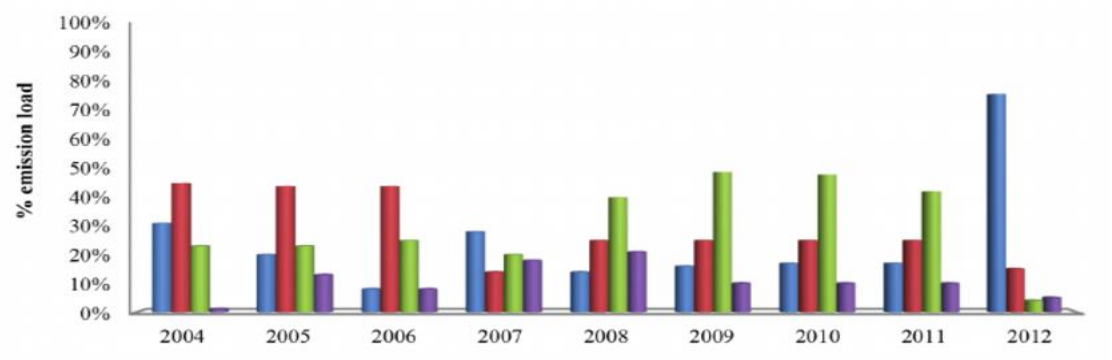
A study by Azmi et al. (2010) shows that severe air quality problems exist in highly urbanised areas on the Malaysian Peninsular. This is true with respect to dust fall-out, suspended particular matter and lead in the ambient air along congested roads. These problems are mainly attributed to motor vehicle emissions (Awang et al., 2000).

Air pollutants are released to the ambient air by two major sources namely, natural sources and anthropogenic activities. Forest fires and windblown dust are among the natural sources that release pollutants to the air (Department of Environment Malaysia 2013). Stated by Salinas, (2013) forest fire is used for land preparation and forest clearance by parties or people involved in plantation which then developed into uncontrollably burning wild-fires. This situation took place between June and November coinciding with drier weather condition.

Ahmat et al, (2015) describes, stationary and mobile sources are the contributors to anthropogenic activities. Among the major local sources of air pollutions in Malaysia are the increasing numbers of motor vehicle usage (mobile sources) and rapid development of industrial sector (stationary sources). Besides that, stated by Afroz and Ibrahim, (2003) stationary sources include power plants, industrial waste incinerators, the emission of dust from urban construction works and quarries, along with open burning. All this activities contribute to the PM_{10} concentration.

Besides that, gaseous parameters (sulphur dioxide (SO_2), nitrogen dioxide (NO_2), ozone (O_3) and carbon monoxide (CO) give impact on human health. Stated on (Local Government Air Quality Toolkit, n.d.) the effect of SO_2 on human health is to irritate the nose, throat and airways so as to cause coughing and shortness of breath. It aggravate asthma and chronic bronchitis and increased susceptibility to respiratory tract infections. The molecules can attach themselves to particles and if these are then inhaled they can cause more serious effects such as emphysema from long-term exposure. A concern with SO_2 is people's short-term exposures to high concentrations in the near vicinity of activities that emit significant amounts of the gas, which under some meteorological conditions may be brought to ground level in high concentrations. Infrequently, high concentrations may also occur due to abnormal conditions within a facility through industrial accidents or breakdown of poorly maintained equipment.

Hence, when referring the Figure 1.1, for the past few years, power and industrial plants have been the major source particulate matter in Malaysia. However, emission of pollution from motor vehicles has also increased annually. Emission of pollution from motor vehicle has increased started in 2008 (14%) until 2012 (76%). It can seen the main emission contribute in the air are motor vehicle followed by power plant, industrial and others as shown in Figure 1.



