FILLING THE GAPS OF MISSING VALUES IN AIR POLLUTION DATASET BY USING VARIOUS IMPUTATION METHODS

FAHREN FAZZER BIN SUKATIS

SCHOOL OF ENVIRONMENTAL ENGINEERING
UNIVERSITI MALAYSIA PERLIS
2019

FILLING THE GAPS OF MISSING VALUES IN AIR POLLUTION DATASET BY USING VARIOUS IMPUTATION METHODS

by

FAHREN FAZZER BIN SUKATIS

Report submitted in partial fulfillment
of the requirements for the degree
of Bachelor of Engineering

JUNE 2019

ACKNOWLEDGMENT

First of all, praise to Allah for giving me strength, guidance and believing to accomplish this study.

I would like to express my deepest gratitude to my supervisor, Dr. Norazian Binti Mohamed Noor for the guidance, materials, advices and support throughout this work.

A lot of thanks for my friends Muhmmad Faizzul Ahmam and Zarrul Azwan Mohd Rasid for the help and supports during accomplish this study.

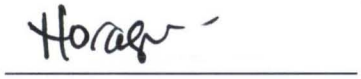
I also wish to express my deepest thanks my beloved mother, Norsiah Hj Tuhalus, my father, Ab Rahman Md Jani, and my sister, Maizatul Akmal Sukatis for their love, courage and patience through the hard time I had during finishing this study.

Last but not least, my sincere thanks to Department of Environment (DOE) Malaysia for the air pollutant data. Without the data, this research may cannot be done.

APPROVAL AND DECLARATION SHEET

This project report titled Filling The Gaps Of Missing Values In Air Pollution Dataset By Using Various Imputation Methods was prepared and submitted by Fahren Fazzer Bin Sukatis (Matrix Number: 151130568) and has been found satisfactory in terms of scope, quality and presentation as partial fulfillment of the requirement for the Bachelor of Engineering (Hons) (Environmental Engineering) in Universiti Malaysia Perlis (UniMAP).

Checked and Approved by



(DR. NORAZIAN BINTI MOHAMED NOOR)
Project Supervisor

School of Environmental Engineering
Universiti Malaysia Perlis

June 2019

MENGISI KEKOSONGAN NILAI YANG HILANG DALAM DATASET PENCEMARAN UDARA DENGAN MENGGUNAKAN PELBAGAI KAEDAH IMPUTASI

ABSTRAK

Hampir kesemua data yang diperoleh daripada stesen pemantauan kualiti udara ambien berterusan (CAAQM) mengandungi data yang hilang. Biasanya, ini berlaku kerana kegagalan mesin, penyelanggaraan, perubahan tempat pemantauan, dan kesilapan manusia. Dataset yang tidak lengkap boleh menyebabkan kecenderungan berat sebelah kerana disebabkan oleh perbezaan sistematik antara data yang diperhatikan dan data yang tidak dapat dilihat, selain mengurangkan kuasa analisa statistik. Dalam kajian ini, sepuluh kaedah imputasi seperti Min Siri, Interpolasi Linear, Min Jiran Terdekat, Pengoptimuman Jangkaan, "Markov Chain Monte Carlo", 24-Jam Purata Bergerak, 12-Jam Purata Bergerak, Dan Pelepasan Eksponen (0.2, 0.5, dan 0.8) digunakan untuk mengisi nilai-nilai yang hilang dalam data pencemaran udara. Data pemantauan setiap jam untuk suhu ambien, kelajuan angin, kelembapan, SO_2, NO_2, O_3, CO, dan PM_{10} daripada Petaling Jaya dan Shah Alam (2012-2016) dihuraikan menggunakan penjelasan statistik. Data pada tahun 2012 dipilih sebagai data rujukan kerana mengandungi data hilang yang sedikit berbanding pada tahun-tahun yang lain. Seterusnya, dataset ini disimulasikan kepada tiga jenis pola data hilang yang bervariasi dari segi tempoh jurang data yang hilang i.e. pola mudah (jurang yang hilang < 24 jam), pola sederhana (jurang yang hilang antara 24 hingga 168 jam), dan pola kompleks (kombinasi pola mudah dan sederhana dengan kadaran 1 kepada 1). Setiap pola disimulasi kepada dua peratusan kehilangan data i.e. 10% dan 20%. Prestasi kaedah imputasi tersebut dinilai dengan menggunakan empat petunjuk prestasi iaitu kesilapan mutlak maksimum, ralat maksimum kudrat kuasa, ketepatan ramalan, dan indeks perjanjian. Kesemua petunjuk prestasi ini dikira untuk menggambarkan kesesuaian kaedah untuk semua kaedah imputasi. Secara keseluruhannya, kaedah Pengoptimuman Jangkaan dipilih sebagai kaedah imputasi terbaik untuk mengisi kehilangan data yang disimulasi kepada pola mudah, sederhana, dan kompleks untuk kesemua parameter kecuali dataset PM_{10} kerana mengandungi outlier yang ekstrim dan nilai kepelbagaian data yang tinggi.

FILLING THE GAPS OF MISSING VALUES IN AIR POLLUTION DATASET BY USING VARIOUS IMPUTATION METHODS

ABSTRACT

Almost all data obtained from continuous ambient air quality monitoring (CAAQM) station contains missing data. Usually, this happens due to machine failure, maintenance, change in the siting monitors and human error. Incomplete dataset can cause a bias due to systematic differences between observed and unobserved data, besides reducing the power of statistical analysis. In this study, ten imputation methods such as Series Mean, Linear Interpolation, Mean Nearest Neighbour, Expectation Maximization, Markov Chain Monte Carlo, 12-hour Moving Average, 24-hour Moving Average, and Exponential Smoothing (0.2, 0.5, and 0.8) were applied to fill in the missing values in air pollution dataset. Annual hourly monitoring data for ambient temperature, wind speed, humidity, SO_2, NO_2, O_3, CO, and PM_{10} from Petaling Jaya and Shah Alam (from year 2012 until 2016) were described using descriptive statistics. Dataset in 2012 were selected as reference data as it has less missing data. Next, these dataset were simulated into three types of missing data pattern which vary in terms of length of gaps in missing data i.e. simple pattern (missing gaps less than 24 hours), medium pattern (missing gaps in between 24 to 168 hours), and complex pattern (combination of simple and medium patterns in proportion of 1 to 1). Each pattern was simulated into two percentages of missing data i.e. 10% and 20%. The performances of these imputation methods were evaluated by using four performance indicators namely Mean Absolute Error, Root Mean Squared Error, Prediction Accuracy, and Index of Agreement. All performance indicators were calculated to describe the goodness of fit for all imputation methods. Overall, Expectation Maximization method has been selected as the best imputation method to fill in simple, medium and complex patterns of simulated missing data for all parameters except PM_{10} as the dataset contains extreme outliers and high variability of data.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBRIVIATION

AT	Ambient Temperature
H	Humidity
WS	Wind Speed
UV	Ultra Violet
SO_2	Sulphur Dioxide
NO_2	Nitrogen Dioxide
O_3	Ozone
CO	Carbon Monoxide
PM_{10}	Particulate Matter (10μm in diameter)
PM	Particulate Matter
$PM_{2.5}$	Particulate Matter (2.5μm in diameter)
VOC	Volatile Organic Compound
BC	Black Carbon
TSP	Total Suspended Solid
PI	Performance Indicator
MAE	Mean Absolute Error
RMSE	Root Mean Squared Error
PA	Prediction Accuracy
d_2	Index of Agreement
SM	Series Mean
LI	Linear Interpolation

MNN	Mean Nearest Neighbour
EM	Expectation Maximization
MCMC	Markov Chain Monte Carlo
MA	Moving Average
ES	Exponential Smoothing
MAR	Missing at Random
MCAR	Missing Completely at Random
MNAR	Missing Not at Random
DOE	Department of Environment
ASMA	Alam Sekitar Malaysia
CAAQM	Continuous Ambient Air Quality Monitoring
API	Air Pollution Index
RMAAQG	Recommended Malaysia Ambient Air Quality Guidelines
IT	Interim Target
SPSS	Statistical Package for Social Sciences
PJ	Petaling Jaya
SA	Shah Alam

CHAPTER 1

INTRODUCTION

1.1 Air Pollution

Air pollution can be defined as the presence of unwanted material in atmosphere in amounts sufficiently substantial to cause harmful impacts to human health, vegetation, property, and climate (Nevers, 2017). These unwanted material or air pollutant can be grouped into four categories: 1) gaseous (e.g. SO_2, NO_2, CO, VOC, and O_3); 2) persistent organic pollutant (e.g. dioxin); 3) heavy metal (e.g. lead, mercury); and 4) particulate matter in which the size of the particles varies in diameter (e.g. $PM_{2.5}$, PM_{10}) (Kampa & Castanas, 2008).

Rapid development and urbanization has been identified as one of the world's major contributors to air pollution, including factories, power plants, and vehicles in Malaysia. According to Abdullah (2012), there are three major sources of air pollution in Malaysia which is about 70% from mobile sources, 20% from stationary sources, and 10% from open burning sources. However, meteorological factors such temperature, wind profile, atmospheric pressure, and others also contribute to the air pollutant transformation into secondary pollutants like smog, acid rain, and many more (Abdullah, 2012).

Health effects are the major negative effects of air pollution, as it has been reported that most people with asthma will experience symptoms like irritation of the nose and throat after getting exposed long-term to low air pollution concentration (Kampa & Castanas, 2008). Epidemiological confirmation suggests that exposure concentration and duration of exposure are factors that contribute to adverse health effects, and long-term exposures to air pollutants have more significant effects compared to short-term exposures (Khallaf, 2011).

1.2 Air Quality Monitoring

The Department of Environment (DOE) is responsible for monitoring the air quality status in Malaysia, and Alam Sekitar Sdn. Bhd. (ASMA) has privatized this operation. The Air Pollution Index (API) is used to measure air quality based on the concentration of five major such as ground level ozone (O_3), carbon monoxide (CO), nitrogen dioxide (NO_2), sulphur dioxide (SO_2) and particulate matter (PM_{10}). There are five categories of air quality that are good, moderate, unhealthy, very unhealthy and hazardous (Department of Environment, 2016). Table 1.1 shows the air quality status based on API scale in Malaysia.

Table 1.1: Air quality status based on API scale in Malaysia (Department of Environment, 2016)

API	Air Quality
0 – 50	Good
51 – 100	Moderate
101 – 200	Unhealthy
201 – 300	Very Unhealthy
> 300	Hazardous

Department of Environment has established the new Malaysia Ambient Air Quality Standard (MAAQS) which shows the maximum requirement for air pollutant concentration that should be present in atmosphere. Previously, in the Recommended Malaysia Ambient Air Quality Guideline (RMAAQG) used from 1989 to 2014, only 5 existing air pollutants concentrations of such particulate matter that were measured in the diameter of less than 10 μm (PM_{10}), sulphur dioxide (SO_2), nitrogen dioxide (NO_2), ozone (O_3), and carbon monoxide (CO). However in the new MAAQS, an additional parameter added is particulate matter with a diameter of less than 2.5 μm ($PM_{2.5}$). Table 1.2 shows the New Malaysia Ambient Air Quality Standard. This new standard will be strengthened in three stages until 2020 (Department of Environment, 2017). There are 3 interim targets in the New Malaysia Ambient Air Quality Standard, which include interim target 1 (IT-1) in 2015, interim target 2 (IT-2) in 2018 and full implementation of the standard (IT-3) in 2020 (Department of Environment, 2017).

Table 1.2: New Malaysia Ambient Air Quality Standard (Department of Environment, 2017)

Pollutant	Averaging time	Ambient Air Quality Standard		
		IT – 1 (2015) $\mu g/m^3$	IT – 2 (2018) $\mu g/m^3$	IT – 3 (2020) $\mu g/m^3$
Ozone (O_3)	1 Hour	200	200	180
	8 Hour	120	120	100
Carbon Monoxide (CO)	1 Hour	35 mg/m^3	35 mg/m^3	30 mg/m^3
	8 Hour	10 mg/m^3	10 mg/m^3	10 mg/m^3
Nitrogen Dioxide (NO_2)	1 Hour	320	300	280
	24 Hour	75	75	70
Sulphur Dioxide (SO_2)	1 Hour	350	300	250
	24 Hour	105	90	80
Particulate Matter (PM_{10})	1 Year	50	45	40
	24 Hour	150	120	100
Particulate Matter ($PM_{2.5}$)	1 Year	35	25	15
	24 Hour	75	50	35

Table 1.3 shows the location of CAAQM location in Malaysia. The air pollution concentration is measured by 52 continuous ambient air quality monitoring (CAAQM) stations throughout Malaysia (Department of Environment, 2016). From 52 stations, 13 stations are in industrial area, 15 stations are in urban area, 20 stations are in rural area, and 4 stations for background (Department of Environment, 2016).

Table 1.3: The location of CAAQM location in Malaysia (Department of Environment, 2016)

State	Location	Type
Johor	Larkin and Pasir Gudang Kota Tinggi Muar	Industrial Urban Sub Urban
Kedah	Alor Setar Langkawi and Sungai Petani	Urban Sub Urban
Kelantan	Tanah Merah Kota Bharu	Industrial Urban
Melaka	Bukit Rambai Bandaraya Melaka	Industrial Urban
Negeri Sembilan	Nilai Seremban and Port Dickson	Industrial Urban
Pahang	Balok Baru Kuantan Jerantut	Industrial Sub Urban Background
Perak	Taiping and Tasek Pagoh Seri Manjung	Industrial Urban Sub Urban
Perlis	Kangar	Sub Urban
Pulau Pinang	Perai USM and Seberang Jaya	Industrial Sub Urban
Sabah	Kota Kinabalu and Tawau Sandakan and Keningau	Urban Sub Urban
Sarawak	Kuching Sri Aman, Sarikei, Sibu, Bintulu, Miri, and Limbang Samarahan, Kapit, and ILP Miri	Industrial Sub Urban Background
Selangor	Petaling Jaya Shah Alam and Pelabuhan Klang Banting, Kuala Selangor, and Tanjung Malim	Industrial Urban Sub Urban
Terengganu	Paka and Kemaman Kuala Terengganu	Industrial Urban
Kuala Lumpur	Batu and Cheras	Urban
Putrajaya	Putrajaya	Urban
Labuan	Labuan	Sub Urban

1.3 Problem Statement

Air pollution is a major social concern that needs to be adequately monitored. To solve the problem, an accurate dataset of air pollution concentration is required, this data

provides the idea of recognizing the problem and its sources, and therefore the control strategies can be prepared and formulated. Therefore, in order to provide these data, nations around the world must establish air quality monitoring systems (Wang et al., 2018). The purpose of air quality monitoring is to measure the ambient air quality, if there is a significant change in the air quality level, it should be told to the public to be prepared for the situation.

However, air pollution data collected by air quality monitoring stations are mostly incomplete due to several factors. According to Gómez-Carracedo et al. (2014), air pollution data can be missing because of too many uncontrollable conditions such as malfunctioning of the instruments, maintenance, and calibration. Missing data can be problematic on performing several mathematical analysis such as time series analysis, principal component analysis (PCA) and multivariate analysis, due to the complete and continuous data are needed (Zainuri et al., 2015).

There are three major issues that may emerge when managing incomplete data. Firstly, there is lost data and, as an outcome, loss of proficiency. Secondly, there are a few complication related with data handling, computation and analysis, because of the anomalies in information structure and the inconceivability of utilizing standard software. Thirdly, the outcomes might be bias because of deliberate contrasts among observed and unobserved data (Noor et al., 2015).

In environmental studies, missing data is a problem repeatedly encountered by researchers (Junninen et al., 2004). Discontinuities of data pose a significant obstacle for time-series forecast schemes, which for most of the parts require continuous information as a condition for their application. However, there are many imputation methods can be used to fill in the missing data in air pollution dataset (Junninen et al., 2004). According to Moshenberg et al. (2015), the length of the missing gaps and the type of study conducted must be considered in determining the best method of imputation. Hence, this study focuses on applying various imputation methods and selecting the best methods for several complexity gaps of the simulated missing observation in air quality dataset.

1.4 Objectives of Study

There are three objectives that will be studied to achieve thoroughly the scope of this research. The objectives are:

i. To study the characteristics of missing data in air pollution dataset.
ii. To simulate the incomplete dataset according to selected percentage of missing data.
iii. To use various imputation methods for filling the simulated missing gap and determine the best imputation methods by using performance indicators.

1.5 Scope of Study

In this study, the data set for hourly air pollution from 2012 to 2016 was used in Petaling Jaya and Shah Alam. The data was obtain from Department of Environment Malaysia. These locations were selected because Petaling Jaya is an industrial area, meanwhile Shah Alam is an urban area in Klang Valley, Selangor. Hence it is expected that different characteristics of air quality can be observed in the both locations. There are five air pollution data which are carbon monoxide (CO), nitrogen dioxide (NO_2), sulphur dioxide (SO_2), particulate matter (PM_{10}), ozone (O_3) and three meteorological data such as ambient temperature (AT), wind speed (WS), and humidity (H) for each dataset. In order to carry out the analysis, SPSS 25 for Windows was used.

Descriptive analysis were done on the dataset to study the pattern of missing data and to determine reference data. A reference data should be a dataset that has the most complete data (Noor et al., 2015). Then, the presence of missing values in the reference data were firstly treated by using known imputation method such as Nearest Neighbour in order to have a complete set of air pollution data. In this study, the reference data would be used in simulation process and performance measure. The design of simulation missing data was conducted based on three levels of missing gaps such as simple, medium, and complex for each pollutant under two different missing percentages of 10% and 20%.

The simulated missing data were later filled by various methods of imputation. There were several imputation methods used in this study namely Series Mean (SM), Expectation Maximization (EM), Mean Nearest Neighbour (MNN), Markov Chain Monte Carlo (MCMC), Linear Interpolation (LI), Moving Average (MA), and Exponential Smoothing (ES). Finally, four performance indicators such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Prediction Accuracy (PA), and Index of Agreement (d_2) were calculated to determine the most reliable imputation method. Reliable imputation method for estimating missing values is important to provide better data quality for advance analysis such as statistical performance, modelling, and forecasting analysis.

1.6 Thesis Layout

Chapter 1 described Malaysia's general idea of air pollution and air quality monitoring. In addition, this chapter also explained the problem statements, objectives, scope of study, and the significance of the study.

Chapter 2 presented the air pollution sources and their effects. The missing data mechanism and common methods for handling the missing data were also explained in this chapter.

Chapter 3 outlined the step-by-step methods used to conduct the study area, simulating missing data, imputation missing data, and performance indicators.

Chapter 4 discusses the outcome of this study. First, data characteristics of raw data and simulated missing data were explained respectively. Then the performance indicator for each method of imputation was discussed and the best method of imputation was selected.

Chapter 5 concludes this study's finding according to the set of objectives. Several suggestions for the upcoming study related to this study were given in this chapter.

CHAPTER 2

LITERATURE REVIEW

2.1 Air Pollution Sources

In general, air pollution sources in Malaysia come from three different types of sources, which are stationary sources, mobile sources and open burning (Halim et al., 2018). According to Department of Environment (2012), Industrial sectors such as power station were one of the main stationary sources of air pollution. Dominick et al. (2012), stated that stationary sources in Malaysia may include energy power plants, incinerators and industries, as well as from urban construction, and quarries. According to Azid et al. (2014), industrial emissions from fuel burning, power stations, production processes, domestic and commercial incinerators and open burning at solid waste disposal sites contributed approximately 0.8% to 9% of the total air pollutant emission load. In Malaysia, the urban area of Klang Valley has undergone extensive physical development of urbanization, infrastructure and industrialization, which has significantly deteriorated air quality (Rahman et al., 2015).

Malaysia economy which is growing speedily at the present time, as well as the private motor vehicle number rising in an accelerating way, and can contribute to the air pollution largely in the form of hydrocarbon, lead, and nitrogen oxides (Almselati et al., 2011). The motor vehicles are the main source of mobile source of air pollution emission in Malaysia which are contributed to 82% of the air emission load (Azid et al., 2014). Furthermore, the increasing number of transportation also raise the heavy traffic flow in the urban area which later would contribute of high concentrations of PM_{10}, NO_2, SO_2, and CO in the urban area such Klang Valley (Azmi et al., 2009).

Open burning sources in Malaysia mainly come from Indonesia which causes the transboundary pollution (haze) due to the fires produced by the combustion of peat and

deteriorates the local air quality (Mahmud, 2013). According to Othman et al. (2014), Malaysia was affected with substantial quantities of particle matter, severe smoke haze episodes occurred in April 1983, August 1990, June 1991, October 1991, and August 1994. However, the most horrible episode of haze occurred in 1997 which almost the whole country was surrounded by thick smog-like particles for almost six months. Haze can be regarded as smog-like small suspended solid or liquid particles (Cheng et al., 2013).

2.2 Air Pollutants and Their Effects

Figure 2.1 shows the trend of the annual average levels of PM_{10} concentration in the ambient air quality for year 2000 until 2016 complied with the Recommended Malaysian Ambient Air Quality Guidelines (RMAAQG). The annual average concentration of PM_{10} concentration was below than 50 $\mu g/m^3$ except during year 2015. This is because of the local and transboundary haze pollution with help of tropical cyclone from Sumatra and Kalimantan, Indonesia increase the PM_{10} concentration records in most of the areas in Malaysia (Department of Environment, 2015). Haze in 2015 was reported as one of the worst episodes since 1997 that lasted for more than two months and affected the whole country (Department of Environment, 2015).

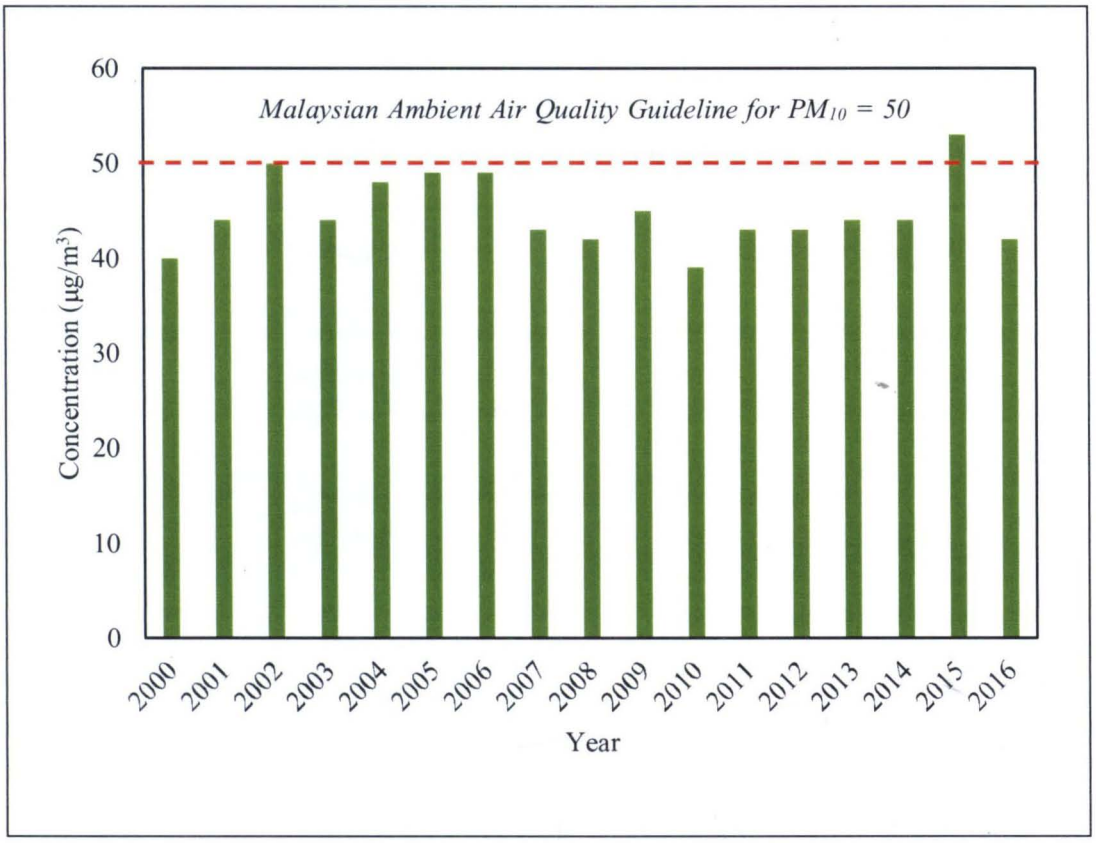


Figure 2.1: The trend of the annual average levels of PM_{10} concentration in the ambient air for 2000 until 2016 complied with the Malaysian Ambient Air Quality Guidelines (Department of Environment, 2016).

Haze smoke mainly contains high concentration of PM_{10} and can affect the human health (Norela et al., 2013). Based on the study conducted by Zheng et al. (2015), it was found that during haze event, the levels of TSP, $PM_{2.5}$, BC, SO_2, and NO_2, are higher while the levels of ozone was lower compare to the non-haze days. There are five major of air pollutants measured by Department of Environment which can harms to human health, environment, and property including sulphur dioxide (SO_2), nitrogen dioxide (NO_2), carbon monoxide (CO), particulate matter (PM_{10}), and ground-level ozone (O_3) (Mabahwi et al., 2015). Table 2.1 shows the five major air pollutants with their description of sources and health effects.

Table 2.1: The five major air pollutants with their sources and effects (World Health Organization, 2019)

Air pollutant	Sources	Health effect
Particulate Matter	Combustion engines, solid-fuel, and combustion for energy production	Risks to health, capable of penetrating peoples' lungs and entering their bloodstream
Ground Level Ozone	Motor vehicle exhaust, industrial facilities, and chemical solvents	Health risk linked to breathing problems, asthma, reduced lung function and respiratory diseases
Nitrogen Dioxide	Power generation, industrial and traffic sources	Increase symptoms of bronchitis and asthma, respiratory infections and reduced lung function and growth
Sulphur Dioxide	Burning of fossil fuels and the smelting of mineral ores that contain sulphur	Affects the respiratory system and the function of the lungs, and causes irritation of the eyes
Carbon Monoxide	Motor vehicle exhaust and machinery that burn fossil fuels	Ruining the amount of oxygen transported in the bloodstream to critical organs

2.2.1 Sulphur Dioxide

Sulphur dioxide (SO_2) is a highly toxic, colourless, non-flammable gas, readily dissolves with water, and has a pungent odour, which is also an environmental pollutant (Pubchem, 2019). Sulphur dioxide can mostly be found in an enormous amount of small solid particles and solute flue gases formed resulting of extended combustion of fossil fuels and human activities (Wang et al., 2018). According to Brown et al. (2017), the power sector contribute to about 64% of the SO_2 emission which is detrimental to human health, environments, harvest, timber production, and the building. However, over the last two decades, SO_2 emissions from the electric power industry has declined swiftly, while electricity generation was increased (Brown et al., 2017). The decreasing of SO_2 emission was due to the use of healthier fuel of EURO-2M in Malaysia starting

from September 2009 in compliance to the stricter enforcement by the DOE (Department of Environment, 2012).

