PREDICTION OF PM_{10} CONCENTRATION USING MULTIPLE LINEAR REGRESSION AND SUPPORT VECTOR MACHINE

By

MASEZATTI BINTI ZAILAN

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 2018

ACKNOWLEDGEMENT

Alhamdulillah and thanks to Allah s.w.t for giving me the knowledge and strength to complete this dissertation after I pass through all the challenges while completing it. I would like to express my deepest appreciation to my supervisor, Professor Ahmad Shukri Yahaya for giving me the opportunity to study about this research. All his guidance and advice are valuable for me to complete this research. Without his guidance and persistent help, this dissertation would not have been complete.

Besides that, I would like to thank all lectures of School of Civil Engineering who gave me detail explanations, briefings and guidelines on how to do this research during FYP class for me to complete this dissertation. Futhermore, deepest appreciation and special thanks also goes to all my friends who are involved in helping me to pass through all the difficulties and challenges while completing this dissertation.

Last but not least, I would like to thank my dear family for always supporting me, give advice and moral support for me to complete my research and pass through all the challenges and difficulties while completing this degree studies in Universiti Sains Malaysia.

ABSTRAK

Zarah berdiameter aerodinamik kurang daripada 10μm (PM_{10}) adalah salah satu bahan pencemar udara yang boleh memberi kesan negatif kepada kesihatan terhadap manusia dan alam sekitar. Tujuan kajian ini adalah untuk meramalkan kepekatan bahan zarah untuk hari berikutnya (PM_{10D1}) dengan menggunakan model Regrasi Linear Berganda (MLR) dan Mesin Vektor Sokongan (SVM). Parameter meteorologi dan gas yang digunakan dalam kajian ini adalah zarah terampai hari ini (PM_{10D0}), kelajuan angin (WS), suhu (TEMP), kelembapan relatif (RH), sulfur dioksida (SO_2), nitrogen dioksida (NO_2), ozon (O_3) dan karbon monoksida (CO). Data purata harian yang digunakan dalam kajian ini dibahagikan kepada data latihan (70%) dan data pengesahan (30%) dan digunakan dari tahun 2013 hingga 2015. Empat stesen pengawasan telah dipilih dalam kajian ini untuk meramalkan kepekatan PM_{10} untuk hari seterusnya (PM_{10D1}) iaitu Jerantut yang bertindak sebagai stesen latar belakang, Nilai (kawasan perindustrian), Seberang Jaya (kawasan sub urban) dan Shah Alam (kawasan bandar). Hasil keseluruhan data yang diperolehi dari kajian ini menunjukkan bahawa stesen pemantauan Nilai menyumbang kepekatan PM_{10} paling tinggi berbanding stesen pemantauan yang lain. Ini menunjukkan bahawa Nilai adalah kawasan yang lebih tercemar kerana ia dikenali sebagai kawasan yang sangat maju. Hasilnya menunjukkan bahawa Regrasi Linear Berganda (MLR) adalah model terbaik dalam meramalkan kepekatan PM_{10} untuk hari berikutnya berbanding dengan model Mesin Sokongan Vektor (SVM).

ABSTRACT

Particulate matter with an aerodynamic diameter less than 10μm (PM_{10}) is one of the most air pollutants that can give negative effect on human health and environment. The purpose of this research is to predict the particulate matter concentration for the next day (PM_{10D1}) by using Multiple Linear Regression (MLR) and Support Vector Machine (SVM) models. The meteorological and gaseous parameters that are used in this study are particulate matter for today (PM_{10D0}), wind speed (WS), temperature (TEMP), relative humidity (RH), sulphur dioxide (SO_2), nitrogen dioxide (NO_2), ozone (O_3) and carbon monoxide (CO). The daily mean data that are used in this study are divided into training data (70%) and validation data (30%) and are used from 2013 until 2015. Four monitoring stations were selected in this study to predict the PM_{10} concentration for the next day (PM_{10D1}) which are Jerantut which act as background station, Nilai (industrial area), Seberang Jaya (sub-urban area) and Shah Alam (urban area). The results of overall data that are obtained from this study has shown that Nilai monitoring stations contributed the highest mean value of PM_{10} concentration compared to the other monitoring stations. This indicated that Nilai is a more polluted area as it is known as a highly industrialised area. The results shows that Multiple Linear Regression (MLR) is the best model in predicting PM_{10} concentration for the next day compared to Support Vector Machine (SVM) model.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

