TIME SERIES MODEL FOR PREDICTING LONG-INTERVAL CONSECUTIVE MISSING VALUES IN AIR POLLUTION DATASET

DANIEL BONG KIM BOON

FACULTY OF CIVIL ENGINEERING TECHNOLOGY

UNIVERSITI MALAYSIA PERLIS

2022

TIME SERIES MODEL FOR PREDICTING LONG-INTERVAL CONSECUTIVE MISSING VALUES IN AIR POLLUTION DATASET

by

DANIEL BONG KIM BOON

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

JULY 2022

ACKNOWLEDGEMENT

First, I want to thank the Lord for keeping me healthy and allowing me to finish this final year project and research.

Next, I'd like to thank my project supervisor, PM. Ts. Dr. Norazian Mohamed Noor, for giving me the direction, advice, and help I needed to finish this project for my final year.

Then, I would like to thank the Department of Environment (DOE) for giving me the data I needed for this research.

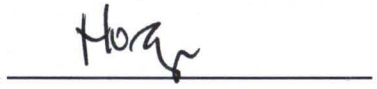
I also want to thank my classmates for the advice and information they gave me about the title of my final year project.

Lastly, I would like to thank my parents, Mr. Bong Ah Luk and Mrs. Limang ak Ajoh, for their love, support, and finances, which helped me finish my final year project and earn a degree.

APPROVAL AND DECLARATION SHEET

This project report titled Time Series Model for Predicting Long-Interval Consecutive Missing Values in Air Pollution Dataset was prepared and submitted by Daniel Bong Kim Boon (Matrix Number: 181133165) and has been found satisfactory in terms of scope, quality and presentation as partial fulfillment of the requirement for the Bachelor of Environmental Engineering with Honours in Universiti Malaysia Perlis (UniMAP).

Checked and Approved by



(PM. Ts. DR. NORAZIAN MOHAMED NOOR)
Project Supervisor

Faculty of Civil Engineering Technology
Universiti Malaysia Perlis

July 2022

MODEL SIRI MASA UNTUK MERAMALKAN NILAI HILANG SELANG PANJANG DALAM SET DATA PENCEMARAN UDARA

ABSTRAK

Data yang dijana daripada stesen pemantauan kualiti udara ambien berterusan (CAAQM) sudah pasti akan mengandungi jurang data yang hilang yang panjang dan pendek. Biasanya, ini disebabkan oleh kegagalan mesin, penyelenggaraan, perubahan dalam penempatan peralatan pemantauan, atau kesilapan manusia. Apabila set data tidak lengkap, ia boleh menyebabkan kehilangan data dan melemahkan kekuatan analisis statistik. Objektif utama penyelidikan ini adalah untuk menilai prestasi kaedah Siri Masa – Auto Regression Integrated Moving Average (ARIMA) - untuk mengira jurang panjang data yang hilang dalam dataset pencemaran udara. Markov Chain Monte Carlo (MCMC) dan Expectation-Maximization (EM) digunakan untuk membandingkan prestasi ARIMA dalam jurang panjang yang tiada nilai. Data pemantauan setiap jam Pegoh dan Kota Kinabalu (2018) untuk PM_{10}, SO_2, NO_2, O_3, CO, kelajuan angin, kelembapan relatif, dan suhu persekitaran dicirikan menggunakan statistik deskriptif. Set data ini telah diprarawat dengan interpolasi linear untuk mengimbangi nilai sedia ada yang hilang. Seterusnya, set data yang dirawat telah disimulasikan dengan tiga peratusan (5%, 10%, dan 15%) corak data yang hilang dengan pelbagai panjang jurang data yang hilang antara 24 hingga 120 jam. Empat ukuran prestasi digunakan untuk menilai kesesuaian kaedah imputasi, iaitu Ralat Min Mutlak, Ralat Purata Purata Purata, Ketepatan Ramalan, dan Indeks Perjanjian. Secara keseluruhannya, pendekatan Expectation-Maximization ditentukan sebagai kaedah imputasi terbaik untuk mengisi semua peratusan simulasi data hilang, manakala ARIMA adalah kaedah imputasi yang kurang sesuai dalam kajian ini.

TIME SERIES MODEL FOR PREDICTING LONG-INTERVAL CONSECUTIVE MISSING VALUES IN AIR POLLUTION DATASET

ABSTRACT

The data generated from a Continuous Ambient Air Quality Monitoring (CAAQM) station will undoubtedly contain long and short gaps of missing data. Typically, this is due to machine failure, maintenance, a change in the placement of the monitoring equipment, or human mistake. When a dataset is incomplete, it can cause a bias and weaken the strength of the statistical analysis. The primary objective of this research is to evaluate the performance of the Time Series method - Auto Regression Integrated Moving Average (ARIMA) - to impute long gaps of missing data in an air pollution dataset. Multiple Imputation - Markov Chain Monte Carlo (MCMC) and Expectation-Maximization (EM) were used to compare the performance of ARIMA in the long gaps missing values. Pegoh and Kota Kinabalu hourly 2018monitoring data for PM_{10}, SO_2, NO_2, O_3, CO, wind speed, relative humidity, and ambient temperature were characterized using descriptive statistics. This dataset was pre-treated with linear interpolation to compensate existing for missing values. Next, the treated datasets were simulated with three percentages (5%, 10%, and 15%) of missing data patterns with various lengths of missing data gaps ranging from 24 to 120 hours. Four performance measures were used to assess the suitability of the imputation methods, namely Mean Absolute Error, Root Mean Squared Error, Prediction Accuracy, and Index of Agreement. Overall, the Expectation-Maximization approach was determined to be the best imputation method for filling in all the percentages of simulated missing data, whereas the ARIMA was the least desirable imputation method in this study.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

ACF	Auto-correlation Function
AR	Auto Regression
ADF	Augmented Dickey-Fuller
ARIMA	Auto-Regression Integrated Moving Average
AT	Ambient Temperature
CAAQM	Continuous Ambient Air Quality Monitoring
CO	Carbon Monoxide
d_2	Index of Agreement
DOE	Department of Environment
EM	Expectation-Maximization
ES	Exponential Smoothing
LI	Linear Interpolation
MA	Moving Average
MAAQG	Malaysia Ambient Air Quality Guidelines
MAE	Mean Absolute Error
MAR	Missing at Random
MCAR	Missing Completely at Random
MCMC	Markov Chain Monte Carlo
MI	Multiple Imputation
MLE	Maximum Likelihood Estimate
MNAR	Missing Not at Random

MNN	Mean Series Neighbour
NO_2	Nitrogen Dioxide
O_3	Ozone
PA	Prediction of Accuracy
PACF	Partial Auto-correlation Function
PM_{10}	Particulate Matter 10
PMM	Predictive Mean Matching
RMSE	Root Mean Squared Error
R^2	Squared Correlation Coefficient
RH	Relative Humidity
SM	Series Mean
SO_2	Sulfur Dioxide
WS	Wind Speed

CHAPTER 1

INTRODUCTION

1.1 Background History

Air is one of the most valuable natural resources because it is required for the existence of humans, animals, plants, and the overall ecosystem's management. It is a gaseous combination that acts as a spacesuit for the biosphere and is maintained in place by gravity. Air contributes to the preservation of heat, which heats the earth's surface and reduces day-night temperature extremes and the absorption of ultraviolet solar radiation. The concentration of gaseous pollutants and the size or number of particulate materials generated by natural or artificial sources determine air quality. Polluted air happens when the concentration of harmful particulate matter and gases emitted into the atmosphere exceeds a safe level.

Air quality monitoring is done to determine the level of pollution that can harm humans. Air quality monitoring aims to measure ambient air quality and notify the public if there is a substantial change. Continuous air quality monitoring stations must be serviced regularly to guarantee an accurate dataset. The air pollution dataset will be incomplete due to maintenance and calibration. Uncontrollable factors such as device breakdown, maintenance, and calibration cause missing measurement of air pollutants. When undertaking mathematical analyses such as time series analysis, principal component analysis, and multivariate analysis, missing data might pose a series problem.

One solution to this problem is to discard data with missing data. However, discarding data will resulting an uncompleted data which will lead to series problem for the dataset. A complete data set enables us to evaluate the efficiency of a specific

strategy. When strategies are implemented to solve an issue, gathering data will help us understand how effective our solution is and whether our strategy requires long-term improvement. This allows us to obtain the desired outcomes from our investigation. In addition, it enables us to make accurate decisions on our next wayfinding. Good data gives conclusive proof. However, personal experience, assumptions, or speculative observation may result in the waste of resources due to taking action based on an inaccurate conclusion (Dekker, 2006). Assumption will only lead to risk of losing important data. Instead of assuming the data, we can apply series of imputation method. The imputation approach can fill in the missing data (Little & Rubin, 1989). The missing data can be imputed, which is a more appropriate strategy. Numerous imputation method to applied to fill in the missing values, include Series Mean (SM), Mean Nearest Neighbour (MNN), Expectation-Maximization (EM), Markov Chain Monte Carlo (MCMC), Linear Interpolation (LI), and Exponential Smoothing (ES). This strategy was successful in filling in missing data for the dataset.

In this study, the main focus is filling in long gaps of missing data in the air pollution dataset. These long gaps of missing data will be filled using the time series imputation method Auto Regression Integrated Moving Average (ARIMA). Other imputation methods to fill the missing data are Expectation Maximized (EM) and Multiple Imputation - Markov Chain Monte Carlo (MCMC), which are used to compare with the time series method in the study.

1.2 Statement of the Problem

Air quality monitoring identifies rising levels of pollutants that may be harmful to human health. However, the monitoring measurement dataset frequently contain significant proportions of missing observations due to machine failure, routine maintenance, monitor relocation, human error, or other factors (Libasin, Fauzi, Ul-Saufie, Idris, & Mazeni, 2021). Missing data in air pollution dataset may become problematic for further research.

Missing data, as well as incomplete data set, are common problems in environmental scientific research. A lack of proper sample, measurement issues, or mistakes in data collection could all contribute to the circumstance. Regardless of the reason, inconsistencies are a significant barrier to applying time-series prediction models, which require continuous data to be effective (Ibrahim, Zailan, Ismail, & Lola, 2009).

Incomplete datasets might yield results that differ from those collected if the dataset had been completed. It is a rare occasion that the collected data does not have any missing data in it. Long gaps of missing data are one major problem in any dataset. It may lead to several significant problems in real life when facing long gaps of missing data in the dataset. The first problem is that there will be a loss of knowledge for the study or research, leading to a loss of efficiency. Second, there are several obstacles in data processing, computing, and analysis due to the irregularity in the data format and the difficulty of using standard software. Some software is unable to treat long gaps of missing data. Even though it might be able to treat the data, there will be a loss of important information or data in the dataset due to long gaps of missing data. Finally, and most importantly, the results may be skewed due to systematic differences between observed and unseen data sets (Zakaria & Noor, 2018).

In this study, three imputations can be the option for solving the problem of long gaps of missing data. Time-series methods were Autoregressive Integrated Moving Average (ARIMA) methods will fill in the missing data on air pollution. The various methods will be compared using four performance metrics to choose the optimal approach.

1.3 Objectives of the Study

The main aim of this research is to estimate the long-interval consecutive missing observations in air pollutant data. To achieve these objectives, the following objectives of the study are planned in several phases along the course of research:

I. To study the characteristics of air pollutant data in the study area.

II. To evaluate the characteristic of the simulated missing data (5%, 10% and 15%).

III. To assess the performances of the time series model to estimate the simulated missing data using performance indicators.

1.4 Scope of the Study

In this study, the data set for hourly air pollution for 2018 was used in Pegoh, Perak, and Kota Kinabalu, Sabah. The data was obtained from the Department of Environment Malaysia. Pegoh is a crowded residential area. Meanwhile, Kota Kinabalu is an urban area filled with skyscrapers and modernized buildings. Hence, different air quality characteristics can be observed in both locations. There are five air pollution data which are particulate matter (PM_{10}), sulfur dioxide (SO_2), nitrogen dioxide (NO_2), ozone (O_3), and carbon monoxide (CO). They are also three meteorological data for this study that are wind speed (WS), relative humidity (RH), and ambient temperature (AT) for both locations. SPSS, E-views, and Microsoft Excel were used to conduct the analysis.

A descriptive analysis was performed on the raw dataset to analyse the pattern of missing data and establish the reference data. All data and information were included in the dataset. The dataset is then pre-treated using linear interpolation and acts as reference data for the study. The cumulative data for

missing values in the reference data will be generated to provide a complete set of air pollution data.

The complete dataset was simulated into several percentages of missing data. The data simulation design was based on three missing percentages of 5%, 10%, and 20%. The simulated missing data were then filled in using three methods that were times series method Autoregressive Integrated Moving Average (ARIMA), Expectation-Maximization (EM), and Multiple Imputation - Markov Chain Monte Carlo (MCMC). Finally, four performance measures will be utilized to evaluate whether the imputation method is the most suitable to fill the long gaps missing data of air pollution data: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Prediction Accuracy (PA), and Index of Agreement (d_2).

1.5 Significant of the Study

Missing data was regularly discovered in the Continuous Air Quality Monitoring (CAAQM) station's air quality measurements. The most typical sources of missing data include machine failure, normal maintenance, and human error. Missing data in the dataset have always restricted accurate prediction. It could lead to a misinterpretation of the current state of air pollution. This paper focuses on the use of ARIMA imputation to replace missing data. Four performance factors were used to assess how effectively this technique matched the needs of the study.

This model should increase the accuracy of filling in missing data, and the model's excellent prediction of air pollution data shows that it has good universal applicability. This system has the potential to play a significant role in the development and promotion of compact air quality monitors and grid-based monitoring of pollutant concentrations. Researchers can utilise this strategy to conduct correlation and modelling studies with high-quality data that can be used for further investigation.

CHAPTER 2

LITERATURE REVIEW

2.1 Continuous Air Quality Monitoring

Continuous Air Quality Monitoring Stations (CAQMS) are a cost-effective method of monitoring compliance. CAQMS have been shown to be more accurate and reliable than Continuous Emissions Monitoring Systems (CEMS) in independent field tests conducted in accordance with US EPA test methods and procedures. The Department of the Environment (DOE) Malaysia maintains a network of 65 stations to monitor the country's ambient air quality. These monitoring stations are strategically placed in residential, commercial, and industrial areas to detect any significant change in air quality that could be detrimental to human health or the environment. CAQM stations are classified into five (5) distinct categories. In Malaysia, 26% of stations are industrial, 57% are residential, 2% are traffic, 2% are background, and 13% are PM_{10} stations (Department of Environment, 2009). Parameters measured in 4 categories of CAAQM stations:

Table 2.1: Air pollutants parameters according to category in CAAQM (DOE, 2019)

Category	Sulphur Dioxide	Nitrogen Oxides	Carbon Monoxides	Ozone	Hydrocarbon	PM_{10}	UV
Industrial	X	X	-	-	X	X	-
Residential	X	X	X	X	X	X	X
Traffic	X	X	-	X	X	X	-
Background	X	X	X	X	X	X	X
PM_{10}	-	-	-	-	-	X	-

The Air Pollution Index (API) is a method for determining the quality of the air based on the concentrations of five primary pollutants: ground-level ozone (O_3), carbon monoxide (CO), nitrogen dioxide (NO_2), sulphur dioxide (SO_2), and particulate matter (PM_{10}). Air quality is classified into five categories: good, moderate, unhealthy, extremely unhealthy, and hazardous which is set by Department of Environment Malaysia. The air quality status using API scale in Malaysia is depicted in Table 2.2 below.

Table 2.2: Air quality status based on API scale in Malaysia (Department of Environment)

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

Ambient air quality is getting concerning and poses an environmental problem that significantly influences human health. Due to this problem, Malaysia enacted the Malaysian Ambient Air Quality Guidelines (MAAQG) in 1989. These guidelines emphasized the concentration limit for major pollutants such as ozone, carbon monoxide, nitrogen dioxide, Sulphur dioxide, particulate matter, total suspended solids, and lead over an average period. These rules have been established to preserve air quality and safeguard public health. Since 1989, there has been no adjustment to these rules until Malaysia introduced new Ambient Air

Quality Standard guidelines in 2013. The new Ambient Air Quality Standard was designed to replace the 1989 Malaysian Ambient Air Quality Guideline. The New Ambient Air Quality Standard adopts six criteria for air pollutants, including five existing ones: particulate matter less than 10 microns (PM_{10}), sulphur dioxide (SO_2), carbon monoxide (CO), nitrogen dioxide (NO_2), and ground-level ozone (O_3), as well as one new one: particulate matter less than 2.5 microns ($PM_{2.5}$). The limit for air pollutants will be gradually increased until 2020. Three interim targets have been established: interim target 1 (IT-1) in 2015, interim target 2 (IT-2) in 2018, and full standard implementation in 2020. Table 2.3 below showing new Malaysia Ambient Air Quality Standard.

Table 2.3: New Malaysia Ambient Air Quality Standard (Department of Environment)

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

2.2 Missing Data in Air Pollution Dataset

Accurate prediction has always been hampered by missing values in the dataset. This may result in a misleading understanding of the air pollution scenario. Each problem may have a limited number of incomplete solutions; however, the missing details may vary. This study focuses mostly on resolving long-term data gaps. The term "single missing value imputation" refers to the process of replacing blank space in a monitoring dataset from a selected Department of the Environment

(DOE) monitoring station with the computed value from the best technique for long gap hours (Libasin, Fauzi, Ul-Saufie, Idris, & Mazeni, 2021).

Simple missing data or short gaps missing data can be easily treated using a simple imputation method. However, treating the data for long gaps in missing data is more complex. More missing data mean more critical data is absent from the data, and such gaps are often the only sign of a massive change in the data series. Other possible consequences of long gaps in missing data include a decrease in the sample size and statistical power and a decrease in the precision and accuracy of parameter estimations (Kellermann, 2018). Loss of accuracy results in incorrect conclusions or skewed judgments regarding outcomes and relationships of interest decreased precision of estimates resulting in decreased performance of confidence intervals, and rising standard errors. Therefore, randomly assuming the long gaps of missing data appears risky. Air pollution data may be missing for significant reasons, such as the finding that a gauge was malfunctioning or was damaged by a storm, or the equipment may recently be replaced with a new machine. It is better to check for the equipment first before we want to impute the missing data.

2.3 Types of missing data

Generally, there are three types of missing data; namely missing complete at random (MCAR), missing at random (MAR) and missing not at random (MNAR) (Little & Rubin, 2002).

2.3.1 Missing completely at random (MCAR)

In theory, MCAR data can be found, but only if you build an operation that randomly eliminates a small percentage of the datasets. Randomly missing data indicates that we can do analyses using just variables with complete data if we have enough features (Zakaria N. A., 2018). In most cases, the MCAR assumption is not valid. It is only likely to be correct if the data is missing due to random error.