2.2.2 Nitrogen Dioxide

Nitrogen dioxide (NO_2) is a very poisonous gas, reddish brown in colour, heavier than air vapour, non-combustible but can accelerate the burning of combustible materials, and NO_2 was one of the major air pollutants which is able to absorb UV light which cannot reach to the earth's surface (Pubchem, 2019). According to the study conducted by Casquero-Vera et al. (2019), because of the emissions of nitrogen oxide (NO) from road transport, the emission of NO does not decreased sufficiently as well as NO_2 ambient concentration does not reduce as expected, and in worst case NO_2 concentration rose up. According to Department of Environment (2012), high number of motor vehicles and combustion processes in Malaysia were the factors of insignificant changes in NO_2 concentration in recent years. Strangely, in Malaysia, most of the sources of NO_2 was quite similar with the SO_2 which is majorly produced from motor vehicle followed by power station, and industrial (Department of Environment, 2012).

2.2.3 Carbon Monoxide

Carbon monoxide (CO) was described as poisonous gas, colourless, odourless, tasteless, and formed from incomplete combustion of carbon (Pubchem, 2019). According to Colvile et al. (2002), in a complete combustion of fuel, carbon dioxide (CO_2) together with water vapour (H_2O) are the most major transport gas emissions to the atmosphere. However, transport power sources only achieve just about complete combustion (Colvile et al., 2002). Carbon monoxide are hazardous to human health because CO molecules can combine with haemoglobin to form carboxyhaemoglobin (no oxygen carrying capacity) during inhalation process, and resulting of oxygen deficiency which can cause a headache, dizziness, decreased respiratory rates, decreased pulse, unconsciousness, and death (National Centre for Biotechnology Information, 2019).

2.2.4 Ozone

Ozone (O_3) was described as the unstable triatomic form of oxygen which is very oxidant, colourless to bluish gas (Pubchem, 2019). Ozone exist about 90% in the atmosphere of stratosphere which also known as stratospheric ozone (Pubchem, 2019). Based on Kanagendran et al. (2018), and Sicard et al. (2017), O_3 in tropospheric ozone (below than stratospheric ozone) is a key pollutant which can cause plant stress, because O_3 could cause biochemical alterations and physiological damage in plants. Besides that, the ground level O_3 (tropospheric ozone) can develop asthma in children and potentially worsen people with existing asthma (Mabahwi et al., 2015). Ozone molecule normally formed in the troposphere with the presence of sunlight while NO reacted with reactive volatile organic compounds (VOC) (Reyrson et al., (2003). According to Jamal et al. (2004), the source of O_3 is mainly generated from industrial, gasoline vapour, motor vehicle, and chemical solvent processes.

2.2.5 Particulate Matter

According to Poeschl (2006), particulate matter (PM) is the standard word used for a type of air impurities which can be made up of complex and varying mixtures of particles suspended in the air which differ in size and composition. The size of PM varies and usually measured according to their aerodynamic diameter, commonly there are two type size of PM such as particulate matter with diameter less than 10 μm (PM_{10}) and less than 2.5 μm ($PM_{2.5}$) (Kampa & Castanas, 2008). The main sources of PM can be divided into two sources; first, the anthropogenic source including the combustion of fossil fuels, exhaust discharge from vehicles, industrial activities, energy production, construction activities, and waste incineration (Mohammed et al., 2017). The second main source of PM are from the natural source which including the volcanic activity, wind eroded soil dusts, forest fires, and sea salt spray (Mohammed et al., 2017). Mohammed et al. (2017), also stated that PM is considered as the most harmful and hazardous pollutant to human health even at the low concentration of exposure, and it was reported that there are link between PM and a number health problems which includes asthma, bronchitis, acute and chronic respiratory symptoms.

2.3 Missing Data in Air Pollution Dataset

Continuous Air Quality Monitoring Station (CAAQM) requires a frequent maintenance to ensure that the air pollution data obtained from this station is accurate. However, maintenance process will cause air pollution data from the station to be incomplete (Moshenberg et al., 2015). According to Noor et al. (2015), missing data is a very common problem that happen in numerous scientific fields including in environmental engineering fields and it can cause prejudice and bias due to systematic dissimilarities between observed and unobserved data. According to Gómez-Carracedo et al. (2014), missing data problems arise because of too many uncontrolled circumstances such as failing of the instruments, maintenance, calibration, and human error.

Discontinuities of data pose a significant difficulty for time-series prediction, because such analysis require an uninterrupted data as a condition for their use (Junninen et al., 2004). Missing data also obstruct the ability to make accurate conclusion or interpretations about the observation (Noor eat al., 2015). Therefore, the missing data needs to be treated, because complete data are required to carry out statistical analysis, for example in time series analysis, principal component analysis (PCA) and multivariate analysis, this analysis needs continuous data in order to perform estimation (Zainuri et al., 2015).

2.4 Mechanisms of Missing Data

There are three kinds of missing data mechanisms which are missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) (Liu and Brown, 2013). Classifying the types of missing data are important in order to determine suitable and reliable imputation methods to fill in the missing observation. This is because the quality of prediction by using specific imputation method are affected by the types of missing data (Garciarena & Santana, 2017). Furthermore, the imputation performance depends on the amount of missing data and the characteristics of missing data patterns (Junninen et al., 2004).

Missing completely at random (MCAR) occur when the distribution of missing data for a variable does not influenced by observed or any missing data (Liu & Brown, 2013). MCAR usually occur for unknown reasons (Gómez-Carracedo et al., 2014). Moreover, MCAR mechanism of the missing data must not due to a structural reasons (Kaiser, 2014). The example of MCAR mechanism is dropped laboratory sample, which causes observation or results missing completely (Kaiser, 2014).

For missing data which is categorized as missing at random (MAR), missing happens when the distribution of missing data for a variable only depends on observed data, but does not influenced by the missing data itself (Liu & Brown, 2013). For example, the data are missing because of a system shutdown, failure in power supply and it is not because of the values themselves (Gómez-Carracedo et al., 2014). Kaiser (2014), describe the MAR mechanism as an organization seeking to collect data on income and property tax. Normally, those with high incomes may possibly be less willing to reveal the income and property tax. If the organization has each person's property tax and given property tax non-answer to the income question is random, then the income data is missing at random. The reason for data being missing depends on property tax but does not depend on income itself (Kaiser, 2014).

Meanwhile, missing not at random (MNAR) occur when the distribution of missing data for a variables is influenced by both observed and the missing data (Liu & Brown, 2013). This happens when the values cannot be recorded because of the values are below the limits of detection, this missing of data occur because of the data itself (Gómez-Carracedo et al., 2014). According to Kaiser (2014), MNAR are considered as non-ignorable missing data and this problem can be solved by returning to the source of data and obtain a complete data set by tracking down more facts, information, and evidence about the mechanism. However, in actual it is unusual to identify the suitable model for the missing data mechanism (Kaiser, 2014). The missing data mechanism of air quality data is commonly random at MAR mechanism, which in wisdom that the possibility that a value is missing does not rely on the missing value (Junninen et al., 2004).

2.5 Descriptive Statistics Analysis on Missing Data

Descriptive statistics is a brief expressive of numbers that summarize a given data set, which can be either a portrayal of the whole or a sample of a datasets (Kenton, 2018). Descriptive statistics can be divided into two measures such as measures of central tendency (mean, median, and mode) and measures of variability (standard deviation, variance, the minimum and maximum variables, the kurtosis and skewness) of the data (Kenton, 2018). In the missing data study conducted by Quinteros et al. (2019), descriptive analysis were performed for all variables such as mean, median, percentiles and measures of dispersion along with boxplots by year to analysis the characteristics of missing data. Descriptive analysis is normally used to present the data quantity descriptions in manageable form (William, 2006).

2.6 Simulation of Missing Data

The purpose of simulation is to assess the proposed imputation method and compare its performance with standard methods, usually a simulation study was based on different missing data configurations (Junger & Leon, 2015). A mixture of distributions of gaps has been considered to randomly create missing data that applied to the observed data set which create the artificial missing data (actually the real values are known), in order to replicate the actual design of missing data, this allows to compute the value of the performance indicators to measure the goodness of the imputation methods (Plaia & Bondi, 2006).

According to Junninen et al. (2004), the imputation performance does not depend simply on the amount of missing data but also depend on the characteristics of missing data forms like the gaps of missing data. Generally, simulation procedures are very vital in missing data study because it provides additional insights that are often unrealistic or impossible to learn through real-world experimental and theoretical analysis alone (Department of Energy, 2013).

2.7 Imputation Methods to Fill in the Missing Data

The most popular method to handle the missing observation in dataset is by deleting those observations (Razak et al., 2014). According to Little and Rubin (2002); and Zainuri et al. (2015), eliminating the missing values using deletion method can introduce a significant biases, especially when the mechanism of the missing data are not randomly distributed such Missing Not at Random (MNAR). However, deletion method might be a reasonable choice if the total amount of missing observation is small compared to the number of remaining samples and a random pattern for missing such Missing at Random (MAR) or Missing Completely at Random (MCAR) occurred (Gómez-Carracedo et al., 2014).

Other than deletion method, mean substitution is also among the popular imputation method because this method is easy to be used. Mean imputation method replace the missing value by the mean value of each variable on the particular missing variables as an estimate of the missing value (Allison, 2001 & Zainuri et al., 2015). According to Gómez-Carracedo et al. (2014), the mean imputation underestimates the variance in the dataset and may change any other derived chemometric study. Besides, this method can lead to a problem of bias and large errors (Zainuri et al., 2015). However, Noor (2006), found that simple modification to mean method such mean top bottom method performed slightly well if the number of missing values is small.

According to Junninen et al. (2004), imputation methods used to replace the missing values can be divided into two methods which are single imputation and multiple imputation. Single imputation is fill in the specific one value for each missing one, subconsciously has a lot of pleasing features, this methods can be used directly and considerable effort is required to create imputations and needs to be carried out only once (Junninen et al., 2004). Basically, single imputation is the method where each missing point is replaced by only one estimated value (Razak et al., 2014). Meanwhile, multiple imputation created multiple simulated values for each missing point, to reflect properly the uncertainty in the missing data (Junninen et al., 2004). According to Mcginniss and Harel (2016), multiple imputation method is acclaimed for the capability to use complete data investigative methods on each data set plus for keeping the variability seen in the observed data to reach at estimates which are not biased by the imputation method.

2.8 Performance Measure

Performance measure is commonly used to describe the goodness of fit for each of the imputation methods (Junninen et al., 2004). In the recent study by Zakaria (2018), two types of performance indicators were used to measure the goodness of fit for each imputation methods i.e. error measure (root mean squared error and mean absolute error) and performance measure (prediction accuracy and index of agreement).

The root mean squared error (RMSE) is used as a standard statistical metric to measure a model performance in air quality, meteorology, and climate research studies, while the mean absolute error (MAE) is another useful error measure which is commonly used in model assessments (Chai & Draxler, 2014). Predictive accuracy (PA) would be measured based on the dissimilarity between the experimental values and estimated values. Though, the estimated values would be able to refer to different information (Li, 2016). Index of agreement (d_2) is another statistical measures for model performance which is usually used to compare the model estimations or predictions (P) with observations (O) that are judged to be reliable and the units of (P) and (O) are the identical. The set of model-prediction errors normally is composed of the difference between prediction and observed values, with most dimensioned measures of model performance being based on the central tendency of the datasets (Legates & Mccabe, 2012).

2.9 Researches on Imputation Method

Junninen et al. (2004), conducted a study about the applicability of a few imputation methods on an air quality data set. This study evaluated univariate methods (linear, nearest neighbour interpolation, and spline), multivariate methods (regression-based imputation (REGEM), nearest neighbour (NN), self-organizing map (SOM), multi-layer perceptron (MLP)), and hybrid methods (combination of univariate methods and multivariate methods). Several simulation pattern such as simple, medium, complex, and blended missing pattern were simulated under two missing percentage; 10% and 25%. In order to test the performance for each imputation methods, a few performance measures were used i.e. the index of agreement, the squared correlation coefficient (R^2), the root

mean squared error and the mean absolute error with bootstrapped standard errors. Generally, the result of the study show that hybridisation of multivariate method (REGEM, NN, SOM, and MLP) show a good performance to replace the missing data in dataset compared to the single imputation because it underestimates the error of variance of missing data, yet the accurateness can be improved by using multiple imputation.

Razak et al. (2014), conducted a study to compare several types of imputation methods such as Mean Substitution, Expectation Maximization (EM), and Hot Deck Method on two groups of PM_{10} data during the southwest monsoon (May 2009 until September 2009) and during the northeast monsoon (November 2009 until March 2010) in Petaling Jaya, Selanor and Seberang Prai, Pulau Pinang. In this study, PM_{10} concentration was simulated into three percentages of missing data i.e. 10%, 30%, and 50%. Then, the simulated missing data was imputed and fitted to the lognormal, gamma, Weibull, and Gumbel distributions. Four performance indicators was used to measure fits of distribution such as root mean square error (RMSE), Akaike information criterion (AIC), mean absolute error (MAE), and coefficient of determination (R^2). It was concluded that the EM imputation is shown to have great performance especially when the percentage of missing data is high as it continually gives low RMSE as compared with other methods. Weibull distribution gives an excellence fit for PM_{10} concentration during the southwest monsoon in Petaling Jaya, while the lognormal distribution outdone the others distribution in defining the PM_{10} concentration during the southwest monsoon in Seberang Perai, and gamma distribution is the finest distribution to define the PM_{10} data for the northeast monsoon in both locations.

Noor et al. (2015), used three different types of mean imputation methods such as mean, mean above, and mean above below to estimate the missing observation in annual hourly monitoring records for PM_{10} concentration in Seberang Perai, Penang. In this study, there were three degree of complexities of random simulated missing data were generated in term of percentages of missing values, for instance small (5% and 10%), medium (15% and 25%), and large (40%). Performance indicators used to calculate and describe the goodness of fit for each of the method in this research are coefficient of determination (R^2), mean absolute error (MAE), and root mean squared error (RMSE). This study concluded that mean above below imputation method was selected as a best imputation method to estimate the all types of simulated missing values.

A study to discover imputation based method suitable for multivariate time series of air quality dataset was done by Junger and Leon (2015). In this study, PM_{10} concentration data from ten monitoring air quality stations located at Sao Paulo in 2004 were used. A simulation process was performed to evaluate validity and performance of proposed method. Three missing mechanisms such as missing at random (MAR), missing not at random (MNAR), and missing completely at random (MCAR) were implemented with four pattern configuration (dispersed, in-column, in-row and sparse), and five percentages of simulation missing values (5%, 10%, 20%, 30%, and 40%). Several imputation methods namely complete case analysis, unconditional mean, median, nearest neighbour, conditional mean, expectation maximization (EM), EM spline, EM Arima, EM-GAM (generalized additive model), EM-MV (moving variance) spline, EM-MV Arima, and EM-MV GAM were applied to estimate the missing data. Besides, six performance indicators were used to define the goodness of fit of the imputed values such as root mean square error, mean absolute error, index of agreement, proportional variance, Pearson's correlation coefficient, and bias. The simulation in this study showed that when the amount of missing data was as low i.e. 5%, the complete data analysis produced reasonable results irrespective of the generating mechanism of the missing data, however when the proportion of missing values were exceeding 10%, the validity start to corrupt. The proposed imputation method displayed a decent precision and accuracy in different sets with respect to the patterns of missing values. Lastly, almost of the imputations methods obtained the valid results, although the missing distribution was MNAR.

CHAPTER 3

METHODOLOGY

3.1 Research Flow

In this study, the characteristics of the air pollution and meteorological dataset of Petaling Jaya (PJ) and Shah Alam (SA) monitoring station from 2012 to 2016 were analysed to obtain the reference data. The reference data were simulated into two percentages of missing data i.e. as 10% and 20%. The pattern of missing observation gaps for each percentages were designed with three different levels and each pattern comprises the various range of missing data gaps. The patterns used in this study were simple, medium, and complex. After that, seven imputation methods were applied to fill in the simulated missing data. The proposed imputation methods were compared to each other by using four performance indicators. The flow chart of research is shown in Figure 3.1.

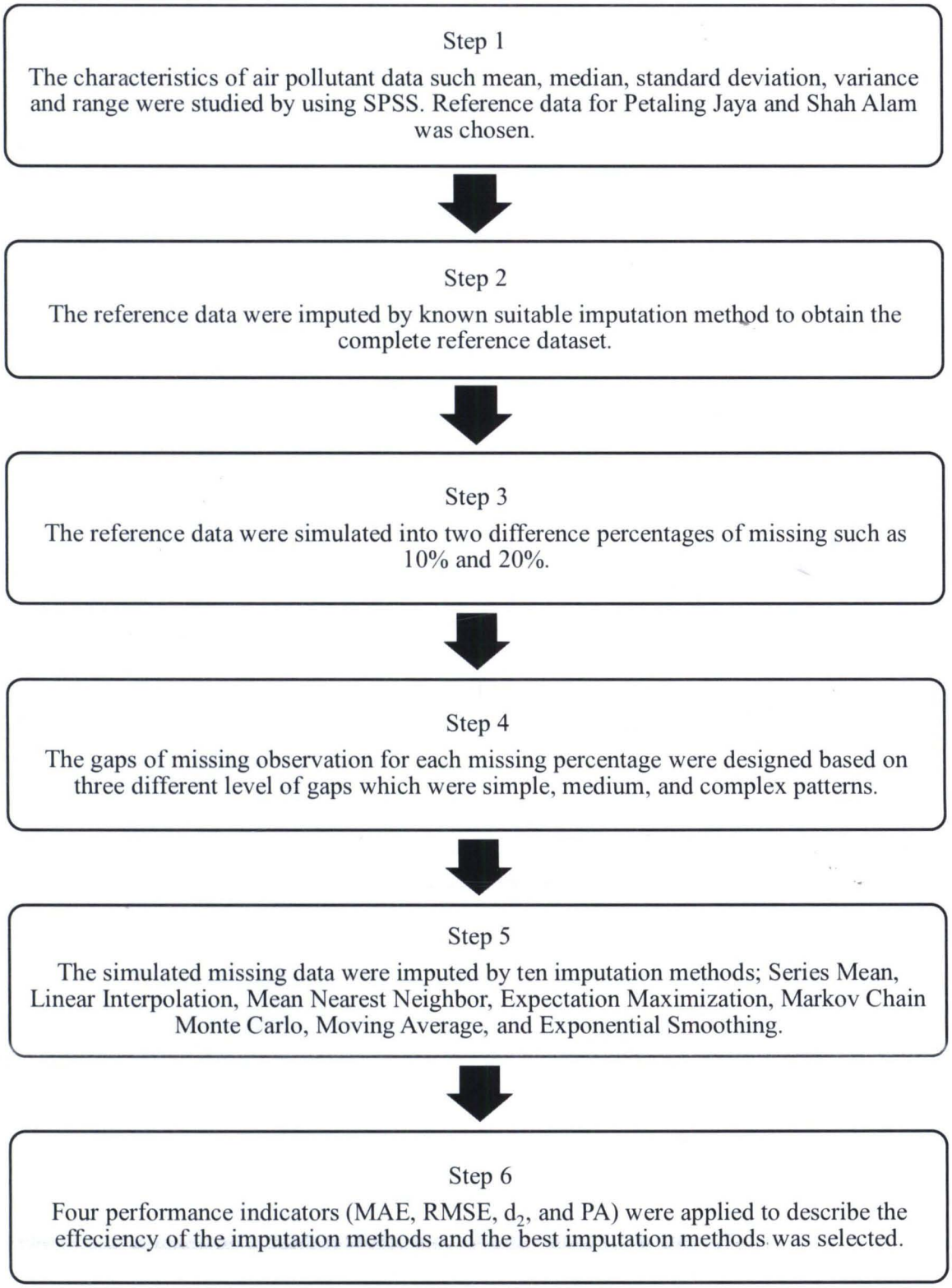


Figure 3.1: The research flowchart

3.2 Data

There are eight parameters from air quality monitoring records used in this study. These data were obtained from the Department of Environment Malaysia. Firstly, the pattern and characteristics of the missing data were analysed using descriptive statistics. The eight parameters of air pollutants were listed:

i. Particulate Matter, PM_{10} ($\mu g/m^3$)
ii. Carbon Monoxide, CO (ppm)
iii. Sulphur Dioxide, SO_2 (ppm)
iv. Nitrogen Dioxide, NO_2 (ppm)
v. Ozone, O_3 (ppm)
vi. Ambient temperature, AT (^oC)
vii. Humidity, H (%)
viii. Wind Speed, WS (km/hr)

3.3 Descriptive Statistics Analysis

The descriptive statistics for reference data were studied to analyse the characteristics for each of the eight parameters. The statistical parameter included in this descriptive statistics were mean, median, mode, standard deviation, variance, range, maximum, minimum, and percentiles. For the missing data percentages (%) in each stations and years, the gaps (in hour) of missing data for each parameters were analysed by using SPSS software and presented in a suitable graph. After that, the boxplot graph was sketch, the main function of the boxplot is to represent the distribution of the data and identify a typical observations in a univariate dataset (Bruffaerts et al., 2014). According to Hoffmann (1981), the distribution of data displayed in boxplot are minimum, first quartile, median, third quartile, maximum, and outliers. Air quality and meteorological dataset from 2012 until 2016 (5 years) were used to construct the boxplot graph to analyse the shape of the dataset.

3.4 Reference Data

Reference data is a data which contain the lowest missing data or the most complete dataset. According to Junninen et al. (2004), the most complete dataset can provide the most reasonable frame of reference. This data were used to generate the simulated missing data in simulation process i.e. the dataset were simulated into difference percentage of missing data and different pattern of gaps (in hour) of missing data. In this study, reference data is important to evaluate the similarity between predicted values and original values (Zakaria, 2018). The dataset which containing the lowest missing data were chosen as the reference data for both stations. The number and characteristics of missing gaps in reference data were counted and analysed to identify the extreme gaps (240 hours). Extreme gaps would be excluded in this study because its contain uncertainty which not be covered by common imputation methods such Mean Nearest Neighbour method. In addition, the least missing observations in dataset can help in minimizing the error because only a few of missing data needed to be filled before simulation process (Zakaria, 2018). In order to obtain the complete reference dataset, the missing gaps were imputed by the Mean Nearest Neighbour methods (Junninen et al., 2004).

3.5 Simulation of Missing Data

In this study, the simulation design can be described as to design the range of interval of missing data (gap) at different percentages of simulated missing data and different patterns of missing gaps. The simulation was conducted based on the three different types of patterns of missing gaps. Simulation of missing data are purposely done to evaluate how the parameters are affected by applying the different imputation methods to handle the missing data (Razak et al., 2014). Three random simulated missing pattern were used in this study include simple, medium, and complex. The patterns were different in term of the length of missing gaps (hour). Simple pattern contained the missing data gaps that are less than 24-hours, while medium have gaps of missing data in between 24-hours to 168-hours, and complex was a combination of simple and medium patterns in proportion of 1:1 respectively. Moreover, two percentages of missing data were applied for each patterns i.e. 10% and 20%. Table 3.1 show the length of missing gaps for each

missing pattern. This model patterns were modified from the study of (Junninen et al., 2004).

Table 3.1: The length of missing gaps for each missing pattern

Pattern	Missing Data Gaps
Simple	$l < 24$ hours
Medium	24 hours $< l \leq 168$ hours
Complex[a]	1 hours $< l \leq 168$ hours

l – The length of the gaps in hour.

[a] Simple and medium patterns are mixed in proportion of 1:1

The summary steps of simulation involved were as follows; (1) for each pattern, the minimum and maximum gaps (hour) were set, all the gaps were listed in sequence; (2), the listed set of hour from minimum to maximum for each pattern were positioned randomly by using SPSS software; (3), the randomized set of hour for each pattern were randomly chosen by the software in order to achieve the summation value (hour) of 10% and 20% percentage of missing, and; (4), the missing gaps position in reference data were randomly chosen by SPSS software and the sequence of gaps must follow the outcome of step (3). As an example, the simulation process of missing data of CO for simple pattern gaps under 10% of missing data is shown in Figure as follows;

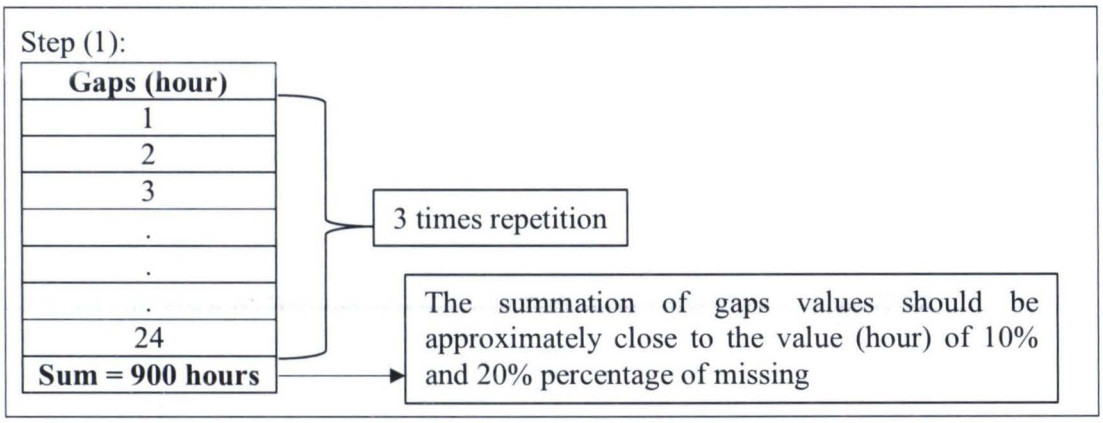


Figure 3.2: Simulation process of missing data of CO for simple pattern gaps under 10% missing data

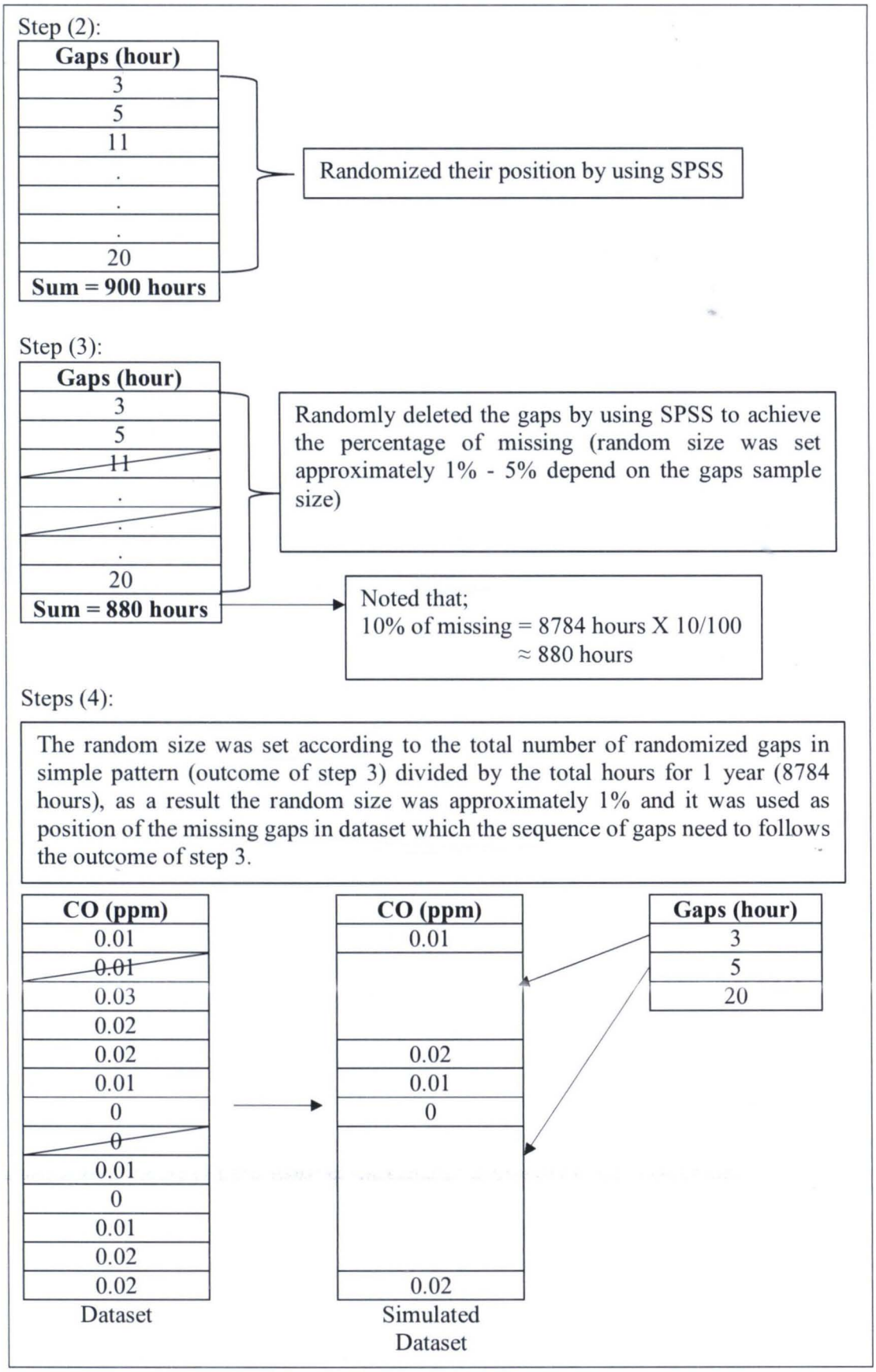


Figure 3.2: Continued

3.6 Imputation Methods

In this study, there are several imputation methods used to fill the missing values of two percentages of simulated missing data. The imputation methods used were Series Mean (SM), Mean Nearest Neighbour (MNN), Expectation Maximization (EM), Markov Chain Monte Carlo (MCMC), Linear Interpolation (LI), Exponential Smoothing (ES) with 0.2, 0.5, and 0.8 for the values of α, 12-hour Moving Average (12MA), and 24-hour Moving Average (24MA).