API	Air Pollution Index
CO	Carbon monoxide
DoE	Department of Environment
MAAQS	Malaysian Ambient Air Quality Standard
MLR	Multiple Linear Regression
NO_2	Nitrogen dioxide
O_3	Ozone
PM_{10}	Particulate Matter with an aerodynamic diameter less than 10μm
PM_{10D0}	Particulate matter for today
PM_{10D1}	Particulate matter for the next day
RH	Relative humidity
SO_2	Sulphur dioxide
SPSS	Statistical Package and Services Solution
SVM	Support Vector Machine
TEMP	Temperature
WS	Wind speed

CHAPTER 1

INTRODUCTION

1.1 Background

Air pollution are contamination of harmful gases and particles in the atmosphere that harm organisms and affect climate. It is one of the most serious environmental problems in the world and will pose enormous health risks to billions of people around the world every day as it will affect the human lives, contributing to heart diseases, lung failure, respiratory disease, cancer and many other fatal human diseases. Most often, human activities are the major caused of air pollution to occur as it contribute to global pollution of the air. Such human activities are construction, transportation, industrial work, fuel burning processes and power stations. However, natural processes such as volcanic eruptions and forest fires may also pollute the air but their occurrence is rare compared to human activities (Sotomayor-Olmedo et al., 2013).

Besides that, air pollution may be described as contamination of the atmosphere by gaseous, liquid, or solid wastes or by-products that can endanger human health and welfare of plants and animals, attack materials, reduce visibility or produce undesirable odors (Li et al., 2018). According to the World Health Organization (WHO), outdoor air pollution is classified into four main categories which are particulate matter, ozone, nitrogen dioxide, and sulphur dioxide. This air pollutants differ in their chemical

composition, reaction properties, emission, time of disintegration and ability to diffuse in long or short distances.

PM_{10} is particulate matter with an aerodynamic diameter less than 10µm, which can be a suspension of solid, liquid or a combination of solid and liquid particles in the air. Particulate matter is composed of many components such as acids, organic chemicals, metals, soil, wood, and dust. Particulate matter pollution is classified primarily by particle size, with smaller particles being the cause of most health problems. The types of health problems are related to the relative size of the contaminant. Particles originate from a variety of stationary and mobile sources and may be directly emitted (primary emissions) or formed in the atmosphere (secondary emissions) by transformation of gaseous emissions (De Rooij et al., 2017).

Particulate matter, PM_{10} are emitted from motor vehicle exhausts, heat and power generation plants, industrial process and open burning activities. However, the most sources are industry and heavy traffic.

Air pollution index (API) is a numerical scale used for reporting day to day air quality with regard to human health and the environment. The daily results of the index are used to convey to the public an estimate of air pollution level. An increase in air quality index signifies increased air pollution and severe threats to human health. API calculations focus on major air pollutants including particulate matter, ground-level ozone, sulfur dioxide (SO_2), nitrogen dioxide (NO_2), and carbon monoxide (CO). Particulate matter and ozone pollutants pose the highest risks to human health and

the environment. 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 range of the API value is shown in Table 1.1.

Table 1-1 : Malaysia Air Pollution Index (API)
(Source : Department of Environment, 2017)

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

As the main pollutant involved in this research is particulate matter with an aerodynamic diameter less than 10µm, the Department of Environment has used new standard in the Malaysian Ambient Air Quality Guidelines (MAAQG) Interim Target 1 (IT-1) in 2015 which state that the average threshold limit concentration for particulate matter, PM_{10} is at 150 µg/m³ for a 24-hours period and 50 µg/m³ for annual. If the concentrations exceed the limit values, health problems may occur to people. Malaysian Ambient Air Quality Standard (MAAQS) were issued and target values for annual and daily mean mass concentrations for various of air pollutant.

Table 1-2 : Malaysian Ambient Air Quality Standard (MAAQS)
(Source : Department of Environment, 2017)

Pollutants	Averaging Time	Ambient Air Quality Standard		
		IT-1 (2015) µg/m³	IT-2 (2018) µg/m³	Standard (2020) µg/m³
Particulate Matter with the size of less than 10 micron (PM_{10})	1 Year	50	45	40
	24 Hour	150	120	100
Particulate Matter with the size of less than 2.5 micron ($PM_{2.5}$)	1 Year	35	25	15
	24 Hour	75	50	35
Sulfur Dioxide (SO_2)	1 Hour	350	300	250
	24 Hour	105	90	80
Nitrogen Dioxide (NO_2)	1 Hour	320	300	280
	24 Hour	75	75	70
Ground Level Ozone (O_3)	1 Hour	200	200	180
	8 Hour	120	120	100
*Carbon Monoxide (CO)	1 Hour	35	35	30
	8 Hour	10	10	10