	2004	2005	2006	2007	2008	2009	2010	2011	2012
Motor vehicles	31%	20%	8%	28%	14%	16%	17%	17%	76%
Power plants	45%	44%	44%	14%	25%	25%	25%	25%	15%
Industrial	23%	23%	25%	20%	40%	49%	48%	42%	4%
Others	1%	13%	8%	18%	21%	10%	10%	10%	5%

Figure 1.1 : The Particulate matter (PM) emission load by sources (in percentage), started 2004 until 2012 (Source: Malaysia Environmental Quality Report 2012, 2011, 2010, 2009, 2008, 2007, 2006, 2005 & 2004)

Different problems connected with protection of our environment can successfully be treated by using large mathematical models as preliminary calculations of estimating the pollutant concentration. There are many researchers using multiple linear regression (MLR) (Ul-Saufie et al., 2012; Elbayomi et al., 2015; Sayegh et al., 2014), artificial neural network (ANN) (Ul-Saufie et al., 2012; McKendry, 2002); Chelani et al., 2002) and principal component regression (PCR) (Jamaludin, 2016; Viana et al., 2006; Sousa et al., 2007) models to predict the PM_{10} concentration. From that, to improve the prediction with better model, this study using Bayesian model averaging (BMA) model to predict the PM_{10} concentration.

1.3 Objectives

The objective of this study are :

1. To determine the characteristics of the PM_{10} with weather parameters (relative humidity, temperature and wind speed) and gaseous parameters

(sulphur dioxide (SO_2), nitrogen dioxide (NO_2), ozone (O_3) and carbon monoxide (CO)).

2. To predict PM_{10} concentration using Multiple Linear Regression and Bayesian Model Averaging modelling.
3. To determine the best model to predict the PM_{10} concentrations.

1.4 Scope of Work

This research was conducted at four stations which are Nilai monitoring stations (industrial area), Seberang Jaya (Sub-urban area), Shah Alam (urban area) and Jerantut (reference station). The duration taken for the data is based on the daily data and started from 2013 until 2015 that was obtained from the Department of Environment (DOE) Malaysia. In Malaysia, the Department of Environment (DOE) monitors country air quality through Continuous Air Quality Monitoring (CAQM).

The parameters selection are weather parameters (relative humidity in percentage, temperature in Celsius and wind speed in meter per second) and gaseous parameters (sulphur dioxide (SO_2), nitrogen dioxide (NO_2), ozone (O_3) and carbon monoxide (CO) in parts per million). Statistical modelling that will be used is Multiple Linear Regression and Bayesian Model Averaging model to predict the concentration of PM_{10}.

1.5 Outline of Research

The thesis consists of five chapters which are Chapter 1(Introduction), Chapter 2 (Literature Review), followed by Chapter 3 (Methodology), Chapter 4 (Results and Analysis) and the last is Chapter 5 (Discussion). The brief outline for every chapter is as follows:

Chapter 1 explained the introduction about air pollution, the sources of air pollution and the impact of PM_{10}. This chapter also includes problem statement, objectives, scope of work and outline of research.

Chapter 2 discussed the literature review for particulate matter (PM) and the summary of the application of statistical analysis. This chapter also include the explanation of multiple linear regression and Bayesian model averaging analysis that has been used by other researches in Malaysia or worldwide in terms of environmental engineering studies.

Chapter 3 described the methods that were used in this study. Parameter selection, descriptive analysis and two models used are multiple linear regression and Bayesian model averaging were explained. Besides that, to get the best model, performance indicator which are error measures and accuracy measures were also discussed.

Chapter 4 represents the results obtained from data analysis which is the characteristics of PM_{10}, weather parameters and gaseous parameters for all monitoring sites. Besides that, all findings for this analysis are used for determining the best model by using performance indicator.

Chapter 5 concludes the research that has been carried out from the results in Chapter 4 and some recommendations for future research are also included.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

Generally, this chapter will discuss in detail about particulate matter and explain about the previous researches that have been done regarding the implementation of multiple linear regression and prediction in the environment field. Other than that, previous studies on the application of Bayesian model averaging in the engineering field were reviewed as additional knowledge of this research.