3.6.1 Series Mean (SM)

Series Mean method is the mean of all subjects related to a certain variable, and it is the default value in the program which is SPSS (Çokluk & Kayri, 2011). The equation of series mean as follows (Altin & Er, 2016):

$$\overline{x} = \frac{1}{N}\sum_{n=1}^{N} x_n \tag{3.1}$$

Where $\overline{x}$ the series is mean of air pollutant, N is the total number of hourly measurements for air pollutant, and x_n is the air pollutant measurements.

3.6.2 Mean Nearest neighbour (MNN)

Mean Nearest Neighbour is a method to replace the missing value with the value of its nearest and usually the upper nearest the value will be selected (Zakaria, 2018). Arithmetical mean is computed by using complete observation values under and above the missing data, then that value will be imputed instead of the missing data (Çokluk & Kayri, 2011). The equation of MNN as follow (Junninen et al., 2004):

$$y = y_1 \; if \; x \leq x_1 + \left[\frac{x_2 - x_1}{2}\right] \tag{3.2}$$

$$y = y_2 \; if \; x > \; x_1 + \left[\frac{x_2 - \; x_1}{2}\right] \tag{3.3}$$

Where y is the interpolant, x is time point of the interpolant, while x_1 and y_1 are the coordinates of the starting point of the gap, x_2 and y_2 are the coordinates of the end point of the gap.

3.6.3 Linear Interpolation (LI)

Linear Interpolation method fill the gaps of missing data by replacing the missing value with average value of the before and after data in sequential pattern (Zakaria, 2018). This method performed better for short gap of missing data (Noor, 2006). The equation of LI is written as follow (Noor et al., 2015):

$$y^* \; = \; y_1 + \left(\frac{y_2 - y_1}{x_2 - x_1}\right)(x^* - x_1) \tag{3.4}$$

Where y^* is the missing observation, $x^*$is the time of point of missing observation, x_1 and y_1 are the coordinates of the starting point of the gap, x_2 and y_2 are the coordinates of the end point of the gap.

3.6.4 Exponential Smoothing (ES)

Exponential Smoothing is a method that merge a linear trend with multiple seasonal component so that the seasonal effect are proportional to the current level of series (Akram et al., 2009). The smoothing coefficient of 0.1 until 0.9 will be used in this study. The equation of ES as follow (Glen, 2017):

$$F_t \; = \; F_{t-1} + \; \alpha(A_{t-1} - \; F_{t-1}) \tag{3.5}$$

Where F_t is a forecast for that period, F_{t-1} is a forecast for previous period, A_{t-1} is an actual demand for that period, and α is a weight (the range is between 0 and 1).

3.6.5 Moving Average (MA)

Moving Average is a method by averaging a number of points from the input signal to produce each number in the output signal (Luo et al., 2012). In this study, Zakaria (2018), suggested to use the moving average point of 12 and 24 hours to improve the estimation of missing values. The equation of MA is written as (Luo et al., 2012):

$$y_i = \left(\frac{1}{m}\right) \sum_{j=-k}^{k} x_{i+j} \tag{3.6}$$

Where x_{i+j} is the total number of hourly measurement for air pollutant based on the average will be used, y_i is the mean of air pollutant, and m is the number of point that used in moving average.

3.6.6 Expectation Maximization (EM)

Expectation Maximization method replaced the missing data with the value from the parameters estimation of incomplete dataset which obtained from maximizing the likelihood on known data (Liu and Brown, 2013). This method involves two steps which are prediction and estimation by iterative calculation (Zainuri et al, 2015). To undertake this method SPSS execute the several steps: 1) the mean, variance, and covariance are estimated from the individual complete data, 2) the maximum likelihood procedures will be used to estimate a regression equations that relate each variable to each other variable which generate the formula, and 3) the formula are used to estimate the missing values (Moss, 2016).

3.6.7 Markov chain Monte Carlo (MCMC)

Markov Chain Monte Carlo method were used to fill in each of missing data by averaging or pooling multiple simulated value that were generated for each of the missing data, the purpose of generating multiple simulated is to reflect the uncertainty that attach

to missing value (Junninen et al., 2004). The data are assumed from a multivariate normal distribution, then data augmentation will be applied to Bayesian inference with missing data by repeating several steps such as the imputation I-step and posterior P-step, these two steps are iterated long enough to produce the results to be reliable for a multiply imputed data set (Schafer, 1997). This method objective is to have iterates converge to stationary distribution and then to simulate an approximately independent draw of the missing values (Yuan, 2005).

3.7 Performance Indicator

There are four performance indicators to measure the goodness of fit of the imputation methods used in this study such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Index of Agreement (d_2), and Prediction Accuracy (PA). The performance for each imputation methods were displayed in the form of rank and the best imputation methods for overall and for each pattern of simulated missing data were selected. Table 3.2 show the performance indicators formulae.

Table 3.2: Performance indicators formulae (Noor et al., 2015)

Performance Indicator	**Formula**
Mean Absolute Error (MAE)	$MAE = \frac{1}{N}\sum_{i=1}^{N}\lvert P_i - O_i\rvert$
Root Mean Squared Error (RMSE)	$RMSE = \left(\frac{1}{N}\sum_{i=1}^{N}[P_i - O_i]^2\right)^{\frac{1}{2}}$
Index of Agreement (d_2)	$d_2 = 1 - \left[\frac{\sum_{i=N}^{N}(P_i - O_i)^2}{\sum_{i=N}^{N}(\lvert P_i - \bar{O}\rvert + \lvert O_i - \bar{O}\rvert)^2}\right]$
Prediction Accuracy (PA)	$PA = \sum_{i=1}^{N}\frac{[(P_i - \bar{P})(O_i - \bar{O})]}{(N-1)\sigma_P\sigma_O}$

Where N is the number of imputations, O_i is the observed data points, P_i is the imputed data points, $\bar{P}$ is the average of imputed data, $\bar{O}$ is the average of observed data, σ_P is the standard deviation of the imputed data, and σ_O is the standard deviation of observed data.

3.8 The Ranking Model of the Method

The results of the performance for each imputation methods were ranked in the ranking model which is based on the performance measure by Index of Agreement (d_2). The rank of imputation methods from the best to worst were based on the agreement between observed and predicted values for each parameters used in this study (Zakaria, 2018). This rank were represented by the level of the number from 1 to 10 since there are 10 imputation methods used to fill in the various simulated missing data. The best imputation method would be rank as number 1, meanwhile the worst as the number 10.

3.9 Scatter Plot

In this study Scatter plot was used to evaluate the predicted values from the best imputation method by the observed values (reference data) for all parameters. The correlation coefficient (R^2) values in these graphs were used to describe the variability of predicted data (y-axis) that has been clarified by observed data (x-axis) (Zakaria, 2018). According to Noor et al. (2008), the R^2 value describe the agreement between the predicted data and the observed data. The closer the value R^2 to 1 the better the prediction due to the stronger relationship between x-axis with y-axis (Siegel, 2011). Zakaria (2018), randomly chose one of the pattern gaps and percentages of missing data to evaluate the strength of relationship between predicted values (best imputation methods) and the observed values (reference data). In this study complex pattern for 10% - simulated missing data was used.

CHAPTER 4

RESULTS AND DISCUSSION

The results obtained from this study are presented and discussed in this chapter. This chapter is mainly divided into three sections. In the first section of this chapter, the characteristics of raw air quality data and selection of the reference data were discussed. In the next section, the characteristics of the simulated missing data were explained. Lastly, the performance of imputation methods applied to the simulated missing data were discussed.

4.1 Characteristics of Air Pollution Datasets

In this study, the characteristics of hourly concentration of 5 air pollutant datasets and 3 parameters of meteorological dataset were analysed. Figures 4.1 (a) until 4.1 (h) show the boxplot plot of meteorological parameters and air pollutants concentration in Petaling Jaya and Shah Alam from 2012 to 2016. Boxplots often provide the information about the shape of a data set. In this study, all the data distribution except for humidity were skewed to the right, which has the mean value higher than the median value. Normally when the data were skewed to the right it indicates that there were extreme events occurred (Md Yusof et al., 2010). Annual dry season in Malaysia is the factors that contribute to the distribution of humidity data skewed to the left.

The interquartile range is indicated by the length of the box. The upper interquartile in the boxplot indicates the maximum value of air pollution concentration. Figure 4.1 shows that the upper interquartile range (for wind speed, ambient temperature, humidity, O_3, and PM_{10}) in Shah Alam was higher than Petaling Jaya. Meanwhile in Petaling Jaya the upper interquartile range (for SO_2, NO_2, and CO) were higher than Shah Alam. Shah Alam was recognized as urban area and Petaling Jaya as industrial area

(Department of Environment, 2017). For instance, the highest air pollutant concentration recorded in Petaling Jaya (industrial area) was SO_2 and NO_2 which are contributed majorly produced from motor vehicle followed by power station, and industrial activities (Department of Environment, 2012).

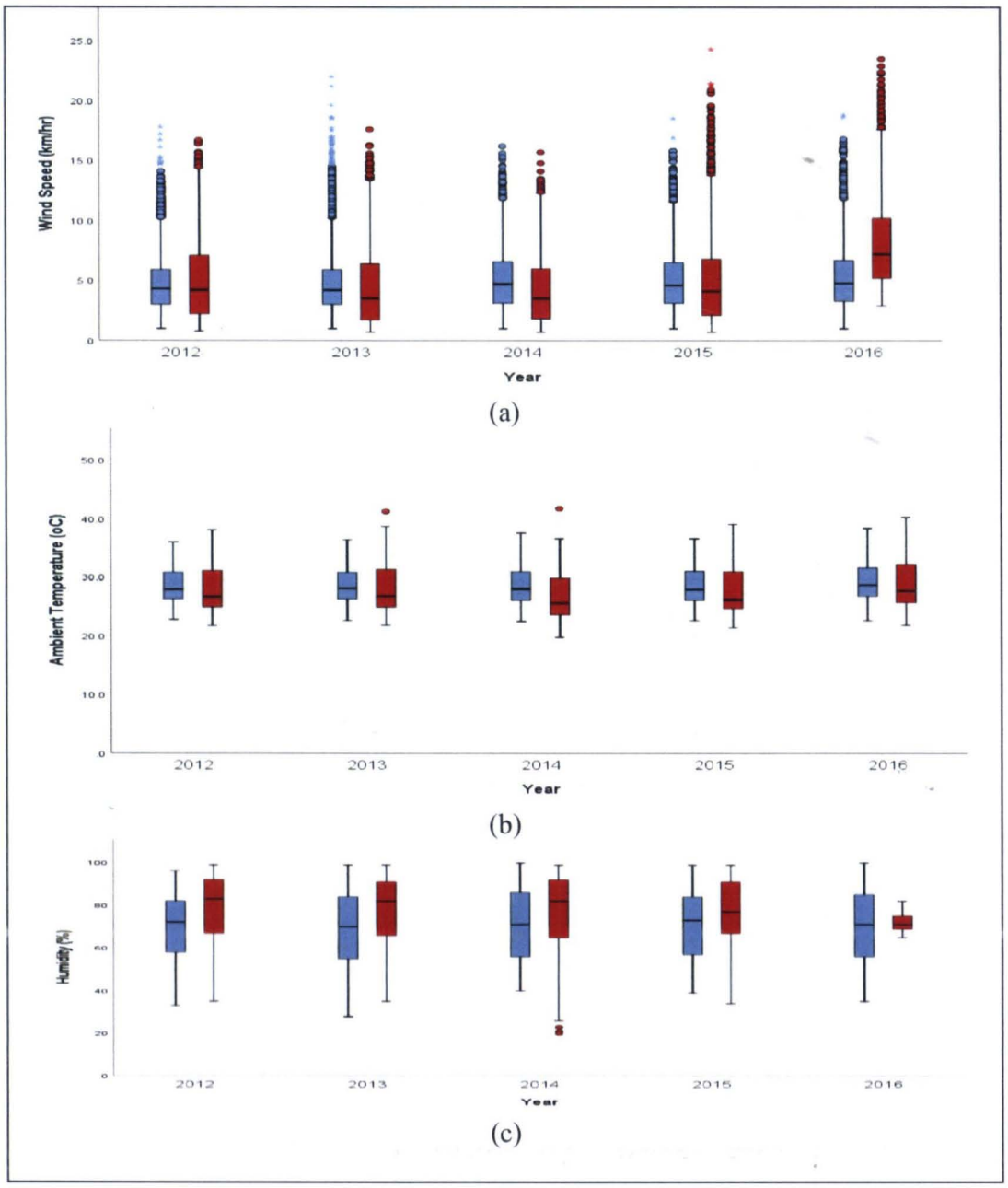


Figure 4.1: The boxplot for (a) wind speed, (b) ambient temperature, (c) humidity, (d) SO_2, (e) NO_2, (f) CO, (g) O3, and (h) PM_{10} in Petaling Jaya and Shah Alam from 2012 to 2016

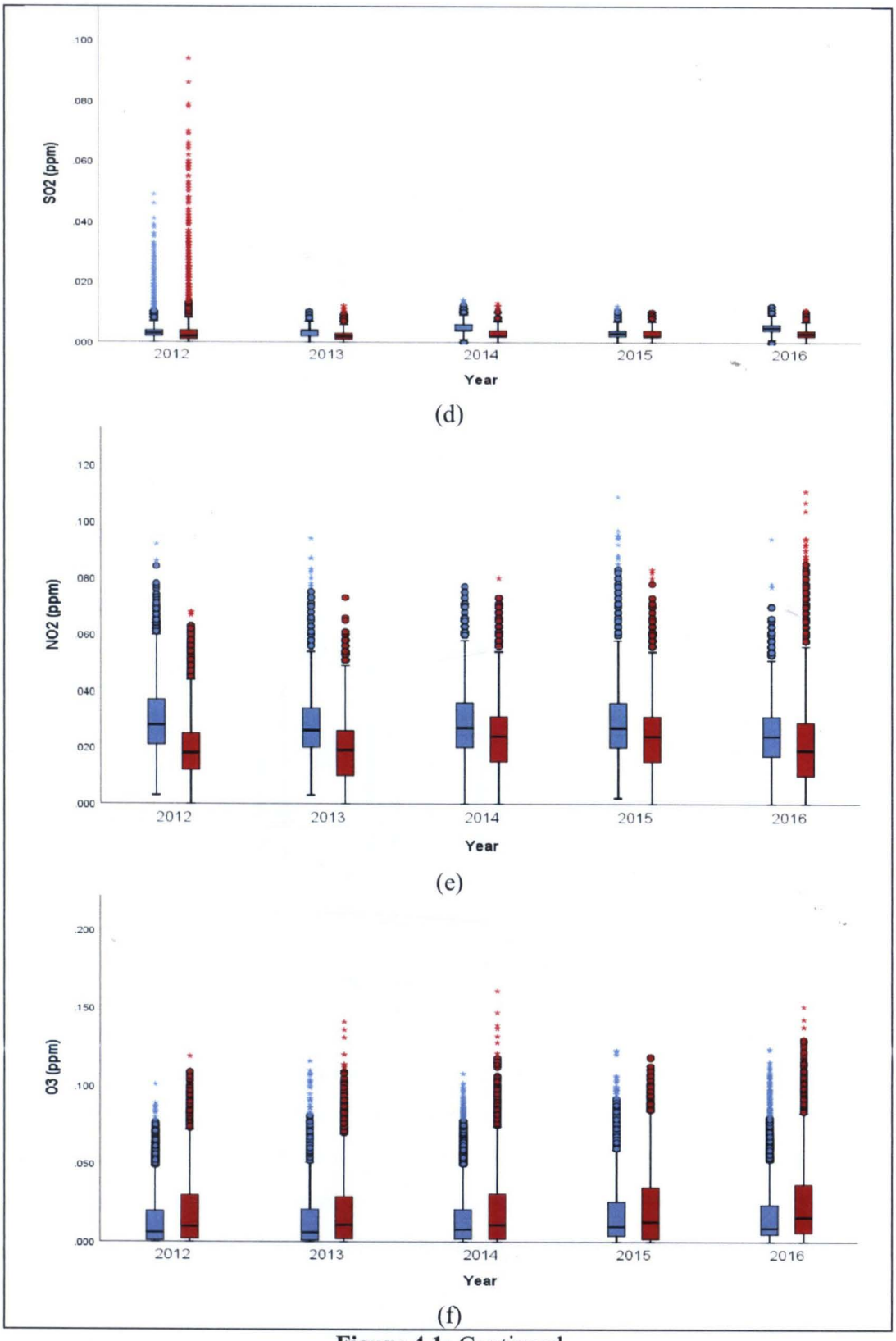


Figure 4.1: Continued

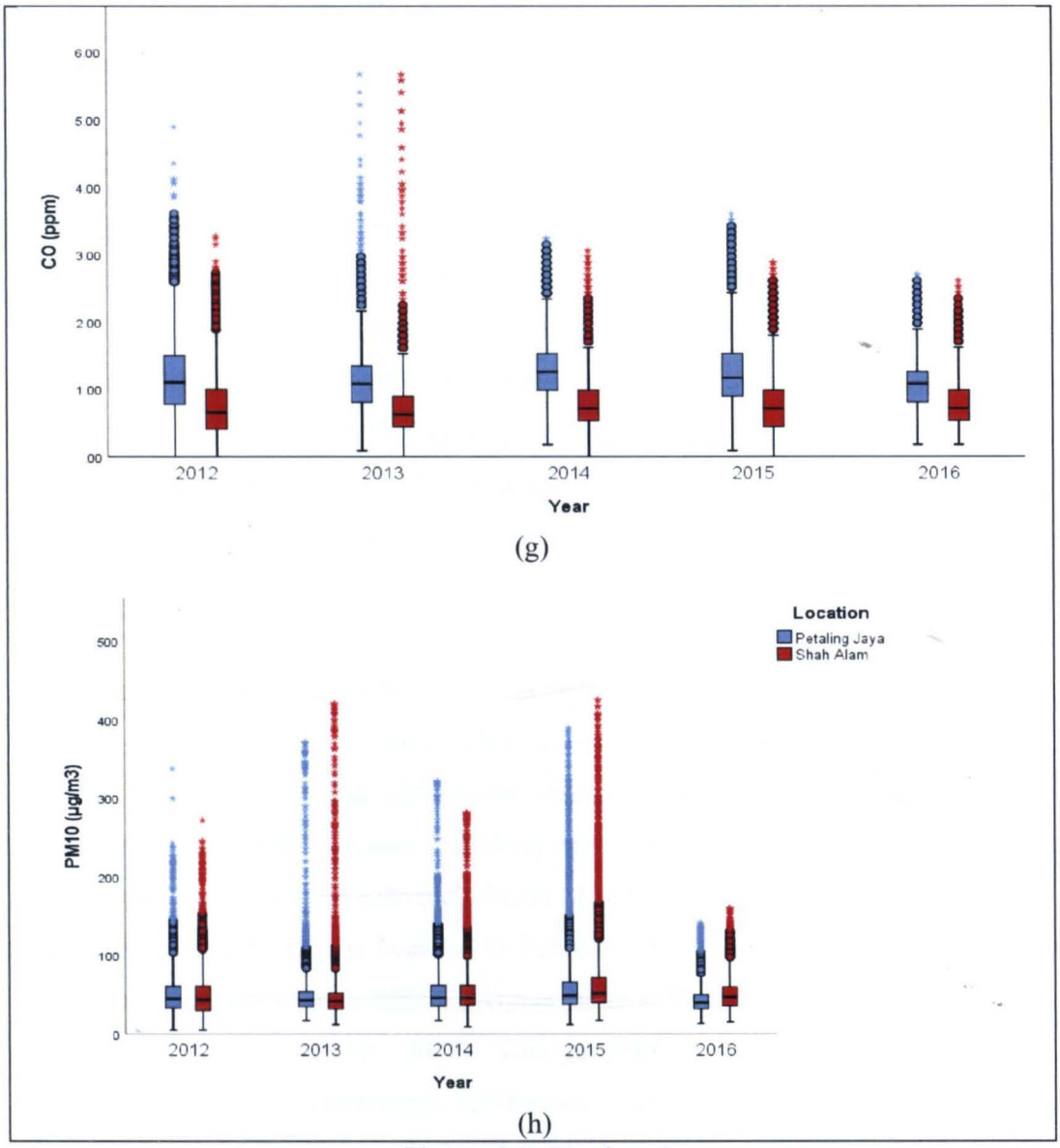


Figure 4.1: Continued

The length of the box also can describe the variability of each air quality parameter. The longer length of box show the higher variability of the data. The distribution of air pollutant concentration in Shah Alam was more varied compared to Petaling Jaya. This is because the length of boxplot of air pollutant concentration in Shah Alam was greater than Petaling Jaya. The location of Shah Alam which is near to the Klang compared to the Petaling Jaya was affected by the variability of air quality data in Klang. This is because Klang is the centre of Selangor and the hub of many industrial and activities (Rahman et al., 2015). Hence, air pollutants in Klang has a high chance to

be transported to Shah Alam area, resulting in the increase the variety of air pollutants concentration in Shah Alam.

The round shapes in the boxplot are indicated as data points with value more than maximum or minimum value, meanwhile the star shapes indicate the points that are more extreme than the outlier, these star shapes are also known as extreme outlier. Based on Figure 4.1 (h), observation on PM_{10} data for both locations contained the highest reading of extreme outliers in 2013 and 2015 which is expected due to haze episodes that occurred from 15 to 27 June 2013 and 22 August to 26 October 2015. Haze in 2015 was reported as one of the worst episodes since 1997 that lasted for more than two months and affected the whole country due to tropical cyclones on that time (Department of Environment, 2015).

SO_2 concentration records show the highest extreme outliers in 2012 compared to the other years for both locations. There are no concrete reasons for this occurrence. However, the best speculation of this event is the potential source of SO_2 emissions from the industrial area, as the industry frequently emits the gas effluent into the atmosphere. Furthermore, the number of extreme outliers of SO_2 concentration from 2013 onwards was drastically drop. This is because in February 2012, Department of Environment implemented a voluntary registration exercise for air pollution control (APC) consultants (Department of Environment, 2012). This program covered the legal framework, understanding Guidance Documents, submission requirements, performance monitoring, record keeping and performance education of air quality management. Therefore, no extreme SO_2 concentration was recorded since 2013.

Overall, the concentration of NO_2 in Petaling Jaya was higher compared to Shah Alam. Since Petaling Jaya is an industrial area, it is no doubt that the area was experiencing heavy traffic congestion due to vehicles such as lorry from industry and vehicle from workers. According to Casquero-Vera et al. (2019), NO_2 is normally emitted by the transportation. However, in year of 2016 the maximum concentration of NO_2 was recorded in Shah Alam, possibly due to the increasing population and development in Shah Alam which may increase the number of transportations.

On the other hand the CO concentration in 2013 shows the highest number of outliers compared to the other years for both locations. According to Colvile et al. (2002), CO emission mainly originated from transport power sources due to incomplete combustion process. In this case, the CO concentration rose up in 2013, which is suspected due to the haze event. Incomplete combustion of forest from neighbour country increase the level of CO concentration in 2013 for both locations. However, CO concentration level seem to be unaffected by the haze event in 2015, possibly because the tropical cyclone assisted in completing the combustion of forest in neighbour country during year 2015.

4.2 Selection of Reference Data

In order to select the reference data, the amount of missing data from 2012 to 2016 in Petaling Jaya and Shah Alam were analysed. This process is important to obtain a complete dataset for as performance evaluation between the predicted and observed values (Zakaria, 2018). Figure 4.2 shows the percentage of total missing observation (%) for all datasets from 2012 until 2016 for Petaling Jaya and Shah Alam. The highest missing data in Shah Alam was 29.68% (2015) and the lowest was 4.88% (2012), meanwhile the highest missing data in Petaling Jaya was 3.14% (2012) and the lowest in 2016 which is 2.11%. The percentage of missing data between Shah Alam and Petaling Jaya shows the large significant difference in all of 5 years except in 2012. This significant difference show that the dataset in Shah Alam from 2013 to 2016 are not suitable for used as reference data.