*mg/m³

1.2 Problem Statement

Particulate Matter (PM) also known as particle pollution, is a complex mixture of extremely small particles and liquid droplets that get into the air. Once inhaled, these particles can affect the heart and lungs and cause serious health effects. These particles originate from a variety of sources such as industrial sources, motor vehicles, construction and road dust and they are formed in the atmosphere by transformation of gaseous emissions. Particulate Matter (PM) is one of the six criteria pollutants and the most important in terms of adverse effects on human health.

Components of particulate matter (PM) include finely divided solids or liquids such as dust, fly ash, soot, smoke, aerosols, fumes, mists and condensing vapour that can be suspended in the air for extended periods of time. There are two types of particulate matter emitted, which are primary emissions and secondary emissions. Primary particulate matter sources are derived from both human and natural activities. A significant portion of particulate matter sources is generated from a variety of human activity. These types of activities include agricultural operations, industrial processes, combustion of wood and fossil fuels, and construction activities, and entrainment of road dust into the air. Natural sources also contribute to the overall PM_{10} problem. These include windblown dust and wildfires (Maenhaut et al., 2016).

Secondary PM_{10} sources directly emit air contaminants into the atmosphere that form or help form PM_{10}. Hence, these pollutants are considered precursors to PM_{10} formation. These secondary pollutants include sulphur dioxide, nitrogen dioxide, ozone,

and ammonia. Control measures that reduce PM_{10} precursor emissions tend to have a beneficial impact on ambient PM_{10} levels (Sotomayor-Olmedo et al., 2013).

Many scientific studies have been published that long exposure to wood smoke and PM_{10} will contributes to serious public health effects, such as higher instances of asthma, decreased lung function in children, and increased morbidity and mortality rates (De Rooij et al., 2017). United Stated Environmental Protection Agency (2017) also reported PM_{10} can also cause coughing, difficulty in breathing, chronic bronchitis, irregular heartbeat, non-fatal heart attacks and some cancer.

1.3 Objectives

The objective of this study are :

1. To determine the characteristics of the PM_{10} concentration 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 Support Vector Machine.
3. To determine the best model to predict the PM_{10} concentration.

1.4 Scope of Work

This research used data obtained for four monitoring stations which are Shah Alam (urban area), Nilai (industrial area), Seberang Jaya (Sub-urban area) and Jerantut (reference station) for three years period from 2013 until 3015.

The parameters selection are weather parameters and gaseous parameters. Weather parameters are relative humidity in percentage, temperature in Celcius and wind speed in meter per second while gaseous parameters are sulphur dioxide (SO_2), nitrogen dioxide (NO_2), ozone (O_3) and carbon monoxide (CO) and all are measured in parts per million (ppm). The statistical modelling that will be used in this study are Multiple Linear Regression and Support Vector Machine.

1.5 Outline of Research

This thesis consists of five chapters and the brief outline for every chapter is as follows :

Chapter 1 is an introduction of air pollution, the sources of air pollution in Malaysia and the impact of air pollution. This chapter also discussed about the problem statement, objectives, scope of work and outline of research.

Chapter 2 discussed the literature review for related previous study of particulate matter (PM_{10}) and the summary of the application of statistical analysis. This chapter also include the explanation of multiple linear regression and support vector machine that has

been used by other researches on worldwide in terms of environmental engineering studies.

Chapter 3 described the methodology that were used in this study. Parameter selection, descriptive analysis and two models used that are multiple linear regression and support vector machine were explained in this chapter. Besides that, performance indicators which are error measures and accuracy measures were also discussed in order to obtain the best model.

Chapter 4 discussed about the results that were obtained by using Statistical Package and Service Solution for multiple linear regression and R Studio for support vector machine and performance indicators in order to determine the best model to predict PM_{10} concentration.

Chapter 5 concludes about this study and provide some recommendations for future study.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter will discussed in detail about the multiple linear regression and support vector machine that has been used by the previous researches for the purpose of prediction of air pollutant in the environment field and also for modelling purpose. Moreover, particulate matter and previous studies was also reviewed as an additional knowledge in this study.

