FEATURE SELECTION AND MODEL PREDICTION OF AIR QUALITY USING $PM_{2.5}$

SHARON DING TIEW KUI

UNIVERSITI SAINS MALAYSIA

2018

FEATURE SELECTION AND MODEL PREDICTION OF AIR QUALITY USING $PM_{2.5}$

by

SHARON DING TIEW KUI

Thesis submitted in partial fulfilment of the requirement for the degree of Bachelor of Chemical Engineering

June 2018

ACKNOWLEDGEMENT

First and foremost, I would like to convey my greatest gratitude and appreciation to my supervisor, Associate Professor Ir. Dr. Zainal for his invaluable advice, guidance and encouragement for me throughout final year project.

Besides that, I would also like to extend my gratitude towards all the administrative, academic and technical staffs of the School of Chemical Engineering, USM for giving a helping hand and constructive opinions throughout the project. A special gratitude towards the Dean of our school, Professor Dr. Azlina bt Harun @ Kamaruddin for her support toward all the final year students on our research projects.

Furthermore, I am grateful to my family for their love, care and support throughout the project. Their contributions are very much appreciated.

Last but not least, I would like to thanks my friends for their valuable advice and support on my project. They are willing to share their valuable knowledge which are very helpful for my research project.

SHARON DING TIEW KUI

June 2018

TABLE OF CONTENTS

		PAGE
ACKNOWLEDGEMENT		ii
LIST OF FIGURES		vi
LIST OF TABLES		vii
LIST OF ABBREVIATIONS		viii
LIST OF SYMBOLS		xi
ABSTRAK		xii
ABSTRACT		xiii
CHAPTER 1 INTRODUCTION		1
1.1	Research Background	1
1.2	Problem Statement	3
1.3	Objectives	4
1.4	Scope of Study	4
1.5	Outlines of Thesis	5
CHAPTER 2 LITERATURE REVIEWS		6
2.1	Introduction	6
2.2	Air Quality	6
2.3	$PM_{2.5}$	11
2.3.1	Particulate Matter	11
2.3.2	Why $PM_{2.5}$?	12
2.3.3	Sources of $PM_{2.5}$	13

2.3.4 Health and Environmental Effects of $PM_{2.5}$ 15

2.3.4 (a) Health Effects 15

2.3.4 (b) Environmental Effects 15

2.4 Model Prediction of Air Quality 16

2.4.1 Artificial Neural Network 17

2.5 Feature Selection 21

2.5.1 Random Forest 22

2.6 Summary and Findings of Study 24

CHAPTER 3 METHODOLOGY 26

3.1 Introduction 26

3.2 Overall Study Structure 27

3.3 Case study: $PM_{2.5}$ of Five Chinese Cities 28

3.4 Data Screening 30

3.5 Input Feature Selection for Modelling: Random Forest 31

3.6 Artificial Neural Network Prediction Model Development 31

3.6.1 Model Validation 33

CHAPTER 4 RESULTS AND DISCUSSION 35

4.1 Introduction 35

4.2 Removal of Outliers 35

4.3 Artificial Neural Network Prediction Model 40

4.4 Artificial Neural Network Prediction Model with Feature Selection 44

4.5 Comparison of Prediction Model With and Without Feature Selection 47

CHAPTER 5 CONCLUSION AND RECOMMENDATION 49

5.1 Conclusion 49

5.2 Recommendation 50

REFERENCE 51

APPENDIX 61

Appendix A Matlab Code for Removal of Outlier 61

Appendix B Matlab Code for Development of Regression Trees 63

LIST OF FIGURES

		PAGE
Figure 2.1	Size comparison of PM particles	11
Figure 2.2	Concentration of $PM_{2.5}$ in Malaysia	13
Figure 2.3	Sources of $PM_{2.5}$ in urban area	14
Figure 2.4	Sources of $PM_{2.5}$ in suburban area	14
Figure 2.5	Sources of $PM_{2.5}$ in rural area	14
Figure 2.6	The typical structure of a multilayer neural network	20
Figure 3.1	Overall study structure	27
Figure 3.2	Geographical location of the US diplomatic posts (blue), China's MEP sites (red), and meteorological sites (black)	29
Figure 4.1	Trend of dew point (a) before and (b) the removal of outliers	37
Figure 4.2	Trend of humidity (a) before and (b) the removal of outliers	37
Figure 4.3	Trend of pressure (a) before and (b) the removal of outliers	38
Figure 4.4	Trend of temperature (a) before and (b) the removal of outliers	38
Figure 4.5	Trend of wind speed (a) before and (b) the removal of outliers	39
Figure 4.6	Graph of MSE vs number of neuron	41
Figure 4.7	Graph of R^2 vs number of neuron	41
Figure 4.8	Graph of predictor importance estimates	46
Figure 4.9	Trend of MSE with different category of predictors	46
Figure 4.10	Trend of R^2 with different category of predictors	47

LIST OF TABLES

		PAGE
Table 2.1	Malaysian air quality guidelines	8
Table 2.2	Indicator of API value with level of pollution and health measures	9
Table 2.3	New Malaysian air quality guidelines	10
Table 2.4	Summary of previous study in $PM_{2.5}$ prediction	18
Table 2.5	Summary of used of feature selection in previous air quality study	23
Table 3.1	Phase of study	28
Table 3.2	List of input and target variables	34
Table 3.3	Setting parameter for FANN prediction model	34
Table 4.1	Analysis for $PM_{2.5}$ before and after removal of outliers	40
Table 4.2	Performance of prediction model with different number of neurons	43
Table 4.3	Category of predictors after apply feature selection	45
Table 4.4	Performance of prediction model with feature selection	45
Table 4.5	Comparison of performance for prediction model with and without feature selection	47