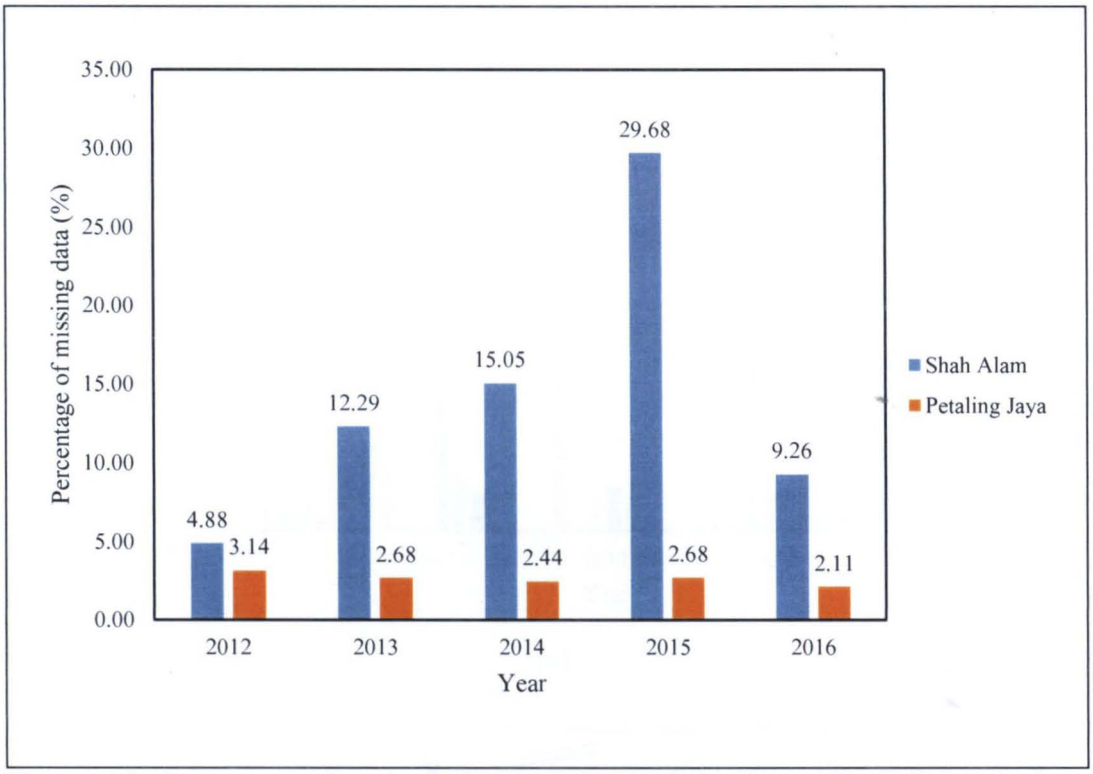


Figure 4.2: Percentage of total missing data (%) at Shah Alam and Petaling Jaya from 2012 to 2016

Figures 4.3 (a) and (b) show the longest gap of missing observation of air pollution (in hour) for Shah Alam and Petaling Jaya from 2012 to 2016. The longest gap of missing data in Shah Alam was 7963 hours (2014) for ambient temperature measurements records and that gaps of missing values are equal to about 332 days which was 90.7% of days in a year. The smallest gap of missing data observed were in humidity measurement at 46 hours (2012) and 46 hours (2014) of PM_{10} measurement. In Petaling Jaya (Figure 3.3 (b)), NO_2 records contains the longest gaps of missing data which is 743 hours (2013) and it is equivalent to a month. Meanwhile, the smallest gap of missing data recorded was 3 hours (2016) in CO measurement records. However, in 2014 there was no missing data recorded in SO_2 concentration records.

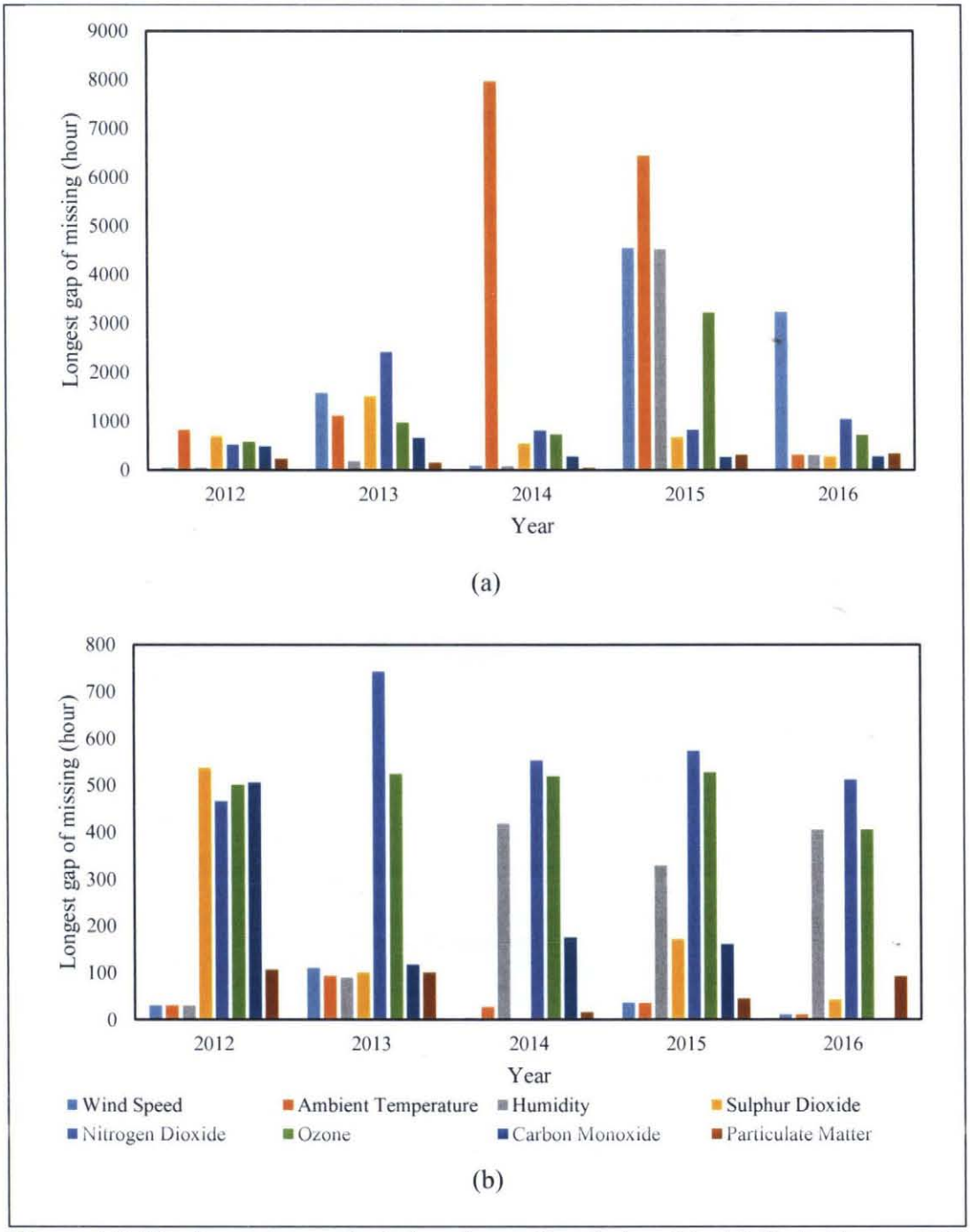


Figure 4.3: The longest gap of missing observation (in hour) in (a) Shah Alam and (b) Petaling Jaya from 2012 to 2016

All air quality parameters in Shah Alam (Figure 4.3 (a)) contained more than 1000 hours gaps of missing except in 2012. Therefore, air quality dataset in 2012 contained the least missing data compared to other year in Shah Alam. Furthermore, the longest gap of missing data was observed in ambient temperature monitoring records of 2012 that was 824 hours. In Petaling Jaya, all air quality parameters contain less than 1000 hours of

missing gaps. According to Noor et al. (2015), the reference data was defined as the dataset that have the most complete data. As for Petaling Jaya the most complete data which containing the lowest total missing data was in 2016. However, in order to select the most suitable dataset as reference data for this research, dataset in 2012 was selected as the most suitable reference data for both location since that dataset only contained the lowest total missing data in Shah Alam and the total missing data in Petaling Jaya slightly below than Shah Alam in 2012.

4.3 The Descriptive Statistics of Reference Data

Table 4.1 shows the summary of descriptive statistics for all measurements of air pollutants at Shah Alam and Petaling Jaya in 2012. The mean value for all parameters except for humidity was higher than the median value. Therefore, the distribution of these measurements except for humidity were skewed to the right, indicating that there were several observations of high concentration of air pollutant occurred in this year. Meanwhile, the mean value for humidity monitoring records is lower than the median value in Petaling Jaya and Shah Alam which indicates the distribution of data were skewed to the left. This result show that the weather in Petaling Jaya and Shah Alam mostly was hot and dry because the distribution of data skewed to the left which means the humidity observation tends to the less humid in this year.

The highest amount of missing observation in Petaling Jaya was 537 hours missing data of SO_2, whereas for Shah Alam ambient temperature monitoring was missing for 824 hours. The lowest number of missing data recorded at Petaling Jaya was 30 observations for wind speed, ambient temperature, and humidity monitoring records. Meanwhile, only 46 missing data of humidity monitoring was recorded as the lowest in Shah Alam. The standard deviation represent the variability of the data in air pollution monitoring dataset. Based on Table 4.1, for both locations, PM_{10} concentration were recorded having the highest standard deviations of 25 $\mu g/m^3$ and 28 $\mu g/m^3$ in Petaling Jaya and Shah Alam respectively. PM_{10} concentration in Shah Alam was slightly more variable compared to the Petaling Jaya, therefore persistent values of PM_{10} most likely recorded in the Petaling Jaya compared to Shah Alam. Since, PM_{10} concentration

distribution varies more compared to the other parameters for both locations, hence, the range of the PM_{10} concentration was the highest.

Table 4.1: Descriptive statistics for all parameters Shah Alam and Petaling Jaya (2012)

	Wind Speed		Ambient Tamperature		Humidity		Sulphur Dioxide		Nitrogen Dioxide		Ozone		Carbon Monoxide		Particulate Matter	
	PJ	SA	PJ	SA	PJ	SA	PJ	SA	PJ	SA	PJ	SA	PJ	SA	PJ	SA
Total N	8784	8004	8784	8004	8784	8004	8784	8004	8784	8004	8784	8004	8784	8004	8784	8004
Valid N	8754	7960	8754	7960	8754	7960	8247	7342	8318	7554	8283	7466	8278	7563	8677	7796
Missing	30	44	30	44	30	44	537	662	466	450	501	538	506	441	107	208
Mean	4.6	5	28.5	28	70	79	0.003	0.003	0.03	0.019	0.013	0.018	1.178	0.753	49	49
Median	4.3	4.2	27.9	26.7	72	83	0.003	0.002	0.028	0.018	0.006	0.01	1.1	0.66	45	44
Mode	3.5	1	26.9	25.1	87	95	0.002	0.002	0.025	0.014	0	0.001	0.99	0.43	38	34
Std Deviation	2.3	3.3	2.8	3.7	15	14	0.003	0.005	0.012	0.01	0.015	0.02	0.563	0.462	25	28
Variance	5.2	10.9	7.8	14.1	213	210	0	0	0	0	0	0	0.317	0.213	649	761
Range	16.8	15.8	13.2	16.4	63	64	0.049	0.094	0.089	0.068	0.101	0.119	4.89	3.28	333	267
Maximum	17.8	16.7	36	38.1	96	99	0.049	0.094	0.092	0.068	0.101	0.119	4.89	3.28	338	272
Minimum	1	0.9	22.8	21.7	33	35	0	0	0.003	0	0	0	0	0	5	5
Percentile 25	3	2.2	26.3	24.9	58	67	0.002	0.001	0.021	0.012	0.001	0.002	0.78	0.41	33	30
Percentile 50	4.3	4.2	27.9	26.7	72	83	0.003	0.002	0.028	0.018	0.006	0.01	1.1	0.66	45	44
Percentile 75	5.9	7.3	30.8	31.1	82	92	0.004	0.004	0.037	0.025	0.02	0.03	1.5	1	61	61

Where: PJ – Petaling Jaya, SA – Shah Alam

Tables 4.2 and 4.3 show the percentages for length of missing gaps (in hour) for all air pollutants parameters at Petaling Jaya and Shah Alam in 2012. The gap of missing data was presented in percentage and the gap interval of this table was 3 hours. About 82 % and 65 % of the missing gaps recorded in Petaling Jaya and Shah Alam were 1 hour respectively. The longest gap of missing data at Petaling Jaya was recorded by carbon monoxide (0.24%) with the gaps of 69 hours to 72 hours, while in Shah Alam, ambient temperature (9.09%) was recorded as the longest missing gap at more than 99 hours. In order to simulate the missing data, the missing values in the reference dataset need to be filled, for this study the nearest neighbour method was used to complete the missing values (Zakaria, 2018). For Shah Alam, the length gap of missing data was more than 99 hours (actual of missing hours was 768 hours or 32 days) referring to ambient temperature records. Therefore, the whole row of missing observation gaps in Shah Alam (768 hours) including other parameters were deleted to prevent error. According to Noor et al. (2013), such long gap of missing data can cause error in prediction due to the uncertainties are not covered by imputation methods.

Table 4.2: The percentages for length of gaps for the missing data in Petaling Jaya 2012

Length of Gap (hour)	Missing Data in Gaps (%)								
	WS	**AT**	**H**	**SO_2**	**NO_2**	**O_3**	**CO**	**PM_{10}**	**Mean**
$l = 1$	71.43	71.43	71.43	92.19	94.19	94.00	94.93	65.71	81.91
$1 < l \leq 3$	14.29	14.29	14.29	6.30	5.33	4.80	4.83	22.86	10.87
$3 < l \leq 6$	-	-	-	0.50	0.24	0.48	-	-	0.41
$6 < l \leq 9$	-	-	-	0.25	-	0.48	-	2.86	1.20
$9 < l \leq 12$	-	-	-	0.25	-	-	-	2.86	1.55
$12 < l \leq 15$	-	-	-	-	-	-	-	-	
$15 < l \leq 18$	-	-	-	-	-	-	-	-	
$18 < l \leq 21$	-	-	-	-	-	-	-	-	
$21 < l \leq 24$	14.29	14.29	14.29	-	0.24	-	-	5.71	9.76
$24 < l \leq 27$	-	-	-	-	-	-	-	-	
$27 < l \leq 30$	-	-	-	-	-	-	-	-	
$30 < l \leq 33$	-	-	-	-	-	-	-	-	
$33 < l \leq 36$	-	-	-	-	-	-	-	-	
$36 < l \leq 39$	-	-	-	0.25	-	-	-	-	0.25
$39 < l \leq 42$	-	-	-	-	-	-	-	-	
$42 < l \leq 45$	-	-	-	0.25	-	0.24	-	-	0.25
$45 < l \leq 48$	-	-	-	-	-	-	-	-	
$48 < l \leq 51$	-	-	-	-	-	-	-	-	
$51 < l \leq 54$	-	-	-	-	-	-	-	-	
$54 < l \leq 57$	-	-	-	-	-	-	-	-	
$57 < l \leq 60$	-	-	-	-	-	-	-	-	
$60 < l \leq 63$	-	-	-	-	-	-	-	-	
$63 < l \leq 66$	-	-	-	-	-	-	-	-	
$66 < l \leq 69$	-	-	-	-	-	-	-	-	
$69 < l \leq 72$	-	-	-	-	-	-	0.24	-	0.24
TOTAL	100	100	100	100	100	100	100	100	100

Where: l – the length gap in hour, WS – wind speed, AT – ambient temperature, H – humidity, SO_2 – sulphur dioxide, NO_2 – nitrogen dioxide, O_3 – ozone, CO – carbon monoxide, PM_{10} – particulate matter.

Table 4.3: The percentages for length of gaps for the missing data in Shah Alam 2012

Length of Gap (hour)	Missing Data in Gaps (%)								
	WS	**AT**	**H**	**SO_2**	**NO_2**	**O_3**	**CO**	**PM_{10}**	**Mean**
$l = 1$	41.67	36.36	36.36	86.71	86.63	89.94	93.15	52.38	65.40
$1 < l \leq 3$	50.00	45.45	54.55	9.37	9.73	8.18	4.36	23.81	25.68
$3 < l \leq 6$	-	-	-	1.51	1.52	0.31	1.56	12.70	3.52
$6 < l \leq 9$	-	-	-	0.30	0.91	0.31	-	-	0.51
$9 < l \leq 12$	-	-	-	-	-	-	-	3.17	3.17
$12 < l \leq 15$	-	-	-	0.30	0.30	-	-	-	0.30
$15 < l \leq 18$	-	-	-	0.91	-	0.31	-	4.76	1.99
$18 < l \leq 21$	-	-	-	-	0.30	-	-	-	0.30
$21 < l \leq 24$	-	-	-	-	-	-	-	-	
$24 < l \leq 27$	8.33	9.09	9.09	-	0.30	0.31	0.31	1.59	4.15
$27 < l \leq 30$	-	-	-	-	-	-	-	1.59	1.59
$30 < l \leq 33$	-	-	-	-	-	-	-	-	
$33 < l \leq 36$	-	-	-	-	-	-	0.31	-	0.31
$36 < l \leq 39$	-	-	-	-	-	-	-	-	
$39 < l \leq 42$	-	-	-	-	-	-	-	-	
$42 < l \leq 45$	-	-	-	-	-	-	-	-	
$45 < l \leq 48$	-	-	-	-	0.30	-	-	-	0.30
$48 < l \leq 51$	-	-	-	-	-	-	-	-	
$51 < l \leq 54$	-	-	-	-	-	-	-	-	
$54 < l \leq 57$	-	-	-	-	-	-	-	-	
$57 < l \leq 60$	-	-	-	-	-	-	-	-	
$60 < l \leq 63$	-	-	-	-	-	-	-	-	
$63 < l \leq 66$	-	-	-	-	-	-	-	-	
$66 < l \leq 69$	-	-	-	-	-	-	0.31	-	0.31
$69 < l \leq 72$	-	-	-	0.30	-	-	-	-	0.30
$72 < l \leq 75$	-	-	-	-	-	-	-	-	
$75 < l \leq 78$	-	-	-	-	-	-	-	-	
$78 < l \leq 81$	-	-	-	-	-	0.31	-	-	0.31
$81 < l \leq 84$	-	-	-	-	-	-	-	-	
$84 < l \leq 87$	-	-	-	-	-	-	-	-	
$87 < l \leq 90$	-	-	-	-	-	-	-	-	
$90 < l \leq 93$	-	-	-	-	-	-	-	-	
$93 < l \leq 96$	-	-	-	-	-	-	-	-	
$96 < l \leq 99$	-	-	-	-	-	0.31	-	-	0.31
$l > 99$	-	9.09	-	-	-	-	-	-	9.09
TOTAL	100	100	100	100	100	100	100	100	100

Where: l – the length gap in hour, WS – wind speed, AT – ambient temperature, H – humidity, SO_2 – sulphur dioxide, NO_2 – nitrogen dioxide, O_3 – ozone, CO – carbon monoxide, PM_{10} – particulate matter.

4.4 Characteristics of the Simulated Missing Data

Table 4.4 show the percentage of missing data gap (in hour) for each of the missing gap patterns of simulated missing data. The number gap of missing data was presented in percentages and the gap intervals for simple pattern was 6 hours and for both medium and complex pattern were 24 hours. The minimum and maximum range gap of the each patterns were designed according to simulation design in Table 3.1. However, the results after simulation process, the range of maximum and minimum gap length were slightly different but still in the set range. This is due to the simulation process that was generated randomly by the SPSS software (Zakaria, 2018).

Table 4.4: Percentage of missing data gap (in hour) for each of the missing gap patterns of simulated missing data

Simple					
Length of gap, l (hour)	Missing data in gaps length (hour) for simulated missing data (%)				Mean
	PJ		SA		
	10%	20%	10%	20%	
$1 < l \leq 6$	30.50	24.67	28.60	24.27	27.01
$6 < l \leq 12$	24.00	25.02	25.83	25.54	25.10
$12 < l \leq 18$	22.17	25.29	22.88	25.34	23.92
$18 < l \leq 24$	23.33	25.02	22.69	24.85	23.98
Total	100	100	100	100	100
Medium					
$24 < l \leq 48$	12.86	19.86	22.86	15.15	17.68
$48 < l \leq 72$	14.29	15.07	20.00	15.91	16.32
$72 < l \leq 96$	21.43	13.01	12.86	16.67	15.99
$96 < l \leq 120$	8.57	15.07	17.14	21.21	15.50
$120 < l \leq 144$	25.71	21.23	15.71	15.15	19.45
$144 < l \leq 168$	17.14	15.75	11.43	15.91	15.06
Total	100	100	100	100	100
Complex					
$l \leq 24$	50.00	49.83	50.34	48.88	49.76
$24 < l \leq 48$	13.75	13.86	11.56	13.06	13.06
$48 < l \leq 72$	11.88	11.88	14.97	13.06	12.95
$72 < l \leq 96$	7.50	6.93	8.16	5.60	7.05
$96 < l \leq 120$	6.25	5.94	7.48	7.46	6.78
$120 < l \leq 144$	4.38	6.27	4.08	6.34	5.27
$144 < l \leq 168$	6.25	5.28	3.40	5.60	5.13
Total	100	100	100	100	100

Where: l – the length of gap, PJ – Petaling Jaya, SA – Shah Alam

Based on Table 4.4, the distribution of gaps in simple and medium patterns were slightly equal. The highest distribution of missing gaps in simple pattern was about 27.01% of mean gaps for 1 to 6 hours and for the lowest was 23.92% for 12 to 18 hours. Meanwhile in the medium patterns the highest was about 19.45% of the missing gaps distributed in 120 to 144 hours of missing and the lowest was 15.06% for the gaps of 144 to 168 hours. In complex pattern, the distribution of simulated missing gaps with the gaps of more than 24 hours was equal to 50.24% and 49.76% for the gaps not more than 24 hours. These mean distribution percentages of gaps were consistent with the proportion of simple and medium patterns in the complex pattern of simulation design (Table 3.1) in which the proportion was 1:1 between simple and medium patterns.

Tables 4.5 and 4.6 show the descriptive statistics of simulated missing data for Petaling Jaya and Shah Alam. Generally, the pattern of descriptive statistics for 10% and 20% simulated missing data did not vary much from one to another percentages of missing data although the pattern of gaps for each percentage of missing was varied. For instance, the mean of humidity observation in Petaling Jaya was 70% for simple, medium, and complex patterns of gaps under 10% and 20% of percentages missing data. According to Noor (2006), this occurrence was due to the random number generated in producing the simulated missing values patterns and the availability of large number of observation with the same range.

Based on both of Tables, the mean values for all parameters except humidity was higher than the median values and PM_{10} measurement records contained the highest standard deviation values for all percentages of simulated missing data in Petaling Jaya and Shah Alam. These results were consistent with the descriptive statistics of the reference data (Table 4.1). Therefore, from the descriptive statistics it can be seen that the main structure of the reference data was not disturbed by the simulation process. This finding is consistent with the other the study e.g. by Zakaria (2018), who also found that the structure of the simulated of missing data not interrupted after simulation process.

Table 4.5: Descriptive statistics of the simulated missing data at Petaling Jaya

Petaling Jaya														
		%	Valid N	Missing	Mean	Median	Mode	Standard Deviation	Variance	Range	Min	Max	Percentile 25	Percentile 75
WS	Simple	10	7899	885	4.6	4.2	3.5	2.3	5.2	16.8	1	17.8	3	5.9
		20	7012	1772	4.6	4.2	3.5	2.3	5.3	16.8	1	17.8	2.9	5.9
	Medium	10	7894	890	4.6	4.3	3.5	2.3	5.2	16.8	1	17.8	3	5.9
		20	7028	1756	4.6	4.2	3.5	2.3	5.3	16.8	1	17.8	2.9	5.8
	Complex	10	7900	884	4.6	4.2	3.5	2.3	5.2	16.8	1	17.8	3	5.9
		20	7024	1760	4.7	4.3	3.5	2.3	5.2	16.8	1	17.8	3	6
AT	Simple	10	7907	877	28.5	28	26.9	2.8	7.8	13.2	22.8	36	26.3	30.8
		20	7025	1759	28.5	27.9	26.9	2.8	7.7	13.2	22.8	36	26.3	30.8
	Medium	10	7890	894	28.5	27.9	26.9	2.8	7.8	13.2	22.8	36	26.2	30.8
		20	7028	1756	28.6	28	26.9	2.8	7.8	12.9	22.8	35.7	26.3	30.9
	Complex	10	7899	885	28.5	28	26.9	2.8	7.8	13.2	22.8	36	26.3	30.8
		20	7025	1759	28.6	28	26.9	2.8	7.9	13.2	22.8	36	26.3	30.8
H	Simple	10	7907	877	70	72	87	15	214	63	33	96	58	82
		20	7028	1756	70	72	80	14	208	63	33	96	58	82
	Medium	10	7896	888	70	72	87	15	215	63	33	96	58	82
		20	7022	1762	70	71	80	15	220	63	33	96	57	82
	Complex	10	7898	886	70	72	87	15	213	63	33	96	58	82
		20	7023	1761	70	72	87	14	210	63	33	96	58	82
SO_2	Simple	10	7890	894	0.003	0.003	0.002	0.003	0	0.046	0	0.046	0.002	0.004
		20	7031	1753	0.003	0.003	0.002	0.003	0	0.049	0	0.049	0.002	0.004
	Medium	10	7892	892	0.003	0.003	0.002	0.003	0	0.049	0	0.049	0.002	0.004
		20	7011	1773	0.003	0.003	0.002	0.003	0	0.049	0	0.049	0.002	0.004
	Complex	10	7900	884	0.003	0.003	0.002	0.003	0	0.049	0	0.049	0.002	0.004
		20	7024	1760	0.003	0.003	0.002	0.003	0	0.049	0	0.049	0.002	0.004
NO_2	Simple	10	7905	879	0.03	0.028	0.025	0.012	0	0.089	0.003	0.092	0.022	0.037
		20	7019	1765	0.03	0.028	0.024	0.012	0	0.089	0.003	0.092	0.021	0.037
	Medium	10	7901	883	0.03	0.029	0.024	0.011	0	0.083	0.003	0.086	0.022	0.037
		20	7021	1763	0.03	0.028	0.024	0.012	0	0.089	0.003	0.092	0.021	0.037
	Complex	10	7902	882	0.03	0.028	0.027	0.012	0	0.089	0.003	0.092	0.021	0.037
		20	7022	1762	0.03	0.028	0.025	0.012	0	0.088	0.004	0.092	0.021	0.037
O_3	Simple	10	7903	881	0.013	0.006	0	0.015	0	0.101	0	0.101	0.001	0.021
		20	7024	1760	0.013	0.007	0	0.015	0	0.101	0	0.101	0.001	0.021
	Medium	10	7902	882	0.013	0.006	0	0.015	0	0.101	0	0.101	0.001	0.021
		20	7024	1760	0.013	0.006	0	0.015	0	0.101	0	0.101	0.001	0.021
	Complex	10	7899	885	0.013	0.006	0	0.015	0	0.101	0	0.101	0.001	0.021
		20	7023	1761	0.013	0.007	0	0.016	0	0.101	0	0.101	0.001	0.021
CO	Simple	10	7900	884	1.173	1.09	0.5	0.564	0.318	4.89	0	4.89	0.77	1.49
		20	7028	1756	1.157	1.07	0.5	0.558	0.311	4.89	0	4.89	0.76	1.47
	Medium	10	7891	893	1.179	1.09	0.5	0.564	0.318	4.89	0	4.89	0.78	1.49
		20	7026	1758	1.197	1.11	0.93	0.569	0.324	4.35	0	4.35	0.8	1.52
	Complex	10	7902	882	1.171	1.09	0.5	0.566	0.32	4.89	0	4.89	0.77	1.49
		20	7025	1759	1.173	1.09	0.5	0.566	0.32	4.89	0	4.89	0.77	1.5
PM_{10}	Simple	10	7897	887	50	46	38	26	652	333	5	338	33	61
		20	7024	1760	50	46	38	25	631	333	5	338	33	62
	Medium	10	7905	879	50	45	38	26	677	333	5	338	33	61
		20	7025	1759	49	45	38	25	623	295	5	300	33	61
	Complex	10	7899	885	50	46	45	26	674	333	5	338	33	62
		20	7022	1762	49	45	38	24	595	333	5	338	32	61