2.2 Sources of Particulate Matter (PM_{10}) Concentration and Its Effect

Particulate matter is the major pollutant in ambient air especially in urban areas. It is tiny particles of solid or semi-solid material that are found in the atmosphere and usually originates from natural and human sources such as road dust, motor vehicle, industrial activities, domestic activities, and open burning (Shaadan et al., 2015).

Particulate matter with an aerodynamic diameter of less than 10 μm (PM_{10}) has been identified as a major atmospheric pollutant in major cities in Southeast Asia, including in Klang Valley, Malaysia. It is believed that particulate matter will give an effect on human respiratory system which in turn may result in chronic health disease and asthma. In Malaysia, PM_{10} is one of the major air pollutants and is under the computation of Malaysian Air Pollution Index (MAPI) (Juneng et al., 2011).

Futhermore, previous studies have found that human respiratory health is most affected by particulate matter and have been associated with increasing mortality and morbidity (Rashid et al., 2014). The distribution of particulate matter is influenced by wind and geographical factors. This is because particulate matter are able to spread far away due to their long atmospheric lifetime. 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. Monitoring data and studies on ambient air quality show that some of the air pollutants in several large cities are increasing with time and are not always at acceptable levels according to the MAAQG (Mohamed et al., 2011).

Most often, human activities are the major caused of air pollution to occur as it contribute to global pollution of the air. Such human activities are construction, transportation, industrial work, fuel burning processes and power stations. However, natural processes such as volcanic eruptions and forest fires may also pollute the air but their occurrence is rare compared to human activities (Sullivan et al., 2018).

In Malaysia, the concentration of particulate matter is usually influenced by the southwest monsoon wind and the occurrence of biomass burning. While during the normal period which is without haze, the level of particulate matter is mostly influenced by motor vehicles and industry (Samsuddin et al., 2018).

According to scientific and previous studies, long exposure to particulate matter (PM_{10}) can cause premature death, adverse cardiovascular effects, asthma attacks, heart attacks, strokes, lung cancer and increased respiratory symptoms such as coughing, wheezing, and shortness of breath. This will increased the number of hospital admissions and emergency department visits by year (Samsuddin et al., 2018). Older people, children, and individual with cardiovascular disease such as asthma and congestive heart disease are particularly at risk for experiencing adverse health effects that are related to particulate matter (PM_{10}) exposure.

2.3 Air Quality Data

The air quality data used for this study were obtained from the Air Quality Division of the Department of Environment, Malaysia (DoE) through long-term monitoring by Department of Environment. Jerantut, Pahang Station has been established as a background air monitoring station by the Department of the Environment (DOE), Malaysia. It is monitored continuously and manually to detect any changes in the ambient air quality status that may cause harm to human health and the environment.

The Department of Environment (DOE) monitors the country's ambient air quality through a network of 52 stations. Generally, it is strategically located in residential, traffic and industrial areas to detect any significant change in the air quality which may be harmful to human health and the environment. This 52 monitoring stations are categorized into four categories which is background station, urban, sub-urban, and industrial station.

There are five major air pollutants that are listed by Department of Environment Malaysia (DOE) for the research which are sulphur dioxide (SO_2), nitrogen dioxide (NO_2), carbon monoxide (CO), and particulate matter with aerodynamic diameter of 10 μm (PM_{10}) and ground level ozone (O_3) (Mabahwi et al., 2015) while there are three meteorological parameter that are used which are relative humidity, temperature and wind speed.

2.4 Parameters

Parameters that are used in this study are particulate matter with an aerodynamic diameter less than 10μm (PM_{10}), wind speed (WS), temperature (TEMP), relative humidity (RH), sulphur dioxide (SO_2), nitrogen dioxide (NO_2), ozone (O_3) and carbon monoxide (CO). All this parameters will be used in order to determine the air pollutant value and indicated the air quality status at particular area.

2.4.1 Sulphur Dioxide (SO_2)

Sulphur dioxide is one of the gases with atmospheric acidifying effect. It is a highly reactive gas and flammable with a pungent odour which can irritates the eyes and the respiratory system. It appears mainly from burning sulphur fossil fuels (coal, oil) and the exhaust of vehicles' engines which can give significant negative effects on the human health and environment (Barbulescu and Barbes, 2017).

2.4.2 Nitrogen Dioxide (NO_2)

Nitrogen dioxide is a nasty-smelling gas and it is formed naturally in the atmosphere by lightning, plants, soil and water. The sources of nitrogen dioxide comes

from motor vehicle exhaust, petrol and metal refining, electricity generation from coal-fired power stations, manufacturing industries and food processing. Nitrogen dioxide is one of the gaseous pollutant generated from fossil fuel combustion which can contributes to the formation of photochemical smog and have significant negative impacts on human health (Gaffin et al., 2017).