LIST OF ABBREVIATIONS

AFNN	Adaptive Fuzzy Neural Network
ANN	Artificial Neural Network
API	Air Pollution Index
AQI	Air Quality Index
AQMS	Air Quality Monitoring Station
ARIMA	Auto Regressive Integrated Moving Average
ARMA	Auto Regression Moving Average
BPNN	Back Propagation Neural Network
BRT	Boosted Regression Trees
BTs	Boosted Trees
CMA	Central Metrological Agency
CMAQ	Community Multiscale Air Quality
CO	Carbon Monoxide
DAL	Deep Air Learning
DOE	Department Of Environment
EEMD-GRNN	Ensemble Empirical Mode Decomposition-General Regression Neural Network
FANN	Feed-Forward Artificial Neural Network
FFNN	Feed-Forward Neural Network
GA	Genetic Algorithms
GRNN	General Regression Neural Network
HMM	Hidden Markov Models
IQR	Interquartile Rule
LASSO	Least Absolute Shrinkage And Selection Operator

LS-SVM	Least Squares Support Vector Machines
MAAGs	The Malaysian Air Quality Guidelines
MEP	China's Ministry Of Environmental Protection
MISO	Multi-Input Single-Output
MLP	Multilayer Perceptron
MLR	Multilinear Regression
MSE	Mean Of Square Error
NLR	Nonlinear Regression
NN	Neural Network
NO_2	Nitrogen Dioxide
O_3	Ground Level Ozone
PCA	Principal Component Analysis
PCR	Principal Component Regression
PM	Particulate Matter
PM_{10}	Particulate Matter With The Size Less Than 10 Micron
$PM_{2.5}$	Particulate Matter With The Size Less Than 2.5 Micron
PMI	Partial Mutual Information
RBF	Radial Basis Function
RBFNN	Radial Basis Function Neural Network
RF	Random Forest
SMLP	Square Multilayer Perceptron
SO_2	Sulphur Dioxide
SVM	Support Vector Machines
SVR	Support Vector Regression
TSP	Total Suspended Particulate

USEPA	United States Environmental Protection Agency
WHO	World Health Organization
WRF	Weather Research And Forecasting
WRF-Chem	Weather Research and Forecasting model coupled with Chemistry
WRF-CMAQ	Weather Research And Forecasting - Community Multiscale Air Quality

LIST OF SYMBOLS

R^2	Coefficient of Determination
N	The Number of Neuron
n_i	Number of Input
x_i	Original Data
x_{max}	Maximum Value Among Original Data
x_{min}	Minimum Value Among Original Data
x_n	Normalized Data
$y_{meas,i}$	Measured Response
$y_{pred,i}$	Predicted Response

PEMILIHAN CIRI DAN RAMALAN PEMODELAN KUALITI UDARA DENGAN $PM_{2.5}$

ABSTRAK

Kajian ini adalah untuk menjana model peramalan rangkaian neural suap depan (FANN) yang sesuai untuk meramalkan kualiti udara dengan menggunakan $PM_{2.5}$. Kini, Malaysia masih belum mempunyai model peramalan untuk kepekatan $PM_{2.5}$. Jadi, dengan model peramalan yang dijana, kepekatan $PM_{2.5}$ dalam udara dapat diramalkan dengan menggunakan parameter meteologi. Kaedah utama yang diselidik dalam kajian ini ialah bilangan neuron lapisan tersembunyi. Prestasi model peramalan telah dianalisa dan dinilai dengan menggunakan nilai min ralat kuasa dua (MSE) dan pekali penentuan (R^2). Dengan menambahkan bilangan neuron dalam lapisan tersembunyi, nilai MSE dapat dikurangkan manakala nilai R^2 ditambahkan. 10 neuron lapisan tersembunyi memberikan prestasi yang terbaik antara bilangan neuron yang diselidik. Oleh sebab prestasi model peramalan yang rendah, pemilihan ciri telah diperkenalkan untuk menyingkirkan parameter yang tidak berkaitan dalam set data. Hutan rawak (RF) telah ditumbuhkan dengam 200 pokok regresi untuk menentukan peramal yang paling penting. Peramal yang kurang penting telah disingkirkan daripada peramal yang lain. Dengan penyingkiran parameter yang tidak berkaitan, kejituan model peramalan telah dipertingkatkan dengan peningkatan prestasi model. Selain itu, keserasian model peramalan juga dapat dikurangkan dengan mengurangkan latihan masa model peramalan. Peramal yang disingkirkan oleh pemilihan ciri dalam kajian ini ialah tekanan, titik embun, curah hujan setiap jam dan curah hujan kumulatif. Maka, jelasnya bahawa prestasi model peramalan dengan pemilihan ciri adalah lebih baik daripada prestasi model peramalan tanpa pemilihan ciri.

FEATURE SELECTION AND MODEL PREDICTION OF AIR QUALITY USING $PM_{2.5}$

ABSTRACT

This study was to develop a feed-forward artificial neural network (FANN) prediction model to predict the air quality using $PM_{2.5}$. Currently, Malaysia does not have any prediction model for concentration of $PM_{2.5}$. Thus, with the prediction model developed, the concentration of $PM_{2.5}$ in air can be predicted by using meteorological variables. The main parameter that investigated in this study was the number of neuron of hidden layer. The performance of the prediction model was analysed and evaluated by using mean square error (MSE) and Coefficient of Determination (R^2) values. With the increasing of the number of neuron of hidden layer, MSE decreased and R increased. 10 neuron of hidden layer gave the best performance among the number of neuron investigated. Due to the low performance of the prediction model, feature selection was introduced to remove irrelevant variables in data set. Random forest (RF) was grew with 200 regression trees to decide the importance of the predictors. The predictors which was less important were removed from the predictors. With the removal of the irrelevant variables, the precision of the prediction model increased with increased of the performance of the model. Besides that, the complexity of the prediction model also reduced by decreasing training time of the prediction model. The predictors removed by feature selection in this study were pressure, dew point, hourly precipitation and cumulated precipitation. Thus, it was clearly seen that the performance of prediction model with feature selection was better than prediction model without feature selection.