Where: WS – wind speed, AT – ambient temperature, H – humidity, SO_2 – sulphur dioxide, NO_2 – nitrogen dioxide, O_3 – ozone, CO – carbon monoxide, PM_{10} – particulate matter

Table 4.6: Descriptive statistics of the simulated missing data at Shah Alam

Shah Alam														
		%	Valid N	Missing	Mean	Median	Mode	Standard Deviation	Variance	Range	Min	Max	Percentile 25	Percentile 75
WS	Simple	10	7196	808	5	4.2	1	3.3	11.1	15.8	0.9	16.7	2.2	7.3
		20	6404	1600	5	4.2	1	3.3	10.9	15.8	0.9	16.7	2.2	7.2
	Medium	10	7202	802	5	4.2	1	3.3	10.8	15.8	0.9	16.7	2.2	7.3
		20	6403	1601	5	4.3	1	3.3	10.8	15.8	0.9	16.7	2.2	7.3
	Complex	10	7204	800	5	4.2	1	3.3	10.9	15.8	0.9	16.7	2.2	7.3
		20	6401	1603	5.1	4.2	1	3.3	11.2	15.8	0.9	16.7	2.2	7.4
AT	Simple	10	7197	807	28	26.8	25.1	3.7	14	16.2	21.7	37.9	24.9	31.2
		20	6402	1602	28	26.8	25.1	3.7	14.1	16.4	21.7	38.1	24.9	31.1
	Medium	10	7189	815	28.1	26.8	25.1	3.8	14.3	16.4	21.7	38.1	24.9	31.2
		20	6394	1610	28	26.8	25.3	3.7	14	16.4	21.7	38.1	24.9	31
	Complex	10	7201	803	28	26.7	24.6	3.8	14.1	16.4	21.7	38.1	24.9	31.1
		20	6402	1602	28	26.7	25.1	3.7	13.8	16.2	21.7	37.9	24.9	31
H	Simple	10	7203	801	79	83	95	15	212	63	36	99	67	92
		20	6400	1604	79	83	95	14	208	64	35	99	67	92
	Medium	10	7203	801	79	83	95	15	211	64	35	99	67	92
		20	6402	1602	79	83	95	14	207	63	36	99	67	92
	Complex	10	7198	806	79	83	95	14	207	64	35	99	68	92
		20	6401	1603	79	83	95	14	209	63	36	99	67	92
SO_2	Simple	10	7198	806	0.003	0.002	0.002	0.005	0	0.094	0	0.094	0.001	0.004
		20	6402	1602	0.003	0.002	0.002	0.005	0	0.094	0	0.094	0.001	0.004
	Medium	10	7201	803	0.003	0.002	0.002	0.005	0	0.094	0	0.094	0.001	0.004
		20	6399	1605	0.003	0.002	0.002	0.006	0	0.094	0	0.094	0.001	0.004
	Complex	10	7197	807	0.003	0.002	0.002	0.005	0	0.094	0	0.094	0.001	0.004
		20	6400	1604	0.003	0.002	0.002	0.005	0	0.094	0	0.094	0.001	0.004
NO_2	Simple	10	7202	802	0.019	0.018	0.014	0.01	0	0.068	0	0.068	0.011	0.025
		20	6399	1605	0.019	0.018	0.014	0.01	0	0.068	0	0.068	0.011	0.025
	Medium	10	7201	803	0.019	0.018	0.014	0.01	0	0.068	0	0.068	0.011	0.025
		20	6395	1609	0.019	0.018	0.014	0.011	0	0.068	0	0.068	0.011	0.025
	Complex	10	7204	800	0.019	0.018	0.014	0.01	0	0.068	0	0.068	0.011	0.025
		20	6395	1609	0.019	0.018	0.014	0.011	0	0.068	0	0.068	0.012	0.026
O_3	Simple	10	7195	809	0.019	0.01	0.001	0.021	0	0.119	0	0.119	0.002	0.031
		20	6399	1605	0.018	0.01	0.001	0.02	0	0.109	0	0.109	0.002	0.03
	Medium	10	7203	801	0.019	0.01	0.001	0.021	0	0.119	0	0.119	0.002	0.031
		20	6393	1611	0.019	0.01	0.001	0.021	0	0.109	0	0.109	0.002	0.031
	Complex	10	7201	803	0.018	0.01	0.001	0.021	0	0.119	0	0.119	0.002	0.03
		20	6403	1601	0.019	0.011	0.001	0.021	0	0.109	0	0.109	0.002	0.032
CO	Simple	10	7205	799	0.74	0.65	0.96	0.456	0.208	3.24	0	3.24	0.4	0.98
		20	6401	1603	0.743	0.66	0.96	0.456	0.208	3.24	0	3.24	0.4	0.99
	Medium	10	7199	805	0.75	0.66	0.96	0.467	0.218	3.28	0	3.28	0.4	1
		20	6399	1605	0.743	0.66	0.96	0.455	0.207	3.28	0	3.28	0.4	0.99
	Complex	10	7196	808	0.757	0.67	0.35	0.463	0.215	3.28	0	3.28	0.41	1
		20	6400	1604	0.765	0.68	0.96	0.465	0.216	3.24	0	3.24	0.42	1.01
PM_{10}	Simple	10	7197	807	48	44	37	28	766	267	5	272	30	61
		20	6399	1605	49	44	37	28	786	267	5	272	30	61
	Medium	10	7202	802	48	43	34	28	789	267	5	272	30	61
		20	6400	1604	46	43	37	23	548	190	5	195	29	59
	Complex	10	7195	809	49	44	34	28	790	267	5	272	30	62
		20	6399	1605	48	43	37	28	764	267	5	272	30	61

Where: WS – wind speed, AT – ambient temperature, H – humidity, SO_2 – sulphur dioxide, NO_2 – nitrogen dioxide, O_3 – ozone, CO – carbon monoxide, PM_{10} – particulate matter

Figure 4.4 shows the percentile for complex patterns of simulated missing data (PM_{10}) from Petaling Jaya. From the figure, it can be seen that there are not much differences for values of every percentiles even though the percentage of missing increases. When the structure of simulated missing data change from its original, it would be considered as a different dataset which is not the same as its original. Disturbed structure of simulated missing data may affect the performance of imputation process. This is because this process depends on existing data to estimate the missing values.

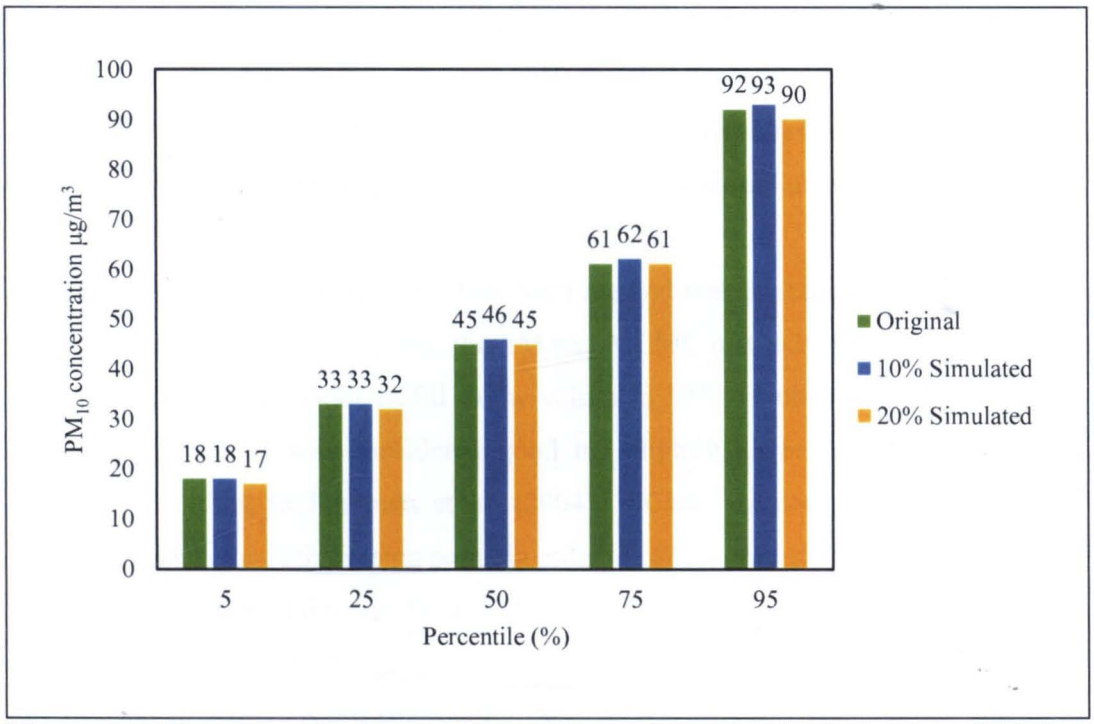


Figure 4.4: The percentile for complex patterns of simulated missing data (PM_{10}) in Petaling Jaya

4.5 The Performance of Imputation Method

In this study, there are two types of performance indicators used i.e. the error measure and the performance measure. The error measures used were Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE), meanwhile the performance measures used were Index of Agreement (d_2) and Prediction Accuracy (PA). Detailed performance indicator results for each imputation methods based on the parameters can be referred in Appendix A.

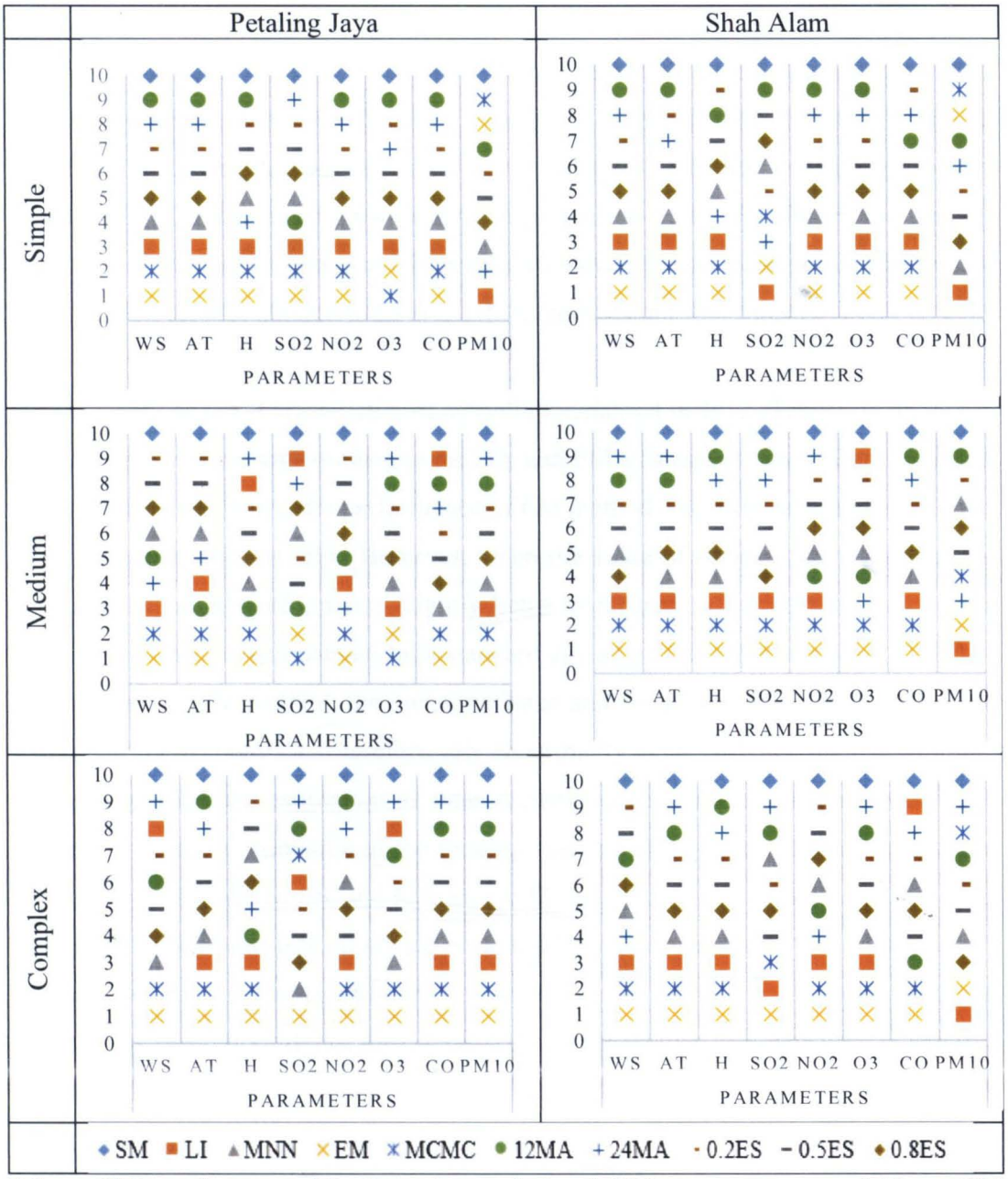
4.5.1 Ten Percent (10%) of Simulated Missing Data

Table 4.7 shows the ranking of all imputation methods for 10% missing data in simple, medium, and complex simulated missing data in Petaling Jaya and Shah Alam. The performances of each of the imputation methods were presented by using Index of Agreement (d_2). Referring to the Table 4.7, most of the parameters agree that Expectation Maximization (EM) method is the best imputation method. This is because EM method was listed as the first in ranking for most of the air quality parameters except for PM_{10}. Therefore, EM method is the most suitable imputation method to replace the 10% simulated missing data. This result is consistent with the others researchers such Zakaria (2018), Razak et al. (2014), Junger and Leon (2015), who also found that EM imputation was the suitable method to impute the missing values in air pollution dataset.

Markov Chain Monte Carlo (MCMC) method was the second most appropriate imputation method. In many cases, the EM and MCMC methods were competing to be the best imputation methods to fill in the values in 10% simulated missing data. This method performance was considered good but slightly lower compared to the EM method. According to Junninen et al. (2004), MCMC methods can create multiple simulated values for each missing point, in order to reflect properly the uncertainty in the missing data. Meanwhile, the Series Mean (SM) method was the worst imputation methods because this methods was ranked as the last for all parameters. Hence, SM method was indicated as not suitable for estimate the missing data in simple, medium, and complex simulated missing pattern. According to Gómez-Carracedo et al. (2014), and Zainuri et al. (2015), the mean imputation underestimates the variance and lead to the error in simulation of missing data.

However, for PM_{10} dataset, Linear Interpolation (LI) method was ranked to have the highest performance in replacing the 10% simple patterns of simulated missing data for Petaling Jaya, and 10% medium and complex patterns for Shah Alam. This is due to the extreme outliers and standard deviation in PM_{10} dataset itself was the largest compared to the other parameters (Table 4.1). The performance of some of EM method in SO_2 was slightly better compared to PM_{10} parameter dataset, although SO_2 dataset contains the largest extreme outliers but the value of standard deviation (variability) in SO_2 was lower compared to PM_{10} data (Table 4.1).

Table 4.7: The ranking of all imputation methods for 10% simulated missing data in Petaling Jaya and Shah Alam



Where: SM – series mean, LI – linear interpolation, MNN – mean nearest neighbour, EM – expectation maximization, MCMC – markov chain monte carlo, MA – moving average, ES – exponential smoothing, WS – wind speed, AT – ambient temperature, H – humidity, SO_2 – sulphur dioxide, NO_2 – nitrogen dioxide, O_3 – ozone, CO – carbon monoxide, PM_{10} – particulate matter

This finding was consistent with the result from other researchers such as Zakaria (2018), which stated that EM method is the best imputation to replace the long gaps of missing data. This was proven when the most of parameters shows that the EM method was in the first rank for 10% - medium patterns of simulated missing data. The gaps of missing data in medium pattern can be considered as long gap of more than 24 hours. Furthermore, EM method performed very well to estimate the missing data in the pattern of complex gaps (length of gaps missing in proportion of 1:1 for simple and medium patterns) for 10% simulated of missing data. However, the performance of EM method was slightly below compared to LI and 24MA methods in PM_{10} dataset.

Based on the characteristics of air pollution dataset in 2012 (Table 4.1), there was a large number of extreme outliers in the SO_2 and PM_{10} dataset for Shah Alam compared to Petaling Jaya. Clearly, the performance of EM method was dropped when the number of extreme outliers was high. However, when the value of standard deviation was too small, EM method performance slightly better compared to the performance of EM method in dataset which contain high standard deviation. According to Moss (2016), in the first step of EM method, the mean, variance and covariance are estimated from the individual completed data. Therefore, any abnormality in the current complete data such outlier may affect the estimation of mean, variance and covariance. As a consequence, the performance of estimation of the missing values by EM method reduced. The other methods such LI, MNN, ES, and MA have different mechanism to estimate the missing values. These methods rely on several complete data before and after the missing data to estimate the missing values and does not depend on the whole complete data in dataset to make any prediction. Hence the performance of EM methods were reduced in presence of outliers.

Table 4.8 shows the average results of performance indicators of the 10% for simple, medium, and complex simulated missing data for Petaling Jaya and Shah Alam respectively. The high performance measures and low value of errors measure indicate that the imputation was the most suitable for estimation of missing values. Based on Table 4.8, EM method was the best imputation method to replace the missing values in 10% for simple, medium, and complex simulated missing data for both locations except for PM_{10} for the reasons discussed earlier, while MCMC method was the second best. This is because, EM method gave the smallest values of RMSE and MAE, and the highest values

of PA and d_2. Series Mean (SM) method was the worst imputation method for estimation of the 10% simulated missing data. This is because, SM methods showed large error and low performance between the predicted and observed values. Overall, the performance of all imputation methods used for 10% simple, medium, and complex simulated missing data from good to worst for both locations were in order of EM; MCMC; LI; MNN; 0.8ES; 0.5ES; 0.2ES; 24MA; 12MA; and SM.

Table 4.8: The average results of performance indicators of the 10% simulated missing data for Petaling Jaya and Shah Alam

Patterns	Methods	Performance Indicators			
		MAE	RMSE	PA	d_2
Simple	EM	2.7161	3.7195	0.7206	0.7988
	MCMC	3.1705	4.3018	0.6429	0.7676
	LI	3.9319	5.3079	0.3873	0.6285
	MNN	4.7496	6.2644	0.2274	0.5246
	12MA	4.9040	6.1369	0.1081	0.4280
	24MA	4.2401	5.3014	0.2740	0.4806
	0.8ES	4.7578	6.2687	0.2237	0.5221
	0.5ES	4.7969	6.2850	0.2091	0.5118
	0.2ES	4.8794	6.2387	0.1523	0.4687
	SM	4.5586	5.7637	0.0000	0.1025
Medium	EM	2.5774	3.2658	0.7098	0.7974
	MCMC	3.0130	3.9090	0.6412	0.7705
	LI	5.0944	6.4377	0.1686	0.4714
	MNN	5.7321	7.1644	0.0898	0.4330
	12MA	5.3869	6.7704	0.0728	0.4089
	24MA	4.7776	5.9515	0.1397	0.4059
	0.8ES	5.7328	7.1645	0.0895	0.4328
	0.5ES	5.7350	7.1654	0.0878	0.4317
	0.2ES	5.7263	7.1421	0.0792	0.4265
	SM	4.2810	5.1570	0.0000	0.1053
Complex	EM	2.6881	3.4816	0.6854	0.7832
	MCMC	3.1225	4.1444	0.6096	0.7513
	LI	4.1946	5.3751	0.2098	0.5073
	MNN	5.1191	6.6129	0.1391	0.4740
	12MA	4.7922	5.9984	0.1294	0.4397
	24MA	4.4677	5.5336	0.1749	0.4269
	0.8ES	5.1205	6.6132	0.1393	0.4743
	0.5ES	5.1298	6.6172	0.1358	0.4720
	0.2ES	5.1605	6.6073	0.1186	0.4593
	SM	4.4132	5.3167	0.0000	0.1609

Where: MAE – mean absolute error, RMSE – root mean squared error, PA – prediction accuracy, d_2 – index of agreement, EM – expectation maximization, MCMC – markov chain monte carlo, LI – linear interpolation, MNN – mean nearest neighbour, MA – moving average, ES – exponential smoothing

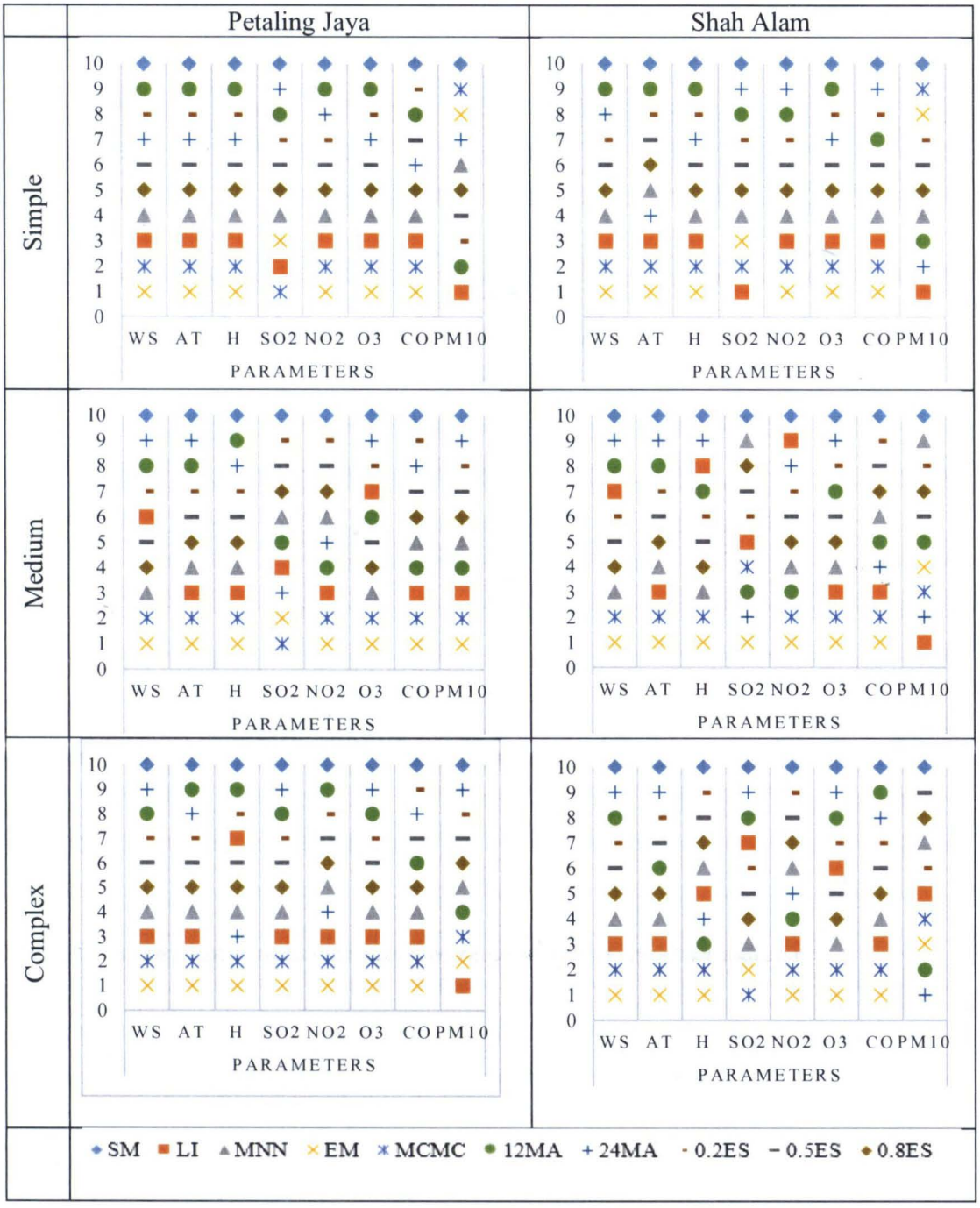
4.5.2 Twenty Percent (20%) of Simulated Missing Data

The performance of all imputation methods for 20% - simple medium, and complex simulated missing data in Petaling Jaya and Shah Alam is shown in Table 4.9. The performances of each of the imputation methods were presented by using Index of Agreement (d_2). Expectation maximization (EM) method was in the first rank for most of the parameters in Petaling Jaya and Shah Alam respectively. For that reason, EM method was chosen the best imputation to fill in the 20% of simulated missing data. Meanwhile, Markov Chain Monte Carlo (MCMC) and Linear Interpolation (LI) methods were in the second and third ranks of the best imputation methods. Series Mean (SM) methods was the worst imputation methods to estimate the missing data in 20% of simple, medium, and complex simulated missing data for both locations because this methods was ranked as the last model for all parameters. Similar to 10% simulated missing data, SM was indicated not suitable for estimation the missing data in simple pattern. Razak et al. (2014), also stated that EM method was superior especially when the percentage of missing values is high. EM method was considered as simple to apply and less time consuming, because it can be easily executed by using SPSS software. Furthermore, this method make a good estimation even though the algorithm of this method are complicated. In this study, EM method shows the best imputation method to fill in the missing values for all parameters except PM_{10} data in Shah Alam and SO_2 data in Petaling Jaya.

MCMC method was also considered as a good imputation method compared to EM method. According to Junninen et al. (2004), MCMC method fill in each of missing data by averaging or pooling multiple simulated value. These process was done by complicated procedures such applying Bayesian inference and repeating several steps such as the imputation I-step and posterior P-step (Schafer, 1997). These complicated procedures would consume time but produce excellent estimation of missing data.

Zakaria (2018), found that MCMC methods was the second best method to impute the missing values especially in long gaps of missing. This method mechanism is quite similar to the EM method which also considers the existing values in dataset to make an estimation of missing values. Therefore, any abnormality or extreme values may reduce the prediction performance of this method.