When the missing data are MCAR, it is acceptable to ignore the dataset. It is okay to ignore the missing data when making sampling distribution assumptions

about the data parameters. In some research (Rubin, 1976), MCAR is considered ignorable, which means the dataset does not require treatment and may be "ignored." It is challenging to meet MCAR in practical applications because of uncontrollable occurrences during data collection, frequently linked to the research variables. The MCAR assumption is thought to be more reasonable if the missing data are absent on purpose, such as when the data cannot capture data in cold weather.

2.3.2 Missing at random (MAR)

In this case, the probability for a data point to be missing has nothing to do with the missing data itself but rather with some of the dataset's direct observation. As far as MAR is concerned, missing data can be predicted from the values of other variables in a dataset. It is necessary to utilize an advanced imputation method, such as multiple imputations, or a specific analytic method for missing at random data when data is missing (Noor, Yahaya, & Abdullah, 2008).

MAR often occurs in collecting air pollution data and causes small data gaps between the dataset. The unpredictable occurrence of this missing data results in varied lengths for both data segments and sampling breaks. For the air pollution dataset, the missing samples cannot be retrieved or remeasured due to the unique nature of the original studies or the difficulty of acquiring the observations. Consideration must be given to the utilized method for treating the data. MAR predicts missingness based on other factors in the observed data. MAR is also considered an ignorable missing data since the missing data is associated with previously captured data (Schafer & Graham, 2002). In practice, MAR is the most frequent and commonly utilized assumption regarding missing data problems, and numerous computational approaches for processing data under MAR have been effectively established. In general, there is no mechanism to assess whether MAR holds in a data collection other than to collect follow-up information from the previous dataset.

2.3.3 Missing Not at Random (MNAR)

Unobserved events or factors are frequently related to MNAR, which means that missing data is linked to occurrences or factors those researchers have not observed. If participants with severe depression are more likely to refuse to complete the survey regarding depression severity, the depression registry may encounter data that are MNAR. Suppose a complete case analysis of a data set containing MNAR data is biased (Little & Rubin, 2002). In that case, the fact that the sources of missing data are itself unmeasured means that this issue cannot be addressed in analysis and the estimate of effect will likely be biased.

2.4 Methods to Fill in the Missing Data

The most common way to deal with missing data is to discard the missing data. When the basis of the missing data is not random, the Listwise Deletion might introduce severe biases, such as Missing Not at Random (MNAR). However, suppose the total number of missing observations is modest compared to the number of remaining samples. A random pattern for missing such as Missing at Random (MAR) or Missing Completely at Random (MCAR) happened. In that case, Listwise Deletion may be a fair decision (Sukatis, 2019).

2.4.1 Listwise Deletion

Listwise Deletion is an approach for handling values that are missing. This approach requires eliminating all observations with missing values for any variable. Consequently, only the observations containing all their values are included in the analysis. Listwise Deletion has two advantages: it may be used for any statistical analysis and requires no particular computer software for the discarding procedures (Allison, 2002). The Listwise Deletion approach offers more significant statistical advantages if the data are missing completely at random. Since the missingness is entirely random, the observations that will be eliminated are randomly distributed over the dataset. The approach will provide estimators that are impartial and standard errors that are accurate (Allison, 2002).

Most statistical software uses Listwise Deletion by default since it is simple to implement and compute. The downside of this technique is that if the number of missing values is too high relative to the sample size, even for a dataset containing data that is Missing Completely at Random, the precision and accuracy of results of the result may be decreased (Myrtveit, Erik, & Olsson, 2001). Despite these drawbacks, Listwise Deletion is not a terrible method for managing missing data. Even though it does not utilize all available information, it provides acceptable conclusions for data that is Missing Completely at Random the majority of the time (Allison, 2002).

2.4.2 Single Imputation

Single imputation fills in the exact value for each missing item with single imputation. Only one estimate is substituted for each missing item in single imputation (Noor, Yahaya, & Abdullah, Estimation of missing values in air pollution data using single imputation techniques, 2008). Multiple imputations created numerous simulated values for each one. This approach may be used directly and requires considerable work to make imputations and needs to be done just once; it also has many interesting qualities (Sukatis, 2019). In addition to utilizing entire data investigation methods on each data set, the multiple imputation method is commended because it maintains observed data's variability so that estimations are not influenced by the imputation method.

The most common used imputation in single imputation is Mean Imputation. Mean imputation replaces the missing value with the mean value of each variable on the individual missing variables as an approximation of the missing value (Zakaria & Noor, 2018). Analysis methods studies can be influenced by the mean imputation, which underestimates the dataset's variation. Furthermore, this strategy is vulnerable to bias and high mistakes. Finding that small adjustment to mean methods such as mean top-bottom performed slightly better if there were a small number of missing values was also proven correct.

2.4.3 Multiple Imputation

Multiple Imputation (MI) is a Monte Carlo approach established by Rubin in 1976. Every missing observation in the dataset is imputed with a simulated value to build a complete dataset. The data is processed several times in MI to build the dataset. MI was developed in response to the limitations of single imputation approaches for dealing with nonresponse in surveys, and a comprehensive explanation for the method was presented. The main contrast between MI and single imputation is how they handle imputed data. Inferences based on single imputation treat imputed data as observed data, even though we have less confidence in the imputed values than if they were seen. By aggregating the results of multiple imputations, inferences based on MI enable us to reflect the uncertainty in the missing data. However, MI is not limited to survey analysis and may be used in any scenario to impute missing data (Rubin, 1976).

There are various imputation methods available, but the Markov Chain Monte Carlo (MCMC) approach is the most used in statistical software. The Markov Chain Monte Carlo (MCMC) method initially simulates a random independent draw from the conditional distribution of missing values given the observed. Next, the Bayesian approach estimates the Bayesian distribution's parameter values. These values are subsequently employed to inform repeated-imputation inference (Schafer J. L., 1999). Each iteration refines the estimation by constructing new regression equations using the mean and covariance vectors from the previous round of imputation. The residual error is added to each piece at random.

2.4.4 Expectation Maximize

Another well-known method is Expectation Maximize; it is common practice in data analysis to utilize expectation maximization as a means of dealing with missing data. Indeed, expectation maximization overcomes some of the limitations of other techniques, such as mean substitution or regression substitution. There are more accurate ways to estimate the true systematic deviation of a distribution. Expectation maximization overcomes this difficulty. The EM algorithm has several interesting aspects. First, an EM estimation is unbiased and efficient when the missing method is ignorable. Second, the EM method is simple, straightforward to implement and stable. Third, it is straightforward in EM to compare different models using the likelihood ratio test because EM is based on the likelihood function.

2.5 Study on Estimating Missing Data

Missing data can impair the statistical power of an investigation and lead to skewed results and invalid conclusions. This section discusses the issues and types of missing data, as well as the methods for dealing with them. The causes behind missing data are demonstrated, and solutions for dealing with them are described.

In the study run by Suhaimi, Ghazali, Nasir, Mokhtar, & Ramli (2017), they use of Multiple Imputation technique to replace missing values in air quality dataset. Multiple imputation of missing value technique was utilized to cope with selective air quality data by employing Markov Chain Monte Carlo (MCMC). Expectation-maximization (EM) algorithm was used to generate the Maximum Likelihood Estimate (MLE), assuming a multivariate normal distribution for the data. The hourly air quality data was utilized to test the performance of the imputation technique. The relative efficiency was determined to analyses the accuracy technique for substituting missing values. The result demonstrates that the relative efficiency is high. Based on the result of performance indicators shows that the MCMC method produce the good linear connection because of R_2 and PA are high approaches to 1 for both stations and follows with minor mistakes for RMSE

and MAE. Therefore, it can be stated that MCMC approach is adequate way for substituting missing air quality data. MCMC approach is also dependable to replace missing value either low or large proportion of missing data. Missing values are always inevitable, but a proper imputation may help correct the analysis as much as necessary (Suhaimi, Ghazali, Nasir, Mokhtar, & Ramli, 2017).

A study conducted by Ye et al. (2019) was to improve the accuracy of air pollution forecasting in Shenzhen, a hybrid model based on ARIMA (Autoregressive Integrated Moving Average) and prophet was proposed for combining time and space interactions. To begin, the ARIMA and Prophet methods were used to train and weight the data from 11 air quality monitoring stations. Then, complete the weighted impact computation for each air quality monitoring station to obtain the final results. Finally, constructed the hybrid model and evaluated the errors. The trials demonstrated that this hybrid strategy of ARIMA may significantly enhance the forecast of air pollution in Shenzhen (Ye, 2019).

Rumaling, et al., (2020) applied Nearest Neighbor Method (NNM) and Expectation-Maximization (EM) algorithm are the two most extensively utilized methods to fill in missing data in their research. Thus, this research intends to compare both methodologies by imputing missing air quality data in five monitoring stations in Sabah, Malaysia. Missing PM_{10} data is inserted into the datasets at five different levels to facilitate performance measurement (5%, 10%, 15%, 25%, and 40%). The missing data are imputed by employing both NNM and EM algorithms. The performance of the data imputation approaches is evaluated using performance measures (RMSE, MAE, IOA, COD) and regression analysis. Based on performance metrics and regression analysis, NNM performs better compared to EM. This may be due to air quality data missing at random (MAR) at every station. Accuracy study using Mean Absolute Percentage Error\s(MAPE) demonstrates that NNM is a more accurate imputation method for the majority of the scenarios.

Weerasinghe (2010) is studying how to estimate imputed values using the series mean directly, and the Expectation Maximize algorithm in his medical data. His research shows EM has better performances than the series mean in imputing

the missing data in his dataset. The size of missing data imputed using the forecast method favors the current observed values and trends. Despite its computational efficiency, the EM approach biases towards the series mean (Weerasinghe, 2010). Furthermore, existing methods ignore prediction errors, which are critical for estimating rare health events. After a stationary time, series with a random error, the method is demonstrated using one-day average sulfur dioxide levels. Prediction error has not been taken into account by existing imputation methods. The best approach can be found when prediction errors are added to an intermediate prediction method. Comparing performance to values without prediction error shows that the model is valid and works well.

A transformation that allows the exact likelihood function of an ARMA process to be efficiently evaluated computationally by decomposing the covariance matrix using the Cholesky decomposition (Penzer & Shea, 1999). The study stated that a proposed transformation allows the Cholesky decomposition to determine the exact likelihood function of an ARIMA model with missing data in their study on Finite Sample Prediction and Interpolation for ARIMA Models with Missing Data. This ARIMA method has been modified to allow the computation of predictions for future observations using a finite representative sample. In addition, the output of the exact likelihood evaluation can be used to estimate missing values in series. Furthermore, this approach is used to determine the exact likelihood function of an ARMA process when data are insufficient. According to the time data provided by these authors, the Cholesky approach is often preferable to the Kalman filter when the data set does not have a substantial number of missing values distributed across the series. They generalize this approach to the missing data problem in air pollution data.

Table 2.4: Summary on study on estimating missing data

Citation	Method	Data Type	Best Method
(Suhaimi, Ghazali, Nasir, Mokhtar, & Ramli, 2017)	1. Markov Chain Monte Carlo 2. Expectation-Maximization	Air quality data	Markov Chain Monte Carlo
(Ye, 2019)	ARIMA	Air pollution data	ARIMA
(Rumaling, Chee, Dayou, & Chang, 2020)	1. Nearest Neighbor Method 2. Expectation-Maximization	Air pollution data	Nearest Neighbor Method
(Weerasinghe, 2010)	1. Series Mean 2. Expectation-Maximization	Medical data	Expectation-Maximization
(Penzer & Shea, 1999)	1. ARIMA 2. Cholesky method	Air pollution	Cholesky method

2.6 Performances Indicator

The Performance Indicator describes the accuracy of the model for each imputation method. Error measures (root mean squared error and mean absolute error) and performance measures were used to assess each imputation method's goodness of fit (prediction accuracy and index of agreement) (Junninen, Niska, Tuppurainen, Ruuskanen, & Kolehmainen, 2004).

The root means squared error (RMSE) is a regularly used statistical tool to assess model performance in studies on air quality, meteorology, and climate. The Predictive accuracy (PA) is calculated by comparing the experimental and estimated values. However, the estimated values may belong to different data. The index of agreement (d_2) is a statistical measure of model performance that compares model estimations or predictions (P) with reliable observations (O) (Tang, Kassim,

& Abubakar, 1996). Typically, the set of model prediction errors is formed of the difference between predicted and observed values, with the most dimensionally accurate measures of model performance based on the dataset's sample data (Junninen, Niska, Tuppurainen, Ruuskanen, & Kolehmainen, 2004).

CHAPTER 3

METHODOLOGY

3.1 Research Flow

In this study, Pegoh, Perak, and Kota Kinabalu, Sabah, hourly air pollution data from 2018 were used. Particulate matter (PM_{10}), sulfur dioxide (SO_2), nitrogen dioxide (NO_2), ozone (O_3), and carbon monoxide (CO) are the five types of air pollution data (CO). For this study, there are also three pieces of weather information: the wind speed (WS), the relative humidity (RH), and the ambient temperature (AT) for both places. The data set was simulated with three percentages of simulated long gaps missing data: 5%, 10%, and 15%. For every air pollution parameter, the missing gaps will be between 24 hours and 120 hours for every air pollution parameter. In the air pollution dataset, long gaps are considered by missing more than 24 hours. The three percentages of simulated missing data will be treated or imputed using the time series method - ARIMA, Expectation Maximize, and multiple imputations - Markov Chain Monte Carlo (MCMC). Four performance measures were used to judge the imputation methods: Prediction Accuracy (PA), Index of Agreement (d_2), Mean Absolute Error (MAE), and Root Mean Square Error (RMSE). These performances indicators were performed to test the accuracy of an imputation method. SPSS, E-views, and Microsoft Excel were used to conduct the analysis in this research.

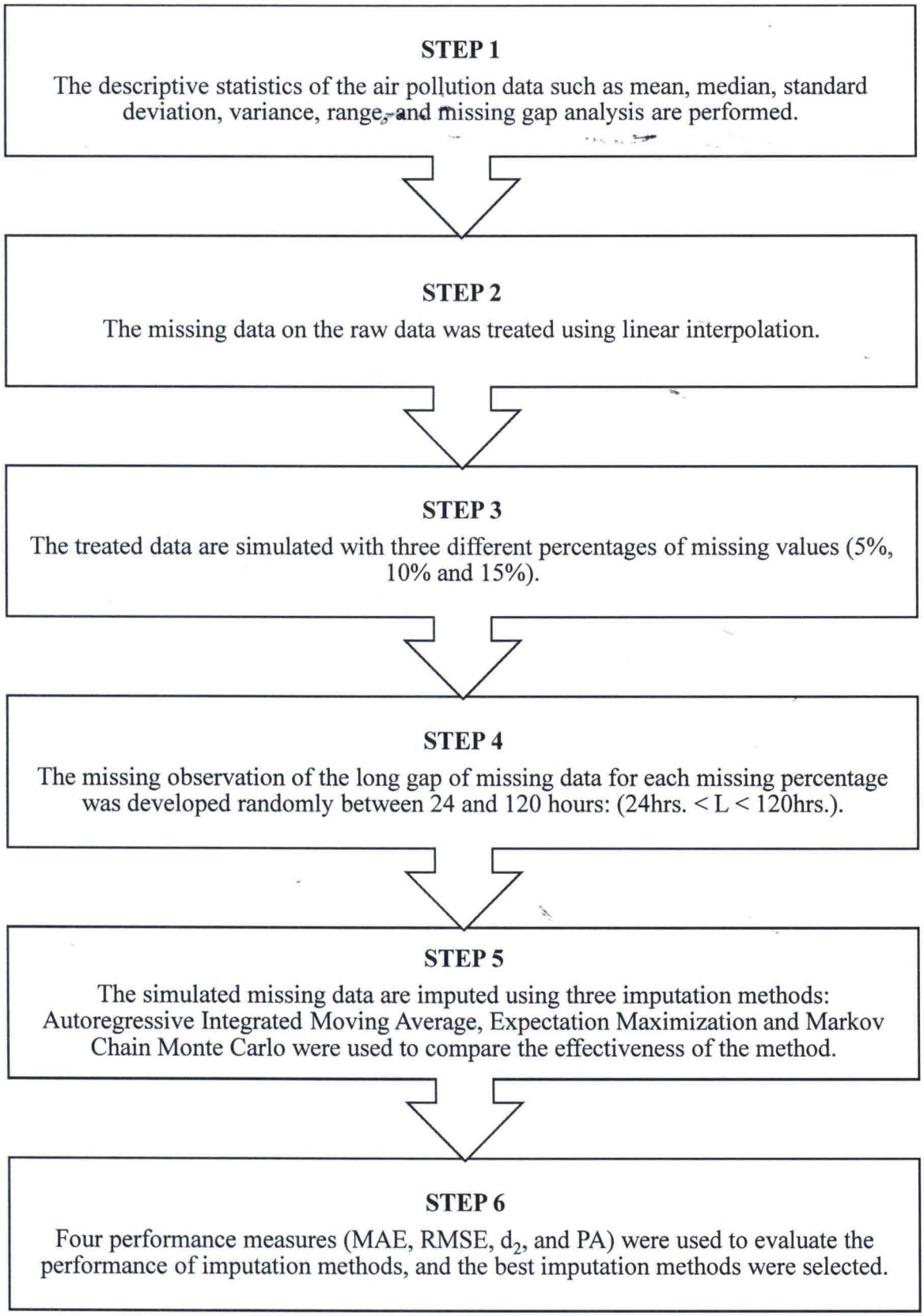


Figure 3.1: The research flowchart

3.2 Data

In this study, eight parameters from air quality monitoring data were used. These data were acquired from the Department of Environment, Malaysia. There were two locations for this study, Pegoh, Perak and Kota Kinabalu, Sabah. The following are the eight parameters of air pollutants:

I. Particulate Matter ($\mu g/m^3$)
II. Sulfur Dioxide (ppm)
III. Nitrogen Dioxide (ppm)
IV. Ozone (ppm)
V. Carbon Dioxide (ppm)
VI. Windspeed (m/s)
VII. Relative Humidity (%)
VIII. Ambient Temperature (°C)

3.2.1 Descriptive Statistics

Descriptive statistics is the collection of short descriptive coefficients for a data set that represents the total population or a sample. The main goal is to summaries the samples and data collected in a study. Combined with numerous visual analyses, descriptive statistics is an essential component of all quantitative data analyses (Sharma, 2019). It is also used to perform visual analyses of the data. In general, descriptive statistics is used to describe how a sample of data behaves. It is used to demonstrate the quantitative analysis of a data set. Because multiple variables must be measured in a study, descriptive statistics are used to summaries this vast amount of data.

There are several descriptive statistical measures. However, in this study, the descriptive statistic is mean, median, standard deviation, minimum value, and the maximum value in the dataset.