CHAPTER 1

INTRODUCTION

1.1 Research Background

Air is always around us and all living things need air for survive. Human and animals need oxygen in air for respiration to generate life energy, while plant needs carbon dioxide in air for photosynthesis to produce energy.

But recently, environment pollution such as air pollution, water pollution, noise pollution and land resources shortage have attracted increasing attention with economic and population increase in cities. Among the problems, air pollution increased public awareness in both developed and developing countries as air pollution have direct impact to human's health through the short term and long term exposures to air pollutants (Kim et al., 2013; Kurt and Oktay, 2010; Gordon, 2003). World Health Organization (WHO) estimated that 6.5 million deaths were associated with air pollution in 2012 as the result of increased mortality from chronic obstructive pulmonary disease, lung cancer, heart disease and stroke. This is 11.6% of all global deaths. Thus WHO has convened a Global Platform on Air Pollution and Health with experts across academia and government to improve methods of monitoring and surveillance of air pollution exposures, ensuring open-across to air quality data (W.H.O., 2013).

In Malaysia, Air Pollution Index (API) is used to report the status of air pollutants and API was widely accepted as important determinants of adverse health effects (Hajek and Olej, 2015). Particulate matter normally is the dominant air pollutant with highest concentration among all the air pollutants in Malaysia. Previously, PM_{10} is usually used as a standard to measure air pollution, but since recent studies shown that smaller particles have greater impact on human's health (Spurny, 1998), control of the particles become very urgent. Most of the secondary particles which produced by chemical reactions in the atmosphere, have the size less than 2.5 micron which known as $PM_{2.5}$ (Harrison et al., 1997).

Therefore, monitoring and predicting of air quality was very important due to health impacts cause by air pollution especially $PM_{2.5}$ which is very harmful and dangerous. Prediction of air quality play an important role in air quality management system. The air quality predictors usually apply for health alert, supplementing existing emission control program, operational planning and emergency response (C.E.N.R., 2001). Besides that, the effects $PM_{2.5}$ can be effectively controlled by providing adequate and efficient air quality control and mitigation measures that can be designed and tested with the aid of air quality models. Air quality regulatory agencies have to complement measurements of air quality with models that can accurately predict pollutant concentrations and determine the cause of the air quality problems.

Artificial Neural Network (ANN) is one of the most popular model in air quality prediction. The model is inspired from the neurological system of humans and used to mimic the human neurological system. ANN is a mathematical model of a natural neural network. It uses a computational or mathematical model based on connectionist approach for solving problems. After all, it shows a remarkable success in the modeling and prediction of higher nonlinear systems including air quality prediction case.

To develop a prediction model, suitable input variables are very important as the condition of input data may affect the network of the model. Hence, feature selection are needed to reduce the the irrelevant input variables so that the precision of the model can be improved. The less complex prediction model with feature selections is more cost-effective and faster.

1.2 Problem Statement

Recently, as the increased concern about environment issues has encouraged researcher from each country to focus on monitoring, predicting and controlling the environmental quality such as air quality.

Currently in Malaysia, there are no any prediction methods used by Department of Environment (DOE) for air quality yet (APIMS, 2018). In Malaysia, The API value announced by DOE were mostly calculated in a complex method which involve the sub-index calculation. Moreover, the API presented only include the highest sub-API value which neglect the effects of other pollutants.

Besides that, Malaysia also have not included the calculation of $PM_{2.5}$ into API as one of the pollutant. Most of the developed countries had include $PM_{2.5}$ as one of the pollutant in determine the air quality. For example, our neighbour country, Singapore already included $PM_{2.5}$ since 2014. While Malaysia still in midst of finalizing a new guideline to include $PM_{2.5}$ as one of the air pollutant in API calculation. Compared with PM_{10}, $PM_{2.5}$ is much more dangerous due to the smaller in particles which will cause serious damage in respiratory system. Therefore, it is important to know the concentration of $PM_{2.5}$ in air to prevent the health effect caused by $PM_{2.5}$.

Most of the researchers use meteorological variables in predicting the API such as temperature, humidity, wind speed and etc. But among the meteorological variables, some variables are irrelevant in predict the API. The present of irrelevant variables will decrease the precision of the prediction model. So, feature selection is introduced to remove the irrelevant variables in this study to investigate and study the feasibility of feature selection in increased the precision of prediction model.

Thus, this study was to develop a model prediction by using FANN with feature selections to predict the concentration of $PM_{2.5}$ in air. The effect of an air pollution peak can be reduced on the surrounding population and ecosystem by an accurate air quality predicting (Peng, 2015).

1.3 Objectives

The objectives of the thesis are defined as follows:

1. To apply Random Forest as features selection for determining suitable predictor variables for prediction models.
2. To develop predictions model for predicting the concentration of $PM_{2.5}$ by using Artificial Neural Network with and without feature selection.
3. To compare the performances of FANN model with and without feature selection.

1.4 Scope of Study

In this thesis, Artificial Neural Network (ANN) was used to predict the air quality by using $PM_{2.5.}$ The ANN model was built in Matlab automatically. Among the type of ANN, feed-forward propagation artificial neural network (FANN) is chosen in develop the prediction model.

The input variables are reduced by using Random Forest (RF) to reduce the complexity and increase the precision of the prediction model.

The performance of the model was analysed by using mean square error (MSE) and Coefficient of Determination value (R^2). The performance of the ANN and ANN with feature selections were then compared to determine which gave the best performance.

1.5 Outlines of Thesis

The following are the contents for each chapter in this thesis:

Chapter 1 introduces the research background, problem statement, research objective and the scope of study of this thesis.

Chapter 2 presents a review of this study including air quality, $PM_{2.5}$, air quality prediction model, Artificial Neural Network, feature selection and Random Forest.