2.2 Particulate Matter (PM), Weather Parameters and Gaseous Parameters

Based on World Health Organization, (2003), particulate matter (PM) is a mixture of solid particles and liquid droplets found in the air which is made up of a number of components, including acids such as nitrates and sulphate, organic metal and soil or dust particles. The size of particles has been directly linked to their potential for causing health problems. Small particles of concern include "inhalable coarse particles" with a diameter of 2.5 to 10 μm and "fine particles" that is smaller than 2.5 μm in diameter.

As reported by World Health Organization (2003), a short-term increase of 10 $\mu g/m^3$ of particulate matter (PM) for several days is associated with more coughing, lower respiratory symptoms, and hospital admissions increments due to respiratory problems. It has been reported that most of the probable human carcinogens have been found to be associated with suspended particle matter in ambient air of urban areas (Hassan and Khoder 2012). In the urban environment, the main source of PM_{10} is

motor vehicle emissions. It would be influenced by various factors such as wind-speed, air temperature and relative humidity and others (Sulaiman et al., 2005).

Dominick et al., (2012) describe, the changes in wind flow patterns and rainfall distinguish the seasons in this country. The wind throughout the country is generally light and variable as the country is located near the equator. However, uniform periodic changes in wind flow patterns determine the country's four seasons which are the Northeast Monsoon (November to March), a transitional period (April to May), the Southwest Monsoon (June to September) and a second transitional period (October to November). Additionally, the Peninsular Malaysia is characterized by quite high but uniform temperatures (between 23 and 31 °C), along with high relative humidity and high rainfall (± 2500 mm annually).

The movement of air pollutants usually follows the pattern of wind direction base on the northeast monsoon and southwest monsoon. The southwest monsoon usually brings the high amount of particulate matter to Malaysia due to biomass burning in Sumatera and Kalimantan, Indonesia. During the northeast monsoon there are also indicators of the influence of biomass burning particularly in Peninsular Malaysia from Indochina region (Juneng, 2009).

According the Department of Environment, DoE (2010), the main pollutants recorded at the Malaysian air quality monitoring stations are given as particulate matter (PM), NO_2, SO_2 , CO and O_3 . Several recent studies for example Azmi et al., (2010) and Latif et al., (2011) show that the concentrations of NO_x and PM_{10} are showing an increasing trend due to the complete combustion of motor vehicles and biomass burning from inter-boundary sources. The oxides of nitrogen (NO_x) is the generic term for a group of highly reactive gases such as NO and NO_2 as well as other gases which

contain nitrogen and oxygen in varying quantities. Many of the oxides of nitrogen are colourless and odourless.

However, NO_2, one of common pollutant along with particles in the air, can often be seen as a reddish-brown layer over many urban areas. Breathing in gaseous pollutants such as NO_2 and suspended particulate PM_{10} are known to have detrimental effects on human health (Lau et al., 2009). Inhalation of NO_2 can irritate the upper respiratory tract and lungs even at low concentrations and will cause cardiovascular diseases (Chang et al., 2005).

2.3 Application of Statistical Modelling For Prediction of Particulate Matter

Statistics are important in the analysis and interpretation of data in which the outcomes from the analysis can be utilized as prediction tools that have become the major aim in environmental engineering.

Ul-Saufie et al., (2012) applied multiple linear regression as statistical analyses to predicting performance indicator for future (next day, next 2 days and next 3 days) PM_{10} concentration levels in Seberang Perai, Malaysia. Performance indicators are used for this study such as Prediction Accuracy (PA), Coefficient of Determination (R^2), Index of Agreement (IA), Normalized Absolute Error (NAE) and Root Mean Square Error (RMSE) were used to measure the accuracy of the models. The result of this study by Ul-Saufie et al., (2012) and the assessment of model performance indicated that multiple linear regression method can be used for long term PM_{10} concentration prediction with next day for next day where performance indicator shows next day (RMSE = 11.211, NAE = 0.124, PA = 0.927, IA = 0.960, R^2 = 0.858,) and

next 2-day (RMSE = 14.652, NAE = 0.155, PA = 0.881, IA = 0.925, R^2 = 0.775) and next 3-day (RMSE = 15.611, NAE = 0.167, PA = 0.849, IA = 0.912, R^2 = 0.720).

