SHORT-TERM PREDICTION MODELS OF PM_{10} CONCENTRATIONS IN PENINSULAR MALAYSIA USING MULTIVARIATE TIME SERIES AND MACHINE LEARNING METHODS

NORAZRIN BINTI RAMLI

UNIVERSITI SAINS MALAYSIA

2021

SHORT-TERM PREDICTION MODELS OF PM_{10} CONCENTRATIONS IN PENINSULAR MALAYSIA USING MULTIVARIATE TIME SERIES AND MACHINE LEARNING METHODS

by

NORAZRIN BINTI RAMLI

Thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

May 2021

ACKNOWLEDGEMENT

Bismillahirrahmanirrahim. Alhamdulillah. Firstly, I praise Allah S.W.T. and His Almighty for giving me this opportunity and granting me the capability to successfully fulfill my doctorate research.My greatest appreciation goes to my helpful supervisor, Dr. Hazrul Bin Abdul Hamid for letting me be under his supervisions. I would like to sincerely thank him for all the support, guidance, suggestion and advice given to encourage me in completing my research. Not forgotten, I would like to express my special thanks to my respectful co-supervisor, Professor Ahmad Shukri Bin Yahaya for the valuable comments and guidance throughout the semester to enable me to complete my research and thesis. I would like to dedicate this thesis to my beloved parents, Ramli Bin Mahmod and Noriah Binti Awang Mat for their prayers, support and words of encouragement. For my husband, Muhammad Zayumi Bin Mohd Yusof, thank you for the love, concern and for always being there for me and for being supportive. For my beloved daughters, Nur Arissa Ariana and Nur Adelia Adriana who inspired me to face the challenges with courage, this thesis is made possible by your love. To my sisters, post-graduate friends and colleagues especially Norhidayah, Norsalimah, Dr. Nik Zainab, Dr. Mohd Shahidan Shaari, Mohd Khairul Nizam, Aida Jasmin and Nur Farhah Naaim, thanks for your support and cooperation during my studies.Finally, my profound acknowledgement to the Ministry of Higher Education Malaysia and Universiti Malaysia Perlis for the schorlarship and sponsorship, Department of Environment Malaysia for providing air monitoring data and also to the School of Distance Studies and School of Civil Engineering, Universiti Sains Malaysia.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Page

LIST OF ABBREVIATION

ADF	Augmented Dickey-Fuller
AIC	Akaike Information Criteria
ANN	Artificial Neural Network
API	Air Pollution Index
AR	Autoregressive
ARIMA	Autoregressive Integrated Moving Average
ARMA	Autoregressive Moving Average
BMA	Bayesian Model Averaging
BRT	Boosted Regression Tree
CIMA	Cement Industries of Malaysia Berhad
CO	Carbon Monoxide
CV	Coefficient of Variation
D_0	At that day
D_1	Day 1
DOE	Department of Environment
ELITE	North-South Expressway Central Link
EPA	Environmental Protection Agency
GCE	Guthrie Corridor Expressway
HQ	Hannan-Quinn information criteria
IA	Index of Agreement
KESAS	Shah Alam Expressway
KLIA	Kuala Lumpur International Airport
KTMB	Malayan Railways Limited
LEKAS	Kajang-Seremban Highway
LKSA	Kemuning-Shah Alam Expressway
LM	Lagrange Multiple
MAAQS	Malaysian Ambient Air Quality Standard
MK	Mann-Kendall
MLE	Maximum Likelihood Estimator
MLR	Multiple Linear Regression

MTS	Multivariate Time Series
NAAQS	National Ambient Air Quality Guideline Standards
NAE	Normalized Absolute Error
NKVE	New Klang Valley Expressway
NO_2	Nitrogen Dioxide
OLS	Ordinary Least Squares
O_3	Ozone
PLUS	North-South Expressway Project
PM_{10}	Particulate Matter with aerodynamic diameter less than 10 µm
$PM_{2.5}$	Particulate Matter with aerodynamic diameter less than 2.5 µm
PPIC	PETRONAS Petrochemical Integrated Complex
PSI	Pollutant System Index/ Pollution System Index
RMSE	Root Mean Square Error
RH	Relative Humidity
SD	Standard Deviation
SIC	Schwarz Information Criterion
SO_2	Sulphur Dioxide
Temp	Temperature
USEPA	United State Environmental Protection Agency
UTS	Univariate Time Series
VAR	Vector Autoregressive
VARMA	Vector Autoregressive Moving Average
VMA	Vector Moving Average
WHO	World Health Organization
WS	Wind Speed

LIST OF APPENDICES

MODEL PERAMALAN JANGKA PENDEK KEPEKATAN PM_{10} DI SEMENANJUNG MALAYSIA MENGGUNAKAN KAEDAH SIRI MASA MULTIVARIAT DAN PEMBELAJARAN MESIN

ABSTRAK

Zarah terampai dengan diameter aerodinamik kurang daripada 10μm (PM_{10}) telah dikenal pasti sebagai salah satu bahan pencemar udara yang berbahaya kepada kesihatan manusia dan kepekatan PM_{10} di bandar-bandar di Asia dan Pasifik kekal sebagai isu pencemaran udara yang paling bermasalah. Objektif kajian adalah untuk menentukan ciri dan trend kepekatan PM_{10} di Malaysia dari tahun 1999 hingga 2015, mencadangkan analisis Siri Masa Multivariat (MTS) menggunakan Autoregressif Vektor (VAR) untuk meramalkan kepekatan jangka pendek PM_{10} di Malaysia dan mentafsir hubungan antara kepekatan PM_{10} dan parameter meteorologi menggunakan pandangan grafik kausal. Tiga buah model untuk peramalan jangka pendek PM_{10} iaitu Model Regresi Linear Berganda (MLR), Model Purata Bayesian (BMA) dan Model Peningkatan Pokok Regresi (BRT). Petunjuk prestasi (R^2, IA, MAE, RMSE dan MAPE) digunakan untuk mendapatkan model terbaik. Kajian menggunakan data tujuh belas tahun pemantauan kualiti udara daripada Jabatan Alam Sekitar Malaysia (DOE) yang mengandungi lapan parameter (PM_{10}, NO_2, SO_2, CO, O_3, kelajuan angin, suhu dan kelembapan relatif) dan sembilan stesen pemantauan telah dipilih iaitu Kangar, Perai, Shah Alam, Nilai, Larkin, Pasir Gudang, Kertih, Kota Bharu dan Jerantut yang mewakili zon Utara, Tengah, Selatan dan Timur Semenanjung Malaysia. Analisis tren menggunakan Ujian Mann-Kendall untuk pengesanan tren dan penganggar cerun Sen untuk penganggaran tren menggunakan purata dan maksimum bulanan kepekatan PM_{10}. Data purata bulanan pemantauan