Chapter 3 outlines the methodology of this research. Case study of this study, process modelling and performance criteria are covering.

Chapter 4 discusses about the result and evaluation of FANN prediction model performance in detail.

Chapter 5 concluded all the findings in this study. Recommendation and suggestion are included as well.

CHAPTER 2

LITERATURE REVIEWS

2.1 Introduction

In this chapter, the overview and background of air quality in term of $PM_{2.5}$ are firstly looked into. The source of $PM_{2.5}$and the effect of $PM_{2.5}$ to human health and environment is studied in this study.

Besides that, the prediction model with the implementation of artificial neural network (ANN) in air quality prediction model is also studied. A summarized of previous study of prediction model of $PM_{2.5}$ is shown in Table 2.4.

Furthermore, feature selection is studies in term of importance and classes of feature selection. Some feature selection used in previous study air quality prediction models are summarized in Table 2.5. Besides that, the selected feature selection which are Random Forest (RF) also been introduced.

2.2 Air Quality

In Malaysia, air quality is monitored manually and continuously via 52 Air Quality Monitoring Station (AQMS) throughout Malaysia. AQMS used to monitor continuously 5 major pollutants which are particulate matter, ozone, carbon monoxide, nitrogen dioxide and sulphur dioxide (Essays, 2013).

Department of Environment Malaysia (DOE) reports the air quality status in Malaysia in term of air pollution index (API). API is developed nearly follows the

United States Environmental Protection Agency (USEPA) Pollution Standards Index. It provides an easily comprehensible information about the air pollution level as shown in Table 2.2.

API is calculated based on The Malaysian Air Quality Guidelines (MAAGs). Table 2.1 shows the guidelines which derived from available scientific and human health data. MAAGs adopted 5 pollutants criteria which are particulate matter with the size less than 10 micron (PM_{10}), sulphur dioxide (SO_2), carbon monoxide (CO), nitrogen dioxide (NO_2), and ground level ozone (O_3). The API calculated is based on the average of concentration of air pollutants including PM_{10}, SO_2, NO_2, CO and O_3. The dominant air pollutant with highest concentration will determine the API value. In Malaysia, PM_{10} normally is the dominant air pollutant (APIMS, 2018).

Most of the developed countries had included particulate matter with size less than 2.5 micron ($PM_{2.5}$) in air quality indicator as researcher from USEPA found that fine particulate which refer to $PM_{2.5}$ is more dangerous compared to PM_{10}. Air with high concentration of $PM_{2.5}$ will cause lung and cardiovascular diseases. For example, USEPA has set its National Ambient Air Quality Standard limit for $PM_{2.5}$ at 15 $\mu g/m^3$ for annual average and 65 $\mu g/m^3$ for 24-hour average. While the Europe Union targeted annual average of $PM_{2.5}$ at 25 $\mu g/m^3$ (Choong, 2012).

Currently, Malaysia have included $PM_{2.5}$ in API. But, the new Malaysian Air Quality Guidelines is in the midst of finalising by DOE to include the standard limit of $PM_{2.5}$ in the ambient air. The guideline which establish to replace the older MAAGs that have been used seen 1989 is based on World Health Organisation (WHO) 2006 Guidelines.(D.O.E, 2015) DOE is currently coming up with $PM_{2.5}$ Air Quality Index

System and data integration with the existing system in Environment Data Centre prior to including $PM_{2.5}$ in API calculation (APIMS, 2018).

The air pollution concentration limit including $PM_{2.5}$ will be strengthen in stages until 2020. 3 interims targets are set which include interim target 1 (IT-1) in 2015, interim target 2 (IT-2) in 2018 and the full implementation of standard in 2020 (D.O.E, 2015). Table 3 shows the air pollution concentration limit in new Malaysian Air Quality Guidelines.

Table 2.1 Malaysian air quality guidelines (APIMS, 2018)

Pollutants	**Average Time**	**Malaysia Guidelines**	
		ppm	µg/m^3
PM_{10}	24 hours	-	50
	1 year	-	150
SO_2	10 minutes	0.19	500
	1 hour	0.13	350
	24 hours	0.04	105
NO_2	1 hour	0.17	320
	8 hours	0.04	75
O_3	1 hour	0.10	200
	8 hours	0.06	120
CO*	1 hour	30	35
	8 hours	9	10

*mg/m^3

Table 2.2 Indicator of API value with level of pollution and health measures (D.O.E, 1997)

API	Condition	Level of Pollution	Health Measures
0-50	Good	Pollution low and has no ill effects on health.	No restriction of activities for all groups of people.
51-100	Moderate	Moderate pollution and has no ill effects on health.	No restriction of activities for all groups of people.
101-200	Unhealthy	Mild aggravation of symptoms among high risk persons, like those with heart or lung disease.	Restriction of outdoor activities for high-risk persons. General population should reduce vigorous outdoor activity.
200-300	Very Unhealthy	Significant aggravation of symptoms and decreased exercise tolerance in person with heart or lung disease.	Elderly and persons with known heart or lung disease should stay indoors and reduce physical activity.
More than 300	Hazardous	Severe aggravation of symptoms and endangers health.	Elderly and persons with known heart or lung disease should stay indoors and reduce physical activity. General population should reduce vigorous outdoor activity.

Table 2.3 New Malaysian air quality guidelines (D.O.E, 2015)

		Malaysia Guidelines		
Pollutants	**Average Time**	IT-1 (2015)	IT-2 (2018)	Standard (2020)
		μg/m^3	μg/m^3	μg/m^3
PM_{10}	1 year	50	45	40
	24 hours	150	120	100
$PM_{2.5}$	1 year	35	25	15
	24 hours	75	50	35
SO_2	1 hour	350	300	250
	24 hours	105	90	80
NO_2	1 hour	320	300	280
	24 hours	75	75	70
O_3	1 hour	200	200	180
	8 hours	120	120	100
CO*	1 hour	35	35	30
	8 hours	10	10	10

*mg/m^3

2.3 $PM_{2.5}$

2.3.1 Particulate Matter

Particulate Matter (PM) also called as particle pollution is defined as a mixture of liquid droplets and solid particles found in air. Some of the particles are large, while some of the particles are very tiny. Large particles such as dirt, dust, smoke and soot can be seen with the naked eyes. While tiny particles such as particle-bound water, metals, allergens and microbial compounds can only be detected by using an electron microscope.