Table 3.1: The formula for descriptive statistic uses in this research

Descriptive Statistics	**Formula**	**References**
Mean	$mean, \bar{x} = \frac{\sum xi}{N}$	(Chapra & Canale, 1998)
Median	$median, m = \frac{(N+1)}{2}$ term; when N is odd Or $median, m = \frac{\frac{N}{2} term + (\frac{N}{2}+1) term}{2}$; when N is even	
Standard Deviation	$s = \sqrt{\frac{\sum_{i=1}^{n}(xi + \bar{x})^2}{n-1}}$	
Minimum Value	The minimum value in the dataset	-
Maximum Value	The maximum value in the dataset	-

Note: xi = each of the values of the data, N = number of variables in the dataset, n = the number of the data points

3.2.2 Data Pre-Treatment

In this study, the raw data was undergone pre-treatment first due to missing data in the dataset. This missing data may be due to machine failure, routine maintenance, monitor relocation, human error, or other factors causing short gaps of missing data (Libasin, Fauzi, Ul-Saufie, Idris, & Mazeni, 2021). This raw data must be treated first to obtain complete data before simulating the data into three different percentages of missingness. This raw data was treated using linear interpolation. In most air pollution data, linear interpolation is the most common imputation method to treat or impute the short gaps in missing data in the air pollution dataset. The study run by (Noor, Abdullah, Yahaya, & Ramli, 2014)

concludes that linear interpolation is the best method to fill in short gaps of missing data in the air pollution dataset.

Linear interpolation means estimating a missing value by connecting points in ascending order on a straight line. In short, it estimates the unknown value in the same ascending order as the previous values.

The air pollution raw data set was pre-treated using this method. The linear interpolation function's equation is (Chapra & Canale, 1998):

$$f(x) = f(x)_0 + \frac{f(x)_1 - f(x)_0}{x_1 - x_0}(x - x_0) \quad (3.1)$$

3.4 Simulation of Missing Data

The simulation of missing data mainly aims to investigate the efficiency of the time series methods used in this study. The simulation data were divided into three groups of missing percentages: 5%, 10%, and 15%. The data was used to compare four individual imputation strategies. The range of missing hours used in this study is (24 hours < L < 120 hours) (Zakaria N. A., 2018). The real goal is finding the best method to impute long-hour missing data gaps. Other methods such as Expectation Maximize and Multiple Imputation - Markov Chain Monte Carlo were used to compare the efficiency of time series methods, ARIMA. The simulations of the missing data are performed using the statistical software SPSS, E-views, and Microsoft Excel for Windows.

The simulation procedure involved several steps. The first step was using Microsoft Excel to juggle the simulation design numbers, ranging from minimum to maximum (24 hr $\leq$ L $\leq$120 hr). For example, the gap for the particulate matter data set in Pegoh ranges from 24 to 120 hours. At this stage, we need to list the hours from 24 (minimum) to 120 (maximum) and then use the command (=RANDBETWEEN(24,120)) in Microsoft Excel to randomize the range. This is

necessary to ensure that the numbers are not consecutive. The step is shown on figure 3.2 below.

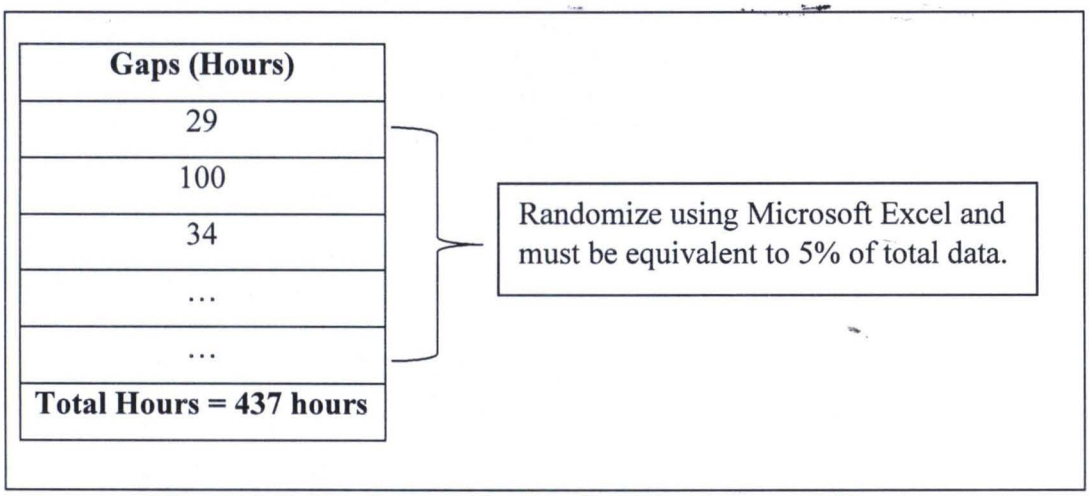


Figure 3.2: Hour gap randomization using Microsoft Excel

The next step is to add the randomized values and ensure they do not exceed the percentage of simulated missing data. For instance, 5% of simulated missing data will have approximately 433 to 438 hours were randomly deleted from the dataset. The third step is to use SPSS to randomize the order of the missing values in the data set. This mimics the random missing data in the air pollution data set. In this step, by using the feature Select Cases in SPSS to randomly select 1% of the total data of each parameter to select which data will be deleted. SPSS was randomly selected data throughout the dataset of each parameter, and the selected data was manually deleted. The deleted data was between 24 and 120 hours which was randomly selected using Microsoft Excel earlier which shown on figure 3.3 below. This step will be repeated for other parameters and percentages (10% and 15%).

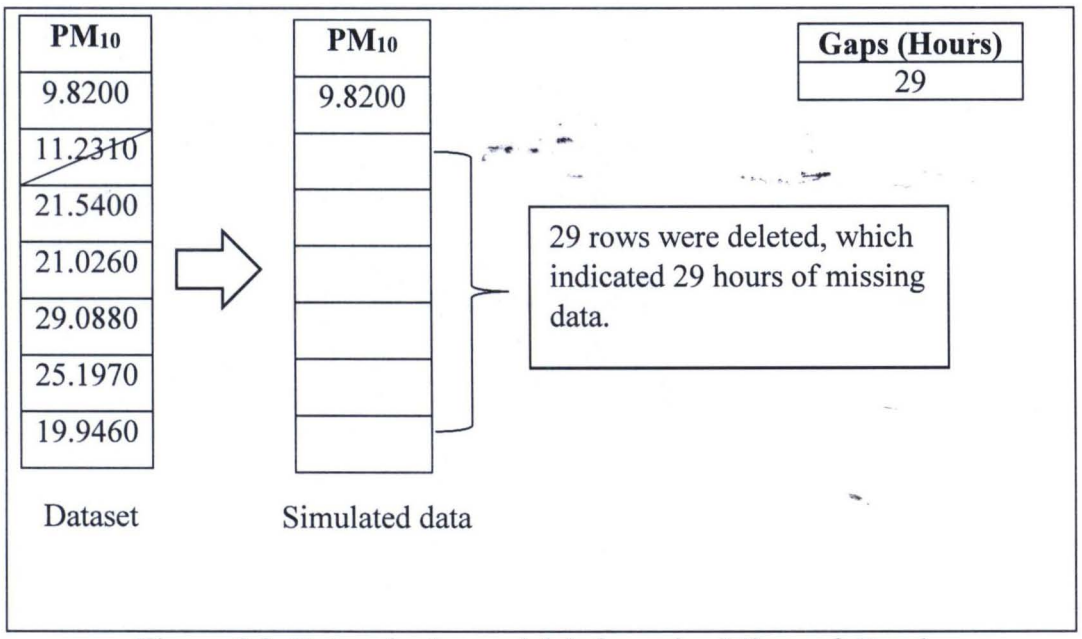


Figure 3.3: Data selection and deletion using Microsoft Excel

3.5 Imputation Method

Four imputation methods were used to fill in the simulated missing data. The methods used were Expectation-Maximization (EM), Multiple Imputation (MI), Autoregressive Moving Average (ARMA), and Autoregressive Integrated Moving Average (ARIMA). The multiple imputation and expectation-maximization methods were performed to compare the performances of the ARMA and ARIMA methods.

3.5.1 Autoregressive Integrated Moving Average (ARIMA)

Using the Box-Jenkins ARIMA model, the forecasting trend was generated by modelling the time series behavior. Autoregressive Integrated Moving Average (ARIMA) and each term represent one stage in the model generation process until only random noise remains (Guarnaccia, et al., 2018). A four-step iterative procedure was utilized in this study's methodology. In the first stage, the historical data are utilized to identify an appropriate Box-Jenkins model provisionally. It is then followed by the estimation of the model's parameters. After that, a diagnostic check must be performed to ensure that the detected model is adequate before

deciding on the finest one. If the model is insufficient, a new one should be found. It is then used to calculate the value of a time series forecast (Ye, 2019).

It is possible to anticipate the next few points in time using the AR(p) model because of the correlation between time series variables, which is why this model is named AR(p) (Lee, Rahman, Suhartono, Nor, & Kamisan, 2012).

$$Y_t = \epsilon_t + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots + \beta_p Y_{t-p} \quad (3.2)$$

AR model can be replaced with the lower order MA(q) model in this case (Lee, Rahman, Suhartono, Nor, & Kamisan, 2012).

$$Y_t = \epsilon_t + \beta_1 Y_{t-1} \quad (3.3)$$

It is essential to determine the best ARIMA order (p, d, q) and seasonal ARIMA order for air pollution prediction (P, D, Q, S). Using the grid search approach, different combinations of parameters can be tested iteratively.

3.5.1.1 Stationary Test

Checking if the time series variables are stationary is crucial before estimating any model. If our series are non-stationary, the models we estimate may result in spurious results and wrong conclusions. The stationary test was tested using a correlogram in Eviews. Correlograms can provide a decent indication of whether or not data sets indicate autocorrelation. A correlogram provides an overview of correlation at different time intervals. Figure 3.4 below shows the example for non-stationary data using the 5% simulated missing data of PM_{10} in Kota Kinabalu. There is fluctuation in the Autocorrelation and the partial correlation graph, which indicate the data is non-stationary. The second test uses the Augmented Dickey-Fuller unit root test (ADF) to see whether the data has a unit root or not. If the data has a unit root, as shown in figure 3.5 below, the PM_{10} data has a unit root, meaning the data is non-stationary.

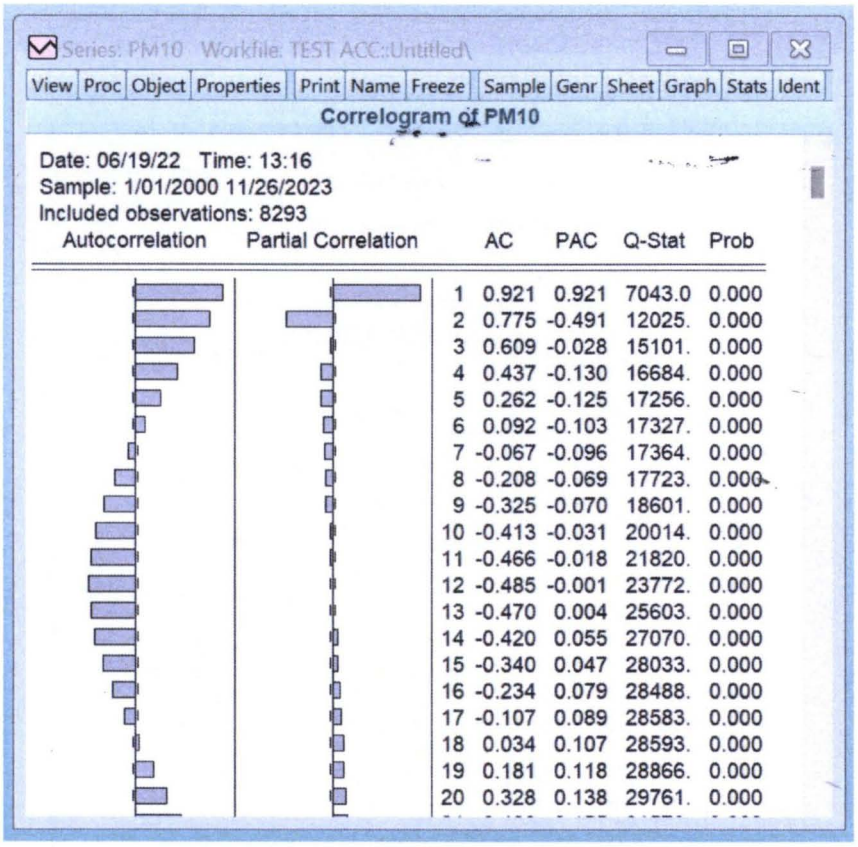


Figure 3.4: Correlogram test for PM_{10} data for 5% simulated missing data in Kota Kinabalu.

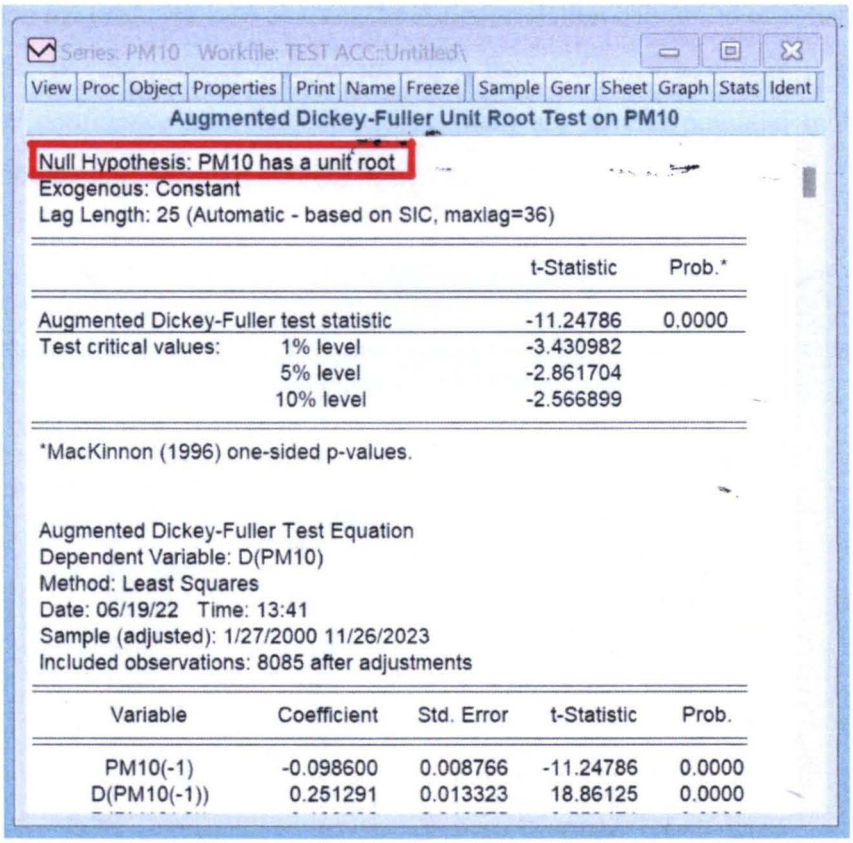


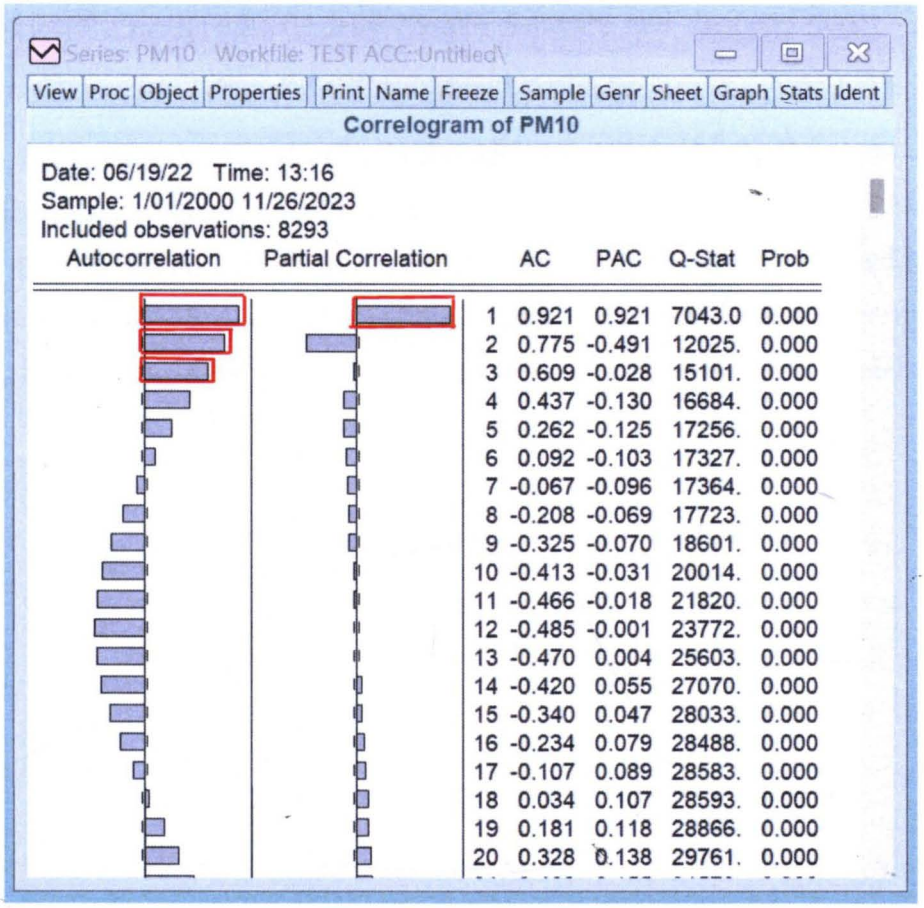
Figure 3.5: Augmented Dickey-Fuller unit root test (ADF) for 5% simulated missing data in Kota Kinabalu.

3.5.1.2 ARIMA: Determine the Value of "p" and "q"

This stage seeks to identify all potential estimation models. Focus on the correlogram of the initial differences to establish the order of the autoregressive and moving average components. These differences present the correlogram because they are stationary in the initial differences.

Next is to observe the partial autocorrelation function (PACF) ("p") column in order to establish the order of the autoregressive component. On the sides of the column is a confidence interval. The numbers that surpass the range indicate the likely order of the autoregressive component. The first lag represents a critical AR(1) component. In contrast, the second and third delays are on the line and may be evaluated. In figure 3.6 below, this study used an AR(1) component.

Then the next step is to observe the Autocorrelation column to establish the order of the moving average component ("q"), then (ACF). Where lags 1 and 3 surpass the confidence intervals. Therefore, there are several potential moving average components, MA (1), MA (2) or MA (3) and so on.