udara juga digunakan untuk model VAR, termasuk empat parameter: PM_{10},suhu, kelembapan relative dan kelajuan angin. MLR, BMA dan BRT menggunakan data purata harian pemantauan udara untuk lapan parameter (PM_{10}, NO_2, SO_2, CO, O_3, kelajuan angin, suhu dan kelembapan relatif). Lapan puluh peratus data pemantauan udara digunakan untuk latihan dan dua puluh peratus digunakan untuk pengesahan model. Keputusan analisis menunjukkan bahawa kepekatan harian PM_{10} di semua stesen berada dalam julat antara 37.34 $\mu g/m^3$ hingga 59.19 $\mu g/m^3$. Analisis trend menunjukkan bahawa trend monotoni adalah signifikan di stesen Nilai, Larkin, Kota Bharu dan Jerantut iaitu trend monotoni purata bulanan PM_{10} di stesen-stesen ini menunjukkan kepekatan PM_{10} meningkat secara konsisten mengikut masa. Model (VAR) 2 adalah model yang paling sesuai untuk meramal kepekatan PM_{10} di stesen Kangar dan Perai. Hubungan kausal menunjukkan bahawa hubungan kausal adalah signifikan di kebanyakan stesen pemantauan antara kepekatan PM_{10}, kelajuan angin dan kelembapan relatif. Model BRT memberikan hasil yang baik untuk meramalkan kepekatan PM_{10} di setiap stesen. Secara keseluruhannya, penemuan kajian ini memberikan gambaran tentang model peramalan baru menggunakan BRT terutamanya berhubung dengan kajian kualiti udara dan hasil penyelidikan ini menyokong idea bahawa model ini boleh diterima pakai dan digunakan dalam pelbagai bidang seperti kawalan pencemaran udara dan kajian pemantauan. Model ini akan menjadi relevan bagi Jabatan Alam Sekitar Malaysia untuk bekerjasama membangunkan dan melaksanakan peramalan kualiti dan pemantauan, penilaian dan sistem amaran awal untuk mencegah, memantau dan mengurangkan pencemaran udara di Malaysia.

SHORT-TERM PREDICTION MODELS OF PM_{10} CONCENTRATIONS IN PENINSULAR MALAYSIA USING MULTIVARIATE TIME SERIES AND MACHINE LEARNING METHODS

ABSTRACT

The particulate matter with an aerodynamic diameter less than 10 µm (PM_{10}) is identified as one of the dangerous air pollutants to human health and the concentrations of PM_{10} in Asian and Pacific cities remain as the most problematic local air pollution issues. The objectives of the research are to determine the characteristics and trend of PM_{10} concentrations in Malaysia from 1999 to 2015, to propose a Multivariate Time Series (MTS) analysis using Vector Autoregressive (VAR) to predict the short-term PM_{10} concentrations and interpret the relationship between PM_{10} concentrations and meteorological parameters using the graphical view of causality. Three models for short-term prediction of PM_{10} using Multiple Linear Regression (MLR), Bayesian Model Averaging (BMA) and Boosted Regression Tree (BRT) model. The performance indicators (R^2, IA, MAE, RMSE, and MAPE) are applied to obtain the best model. A study using seventeen years of air quality monitoring data from the Department of Environment Malaysia (DOE) was used with eight parameters (PM_{10}, NO_2, SO_2, CO, O_3, wind speed, temperature, and relative humidity) and nine monitoring stations were selected which included Kangar, Perai, Shah Alam, Nilai, Larkin, Pasir Gudang, Kertih, Kota Bharu and Jerantut to represent the Northern, Central, Southern and East of Peninsular Malaysia. The trend analysis used the Mann-Kendall test for trend detection and Sen's slope estimator for trend estimation using monthly average and maximum monthly of PM_{10} concentrations. The monthly average of air monitoring data was also used for the VAR model, which includes four parameters: PM_{10}, temperature, relative humidity, and wind speed. The MLR, BMA and BRT used daily average of air monitoring data for eight parameters (PM_{10}, NO_2, SO_2, CO, O_3, wind speed, temperature, and relative humidity). Eighty percent of the air monitoring data was used for training and twenty percent was used for validation of the models. The results showed that the daily mean PM_{10} concentrations at all the stations were in the range between 37.34 $\mu g/m^3$ to 59.19 $\mu g/m^3$ for all stations. The trend analysis showed that the monotonic trend was significant at Nilai, Larkin, Kota Bharu and Jerantut which indicated that the mean monthly PM_{10} concentrations at these stations had a monotonic trend where the PM_{10} concentrations consistently increased through time. The VAR (2) model is the most suitable model for predicting PM_{10} concentrations at Kangar and Perai monitoring station. The causality relationship shows that the significant causal relationship at most monitoring stations is between PM_{10} concentrations, wind speed and relative humidity. The BRT model gave good results to predict the PM_{10} concentrations for each station. Overall, the finding of the research provides an insight for a new prediction model using BRT especially in relation to the air quality studies and the results support the idea that this models can be adopted and applied in various fields such as air pollution control and monitoring studies. The model will be relevant for Department of Environment Malaysia to cooperatively develop and implement air quality forecasting and monitoring, assessment and early warning system to prevent, monitor and mitigate the air pollution in Malaysia.

CHAPTER 1

INTRODUCTION

1.1 Background

The particulate matter with an aerodynamic diameter less than 10 microns in size (PM_{10}) become one of the most problematic issues in the Asian and Pasific cities although from 1990 to 2006 the PM_{10} concentrations declined by 38% (International Energy Agency et al., 2011). However, the annual average PM_{10} concentrations in 230 cities monitored in 2008 was 4.5 times over the World Health Organization's (WHO) standard (International Energy Agency et al., 2011). Recently, the World Health Organization (WHO) reported that more than 80% of people living in urban areas are exposed to air quality levels that exceed WHO guideline limits, with low- and middle-income countries suffering from the highest exposures, both indoors and outdoors. The data shows that 9 out of 10 people breathe air containing high levels of pollutants (World Health Organization, 2020[a]).