Generally, particulate matter can categorize into two groups in term of the particle size. The two groups are PM_{10} and $PM_{2.5}$. PM_{10} which also called as coarse particles refer to particulate matter with size less than 10 micrometres while $PM_{2.5}$ which also called as fine particles are particulate matter with size less than 2.5 micrometres. Figure 2.1 shows the size comparison of particulate matter with human hair. Human hair is about 70 micrometres averagely in diameter (E.P.A, 2018).

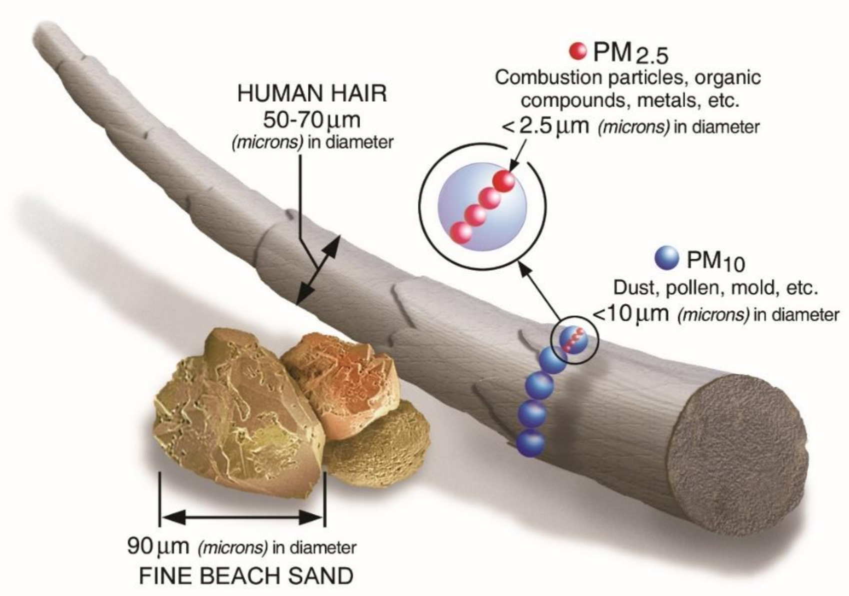

Figure 2.1 Size comparison of PM particles (E.P.A, 2018)

2.3.2 Why $PM_{2.5}$?

As the size of the particles decrease, the particles are more dangerous to humans. USEPA officially recognized this statement in July, 1987. The older EPA standard for total suspended particulate (TSP) is replaced by PM_{10} and $PM_{2.5}$ which specify the size of the particles (Smoke, 2010).

$PM_{2.5}$ is much more dangerous compared to PM_{10}. But, currently much of the countries only include PM_{10} in air quality standard but no $PM_{2.5}$. Thus, WHO pushed all countries to have standards for $PM_{2.5}$.

The smaller the particle, the particles tend to stay longer in the air, the higher the chances of human inhaling the particles (Miettinen, 2018). PM_{10} which filtered by nose and throat can pass into the large airways, while $PM_{2.5}$ can penetrate deeply into the lungs and move directly into blood stream.

The standards for $PM_{2.5}$ have been implemented as a guide for clean air. Europe Union air quality standards limit the concentration of $PM_{2.5}$ at 25 $\mu g/m^3$ annually, while UN'S WHO has guidelines recommending annual exposure limits at 10 $\mu g/m^3$ for $PM_{2.5}$. Singapore's annual target for $PM_{2.5}$ target for $PM_{2.5}$ is also 10 $\mu g/m^3$. Furthermore, the USEPA's standards for average annual $PM_{2.5}$ level is 10 to 15 $\mu g/m^3$.

Although $PM_{2.5}$ is much more important and dangerous than PM_{10}, but PM_{10} measurement is still needed as the coarser fraction in PM_{10} which between 2.5 micron to 10 microns is also main cause of air pollution. The coarser fraction usually is from dust resuspension which from dust storms and road dust.

2.3.3 Sources of $PM_{2.5}$

Particles can categorize into primary PM and secondary PM. Primary PM is the particles emitted directly into the air or formed in the atmosphere from gaseous precursors. Primary PM can have both anthropogenic and non-anthropogenic sources. Anthropogenic sources include combustion of engines, solid-fuel, combustion of energy production and industrial activities (W.H.O., 2013).

Secondary particles are formed through chemical reactions of gaseous pollutants. They are products of SO_2 resulting from the combustion of sulphur-containing fuels and atmospheric transformation of NO_x. Secondary particles are mostly found in $PM_{2.5}$ (W.H.O., 2013).

In Malaysia, the main sources of $PM_{2.5}$ is anthropogenic activities such as open burning and traffic emission. The concentration of $PM_{2.5}$ is highest at urban area, followed by suburban area and rural area as shown in Figure 2.2. The main sources for urban, suburban and rural areas are motor vehicles or soil dust, domestic waste combustion and biomass combustion. The main sources of $PM_{2.5}$ for each area are summarized in Figure 2.3, 2.4 and 2.5 respectively (Ee-Ling et al., 2015).