2.4.3 Ozone (O_3)

Ozone is an allotrope of oxygen in which the molecule is composed of three oxygen atoms. It is known as one of the most powerful oxidizing agents which can be beneficial or harmful depending on where it is found in the atmosphere. Ozone gas in the stratosphere act as a protective barrier against harmful radiation from the sun by absorbing ultraviolet radiation. However, ozone in the troposphere is considered as a pollutant and is harmful to human health and environment (Olewnik-Kruszkowska et al., 2016).

2.4.4 Carbon Monoxide (CO)

Carbon monoxide is a colourless, odourless and tasteless gaseous pollutant that is generally produced by incomplete combustion of fossil fuels such as motor vehicles. Major sources of carbon monoxide include the oxidation of methane (CH_4), non-methane hydrocarbons (NMHC) and direct emissions from burning biomass and human activities (fossil fuel and biofuel use). Study has shown that carbon monoxide have an adverse health effects if exposed in a long or short term as it is associated with mortality and morbidity from cardiovascular diseases (Liu et al., 2018)

2.5 Multiple Linear Regression (MLR)

Multiple Linear Regression (MLR) models have been developed for forecasting urban air quality. This models are commonly applicable for predicting PM_{10} concentration distributions in urban areas except for some specific model types, such as the land use regression models and for evaluating air pollution scenarios for many years. Multiple Linear Regression (MLR) model establish the relationships between dependent variable and independent variables for predicting air pollutant concentrations with remarkable success (Niazian et al., 2018). The difference of multiple linear regression (MLR) model with the other model is it provide significant improvements of the forecasting ability to predict daily concentration of air pollutant.

Stadlober et al., (2008) show how multiple linear regression models combining information of the present day with meteorological forecasts of the next day can help forecasting daily PM_{10} concentrations for sites located in the three cities. Juneng et al., 2011 applied multiple linear regression method to establish statistical relationships between the summer monsoon PM_{10} concentrations and that of the weather elements both on local and synoptic scales.

Pires et al. (2008) compared the performance of five linear models to predict the daily average PM_{10} concentrations. The linear models implemented were multiple linear regression (MLR), principal component regression (PCR), independent component regression (ICR), quantile regression (QR), and partial least squares regression (PLSR).

These models were applied to two datasets with different sizes (first set with data from three years and the second with data from six months). The results showed that the prediction of the daily mean PM_{10} concentrations was more efficient when using ICR for the smaller dataset and PLSR for the larger dataset. As presented in the referred studies, the structure definition is an important step for the model success, being followed by the optimization of its parameters. All of these studies lead to the conclusion that it is practically impossible to establish a ranking of models for predicting the daily average PM_{10} concentrations on the next day. Thus, due to the high number of variables of different nature involved in this process and the specificities of the monitoring sites, the procedure to be applied is to test simultaneously different model structures to find the one/ones with the best performances (Pires et al., 2010).

Ul-Saufie et al., (2013) defined Multiple regression analysis (MRA) is a popular methodology to express response of a dependent variable of several independent variables. In spite of its success, MRA presents problems in indentifying the most important contributors when multicollinearity or high correlation between the independent variables in regression equation are present. Multivariate data analysis techniques such as multiple linear regression (MLR) and principle component analysis (PCA) have been proven to be effective tools to study the relationship between voluminous data such as air pollution and meteorological records.

Sousa et al. (2007) used Multiple linear regression (MLR) and feedforward artificial neural network (FANN) were used to predict the next day hourly ozone

concentration using as predictors air pollutant concentrations (NO, NO2 and O3) and meteorological parameters (T, RH and WV). The same models, but based on principal component analysis (PCA), were also used, being referred to as principal component regression (PCR) and feedforward artificial neural network based on PC (PC-FANN), respectively. The results showed that the use of FANN led to more accurate results than linear models (MLR and PCR), due to the account of non-linearities. The application of PC in this model was considered better than using the original data, because it reduced the number of inputs and therefore decreased the model complexity. The performance indexes were similar using both approaches.

2.6 Support Vector Machine (SVM)

Support vector machine (SVM) has been used widely in predicting the particulate matter concentration and any other air pollutant. It is because it shows a good generalization performance for high dimensional data due to its convex optimization problem. The incorporation of prior knowledge about the data leads to a better optimized classifier.