Ul-Saufie et al., (2011) had carried out a research to determine the best technique between Multiple Linear Regression (MLR) and Feedforward Backpropagation Artificial Neural Network (ANN) models for predicting concentration in Pulau Pinang. Relative humidity (RH), wind speed (WS), nitrogen dioxide (NO_2), temperature (T), carbon monoxide (CO), sulphur dioxide (SO_2), ozone (O_3) and previous day $PM_{10,t-1}$ were used as independent variables. The quality and reliability of the developed models were evaluated via performance indicators (NAE, RMSE, PA, IA and R^2). Assessment of model performance indicated that neural network can predict particulate matter better than multiple regressions. However models adequacy checked by various statistical methods showed that the developed multiple regression models can also be used for prediction of PM_{10}.

Elbayomi et al., (2015) conducted a research to evaluate the influence of seasons on the concentrations of indoor $PM_{2.5-10}$ and $PM_{2.5}$ in ventilated schools located in Gaza Strip, Palestine. The types of model used are a combination of multivariate statistical methods, including multiple linear regression (MLR) and feedforward back propagation (FFBP). Samples were collected by using hand held particulate matter sampler during fall, winter and spring from 2011 to 2012. Statistical results revealed that MLR models agree fairly well with the measured data with reasonable coefficients of determination (R^2) for indoor $PM_{2.5-10}$ and $PM_{2.5}$ during fall, winter and spring, respectively. Then, the accuracy (R^2) of the FFBP model results performed better and showed improvement compared than the MLR analysis in determining indoor $PM_{2.5-10}$ and $PM_{2.5}$. The result of the research showed that the artificial neural network approach

can be capable of accurately modelling indoor air quality in naturally ventilated buildings.

Abdullah et al., (2016) applied two statistical analyses to evaluation for long term PM_{10} concentration forecasting using multiple linear regression (MLR) and principal component regression (PCR) models. Daily observations for PM_{10} in Kuala Terengganu, Malaysia from January 2003 till December 2011 were utilized to forecast PM_{10} concentration levels. MLR and PCR (using PCs input) models were developed and the performance was evaluated using RMSE, NAE and IA. The result of this study by Samsuri et al., (2013) revealed that PCR performed better than MLR due to the implementation of PCA which reduce intricacy and eliminate data multi-collinearity.

Jamaludin, (2016) discussed about improving the prediction model of Multiple Linear Regression (MLR) by combining with the Principle Component Analysis (PCA) to predict future (next day, next two-day and next three-day) of the PM_{10} concentration. The site selection for this study is at Pasir Gudang and Paka. Both of these places are industrialization areas. The analysis used annual hourly observations for PM_{10} concentration in Pasir Gudang and Paka from 2005 until 2009 to predict PM_{10} concentration level. The analysis started with the analysis of descriptive statistics of PM_{10} concentration and weather parameters. Then, Principal Component Analysis (PCA) was used to correlate the PM_{10} concentration and weather parameters. To develop the model of PM_{10} concentration, the Multiple Linear Regression (MLR) and Multiple Linear Regression (MLR) by combining with the Principle Component Analysis (PCA) was applied. Next, the performance indicators were used for validation the model which are two accuracy measures which are Prediction Accuracy (PA) and Coefficient of Determination (R^2) then for the error measurement are Normalized Absolute Error (NAE), Mean Absolute Error and Root Mean Square Error (RMSE).

The result of this study byJamaludin, (2016) shows that the modelling of MLR-PCA is the better compared to MLR modelling.

Sayegh et al., (2014) compared Multiple Linear Regression Model (MLRM), Quantile Regression Model (QRM), Generalised Additive Model (GAM), and Boosted Regression Trees 1-way (BRT1) and 2-way (BRT2) to predict PM_{10} concentration. Several meteorological parameters and chemical species measured during 2012 are used as covariates in the models. Various statistical metrics, including the Mean Bias Error (MBE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), the fraction of prediction within a Factor of Two (FACT2), correlation coefficient (R), and Index of Agreement (IA) are calculated to compare the predictive performance of the models. The result shown the statistical model QRM showed better performance in predicting hourly PM_{10} concentrations.