In urban cities, PM_{10} was emitted from road vehicles which can be separated into two primary sources which are exhaust and non-exhaust emission. The exhaust emission was emission from carbonaceous particles and non-exhaust emission were from tires, clutch, brakes wear and tear producing a by-product of particulate (Lawrence et al., 2016; Samiksha et al., 2017). The air quality, as determined by the concentrations of the major pollutants in the atmosphere, has further deteriorated in recent years and is projected to further decline over the next 40 years due to the rapid rate of economic development in Asia (United States Environmental Protection Agency, 2017).

According to the World Health Organization (WHO) air quality is deteriorating in the region due to industrialization and urbanization. The motor vehicle emissions contribute to the largest proportion of air pollution in urban areas. A survey in Metro Maanila in the Philippines shows that road traffic contributes 50% to 85% of the particulate matter pollution at various sampling locations in the metropolitan area (World Health Organization, 2008). The situation is more serious in the urban cities where the level of particulate matter with an aerodynamic diameter of less than 10 µm (PM_{10}) had significantly exceeded the World Health Organization (WHO) annual mean guideline which is 20 $\mu g/m^3$ as 24-hour mean in the past decade (United States Environmental Protection Agency, 2017).

The World Health Organization (WHO) estimated that seven million premature deaths annually are linked to air pollution thereby confirming that air pollution is now the world's largest single environmental health risk. Based on the projections by the United Nations, the total population in Asia will exceed five billion by the year 2050, and approximately 65% of the population will be living in the cities. This will result in a growing number of people at risk of premature death (United States Environmental Protection Agency, 2017).

Emissions of air pollutants from the combustion of fossil fuels continue to increase in Asia due to the increasing energy demands. The substantial increases in the combustion of fossil fuels for power generation and transportation can improve economic conditions but can also, if not controlled, cause significantly negative consequences for human health and environmental quality through the transboundary transport of pollutants. Effective approaches to control and reduce pollution are crucial to avert increased environmental degradation and associated health impacts while reducing poverty and providing economic stability for the population (Acid Deposition Monitoring Network in East Asia, 2014).

According to Environmental Protection Agency (EPA), there are two kinds of air pollution trends. Firstly, the air concentrations based on actual measurements of pollutant concentration in the ambient air at the monitoring sites/stations throughout the country (United States Environmental Protection Agency, 2012). The purpose of trends is to assess the level of air pollution and provide information about local air quality changes over time and revealed the current state of air quality status. Secondly, the emission based on engineering estimation of the total tons of pollutants released into the air each year (United States Environmental Protection Agency, 2012).

The PM_{10} emitted from vehicular exhaust contains an organic compound (OC) derived from the primary emission sources and the OC can form the secondary PM_{10} emission sources via conversion of volatile organic compound (VOCs) into particles (Wang et al., 2014). Besides, the active movement of vehicles in the urban area also emit the re-suspension of dust to the atmosphere (Suleiman et al., 2016). According to

Department of Statistics Malaysia (2017) in 2012, the number of vehicles registered was 22.6 million, and the cumulative registered motor vehicles had shown an increase of 22.1% in 2016 with the total of 27.6 million vehicles respectively (Department of Statistics Malaysia, 2017). In June 2017, the number of vehicles registered in Malaysia was 28.2 million which increased by approximately 1.23 million vehicles a year (Department of Statistics Malaysia, 2018).

Air Pollution Index (API) for Malaysia is used to compare the air quality with other regional countries after the Department of Environment (DOE) Malaysia revised its index system in 1996. The API for Malaysia is shown in Table 1.1. The indexes closely follow the United States Pollutant System Index (PSI) and it is based on six pollutants and reported as a number, on a scale of 0 to 500. The index enables the public to determine the air pollution level. With effect from 1 April 2014, Singapore included $PM_{2.5}$ into the current Pollutant System Index (PSI) as its sixth pollutant parameter (National Environment Agency Singapore, 2014) and the DOE Malaysia has improved the new calculation of API by using $PM_{2.5}$ starting on 16 August 2018 (Department of Environment Malaysia, 2019).

Table 1.1 Malaysia Air Pollution Index (API) (Department of Environment Malaysia, 2018[b])

Air Pollution Index (API)	**Air Quality Index (AQI)**
0-50	Good
51-100	Moderate
101-200	Unhealthy
201-299	Very Unhealthy
300-500	Hazardous
>500	Emergency

The calculation of API value is calculated based on the average concentration of air pollutants namely SO_2, NO_2, CO, O_3, PM_{10} and $PM_{2.5}$ and the dominant pollutant (the highest concentration) will determine the API value. The calculation of API is presented in Figure 1.1.