A study by Arampongsanuwat and Meesad (2011) have developed a support vector regression (SVR) model for the PM_{10} forecasting in Bangkok. It have developed to establish the relationships of PM_{10} with meteorological variables including globe radiation, net radiation, air pressure, rainfall, relative humidity, temperature and wind direction. To design the SVR model, some parameters must be determined before

running the particular algorithm. These parameters are error acceptance (ε), constant (C) and kernel specific parameters. In this work, radial basis function was used where σ is the parameter that determines performance in the learning of the kernel function.

Support vector machine (SVM) methods are supervised machine learning algorithms that can be used for regression and classification problems. As a general rule, the SVM regression are also called penalty function with value zero if the predicted value y_i is within a distance of less than ε from the observed value t_i. Another modification to the penalty function is that the output variables falling out of the tube are supplied through two penalizations in the form of slack variables that depend on the position in relation to the tube. For sufficiently small values of C, the soft-margin SVM will behave identically to the hardmargin SVM if the input data are linearly regressable. The effectiveness of SVM models depends on the selection of kernel, the kernel's hyperparameters, and soft margin parameter C. The best combination of these hyperparameters is determined here by the grid search technique or a parameter sweep, which is simply an exhaustive searching through a manually specified subset of the hyperparameter space of a learning algorithm. A grid search algorithm must be guided by some performance metric, typically measured by cross-validation on the training set or evaluation on a held-out validation set. Indeed, each combination of parameters is checked using cross-validation, and the parameters with best cross validation accuracy are picked up as the optimal hyperparameters of the SVM model (Garcia Nieto et al., 2018).

Weizhen et al. (2014) have developed a successive relaxation support vector regress (SOR-SVR) model for the PM_{10} and $PM_{2.5}$ prediction, based on the daily average aerosol optical depth (AOD) and meteorological parameters (atmospheric pressure, relative humidity, air temperature, wind speed), which were all measured in Beijing during the year of 2010-2012. The Gaussian kernel function, as well as the k-fold crosses validation and grid search method, are used in SVR model to obtain the optimal parameters to get a better generalization capability. The result shows that predicted values by the SOR-SVR model agree well with the actual data and have a good generalization ability to predict PM_{10} and $PM_{2.5}$.

Sotomayor-Olmedo et al. (2013) predict pollution model of ozone (O_3), particulate matter (PM_{10}) and nitrogen dioxide (NO_2) using support vector machines and kernel functions which are Gaussian, Polynomial and Spline. The use of an appropriate kernel is the key feature in support vector applications, since it provides the capability of mapping non-linear data into feature spaces that in essence are linear, then an optimization process can be applied as in the linear case. The Gaussian kernel process delivers an estimate for the reliability of the prediction in the form of the variance of the predictive distribution and the analysis can be used to estimate the evidence in favor of a particular choice of covariance function and the covariance or kernel function can be seen as a model of the data. There are other considerations when working with SVMs on regression mode. The most important are Bias Analysis, Free parameters and the quadratic problem. The inclusion of a bias within the kernel function generally leads to a more efficient implementation and a slightly better accuracy model. Conversely the

solutions achieved with an implicit or explicit bias are not the same. This dichotomy emphasizes the difficulties with the interpretation of generalization in high dimensional feature spaces. In this work the explicit bias approach is used.

CHAPTER 3

METHODOLOGY

3.1 Introduction

This chapter will discussed and review in details about the methodology process for this research. Methodology is an important part of this study in order to achieve all the objectives as discussed in Chapter 1. Therefore, research methods and approaches that are applied to obtain the results are also explained in this chapter. Figure 3.1 shows the flowchart of the methodology for this study.

