MODEL PREDICTION OF $PM_{2.5}$ AND PM_{10} USING MACHINE LEARNING APPROACH

NORFARHANAH BINTI HAMID

UNIVERSITI SAINS MALAYSIA

2021

MODEL PREDICTION OF $PM_{2.5}$ AND PM_{10} USING MACHINE LEARNING APPROACH

by

NORFARHANAH BINTI HAMID

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

July 2021

ACKNOWLEDGEMENT

The completion of this report could not have been made possible without the participation and assistance of so many people be it voluntarily or involuntarily. Their names may not all be mentioned but their contributions and sincerity are very much appreciated and acknowledged. However, a special expression of gratitude is in order, to demonstrate my deep appreciation and indebtedness particularly to the following. First and foremost, I would like to thank the supreme power Allah SWT for He is the one constantly guiding and providing a blessing for me to push forward. It is for His never-ending grace that gives me the strength and ultimately led to where I am destined to be. Next, special gratitude to my supervisor, Professor Ir. Dr. Zainal Ahmad, who's not only kind but his contribution in providing top-notch expert advice as well as stimulating suggestions and positive encouragement had helped me in continuing with this research in the first place. A special acknowledgement for Kementerian Pendidikan Malaysia (KPM) through Fundamental Research Grant Scheme (FRGS) grant number PJKIMIA/6071414 for their assistance and support for the successful completion of this study. Next is my loving family that is always there for me in lending a helping hand as well as a shoulder to cry on. Nobody has been more important to me in pursuit of completing this project. Thank you for continuously supporting and being there no matter what. Without their unconditional love and support I would not be able to make it thus far. Furthermore, I would also like to acknowledge with much appreciation to my dearest friends who provided help in sharing information and knowledge while constantly inspiring me to always be optimistic in achieving my goals.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF SYMBOLS

b_1	Slope of the regression line
b_o	Intercept of the regression line
e	Error Term
f	Nonlinear function representing ANN
M	Number of observations value
m/s	Meter per Seconds
mm	Millimeter
n	Number of process input
NAE	Normalized Absolute Error
NO	Nitrogen Oxides
NO_2	Nitrogen Dioxide
Ns	Normalized by Scaling
O_3	Ozone
R	Regression Number
R^2	Correlation of Determination
RMSE	Root Mean Square Error
SO_2	Sulfur Dioxide
u(t)	Process input at time t
ug/m^3	Micro-gram per Meter Cubic
Y(t)	Predicted process output at time t
Y_j	Observed values
σ_x	Standard deviation of the observation X
σ_y	Standard deviation of the observation Y

LIST OF ABBREVIATIONS

ANN	Artificial Neural Network
AOD	Aerosol Optical Depth
API	Air Particel Index
AR	Autoregressive
ARIMA	Autoregressive Integrated Moving Average
BP	Back Propagation
CTM	Chemical Transport Model
DOE	Department of Environment
FNN	Feed Forward Network
FTS	Fuzzy Time Series
GAM	Generalized Additive Model
GUI	Graphical User Interface
MI	Myocardial Infarction
MIMO	Multiple Input Multiple Output
MISO	Multiple Input Single Output
MLR	Multiple Linear Regression
NA	Not Available
NAN	Not a Number
NARX	Nonlinear Autoregressive with Exogenous Inputs
PCA	Principal Component Analysis
PCR	Principal Component Regression
PLS	Partial Least-Squared
PM	Particulate Matter
PM10	Particulate Matter 10
PM2.5	Particulate Matter 2.5
RTC	Real Time Correction
SDG	Sustainable Development Goals
SLP	Sea-Level Pressure
SOR-SVR	Successive Over Relaxation Support Vector Regress
SVM	Support Vector Machine
TBATS	Trigonometric regressors, Box-Cox transformation, ARMA errors, trend, and seasonality component

UCI	University of California Irvine
USM	Universiti Sains Malaysia
VIP	Variable Importance in Projection
WHO	World Health Organization

LIST OF APPENDICES

MODEL RAMALAN KEPEKATAN $PM_{2.5}$ DAN PM_{10} MENGGUNAKAN KAEDAH MESIN PEMBELAJARAN

ABSTRAK

Kajian ini dilakukan untuk menjana model berbilang-masukan-keluaran-tunggal (MISO) dan berbilang-masukan-berbilang-keluaran (MIMO) menggunakan rangkaian neural oleh perisian MATLAB bagi meramalkan kepekatan $PM_{2.5}$ dan PM_{10}, berdasarkan faktor meteorologi. Untuk tujuan kajian ini, sejarah set data telah diambil daripada Pusat Pemantauan Persekitaran Perbandaran Beijing sebagai kajian kes. Model telah dibina sebagai kegunaan generik dimana pra-pemprosesan data menggunakan dua kaedah berbeza dengan mengira pekali sekaitan dan kepentingan pemboleh ubah dalam skor unjuran (VIP), berjaya memilih input yang signifikan terhadap keluaran untuk menjana model. Kedua-dua kaedah pemilihan masukan telah memberikan keputusan yang sejajar di mana pencemar gas karbon monoksida (CO), nitrogen dioksida (NO_2) dan sulfur dioksida (SO_2) menunjukkan kaitan tertinggi dengan sasaran keluaran. Berdasarkan pemilihan masukkan, penjanaan model dibina dengan mengambil kira semua masukan dan masukan terpilih menggunakan model jaringan saraf tak auto mundur tak lelurus luar kawalan (NARX) yang mengaplikasikan 10 saraf tersembunyi dan 2 jumlah kelewatan serta menerapkan Levenberg-Marquardt sebagai algoritma latihan. Prestasi model ramalan telah dinilai dengan mengukur nilai ralat min kuasa dua (MSE), ralat punca min kuasa dua (RMSE), nombor mundur (R) dan penentuan sekaitan (R^2) Model yang mengambil kira semua masukan dan masukan terpilih telah dikaji dan dibandingkan di mana MISO Model 1 dengan semua masukan telah mendapat prestasi terbaik dengan nilai MSE, RMSE, R and R^2 sebanyak 0.0594, 0.2437, 0.9704 dan 0.9417 masing-masing untuk ujian. Sementara itu, dengan masukan terpilih nilai 0.05894, 0.2428, 0.9709 dan

0.9427 telah diperolehi. Didapati bahawa mengambil kira penghapusan pemboleh ubah yang tidak berkaitan, tidak meningkatkan ketepatan secara signifikan atau menurunkan prestasi. Sebaliknya, mengetahui faktor utama yang mempunyai hubungan berkait rapat dengan $PM_{2.5}$ dan PM_{10} menjamin ramalan kepekatan yang lebih baik. Ramalan kepekatan $PM_{2.5}$ dan PM_{10} menggunakan pembelajaran mesin tercapai akan berguna bukan sahaja untuk meningkatkan kesedaran masyarakat tetapi menambah baik pengurusan kualiti udara di Malaysia dan juga bahagian lain di dunia.