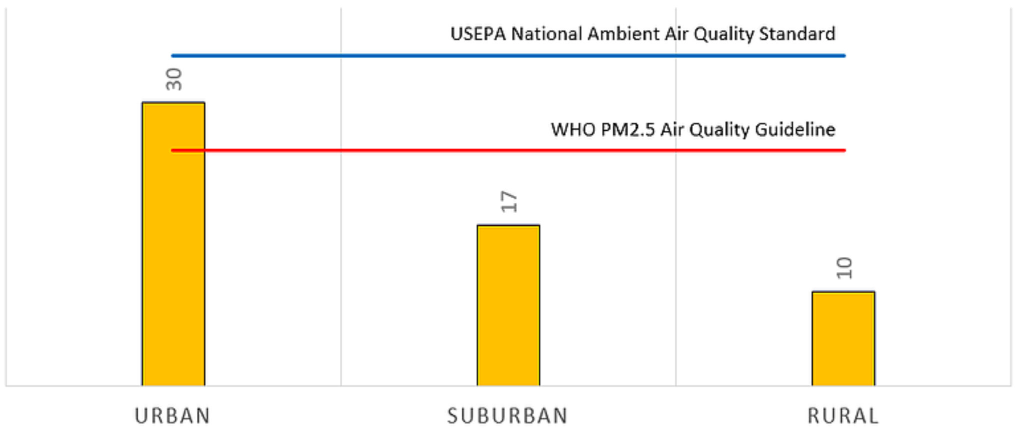

Figure 2.2 Concentration of $PM_{2.5}$ in Malaysia (Ee-Ling et al., 2015)

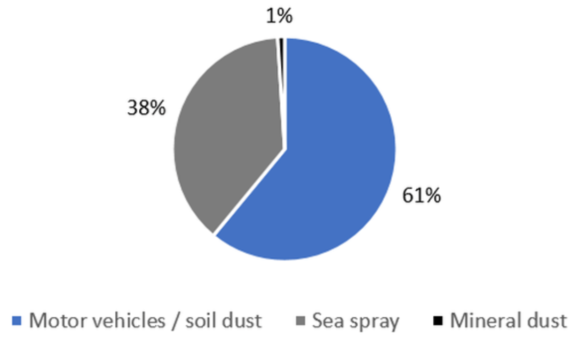

Figure 2.3 Sources of $PM_{2.5}$ in urban area (Ee-Ling et al., 2015)

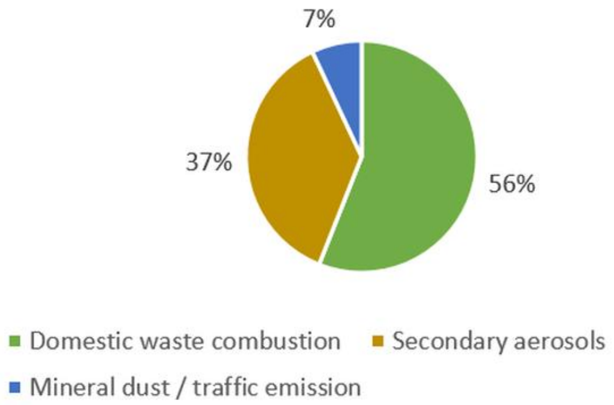

Figure 2.4 Sources of $PM_{2.5}$ in suburban area (Ee-Ling et al., 2015)

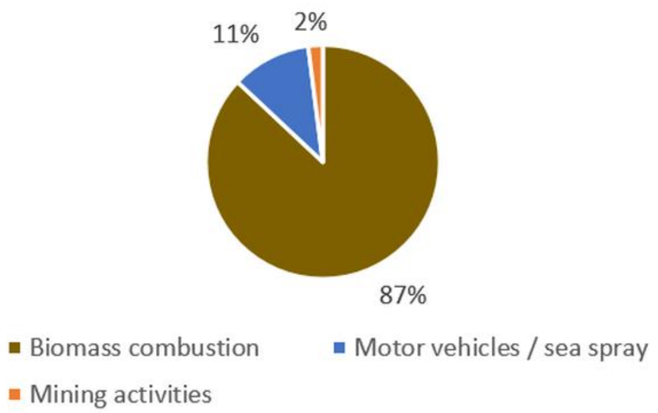

Figure 2.5 Sources of $PM_{2.5}$ in rural area (Ee-Ling et al., 2015)

2.3.4 Health and Environmental Effects of $PM_{2.5}$

2.3.4 (a) Health Effects

$PM_{2.5}$ include inhalable particles which are tiny enough to penetrate deeply into respiratory system. Short term exposures to $PM_{2.5}$ can cause respiratory and cardiovascular morbidity, such as aggravation of asthma and respiratory symptoms while long term exposures can cause mortality from cardiovascular and respiratory disease and form lung cancer. These have been proven by scientists from Canada and United States (Franklin et al., 2008; Schwartz, 2000).

In Europe Union countries, average life span decreased for 8.6 months due to exposures to $PM_{2.5}$. 5% of lung cancer deaths and 3% of cardiopulmonary deaths are estimated due to PM globally. Lim SS et al. estimated that in 2010, annual $PM_{2.5}$ accounted for around 3.1% of global disability-adjusted life years and 3.1 million deaths (Lim and Vos, 2012).

2.3.4 (b) Environmental Effects

$PM_{2.5}$ can be carried by wind over a long distances and then settle on water or ground. These settling may cause some environmental effects depends on their chemical composition include acidify lakes and streams, affect the nutrient balance of river, deplete the nutrients in soil, affect diversity of ecosystem and cause acid rain. $PM_{2.5}$ also cause haze and reduce the visibility (E.P.A, 2018).

2.4 Model Prediction of Air Quality

An air pollution prediction conducted for better reflection of the changing trend of air pollution to provide efficient and complete environmental quality information for environment decisions to avoid severe air pollution accidents (Chen et al., 2013). It is due to adverse health impact due to air pollution protected air pollution control by obtaining real-time air quality information (Zheng et al., 2013).