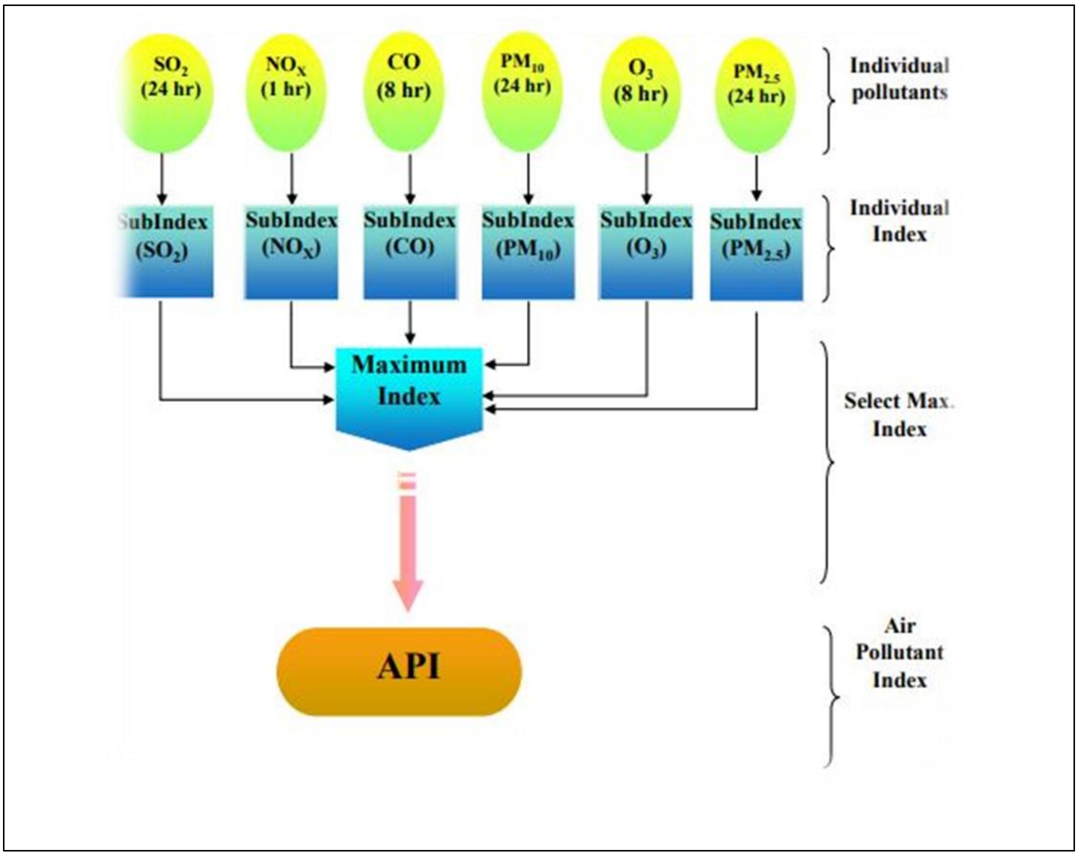


Figure 1.1 The calculation of the Air Pollution Index
(Department of Environment Malaysia, 2018[b])

The average calculation for individual pollutants is taken at different range of periods such as SO_2 (1 hour), NO_2 (1 hour), CO (8 hours), O_3 (8 hours), PM_{10} (24 hours) and $PM_{2.5}$ (24 hours) because the pollutants have different exposure periods that are acceptable to humans. The average concentrations of individual pollutants over a given period is standardized to produce the sub index which is non-unitary value using a specific mathematical formula. Each individual pollutant will produce its sub -index and the maximum sub-index will be considered as API reading. Usually, the concentrations of particulate matter is the dominant pollutant most of the time especially during haze period in Malaysia (Department of Environment Malaysia, 2018[b]). Particulate matter (PM) once inhaled, can affect the heart and lungs and cause serious health effects. The particulate matter with an aerodynamic diameter less than 10 microns in size (PM_{10}) is one of the most dangerous air pollutants to human health because its size range overlaps with the range of respirable particles (Beatrix et al., 2013; Chiou and Tsai, 2001; Zobeck and Van Pelt, 2006).

The Air Pollution Index (API) tells how clean or polluted an outdoor air is, along with associated health effects that may be of concern. The API translates air quality data into numbers and colours that help people to understand when to act to protect the health. The API reading of 0 to 50 is regarded as a ‘Good’ status and no health impacts are expected. The ‘Moderate’ status refers to moderate pollution which does not pose bad effects on the health while the ‘Unhealthy’ status appears when the API reading is between 101 to 200. The ‘Unhealthy’ status pollution level may affect the health conditions of high-risk people who are people with heart and lung complications such as asthma. When the API reading is recorded at 201 to 299, the status of air quality is ‘Very Unhealthy’ which can worsen health conditions especially for people with heart and lung complications (Department of Environment Malaysia, 2018[a]; Department of Environment Malaysia, 2017[b]).

The ‘Hazardous’ status occurs when the API reading is between 300 to 500 and the air pollution is considered hazardous to high-risk people especially the elderly, children and public health. Hence, the public should limit prolonged outdoor exertion during this condition. An ‘Emergency’ status will be announced when the API reading is over 500 which is hazardous to high-risk people (respiratory disease such as asthma), and public health and all outdoor exertion and activities must be avoided during emergency period (Department of Environment Malaysia, 2018[a]). Due to the public concern, the DOE Malaysia has launched the application of MyIPU for smartphones and Android users to inform and update the air quality status throughout the country (Department of Environment Malaysia, 2018).

In Malaysia, from 1995 to 2016 the PM_{10} concentrations data are recorded and validated by the Department of Environment (DOE) Malaysia using the *Beta* Attenuation Mass Monitor (BAM-1020) as part of a Malaysian Continuous Air Quality Monitoring (CAQM) program, which is manufactured by Met One Instruments Inc. (Afroz et al., 2003). The DOE monitors the air pollutants concentrations in Malaysia by observing 52 Continuous Air Quality Monitoring Stations (CAQMS) which are located throughout the country. Five major air pollutants measured such as ground-level ozone (O_3), carbon monoxide (CO), nitrogen dioxide (NO_2), sulphur dioxide (SO_2), particulate matter with a diameter size below than 10μm (PM_{10}) and meteorological parameters such as ambient temperature, relative humidity and wind speed. These stations have been categorized into four categories which are industrial, urban, sub-urban and background stations (Department of Environment Malaysia, 2017, 2018, 2019).

In June 2016, Pakar Scieno TW Sdn. Bhd. was awarded a 15 year contract to establish, develop and implement the National Environmental Quality Monitoring Program (NEQMP) in Malaysia (Pakar Scieno TW, 2018). Starting mid-April 2017, DOE has upgraded to 65 Continuous Air Quality Monitoring Stations (CAQMS), 14 Manual Air Quality Monitoring Stations (MAQMS) and three Mobile Continuous Air Quality Monitoring Stations (MCAQM) under new Environmental Quality Monitoring Programme (EQMP). Particulate matter with a diameter size below 2.5 microns in size ($PM_{2.5}$) was added into the measurement in selected areas starting on 16 August 2018 (Department of Environment Malaysia, 2017, 2018, 2019). From 2017 until present, the PM_{10} and $PM_{2.5}$ concentrations were measured using Tapered Element Oscillating Microbalance (TEOM) which is the particle analyzer TEOM 1405DF (Pakar Scieno TW, 2018).