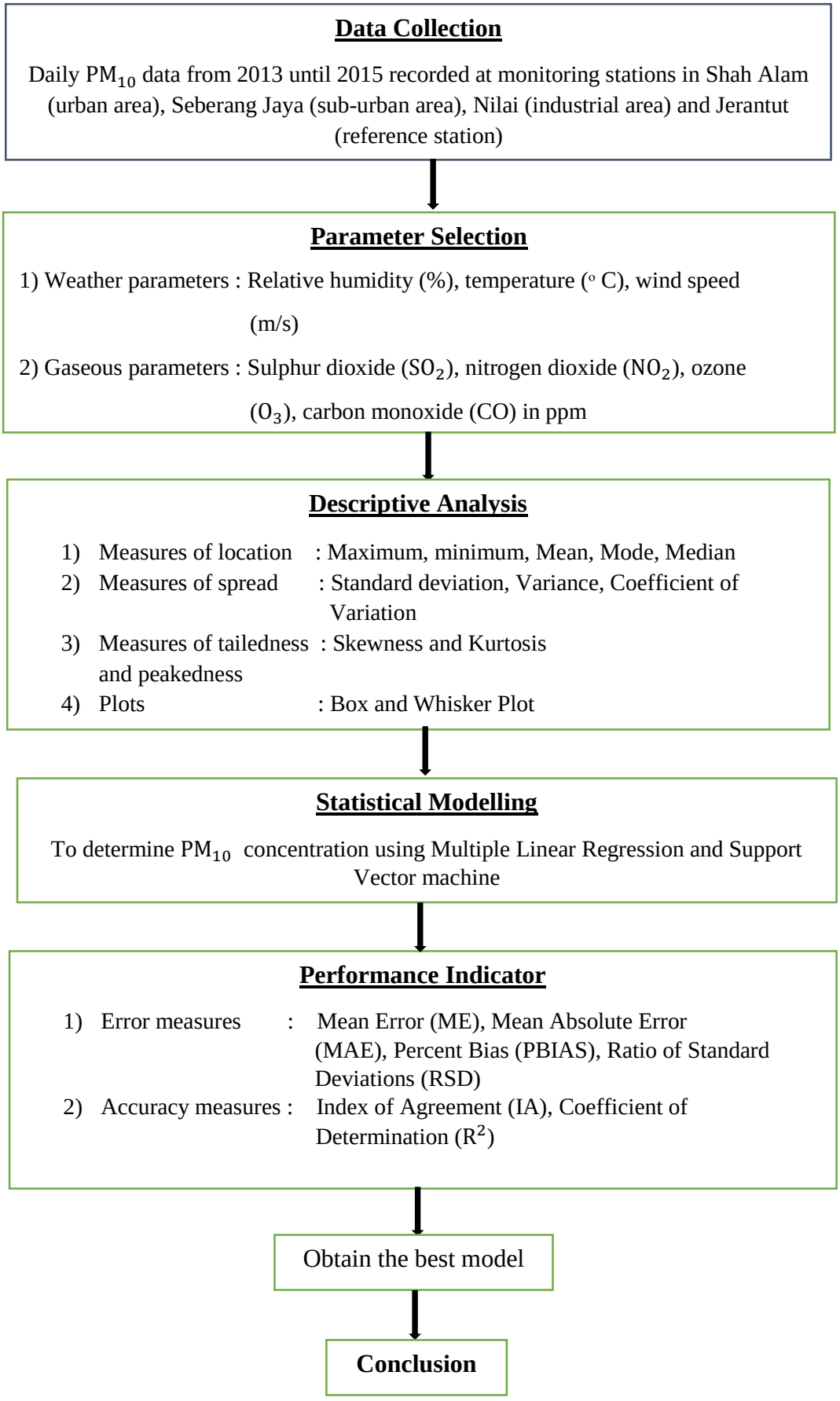

Figure 3-1: Research Flowchart

3.2 Data Collection

The Department of Environment (DOE) continuously monitors ambient air quality to detect any significant change in the air quality which may cause harm to human health and the environment. The air quality data used for this study were obtained from the Air Quality Division of the Department of the Environment, Malaysia, (DOE) through long-term monitoring by a private company, Alam Sekitar Sdn Bhd (ASMA). Four monitoring stations in Malaysia are selected in monitoring air quality which are located in Nilai (industrial area), Seberang Jaya (Sub-urban area), Shah Alam (urban area) and Jerantut (reference station). The data used in this study is daily PM_{10} concentration ($\mu g/m^3$) data obtained from Continuous Air Quality Monitoring (CAQM) stations recorded from 2013 until 2015 in order to gain a better understanding of PM_{10} variability.

The air pollutant parameters used in these studies were particulate matter with a diameter size of below 10 μm (PM_{10}), carbon monoxide (CO), sulphur dioxide (SO_2), nitrogen dioxide (NO_2) and ozone (O_3). In addition, meteorological parameters such as wind speed, temperature, and humidity were also recorded at the stations. This are measured by 52 Continuous Air Quality Monitoring Stations (CAQMS) in Malaysia that are monitored by Alam Sekitar Malaysia (ASMA) which is the authorized agency for DoE. This instrument automatically measure and record the data continuously 24 hours per day by using Beta Attenuation Mass Monitor (BAM-1020) from Met One Instrument, Inc.