2.4 Bayesian Approach

In assembling information for Bayesian analysis, data collected in the traditional manner in supplemented with prior knowledge. In theory, the existing information about the data (prior distribution) combined with the probability that the data can be generated by the model (likelihood) will result in the set of belief or new theory of the parameter when the data is taken into consideration (posterior).

Theoretically, 'prior' is the best set of prior information about the data. Often, if there is no such knowledge, the arbitrary distributions with a large variance are adopted to reflect this prior ignorance (Coles et al., 2003). This is called non-informative prior. The Bayesian approach specified the probability model for the observed data and the unknown parameter. Given the observed data, Bayes' Theorem (likelihood) is used to

determine the posterior distribution of the parameters. Posterior distribution recapitulates the uncertainly of the parameters based on the prior information and the observed data and it is used for future inferences and decision related to the parameter (Lu Fang, 2003 and Kruschke, 2010). In short, the posterior is proportional to prior multiplied by the likelihood of the distribution.

2.4.1 Bayesian Model Averaging Approach

Duan et al., (2007) studied the use of Bayesian model averaging (BMA) scheme to develop more skilful and reliable probabilistic hydrologic predictions from multiple competing predictions made by several hydrologic models. BMA is a statistical procedure that infers consensus predictions by weighing individual predictions based on their probabilistic likelihood measures, with the better performing predictions receiving higher weights than the worse performing ones. Furthermore, BMA provides a more reliable description of the total predictive uncertainty than the original ensemble, leading to a sharper and better calibrated probability density function (PDF) for the probabilistic predictions. In this study, a nine-member ensemble of hydrologic predictions was used to test and evaluate the BMA scheme. This ensemble was generated by calibrating three different hydrologic models using three distinct objective functions. These objective functions were chosen in a way that forces the models to capture certain aspects of the hydrograph well (e.g., peaks, mid-flows and low flows). Two sets of numerical experiments were carried out on three test basins in the US to explore the best way of using the BMA scheme. In the first set, a single set of BMA weights was computed to obtain BMA predictions, while the second set employed multiple sets of weights, with distinct sets corresponding to different flow intervals. In both sets, the stream flow values were transformed using Box–Cox transformation to

ensure that the probability distribution of the prediction errors is approximately Gaussian. A split sample approach was used to obtain and validate the BMA predictions. The test results showed that BMA scheme has the advantage of generating more skilful and equally reliable probabilistic predictions than original ensemble. The result by Duan et al., (2007) shows the performance of the expected BMA predictions in terms of daily root mean square error (DRMS) and daily absolute mean error (DAM) is generally superior to that of the best individual predictions. Furthermore, the BMA predictions employing multiple sets of weights are generally better than those using single set of weights.

Zhang et al., (2009) has done the study a Genetic Algorithms (GA) and Bayesian Model Averaging (BMA) that were used to simultaneously conduct calibration and uncertainty analysis for the Soil and Water Assessment Tool (SWAT). In this combined method, several SWAT models with different structures are first selected with next GA is used to calibrate each model using observed stream flow data. Finally, BMA is applied to combine the ensemble predictions and provide uncertainty interval estimation. This method was tested in two contrasting basins, the Little River Experimental Basin in Georgia, USA, and the Yellow River Headwater Basin in China. The results obtained in the two case studies show that this combined method can provide deterministic predictions better than or comparable to the best calibrated model using GA. The 66.7% and 90% uncertainty intervals estimated by this method were analyzed. The differences between the percentage of coverage of observations and the corresponding expected coverage percentage are within 10% for both calibration and validation periods in these two test basins. This combined methodology provides a practical and flexible tool to attain reliable deterministic simulation and uncertainty analysis of SWAT.