1.2 Problem Statement

Air pollution has been widely recognized as a problem of the last five decades which impacts human health, well-being and the environment (Hussein and Abdullah, 2018). According to World Health Organization (WHO), outdoor air pollution causes 4.2 million death each year (World Health Organization, 2020[a]) principally in large cities where the major outdoor pollution sources include vehicles, power generation, building heating systems, agriculture/waste incineration and industry (World Health Organization, 2020[b]). In 2020, the WHO reported that air pollution cause about seven million premature deaths every year, largely as a result of increased mortality from stroke, heart disease, chronic obstructive pulmonary disease, lung cancer and acute respiratory infections (World Health Organization, 2020[a]; Fang et al., 2017; Kanniah et al., 2016) when the public had been exposed to high level of PM_{10} pollution (Dotse et al., 2016).

The increase of transportation and combustion of fossil fuels for power generation can improve the economic condition. However, one of the greatest challenges is to control and monitor emissions (Jamalani et al., 2018). The concentration of PM_{10} in Asian and Pacific cities remain as the most problematic local air pollution issues (International Energy Agency et al., 2011; Zhou et al., 2014) which has been classified as the most significant pollutant in Southeast Asia and Peninsular Malaysia (Mohamed Noor et al., 2015; Latif et al., 2014). The high amount of PM_{10} emission was significantly proportional to the increase of industry and the number of vehicles on-road which resulted in an increase in air pollution (Jamalani et al., 2018). The high level of PM_{10} concentrations has been shown to be related to adverse effect in agricultural, degradation of the environment and biodiversity (Sulong et al., 2017; Fotourehchi, 2016; Hassan et al., 2015).

Air quality in Malaysia is also affected by transboundary pollution or haze where several areas were struck by haze especially in the West Coast of Peninsular Malaysia. The sources of haze generally came from the land-use changes, slash and burn, burning within the oil palm plantation, peat combustion and local open burning activities (Department of Environment Malaysia, 2016a). The agricultural and tourism sectors also experienced heavy losses due to high concentrations of PM_{10}. The other impacts include the reduction in plant yield due to the limitation of light level (Sulong et al., 2017). Towards the Sustainable Development Goals (SDGs), the government holds the promise of a path to environmental sustainability and as well as the improvement of air quality status. Sustainable consumption and production (SCP) were introduced to achieve environmental sustainability.

In 2016, the Malaysian Carbon Reduction and Environmental Sustainability Tool (MyCREST) was adopted to quantify carbon emissions and sustainable impacts of the built environment (Malaysia Prime Minister's Department, 2017). The air pollutants concentration limit will be strengthened through the interim target and the full implementation of the standard is in 2020. One of the challenges is to ensure that the air quality in Malaysia is in good condition and the environmental pollution can be reduced through the new standards (Department of Environment Malaysia, 2018). There is increasing concern due to rapidly industrial planning, projected economic growth, and development will increase the number of people, vehicles and industries and this in turn will create environmental challenges and may deteriorate the air quality in Malaysia (Abdullah et al., 2019; Department of Statistics Malaysia, 2018; Jamalani, 2018).

The statistical modelling is required to predict the future PM_{10} concentrations in Malaysia (Ul –Saufie et al., 2015). There are numerous methods and model for PM_{10} prediction such as principle component regression (PCR) (Fong et al., 2018; Abdullah et al., 2016), principle component analysis (PCA) (Abdullah et al., 2016; Ul-Saufie et al., 2013; Dominick et al., 2102), multiple linear regression (MLR) (Abdullah et al., 2019; Fong et al., 2018; Abdullah et al., 2017; Abdullah et al., 2016; Ul-Saufie et al., 2013; Dominick et al., 2012), feedforward backpropagation (FFBP) (Ul-Saufie et al., 2015, 2013), probabilistic and distribution modelling (Hamid, 2013), hybrid model (Ul-Saufie et al., 2013) and so on. The prediction models are an important tool because the prediction model is developed to minimize the autocorrelation or error in the model. The statistical modelling has the potential for high accuracy for PM_{10} concentrations prediction (Shahraiyni and Sodoudi, 2016).

The short-term prediction is a short period of prediction such as daily prediction (the next day), monthly prediction (next month) or yearly prediction (next year) of PM_{10} concentration. The public must be informed when high PM_{10} concentration conditions are present (Shahraiyni and Sodoudi, 2016) and the administrations must attempt to reduce pollutant concentrations by limiting vehicular traffic on some days (Brunelli et al., 2007; Stadlober et al., 2008), industrial emission restriction, and urban planning (Paschalidou et al., 2011). To prevent the risk of critical concentration levels, abatement actions such as traffic reduction should be planned at least one or two days in advance (Baklanov et al., 2007). Therefore, a short-term prediction must be developed and used as a rapid alert system to inform the public of harmful air pollution events, as well as to adapt air pollution control strategies (Brunelli et al., 2007).

Thus, it is important to predict the short-term PM_{10} concentrations in Malaysia using statistical model prediction. There is also an urgent need to address the inter-relationship among the air pollutants and their negative consequence to the air quality through the statistical modelling and prediction strategies. The research proposed a

Multivariate Time Series (MTS) analysis using Vector Autoregressive (VAR) model to predict the short-term PM_{10} concentration in Malaysia. So far, this method has been widely applied in econometrics studies describing the dynamic behavior of economic and financial time series and forecasting. Although the MTS is widely used with economic and financial data, its use on environmental data is limited. Thus, the research applied this method for examining the air quality data for short-term prediction and find the interaction of air pollutants especially the particulate matter (PM_{10}) with the meteorological parameters.

Dealing with air pollution data, many uncertainties needs to be considered because of the dynamic nature of the system. In recent years, the Bayesian approach has gained popularity to fit statistical models, and the Bayesian methods offers an alternative modeling strategy because the approach has the ability to take account of all parameter uncertainties (Dongen and Geuens, 1998; Evans, 2012). This research also considered the uncertainties in air pollution studies, so the Bayesian Model Averaging (BMA) was suggested. A new method in air pollution especially for PM_{10} concentrations using Boosted Regression Tree (BRT) method was applied to develop a model for predicting PM_{10} concentrations in Malaysia.