3.3 Description of study area

Based on Figure 3.2, four monitoring stations in Malaysia were selected which are located in Nilai (industrial area), Seberang Jaya (Sub-urban area), Shah Alam (urban area) and Jerantut (reference station). The air quality data is used to predict PM_{10} concentration based on weather parameters and gaseous parameters.

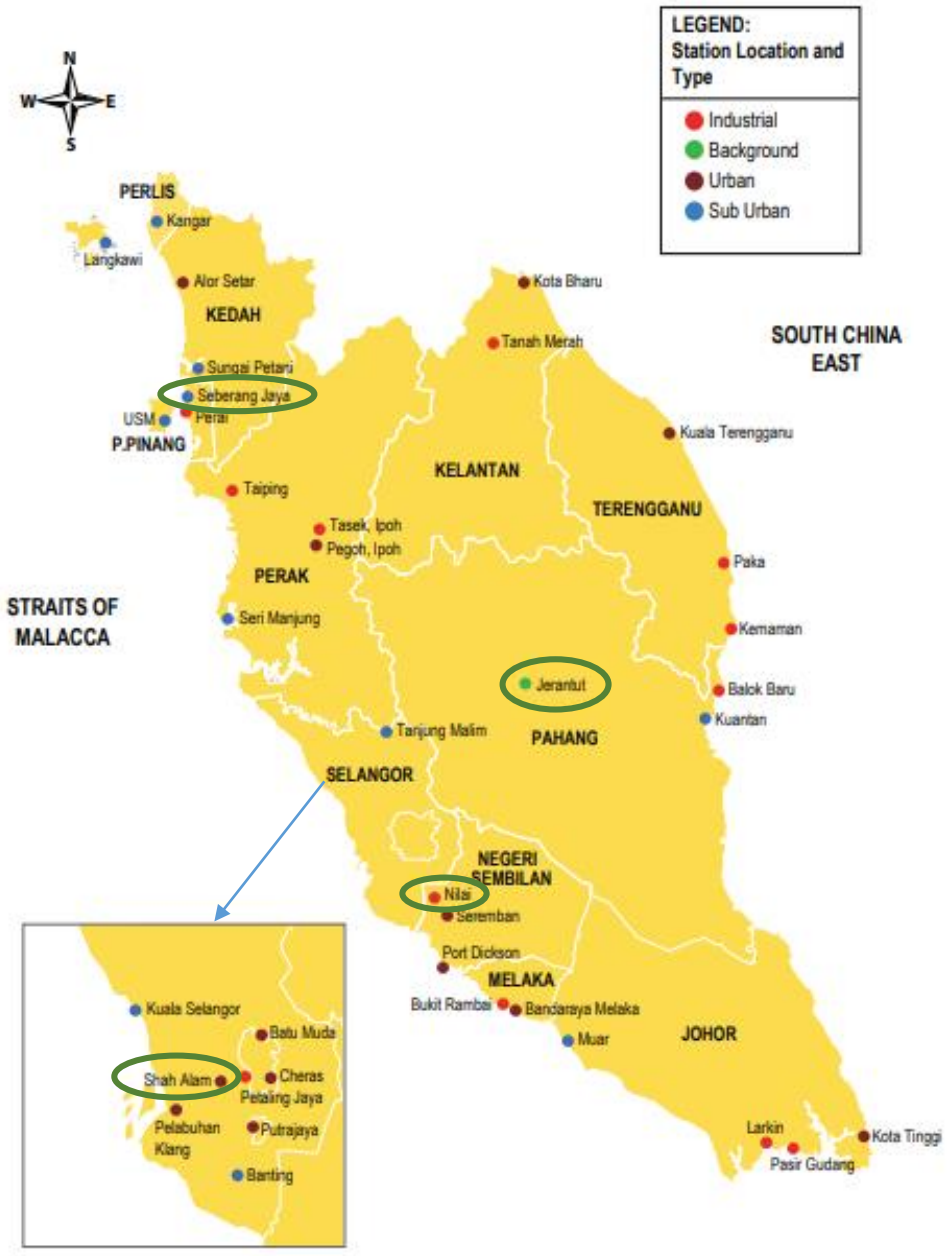

Figure 3-2 Location of continuous monitoring stations in Peninsular Malaysia (Source : Department of Environment Malaysia, 2015)