Bayesian Model Averaging (BMA) has recently been proposed as a method for statistical post processing of forecast ensembles from numerical weather prediction models (Vrugt et al., 2006). The BMA predictive probability density function (PDF) of any weather quantity of interest is a weighted average of PDFs centered on the bias-corrected forecasts from a set of different models. However, current applications of BMA calibrate the forecast specific PDFs by optimizing a single measure of predictive skill. They proposed a multi-criteria formulation for post processing of forecast ensembles such as multi-criteria framework are implemented in different diagnostic measures to reflect different but complementary metrics of forecast skill, and uses a numerical algorithm to solve for the Pareto set of parameters that have consistently good performance across multiple performance. Two illustrative case studies using 48-hour ensemble data of surface temperature and sea level pressure, and multi-model seasonal forecasts of temperature, show that a multi-criteria formulation provides a more appealing basis for selecting the appropriate BMA model.

Raftery et al., (2005) proposed a statistical method for post processing ensembles based on Bayesian model averaging (BMA) which is a standard method for combining predictive distributions from different sources. The BMA predictive probability density function (PDF) of any quantity of interest is a weighted average of PDFs centered on the individual bias-corrected forecasts, where the weights are equal to posterior probabilities of the models generating the forecasts and reflect the models' relative contributions to predictive skill over the training period. The BMA weights can be used to assess the usefulness of ensemble members, and this can be used as a basis for selecting ensemble members; this can be useful given the cost of running large ensembles. The BMA PDF can be represented as an unweighted ensemble of any desired size, by simulating from the BMA predictive distribution. The BMA predictive

variance can be decomposed into two components, one corresponding to the between-forecast variability, and the second to the within-forecast variability. Predictive PDFs or intervals based solely on the ensemble spread incorporate the first component but not the second. Thus BMA provides a theoretical explanation of the tendency of ensembles to exhibit a spread-error correlation but yet be under dispersive. The method was applied to 48-h forecasts of surface temperature in the Pacific Northwest in January–June 2000. The predictive PDFs were much better calibrated than the raw ensemble, and the BMA forecasts were sharp in that 90% BMA prediction intervals were 66% shorter on average than those produced by sample climatology. As a by-product, BMA yields a deterministic point forecast, and this had root-mean-square errors 7% lower than the best of the ensemble members and 8% lower than the ensemble mean. Similar results were obtained for forecasts of sea level pressure. Simulation experiments show that BMA performs reasonably well when the underlying ensemble is calibrated, or even over dispersed.

Patrik et al. (2010) applied the Bayesian Model Averaging (BMA) approach which accounts for model uncertainty by combining information from all possible models of Generalized Additive Models (GAMs) and Natural Cubic Splines (NS) were used. The research conducted a sensitivity analysis with simulation studies for Bayesian Model Averaging with different calibrated parameters contained in the posterior model probabilities. Results showed that the posterior means of the relative risks and 95% posterior probability intervals were close to each other under different choices of the prior distributions. Simulations results were consistent with these findings. It was also found that using lag variables in the model when there is only same day effect, may underestimate the relative risk attributed to the same day effect.

Yuen (2010) studied the recent developments of Bayesian model class selection and applications in civil engineering. Bayesian model class selection has attracted substantial interest in recent years for selecting the most plausible/suitable class of models based on system input–output data. The Bayesian approach provides a quantitative expression of a principle of model parsimony or of Ockham's razor which in engineering applications can be stated as simpler models are to be preferred over unnecessarily complicated ones. Some recent developments are reviewed for this research. Linear and nonlinear regression problems are considered in detail. Bayesian model class selection is particularly useful for regression problems since the regression formula order is difficult to be determined solely by physics due to its empirical nature. In this research, some recent developments and civil engineering applications of Bayesian model class selection are reviewed. It is illustrated for globally identifiable case with asymptotic expansion and for general case with the transitional Markov chain Monte Carlo (MCMC) method. Bayesian model class selection has attracted substantial interest in recent years for selecting the most plausible/suitable class of models among some specified model classes, based on system measurements. Asymptotic expansion and Monte Carlo method can be used to compute the evidence of a model class. Special cases of linear and non-linear regression formula are considered and Bayesian model class selection is particularly useful in this case.