The Multiple Linear Regression (MLR) has been widely used for PM_{10} forecasting in urban areas (Shahraiyni and Sodoudi, 2016). The comparison between the results of MLR and other techniques demonstrates the weakness of the MLR approach. The stepwise input variable selection technique is often used in MLR for the determination of suitable explanatory variables for regression. In different MLR studies, the collinearity among the input parameters has often occurred and sometimes the Principal Component Analysis (PCA) is used to overcome the problem of collinearity (Paschalidou et al., 2011). The MLR model as a short- term predicting tool is also included in the research to compare the results with the other models in this research.

1.3 Objectives

The objectives of this research are:

1. To determine the characteristics and trend of PM_{10} concentrations in Malaysia from 1999 to 2015.

2. To develop a short-term prediction model of PM_{10} concentrations in Malaysia using Multivariate Time Series Analysis and interpret the relationship between PM_{10} concentrations and meteorological parameters.

3. To develop the short-term prediction models to predict PM_{10} concentrations using Multiple Linear Regression Model (MLR), Bayesian Model Averaging (BMA) and Boosted Regression Tree (BRT).

4. To compare the performance of the Multiple Linear Regression Model (MLR), the Bayesian Model Averaging (BMA) and the Boosted Regression Tree (BRT) model and obtain an appropriate short-term prediction model to predict PM_{10} concentrations in Malaysia.

1.4 Scope of Research

The research only concentrate on one specific pollutant which is the particulate matter with an aerodynamic diameter less than 10 microns in size (PM_{10}). The real data of ground level air monitoring records were obtained from the Department of Environment Malaysia for the period between 1999 to 2015 from nine monitoring stations: Kangar, Perai, Shah Alam, Nilai, Larkin, Pasir Gudang, Kertih, Kota Bharu and Jerantut. 80% of the data were used as training data, and another 20% of the data were used for validation.

For trend analysis the data used were monthly average and maximum monthly of PM_{10} concentrations. For short-term prediction PM_{10} concentrations using Vector Autoregressive (VAR) model the data used was monthly average data. The daily average of PM_{10} concentrations data were used for statistical models to predict the short-term PM_{10} concentrations using Multiple Linear Regression (MLR), Bayesian Model Averaging (BMA) and Boosted Regression Trees (BRT).

For trend analysis, two test statistics were utilized which are the Mann-Kendall test and Sen's slope test. For multivariate time series (MTS) model, the short-term prediction of PM_{10} concentrations which focused on the PM_{10} as the dependent variable and the meteorological parameters (wind speed, temperature, and relative humidity) as independent variables. The statistical models to predict the PM_{10} concentrations included Multiple Linear Regression, Bayesian Model Averaging and Boosted Regression Trees were used eight parameters (PM_{10}, NO_2, SO_2, CO, O_3, wind speed, temperature, and relative humidity). The performance indicators will be used to obtain a good model to predict PM_{10} concentrations. The methods and flowchart of the methods will be discussed in Chapter 3.

1.5 Research Layout

The overall structure of the research takes the form of five chapters, including the following:

Chapter 1 is an overview of air pollution in cities, problems statement, objectives of this research, the scope of the research and research layout.

Chapter 2 discusses and summarizes the literature review about air pollution, particulate matter (PM_{10}), sources and effects of PM_{10} concentration, other pollutants such as ground-level ozone (O_3), carbon monoxide (CO), sulphur dioxide (SO_2), and nitrogen dioxide (NO_2). This chapter also includes the literature review of trend analysis studies, Multiple Linear Regression (MLR), Multivariate Time Series (MTS), Boosted Regression Tree (BRT) and Bayesian Model Averaging (BMA).

Chapter 3 describes the procedures and methodology applied in this research. The trend analysis, Multivariate Time Series (MTS) analysis, Multiple Linear Regression (MLR), Boosted Regression Tree (BRT) and Bayesian Model Averaging (BMA) are presented with the flowchart of the analysis and test procedure which are undertaken during the research. The performance indicators such as the coefficient of determination (R^2), index of agreement (IA), normalized absolute error (NAE), root mean square error (RMSE) and Mean Absolute Percentage Error (MAPE) were used to find an appropriate short-term prediction model that fit the actual monitoring records of PM_{10} concentrations and to validate the short-term prediction models.

Chapter 4 presents the findings of the research, focusing on the four objectives of this research. The characteristics and trend of the PM_{10} concentrations are presented for each monitoring stations. The short-term prediction model of the PM_{10} concentrations that uses the Multivariate Time Series (MTS) analysis and the Granger causality is illustrated in graphical view of causality. Multiple Linear Regression (MLR), Boosted Regression Tree (BRT) and Bayesian Model Averaging (BMA) model is also presented. The obtained performance indicators for each model will also be presented. Finally, this chapter proposes an appropriate model for the short-term prediction of the PM_{10} concentrations.

Chapter 5 provides the conclusion for this research and its significant findings. The contribution of this research will also be summarized, and the recommendations for the future work are included in this chapter.

CHAPTER 2

LITERATURE REVIEW

2.1 Air Pollution

In recent decades, air pollution studies have been one of the major research interests due to the increasing emission discharged to the atmosphere. The air pollution has become a major issue for many nations. The history of air pollution control actions started in the United States in 1945. In 1930 until 1940, a factory smokestack issuing a thick plume of smoke was considered a sign of prosperity. However, between 1945 and 1970, an awareness of air pollution problems gradually increased in the United States. The Clean Air Act of 1970 came as an immense surprise to most major industries. In the late 1980s, air pollution became an issue caused by problems involving the longer-lived pollutants and the transported trans-boundary pollutants. Due to the increase in air pollution problems, developing countries subsequently responded to control air pollution by stringent the act and standard emissions. The regional countries revised the standard and closely followed the Pollutant System Index (PSI) from the United States and the World Health Organization (WHO), which seek similar goals to control the air pollution (Nevers, 2000).

According to the United States Environmental Protection Agency (US EPA), about 78000 million kilograms of pollution were emitted in the United States atmosphere in 2016, and 79000 million kilograms in 2017. The emission has increased approximately one million kilograms each year that mostly contributes to the deposition of acids, visibility problem and ozone formation (United States Environmental Protection Agency (USEPA), 2017; 2018).