MODEL PREDICTION OF $PM_{2.5}$ AND PM_{10} USING MACHINE LEARNING APPROACH

ABSTRACT

This study was done to develop a multi-input-single-output (MISO) and multi-input-multi-output (MIMO) models using an artificial neural network by MATLAB software to predict the concentrations of $PM_{2.5}$ and PM_{10} respectively based on meteorological parameters. For the purpose of this research, the historical dataset is obtained from the Beijing Municipal Environmental Monitoring Centre to be used as the case study. The model was developed as a generic use where data pre-processing using two separate methods of calculating a correlation coefficient and variable importance in projection (VIP) scores managed to select significant input toward output for model development. Both methods of feature selection produced similar results where gaseous pollutants of Carbon Monoxide (CO), Nitrogen Dioxide (NO_2) and Sulfur Dioxide (SO_2) demonstrated the highest correlation towards the output target. Based on the feature selection, model development was built with and without input selection using the Nonlinear Autoregressive with Exogeneous Input (NARX) neural network model which made use of 10 number of hidden neurons and 2 number of delays, implementing Levenberg-Marquardt as training algorithm. The performance of the prediction model was evaluated by measuring Means Square Error (MSE), Root Mean Square Error (RMSE), Regression Number (R), and Coefficient of Determination (R^2) values as a performance validation. Models developed with and without input selections were studied and compared where MISO Model 1, without input selection obtained the best performance having MSE, RMSE, R and R^2 with values of 0.0594, 0.2437, 0.9704 and 0.9417 respectively for testing. Meanwhile, with input selection the values obtained 0.0589, 0.2428, 0.9709 and 0.9427. It was found

that taking into account the removal of the irrelevant variables does not increase precision significantly nor does it reduce the performance tremendously. Instead, knowing the key parameters with the most relation with $PM_{2.5}$ and PM_{10} would guarantee a better predicament of the concentration. Prediction of $PM_{2.5}$ and PM_{10} concentration using machine learning is achieved and useful not only to improve public awareness but the air quality management in Malaysia as well as other parts of the world.

CHAPTER 1

INTRODUCTION

1.1 Introduction

This research focuses on the development of the Artificial Neural Network (ANN) model to forecast the concentration of PM_{10} and $PM_{2.5}$ presence in the air as a generic usage. In this case the area of Beijing, China will be used as the case study to achieve the purposed objectives. Both correlation coefficient and partial least-squared method will be compared in selecting the appropriate parameters as an input for the developed model. The large data were pre-processed through normalization and managing of NAN values prior to the input selection. Three ANN models were developed using all candidate input and selected input each. The models were validated using various statistical analysis methods. Through the result of forecasted pollutant's concentration, a better decision with respect to the air quality is expected to be made not only in Beijing, China but also throughout the world.

1.2 Background

Having the advantage of breathing fresh air is truly a blessing, as air pollution is climbing up the ladder of becoming one of the biggest environmental concerns that the world is currently facing. Air pollution is a physical or chemical changes brought by natural processes or human activities that result in air quality degradation (Mabahwi et al., 2015). The causes of air pollution may vary based on the location of occurrence. However, the major contributor to it typically comes from industrial machinery, power producing stations, combustion engines and transportation (Manisalidis et al., 2020). Few places may also experience air pollution caused by natural sources such as volcanic or soil eruptions and forest fires (Manisalidis et al., 2020). Although it could

affect anywhere from developed to the developing country, low- and middle-income countries is usually the most affected by it. As of mid-2020, Afghanistan is known to be the most polluted country with the pollution index of 93.47 (*Pollution Index by Country 2020 Mid-Year*, n.d.)

According to the World Health Organization (WHO), air pollution kills approximately seven million people worldwide every year in which the data portrays that 9 out of 10 people breathe air that exceeds the guideline limits containing high levels of pollutants (*Air Pollution*, n.d.). In 2005, WHO released an Air Quality Guidelines presented with a clear assessment of thresholds for health harmful pollution levels along with its effect on human's health. The guidelines are based on the presence of air pollution indicators which includes particulate matter (PM), ozone (O_3), nitrogen dioxide (NO_2) and sulfur dioxide (SO_2). These are also the major pollutants listed by the Malaysian Department of Environment (DOE) to possess a high probability of causing significant damages to health, environment and property (Mabahwi et al., 2015).

PM which is one of the main indicators affecting air quality comes in a variety of sizes. PM_{10} and $PM_{2.5}$ are referring to particulate matter with a diameter smaller than about 10 and 2.5 microns, respectively. Due to its nature of having a small size, both PM_{10} and $PM_{2.5}$ have been recognized by many experts and scholars to possess the capability of penetrating inside the bronchi and lungs thus causing health related problems. Moreover, $PM_{2.5}$ demonstrated an even more damaging effect as it can infiltrate inside the lung barrier and enter the blood system. As a result, exposure to these particles is detrimental to human's health as it increases the risk of developing

cardiovascular and respiratory diseases, as well as lung cancer (*Air Pollution*, n.d.) which eventually may lead to fatality.

Therefore, considering the development of science and technology nowadays has led to the discovery of machine learning. Through the application of machine learning, the concentration of PM_{10} and $PM_{2.5}$ in air pollution can be forecasted to not only improve public awareness but air quality management as well (Chaloulakou et al., 2003). Other than that, having prompt and complete environmental quality information from the prediction will allow the government to take timely action for the environment (Zhang et al., 2020). Machine learning uses the technique of recognizing patterns to predict the outcome based on previously available data. The data input can be in regard to the meteorological, persistence, and co-pollutant presence (McKendry, 2002).