2.5 Summary of Literature Review

Based on Literature Review, the modelling of PM_{10} concentration was conducted using various statistical analyses. However, the previous researches have never conducted modelling PM_{10} concentration or in the field of air pollution using Bayesian model averaging. There are some researches that used Bayesian model averaging in the water, soil, air in environmental engineering field as discussed in this chapter. Most of these studies indicated that Bayesian model averaging modelling gives good and appropriate result of threshold classification. This show the Bayesian model averaging analysis is also suitable to be used to predict the PM_{10} concentration.

CHAPTER 3

METHODOLOGY

3.1 Introduction

This chapter will review the procedure that need to be processed in order to achieve all the objectives in Chapter 1. Generally, the methods that are involved to analyse the parameter are using descriptive analysis, statistical modelling, and performance indicator are explained in this chapter. The flowchart of the methodology for this study is summarized in Figure 3.1. There are four stages to achieve all objectives.

The first stage in this study is called pre-stage. The pre-stage in this study includes the data analysis, parameter selection and monitoring data screening. After the parameter selection, the next stage is to determine the first objective which is to determine the descriptive analysis using the software (SPSS).

Therefore the next step is to analyze data using Multiple Linear Regression (MLR) model and Bayesian Model Averaging (BMA) to predict the PM_{10} concentration for the next day (Day 1) to achieve the second objective. Then, the last stage of this study is to obtain the best model to predict concentration of PM_{10} using performance indicators which are error measures (mean bias error (MBE), mean absolute error (MAE) and mean absolute percentage error (MAPE)) and accuracy measures (index of agreement (IA), prediction accuracy (PA) and FACT2) to achieve the third objective.