	AC	PAC	Q-Stat	Prob
1	0.921	0.921	7043.0	0.000
2	0.775	-0.491	12025.	0.000
3	0.609	-0.028	15101.	0.000
4	0.437	-0.130	16684.	0.000
5	0.262	-0.125	17256.	0.000
6	0.092	-0.103	17327.	0.000
7	-0.067	-0.096	17364.	0.000
8	-0.208	-0.069	17723.	0.000
9	-0.325	-0.070	18601.	0.000
10	-0.413	-0.031	20014.	0.000
11	-0.466	-0.018	21820.	0.000
12	-0.485	-0.001	23772.	0.000
13	-0.470	0.004	25603.	0.000
14	-0.420	0.055	27070.	0.000
15	-0.340	0.047	28033.	0.000
16	-0.234	0.079	28488.	0.000
17	-0.107	0.089	28583.	0.000
18	0.034	0.107	28593.	0.000
19	0.181	0.118	28866.	0.000
20	0.328	0.138	29761.	0.000

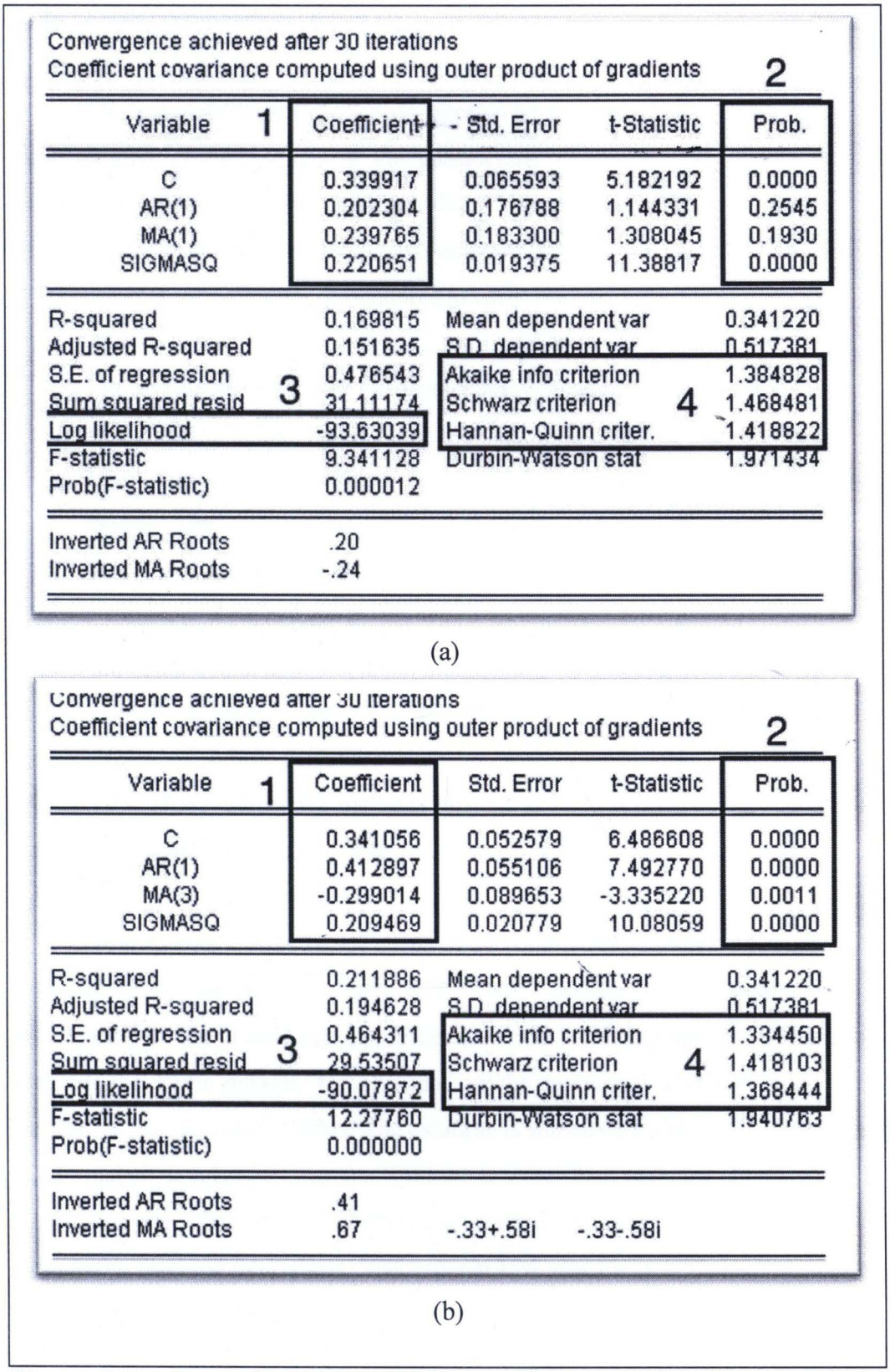
Figure 3.6: Correlogram to determine the "p" and "q"

3.5.1.3 Estimation

Once the potential ARIMA models have been identified, the next step is estimating and selecting the best appropriate model. For example, ARIMA (1,1,1) and ARIMA (1,1,3) were the models to estimate. The Box Jenkins Method's first step is to estimate the identified models. Then, choose a model based on the importance of the estimated coefficients and model criteria such as Schwartz, Akaike, and Hannan-Quinn. The model with the fewest model criteria values and the highest value coefficient will be the most suitable. Several things should be kept in mind:

- Significance of the ARMA definitions: Select the model with the highest proportion of significant terms (p-values 0.05).
- SigmaSQ: is a volatility measure. Choose the smallest option.
- Log-Likelihood: choose the highest number to optimize the log-likelihood function. (The bigger value is the least negative)
- Select the model with the least Akaike, Schwarz, and Hannan-Quinn distances.

Figure 3.7 compares the ARIMA (1,1,1) and ARIMA models (1,1,3). By examining the autocorrelation function (ACF) and partial autocorrelation function (PACF) plots of the differenced series, it is possible to estimate the number of AR and MA terms. In the figure we can clearly see that model B has better and smaller value compare to model A. As a conclusion, Model B has a better match than Model A.



Convergence achieved after 30 iterations
Coefficient covariance computed using outer product of gradients

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.339917	0.065593	5.182192	0.0000
AR(1)	0.202304	0.176788	1.144331	0.2545
MA(1)	0.239765	0.183300	1.308045	0.1930
SIGMASQ	0.220651	0.019375	11.38817	0.0000

R-squared	0.169815	Mean dependent var	0.341220
Adjusted R-squared	0.151635	S.D. dependent var	0.517381
S.E. of regression	0.476543	Akaike info criterion	1.384828
Sum squared resid	31.11174	Schwarz criterion	1.468481
Log likelihood	-93.63039	Hannan-Quinn criter.	1.418822
F-statistic	9.341128	Durbin-Watson stat	1.971434
Prob(F-statistic)	0.000012		

Inverted AR Roots	.20
Inverted MA Roots	-.24

(a)

Convergence achieved after 30 iterations
Coefficient covariance computed using outer product of gradients

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.341056	0.052579	6.486608	0.0000
AR(1)	0.412897	0.055106	7.492770	0.0000
MA(3)	-0.299014	0.089653	-3.335220	0.0011
SIGMASQ	0.209469	0.020779	10.08059	0.0000

R-squared	0.211886	Mean dependent var	0.341220
Adjusted R-squared	0.194628	S.D. dependent var	0.517381
S.E. of regression	0.464311	Akaike info criterion	1.334450
Sum squared resid	29.53507	Schwarz criterion	1.418103
Log likelihood	-90.07872	Hannan-Quinn criter.	1.368444
F-statistic	12.27760	Durbin-Watson stat	1.940763
Prob(F-statistic)	0.000000		

Inverted AR Roots	.41		
Inverted MA Roots	.67	-.33+.58i	-.33-.58i

(b)

Figure 3.7: Comparison on ARIMA model, (a) ARIMA(1,1,1) and (b) ARIMA(1,1,3)

3.5.2 Expectation-Maximization

The expectation-maximization approach replaces missing data with the value obtained from estimating the parameters of an incomplete data set by maximizing the probability of known data. This method consists of two steps: prediction and estimation by iterative calculation (Ghapor, Zubairi, & Imon, 2017). SPSS performs the following procedures to carry out this method:

1. The mean, variance, and covariance are estimated from the whole data on an individual basis.
2. Maximum likelihood algorithms are used to estimate regression equations that tie each variable and construct the formula.
3. The formula is used to estimate missing values.

Expectation maximize is an efficient method often applied in data analysis to deal with missing data. Indeed, expectation-maximization solves the shortcomings of other methods, such as mean substitution and regression substitution. These alternative methods provide skewed results; in particular, they underestimate standard errors. Maximizing expectations resolves this issue.

To implement this strategy using SPSS for this study, select "Missing Value Analysis" from the Analyze menu first. Then Transfer all quantitative factors associated with the investigation or issue to the box labeled Quantitative Variables. Transfer the PM_{10} dataset, for example, to the numerical variables box. Then, choose the EM checkbox option. The example is shown on Figure 3.8 below. Select Save finished data by pressing the EM button. After that, we must choose to Write a new data file. Press the File button and enter a filename. This new file should include the imputed data. Click "OK" to run EM, and the new dataset of imputed data appears in a new window.

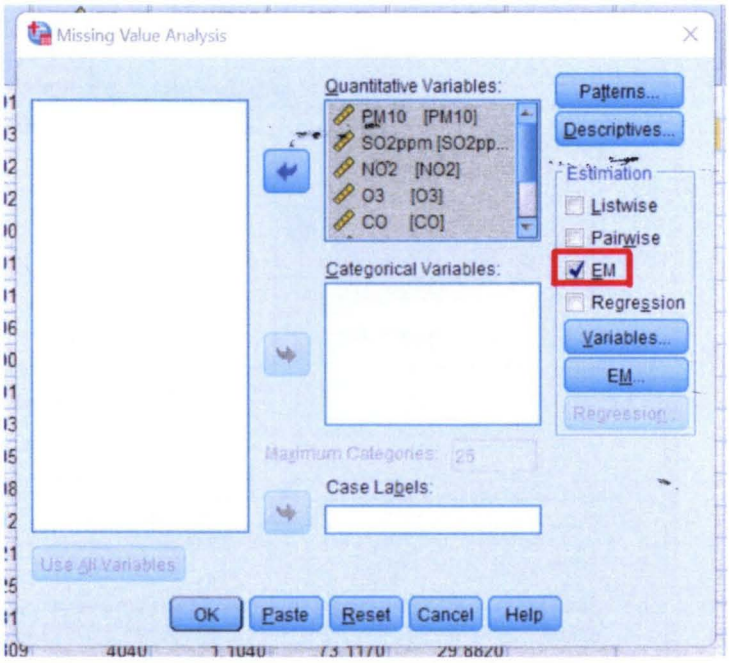


Figure 3.8: EM interface in SPSS

3.5.3 Markov Chain Monte Carlo (Multiple Imputation)

Multiple imputation was performed using the Markov Chain Monte Carlo (MCMC) method due to the assumption of multivariate normality. MCMC is a sequence of random variables whose distributions depend on the value of the previous variable (Suhaimi, Ghazali, Nasir, Mokhtar, & Ramli, 2017). MCMC is a simulation technique that can be used to determine and sample from the likelihood function.

To start performing the multiple imputation of MCMC method involves navigating to Analyze -> Multiple Imputation -> Impute Missing Data Values. The Variables window is the first window. Then, a window with four tabs, Variables, Method, Constraints, and Output, appears as shown on figure 3.9 below. Here, the complete and incomplete variables that need to include in the imputation model must be transferred from the dataset to the "Variables in Model" window—for example, the PM_{10} dataset.

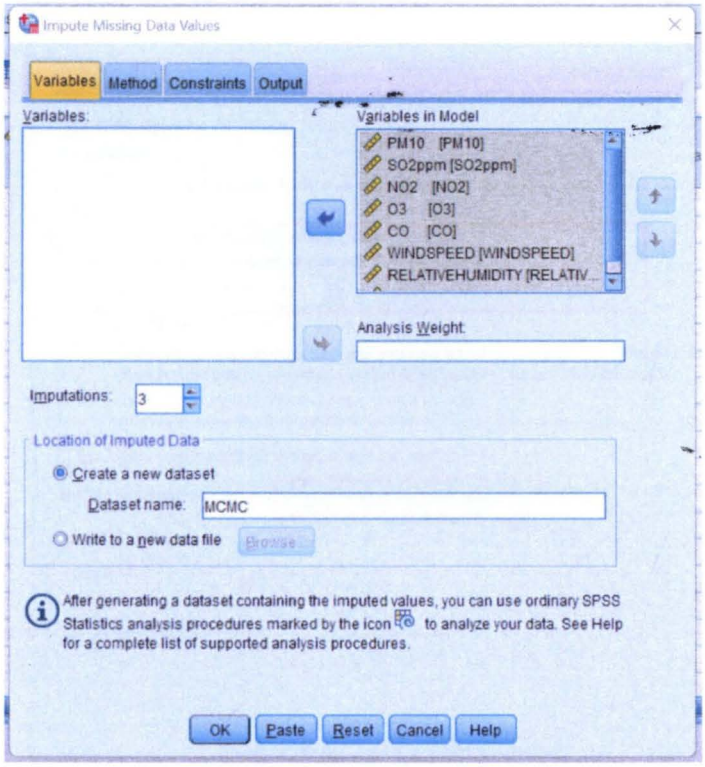


Figure 3.9: Variables tab interface to input data for computing MCMC

After that, the number of imputed datasets may be adjusted to three in the "Imputations" box. Choose a name for the dataset to which the imputed data values will be stored so that the freshly imputed data will be saved in a new file. Select "Custom" under Imputation Method and "Fully conditional specification (MCMC)" next to the Method which shown on figure 3.10. Predictive Mean Matching (PMM) is selected as the Model type for scale (continuous) variables. PMM is the default approach to impute continuous variables in this feature. In SPSS, the default approach is linear regression. It is preferable to replace it with PMM.

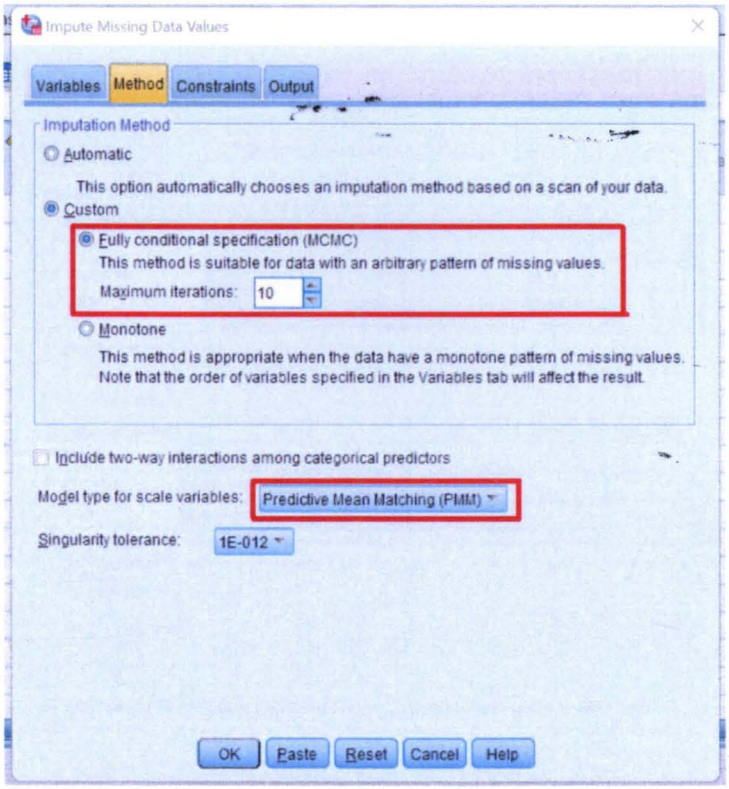


Figure 3.10: Methods tab interface to input data for computing MCMC

Enter the lowest and maximum permissible imputed values for continuous variables in the dataset next to the "Constraints" tab. To acquire the current range of variable values, click the "Scan" button which shown on figure 3.11 below; these values can then be modified. The Constraints tab can be bypassed when the Method Tab selects the PMM method. Do not tick the select "Exclude variables with substantial amounts of missing data" which will limit the analysis to variables with fewer than the maximum percentage of missing values.

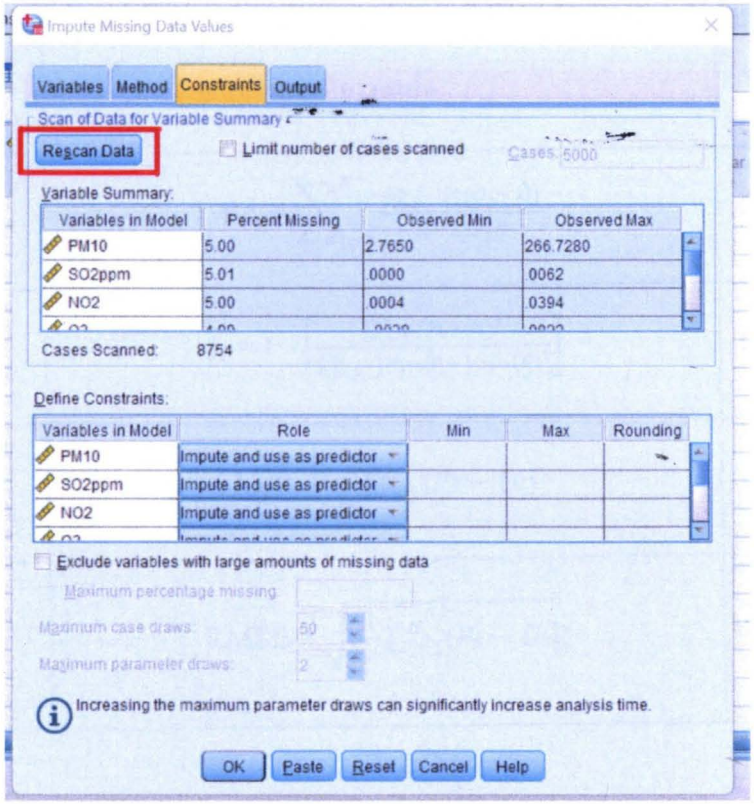


Figure 3.11: Constraints tab interface to input data for computing MCMC

Next on the Output tab, pick the produced output. By choosing "Imputation model" and "the Descriptive statistics for variables with imputed values" in the Output tab, descriptive statistics of imputed variables may be retrieved. At each iteration, the dataset includes the means and standard deviations of the imputed scale variables. Then press "OK" to run the simulation and new window with treated data will appear.

3.6 Performance Indicators

The imputation methods were evaluated using four performance indicators: Prediction Accuracy (PA), Index of Agreement (d_2), Mean Absolute Error (MAE), and Root Mean Square Error (RMSE). Methods for estimating missing values were evaluated using both theoretical and raw data.