Table 4.9: The ranking of all imputation methods for 20% simulated missing data in Petaling Jaya and Shah Alam



Where: SM – series mean, LI – linear interpolation, MNN – mean nearest neighbour, EM – expectation maximization, MCMC – markov chain monte carlo, MA – moving average, ES – exponential smoothing, WS – wind speed, AT – ambient temperature, H – humidity, SO_2 – sulphur dioxide, NO_2 – nitrogen dioxide, O_3 – ozone, CO – carbon monoxide, PM_{10} – particulate matter.

Other methods such as LI, MNN, MA, and ES methods show moderate performance. As they are always ranked between 3rd and 8th. This is due to the solving operation process of all this methods which only considers several current complete data to estimate the missing data. MNN method does not cover the uncertainty completely because this method only bring the previous complete data to fill in the missing gaps. Usually upper nearest the value would be selected to replace the missing values (Zakaria, 2018). Meanwhile, LI method show better performance compared to MNN method because the uncertainty were covered by this method. According to Zakaria (2018), LI method fill in the gaps of missing data by replacing the missing value with average value of data before and after missing data in sequential pattern. In this study, MA and ES methods show moderate performance because they were originally used for forecasting analysis which only consider several past values to predict the value in the future. This methods converge the estimation values when the gaps of missing become larger which the uncertainty would not be covered especially in long gaps of missing data.

Table 4.10 shows the performance for each of the imputation method to estimate 20% simple, medium, and complex simulated missing data for Petaling Jaya and Shah Alam were. Referring to the table, the Expectation Maximization (EM) methods shows the best performance followed by Markov Chain Monte Carlo (MCMC) method. This is because almost all of the parameters showed small error values in root mean squared error (RMSE) and mean absolute error (MAE), while the performance values in index of agreement (d_2) and prediction accuracy (PA) were high for both EM and MCMC methods. Series Mean (SM) method was shown to be the worst imputation method for estimation of all patterns in 20% of simulated missing data. This is because these method contributed to large error and small performance compared to the EM and MCMC methods. Overall, the performance of all imputation methods used for 20% simulated missing data from the good to worst was in the order of EM; MCMC; LI; MNN; 0.8ES; 0.5ES; 0.2ES; 12MA; 24MA; and SM.

Table 4.10: The average results of performance indicators of the 20% simulated missing data for Petaling Jaya and Shah Alam

Patterns	Methods	Performance Indicators			
		MAE	RMSE	PA	d_2
Simple	EM	2.8231	3.9357	0.7036	0.7788
	MCMC	3.2791	4.5528	0.6230	0.7553
	LI	3.9971	5.4612	0.3712	0.6161
	MNN	4.8309	6.4419	0.2135	0.5158
	12MA	4.8134	6.0641	0.1209	0.4392
	24MA	4.1941	5.3123	0.2600	0.4742
	0.8ES	4.8376	6.4425	0.2101	0.5132
	0.5ES	4.8706	6.4475	0.1968	0.5037
	0.2ES	4.9060	6.3422	0.1473	0.4666
	SM	4.6899	5.9688	0.0000	0.0710
Medium	EM	3.2020	4.8371	0.6923	0.7744
	MCMC	3.6398	5.3480	0.6104	0.7478
	LI	5.1497	7.0362	0.1689	0.4805
	MNN	5.7637	7.8420	0.0965	0.4334
	12MA	5.4161	7.2905	0.1086	0.4305
	24MA	5.0083	6.6890	0.1697	0.4104
	0.8ES	5.7666	7.8435	0.0960	0.4330
	0.5ES	5.7746	7.8471	0.0937	0.4315
	0.2ES	5.7854	7.8389	0.0850	0.4295
	SM	5.0583	6.8406	0.0000	0.0951
Complex	EM	2.9292	4.1235	0.6929	0.7735
	MCMC	3.4066	4.7440	0.6131	0.7511
	LI	4.8906	6.3658	0.2169	0.5095
	MNN	5.7993	7.7586	0.1302	0.4559
	12MA	5.0164	6.4806	0.1133	0.4301
	24MA	4.6462	5.9436	0.1744	0.4165
	0.8ES	5.8023	7.7590	0.1291	0.4550
	0.5ES	5.8115	7.7607	0.1247	0.4518
	0.2ES	5.8116	7.7249	0.1078	0.4390
	SM	4.7603	6.1934	0.0000	0.1051

Where: MAE – mean absolute error, RMSE – root mean squared error, PA – prediction accuracy, d_2 – index of agreement, EM – expectation maximization, MCMC – markov chain monte carlo, LI – linear interpolation, MNN – mean nearest neighbour, MA – moving average, ES – exponential smoothing

4.6 Summary

Figure 4.5 shows the overall performance and error measures for 10% and 20% - simulated missing data. Generally, the best imputation method to fill in 10% and 20% - simulated missing data was Expectation Maximization (EM) method followed by MCMC method. This was evidenced when almost all performance indicators in each of the percentages agreed that EM method as the most appropriate for replacing the missing values. It seems that the performance of EM method was very good although the variety of the missing gaps in air pollution dataset were different. EM method was shown to have great performance and low error to impute the 10% and 20% of simple, medium, and complex simulated missing data.

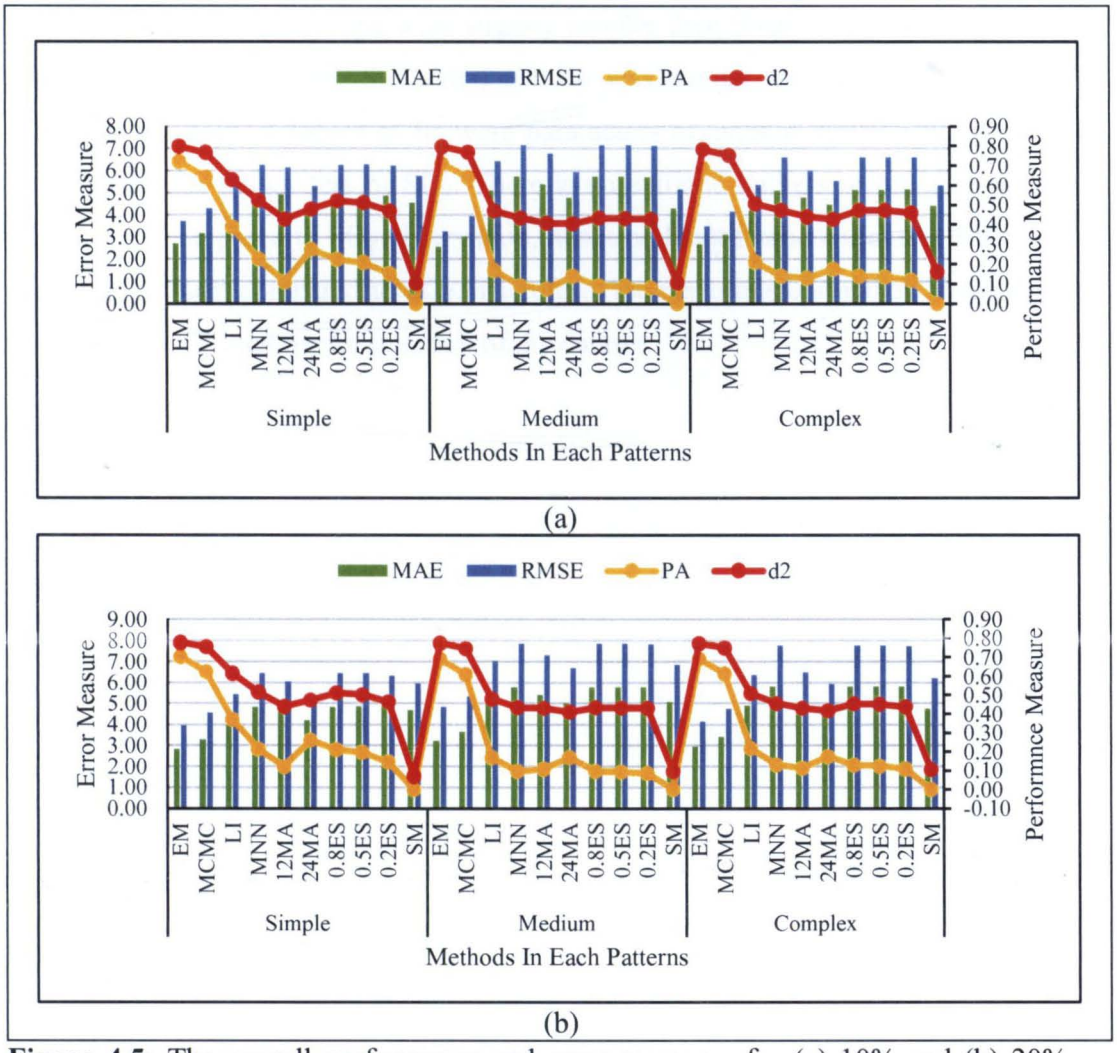


Figure 4.5: The overall performance and error measures for (a) 10% and (b) 20% - simulated missing data

Figures 4.6 and 4.7 show the scatter plot of the observed values and predicted values by using EM (Expectation Maximization) method for all parameters for 10% - complex simulated missing data pattern. Scatter plot are purposely made to evaluate the strength of relationship between predicted (EM estimation) values and observed values (references data). Overall, the values of R^2 for all data in Petaling Jaya and Shah Alam varied. Based on both figures, the values of R^2 for 10% - complex simulated missing data in Shah Alam were slightly larger than Petaling Jaya. Indicating that the predicted values in Shah Alam less deviated from the observed values. The larger value of R^2 indicates strong agreement between the predicted data and the observed data (Noor et al., 2008). In this study, Shah Alam dataset contained lower number of total observation (N = 8004 hours) compared to Petaling Jaya (N = 8784 hours). Based on the results, the prediction was influenced by the sample size of observations of air pollution dataset. Even though, the R^2 values in Petaling Jaya were slightly smaller than Shah Alam, the range for R^2 values for each parameter for both locations was similar. As an example for both locations, the highest values of R^2 were parameters of ambient temperature while the lowest were the parameters of SO_2. There is no extreme outliers recorded in ambient temperature compared to SO_2 observation, hence ambient temperature data distribution are more uniform compared to SO_2. Therefore, the performances and R^2 value of EM method were affected by the structure of the dataset.

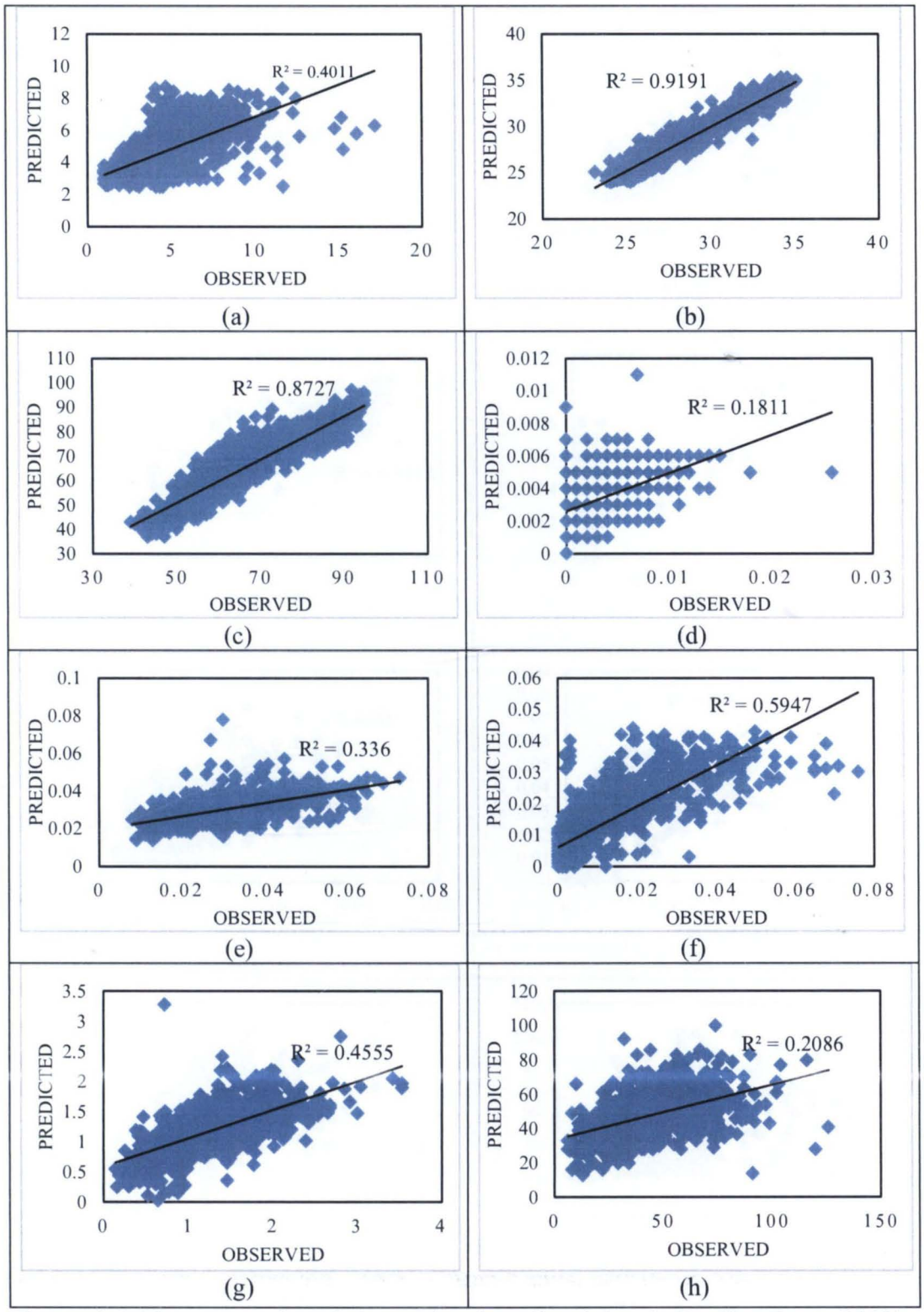


Figure 4.6: The scatter plot of observed and predicted data in Petaling Jaya of 10% - complex simulated missing data for (a) wind speed, (b) ambient temperature, (c) humidity, (d) SO_2, (e) NO_2, (f) O_3, (g) CO, and (h) PM_{10}.

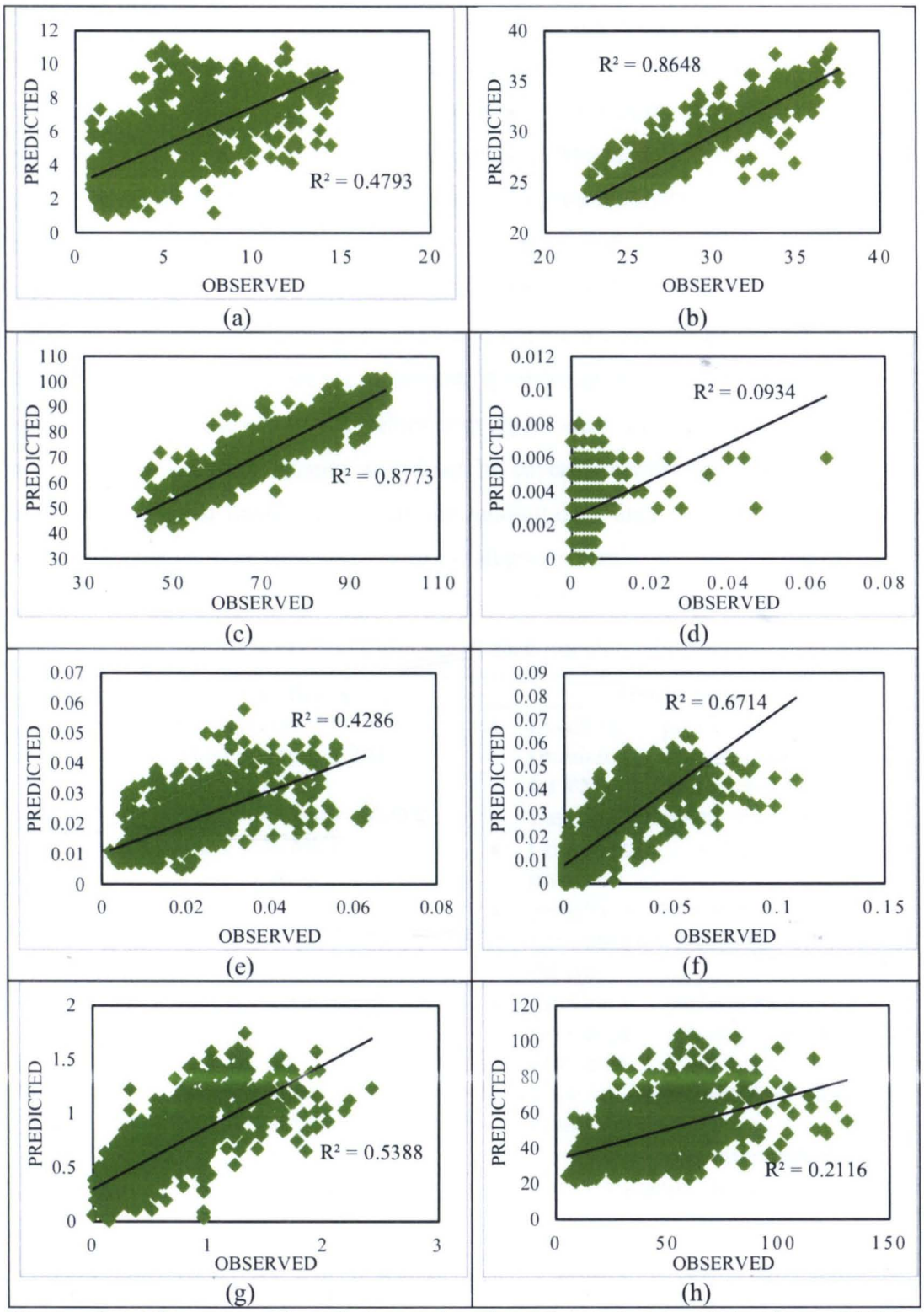


Figure 4.7: The scatter plot of observed and predicted data in Shah Alam of 10% - complex simulated missing data for (a) wind speed, (b) ambient temperature, (c) humidity, (d) SO_2, (e) NO_2, (f) O_3, (g) CO, and (h) PM_{10}.

Table 4.11 shows the summary of each imputation method. The performance of EM methods was superior compared to the other methods to estimate the missing data in various types of missing gaps. MCMC methods was shown as the second best imputation methods competing with EM methods. Observably, EM and MCMC methods considered all current complete data in air pollution dataset for estimation of missing values. Hence, extreme values may reduce the prediction performance of this method. LI method has shown slightly good performance in this study. Other methods such as MNN, MA, and ES methods showed moderate performance to impute the missing values and they also showed inconsistent performance to estimate the missing data in some different patterns of missing gaps. The estimation values converged when the gaps of missing become bigger in which the uncertainty would not be covered. Meanwhile, SM Method was shown as the worst method to impute the missing data since the performance of this method in each patterns gaps was lowest for all parameters.

Table 4.11: The summary of each imputation methods

Tiers	Methods	Description
Good	Expectation Maximization (EM) Markov Chain Monte Carlo (MCMC)	• Excellent performance in all simulated of missing data gaps except for PM_{10} in several cases. • Easy to implement and cost effective. • Complicated solving operation process. • Sensitive to the abnormality in dataset i.e. outliers which reflected by PM_{10} dataset
Moderate	Linear Interpolation (LI) Mean Nearest Neighbour (MNN) Moving Average (MA) Exponential Smoothing (ES)	• Moderate performance in all simulated missing data gaps estimation. • Easy to implement but consuming time. • Easy solving operation process. • The estimation tends to converge when the gaps become large.
Bad	Series Mean (SM)	• Worst performance for all simulated missing data gaps. • Easy to implement and cost effective. • Easy solving operation process. • Lead to bias due to uncertainty are not covered.

CHAPTER 5

CONCLUSION AND RECOMMENDATIONS

5.1 Conclusion

From this study, there are several conclusions that can be obtained. The datasets used in this study were hourly monitoring records of 5 air quality data and 3 meteorological data at 2 different air quality monitoring stations in Klang Valley which are Petaling Jaya and Shah Alam from 2012 to 2016. The 5 air quality data used were sulphur dioxide (SO_2), nitrogen dioxide (NO_2), ozone (O_3), carbon monoxide (CO), particulate matter (PM_{10}), and the 3 meteorological data used were wind speed (WS), ambient temperature (AT), and humidity (H). These dataset were used to study the characteristics of the missing values. Overall, air quality dataset in Shah Alam contains higher missing data compared to in dataset of Petaling Jaya. The highest missing data in Shah Alam was 29.68% (2015) and the lowest missing data was 4.88% (2012), meanwhile Petaling Jaya shows that the highest missing data was 3.14% (2012) and the lowest in 2016 which was 2.11%. In order to choose the reference data, the dataset must be the most complete one and contain the lowest missing data to provide a realistic structure of reference. Dataset in 2012 was selected because the total of missing data in the 2012 for Shah Alam was the lowest. Besides that, the total missing data in Petaling Jaya dataset slightly below than Shah Alam in 2012.

The reference data were simulated based on the simulation design patterns which varies in length of gaps (in hour) of missing observations. There are three patterns of missing used in this study such as simple, medium, and complex patterns. For the simple pattern of simulated missing data, the missing gaps were not more than 24 hours, hence, 1 and 24 hours were set as the minimum and as the maximum gap of missing data respectively. In medium pattern of simulated missing data with gaps more of than 24 hours, 25 and 168 hours (7 days) were set as the minimum and maximum of the gap.

Meanwhile, for the complex pattern of simulated missing data, the simple and medium patterns were blended or mixed with proportion of 1:1 respectively to generate complex simulated missing data, 1 and 168 hours were set as minimum and maximum gap of missing data by limiting the amount of gaps by 50% for the gaps of not more than 24 hours. The distribution of gaps in simple and medium patterns were distributed almost equally. The distribution of gaps which slightly above average in simple patterns was about 27.01% in gaps of 1 to 6 hours, while in medium patterns about 19.45% gaps distributed in 120 to 144 hours. In complex pattern, the percentage of missing gaps distributed more than 24 hours were equal to 50.24% and 49.76% for the gaps of not more than 24 hours. These data were simulated into 10% and 20% of missing. In this study, the main structure and characteristics of the data does not disturbed by the simulation process because the descriptive statistics before and after simulation process does not change too much.

In order to fill in the missing values in simulated missing data, ten imputation methods were applied. The ten imputation methods used in this study were Series Mean (SM), Linear Interpolation (LI), Mean Nearest Neighbour (MNN), Expectation Maximization (EM), Markov Chain Monte Carlo (MCMC), 12 – hours and 24 – hours Moving Average (MA), and Exponential Smoothing (ES) (0.2, 0.5, and 0.8). The goodness of fit of all these imputation methods was described by using four performance indicators such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Prediction Accuracy (PA), and Index of Agreement (d_2). Generally, EM method was selected as the best imputation method to fill in the simulated short, long, and combination of short-long gaps of missing data in air quality monitoring dataset compared to the other methods. In the ranking model, EM method was mainly placed in the first rank for most of parameters although the percentages of missing and type of missing gaps was varied except for PM_{10}. This is because, the EM method show high value of performance and low value of error in the datasets which contained various gaps and percentages of missing observations. On the other hand, PM_{10} dataset has high outliers which affect the performance of EM method in estimation of its missing data.

5.2 Recommendations

This study has discovered the importance of statistical analysis especially in environmental data. The aim of this study was to find the best imputation method to fill in the different types of simulated missing gaps in air quality monitoring dataset. This study can be improved in the future by:

i. Temporarily replace the outliers and extreme outliers before the simulation and imputation process with suitable values to enhance the EM and MCMC methods performance because abnormality present in dataset would affect the performance of EM and MCMC methods.

ii. Using the different distribution of missing gaps in each simulated missing gap patterns such as Uniform distribution, Normal distribution, and Poisson distribution.

iii. Study the modified proportion of simple and medium simulated missing gap patterns in complex simulated missing gaps pattern by (3:1) , (1:3) and (1:1).

iv. Apply the missing mechanism such as Missing at Random (MAR), Missing Not at Random (MNAR), and Missing Completely at Random (MAR) mechanisms in simulation of missing data by adjusting the randomness of missing data.

v. Use 5 new patterns of simulated missing gap such as simple, medium, large, extreme, and complex pattern to reflect the real situation and variety of missing data in Malaysia.

REFERENCES

Abdullah, A. M. (2012). An Overview of the Air Pollution Trend in Klang Valley, Malaysia. *Open Environmental Sciences, 6*(1), 13-19.

Akram, M., Hyndman, R. J., & Ord, J. K. (2009). Exponential Smoothing And Non-Negative Data. *Australian & New Zealand Journal of Statistics, 51*(4), 415-432.

Allison, P.D. 2001. *Missing Data*. Sage Publications, Inc.

Almselati, A. S., Rahmat, R. A., & Jaafar, O. (2011). An Overview of Urban Transport in Malaysia. *The Social Sciences, 6*(1), 24-33.

Altın, C., & Er, O. (2016). Comparison of Different Time and Frequency Domain Feature Extraction Methods on Elbow Gesture's EMG. *European Journal of Interdisciplinary Studies, 5*(1), 35-44.

Ambient air pollution: Pollutants (2017). Retrieved November 2, 2018 from https://www.who.int/airpollution/ambient/pollutants/en/

Azid, A., Juahir, H., Toriman, M. E., Kamarudin, M. K., Saudi, A. S., Hasnam, C. N., Yamin, M. (2014). Prediction of the Level of Air Pollution Using Principal Component Analysis and Artificial Neural Network Techniques: A Case Study in Malaysia. *Water, Air, & Soil Pollution, 225*(8).

Azid, A., Juahir, H., Toriman, M. E., Endut, A., Kamarudin, M. A., Rahman, M. A., Yunus, K. (2014, December 1). *Source Apportionment of Air Pollution: A Case Study In Malaysia* [Scholarly project]. In *Jurnal Teknologi*. Retrieved March 4, 2019, from file:///C:/Users/Acer/Downloads/2934-10410-1-PB.pdf

Azmi, S. Z., Latif, M. T., Ismail, A. S., Juneng, L., & Jemain, A. A. (2009). Trend and status of air quality at three different monitoring stations in the Klang Valley, Malaysia. *Air Quality, Atmosphere & Health, 3*(1), 53-64.

Brown, M. A., Li, Y., Massetti, E., & Lapsa, M. (2017). U.S. sulfur dioxide emission reductions: Shifting factors and a carbon dioxide penalty. *The Electricity Journal, 30*(1), 17-24.

Bruffaerts, C., Verardi, V., & Vermandele, C. (2014). A generalized boxplot for skewed and heavy-tailed distributions. *Statistics & Probability Letters, 95*, 110-117.

Carbon monoxide. (2019). Retrieved February 5, 2019 from https://pubchem.ncbi.nlm.nih.gov/compound/carbon_monoxide

Carbon Monoxide - MeSH - NCBI. (2019). Retrieved February 5, 2019 from https://www.ncbi.nlm.nih.gov/mesh/68002248

Casquero-Vera, J., Lyamani, H., Titos, G., Borrás, E., Olmo, F., & Alados-Arboledas, L. (2019). Impact of primary NO2 emissions at different urban sites exceeding the European NO2 standard limit. *Science of The Total Environment, 646*, 1117-1125.