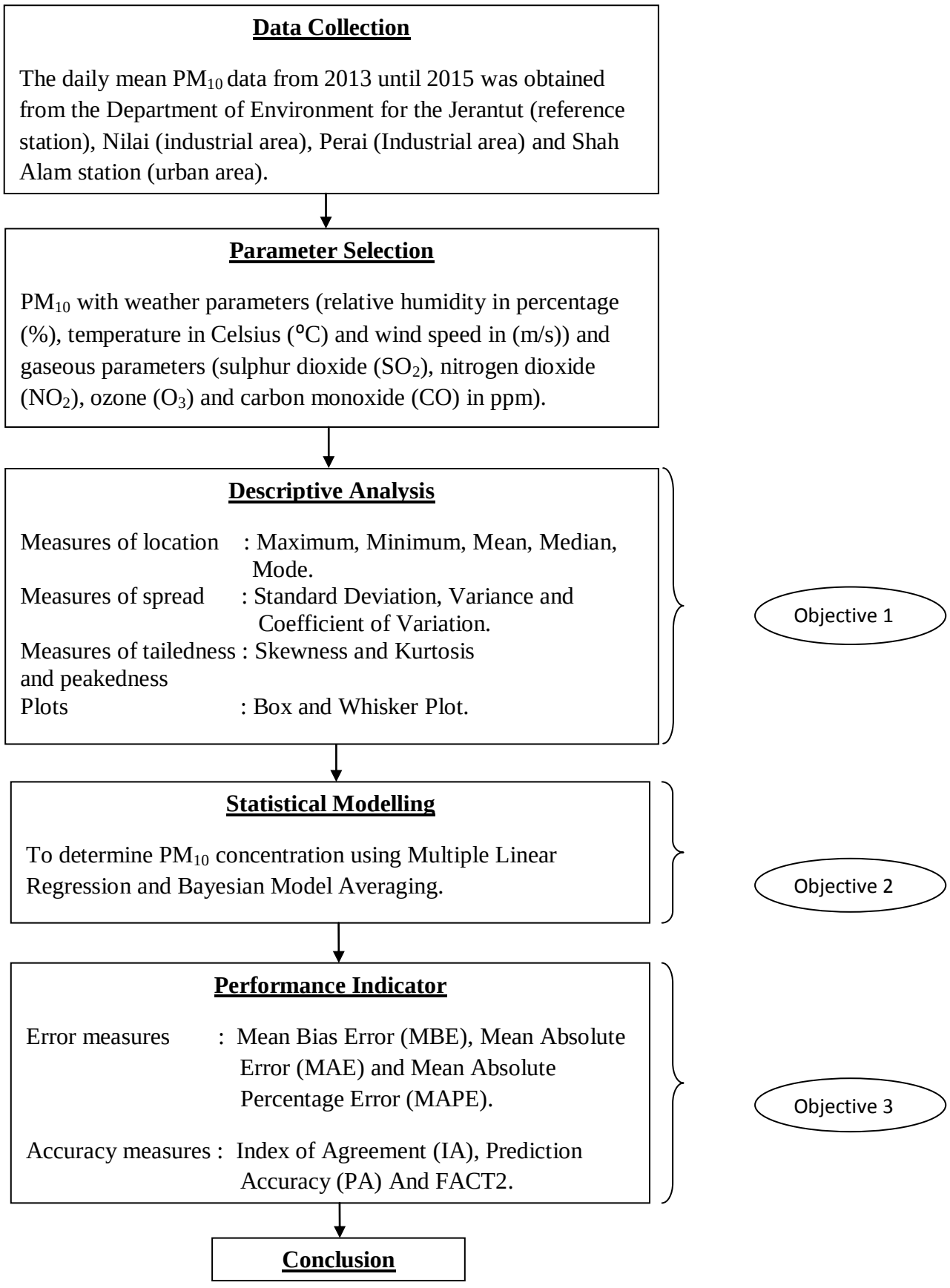


Figure 3.1 : Flow chart for study procedure

3.2 Data Collection

From this study, four stations were selected in monitoring air quality which are located in Jerantut (reference station), Nilai (industrial area), Seberang Jaya (Sub-urban area) and Shah Alam station (urban area). Table 3.1 described the detailed location of the monitoring sites. The parameter selection is weather parameters (relative humidity in percentage, temperature in Celsius and wind speed meter per second) and gaseous parameters (sulphur dioxide (SO_2), nitrogen dioxide (NO_2), ozone (O_3) and carbon monoxide (CO) in part per million).

Table 3.1 : Detailed Location of The Monitoring Sites

Site Id	**Monitoring Station**	**Site Location**	**Category**	**Coordinate**
CA0010	Nilai	Taman Semarak (Phase II), Nilai.	Industrial	02º 49.3001' N 101º 48.6894' E
CA0009	Seberang Jaya	Sek.Keb. Seberang Jaya II, Seberang Jaya, Penang.	Sub Urban	5º 23' 28.00" N 100º 23' 12.45" E
CA0025	Shah Alam	Sek. Keb. Raja Muda, Shah Alam.	Urban	03º 04.4404' N 101º 30.3537' E
CA0007	Jerantut	Meteorology Monitoring Station Batu Embun, Jerantut	Reference	03º 58.2482' N 102º 20.8891'E

The data were collected and monitored by Alam Sekitar Malaysia Sdn.Bhd. (ASMA), authorized agency for Department of Environment (DoE) and the equipment used by ASMA to monitor air quality data was Beta Attenuation Mass Monitor (Azid et al., 2014). The duration of the study in 2013 until 2015 was taken based on the daily mean data of PM_{10}.

In Malaysia, the Department of Environment, DoE (2013), monitors the country's ambient air quality through a network of 52 stations. The continuous Air Quality Monitoring (CAQM) stations are divided into four categories which is industrial, background, urban and sub-urban as shown in Figure 3.2. All the strategically located monitoring stations in urban, sub- urban and industrial areas would record any significant changes in air quality that might have detrimental effects on the global environment, crops and human health (DoE, 2013).

Figure 3.2 : Continuous Air Quality Monitoring Station (CAQMS) in Malaysia.

(Source : DoE,2013)