Table 3.2: Performance Indicators (Noor, Yahaya, & Abdullah, Estimation of missing values in air pollution data using single imputation techniques, 2008)

Performance Indicator	**Formula**	**Best Fit**
Prediction Accuracy (PA)	$PA = \sum_{i=1}^{N} \frac{[(P_i - \dot{P})(O_i - \bar{O})}{(N-1)\sigma_p\sigma_o}$	Close to 1
Index of Agreement (d_2)	$d_2 = 1 - \left[\frac{\sum_{i=N}^{N}(P_i - O_i)^2}{\sum_{i=N}^{N}(\lvert P_i - \bar{O}\rvert + \lvert O_i - \bar{O}\rvert)^2}\right]$	Close to 1
Mean Absolute Error (MAE)	$MAE = \frac{1}{N}\sum_{i=1}^{N}\lvert P_i - O_i\rvert$	Close to 0
Root Mean Square Error (RMSE)	$RMSE = \sqrt{(\frac{1}{N}\sum_{i=1}^{N}\lvert P_i - O_i\rvert)}$	Close to 0

Where,

N = Number of imputations

O_i = Observed data points

P_i = Imputed data points

$\dot{P}$ = Average of imputed data

$\bar{O}$ = Average of observed data

σ_o = Population standard deviation of the observed data

σ_p = Population standard deviation of the imputed data

CHAPTER 4

RESULTS AND DISCUSSION

In this chapter, the outcomes of this investigation are described and discussed. This chapter is subdivided into three sections for the most part. The first section of this chapter covered raw air quality data characteristics. The second part show the characteristic and descriptive statistics analysis of simulated missing data for Pegoh and Kota Kinabalu. Finally, the effectiveness of imputation techniques applied to the simulated missing data was explored.

4.1 Raw Data Characterization

Table 4.1 shows the summary descriptive statistic of raw air pollution data of Pegoh, Perak, and Kota Kinabalu, Sabah in 2018. The mean values for PM_{10} and relative humidity were highest in both locations than in the other parameters. The standard deviation indicates the dispersion of data from the mean value, or, in other words, the variability of the data. PM_{10} concentrations in Pegoh (15.715 $\mu g/m^3$) and Kota Kinabalu (18.0751 $\mu g/m^3$) had the most extensive standard deviations in both study regions. The range represents the difference between the dataset's highest and lowest values. Since PM_{10} concentration variation was the most significant, PM_{10} concentration variation was the greatest.

Figure 4.1 shows the box and whisker plot for air pollution pollutant parameters in Pegoh, Perak, and Kota Kinabalu, Sabah. For instance, the PM_{10} extreme outliers in Kota Kinabalu were much higher than in Pegoh, yet the box shapes were nearly comparable in size. The upper whisker represents the maximum value in the data, while the lower whisker represents the minimum. The length of whiskers varies between the two locations. For example, the upper whisker of O_3

Pegoh shows a higher upper whisker than the upper whiskers of O_3 Kota Kinabalu, which indicates that O_3 Pegoh has a higher maximum value than O_3 Kota Kinabalu. While the lower whisker for O_3 Pegoh was longer to the bottom, O_3 Pegoh has a lower minimum value than Kota Kinabalu.

In figure 4.1 on relative humidity for both areas are skewed to the left, indicating that most of the time was the dryer season during 2018. The average humidity readings for Pegoh and Kota Kinabalu are 77.13 % and 88.15 %, respectively. The leftward skew of the relative humidity box plots for both areas in 2018 suggests a regional dry season. As reported by weatherspark.com, the probability of rainy days fluctuates dramatically throughout the year in Kota Kinabalu. From April 27 to January 15, there is a greater than 39 % probability of precipitation on any day during the wetter season. November is the wettest month in Kota Kinabalu, with 16.9 days with at least 1 millimeter of precipitation on average. The dry season is 3 to 4 months long, lasting from January 15 to April 27. March is the month with the lowest average number of days with at least 1 millimeter of precipitation in Kota Kinabalu (Cedar Lake Ventures, 2018). For PM_{10}, the boxplot in figure 4.1 is skewed to the right, indicating that the mean and median were almost identical, and the box plot will be symmetric. The value distribution is narrow; however, the highest PM_{10} concentrations in Pegoh and Kota Kinabalu are 266.728 $\mu g/m^3$ and 494.278 $\mu g/m^3$. Kota Kinabalu had the highest expression in PM_{10} concentrations. Their geographical location and population density might account for this.

Table 4.1: Descriptive Statistical summary for both Pegoh and Kota Kinabalu (2018)

Location	Parameters	Data Count		Descriptive Statistics				
		Valid Data	Missing Data	Mean	Minimum Value	Maximum Value	Range	Standard Deviation
Pegoh, Perak	PM_{10} ($\mu g/m^3$)	8610	144	33.8698	2.765	266.728	263.963	15.715
	SO_2 (ppm)	8266	488	0.0009	0	0.0062	0.0062	0.0005
	NO_2 (ppm)	8291	463	0.0096	0.0004	0.0394	0.039	0.0049
	O_3 (ppm)	8296	458	0.0189	0	0.0822	0.0822	0.0169
	CO (ppm)	8259	495	0.601	0.042	2.509	2.467	0.2175
	Wind Speed (m/s)	8650	104	1.2423	0	23.976	23.976	0.6814
	Relative Humidity (%)	8662	92	77.1308	32.5	97.883	65.383	13.6794
	Ambient Temperature (°c)	8614	140	28.8384	20.73	38.179	17.449	3.3594
Kota Kinabalu, Sabah	PM_{10} ($\mu g/m^3$)	8435	296	23.2551	0	494.278	494.278	18.0751
	SO_2 (ppm)	7805	926	0.0012	0	0.0027	0.0027	0.00049
	NO_2 (ppm)	8233	498	0.0049	0	0.0528	0.0528	0.00477
	O_3 (ppm)	8312	419	0.0129	0	0.0508	0.0508	0.01049
	CO (ppm)	8053	678	0.5245	0.04	1.915	1.875	0.30168
	Wind Speed (m/s)	8655	76	1.2114	0.032	8.117	8.085	0.87569
	Relative Humidity (%)	8708	23	81.1511	2.407	99.00	96.593	11.4614
	Ambient Temperature (°c)	8576	155	27.8464	21.85	36.439	14.589	3.0392

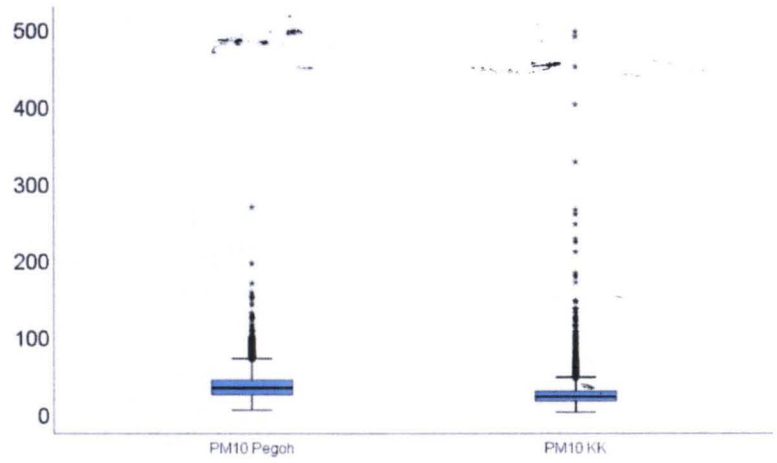
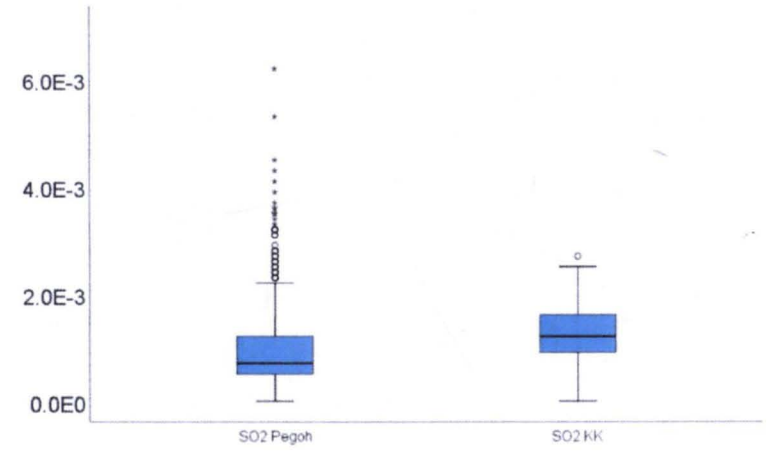
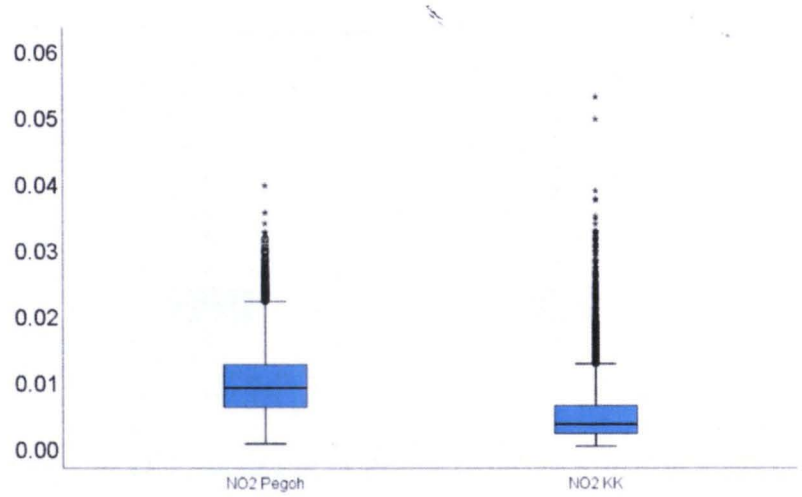
Parameters	Boxplot
PM_{10} ($\mu g/m^3$)	
SO_2(ppm)	
NO_2(ppm)	

Figure 4.1: The boxplot for each parameter in Pegoh and Kota Kinabalu (2018)

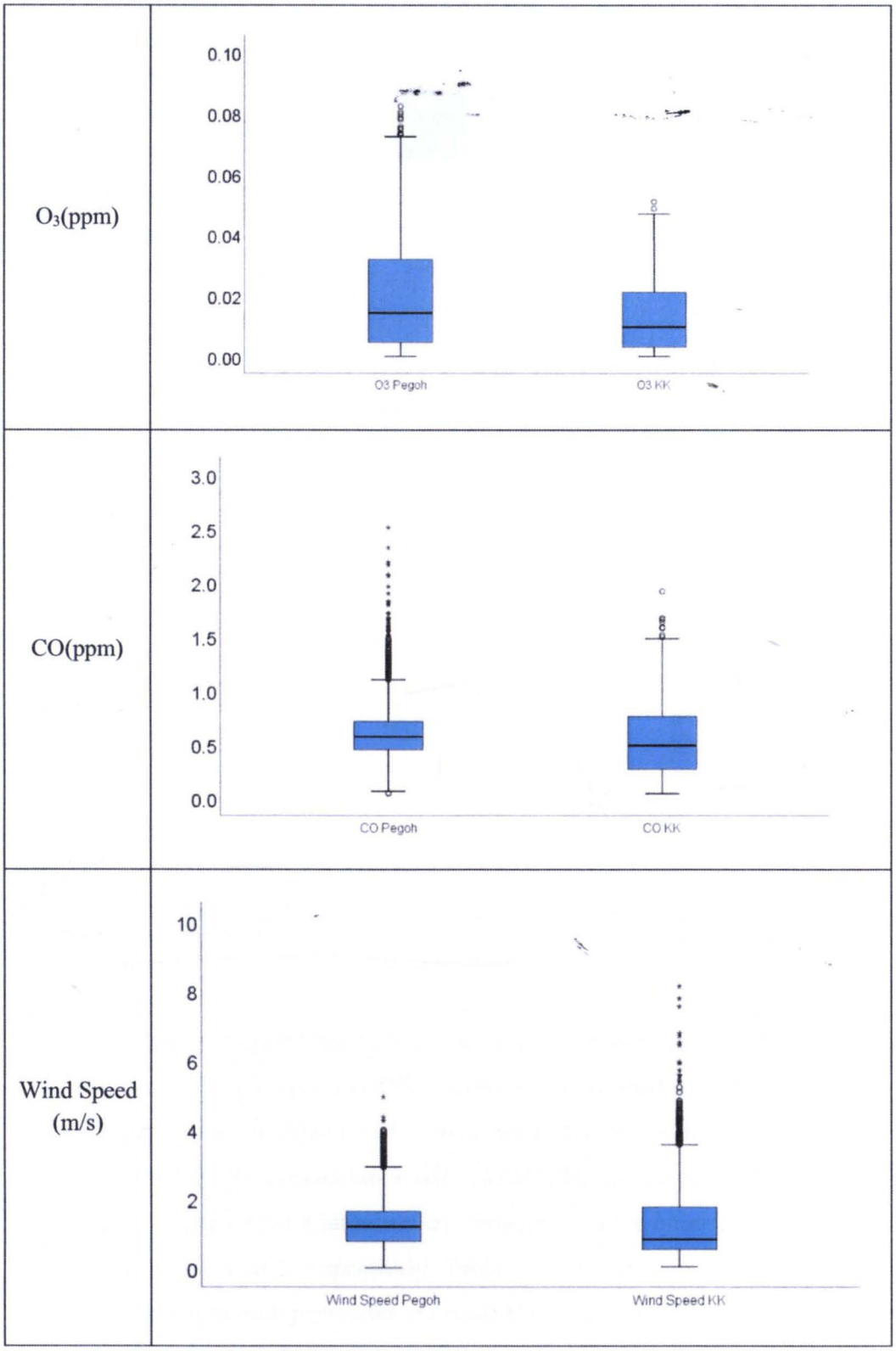


Figure 4.1: The boxplot for each parameter in Pegoh and Kota Kinabalu (2018) (continued)

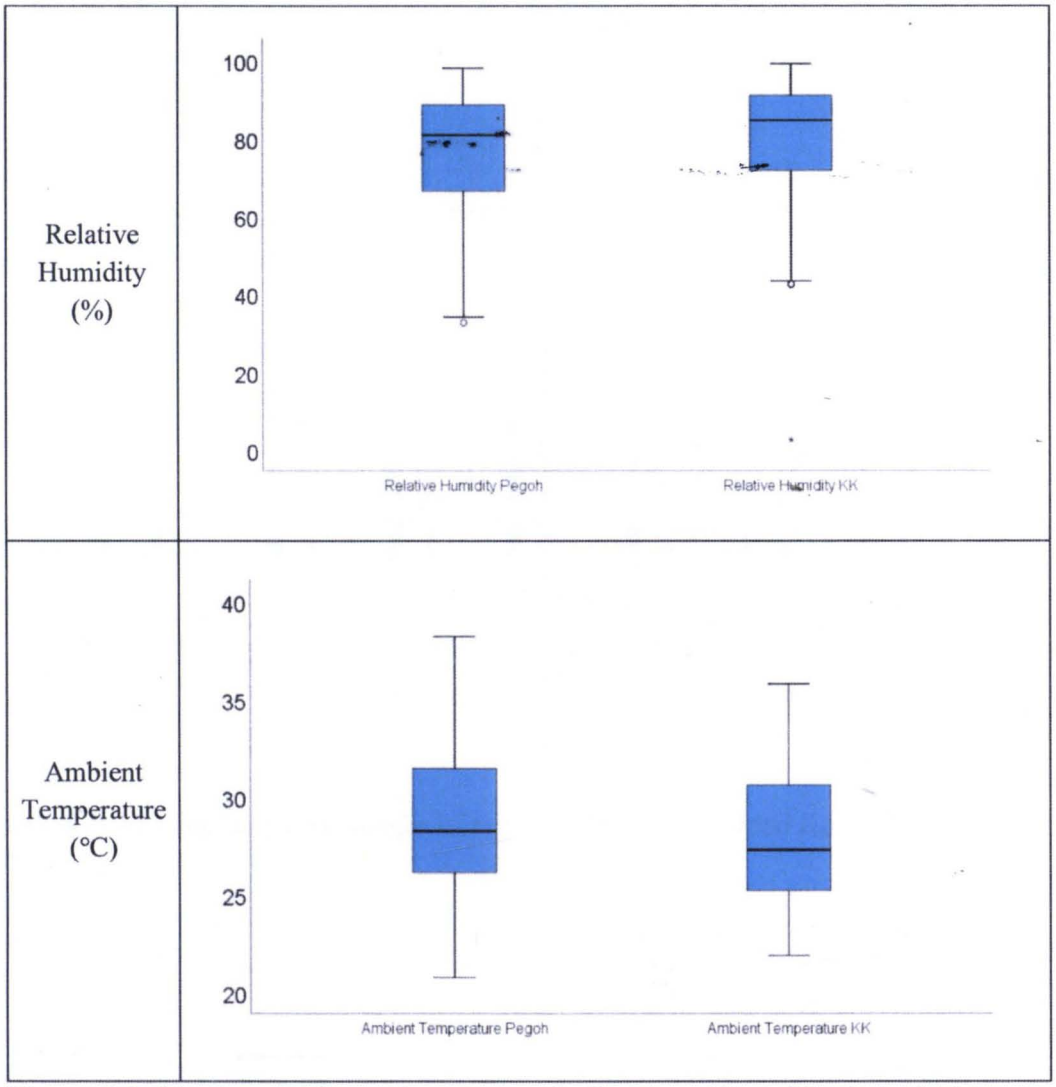


Figure 4.1: The boxplot for each parameter in Pegoh and Kota Kinabalu (2018) (continued)

Figure 4.2 shows the missing percentages of raw data obtained from the Department of Environment (DOE) Malaysia. The highest percentage of missing data at Pegoh was recorded by CO with a value of 5.65%, whereas Kota Kinabalu was recorded by SO_2 concentration with 10.61%. The lowest percentage of missing data at Pegoh and Kota Kinabalu were shown in relative humidity measurements with 1.05% and 0.26%, respectively. Table 4.2 and Table 4.3 show the length of gaps according to each parameter in detail. The data show that the longest gap for Pegoh is 24 hours and 31 hours for Kota Kinabalu. This is insufficient for this study which requires long gaps of missing data that require more than 24 hours and no more than 120 hours for a time series method study. Besides that, most of the missing gaps are 1-hour for both places, 23.78% and 18.45% of total percentages

for Pegoh and Kota Kinabalu, respectively. This data needs to be treated first before running a missing data.