1.3 Problem Statement

There is no denying that air pollution is clearly one of the biggest challenges that the world is currently facing. Equipped with its small size, airborne particulate matter can easily be dispersed in the air that we breathe and pass through the bronchi and lung which ultimately may lead to a serious health condition. Furthermore, it could also be harmful to the environment as it behaves as a substantial influence causing harsh weather and extreme climate change. Therefore, it is agreed upon by many scholars regarding the vitality to supervise and understand the air quality in real time to predict the possibility of pollution due to the presence of PM_{10} and $PM_{2.5}$. Numerous researches has been published with the notion of applying machine learning to be utilized to predict the pollutant's concentration in order to create awareness along with preparing the public for better management of our air quality.

Throughout the years, humans able to witness first-hand the development and evolution of model prediction in the air quality field. As of today, multiple works of literature have proposed and developed potential methods as a solution to forecast the pollutant's concentration. These methods varying from physical deterministic model, statistical, empirical and neural network approach that can be categorized into deterministic and statistical methods (Kök et al., 2017). Due to the nature of the deterministic model which requires high computational time, statistical models have always been preferred for its simplicity approach. For example, the autoregressive integrated moving average model (ARIMA) is one of the most highly applied models to forecast time series due to its statistical properties, adaptability to represent a wide range of processes, and the ability to be extended (Castelli et al., 2020). Other than a single approach, recently hybrid models such as fuzzy models are also being widely developed in order to achieve higher prediction performance (Kök et al., 2017). The hybridization path allows the coupling of different methods into one complete model resulting in a maximize potential.

However, the main concern regarding the implementation of modelling software in predicting PM concentration, is the several limitations that it comes with. In such models, errors and inaccurate results are bound to surface as a result of multiple factors not being considered (Asadollahfardi et al., 2016). Furthermore, with the increasing amount of historical data available for analysis along with the requirement to obtain a model with higher performance and predicting accuracy, a conventional statistical model such as ARIMA is inadequate. Hence, for this research study, Artificial Neural Network is chosen to be the tool for forecasting PM_{10} and $PM_{2.5}$ concentrations. ANN is a model that behaves similarly to a human's biological neuron. It works based on adaptive learning and pattern recognition techniques, which rely on

historical datasets to identify atmospheric parameters which could influence air pollutant concentrations (Agarwal et al., 2020). ANN seems to be the appropriate choice as it holds the capacity to learn from data, works well for a given site, can handle nonlinear and chaotic chemical systems at a site, requires modest expertise along with moderate to high accuracy results which are computationally fast as well (Zhang et al., 2012). It is a well-rounded machine learning model that leverages historical information to learn the hidden relationship between data.

For the purpose of this research, the historical dataset is obtained from the Beijing Municipal Environmental Monitoring Centre. Since the raw data available is at a large scale, Partial Least Square (PLS) and Correlation Coefficient were applied as a medium of pre-processing step to ensure the data is suitable to be input into the ANN model structure. Other than that, a forecasting system with the ability to predict the concentration of PM_{10} and $PM_{2.5}$, simultaneously will be introduced using Multi-Output-Multi-Input (MIMO) system. This will reduce the model required to be developed as MIMO possessed the ability to produce multiple outputs in regard to the problem that requires solving. Thus, the overall aim of this project is to create a learner algorithm using an Artificial Neural Network that will be able to predict the hourly concentration of both PM_{10} and $PM_{2.5}$. A MIMO system will also be generated in order to predict the pollutant's concentration simultaneously. The model prediction method using machine a learning approach will be developed in the hope to improve the air quality sustainability as well as its management for a better lifestyle.

1.4 Research Objectives

i. To select significant input toward output for model development.

ii. To predict the concentration of $PM_{2.5}$ through the utilization of machine learning as multiple input single output (MISO).

iii. To predict the concentration of PM_{10} through the utilization of machine as multiple input single output (MISO).

iv. To model $PM_{2.5}$ and PM_{10} as multiple-input and multiple-output (MIMO).

1.5 Integration of Sustainable Development

Climate change has been one of the most detrimental global challenges caused by urbanization and pandemics, spirals to yet again, environmental degradation. With arises of numerous disasters, building a predictive model as a mean of forecasting the concentration of particulate matter are done based on the integration of sustainable development in mind. Having peace and prosperity for people and the planet has been the sole purpose of 'The 2030 Agenda for Sustainable Development' released by the Department of Economic and Social Affairs under Sustainable Development of the United Nation. This study focuses on achieving the Sustainable Development Goals (SDGs) (*THE 17 GOALS | Sustainable Development*, n.d.) particularly on the third goal of good health and well-being as well as the thirtieth goal which is on climate change.

This study strives to achieve SDG 3 which aspired to promote healthy lives and well-being for all at all ages by forecasting the presence of particulate matter and gaseous pollutants which are known to cause numerous diseases and health related problem. Through the development of predictive modelling, the composition of air quality specifically in the presence of $PM_{2.5}$ and PM_{10} can be identified and forecasted to ensure adequate mitigation strategies in terms of people's health and safety are able

to be implemented wisely. Apart from that, this study also strives to combat the climate change phenomenon through achieving SDG 13. Air pollution which disguised as a multiplier threat with the possibility to worsen some of humanity's greatest challenges is the biggest contributor to climate change. Since it is highly dependent on weather, climate change could significantly impact air quality (M. Heshmati, 2021). Predicting the concentration of particulate matter through the utilization of machine learning could significantly improve the management of air quality in Malaysia apart from pushing forward the notion of data digitalization Thus, integration of sustainable development was applied in the hope as to reduce and control the air pollution from early detection of air pollutant concentration to achieve a healthy lifestyle and better climate for the environment.

1.6 Scope of the Thesis

The scope of the study for this project is to solve the problem stated above in the problem statement through modelling. The development of model prediction is designed as a generic use, however, the case study for this project is specifically focused on the prediction of PM_{10} and $PM_{2.5}$ concentration in Beijing, China.

The project focuses on the feasibility of building a predictive model using ANN along with providing model performance evaluation to ensure a minimize error is obtained. Model prediction of air pollutant is studied in an effort to crush air pollution and provide better awareness and management of air quality to the public to prevent health and environmental risks possibly arise from the situation.