Many researches have focused on air quality predictions where the predictors are generally classified into two type which are deterministic and statistical. A deterministic method employs theoretical chemical models and meteorological emissions (Jeong et al., 2011; McHenry et al., 2004) to stimulate pollutant discharge, pollutants' transfer, diffusion and removal processes using dynamic data of a limited number of monitoring stations in a model-driven way (Kim et al., 2010; Baklanov et al., 2008). Representative methods such as the Community Multiscale Air Quality (CMAQ) model (Chen et al., 2014) and the Weather Research and Forecasting and Community Multiscale Air Quality (WRF-CMAQ) model (Saide et al., 2011) are usually used for urban air quality prediction. However, the prediction results suffer from low prediction accuracy due to incomplete theoretical data, unreliable pollutant emission data and complicated underlying surface conditions (Vautard et al., 2007).

Statistical methods simply use a statistical modelling technique to predict the air quality in data-driver manner compared with complicated deterministic methods (Li et al., 2016). Air quality prediction commonly used straightforward methods such as the auto regression moving average (ARMA) model (Box, 1976) and the multilinear regression (MLR) model (Li et al., 2011). But these methods usually produce limited accuracy as the result of their inability to model non-linear pattern. Therefore, these methods cannot predict extreme concentration of air pollutants (Goyal et al., 2006).

Artificial neural networks (ANN) (Gardner and Dorling, 1998; Hooyberghs et al., 2005; Lal and Tripathy, 2012; Bernardo Sánchez et al., 2013) and support vector regression (SVR) (Garcia Nieto et al., 2013; Hájek and Olej, 2012) are used as promising alternative to these linear models. ANN is more accurate compared to linear models due to the air quality data presented clearer nonlinear pattern than linear pattern (Prybutok et al., 2000). Some studies also combined these models for air quality prediction and shown that the hybrid methods performance better than single models (Díaz-Robles et al., 2008; Bernardo Sánchez et al., 2013; Spurny, 1998; Chen et al., 2013). Table 2.4 shown the summary of the previous study of prediction model on the concentration of $PM_{2.5}$.

2.4.1 Artificial Neural Network

Artificial neural network (ANN) is one of the branches of artificial intelligence. It consists of massively interconnected nonlinear memoryless processing elements which called as nodes or neurons. ANN is a self-adaptive, data driven and black-box method which learns from examples. The network can always correctly estimate on a population when trained with sufficient data even if the underlying relationships are unknown and hard to describe as it is the nonlinear nature of the real-world events generally. Therefore, ANN is frequently included in air quality predicting (Xie et al., 2009).

ANN are designed to imitate the characteristic of the human brain which comprises interconnected synaptic neurons capable of learning and storing information about their environment (Bishop, 1995).

Table 2.4 Summary of previous study in $PM_{2.5}$ prediction

Publications	Input Variables	Target Variables	Models	Location
Kleine Deters et al. (2017)	Meteorological variables and Pollution variables	$PM_{2.5}$	WRF-Chem, CMAQ, NN, BTs and LS-SVM	Quito, Ecuador
Jiang et al. (2017)	Meteorological variables	$PM_{2.5}$	AFNN, LS-SVM	China
Ni et al. (2017)	Meteorological variables	$PM_{2.5}$	BPNN, ARIMA	Beijing, China
Sun and Sun (2017)	Pollution variables	Daily $PM_{2.5}$	GRNN, PCA and LS-SVM	Beijing, Tianjin and Hebei, China
Suleiman et al. (2016)	Meteorological variables and Pollution variables	$PM_{2.5}$, PM_{10}, PNC	ANN and BRT	London
Feng et al. (2015)	Meteorological variables	Daily $PM_{2.5}$	ANN	Beijing, Tianjin and Hebei, China
Fu et al. (2015)	Meteorological variables	Daily $PM_{2.5}$, PM_{10}	FFNN	Hangzhou, Shanghai, Nanjing, China
Chen et al. (2014)	Pollution variables	Seasonally $PM_{2.5}$	CMAQ	California, US

Continued

Zhou et al. (2014)	Meteorological variables	Daily $PM_{2.5}$	EEMD-GRNN, PCR and MLR	Xi'an, China
Haiming and Xiaoxiao (2013)	Meteorological variables and Pollution variables	$PM_{2.5}$	RBFNN	Hebei. China
Sun et al. (2012)	Meteorological variables	24-hour-average $PM_{2.5}$	HMM	Northern California, US
Voukantsis et al. (2011)	Meteorological variables	Daily $PM_{2.5}$, PM_{10}	ANN-MLP, PCA	Thessaloniki, Greece and Helsinki, Finland
Cobourn (2010)	Meteorological variables	Daily $PM_{2.5}$	NLR	Louisville, Kentucky
Ordieres et al. (2005)	Meteorological variables	Daily $PM_{2.5}$	ANN: MLP, RBF and SMLP	US – Mexico border in Texas and Chihuahua
McKendry (2002)	Meteorological variables	Daily $PM_{2.5}$, PM_{10}, O_3	ANN-MLP, MLR	Vancouver, Canada
Pérez et al. (2000)	Meteorological variables	Hourly $PM_{2.5}$	FFNN	Santiago, Chile

A neuron model includes three elements which are a linear combiner which combines the weighted input signals, the connecting links characterized by their strength and an activation function for limiting the amplitude range of the neuron's output to some finite value. The neural network model structure includes three different and interconnected layers of neurons which are input layer, hidden layer and output layer (Nejadkoorki, 2011). The information is processed sequentially in the order as shown in Figure 2.6.

The ANN models are designed to perform a certain task historical data. Besides that, the training's goal is not restricted to learning and precise representation of the sets of training data, but limited to mode; statistically the process that generates the data for generalisation and precise prediction (Bishop, 1995).