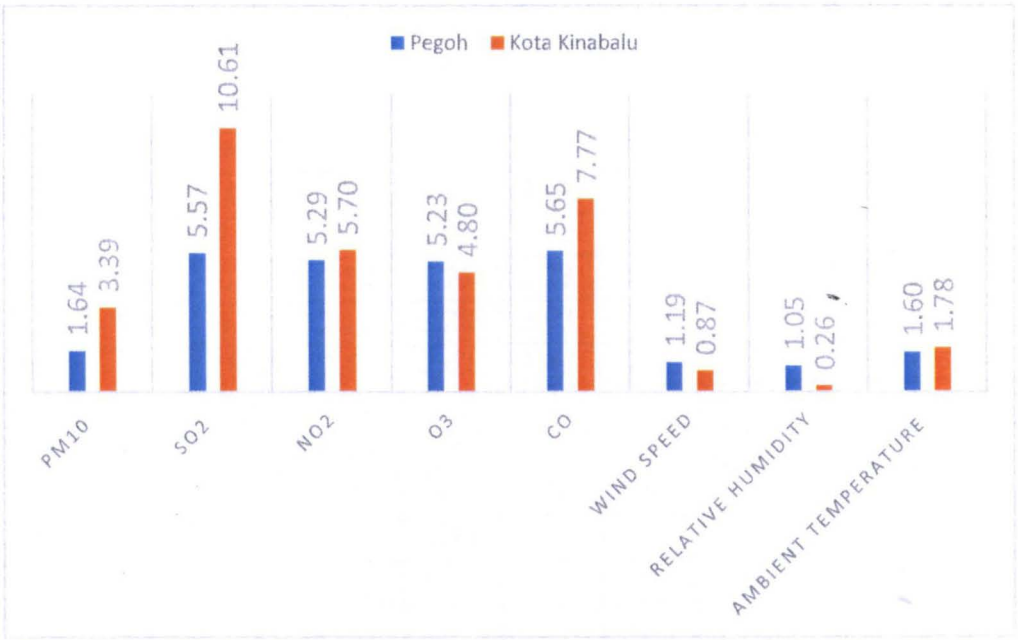


Figure 4.2: Missing data percentages in raw dataset for Pegoh and Kota Kinabalu (2018)

Table 4.2: Percentages of missing data gaps for Pegoh, Perak

Length of Gap (Hours)	Percentages of Missing Data in Gaps for Pegoh, Perak (%)									
	PM_{10}	SO_2	NO_2	O_3	CO	Wind Speed	Relative Humidity	Ambient Temperature	Mean	Total Percentage (%)
1	22.22	68.65	77.97	77.29	73.13	9.62	2.17	2.88	41.74	23.78
2	6.94	18.85	2.16	4.37	4.44	1.92	2.17	1.43	5.29	3.01
3	6.25	1.23	-	1.97	3.03	2.88	-	-	3.07	1.75
4	19.44	0.82	2.59	1.75	3.23	3.85	4.35	5.71	5.22	2.97
5	3.47	1.02	2.16	3.28	3.03	-	-	-	2.59	1.48
6	4.17	18.85	2.59	2.62	2.42	5.77	6.52	4.29	5.90	3.36
7	-	1.43	1.51	1.53	1.41	-	-	-	1.47	0.84
8	5.56	-	1.73	1.75	-	-	-	-	3.01	1.72
9	-	12.50	3.89	-	3.64	-	-	6.43	6.61	3.77
10	-	-	-	2.18	-	-	-	-	2.18	1.24
11	-	-	-	-	-	-	-	-		
12	8.33	-	-	-	-	-	-	-	8.33	4.75
13	-	-	-	-	2.63	-	-	-	2.63	1.50
14	-	-	-	-	-	-	-	10.00	10.00	5.70
15	0.69	10.42	3.24	3.28	3.03	14.42	16.30	10.71	7.76	4.42
16	-	-	-	-	-	-	-	-		
17	-	-	-	-	-	16.35	18.48	12.14	15.66	8.92
18	-	-	-	-	-	-	19.57	12.86	16.21	9.24
19	13.19	-	-	-	-	18.27	-	-	15.73	8.96
20	-	-	-	-	-	-	-	-		
21	-	-	-	-	-	-	-	-		
22	-	-	-	-	-	-	-	-		
23	-	-	-	-	-	-	-	-		
24	-	-	-	-	-	23.08	26.09	17.14	22.10	12.59
Total	100	100	100	100	100	100	100	100	175.5	100

Table 4.3: Percentages of missing data gaps for Kota Kinabalu, Sabah

	Percentages of Missing Data in Gaps for Kota Kinabalu, Sabah (%)									
Length of Gap (Hours)	PM_{10}	SO_2	NO_2	O_3	CO	Wind Speed	Relative Humidity	Ambient Temperature	Mean	Total Percentage (%)
1	22.64	7.24	78.11	95.47	54.57	11.84	26.09	4.52	37.56	18.45
2	19.59	61.12	10.04	1.43	6.49	2.63	-	-	16.89	8.29
3	14.19	3.89	4.22	0.72	2.21	3.95	17.39	2.58	6.14	3.02
4	9.46	2.59	0.8	0.95	7.08	15.79	13.04	1.94	6.46	3.17
5	6.76	1.62	3.01	-	2.21	6.58	-	-	4.04	1.98
6	19.59	2.59	-	1.2	1.77	-	-	-	6.29	3.09
7	-	3.02	1.41	-	2.06	9.21	-	-	3.93	1.93
8	-	0.86	-	-	1.18	-	-	-	1.02	0.5
9	-	-	-	-	1.33	-	-	-	1.33	0.65
10	-	3.24	-	-	1.47	13.16	43.48	6.45	16.95	8.33
11	-	-	-	-	-	14.47	-	-	14.47	7.11
12	-	2.59	2.41	-	-	-	-	-	2.5	1.23
13	4.39	1.4	-	-	-	-	-	-	2.9	1.42
14	-	-	-	-	-	-	-	-		
15	-	-	-	-	-	-	-	-		
16	-	-	-	-	-	-	-	10.32	10.32	5.07
17	-	-	-	-	4.06	-	-	-	4.06	1.99
18	-	-	-	-	4.3	-	-	-	4.3	2.11
19	-	-	-	-	-	-	-	12.26	12.26	6.02
20	-	2.16	-	-	-	-	-	-	2.16	1.06
21	-	-	-	-	-	-	-	-		
22	7.43	-	-	-	-	-	-	-	7.43	3.65
23	-	2.48	-	-	-	-	-	-	2.48	1.22
24	-	2.59	-	-	5.73	-	-	61.94	23.42	11.5
25	-	-	-	-	-	-	-	-		
26	-	-	-	-	6.21	-	-	-	6.21	3.05
27	-	-	-	-	-	-	-	-		
28	-	-	-	-	-	-	-	-		
29	-	-	-	-	-	-	-	-		
30	-	-	-	-	-	-	-	-		
31	10.47	-	-	-	-	-	-	-	10.47	5.14
Total	100	100	100	100	100	100	100	100	203.6	100

4.2 Simulated Missing Data

Table 4.4 displays the missing data gap (hours) for each simulated missing data gap pattern for Pegoh and Kota Kinabalu. The length of gaps in missing data was shown as a percentage. The missing data gap were simulated as 5%, 10%, and 15%, and the gaps in the simulated missing data ranged from 24 to 120 hours. The highest distribution of missing gaps in a 5% simulated missing data pattern for Pegoh, Perak was around 35.97% of mean gaps for 72 to 96 hours, while the lowest distribution was 16.31% for 96 to 120 hours. In the 10% simulated missing data patterns, the maximum percentage of missing gaps was 36.88% for 72 to 96 hours of missing data, and the lowest percentage was 13.23% for 24 to 48 hours of missing data. The highest percentage of simulated missing data is 39.89% for 96 to 120 hours, and the lowest percentage is 15.23% for 24 to 48 hours. For total percentages of missing data for both Pegoh and Kota Kinabalu, the distribution of gaps (5%, 10%, and 15%) was relatively comparable. The highest distribution of missing gaps for Pegoh was around 31.94% of total percentages for the 72 to 96 hours gap and 16.68% for the 24 to 48 hours gap. The maximum percentage of simulated missing gaps for Kota Kinabalu was around 33.67% for 96 to 120 hours, while the lowest percentage was 15.82% for 24 to 48 hours.

Table 4.4: Percentages of length of gap simulated missing data in Pegoh and Kota Kinabalu

Length of Gap (Hour)	Simulated Missing Data Length (%)			Mean (%)	Total Percentage (%)
	5%	10%	15%		
Pegoh					
$24 \leq L \leq 48$	22.78	13.23	14.04	15.23	16.68
$48 < L \leq 72$	24.94	16.21	23.10	21.11	21.42
$72 < L \leq 96$	35.97	36.88	22.96	29.77	31.94
$96 < L \leq 120$	16.31	33.68	39.89	33.89	29.96
Total	100	100	100	100	100
Kota Kinabalu					
$24 \leq L \leq 48$	18.84	16.46	12.16	14.95	15.82
$48 < L \leq 72$	21.03	21.48	20.38	20.89	20.96
$72 < L \leq 96$	30.13	30.99	27.53	29.23	29.55
$96 < L \leq 120$	30.00	31.08	39.93	34.93	33.67
Total	100	100	100	100	100

Figure 4.3 show the difference of histogram in raw data and 5% of simulated missing data for all parameters at Pegoh, Perak. The distribution of data gaps was distinct in the simulated data. This scenario was produced by the random number generated by the SPSS software when simulating missing value patterns and the vast number of observations with the same range (Noor, Yahaya, & Abdullah, 2008). However, the outcomes of the simulation method produced a slightly different but still within the acceptable range. A minimal discrepancy between the histogram indicates that the simulated data is still acceptable. Since the sampling distribution is entirely random, the observations that will be eliminated are randomly distributed over the dataset. The approach will provide estimators that are impartial and standard errors that are accurate (Allison, 2002).

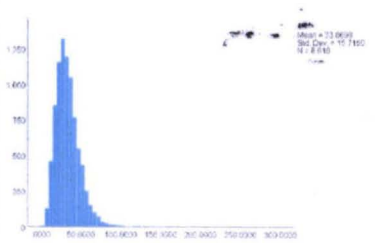
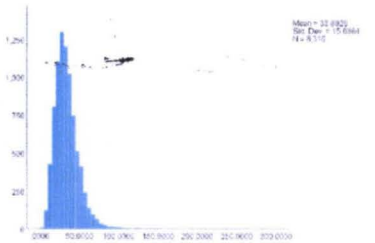
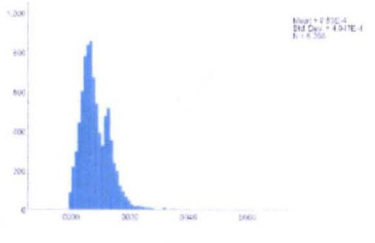
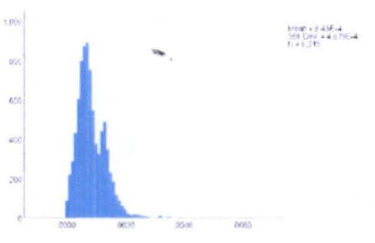
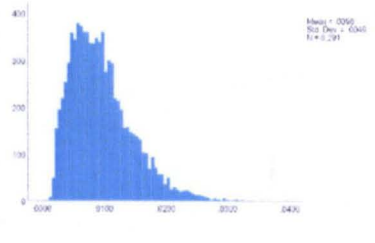
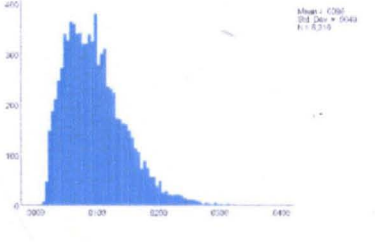
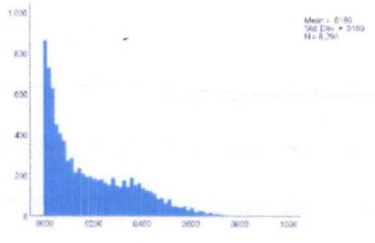
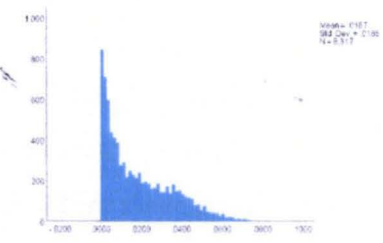
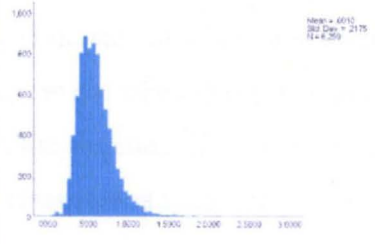
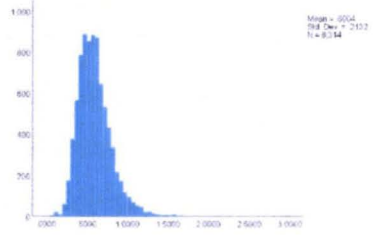
Parameters	Raw data histogram	Simulated Missing data (5%)
PM_{10}		
SO_2		
NO_2		
O_3		
CO		

Figure 4.3: The histogram of raw data and simulated missing data (5%) in Pegoh

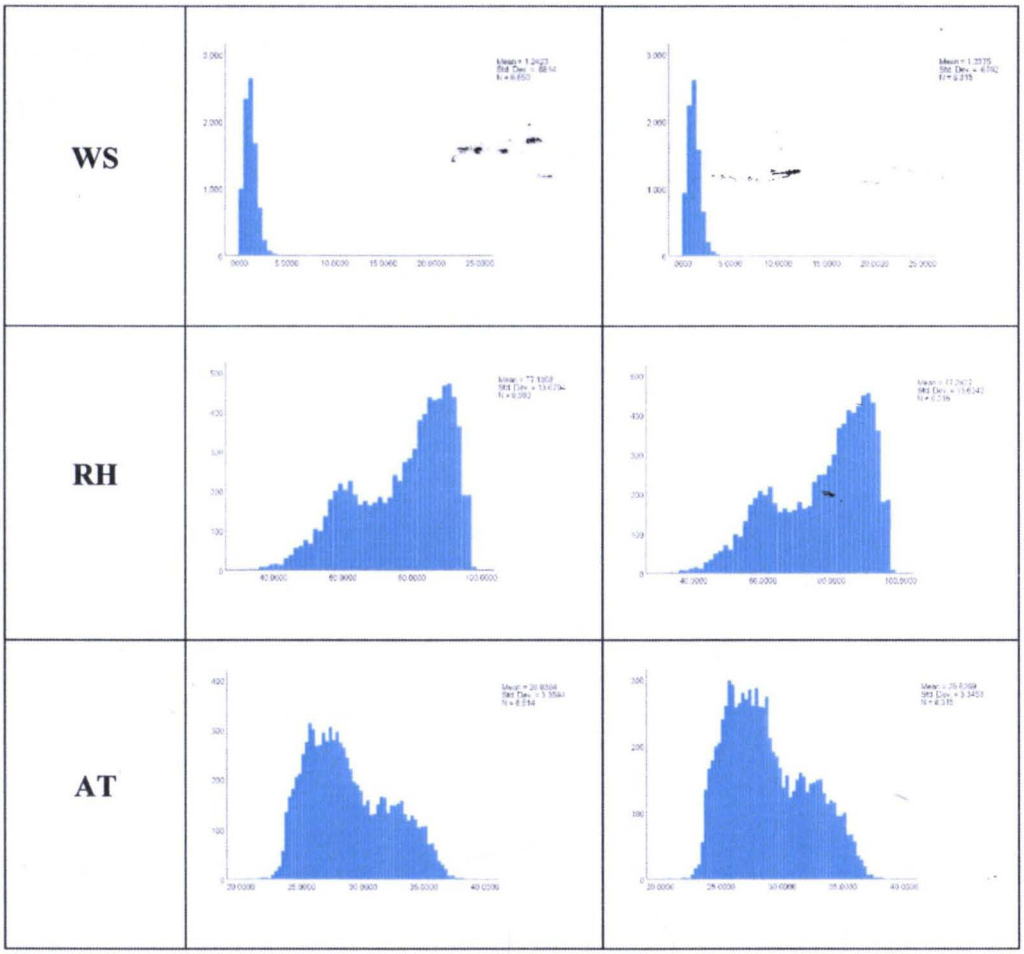


Figure 4.3: The histogram of raw data and simulated missing data (5%) in Pegoh (continued)

The descriptive statistics for the three different percentages of simulated missing data for Pegoh, Perak, and Kota Kinabalu, Sabah are presented in Tables 4.5 and 4.6. Furthermore, the patterns of descriptive statistics for simulated missing data did not show any significance from those raw data. There are relatively minor differences between the raw data and the three unique missing data simulations for both locations. The performance of the imputation process may be influenced by the structure of the simulated missing data. Even when the percentage of missing data increases, there are not many changes in the values of each percentage, as seen in the graph. When the structure of simulated missing data strays from its original structure, it will be classified as a separate dataset that is not identical to the original. This is because this method estimates missing numbers based on existing data.

According to these statistics (Tables 4.5 and 4.6), the mean values for all parameters were slightly greater than the median values. For both Pegoh and Kota

Kinabalu, PM_{10} was the parameter with the most significant value in mean, median, standard variation, minimum value and maximum value among the others parameters.

Table 4.5: Descriptive statistics of the simulated missing data for Pegoh, Perak

Parameter	%	Mean	Median	Std. Deviation	Range	Minimum	Maximum	Valid Data	Total Missing
PM_{10}	5%	33.8928	31.6430	15.6384	263.9630	2.7650	266.7280	8316	438
	10%	33.8672	31.6305	15.7204	263.9630	2.7650	266.7280	7878	876
	15%	33.8022	31.5885	15.6444	263.1340	3.5940	266.7280	7442	1312
SO_2	5%	0.0008	0.0007	0.0005	0.0062	0.0000	0.0062	8315	439
	10%	0.0008	0.0008	0.0005	0.0062	0.0000	0.0062	7886	868
	15%	0.0009	0.0008	0.0005	0.0062	0.0000	0.0062	7440	1314
NO_2	5%	0.0098	0.0091	0.0049	0.0390	0.0004	0.0394	8316	438
	10%	0.0098	0.0091	0.0049	0.0390	0.0004	0.0394	7880	874
	15%	0.0098	0.0091	0.0049	0.0390	0.0004	0.0394	7438	1316
O_3	5%	0.0187	0.0137	0.0165	0.0861	0.0000	0.0822	8317	437
	10%	0.0187	0.0136	0.0166	0.0861	0.0000	0.0822	7876	878
	15%	0.0186	0.0135	0.0166	0.0834	0.0000	0.0803	7439	1315
CO	5%	0.6004	0.5740	0.2132	2.4670	0.0420	2.5090	8314	440
	10%	0.6006	0.5740	0.2113	2.4670	0.0420	2.5090	7875	879
	15%	0.6018	0.5740	0.2145	2.4670	0.0420	2.5090	7438	1316
WS	5%	1.2375	1.1847	0.6762	23.9760	0.0000	23.9760	8315	439
	10%	1.2332	1.1770	0.6804	23.9440	0.0320	23.9760	7876	878
	15%	1.2530	1.1970	0.6849	23.9760	0.0000	23.9760	7438	1316
RH	5%	77.2827	80.8820	13.6242	65.3830	32.5000	97.8830	8316	438
	10%	76.8698	80.3335	13.6426	65.3830	32.5000	97.8830	7878	876
	15%	77.3360	81.1670	13.6229	65.3830	32.5000	97.8830	7439	1315
AT	5%	28.8269	28.2370	3.3453	17.4490	20.7300	38.1790	8315	439
	10%	28.8392	28.2680	3.3596	17.4490	20.7300	38.1790	7879	875
	15%	28.8453	28.2200	3.3583	17.4490	20.7300	38.1790	7440	1314

Where: % - percentage of simulation data, PM_{10} – particulate matter, SO_2 – sulfur dioxide, NO_2 – nitrogen dioxide, O_3 – ozone, CO – carbon monoxide, WS – wind speed, RH – relative humidity, AT – ambient temperature.