Both the MISO and MIMO system were developed by using MATLAB software which was granted access from the license provided by Universiti Sains Malaysia, USM. The software provides a user-friendly interface with pre-programmed

data available where users can practice and experiment as well prior to the model prediction development. The ANN model is available in the deep learning toolbox of the software while the historical data sets retrieved from the Beijing Municipal Environmental Monitoring Centre through the official website of UCI Machine Learning Repository is available in Excel form.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

In the previous chapter, the detrimental effects that air pollution possessed are really concerning and therefore, the need to forecast the concentrations of these air pollutants specifically PM_{10} and $PM_{2.5}$ in the atmosphere have been proposed specifically utilizing the application of machine learning approach. In the view of aforementioned observations, Chapter 2 presents the previous discoveries and reviews available from credible scientific records and references that are related to this final year project topic. This chapter covers the overview of air pollutants which also behaves as the parameters for this research study, related model prediction used from the previous study as a forecasting approach to these pollutants and lastly the utilization of ANN as the prediction model to predict PM_{10} and $PM_{2.5}$ concentration both as MISO and MIMO system, respectively. Other than that, this chapter also provides crucial references and information required for the prediction of PM_{10} and $PM_{2.5}$ in ensuring the modelling can be done appropriately based on the proposed methodology.

2.2 Air Pollutant

Air pollution continuously arises as one of the tops concerning issues surrounding the world. By definition of Chen (2007), air pollution is a combination of substances that includes various types of particulates as well as other pollutants which behaves as a heterogenous mixture (Chen et al., 2007). Sources of air pollutants emission can be identified as a primary emission in which the pollutant is directly released from its source or as a secondary emission in which the pollutant is formed due to a chemical reaction between substances or pollutants. Friedlander (1975), also stated

how sources of air pollution can be classified either as a natural cause or a man-made phenomenon (Friedlander, 1973). A man-made phenomenon is a result of the release of characteristics of chemical elements with their own respective fixed proportions where it can be identified as primary or secondary emission. A study conducted by Lelieveld (2015), denoted how most man-made sources of air pollution is coming from residential and commercial energy used which made up almost 1/3 more of other possible sources. This is followed by agriculture, power generation, industry, biomass burning and land traffic (Lelieveld et al., 2015). Meanwhile, a natural cause of air pollution is contributed by an environmental disasters such as volcanic eruptions, windblown dust, and sea-salt spray.

There are numerous air pollutants being reported especially in the urban area and cities where the most industry would be located along with busy traffics and other potential sources. A slightly different approach as previously mentioned air pollutant characterization is done according to the literature by Kampa (2008). In his study, Kampa denoted that these air pollutants possessed its own unique and distinct chemical composition, reaction properties, emission, persistence in the environment, ability to be transported in various distances and their eventual impacts on human and/or animal health (Kampa & Castanas, 2008). However, they do exhibit few similarities where typically it would be characterized based on four major groups of air pollutants. The four major groups of air pollutants include gaseous pollutants (SO_2, NOx, CO, ozone, Volatile Organic Compounds), persistent organic pollutants (dioxins), heavy metals (lead, mercury) and lastly particulate matter (PM_{10}, $PM_{2.5}$). From this four major group, six of them is known to be the most common and sufficiently harmful for humans and the environment henceforth causing it to be routinely monitored by the United States

Environmental Protection Agency (US EPA), which also has set the National Ambient Air Quality Standards.

Generally, the six most common air pollutants are comprised of carbon dioxide (CO_2), carbon monoxide (CO), nitrogen oxide (NO_2), nitrogen monoxide (NO), ozone (O_3), and fine particulate matter (PM) known as pollutants. Based on the particle diameter size, PM can be characterized into PM_{10} and $PM_{2.5}$ where PM_{10} and $PM_{2.5}$ are particles with diameters <10 and <2.5 μm, respectively. These air pollutants are associated with many health impacts and deterioration to the environment. This is supported through the evidence brought upon the World Health Organization which had estimated that air pollution particularly containing particulate matter (PM) contributes to approximately up to 800,000 premature deaths annually, resulting in it being ranked 13th in the leading cause of mortality worldwide (Anderson et al., 2012). It was brought upon by Ayala (2011) that there are multiple factors that should be focused on in regard to seeking the detrimental health effects that revolve around the exposure of these air pollutants. Two of those factors that are claimed to be most vital is source intensity and emissions characteristics. These two factors affect human exposure to air pollutants and ultimately increase the possibility of adverse health outcomes (Ayala et al., 2012). Thus, air pollution health effects can be determined through source characterization as it showcased the pollutants that come with it.

Usually, these air pollutants are presents in conjunction with human activities where some may differ from one another, and some may be otherwise similar. For example, NO_2 is formed in the atmosphere as a result of nitrogen oxides (NO) instantly reacting with ozone or radical presence. The main source of human activity that correlate to the generation of NO_2 would be mobile and stationary combustion sources.

Similarly, CO is majorly produced from road transportation as a product of incomplete combustion (Kagawa, 1994; Kampa & Castanas, 2008). On the other hand, ozone in the lower atmospheric layers is formed through a series of reactions between NO_2 and volatile organic compounds with the requirement of sunlight present. Meanwhile, SO_2 emission is a result of the combustion of sulfur-containing fossil fuels (principally coal and heavy oils) and the smelting of sulfur containing ores, volcanoes, and oceans. But it was denoted from Kagawa (1994) that SO_2 emission can be reduced and controlled through the importance of desulfurization of heavy oil, and flue-gas. As for the particulate matter, its major contributor would be from factories, power plants, refuse incinerators, motor vehicles, construction activity, fires, and natural windblown dust.

All respective pollutants exhibit an association in impacting the human's health as well as the environment. There have been significant advances in knowledge regarding the effects that it holds where some pollutants may be more damaging than other. The health effects that were put forward may also vary depending on sex and age. Few studies have reported the harmful effects of ambient air pollution is more likely towards young children (Kim, 2004; Makri & Stilianakis, 2008; Salvi, 2007). Because their defence mechanism is still evolving, children breathe in a higher volume of air per body weight than adults causing them to be more vulnerable. Unfortunately, the effect can potentially be developed prior to the birth of the child itself. This is subjected to occur if pregnant mothers are exposed to high levels of ambient air pollution where it can be demonstrated through the presence of these pollutants in their blood. Through the blood streams, these pollutants can enter the fetal circulation via the placenta and umbilical cord blood thus attacking the fetus (Salvi, 2007). Apart from children, the elderly age group also falls under the vulnerable category where this can be supported through the research done by Hong (2002) indicating that the relative risk of stroke

mortality towards the elderly is higher where linear exposure-response relationships is evident as compared to the younger age group (Hong et al., 2002).