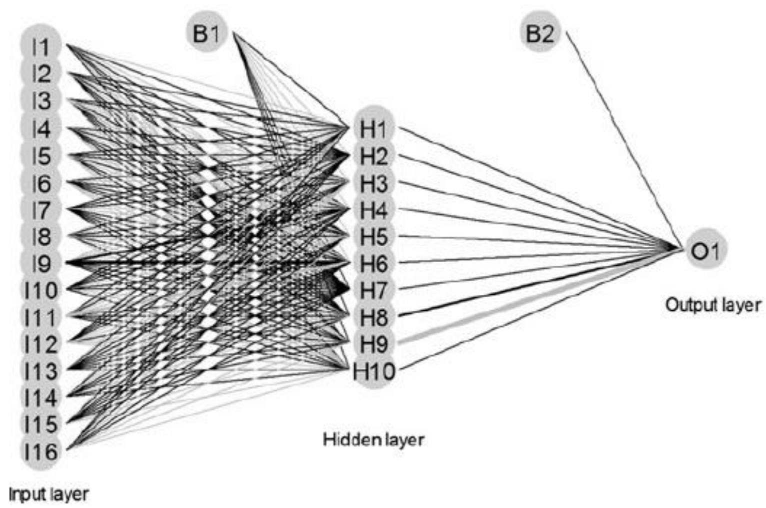

Figure 2.6 The typical structure of a multilayer neural network (Nejadkoorki, 2011)

2.5 Feature Selection

Feature Selection also known as variable selection or attribute selection. It is the automatic selection of attributes in data that are most relevant to the predictive modelling problem. Feature selection is very importance in building a model. It is needed to remove irrelevant model input variable thereby reducing the learning difficulty, computational complexity, model complexity and memory requirements (Bowden et al., 2005; Hastie et al., 2009). It also can improve the prediction accuracy as well as the ability to generalize the model. Guyon and Elisseeff stated that the objective of feature selection is three-fold which are improving the prediction performance of the predictors, providing faster and more cost-effective predictors and providing a better understanding of the underlying process that generated the data (Guyon et al., 2003).

Feature selection algorithms can be categorized into three general classes which are filter methods, wrapper methods, and embedded methods. Filter methods apply a statistical measure to assign a scoring to each feature. The features either selected to be kept or removed from the dataset by the score ranked. The methods are considering feature independently and univariate. The examples of filter methods are Chi squared test, correlation coefficient scores and information gain (Brownlee, 2014).

Wrapper methods is the selection of a set of features as a search problem. In this method, different combinations are prepared, evaluated and compared to other combinations. It is used to evaluate a combination of features and assign a score based on model accuracy in a predictive model. The example of wrapper method is the recursive feature elimination algorithm (Brownlee, 2014).

Embedded methods learn which features best contribute to the accuracy of the model while the model being created. The most common type of embedded methods are regularization methods which also known as penalization methods. These methods introduce additional constraints into the optimization of a predictive algorithm that bias the model toward lower complexity. Examples of the regularization algorithms are the LASSO, Ridge Regression and Elastic Net (Brownlee, 2014).

Table 2.6 shown a summary of previous study on feature selection used in air quality prediction model. From the summary, it is clearly shown that Random Forest and Genetic Algorithm are very common feature selection used in air quality prediction model.

2.5.1 Random Forest

Random forest (RF) is a nonparametric method that builds an ensemble model of decision trees from random subsets of features and bagged samples of the training data. RF have shown excellent performance for both regression and classification problems (Nguyen et al., 2015).

Random forest is one of the most popular machine learning methods due to their relatively good accuracy, robustness and ease of use. They also provide two straightforward methods for feature selection: mean decrease impurity and mean decrease accuracy. For mean decrease impurity, it computed the decrease of weighted impurity of each features in tree when training a tree. For a forest, the impurity decease from each feature can be averaged. While mean decrease accuracy directly measure the impact of each feature on accuracy of the model. It measured by determined the

Table 2.5: Summary of used of feature selection in previous air quality study

Publications	Feature Selection	Target Variable	Model	Location
Qi et al. (2017)	DAL	Air Quality	NN	Beijing, China
Shamsoddini et al. (2017)	RF	Daily $PM_{2.5}$, SO_2, NO_2 and CO	ANN and MLR	Tehran, Iran
Siwek and Osowski (2016)	RF and GA	Daily PM_{10}, SO_2, NO_2 and O_3	NN (MLP, RBF), SVM	Warsaw, Poland
Suleiman et al. (2016)	PCA, LASSO and Elastic-Net Regression	$PM_{2.5}$, PM_{10}, PNC	ANN and BRT	London
Yu et al. (2016)	RF	AQI	RF	Shenyang, China
Mesin et al. (2010)	PMI	Daily PM_{10}	ANN	Goteborg, Sweden
Kalapanidas and Avouris (2003)	GA	NO_2 and O_3	-	Athens, Greece

permutation for each variable in decrease the accuracy of the model (Crossentropy, 2014).

For the random forest, the tree-based strategies naturally ranked by the ability of variables to improve the purity and accuracy of the node. Nodes with the greatest decrease in impurity and with highest accuracy happen at the start of the tree, while the nodes with lowest decrease in impurity and with lowest impurity occur at the end of the trees. Therefore, we can create a subset of the most importance features by pruning trees below a particular node (Alon, 2017). In Table 2.5, same previous studies for example used random forest as feature selection in the prediction of air quality model.

2.6 Summary and Findings of Study

After all the reviews above, it was very clear that the prediction of concentration of $PM_{2.5}$ is very important. This is due to the dangerous of $PM_{2.5}$ to human health and currently, Malaysia does not have the prediction model yet for $PM_{2.5}$.

The overview of usage of ANN in prediction model is studied. FANN is chosen among several type of neuron network. Metrological variables are used as input to train the prediction model.

From the review, we can see that random forest is one of the most common feature selection in air quality prediction which with high accuracy to remove the irrelevant inputs into the prediction model. The performance of the prediction model can be improved by feature selection.