Table 4.6: Descriptive statistics of the simulated missing data for Kota Kinabalu, Sabah

Parameter	%	Mean	Median	Std. Deviation	Range	Minimum	Maximum	Valid Data	Total Missing
PM_{10}	5%	23.1035	19.9325	17.8382	493.6730	0.6050	494.2780	8296	435
	10%	23.2837	19.9920	18.1857	494.2780	0.0000	494.2780	7861	870
	15%	23.1390	19.9027	17.6079	494.2780	0.0000	494.2780	7418	1313
SO_2	5%	0.0012	0.0012	0.0005	0.0027	0.0000	0.0027	8292	439
	10%	0.0012	0.0012	0.0005	0.0027	0.0000	0.0027	7858	873
	15%	0.0012	0.0012	0.0005	0.0027	0.0000	0.0027	7424	1307
NO_2	5%	0.0049	0.0033	0.0047	0.0555	0.0000	0.0528	8294	437
	10%	0.0048	0.0033	0.0046	0.0555	0.0000	0.0528	7859	872
	15%	0.0048	0.0032	0.0047	0.0555	0.0000	0.0528	7422	1309
O_3	5%	0.0129	0.0106	0.0104	0.0539	0.0000	0.0508	8297	434
	10%	0.0128	0.0104	0.0103	0.0516	0.0000	0.0508	7857	874
	15%	0.0129	0.0106	0.0104	0.0516	0.0000	0.0508	7419	1312
CO	5%	0.5221	0.4807	0.2939	1.8750	0.0400	1.9150	8293	438
	10%	0.5085	0.4580	0.2918	1.8750	0.0400	1.9150	7861	870
	15%	0.5218	0.4790	0.2965	1.8750	0.0400	1.9150	7425	1306
WS	5%	1.1811	0.8170	0.8788	8.0850	0.0320	8.1170	8294	437
	10%	1.1853	0.8240	0.8687	8.0850	0.0320	8.1170	7859	872
	15%	1.1776	0.8190	0.8704	8.0850	0.0320	8.1170	7418	1313
RH	5%	81.1022	84.4500	11.4746	96.5930	2.4070	99.0000	8291	440
	10%	81.1601	84.4500	11.4421	96.5930	2.4070	99.0000	7854	877
	15%	80.9726	84.2320	11.4705	96.5930	2.4070	99.0000	7420	1311
AT	5%	27.7960	27.2250	3.0269	22.1718	22.0070	44.1788	8293	438
	10%	27.7639	27.1970	3.0329	22.3288	21.8500	44.1788	7855	876
	15%	27.7580	27.1740	3.0435	22.3300	21.8500	44.1800	7423	1308

Where: % - percentage of simulation data, PM_{10} – particulate matter, SO_2 – sulfur dioxide, NO_2 – nitrogen dioxide, O_3 – ozone, CO – carbon monoxide, WS – wind speed, RH – relative humidity, AT – ambient temperature.

4.3 The AR and MA

Everything outside the PACF plot's perimeter or boundary indicates the order of the Auto Regression (AR) model. Usually, AR "p" has a fixed value throughout the data. In this study, all AR values are 1, as shown in Table 4.7 below. For the value of Integrated (I) "d", it is always constant which value 1 throughout the series. Moving average (MA) is Similar to choosing "p" for the AR model. In order to determine the correct "q" order for the MA model, all values outside the boundary must be analyzed. Unlike the AR model, we may pick the order "q" for the MA(q) model from the ACF if this plot has a sharp cut-off after lag "q." The PACF plot decays more slowly, which is evidence of the MA process (Guarnaccia, et al., 2018).

Finding the correct sequence of "p" and "q" for the ARMA model through analysis of Auto-Correlation Function (ACF) and Partial Auto-Correlation Function (PACF) graphs can be difficult and time-consuming at times. To determine the optimal combination of p and q, an objective function that measures model performance on a validation set is required. For example, Log-Likelihood: choose the highest number to optimize the log-likelihood function. Select the highest value in the boundary. The bigger value is the least negative in the dataset (Myrtveit, Erik, & Olsson, 2001).

Table 4.7: Value of "AR", "I" and "MA" based on ACF and PACF

%	Parameter	Pegoh			Kota Kinabalu		
		AR "p"	I "d"	MA "q"	AR "p"	I "d"	MA "q"
5%	PM_{10}	1	1	**4**	1	1	**8**
	SO_2	1	1	**6**	1	1	**8**
	NO_2	1	1	**4**	1	1	**6**
	O_3	1	1	**4**	1	1	**4**
	CO	1	1	**7**	1	1	**5**
	WS	1	1	**9**	1	1	**12**
	RH	1	1	**10**	1	1	**9**
	AT	1	1	**3**	1	1	**3**
10%	PM_{10}	1	1	**4**	1	1	**7**
	SO_2	1	1	**7**	1	1	**8**
	NO_2	1	1	**3**	1	1	**8**
	O_3	1	1	**4**	1	1	**5**
	CO	1	1	**7**	1	1	**7**
	WS	1	1	**8**	1	1	**10**
	RH	1	1	**9**	1	1	**7**
	AT	1	1	**5**	1	1	**2**
15%	PM_{10}	1	1	**5**	1	1	**6**
	SO_2	1	1	**9**	1	1	**9**
	NO_2	1	1	**5**	1	1	**11**
	O_3	1	1	**5**	1	1	**4**
	CO	1	1	**5**	1	1	**4**
	WS	1	1	**7**	1	1	**7**
	RH	1	1	**10**	1	1	**6**
	AT	1	1	**10**	1	1	**4**

Note: % - Percentage of simulated missing data

4.4 The Performance of Imputation Method

The performances of the suggested imputation method, Time Series - ARIMA, Expectation-Maximize, and Markov Chain Monte Carlo to impute several percentages of missing data were discussed in this section. The performances will be ranked in a ranking model through 4 performances indicator, Prediction of Accuracy, Index of Agreement (d_2), Mean Absolute Error (MAE), and Root Mean Squared (RMSE).

4.4.1 Five Percent Simulated Missing Data

Table 4.8 shows the performance indicators for 5% simulated missing data in Pegoh and Kota Kinabalu. The index of agreement (d_2) of EM shows a higher number of performance indicators closer to the value 1. The closer the value to 1, the better the result will be. The value of d_2 for meteorological data (Wind Speed, Relative Humidity, Ambient Temperature) shows the highest d_2 value among other parameters. The d_2 value in ambient temperature shows the highest among other parameters. For Pegoh which is 0.900, and 0.937 for Kota Kinabalu. This imputation method is suitable for imputing the data like ambient temperature. The lowest d_2 value was at PM_{10}, 0.585, and 0.477 for Pegoh and Kota Kinabalu, respectively. However, the d_2 value for ARIMA in SO_2 shown is the highest in SO_2, 0.780. ARIMA shows it is suitable to impute SO_2 data in 5% of simulated missing data. Overall, EM is superior among other imputations in this percentage.

Table 4.8: The results of performance indicators for 5% simulated missing data in Pegoh and Kota Kinabalu

Method	PI	PM_{10}		SO_2		NO_2		O_3		CO		WS		RH		AT	
		Pegoh	KK	Pegoh	KK	Pegoh	KK	Pegoh	KK	Pegoh	KK	Pegoh	KK	Pegoh	KK	Pegoh	KK
EM	PA	0.408	0.361	0.145	0.399	0.620	0.723	0.826	0.879	0.637	0.839	0.195	0.716	0.725	0.894	0.900	0.937
	d_2	**0.585**	**0.477**	**0.644**	**0.588**	**0.754**	**0.831**	**0.884**	**0.933**	**0.744**	**0.891**	**0.257**	**0.830**	**0.837**	**0.940**	**0.944**	**0.961**
	MAE	10.412	8.944	0.000	0.000	0.003	0.002	0.008	0.003	0.128	0.146	0.545	0.391	7.761	3.712	1.144	0.935
	RMSE	13.982	15.888	0.000	0.000	0.004	0.003	0.010	0.005	0.163	0.180	0.685	0.522	9.268	4.898	1.438	1.221
MCMC	PA	0.261	0.113	0.121	0.249	0.458	0.554	0.727	0.742	0.384	0.747	0.110	0.460	0.644	0.825	0.787	0.870
	d_2	0.549	0.387	0.450	0.528	0.673	0.727	0.849	0.859	0.624	0.861	0.467	0.688	0.798	0.907	0.883	0.930
	MAE	13.991	11.970	0.000	0.000	0.004	0.003	0.009	0.005	0.176	0.171	0.696	0.594	9.375	4.562	1.712	1.323
	RMSE	18.069	19.496	0.001	0.001	0.005	0.004	0.013	0.007	0.231	0.223	0.885	0.830	11.635	6.362	2.132	1.681
ARIMA	PA	-0.168	0.196	0.628	0.264	-0.044	-0.138	-0.019	0.059	0.091	0.527	-0.035	0.005	0.030	0.051	0.180	0.122
	d_2	0.347	0.432	**0.780**	0.549	0.401	0.232	0.414	0.466	0.058	0.719	0.356	0.421	0.423	0.450	0.500	0.508
	MAE	15.829	13.127	0.000	0.000	0.004	0.005	0.019	0.010	5.123	0.227	0.667	0.770	13.308	10.712	2.842	3.365
	RMSE	21.163	19.389	0.000	0.000	0.006	0.007	0.025	0.013	5.377	0.285	0.833	0.967	16.568	13.055	3.634	4.137

Where: PI – Performance indicators, PA – Prediction of Accuracy, d_2 – Index of agreement, MAE – Mean absolute error, RMSE – Root mean squared error, PM_{10} – particulate matter, SO_2 – sulfur dioxide, NO_2 – nitrogen dioxide, O_3 – ozone, CO – carbon monoxide, WS – wind speed, RH – relative humidity, AT – ambient temperature

In figure 4.4, the data shows the ranking of imputation methods for 5% of simulated missing data in Pegoh and Kota Kinabalu. This ranking model was made based on Prediction Accuracy (PA) only. In figure 4.4, most parameters show that the Expectation-Maximization (EM) method is the most effective imputation technique. The EM technique was ranked highest for most air quality criteria. This show that the result is consistent with other researchers, such as Zakaria & Noor (2018) and Zakaria N. A. (2018), stating that EM was the best imputation method to impute air pollution data. Although a Time Series method, ARIMA is a top method in SO_2 with 5% simulated missing data in Pegoh. Overall, EM is the best imputation method throughout this study to impute long gaps in missing data in air pollution data. Therefore, the EM approach is the most appropriate imputation method to replace the long data gaps in this study.

The Markov Chain Monte Carlo (MCMC) method was ranked the second most suitable method for imputation. In this percentage, the MCMC and ARIMA methods competed to be the second-best imputation methods for filling in values for simulated missing data of 5%. Overall, MCMC was the second-best imputation method for this percentage. The performance of this method was good, but it was not as good as the EM method. According to Junninen et al. (2004), MCMC methods can create multiple simulated values for each missing point to reflect the uncertainty in the missing data.

The time series method ARIMA rank as the third-best imputation method to impute long gaps in missing data in this study. The performances of the method are pretty low compared to EM and MCMC. However, in some cases, it can be the second-best method where it is slightly better than MCMC. For example, in PM_{10} and SO_2, Kota Kinabalu air pollution data rank better than the MCMC method which shown in figure 4.4 below.

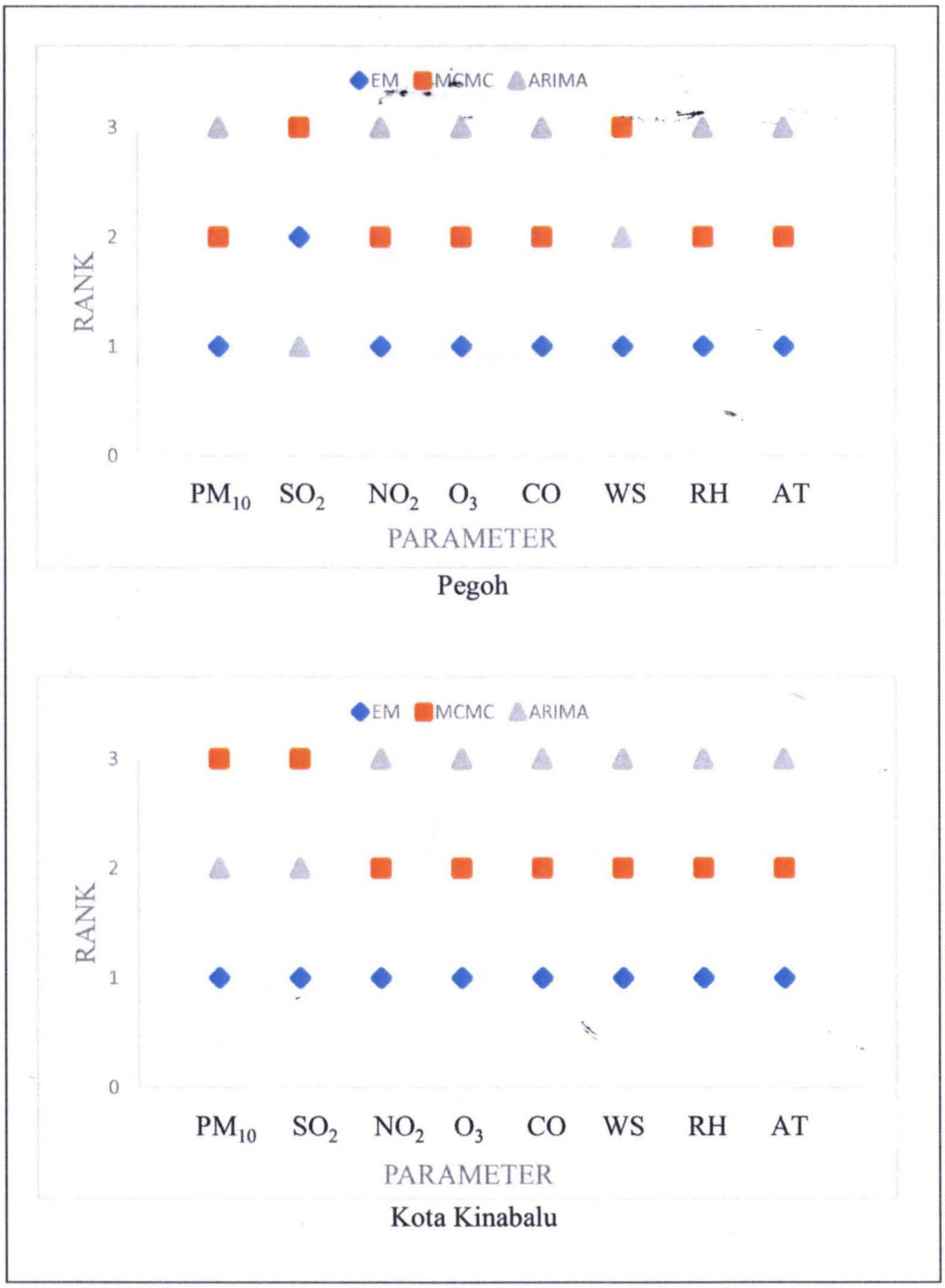


Figure 4.4: The ranking of all imputation methods for 5% simulated missing data in Pegoh and Kota Kinabalu

4.4.2 Ten Percent Simulated Missing Data

Table 4.9 shows the 10% simulated missing data performance indicators in Pegoh and Kota Kinabalu. The high-performance measures and low error values show that imputation was the most appropriate method for a missing value estimate. Based on Table 4.9, the EM method was the best way to impute the 10% of missing

values for Pegoh and Kota Kinabalu. This can be proved that there is a smaller error value in the Mean Absolute Error (MAE) in the EM imputation method. The value of MAE is shown to be the lowest among other parameters. The value was the closest to the value 0; if the value is closer to 0, the slightest error in the imputation method. The highest MAE in EM was in PM_{10}, 9.489, and 6.848 for Pegoh and Kota Kinabalu, respectively. The value for both places was 0 in the SO_2 value in the EM section. EM method proved to be the best imputation method to impute 10% of simulated missing data in the air pollution dataset.

Table 4.9: The results of performance indicators for 10% simulated missing data in Pegoh and Kota Kinabalu

Method	PI	PM_{10}		SO_2		NO_2		O_3		CO		WS		RH		AT	
		Pegoh	KK	Pegoh	KK	Pegoh	KK	Pegoh	KK	Pegoh	KK	Pegoh	KK	Pegoh	KK	Pegoh	KK
EM	PA	0.498	0.507	0.519	0.594	0.638	0.678	0.850	0.876	0.602	0.742	0.196	0.741	0.874	0.926	0.916	0.930
	d2	0.639	0.627	0.557	0.710	0.763	0.740	0.913	0.927	0.716	0.835	0.260	0.814	0.929	0.958	0.955	0.963
	MAE	**9.489**	**6.848**	**0.000**	**0.000**	**0.003**	**0.003**	**0.007**	**0.004**	**0.124**	**0.162**	**0.501**	**0.416**	**4.928**	**3.242**	**1.053**	**0.930**
	RMSE	12.699	11.838	0.001	0.000	0.004	0.004	0.009	0.005	0.182	0.209	0.637	0.617	6.409	4.397	1.283	1.162
MCMC	PA	0.276	0.206	0.271	0.372	0.408	0.455	0.730	0.774	0.413	0.556	0.037	0.503	0.750	0.851	0.800	0.860
	d2	0.558	0.457	0.545	0.617	0.644	0.669	0.852	0.875	0.640	0.744	0.412	0.704	0.863	0.920	0.891	0.925
	MAE	13.353	15.701	0.000	0.000	0.004	0.004	0.009	0.006	0.177	0.229	0.716	0.705	7.320	4.917	1.662	1.338
	RMSE	17.432	20.716	0.001	0.001	0.005	0.005	0.012	0.007	0.244	0.287	0.914	0.919	9.408	6.284	2.083	1.693
ARIMA	PA	0.038	0.206	0.216	0.151	0.045	0.218	0.063	0.096	0.092	0.663	-0.089	0.035	0.008	0.076	0.129	-0.036
	d2	0.400	0.466	0.480	0.496	0.437	0.521	0.467	0.454	0.411	0.791	0.290	0.402	0.402	0.481	0.449	0.416
	MAE	13.062	11.020	0.000	0.000	0.005	0.004	0.017	0.010	0.199	0.190	0.608	0.882	13.295	11.524	2.712	3.177
	RMSE	17.040	16.132	0.001	0.001	0.007	0.006	0.022	0.013	0.279	0.234	0.756	1.155	17.167	14.379	3.691	3.927

Where: PI – Performance indicators, PA – Prediction of Accuracy, d_2 – Index of agreement, MAE – Mean absolute error, RMSE – Root mean squared error, PM_{10} – particulate matter, SO_2 – sulfur dioxide, NO_2 – nitrogen dioxide, O_3 – ozone, CO – carbon monoxide, WS – wind speed, RH – relative humidity, AT – ambient temperature

The performance of all imputation methods for 10% simulated missing data in Pegoh and Kota Kinabalu is shown in figure 4.5. This ranking model was made using Prediction Accuracy (PA). Once again, the Time Series method - ARIMA ranked as the worst imputation method among the three imputation methods to impute long gaps missing data in air pollution data.