In the research done by Bernstein (2004), children playing outdoor sports are more likely to be associated with the risk of asthma development due to ozone pollutant. Ozone has the ability to increase airway inflammation leading to difficulty in breathing. Inflammation can cause limitation of the airway to provide adequate air flows to the breather. It also can trigger the airway response to inhaled allergens which is bound to attract various health risking diseases. Other than ozone, exposure to nitrogen oxides alco could induce the effects of inhaled allergen response thus contributing to the rise of respiratory infection and wheezing. Similarly, a high concentration of sulfur dioxide is a respiratory irritant, provoking airflow limitations which could also lead to asthma (Bernstein et al., 2004). Although different pollutants can trigger different health effects on humans, a study by Hong (2002) reported that it is hard to separate one pollutant's effect from another's as the different pollutant levels tend to be interrelated. Thus, it is noted that these air pollutants are useful to be utilized along with other meteorological parameters in forecasting the concentration of PM_{10} and $PM_{2.5}$ as demonstrated by previous related work.

As previously mentioned, this research will be focusing on forecasting the concentration of PM_{10} and $PM_{2.5}$. Among the six common pollutants, the particulate matter seems to be the main culprit in most cases of health issues reported since it consists of complex and varying mixtures of particles suspended in the breathing air, which vary in size and composition (Kampa & Castanas, 2008). Due to its small diameter size and great carrying capacity, $PM_{2.5}$ can easily penetrate the human's lung through the respiratory pathway, bringing in other pollutants thus it is known to be an

important agent for increased risks of hospitalizations for Parkinson's disease, carcinogenicity, diabetes and risks of myocardial infarction (MI) and more within susceptible individuals from hours to days of exposure (Gholizadeh et al., 2019). Apart from that, PM_{10} also exhibits similar characteristics where it is often associated with respiratory disorders to premature death (Cortina–Januchs et al., 2015; Terziyska et al., 2019). Meanwhile, a study from few scholars also mentions the effects that air pollutants have towards animal, insects and plants (Alstad et al., 1982; Heath, 1980; Jaffe, 1968).

Thus, from the discussed types, sources, and effects of air pollutants, it is clearly important to ensure these pollutants does not exceed the standard allowable emission. With the current technology and knowledge, arise various tools and techniques available to forecast its concentration in ensuring better management of air quality is manifested to preserve the environment as well as improve our everyday lives. This principle should be applied worldwide to reduce and control air pollution from early detection of air pollutant concentration.

2.3 Model Prediction for Air Pollutant

Through the urgency of overcoming health and environmental issues evolved around air pollution, arise a variety of ways that can be utilized as a means of predicting the concentration of air pollutants. The deterioration of air quality pushes academic scholars and researchers to experiment with the utilization of model prediction for a better predicament of PM_{10} and $PM_{2.5}$ concentration. Constantly monitoring the level of air pollutants in the air is crucial in controlling the danger that it brought upon. This means of action is vital especially for an area that is highly populated and continuously developed. This area typically exhibits the major source of emissions for air pollutants

through various factors such as meteorological, traffic, burning of fossil fuels, and industrial sector which all plays a huge role in the occurring of air pollution. It was noted from many studies that although the behaviour of PM_{10} and $PM_{2.5}$ is uncertain and may be flamboyant at times, its concentration is still predictable. Decades of research and developing technology from various studies using the method of model prediction is proven useful for the measuring of air quality sustainability. The effort of forecasting the concentrations of pollutant strike to be helpful especially in creating awareness for the public thus preventing any disaster that bound to happen due to it.

The model prediction is a type of decision-making technique applying a mathematical process to forecast future events in regard to historical information available. According to Al-Janabi (2020), there are three categories of prediction techniques which are viz. traditional that mostly offers accuracy, a self which focuses on speed, and intelligent which offers both speed and accuracy simultaneously (Al-Janabi et al., 2020). Meanwhile, a study by Feng (2015) explained how these predictive models fundamentally can be grouped based on their approach which are deterministic and statistical approaches (Feng et al., 2015). The two approaches demonstrate different requirements, advantages and disadvantages which usually is compared in this area of study. Numerous predictive models have been proposed under both approaches such as the autoregression model (Aditya et al., 2018), time series model, support vector machine (Delavar et al., 2019; Leong et al., 2020), multiple linear regression mode (Ismail et al., 2018), neural networks (Rahman et al., 2017), and so on (Xu & Ren, 2019). These predictive methods can be utilized as a single approach, or some are combined to form a hybrid model in order to obtain better performance and accuracy. Few predictions method was reviewed from related studies to analyse other alternatives prior to model development.

In the study conducted by Ismail (2018), a statistical approach is used where two models consisting of multiple linear regression (MLR) and principal component regression (PCR) are implemented to predict the concentration level of PM_{10}. The research takes into account seven meteorological parameters based on 8 years of data retrieved from 2007 until 2014 in four major industrial areas of Seberang Prai, Pasir Gudang, Kemaman and Nilai located in Peninsular Malaysia. Both MLR and PCR models were utilized, and their performance were compared. Although PCR is a hybrid model resulted from the combination of MLR and Principal Component Analysis (PCA), the final result obtained showed that MLR performed better than PCR through calculating the error and measuring accuracy using root mean square error (RMSE), normalized absolute error (NAE) and correlation coefficient (R^2). Through the study, main sources of air pollutant emissions were able to be determined where the most pollutant is emitted from road traffic and industrial emissions. This is valid as four of the locations is near an industrial area. Meanwhile, the research also mentioned that the wind speed parameter showcased an inversely proportional relationship with the PM_{10} concentrations along with additional information that current weather conditions at that time may contribute to the pollutant concentration level as Malaysia had experienced transboundary haze pollution during the southwest monsoon from June to October in the year of 2010-2014 (Ismail et al., 2018).