The percentage of changes in air quality for the six pollutants across the United States' (U.S.) atmosphere obtained through the national and state regulation and EPA's program is shown in Table 2.1. EPA tracks a range of emissions data that was emitted from various pollution sources and the trend for the six pollutants concentrations in the U.S. has decreased substantially over the years especially for Pb that has decreased by 99% between 1980 and 2017 after it was eliminated from gasoline, and this was followed by the CO concentrations which decreased by 84%. The NO_2 concentration decreased by 63%, and O_3 indicated a 32% decreased. However, the PM_{10} concentrations percentage did not show any changes in air quality nationally between 2010 and 2017, and this is the sign that the PM_{10} concentrations are the most problematic issue that needs to be considered nationally.

Table 2.1 Percent Change in Air Quality (Environmental Protection Agency, 2018)

Pollutants	**1980 vs 2017**	**1990 vs 2017**	**2000 vs 2017**	**2010 vs 2017**
Carbon Monoxide (CO)	-84	-77	-61	-13
Lead (Pb)	-99	-98	-94	-80
Nitrogen Dioxide (NO_2)	-63	-56	-49	-21
Ozone (O_3) (8-hour)	-32	-22	-17	-5
PM_{10} (24-hour)	---	-34	-30	0
$PM_{2.5}$ (24-hour)	---	---	-40	-10
Sulphur Dioxide (SO_2) (1-hour)	-90	-88	-79	-66

The emissions of air pollutants either from the stationary or mobile source is a primary concern due to the fact that the emission will affect the health, longevity, and quality of life. The exposure to air pollution can generally be divided into two types either directly or indirectly, where direct exposure is through physical exposure, and indirect exposure is through the perception of risk and mental health (Fotourehchi, 2016; Nevers, 2000).

Malaysia which has a tropical climate due to its geographical location close to the equator has constant high relative humidity and the seasons are highly dependent on the changes in wind flow and rainfall (Md Yusof et al., 2010). In Malaysia, air pollution is generally transported based on wind direction during the Northeast monsoon (November to March) and Southwest monsoon (June to September) and a large amount of transboundary air pollution in Malaysia is a result of forest fire in Sumatera, Indonesia transported by the Southwest monsoon where the country experiences dry and hot weather conditions (Dominick et al., 2012).

The sources of air contaminants may generally be classified as stationary, mobile, and open burning which are attributed to the point sources such as industrial stack emission, transportation activities such as automobile emissions and uncontrolled sources such as wind-blown dust from stockpiles (Corbitt, 2004). There are three major sources of air pollution in Malaysia which are mobile, stationary and open burning. From the three major sources, the cause of air pollution can be divided into six sources which are from power generation and industries (stationary sources), development activities, motor vehicle (mobile sources), land clearing, forest fire and open burning (Department of Environment, 2017).

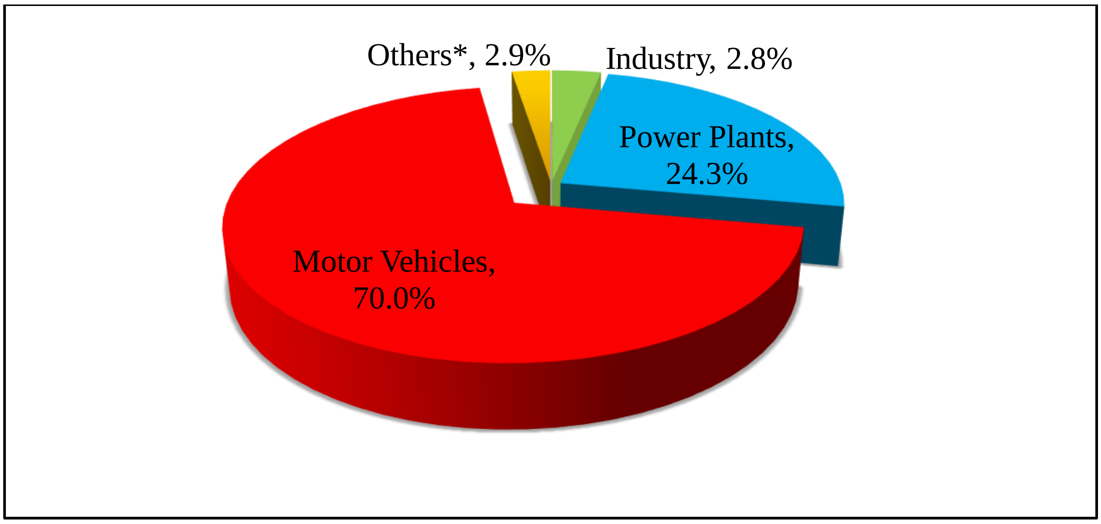
The Department of Statistics Malaysia reported that 70% of air pollution in Malaysia is from motor vehicles emissions while the stationary sources and other sources were 27.8% and 2.6% respectively. The major contributors of PM_{10} emission especially in urban areas are motor vehicles. The increasing number of motor vehicles especially in urban areas, contribute to the major sources of PM_{10} concentrations (Abdullah et al., 2019; Department of Statistics Malaysia, 2017; Rahman, 2013). The number of motor vehicles registered in Malaysia from the period of 2013 until April 2018 is shown in Table 2.2 as reported in the Monthly Statistical Bulletin Malaysia, 2018. It is shown that the number of motor vehicles in 2017 is increasing compared to

2016 especially for motorcycles which are 492,130 in 2017, followed by motorcars (572,615), goods vehicles (34,991) and other vehicles (15,333).

Table 2.2 New registration of motor vehicles in Malaysia (Department of Statistics Malaysia, 2018)

Period	**Number of vehicles registered during the period**	**Motorcycles**	**Motorcars**	**Public transport**	**Goods vehicles**	**Other vehicles**
	Number					
2013	1,202,689	528,508	583,060	9,603	40,765	40,753
2014	1,282,936	541,387	664,335	13,019	43,705	19,490
2015	1,200,760	465,525	671,786	9,355	38,115	15,979
2016	1,075,768	451,302	569,995	8,510	32,735	13,226
2017	1,119,990	492,130	572,615	4,921	34,991	15,333
April 2018	389,594	184,784	184,499	1,893	12,632	5,786

The emission of pollutants by sources to the atmosphere in Malaysia in 2016 is illustrated in Figure 2.1Figure 2.1. Motor vehicles account for the highest contribution of emission to the atmosphere which is 70.0 percent followed by power plants 24.3 percent, industry 2.8 percent and other sources 2.9 percent. These figures are in line with the statistics of vehicles registered in Malaysia which increase each year.