Cheng, Z., Wang, S., Jiang, J., Fu, Q., Chen, C., Xu, B., Yu, J., Fu, X., Hao, J. (2013). Long-term trend of haze pollution and impact of particulate matter in the Yangtze River Delta, China. *Environmental Pollution, 182*, 101-110.

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geoscientific Model Development, 7*(3), 1247-1250.

Çokluk, Ö, & Kayri, M. (2011). The Effects of Methods of Imputation for Missing Values on the Validity and Reliability of Scales. *Educational Sciences: Theory & Practice, 11*(1), 303-309.

Colvile, R., Hutchinson, E., & Warren, R. (2002). Chapter 6 The transport sector as a source of air pollution. *Air Pollution Science for the 21st Century Developments in Environmental Science,* 187-239.

Department of Energy. (2013). Role of Modelling and Simulation in Scientific Discovery. Retrieved March 4, 2019, from https://www.energy.gov/ne/articles/role-modeling-and-simulation-scientific-discovery

Department of Environment. (2012). *Malaysia Environmental Quality Report 2012.* Department of Environment, Ministry of Natural Resources and Environment Malaysia. Kuala Lumpur.

Department of Environment. (2013). *Malaysia Environmental Quality Report 2013.* Department of Environment, Ministry of Natural Resources and Environment Malaysia. Kuala Lumpur.

Department of Environment. (2015). *Malaysia Environmental Quality Report 2015.* Department of Environment, Ministry of Natural Resources and Environment Malaysia. Kuala Lumpur.

Department of Environment. (2016). *Malaysia Environmental Quality Report 2016.* Department of Environment, Ministry of Natural Resources and Environment Malaysia. Kuala Lumpur.

Department of Environment. (2017). *Malaysia Environmental Quality Report 2017.* Department of Environment, Ministry of Natural Resources and Environment Malaysia. Kuala Lumpur.

Dominick, D., Juahir, H., Latif, M. T., Zain, S. M., & Aris, A. Z. (2012). Spatial assessment of air quality patterns in Malaysia using multivariate analysis. *Atmospheric Environment, 60*, 172-181.

Garciarena, U., & Santana, R. (2017). An extensive analysis of the interaction between missing data types, imputation methods, and supervised classifiers. *Expert Systems with Applications, 89*, 52-65.

Glen, S. (2017). Exponential Smoothing: Definition of Simple, Double and Triple. Retrieved September 28, 2018, from https://www.statisticshowto.datasciencecentral.com/exponential-smoothing/

Gómez-Carracedo, M., Andrade, J., López-Mahía, P., Muniategui, S., & Prada, D. (2014). A practical comparison of single and multiple imputation methods to handle complex missing data in air quality datasets. *Chemometrics and Intelligent Laboratory Systems*, 134, 23-33.

Halim, N. D., Latif, M. T., Ahamad, F., Dominick, D., Chung, J. X., Juneng, L., & Khan, M. F. (2018). The long-term assessment of air quality on an island in Malaysia. *Heliyon, 4*(12).

Hoffmann, R (1981). Box Plot: Display of Distribution. Retrieved October 11, 2018 from http://www.physics.csbsju.edu/stats/box2.html

Jamal, H. H., Pillay, M. S., Zailina, H., Shamsul, B. S., Sinha, K., Zaman Huri, Z., Khew, S. L., Mazrura, S., Ambu, S., Rasimah, A., & Ruzita, M. S. (2004). *A Study of Health Impact and Risk Assessment of Urban Air Pollution in the Klang Valley, Malaysia.* UKM Pakarunding Sdn. Bhd, Kuala Lumpur.

Junger, W., & Leon, A. P. (2015). Imputation of missing data in time series for air pollutants. *Atmospheric Environment, 102*, 96-104.

Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J., & Kolehmainen, M. (2004). Methods for imputation of missing values in air quality data sets. *Atmospheric Environment, 38*(18), 2895-2907.

Kaiser, J. (2014). Dealing with Missing Values in Data. *Journal of Systems Integration,* 42-51. doi:10.20470/jsi.v5i1.178

Kampa, M., & Castanas, E. (2008). *Human health effects of air pollution. Environmental Pollution*, 151(2), 362-367.

Kanagendran, A., Pazouki, L., Bichele, R., Külheim, C., & Niinemets, Ü. (2018). Temporal regulation of terpene synthase gene expression in Eucalyptus globulus leaves upon ozone and wounding stresses: Relationships with stomatal ozone uptake and emission responses. *Environmental and Experimental Botany, 155*, 552-565.

Kenton, W. (2018, December 13). Descriptive Statistics. Retrieved October 5, 2018 from https://www.investopedia.com/terms/d/descriptive_statistics.asp

Khallaf, M. K. (2011). *The Impact of Air Pollution on Health, Economy, Environment and Agricultural Sources*. InTech.

Legates, D. R., & Mccabe, G. J. (2012). A refined index of model performance: A rejoinder. *International Journal of Climatology, 33*(4), 1053-1056.

Li, J. (2016). Assessing spatial predictive models in the environmental sciences: Accuracy measures, data variation and variance explained. *Environmental Modelling & Software, 80*, 1-8.

Little, R.J.A. & Rubin, D.B. 2002. *Statistical Analysis with Missing Data*. 2nd ed. New York: Wiley.

Liu, Y., & Brown, S. D. (2013). Comparison of five iterative imputation methods for multivariate classification. *Chemometrics and Intelligent Laboratory Systems, 120*, 106-115.

Luo, P., Zhang, M., Liu, Y., Han, D., & Li, Q. (2012, July). A moving average filter based method of performance improvement for ultraviolet communication system. *In 2012 8th International Symposium on Communication Systems, Networks & Digital Signal Processing (CSNDSP)* (pp. 1-4). IEEE.

Mabahwi, N. A., Leh, O. L., & Omar, D. (2015). Urban Air Quality and Human Health Effects in Selangor, Malaysia. *Procedia - Social and Behavioral Sciences, 170*, 282-291.

Mahmud, M. (2013). Assessment of atmospheric impacts of biomass open burning in Kalimantan, Borneo during 2004. *Atmospheric Environment, 78*, 242-249.

Mcginniss, J., & Harel, O. (2016). Multiple imputation in three or more stages. *Journal of Statistical Planning and Inference, 176*, 33-51.

Md Yusof, N. F. F., Ramli, N. A., Yahaya, A. S., Sansuddin, N., Ghazali, N. A., & Al Madhoun, W. (2010). Monsoonal differences and probability distribution of PM10 concentration. *Environmental Monitoring and Assessment, 163(1-4)*, 655-667.

Mohammed, G., Karani, G., & Mitchell, D. (2017). Trace Elemental Composition in PM10 and PM2.5 Collected in Cardiff, Wales. *Energy Procedia, 111*, 540-547.

Moshenberg, S., Lerner, U., & Fishbain, B. (2015). Spectral methods for imputation of missing air quality data. *Environmental Systems Research, 4*(1).

Moss, S. (2016). Expectation maximization--to manage missing data. Retrieved September 28, 2018, from https://www.sicotests.com/psyarticle.asp?id=267

Nevers, N. D. (2017). *Air pollution control engineering*. Long Grove, IL: Waveland Press.

Nitrogen dioxide. (2019). Retrieved January 22, 2019 from https://pubchem.ncbi.nlm.nih.gov/compound/nitrogen_dioxide

Noor, N. M., Yahaya, A., Ramli, N., & Abdullah, M. M. (2015). Filling the Missing Data of Air Pollutant Concentration Using Single Imputation Methods. *Applied Mechanics and Materials, 754-755*, 923-932.

Noor, N.M. (2006). *The replacement of missing values of continuous air pollution monitoring data using various imputation technique.* Universiti Sains Malaysia, Penang.

Noor, N. M., Ahmad Shukri, Y., Nor Azam, R., & Mohd Mustafa Al Bakri, A. (2008). Estimation of missing values in air pollution data using single imputation techniques. *Science Asia, 34*, 341-345.

Noor, N. M., Yahya, A. S., Ramli, N. A., Yusof, N. F. F. M., & Abdullah, M. M. A. (2013). Roles of Imputation Methods for Filling the Missing Values: A Review. Advances in Environmental Biology, 7(12), 3861-3869.

Norela, S., Saidah, M., & Mahmud, M. (2013). Chemical composition of the haze in Malaysia 2005. *Atmospheric Environment, 77*, 1005-1010. doi:10.1016/j.atmosenv.2013.05.024

Othman, J., Sahani, M., Mahmud, M., & Ahmad, M. K. (2014). Transboundary smoke haze pollution in Malaysia: Inpatient health impacts and economic valuation. *Environmental Pollution, 189*, 194-201.

Ozone. (2018). Retrieved January 22, 2018 from https://pubchem.ncbi.nlm.nih.gov/compound/ozone

Plaia, A., & Bondi, A. (2006). Single imputation method of missing values in environmental pollution data sets. *Atmospheric Environment, 40*(38), 7316-7330.

Poeschl, U. (2006). Atmospheric Aerosols: Composition, Transformation, Climate and Health Effects. *ChemInform, 37*(7).

Quinteros, M. E., Lu, S., Blazquez, C., Cárdenas-R, J. P., Ossa, X., Delgado-Saborit, J., Harrison, R., Ruiz-Rudolph, P. (2019). Use of data imputation tools to reconstruct incomplete air quality datasets: A case-study in Temuco, Chile. *Atmospheric Environment, 200*, 40-49.

Rahman, S. A., Ismail, S. S., Raml, M. F., Latif, M. T., Abidin, E. Z., & Praveena, S. M. (2015). The Assessment of Ambient Air Pollution Trend in Klang Valley, Malaysia. *World Environment, 5*(1), 1-11.

Razak, N. A., Zubairi, Y. Z., & Yunus, R. M. (2014). Imputing missing values in modelling the PM10 concentrations. *Sains Malaysiana*, 43(10), 1599-1607.

Ryerson, T. B. (2003). Effect of petrochemical industrial emissions of reactive alkenes and NOxon tropospheric ozone formation in Houston, Texas. *Journal of Geophysical Research, 08*(8).

Schafer, J. L. (1997). *Analysis of incomplete multivariate data. Monpgraphs on statistics and applied probability.* (C. and Hall, Ed.). London.

Sicard, P., Anav, A., Marco, A. D., & Paoletti, E. (2017). Projected global tropospheric ozone impacts on vegetation under different emission and climate scenarios. *Atmospheric Chemistry and Physics Discussions,* 1-34.

Siegel, A.F. (2011). *Practical business statistics* (Sixth edit). USA: Academic press.

Sulfur dioxide. (2019). Retrieved February 11, 2019 from https://pubchem.ncbi.nlm.nih.gov/compound/sulfur_dioxide

Wang, C., Zhao, L., Sun, W., Xue, J., & Xie, Y. (2018). Identifying redundant monitoring stations in an air quality monitoring network. *Atmospheric Environment, 190*, 256-268.

Wang, Z., Dong, K., Tian, L., Zhan, S., Wang, X., Wang, J., & Tu, J. (2018). Numerical analyses of sulfur dioxide transport by an atmospheric circulating drop. *Atmospheric Pollution Research.*

William M.K. Trochim (2006) Descriptive Statistics. Retrieved December 25, 2018 from http://www.socialresearchmethods.net/kb/statdesc.htm

Yuan, Y. (2005). Multiple Imputation for Missing Data: Concepts and New Development. *25*, 267. Retrieved September 29, 2018, from https://stats.idre.ucla.edu/wp-content/uploads/2016/02/multipleimputation.pdf.

Zainuri, N. A., Jemain, A. A., & Muda, N. (2015). A Comparison of Various Imputation Methods for Missing Values in Air Quality Data. *Sains Malaysiana*, 44(3), 449-456.

Zakaria, N. A. (2018). *Imputation Methods for Filling the Long Interval of Missing Observations in Air Pollution Data and Meteorological Dataset.* Universiti Malaysia Perlis.

Zhang, Q., Shen, Z., Cao, J., Zhang, R., Zhang, L., Huang, R., Zheng, C., Wang, L., Liu, S., Xu, H., Zheng, C., Liu, P. (2015). Variations in PM2.5, TSP, BC, and trace gases (NO2, SO2, and O3) between haze and non-haze episodes in winter over Xian, China. *Atmospheric Environment, 112*, 64-71. doi:10.1016/j.atmosenv.2015.04.03

APPENDIX

The results of performance indicators of the (a) 10% - simple, (b) 20% - simple, (c) 10% - medium, (d) 20% - medium, (e) 10% - complex, and (f) 20% - complex simulated missing data of Petaling Jaya and Shah Alam

METHODS	P.I	SIMPLE (10%)															
		WS		AT		H		SO_2		NO_2		O_3		CO		PM_{10}	
		PJ	SA	PJ	SA	PJ	SA	PJ	SA	PJ	SA	PJ	SA	PJ	SA	PJ	SA
SM	MAE	1.85	2.67	2.40	3.22	12.71	12.26	0.00	0.00	0.01	0.01	0.01	0.02	0.43	0.38	17.21	19.76
	RMSE	2.34	3.18	2.78	3.80	14.70	14.09	0.00	0.00	0.01	0.01	0.02	0.02	0.57	0.51	24.25	25.91
	PA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	d_2	0.05	0.01	0.09	0.04	0.18	0.19	0.09	0.08	0.19	0.14	0.08	0.12	0.01	0.17	0.09	0.12
LI	MAE	1.87	2.55	2.27	3.37	11.74	11.84	0.00	0.00	0.01	0.01	0.01	0.02	0.45	0.30	13.65	14.82
	RMSE	2.43	3.43	3.06	4.39	15.63	15.57	0.00	0.00	0.01	0.01	0.02	0.02	0.59	0.46	18.92	20.37
	PA	0.33	0.36	0.32	0.26	0.38	0.32	0.15	0.49	0.50	0.42	0.25	0.25	0.36	0.49	0.67	0.65
	d_2	0.59	0.64	0.62	0.58	0.64	0.60	0.35	0.67	0.71	0.66	0.55	0.56	0.61	0.68	0.81	0.80
MNN	MAE	2.17	3.11	2.83	4.36	14.87	13.41	0.00	0.00	0.01	0.01	0.01	0.02	0.54	0.39	16.70	17.55
	RMSE	2.73	4.10	3.67	5.56	19.21	17.26	0.00	0.00	0.01	0.01	0.02	0.03	0.73	0.55	23.33	23.00
	PA	0.18	0.17	0.19	0.04	0.17	0.19	0.06	0.18	0.35	0.31	0.06	0.05	0.23	0.35	0.59	0.52
	d_2	0.50	0.51	0.54	0.46	0.53	0.50	0.29	0.36	0.62	0.60	0.44	0.43	0.53	0.60	0.77	0.72
EM	MAE	1.40	1.66	0.69	0.84	3.86	3.97	0.00	0.00	0.01	0.01	0.01	0.01	0.31	0.21	14.03	16.46
	RMSE	1.83	2.20	0.92	1.15	4.96	5.36	0.00	0.00	0.01	0.01	0.01	0.01	0.42	0.29	20.20	22.12
	PA	0.62	0.72	0.94	0.95	0.94	0.92	0.39	0.33	0.74	0.79	0.79	0.83	0.67	0.82	0.55	0.52
	d_2	0.74	0.83	0.97	0.97	0.97	0.96	0.45	0.47	0.81	0.86	0.86	0.90	0.79	0.88	0.67	0.66
MCMC	MAE	1.58	1.83	0.78	0.99	4.31	4.40	0.00	0.00	0.01	0.01	0.01	0.01	0.36	0.25	16.73	19.45
	RMSE	2.05	2.43	1.02	1.35	5.65	6.09	0.00	0.00	0.01	0.01	0.01	0.01	0.49	0.34	23.48	25.88
	PA	0.55	0.67	0.93	0.94	0.92	0.90	0.23	0.19	0.64	0.70	0.77	0.74	0.58	0.75	0.39	0.38
	d_2	0.73	0.81	0.96	0.97	0.96	0.95	0.41	0.39	0.78	0.82	0.87	0.86	0.75	0.85	0.59	0.60
12 MA	MAE	2.20	3.08	3.16	4.41	15.37	15.13	0.00	0.00	0.01	0.01	0.01	0.02	0.47	0.37	16.30	17.92
	RMSE	2.71	3.78	3.68	5.15	17.99	17.71	0.00	0.00	0.01	0.01	0.02	0.03	0.60	0.49	22.62	23.39
	PA	-0.03	-0.11	-0.02	-0.18	0.12	0.06	0.10	0.14	0.19	0.18	-0.14	-0.25	0.20	0.38	0.59	0.49
	d_2	0.32	0.31	0.39	0.29	0.47	0.44	0.31	0.34	0.49	0.46	0.30	0.22	0.47	0.58	0.77	0.69
24 MA	MAE	1.83	2.62	2.44	3.34	11.92	11.88	0.00	0.00	0.01	0.01	0.01	0.02	0.43	0.35	16.07	16.90
	RMSE	2.36	3.22	2.87	3.92	13.79	13.85	0.00	0.00	0.01	0.01	0.02	0.02	0.56	0.45	22.02	21.72
	PA	0.17	0.16	0.18	0.14	0.41	0.34	0.13	0.25	0.36	0.26	0.13	-0.01	0.27	0.45	0.61	0.55
	d_2	0.38	0.37	0.46	0.40	0.59	0.56	0.28	0.44	0.57	0.48	0.34	0.28	0.48	0.58	0.78	0.70
0.2 ES	MAE	2.21	3.14	3.03	4.54	15.66	14.48	0.00	0.00	0.01	0.01	0.02	0.02	0.52	0.38	16.54	17.50
	RMSE	2.73	3.99	3.71	5.52	19.14	17.55	0.00	0.00	0.01	0.01	0.02	0.03	0.68	0.53	23.05	22.83
	PA	0.07	0.05	0.08	-0.09	0.10	0.08	0.05	0.18	0.28	0.23	-0.10	-0.11	0.19	0.34	0.59	0.52
	d_2	0.41	0.42	0.46	0.36	0.47	0.42	0.29	0.37	0.57	0.54	0.33	0.31	0.50	0.58	0.77	0.71
0.5 ES	MAE	2.19	3.12	2.89	4.44	15.13	13.77	0.00	0.00	0.01	0.01	0.01	0.02	0.54	0.39	16.70	17.53
	RMSE	2.74	4.09	3.70	5.59	19.31	17.43	0.00	0.00	0.01	0.01	0.02	0.03	0.72	0.55	23.35	22.99
	PA	0.15	0.15	0.17	0.02	0.15	0.16	0.05	0.17	0.33	0.29	0.02	0.02	0.22	0.34	0.59	0.52
	d_2	0.48	0.50	0.52	0.44	0.52	0.49	0.29	0.36	0.61	0.58	0.42	0.40	0.52	0.59	0.77	0.71
0.8 ES	MAE	2.17	3.12	2.84	4.37	14.91	13.48	0.00	0.00	0.01	0.01	0.01	0.02	0.54	0.39	16.70	17.54
	RMSE	2.73	4.10	3.68	5.57	19.23	17.29	0.00	0.00	0.01	0.01	0.02	0.03	0.73	0.55	23.35	22.99
	PA	0.17	0.16	0.18	0.04	0.16	0.18	0.06	0.17	0.35	0.31	0.05	0.05	0.23	0.35	0.59	0.52
	d_2	0.50	0.51	0.53	0.45	0.52	0.50	0.29	0.36	0.62	0.60	0.44	0.42	0.53	0.60	0.77	0.72

(a)

METHODS	P.I	SIMPLE (20%)															
		WS		AT		H		SO_2		NO_2		O_3		CO		PM_{10}	
		PJ	SA	PJ	SA	PJ	SA	PJ	SA	PJ	SA	PJ	SA	PJ	SA	PJ	SA
SM	MAE	1.75	2.81	2.42	3.26	13.16	12.73	0.00	0.00	0.01	0.01	0.01	0.02	0.45	0.38	18.55	19.49
	RMSE	2.23	3.39	2.84	3.76	15.43	14.91	0.00	0.01	0.01	0.01	0.02	0.02	0.59	0.48	26.51	25.28
	PA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	d_2	0.12	0.07	0.03	0.00	0.10	0.10	0.08	0.04	0.01	0.09	0.02	0.15	0.17	0.05	0.06	0.03
LI	MAE	2.04	3.04	2.32	3.07	10.66	11.92	0.00	0.00	0.01	0.01	0.01	0.02	0.45	0.27	15.15	14.97
	RMSE	2.69	3.94	3.03	4.19	14.39	15.85	0.00	0.01	0.01	0.01	0.02	0.02	0.60	0.37	21.69	20.57
	PA	0.14	0.23	0.35	0.24	0.49	0.35	0.27	0.28	0.35	0.55	0.18	0.21	0.39	0.67	0.63	0.62
	d_2	0.48	0.54	0.62	0.56	0.71	0.63	0.50	0.42	0.61	0.73	0.50	0.53	0.63	0.81	0.80	0.78
MNN	MAE	2.39	3.62	2.98	3.87	13.70	14.60	0.00	0.00	0.01	0.01	0.02	0.02	0.57	0.37	17.23	17.90
	RMSE	3.16	4.69	3.81	4.98	17.75	19.10	0.00	0.01	0.01	0.01	0.02	0.03	0.73	0.52	24.04	24.20
	PA	0.04	0.07	0.11	0.01	0.31	0.14	0.15	0.17	0.24	0.35	0.03	0.02	0.29	0.48	0.52	0.49
	d_2	0.40	0.45	0.49	0.43	0.61	0.52	0.36	0.30	0.55	0.61	0.41	0.43	0.57	0.70	0.72	0.70
EM	MAE	1.27	1.79	0.84	0.87	5.05	3.47	0.00	0.00	0.01	0.01	0.01	0.01	0.32	0.23	14.73	16.57
	RMSE	1.69	2.37	1.21	1.22	6.87	4.98	0.00	0.01	0.01	0.01	0.01	0.01	0.43	0.31	21.68	22.17
	PA	0.65	0.72	0.90	0.95	0.90	0.94	0.39	0.32	0.64	0.75	0.77	0.81	0.68	0.77	0.58	0.49
	d_2	0.77	0.82	0.95	0.97	0.94	0.97	0.49	0.33	0.76	0.81	0.84	0.87	0.78	0.86	0.66	0.63
MCMC	MAE	1.55	1.98	0.93	1.00	5.57	4.09	0.00	0.00	0.01	0.01	0.01	0.01	0.37	0.27	17.56	19.12
	RMSE	2.00	2.64	1.34	1.44	7.62	5.81	0.00	0.01	0.01	0.01	0.01	0.01	0.50	0.36	24.99	26.09
	PA	0.53	0.65	0.88	0.93	0.87	0.92	0.31	0.17	0.54	0.67	0.73	0.74	0.56	0.69	0.42	0.35
	d_2	0.72	0.80	0.94	0.96	0.93	0.96	0.51	0.33	0.72	0.80	0.84	0.85	0.73	0.82	0.61	0.59
12 MA	MAE	2.22	3.62	2.93	4.14	14.77	15.70	0.00	0.00	0.01	0.01	0.02	0.02	0.47	0.35	15.82	16.92
	RMSE	2.77	4.37	3.48	4.94	17.66	18.69	0.00	0.01	0.01	0.01	0.02	0.03	0.61	0.47	21.69	22.27
	PA	-0.11	-0.16	-0.05	-0.17	0.17	-0.02	0.08	0.14	0.10	0.31	-0.11	-0.13	0.29	0.47	0.58	0.53
	d_2	0.30	0.28	0.37	0.31	0.51	0.38	0.30	0.27	0.44	0.56	0.31	0.32	0.55	0.68	0.74	0.71
24 MA	MAE	1.83	2.97	2.41	3.13	12.08	12.70	0.00	0.00	0.01	0.01	0.01	0.02	0.43	0.34	15.16	15.99
	RMSE	2.31	3.59	2.90	3.79	14.46	14.97	0.00	0.01	0.01	0.01	0.02	0.02	0.57	0.44	20.81	21.08
	PA	0.11	0.06	0.18	0.20	0.39	0.22	0.07	0.14	0.23	0.33	0.12	0.11	0.35	0.48	0.62	0.56
	d_2	0.36	0.35	0.44	0.43	0.59	0.46	0.29	0.27	0.49	0.54	0.34	0.36	0.56	0.66	0.72	0.71
0.2 ES	MAE	2.38	3.68	3.10	4.12	14.55	15.40	0.00	0.00	0.01	0.01	0.02	0.02	0.54	0.37	16.68	17.60
	RMSE	3.09	4.62	3.79	5.03	17.97	19.13	0.00	0.01	0.01	0.01	0.02	0.03	0.70	0.51	23.08	23.47
	PA	-0.06	-0.04	0.00	-0.12	0.22	0.05	0.10	0.15	0.17	0.31	-0.07	-0.10	0.26	0.45	0.54	0.49
	d_2	0.33	0.37	0.42	0.34	0.55	0.46	0.32	0.28	0.50	0.57	0.34	0.36	0.55	0.68	0.73	0.69
0.5 ES	MAE	2.39	3.65	3.02	3.95	14.00	14.88	0.00	0.00	0.01	0.01	0.02	0.02	0.57	0.37	17.18	17.84
	RMSE	3.15	4.70	3.83	5.02	17.91	19.22	0.00	0.01	0.01	0.01	0.02	0.03	0.73	0.52	23.90	24.08
	PA	0.02	0.05	0.09	-0.02	0.29	0.12	0.12	0.16	0.22	0.34	0.01	-0.01	0.28	0.47	0.52	0.49
	d_2	0.39	0.43	0.48	0.41	0.59	0.51	0.34	0.30	0.54	0.60	0.39	0.41	0.56	0.69	0.72	0.69
0.8 ES	MAE	2.39	3.63	2.99	3.88	13.76	14.66	0.00	0.00	0.01	0.01	0.02	0.02	0.57	0.37	17.22	17.87
	RMSE	3.15	4.69	3.81	4.99	17.78	19.12	0.00	0.01	0.01	0.01	0.02	0.03	0.73	0.52	24.01	24.17
	PA	0.04	0.07	0.10	0.00	0.30	0.14	0.14	0.17	0.23	0.35	0.03	0.01	0.29	0.48	0.52	0.49
	d_2	0.40	0.45	0.49	0.43	0.60	0.52	0.35	0.30	0.54	0.61	0.41	0.43	0.57	0.69	0.72	0.70

(b)