On the other hand, Weizhen (2019) had presented their study in predicting the concentration of PM_{10} and $PM_{2.5}$ using support vector machines (SVM). SVM is a statistical learning technique operated based on machine learning and the generalization of theories. The study uses the concept of the hybrid model by implementing successive over relaxation support vector regress (SOR-SVR) model. Using the daily average aerosol optical depth (AOD) and meteorological parameters measured in Beijing from

2010-2012 as the historical information, the data is processed using Gaussian kernel function, k-fold crosses validation and grid search method prior to the development of the SVR model in order to obtain the ideal parameters to increase the chances of obtaining better generalization capability. The final result obtained demonstrated that although the result retrieved is similar to the actual $PM_{2.5}$ concentration, the generalization ability of this model is poor as this may be due to the over-fitting. However, due to the utilization of the k-fold cross validation and grid search, this is prevented and had improved the overall generalization ability that the model holds (Weizhen et al., 2014).

A comparison of model prediction tools is done in the research study by Koo (2019) to investigate the most effective model to forecast API values. The study utilizes a variety of models in the forecasting of air pollution index based on Kuala Lumpur, Malaysia, as their case study, specifically in the year 2017. Kuala Lumpur is specially selected as it is the largest city in the country and highly populated. The comparative study is done on different models which include artificial neural network (ANN), autoregressive integrated moving average (ARIMA), trigonometric regressors, Box-Cox transformation, ARMA errors, trend, and seasonality (TBATS) and several fuzzy time series (FTS) models. Prior to modelling, Koo denoted how different model requires different steps of the procedure to achieve the prediction of air pollutant. Some models possessed a simpler procedure for example ARIMA requires three major steps which are parameter estimation, tentative identification, and diagnostic checking, while some may require more complex steps such as the ANN model that called for data training and testing along with recognizing the nonlinear pattern between the input and output layer. Overall, the study concluded that the FTS methods exhibit lower RMSE and MAPE values on average as compared with the other models indicating it is the most

efficient model that was studied. Since the API readings are uncertain and may change depending on several factors, FTS is suitable and efficient in accurately measure the fluctuations and seasonality of the readings (Koo et al., 2020).

Similar to previous studies, research by Aditya (2018) also uses model prediction to forecast the future values of $PM_{2.5}$ based on its previous readings. The modelling is done through the employment of Autoregression. An autoregressive (AR) model utilizes the historical data set from the previous time as its input to predict the future value at the next time step. The model is used as a forecasting method when the correlation between values in a time series and the values that exist before and after the time series is present. Prior to AR, Logistic regression is also implemented in determining whether the air is polluted or not. Since the output is binary (only two possible results), logistic regression is viewed to be the appropriate regression model in conducting the analysis. The experiment is done where the autoregression model is applied on the data set to predict the concentration of $PM_{2.5}$ 7 days prior to the date of the actual observations. The results indicated that the model provides a prediction value with acceptable errors of 27.00 in the calculation through MSE. However, the research had mentioned that the error may be reduced by lowering the number of days required from the $PM_{2.5}$ prediction date and the actual observation date (Aditya et al., 2018).

Furthermore, in the study done by Zeng (2020) to forecast the concentration of $PM_{2.5}$ in Chengdu, China, a generalized additive model, (GAMs) was developed. The study specifically opted for the model as it is suggested to help understand the relationships between the pollutant concentrations and meteorological factors in the duration of five years from January 2013 to December 2017. Zeng stresses the importance of understanding the mechanisms of air pollution formation through the

relationship prior to providing recommendations in battling the air pollution. The study had taken into account the use of statistical models and chemical transport model (CTM) before decided to develop GAM. In this area of study, the statistical model is typically preferred compared to CTMs due to its simplicity and ability to fit the patterns derived from the data observed in a straight manner. Meanwhile, CTMs are often required more time and complex along with needing large scale of input data such as emissions inventories, topography, land use and cover, and meteorological simulations (Zeng et al., 2020) which most of the time may not be available or tedious to be obtained. The study had developed two GAMs models where one included sea-level pressure (SLP) as one of the variables and the other one included ΔSLP for 5 days in advance. The results indicated better performance from the use of ΔSLP. The measuring of ΔSLP showed the relationship between $PM_{2.5}$ and ΔSLP where the dispersion of pollutants decreases with the increase of sea-level pressure due to the decrement of average surface wind speed (Zeng et al., 2020).

2.4 Artificial Neural Network (ANN) Model Prediction for Air Pollutant

In this research project, Artificial Neural Network (ANN) is chosen as the main model prediction used to forecast the concentration of PM_{10} and $PM_{2.5}$ using machine learning approach for MISO and MIMO system respectively. This section will focus on the usage of ANN from previous studies by researchers in this area where it will be reviewed and discussed with respect to each case, and the differences that they may portray. ANN is one of the most widely used tools of a forecasting model that has been applied worldwide in various research fields. It was first invented by McCulloch and Pitts between the year of 1943 to 1947 (Dedovic et al., 2016). The working principle of

ANN is a direct imitation of the human's biological neuron where its basic approach can be presented in three steps which are (Dedovic et al., 2016):

Step 1: An artificial neuron will process and produce an output based on the raw data or the output from another neuron that it received.

Step 2: Network describes input layer, an output layer and hidden layers. The network topology may be different depending on the problem to be solved, the type of input layer, output layers and other factors.

Step 3: Artificial neuron model:

- Inputs – Numerical data which are scaled according to the connection weight.
- Outputs – May be more than one as it is dependent on the problem to be solved.
- Weight – Present the ability of the inputs to stimulate neuron.
- Weighted summation – Weighted inputs are summed.

A study conducted by Afzali et al. (2014) has been a good reference on the issue of forecasting with ANN. The study uses ANN to predict PM_{10} concentration in Pasir Gudang, Johor industrial area. The data sample collected from 2008 to 2010 were used in which the relationship between PM_{10} concentrations and meteorological parameters such as wind speed, relative humidity, solar radiation, and the temperature has been analysed statistically. Two techniques of the neural network were applied and compared which are the Feed Forward Network (FNN) and the Elman Network. It was found that although both demonstrated low percentage error through model validation, Elman Network shines through with a more precise result. Both models showed a valid value

between the predicted and daily observed PM_{10} concentration with R=0.69 and R=0.70 for FNN and Elman networks, respectively.