**Includes hotels, commercial centres, institutions, night markets and open burnings activities*

Figure 2.1 Emission of pollutants by sources to the atmosphere in Malaysia, 2016 (Department of Statistics, 2017)

Figure 2.1 shows the emission of pollutants that have been dispersed into the atmosphere in 2016 throughout Malaysia according to types and sources and they are shown in Table 2.3. The stationary sources of air pollution in Malaysia are from industries, and power plants and the non-stationary source is from motor vehicles. The other sources include hotels, commercial centers, institutions, night markets, and open burnings activities. The major source of pollution in Malaysia is from motor vehicles, followed by power plant emissions, industries, and other sources. This statistics reveals that the highest pollutant emitted to the atmosphere is carbon monoxide; CO emission is over two million tonnes (2,044,142 tons), followed by the nitrogen dioxide, NO_2 (849,249 tons), sulphur dioxide, SO_2 (228,983 tons), and particulate matter, PM (26,581 tons) that are recorded in 2016 (Department of Statistics, 2017).

Table 2.3 Emission of pollutants to the atmosphere according to type and source, Malaysia (Department of Statistics, 2017)

Metric Tons

Pollutants	Sources					
	Stationary sources			Motor vehicles	Others*	Total
	Industry	Power plant	Total			
Particulate Matter**	9,116	9,264	18,380	3,995	4,206	**26,581**
Carbon Monoxide (CO)	9,667	77,911	87,578	1,951,622	4,942	**2,044,142**
Nitrogen Dioxide (NO_2)	59,327	556,603	615,930	222,459	11,040	**849,429**
Sulphur Dioxide (SO_2)	23,761	130,769	154,530	14,204	60,249	**228,983**
Total	**101,871**	**774,547**	**876,418**	**2,192,280**	**80,437**	**3,149,135**

**Includes hotels, commercial centres, institutions, night markets, and open burnings activities*
***Particulate Matter (PM) includes all types of particulate matters*

The Department of Environment, Malaysia (2018) has established a new Ambient Air Quality Standard (AAQS) to replace the older Malaysia Ambient Air Quality Guideline that has been used since 1989. Table 2.4 shows the new AAQS in Malaysia. The standard adopts six air pollutants criteria that include five existing air pollutants which are particulate matter with an aerodynamic diameter less than 10 micron (PM_{10}), sulphur dioxide (SO_2), carbon monoxide (CO), nitrogen dioxide (NO_2), and ground-level ozone (O_3) as well as one additional parameter which is particulate matter with an aerodynamic diameter less than 2.5 micron ($PM_{2.5}$). The pollutants concentrations were reduced respectively until the standard implementation in the year 2020. 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. The air pollutants concentration limit will be strengthened in stages until

2020. The Ambient Air Quality Standard Interim Target 2 (IT-2) for PM_{10} concentrations was reduced to 45 µg/m^3 for a 1-year averaging time in order to follow the standard in 2020 which will not exceed 40 µg/m^3 yearly. The Ambient Air Quality Standard for Interim Target 3 (IT-3) stated that in 2020, the PM_{10} concentrations for a 24-hour averaging time was reduced to 100µg/m^3.

Table 2.4 New Ambient Air Quality Standard (Department of Environment, 2018[b])

Pollutants	**Averaging Time**	**Ambient Air Quality Standard**		
		IT-1(2015)	**IT-2(2018)**	**Standard (2020)**
Particulate Matter with an aerodynamic diameter of less than 10 microns (PM_{10}), µg/m^3	1 Year	50	45	40
	24 Hour	150	120	100
Particulate Matter with an aerodynamic diameter of less than 2.5 microns ($PM_{2.5}$), µg/m^3	1 Year	35	25	15
	24 Hour	75	50	35
Sulphur Dioxide (SO_2), µg/m^3	1 Hour	350	300	250
	24 Hour	105	90	80
Nitrogen Dioxide (NO_2), µg/m^3	1 Hour	320	300	280
	24 Hour	75	75	70
Ground Level Ozone (O_3), µg/m^3	1 Hour	200	200	180
	8 Hour	120	120	100
Carbon Monoxide (CO), mg/m^3	1 Hour	35	35	30
	8 Hour	10	10	10

2.1.1 Air Quality Monitoring Network in Malaysia

Air quality monitoring in Malaysia is supervised by Department of Environment (DOE) Malaysia and air quality monitoring activities implemented by DOE officers until 1995. In 1995 until 2016, the installation, operation and maintenance of Air Quality Monitoring Stations (AQMS) monitoring activities is outsourced to private sector which is performed by Alam Sekitar Malaysia Sdn. Bhd (ASMA) on behalf of the Department of Environment (Afroz et al., 2003; Pakar Scieno TW, 2018). The PM_{10} concentrations was recorded using the β-ray attenuation mass monitor (BAM-1020), as made by Met One Instruments Inc. The BAM-1020 is built-in with a cyclone and PM_{10} head particle trap, fiberglass tape, flow control and a data logger which delivers a fairly high resolution of 0.1 $\mu g/m^3$ at a 16.7 L/min flow rate, with lower detection limits of 4.8 $\mu g/m^3$ and 1.0 $\mu g/m^3$ for 1h and 24 h, respectively (Latif et al., 2014).

In June 2016 until present, the monitoring activities is performed by Pakar Scieno TW Sdn. Bhd. to establish, develop and implement the National Environmental Quality Monitoring Program (NEQMP) in Malaysia (Pakar Scieno TW, 2018). The NEQMP is a Private Finance Initiative which focussing on the monitoring and assessment of the environmental quality. The NEQMP comprises of Air Quality Monitoring, Water Quality Monitoring (River and Marine) and Environmental Data Centre. The monitoring modes include the 65 Continuous Air Quality Monitoring Stations (CAQMS), 14 Manual Air Quality Monitoring Stations (MAQMS) and three Mobile Continuous Air Quality Monitoring Stations (MCAQM) (Pakar Scieno TW, 2018).