METHODS	P.I	MEDIUM (10%)															
		WS		AT		H		SO_2		NO_2		O_3		CO		PM_{10}	
		PJ	SA	PJ	SA	PJ	SA	PJ	SA	PJ	SA	PJ	SA	PJ	SA	PJ	SA
SM	MAE	1.76	2.96	2.36	3.08	12.29	12.47	0.00	0.00	0.01	0.01	0.01	0.02	0.44	0.33	15.49	17.27
	RMSE	2.31	3.55	2.81	3.54	14.26	14.46	0.00	0.00	0.01	0.01	0.01	0.02	0.57	0.41	18.96	21.59
	PA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	d_2	0.00	0.08	0.06	0.16	0.10	0.11	0.13	0.03	0.21	0.15	0.11	0.06	0.13	0.10	0.16	0.09
LI	MAE	2.31	3.42	3.13	3.29	12.80	13.82	0.00	0.00	0.01	0.01	0.01	0.02	0.45	0.40	17.65	24.18
	RMSE	2.99	4.42	3.75	4.33	16.34	17.90	0.01	0.00	0.01	0.01	0.02	0.03	0.58	0.52	22.00	30.08
	PA	0.01	0.06	0.03	0.15	0.22	0.18	0.05	0.34	0.15	0.37	0.15	-0.15	0.26	0.19	0.22	0.47
	d_2	0.38	0.46	0.45	0.50	0.52	0.51	0.24	0.41	0.42	0.63	0.49	0.35	0.49	0.52	0.53	0.66
MNN	MAE	2.85	3.74	3.64	3.51	12.88	17.35	0.00	0.00	0.01	0.01	0.01	0.03	0.50	0.36	20.21	26.60
	RMSE	3.78	4.81	4.37	4.87	16.64	21.86	0.00	0.00	0.02	0.01	0.02	0.03	0.63	0.47	24.83	32.29
	PA	-0.12	0.04	-0.05	0.06	0.26	-0.16	0.19	0.29	-0.03	0.30	-0.03	-0.17	0.27	0.22	0.13	0.24
	d_2	0.28	0.43	0.40	0.47	0.54	0.41	0.28	0.37	0.35	0.59	0.41	0.36	0.55	0.49	0.48	0.52
EM	MAE	1.35	1.97	0.79	0.64	3.51	2.96	0.00	0.00	0.01	0.01	0.01	0.01	0.31	0.21	13.37	16.10
	RMSE	1.84	2.47	1.04	0.81	4.45	3.80	0.00	0.00	0.01	0.01	0.01	0.01	0.40	0.30	16.89	20.21
	PA	0.61	0.72	0.93	0.97	0.95	0.97	0.35	0.41	0.71	0.76	0.78	0.84	0.73	0.68	0.55	0.40
	d_2	0.73	0.81	0.96	0.99	0.97	0.98	0.44	0.54	0.78	0.85	0.87	0.91	0.81	0.81	0.73	0.59
MCMC	MAE	1.48	2.04	0.82	0.76	4.02	3.53	0.00	0.00	0.01	0.01	0.01	0.01	0.36	0.25	15.66	19.26
	RMSE	2.01	2.67	1.09	0.99	5.03	4.41	0.00	0.00	0.01	0.01	0.01	0.01	0.47	0.36	20.24	25.24
	PA	0.53	0.67	0.93	0.96	0.94	0.95	0.28	0.25	0.61	0.71	0.77	0.76	0.62	0.59	0.45	0.25
	d_2	0.71	0.80	0.96	0.98	0.97	0.98	0.45	0.45	0.75	0.83	0.87	0.87	0.77	0.75	0.67	0.53
12 MA	MAE	2.25	3.56	2.54	3.06	12.57	15.78	0.00	0.00	0.01	0.01	0.01	0.02	0.46	0.34	20.00	25.58
	RMSE	2.83	4.55	3.05	3.72	15.60	19.54	0.00	0.01	0.02	0.01	0.02	0.03	0.61	0.46	24.84	33.04
	PA	-0.06	0.02	0.14	0.10	0.25	-0.15	0.10	0.02	0.02	0.33	-0.06	-0.10	0.26	0.15	-0.15	0.30
	d_2	0.35	0.42	0.46	0.36	0.54	0.39	0.33	0.25	0.39	0.59	0.30	0.36	0.50	0.46	0.34	0.52
24 MA	MAE	1.87	3.29	2.31	2.93	11.85	12.99	0.00	0.00	0.01	0.01	0.01	0.02	0.42	0.35	18.09	22.30
	RMSE	2.39	4.06	2.77	3.51	14.08	15.82	0.00	0.00	0.01	0.01	0.02	0.02	0.55	0.46	22.69	28.84
	PA	0.13	0.06	0.24	0.15	0.26	0.07	0.08	0.08	0.15	0.31	-0.02	0.04	0.36	0.16	-0.13	0.30
	d_2	0.37	0.36	0.45	0.32	0.48	0.40	0.28	0.33	0.46	0.52	0.26	0.39	0.54	0.48	0.33	0.54
0.2 ES	MAE	2.83	3.74	3.64	3.54	12.91	17.31	0.00	0.00	0.01	0.01	0.02	0.03	0.50	0.36	20.13	26.60
	RMSE	3.74	4.80	4.34	4.83	16.60	21.72	0.00	0.00	0.02	0.01	0.02	0.03	0.63	0.47	24.77	32.29
	PA	-0.13	0.03	-0.06	0.04	0.26	-0.16	0.18	0.27	-0.05	0.29	-0.05	-0.18	0.27	0.20	0.11	0.24
	d_2	0.27	0.42	0.39	0.46	0.54	0.41	0.28	0.36	0.34	0.58	0.39	0.35	0.55	0.48	0.48	0.52
0.5 ES	MAE	2.85	3.74	3.65	3.51	12.90	17.36	0.00	0.00	0.01	0.01	0.02	0.03	0.50	0.36	20.17	26.65
	RMSE	3.78	4.81	4.37	4.86	16.65	21.86	0.00	0.00	0.02	0.01	0.02	0.03	0.63	0.47	24.82	32.31
	PA	-0.12	0.04	-0.05	0.06	0.26	-0.16	0.19	0.29	-0.03	0.30	-0.04	-0.17	0.27	0.21	0.12	0.24
	d_2	0.28	0.43	0.40	0.47	0.54	0.41	0.28	0.37	0.35	0.58	0.40	0.36	0.55	0.49	0.48	0.52
0.8 ES	MAE	2.85	3.74	3.65	3.51	12.88	17.35	0.00	0.00	0.01	0.01	0.01	0.03	0.50	0.36	20.20	26.62
	RMSE	3.78	4.81	4.37	4.87	16.64	21.86	0.00	0.00	0.02	0.01	0.02	0.03	0.63	0.47	24.82	32.29
	PA	-0.12	0.04	-0.05	0.06	0.26	-0.16	0.19	0.29	-0.03	0.30	-0.03	-0.17	0.27	0.22	0.12	0.24
	d_2	0.28	0.43	0.40	0.47	0.54	0.41	0.28	0.37	0.35	0.59	0.41	0.36	0.55	0.49	0.48	0.52

(c)

METHODS	P.I	MEDIUM (20%)															
		WS		AT		H		SO_2		NO_2		O_3		CO		PM_{10}	
		PJ	SA	PJ	SA	PJ	SA	PJ	SA	PJ	SA	PJ	SA	PJ	SA	PJ	SA
SM	MAE	1.74	2.88	2.39	3.26	11.94	13.06	0.00	0.00	0.01	0.01	0.01	0.02	0.44	0.38	19.48	25.32
	RMSE	2.23	3.43	2.79	3.78	13.75	15.02	0.00	0.00	0.01	0.01	0.02	0.02	0.55	0.49	27.17	40.18
	PA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	d_2	0.19	0.01	0.11	0.07	0.17	0.02	0.05	0.01	0.01	0.08	0.08	0.04	0.25	0.06	0.12	0.27
LI	MAE	2.15	3.49	2.68	3.95	11.89	13.83	0.00	0.00	0.01	0.01	0.01	0.02	0.50	0.37	19.83	23.65
	RMSE	2.75	4.31	3.32	4.94	15.10	17.62	0.00	0.00	0.01	0.01	0.02	0.03	0.65	0.47	27.66	35.67
	PA	0.03	-0.05	0.11	0.02	0.29	0.16	0.18	0.08	0.21	0.23	0.04	0.11	0.09	0.45	0.33	0.41
	d_2	0.41	0.38	0.47	0.44	0.59	0.52	0.43	0.32	0.50	0.53	0.40	0.48	0.42	0.68	0.56	0.56
MNN	MAE	2.24	4.18	3.13	4.31	14.49	13.46	0.00	0.00	0.01	0.01	0.02	0.02	0.69	0.48	22.21	26.97
	RMSE	2.93	5.21	3.96	5.34	18.30	17.84	0.00	0.01	0.02	0.01	0.02	0.03	0.87	0.61	30.02	40.30
	PA	0.05	-0.08	0.08	-0.02	0.16	0.25	0.00	-0.02	0.06	0.30	0.06	0.07	-0.01	0.22	0.22	0.19
	d_2	0.41	0.40	0.47	0.42	0.52	0.56	0.30	0.22	0.41	0.56	0.40	0.43	0.38	0.54	0.51	0.39
EM	MAE	1.29	1.68	0.74	0.90	4.07	3.98	0.00	0.00	0.01	0.01	0.01	0.01	0.31	0.26	15.59	22.39
	RMSE	1.70	2.20	0.96	1.31	5.58	5.93	0.00	0.00	0.01	0.01	0.01	0.01	0.42	0.35	22.72	36.19
	PA	0.65	0.77	0.94	0.94	0.91	0.92	0.42	0.30	0.67	0.70	0.77	0.79	0.63	0.71	0.55	0.42
	d_2	0.77	0.85	0.97	0.97	0.95	0.96	0.53	0.45	0.77	0.81	0.85	0.87	0.75	0.80	0.65	0.45
MCMC	MAE	1.52	1.93	0.85	1.05	4.69	4.67	0.00	0.00	0.01	0.01	0.01	0.01	0.35	0.29	18.15	24.68
	RMSE	1.99	2.56	1.11	1.55	6.38	6.98	0.00	0.00	0.01	0.01	0.01	0.02	0.48	0.38	25.95	38.15
	PA	0.53	0.68	0.92	0.91	0.89	0.89	0.32	0.17	0.51	0.61	0.73	0.71	0.53	0.63	0.41	0.31
	d_2	0.72	0.81	0.96	0.95	0.94	0.94	0.53	0.38	0.71	0.77	0.85	0.84	0.71	0.78	0.61	0.46
12 MA	MAE	2.07	3.00	2.74	3.92	13.17	12.88	0.00	0.00	0.01	0.01	0.02	0.02	0.56	0.45	19.02	28.79
	RMSE	2.64	3.72	3.39	4.68	16.32	16.56	0.00	0.00	0.01	0.01	0.02	0.03	0.72	0.57	26.40	41.58
	PA	-0.07	-0.07	0.06	-0.03	0.01	0.21	0.17	0.18	0.06	0.35	0.01	0.05	0.02	0.27	0.30	0.20
	d_2	0.32	0.28	0.45	0.40	0.38	0.52	0.42	0.42	0.43	0.57	0.40	0.42	0.38	0.57	0.51	0.42
24 MA	MAE	1.83	2.86	2.30	3.49	12.02	12.20	0.00	0.00	0.01	0.01	0.01	0.02	0.49	0.40	17.80	26.69
	RMSE	2.31	3.41	2.79	4.12	14.54	14.56	0.00	0.00	0.01	0.01	0.02	0.02	0.64	0.51	25.15	38.92
	PA	-0.03	0.13	0.20	0.03	0.14	0.28	0.19	0.21	0.10	0.32	0.06	0.08	0.03	0.32	0.38	0.29
	d_2	0.24	0.26	0.43	0.37	0.42	0.45	0.43	0.42	0.42	0.55	0.31	0.36	0.37	0.58	0.49	0.48
0.2 ES	MAE	2.25	4.18	3.14	4.35	14.67	13.62	0.00	0.00	0.01	0.01	0.02	0.02	0.68	0.48	22.09	27.05
	RMSE	2.93	5.18	3.95	5.34	18.34	17.90	0.00	0.00	0.02	0.01	0.02	0.03	0.86	0.61	29.89	40.31
	PA	0.04	-0.09	0.06	-0.04	0.13	0.24	0.00	-0.02	0.05	0.30	0.05	0.04	-0.02	0.22	0.22	0.19
	d_2	0.40	0.39	0.46	0.41	0.51	0.55	0.30	0.29	0.40	0.56	0.39	0.41	0.37	0.53	0.50	0.39
0.5 ES	MAE	2.25	4.18	3.14	4.32	14.56	13.51	0.00	0.00	0.01	0.01	0.02	0.02	0.69	0.48	22.22	27.00
	RMSE	2.93	5.21	3.97	5.35	18.32	17.88	0.00	0.01	0.02	0.01	0.02	0.03	0.87	0.61	30.02	40.30
	PA	0.05	-0.08	0.08	-0.02	0.15	0.25	0.00	-0.02	0.06	0.30	0.06	0.07	-0.01	0.22	0.22	0.19
	d_2	0.41	0.40	0.47	0.42	0.52	0.56	0.30	0.22	0.41	0.56	0.40	0.43	0.37	0.54	0.51	0.39
0.8 ES	MAE	2.24	4.18	3.13	4.31	14.51	13.47	0.00	0.00	0.01	0.01	0.02	0.02	0.69	0.48	22.21	26.98
	RMSE	2.93	5.21	3.96	5.35	18.30	17.86	0.00	0.01	0.02	0.01	0.02	0.03	0.87	0.61	30.02	40.30
	PA	0.05	-0.08	0.08	-0.02	0.16	0.25	0.00	-0.02	0.06	0.30	0.06	0.07	-0.01	0.22	0.22	0.19
	d_2	0.41	0.40	0.47	0.42	0.52	0.56	0.30	0.22	0.41	0.56	0.40	0.43	0.38	0.54	0.51	0.39

(d)

METHODS	P.I	COMPLEX (10%)															
		WS		AT		H		SO_2		NO_2		O_3		CO		PM_{10}	
		PJ	SA	PJ	SA	PJ	SA	PJ	SA	PJ	SA	PJ	SA	PJ	SA	PJ	SA
SM	MAE	1.80	2.85	2.37	3.18	12.86	13.28	0.00	0.00	0.01	0.01	0.01	0.02	0.44	0.37	16.38	17.02
	RMSE	2.38	3.49	2.77	3.75	14.76	15.67	0.00	0.00	0.01	0.01	0.01	0.02	0.55	0.45	19.81	21.37
	PA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	d_2	0.12	0.12	0.10	0.08	0.21	0.28	0.13	0.08	0.17	0.14	0.08	0.17	0.06	0.26	0.33	0.26
LI	MAE	2.06	3.30	2.40	3.41	11.42	13.55	0.00	0.00	0.01	0.01	0.01	0.02	0.45	0.39	16.24	13.83
	RMSE	2.71	4.21	3.13	4.36	15.13	17.01	0.00	0.00	0.01	0.01	0.02	0.03	0.58	0.53	20.26	18.00
	PA	0.11	0.11	0.21	0.21	0.37	0.30	0.31	0.21	0.24	0.18	-0.02	0.02	0.24	0.05	0.32	0.52
	d_2	0.41	0.47	0.54	0.55	0.63	0.61	0.51	0.36	0.55	0.51	0.37	0.43	0.51	0.40	0.58	0.68
MNN	MAE	2.19	4.52	2.98	4.37	12.71	15.69	0.00	0.00	0.01	0.01	0.01	0.03	0.63	0.33	19.59	18.82
	RMSE	2.76	5.58	3.70	5.41	17.12	20.03	0.00	0.00	0.02	0.02	0.02	0.04	0.77	0.44	24.37	25.54
	PA	0.11	-0.06	0.11	0.13	0.27	0.15	0.32	0.10	0.17	0.09	0.01	0.01	0.00	0.26	0.23	0.33
	d_2	0.45	0.39	0.48	0.50	0.58	0.51	0.53	0.30	0.50	0.42	0.41	0.41	0.43	0.50	0.53	0.62
EM	MAE	1.33	1.97	0.61	0.97	4.29	3.94	0.00	0.00	0.01	0.01	0.01	0.01	0.31	0.23	13.80	15.53
	RMSE	1.83	2.51	0.79	1.38	5.46	5.46	0.00	0.00	0.01	0.01	0.01	0.01	0.41	0.30	17.67	19.85
	PA	0.63	0.69	0.96	0.93	0.93	0.94	0.43	0.31	0.58	0.65	0.77	0.82	0.67	0.73	0.46	0.46
	d_2	0.74	0.79	0.98	0.96	0.96	0.97	0.59	0.38	0.70	0.79	0.86	0.89	0.79	0.84	0.64	0.65
MCMC	MAE	1.57	2.24	0.69	1.13	4.60	4.61	0.00	0.00	0.01	0.01	0.01	0.01	0.35	0.26	16.08	18.39
	RMSE	2.12	2.88	0.88	1.62	5.89	6.37	0.00	0.00	0.01	0.01	0.01	0.01	0.46	0.34	21.11	24.58
	PA	0.52	0.60	0.95	0.90	0.92	0.91	0.26	0.19	0.50	0.60	0.76	0.74	0.59	0.66	0.33	0.33
	d_2	0.70	0.76	0.97	0.95	0.96	0.95	0.50	0.35	0.70	0.76	0.86	0.85	0.76	0.80	0.58	0.57
12 MA	MAE	2.04	3.70	2.93	4.31	11.89	15.02	0.00	0.00	0.01	0.01	0.01	0.03	0.49	0.34	18.59	17.31
	RMSE	2.64	4.60	3.51	5.20	14.37	18.25	0.00	0.00	0.01	0.01	0.02	0.03	0.61	0.45	23.19	23.07
	PA	0.13	-0.02	-0.06	0.01	0.37	-0.02	0.33	0.17	0.17	0.20	0.00	0.00	-0.01	0.27	0.21	0.31
	d_2	0.44	0.39	0.37	0.43	0.60	0.40	0.46	0.27	0.48	0.46	0.38	0.38	0.36	0.52	0.50	0.59
24 MA	MAE	1.86	3.10	2.45	3.41	11.13	13.81	0.00	0.00	0.01	0.01	0.01	0.02	0.45	0.35	18.48	16.39
	RMSE	2.44	3.79	2.98	3.98	13.10	16.83	0.00	0.00	0.01	0.01	0.02	0.03	0.57	0.45	22.80	21.53
	PA	0.15	0.08	0.06	0.10	0.46	0.06	0.30	0.16	0.27	0.25	0.09	0.04	0.09	0.21	0.20	0.29
	d_2	0.39	0.42	0.39	0.40	0.59	0.42	0.44	0.24	0.49	0.47	0.34	0.38	0.34	0.48	0.49	0.57
0.2 ES	MAE	2.20	4.53	3.03	4.46	12.86	16.02	0.00	0.00	0.01	0.01	0.01	0.03	0.62	0.32	19.48	18.97
	RMSE	2.77	5.55	3.70	5.41	16.99	20.11	0.00	0.00	0.02	0.02	0.02	0.04	0.76	0.44	24.28	25.62
	PA	0.07	-0.09	0.08	0.08	0.26	0.11	0.31	0.10	0.16	0.09	-0.03	-0.01	-0.03	0.26	0.21	0.32
	d_2	0.42	0.38	0.46	0.47	0.58	0.48	0.52	0.31	0.50	0.42	0.39	0.40	0.40	0.50	0.52	0.61
0.5 ES	MAE	2.20	4.54	3.00	4.41	12.73	15.77	0.00	0.00	0.01	0.01	0.01	0.03	0.63	0.32	19.54	18.85
	RMSE	2.77	5.59	3.71	5.43	17.11	20.06	0.00	0.00	0.02	0.02	0.02	0.04	0.77	0.44	24.34	25.57
	PA	0.10	-0.07	0.10	0.12	0.27	0.14	0.32	0.12	0.17	0.09	0.00	0.01	-0.01	0.26	0.23	0.33
	d_2	0.44	0.39	0.48	0.50	0.58	0.50	0.53	0.32	0.50	0.42	0.40	0.41	0.42	0.51	0.53	0.61
0.8 ES	MAE	2.19	4.53	2.98	4.38	12.71	15.71	0.00	0.00	0.01	0.01	0.01	0.03	0.63	0.33	19.57	18.83
	RMSE	2.77	5.58	3.70	5.41	17.12	20.03	0.00	0.00	0.02	0.02	0.02	0.04	0.77	0.44	24.36	25.54
	PA	0.11	-0.06	0.11	0.13	0.27	0.15	0.32	0.11	0.17	0.09	0.01	0.01	0.00	0.26	0.23	0.33
	d_2	0.45	0.39	0.48	0.50	0.58	0.51	0.53	0.32	0.50	0.42	0.41	0.41	0.42	0.51	0.53	0.62

(e)

METHODS	P.I	COMPLEX (20%)															
		WS		AT		H		SO_2		NO_2		O_3		CO		PM_{10}	
		PJ	SA	PJ	SA	PJ	SA	PJ	SA	PJ	SA	PJ	SA	PJ	SA	PJ	SA
SM	MAE	1.81	2.70	2.32	3.34	13.09	12.59	0.00	0.00	0.01	0.01	0.01	0.02	0.43	0.37	19.32	20.16
	RMSE	2.30	3.19	2.71	3.87	15.15	14.70	0.00	0.01	0.01	0.01	0.01	0.02	0.56	0.45	29.13	26.97
	PA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	d_2	0.16	0.08	0.11	0.12	0.12	0.08	0.02	0.05	0.00	0.15	0.14	0.19	0.00	0.24	0.11	0.13
LI	MAE	2.08	2.98	2.51	3.78	13.50	12.86	0.00	0.00	0.01	0.01	0.01	0.02	0.53	0.33	17.45	22.20
	RMSE	2.62	3.77	3.19	4.82	17.01	16.46	0.00	0.01	0.01	0.01	0.02	0.02	0.68	0.43	22.10	30.69
	PA	0.19	0.15	0.14	0.05	0.21	0.22	0.30	0.17	0.06	0.11	0.12	0.10	0.16	0.45	0.68	0.36
	d_2	0.49	0.49	0.48	0.46	0.54	0.54	0.53	0.27	0.43	0.44	0.45	0.44	0.49	0.67	0.82	0.61
MNN	MAE	2.23	3.37	2.96	3.91	14.29	14.88	0.00	0.00	0.01	0.01	0.01	0.02	0.63	0.35	22.11	28.00
	RMSE	2.82	4.35	3.65	4.95	18.24	19.00	0.00	0.01	0.02	0.01	0.02	0.02	0.82	0.47	30.41	39.35
	PA	0.10	0.13	0.06	0.04	0.20	0.18	0.15	0.09	0.00	0.08	0.07	0.10	0.07	0.34	0.20	0.26
	d_2	0.45	0.49	0.44	0.45	0.54	0.53	0.40	0.28	0.40	0.43	0.44	0.46	0.40	0.60	0.44	0.54
EM	MAE	1.37	1.81	0.85	0.92	4.44	4.29	0.00	0.00	0.01	0.01	0.01	0.01	0.31	0.22	16.13	16.50
	RMSE	1.81	2.30	1.18	1.33	5.92	6.03	0.00	0.01	0.01	0.01	0.01	0.01	0.41	0.30	24.40	22.26
	PA	0.62	0.69	0.91	0.94	0.92	0.91	0.46	0.30	0.61	0.70	0.75	0.76	0.67	0.73	0.55	0.57
	d_2	0.74	0.80	0.95	0.97	0.96	0.95	0.57	0.29	0.73	0.81	0.85	0.86	0.77	0.84	0.63	0.66
MCMC	MAE	1.60	1.99	0.95	1.08	5.11	4.93	0.00	0.00	0.01	0.01	0.01	0.01	0.35	0.25	18.77	19.44
	RMSE	2.08	2.61	1.34	1.54	6.92	6.84	0.00	0.01	0.01	0.01	0.01	0.02	0.47	0.35	27.34	26.36
	PA	0.51	0.63	0.88	0.92	0.89	0.89	0.32	0.14	0.54	0.63	0.72	0.68	0.58	0.66	0.41	0.41
	d_2	0.70	0.78	0.93	0.96	0.94	0.94	0.53	0.30	0.72	0.78	0.84	0.82	0.75	0.80	0.60	0.62
12 MA	MAE	1.93	3.04	3.01	4.01	13.00	13.46	0.00	0.00	0.01	0.01	0.01	0.02	0.52	0.39	20.94	19.90
	RMSE	2.49	3.75	3.60	4.78	15.36	16.70	0.00	0.01	0.01	0.01	0.02	0.02	0.66	0.52	29.60	26.17
	PA	0.12	0.06	-0.14	0.04	0.29	0.25	0.01	0.08	-0.03	0.09	-0.05	-0.01	0.08	0.28	0.26	0.47
	d_2	0.40	0.43	0.34	0.44	0.53	0.57	0.33	0.24	0.36	0.43	0.33	0.37	0.40	0.56	0.48	0.67
24 MA	MAE	1.74	2.71	2.44	3.35	11.76	12.05	0.00	0.00	0.01	0.01	0.01	0.02	0.48	0.39	20.87	18.50
	RMSE	2.27	3.31	2.94	3.95	13.77	14.19	0.00	0.01	0.01	0.01	0.02	0.02	0.61	0.51	29.78	23.69
	PA	0.19	0.12	0.02	0.14	0.43	0.32	0.04	0.11	0.08	0.08	0.00	0.11	0.13	0.28	0.22	0.51
	d_2	0.37	0.40	0.35	0.40	0.59	0.55	0.31	0.24	0.41	0.43	0.27	0.29	0.38	0.57	0.43	0.68
0.2 ES	MAE	2.23	3.38	3.00	4.01	14.46	15.12	0.00	0.00	0.01	0.01	0.02	0.02	0.63	0.35	21.97	27.78
	RMSE	2.81	4.34	3.65	4.98	18.18	19.02	0.00	0.01	0.02	0.01	0.02	0.02	0.81	0.47	30.27	38.99
	PA	0.08	0.10	0.02	0.00	0.19	0.16	0.11	0.09	-0.01	0.05	0.05	0.05	0.04	0.33	0.21	0.26
	d_2	0.43	0.47	0.41	0.42	0.53	0.52	0.36	0.28	0.40	0.41	0.43	0.43	0.38	0.59	0.44	0.54
0.5 ES	MAE	2.23	3.37	2.97	3.94	14.35	14.99	0.00	0.00	0.01	0.01	0.01	0.02	0.63	0.35	22.11	27.98
	RMSE	2.82	4.35	3.66	4.96	18.26	19.04	0.00	0.01	0.02	0.01	0.02	0.02	0.82	0.47	30.39	39.32
	PA	0.09	0.12	0.05	0.03	0.20	0.18	0.14	0.09	0.00	0.07	0.07	0.09	0.06	0.34	0.20	0.26
	d_2	0.44	0.49	0.44	0.44	0.54	0.53	0.39	0.28	0.40	0.42	0.44	0.45	0.39	0.59	0.44	0.54
0.8 ES	MAE	2.23	3.37	2.96	3.92	14.30	14.91	0.00	0.00	0.01	0.01	0.01	0.02	0.63	0.35	22.11	28.01
	RMSE	2.82	4.35	3.65	4.95	18.24	19.01	0.00	0.01	0.02	0.01	0.02	0.02	0.82	0.47	30.40	39.35
	PA	0.10	0.13	0.06	0.04	0.20	0.18	0.15	0.09	0.00	0.08	0.07	0.10	0.07	0.34	0.20	0.26
	d_2	0.45	0.49	0.44	0.45	0.54	0.53	0.40	0.28	0.40	0.43	0.44	0.46	0.40	0.60	0.44	0.54

(f)