A slightly different approach can be reviewed from the study conducted by Chaloulakou et al (2003). In this study, a comparison between two methods of statistical model which were the multiple regression model and ANN model were done. Sample data were collected in Athens, Greece, ranging over a two-year period. Prior to the development of the forecasting model, selection of the predictor variables was made via a stepwise regression analysis procedure in which the meteorological variables were used as inputs. It was found that the neural network approach provides a better predicament of PM_{10} over the regression models where the root means square error values obtained is lower by 8.2–9.4% as compared to the regression model. However, it was noted in the study that the neural networks model does not rely much on persistence information as compared to the regression model. As a result, incorporation of a lagged concentration term gives a higher improvement on the multiple regression models' predictive power.

On the other hand, research from Cortina–Januchs et al (2015) proposed the utilization of Multilayer Perceptron Neural Network which is a type of ANN with the application of clustering algorithm to predict the average concentration of PM_{10} for the next 24 hours in the city of Salamanca, Mexico. The implementation of clustering algorithms is done to find relationships among pollutant and meteorological variables which lead to the extraction of more additional information. The clustering algorithm used in this paper is the K–means and Fuzzy C–Means (FCM). It was observed that the clustering algorithm helps in finding the correlation between the available variables,

grouping it with similar data characteristics which is something that would not be otherwise obtainable with applying ANN alone. The additional information obtained allows ANN to make better predictions, resulting in a better generalization capacity than without it.

Meanwhile, in the study done by Agarwal et al (2020), the Artificial Neural Networks model was developed to forecast pollutant concentrations of PM_{10}, $PM_{2.5}$, NO_2, and O_3 in 32 different locations of Delhi, India. What makes this study unique is the utilization of Real Time Correction (RTC) in the model. It was found that the application of RTC induced the reduction of errors in the forecasted values as compared to ANN alone. RTC works in a way that provides correction on the forecasted value obtained by dynamically adjusting it based on model performance during the past few days. However, the use of RTC does comes with a small drawback as more time is required for the forecast value to be obtained in a case of change in conditions.

In the effort made by Perez (2018) to predict the concentration of PM_{10} and $PM_{2.5}$ in Santiago de Chile, the study had developed a multilayer neural network model. From the aforementioned observations, the study had denoted the benefits of choosing deterministic models as they have higher accuracy than the statistical model. However, many studies in the present years had showcased that both results still comes with errors while the deterministic method may be lacking in term of requiring details of emissions and topography for the region of interest making the statistical method more preferrable. The research had initially developed one model and end up developing a second model with better modification. One of the interesting pieces of information gathered from this research is the utilization of $PM_{2.5}$ and PM_{10} concentrations at the specific time of 6 PM and 7 PM as input of the model. The reasons for the particular input specification are in

regard to the correlation where a significant increase of particulate matter pollutant can be observed during the onset of a night episode as in the afternoon the winds weaken making this correlation as a good predictor to help to forecast the pollutant's concentration. The study mentioned the lack of wind direction used as input may have improved the model performance indicating the cruciality of exploring the use of additional input variables along with higher efficiency of data training (Perez & Menares, 2018).

Similarly, Perez also had published his paper in a different study which compares the performance of the neural network with a linear model making Coyhaique, a southern Chilean city its case study. The study also denoted how simple a statistical model is where no highly detailed information is required to perform a prediction. A method of stripping scheme is utilized to scout for relevant variables which is crucial in the case of having large scale data set. The result had indicated better performance of neural network model where appropriate selected input variables which include past concentrations of $PM_{2.5}$ and gases combined with additional meteorological parameters can induce the contribution of the formation of the secondary particles. Through the allocations of input variables for the predictive model of $PM_{2.5}$, Perez had come to the conclusion that O_3 has a significant anti- correlation with particle concentration particularly between the time where the sun is visible. The theory that Perez had presented is due to the sun presence, solar radiation increases, resulting to an increment of O_3 formation. This showed the importance of parameters correlation and how this predictor helps in choosing input variables. It was denoted that the best results of the neural network model is achieved by using 7 neurons in one hidden layer (Perez et al., 2020).

2.5 Summary/Findings

In general, through the aforementioned observations it can be summarized that there is an abundance of information available on the method of forecasting PM_{10} and $PM_{2.5}$ concentration from previous studies and literature. As days go, these methods and models increases with many publications emerged implementing the use of a variety of model prediction tools to find the most suitable and efficient. The forecasting procedure may vary for each different model however, its processes is thoroughly implemented starting from the beginning step until the end.

Typically, the case study for this research mainly is based on the urban area where industrial activity and traffic may be heavy as compared to a rural area. Each study highlighted the importance of forecasting air pollutant concentration to prevent many possible diseases which may cause fatality. Every forecasting model was categorized either as statistical model or deterministic model where its approaches may be single or hybrid. Based on analysis and observations, a statistical model typically is preferred as it is simpler and suitable for cases where there is much historical data available to develop a predictive model.

In this case, the ANN model which is an artificial intelligence tools seems to be the most favourable model be it for its single application or for hybrid due to the capacity that it holds to learn from data, works well for a given site, can handle nonlinear and chaotic chemical system at a site. Furthermore, it also requires modest expertise which indicates simpler to be learnt along with moderate to high accuracy results which are computationally fast as well (Zhang et al., 2012).

CHAPTER 3

METHODOLOGY

3.1 Introduction

This chapter specifically disclose the overall modelling aspects of the project. It includes the general flowchart diagram of the research, the procedure of using obtained sample historical dataset to be used as modelling parameters, the application of MATLAB to perform input selection followed by the development of the ANN model and lastly the evaluation of model performance using statistical methods in which the equations used is introduced.

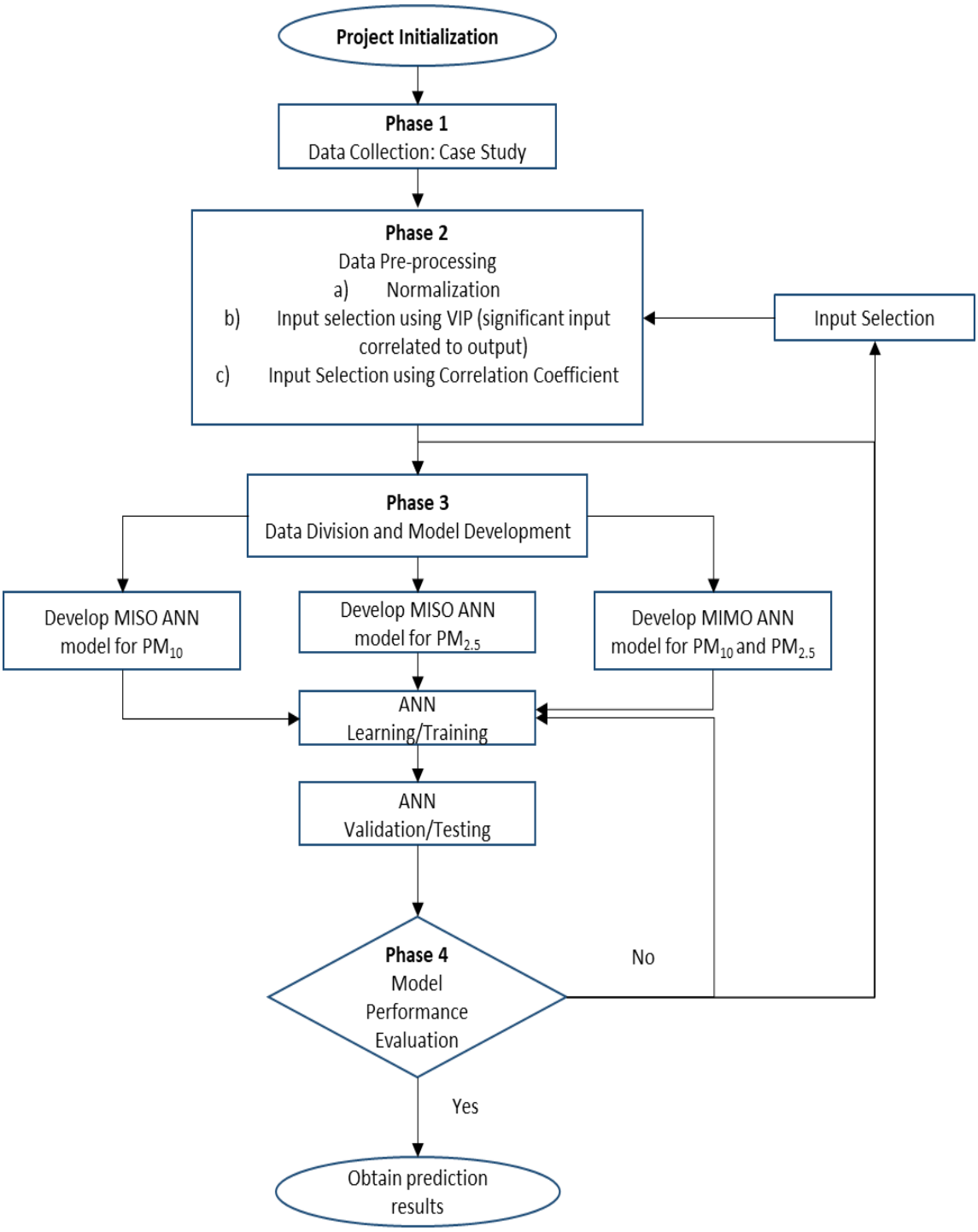

Figure 3.1 Flowchart of the Research Methodology

3.2 Case Study: Beijing Multi-Site Air Quality

Air-quality data set from the Beijing Municipal Environmental Monitoring Centre is retrieved from the official website of UCI Machine Learning Repository (*UCI Machine Learning Repository: Beijing Multi-Site Air-Quality Data Data Set*, n.d.), in reference to the paper written by Shuyi Zhang (Zhang et al., 2017) to be used as the case study, for the purpose of this research. The data includes hourly air pollutants from a total of 12 nationally controlled air-quality monitoring sites. The time period of the data collected is at a large scale ranging from March 1st, 2013, to February 28th, 2017. This hourly data set considers six main air pollutants and six relevant meteorological variables at multiple sites in Beijing. The six main air pollutants are PM_{10}, $PM_{2.5}$, SO_2, NO_2, CO, and O_3 ($\mu g/m^3$), while the six meteorological variables are temperature (°C), pressure (hPa), dew point temperature (°C), precipitation (mm), wind direction and wind speed (m/s). The meteorological data in each air-quality site are matched with the nearest weather station from the China Meteorological Administration. There are missing data observed which is denoted as NA. Table 3.1 showcased the historical data in term of air pollutants and meteorological variables retrieved from the machine learning repository along with its number of samples.

Table 3.1 Air Pollutants and Meteorological Data Variables.

No.	Air Pollutants Variables	Meteorological Variables	No. of Samples
1	$PM_{2.5}$ ($\mu g/m^3$)	Temperature (°C)	35,064
2	PM_{10} ($\mu g/m^3$)	Pressure (hPa)	
3	SO_2 ($\mu g/m^3$)	Dew Point (°C)	
4	CO ($\mu g/m^3$)	Precipitation (mm)	
5	NO_2 ($\mu g/m^3$)	Wind Direction	
6	O_3 ($\mu g/m^3$)	Wind Speed (m/s)	

3.3 Model Development

The methodology of model development is consisting of the following four major phase: phase 1: data collection, phase 2: data pre-processing, phase 3: data division and model development and lastly phase 4: model performance evaluation. Throughout this modelling process, Excel 2002 and Matlab^{TM} 2020 were utilized as computing software.

3.3.1 Data Collection

As previously mentioned, the data is collected based on specific time period at a large scale ranging from March 1st, 2013, to February 28th, 2017. This hourly data set considers six main air pollutants and six relevant meteorological variables at multiple sites in Beijing. The six main air pollutants ($\mu g/m^3$) are PM_{10}, $PM_{2.5}$, SO_2, NO_2, CO, and O_3, while the six meteorological variables are temperature (°C), pressure (hPa), dew point temperature (°C), precipitation (mm), wind direction and wind speed (m/s). Having a large historical data is beneficial in removing "small disjuncts", nuances and other nonlinearities that are otherwise impossible to capture from smaller datasets (Junqué de Fortuny et al., 2013). Through this, preventive measures can be taken, in a necessary situation once the forecasting model obtain an accurate forecast of the pollutant's concentration.

3.3.2 Data Pre-Processing

Since a different combination of input parameters may give a different prediction, the parameters to be made an input in the forecasting model will be selected. This is a means of data pre-processing as all available parameters can be tested, ultimately removing the possibility of having an outlier, encompassing missing data, and normalized it to transform into a form of suitable reliable input data which the model can identify and use for forecasting of PM_{10} and $PM_{2.5}$. This step will scout the