This study demonstrates that the EM approach is the superior imputation technique for filling in missing values for long gaps in missing data for all parameters in Pegoh and Kota Kinabalu. Like 5% simulated missing data, EM was shown to be the most effective method for estimating missing data in long gaps. The EM approach is preferable, mainly when the proportion of missing data is considerable. SPSS software makes the EM approach easier to implement and less time-consuming. Hence it is seen as straightforward and efficient. In addition, considering the complexity of this method's methodology, it provides accurate estimations.

In 10% of simulated missing data, the MCMC method was listed as an second most effective imputation method for air pollution data for Pegoh and Kota Kinabalu. According to research by Junninen et al. (2004), the MCMC approach fills in missing data by averaging or combining many simulated values. Applying Bayesian inference and repeating multiple phases, such as the imputation I-step and posterior P-step, were required to complete this process. These intricate techniques would be time-consuming but yield reasonable estimates of missing data. Another researcher, Zakaria (2018), reported that MCMC approaches were the second-best way to impute missing data, particularly in the case of long gaps, similar to this study showing that MCMC was listed second-best imputation method.

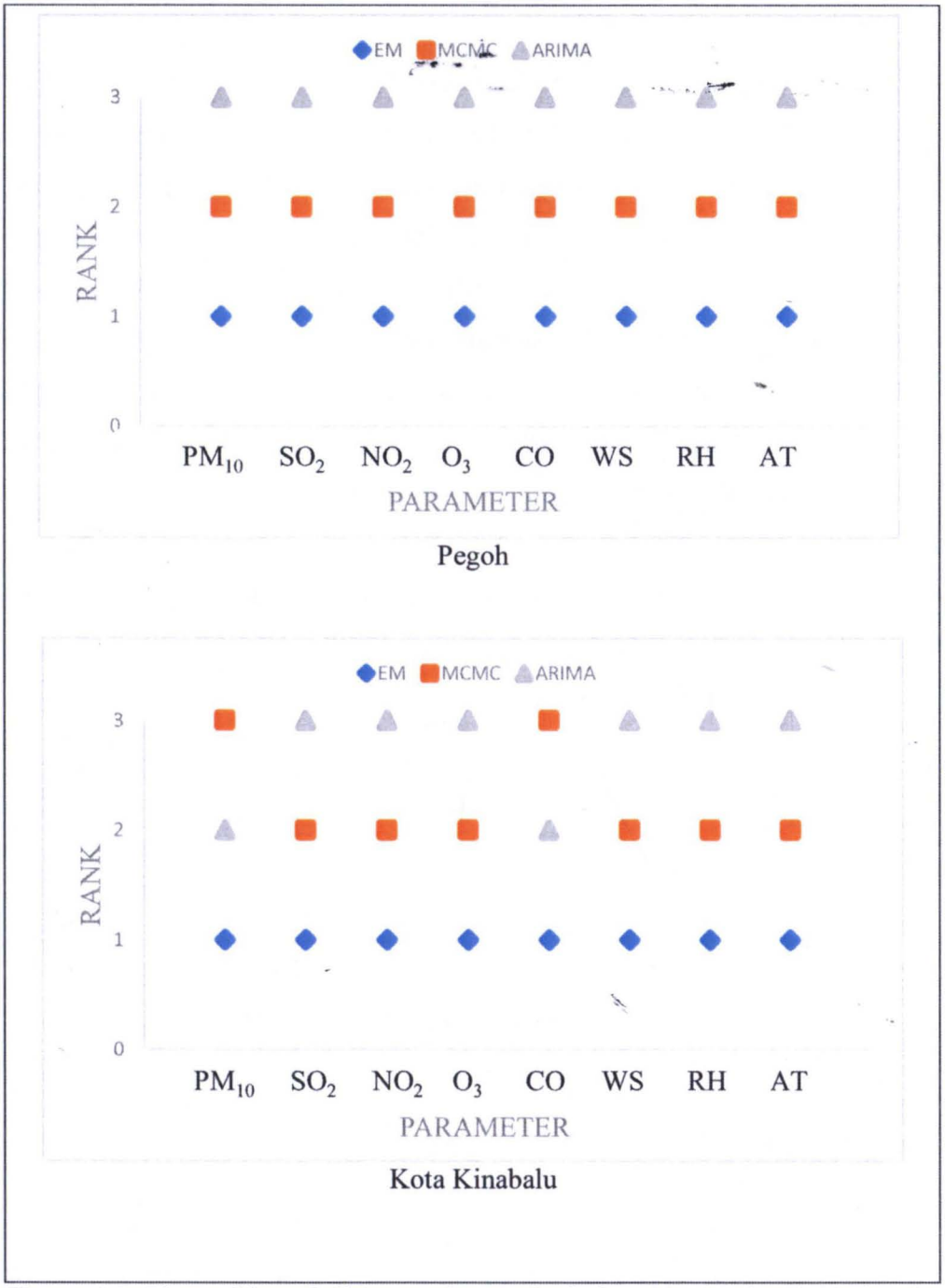


Figure 4.5: The ranking of all imputation methods for 10% simulated missing data in Pegoh and Kota Kinabalu

4.4.3 Fifteen Percent Simulated Missing Data

Table 4.10 shows the 15% simulated missing data performance indicators in Pegoh and Kota Kinabalu. In the table, EM shows the lowest error values in Root Mean Square Error (RMSE) among other imputation methods. Both methods, EM, and ARIMA, have 0 values for RMSE value for SO_2 shows 0 values for both places. Besides EM, ARIMA also has 0 RMSE values for SO_2 at both Pegoh and Kota Kinabalu. However, other parameters such as PM_{10} show the highest values in the EM, which are 13.734 and 17.123 for Pegoh and Kota Kinabalu. This is may due to the high value of extreme outliers in the PM_{10} data. The performance of EM may be dropped due to this problem. The performance of EM is slightly better than MCMC in the PM_{10}. Any abnormality in the data may affect the performance of the EM and cause a high RMSE value in table 4.10. Overall, EM proved the superior method to impute missing data in air pollution data. Then followed by MCMC, then ARIMA.

Table 4.10: The results of performance indicators for 15% simulated missing data in Pegoh and Kota Kinabalu

Method	PI	PM_{10}		SO_2		NO_2		O_3		CO		WS		RH		AT	
		Pegoh	KK	Pegoh	KK	Pegoh	KK	Pegoh	KK	Pegoh	KK	Pegoh	KK	Pegoh	KK	Pegoh	KK
EM	PA	0.464	0.427	0.506	0.502	0.684	0.676	0.828	0.846	0.516	0.772	0.184	0.662	0.803	0.888	0.835	0.895
	d_2	0.571	0.430	0.628	0.650	0.786	0.782	0.897	0.916	0.655	0.844	0.264	0.763	0.882	0.939	0.906	0.941
	MAE	10.006	7.934	0.000	0.000	0.003	0.003	0.007	0.004	0.123	0.141	0.498	0.449	6.653	3.973	1.467	1.064
	RMSE	**13.734**	**17.123**	**0.000**	**0.000**	**0.004**	**0.007**	**0.009**	**0.005**	**0.175**	**0.185**	**0.623**	**0.665**	**8.208**	**5.214**	**1.787**	**1.366**
MCMC	PA	0.225	0.161	0.182	0.371	0.462	0.479	0.686	0.718	0.246	0.620	0.063	0.463	0.669	0.807	0.706	0.807
	d_2	0.518	0.368	0.503	0.619	0.681	0.683	0.826	0.845	0.531	0.786	0.428	0.683	0.814	0.897	0.838	0.896
	MAE	14.083	11.071	0.000	0.000	0.004	0.003	0.010	0.006	0.182	0.195	0.678	0.615	8.541	5.234	1.956	1.442
	RMSE	18.258	21.019	0.001	0.001	0.005	0.005	0.014	0.008	0.248	0.254	0.856	0.903	11.116	7.071	2.482	1.913
ARIMA	PA	0.163	0.022	0.436	0.413	0.008	0.040	0.112	0.072	0.045	0.678	0.016	0.072	0.126	0.008	0.191	0.212
	d_2	0.495	0.249	0.663	0.621	0.346	0.392	0.463	0.461	0.412	0.808	0.400	0.428	0.471	0.438	0.504	0.543
	MAE	15.366	9.899	0.000	0.000	0.004	0.004	0.016	0.010	0.191	0.171	0.646	0.910	13.399	11.354	2.899	2.831
	RMSE	20.236	20.161	0.000	0.000	0.006	0.005	0.020	0.013	0.246	0.215	0.826	1.179	16.537	14.919	3.608	3.433

Where: PI – Performance indicators, PA – Prediction of Accuracy, d_2 – Index of agreement, MAE – Mean absolute error, RMSE – Root mean squared error, PM_{10} – particulate matter, SO_2 – sulfur dioxide, NO_2 – nitrogen dioxide, O_3 – ozone, CO – carbon monoxide, WS – wind speed, RH – relative humidity, AT – ambient temperature

Figure 4.6 show the result of imputation methods for 15% of simulated missing data for Pegoh and Kota Kinabalu. Similar to both previous percentages, EM and MCMC show superior imputation methods for filling in long gaps missing data in air pollution data. EM shows a superior method in all percentages of missing data in air pollution data. This conclusion was similar to the findings of other researchers, such as Zakaria (2018), who concluded that the EM approach is the most effective imputation for imputing long gaps of missing data in air pollution data. This was demonstrated when most parameters indicated that the EM approach was top-ranked in the ranking model for all simulated missing data.

MCMC remained listed as second-ranked in the ranking model; meanwhile, ARIMA ranked third, making it unsuitable for imputing long gaps in missing data in this study. Although ranked last, ARIMA is still a suitable imputation method since it has ranked the second in SO_2 and wind speed for Pegoh and SO_2 and CO in Kota Kinabalu.

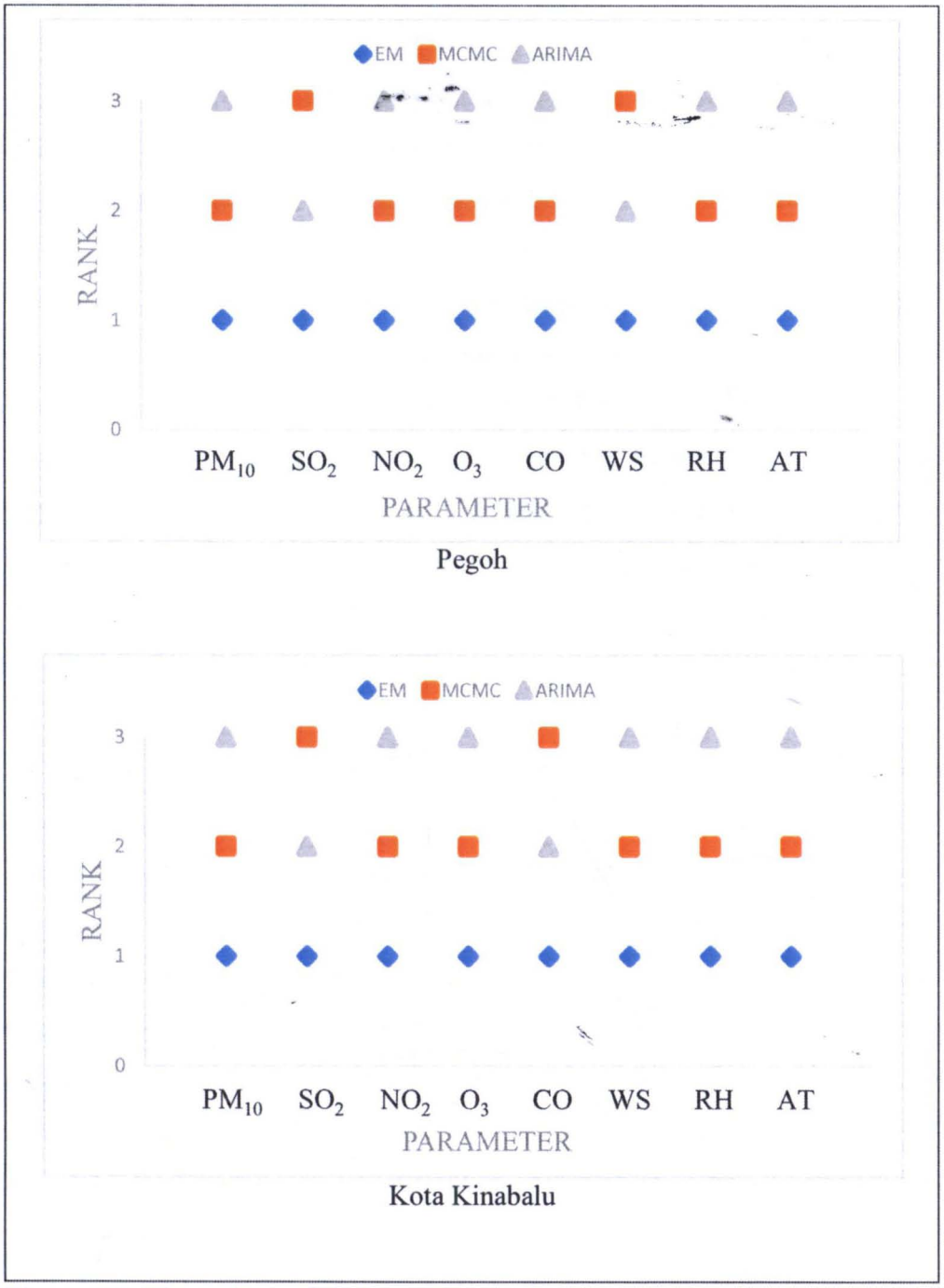


Figure 4.6: The ranking of all imputation methods for 15% simulated missing data in Pegoh and Kota Kinabalu

4.5 Summary

Tables 4.11 and 4.12 display the overall performance and error measurements for the three percentages of simulated missing data for the air pollution dataset for both Pegoh and Kota Kinabalu. The best imputation approach for filling in long gaps of missing data in air pollution was the Expectation-Maximization (EM) method, followed by the Markov Chain Monte Carlo (MCMC) method, and then the ARIMA method. This was proven by the fact that nearly all performance metrics in each percentage acknowledged that the EM approach was the best suitable for filling the long gaps of missing values. It appears that the performance of the EM approach was rather good, despite the significant gaps (24 to 120 hours) in the air pollution dataset. All percentages of simulated missing data (5 %, 10 %, 15 %) were proven to be imputed with high performance and low error by the EM method.

Across all percentages of simulated missing data, the EM method was the best imputation technique for filling in a long gap of missing values in the air pollution dataset. This is demonstrated when all the performance indicators for each percentage of simulated missing data concurred that this imputation method was the most effective. Even for datasets with long-missing hour gaps, this technique's performance was deemed exceptional. This conclusion is consistent with what Abd Razak et al. (2014) observed: the EM approach performed exceptionally well despite the large percentages of missing values. Using the EM approach, both Pegoh and Kota Kinabalu can demonstrate good performance despite a long interval and a significant proportion of missing data simulations.

The EM approach was the best imputation method for filling in the missing data for this study's air pollution dataset. In this dataset, the performance of the EM and MCMC methods was comparable. However, the EM technique was selected as the best imputation method since most of the findings in table 4.12 indicated that the EM approach was superior. EM consisted of iterative estimation, which consisted of two steps: estimate and prediction (Zainuri, Jemain, & Muda, 2015). Even though the proportion of simulated missing data was raised due to the recurrent pattern in the dataset, this strategy performed well. According to

Moshenberg et al. (2015), the performance of this technique tends to be positive when the missing data are part of a repeating pattern; otherwise, this method may fail when the missing data pattern does not repeat in a regular rhythm. As a single imputation approach that is simple, robust, and straightforward to implement, the EM method is more accessible to implement than the MCMC method (Guarnaccia, et al., 2018).

The MCMC approach was the second-best imputation method for computing simulated missing data. This technique performed exceptionally well, even as the proportion of simulated missing data increased with more missing data in the dataset. The performance indicators indicate no significant difference between EM and MCMC for all three simulated missing data in Pegoh and Kota Kinabalu. The MCMC approach was recognized as the best imputation method. This demonstrates that this imputation strategy performed exceptionally well on datasets with a lower percentage of missing data. Junninen et al. (2004), have indicated that MCMC is the ideal approach for imputation due to its complicated procedure that may reflect the uncertainty associated with missing data. In this investigation, however, EM outperformed MCMC due to the linear relationship between missing data and accessible data in air pollution (Moshenberg et al., 2015; Junninen et al., 2004). This is because the primary assumption of the EM approach is that the missing data has a linear relationship with the available data, such as time-series data (Junninen et al., 2004). Air pollution data is one example of a time series.

Table 4.11: The overall average of performances indicators for Pegoh and Kota Kinabalu

Method	PI	5%	10%	15%	Average
EM	PA	0.638	0.693	0.656	0.662
	d_2	0.756	0.769	0.741	0.755
	MAE	2.133	1.732	2.020	1.962
	RMSE	3.017	2.466	3.069	2.851
MCMC	PA	0.503	0.516	0.479	0.500
	d_2	0.699	0.707	0.682	0.696
	MAE	2.787	2.884	2.751	2.807
	RMSE	3.848	3.751	4.010	3.870
ARIMA	PA	0.109	0.119	0.163	0.131
	d_2	0.441	0.460	0.481	0.461
	MAE	4.126	3.544	3.606	3.759
	RMSE	5.341	4.676	5.088	5.035

Table 4.12: Summary of best imputation at every percentage at Pegoh and Kota Kinabalu

Missing Percentage (%)	Best Imputation Method							
	PA		d_2		MAE		RMSE	
	Pegoh	KK	Pegoh	KK	Pegoh	KK	Pegoh	KK
5%	EM	EM	EM	EM	EM	EM	EM	EM
10%	EM	EM	EM	EM	EM	EM	EM	EM
15%	EM	EM	EM	EM	EM	EM	EM	EM

The objective of a scatter plot is to assess the strength of the link between predicted (EM estimate) values and observed values (treated raw data). Figure 4.7 presents the scatter plot of the observed and predicted data using the EM (Expectation-Maximization) method for all parameters to simulate 15% of missing data in Pegoh. The R^2 values in these graphs were characterized as the variance of the y-axis (predicted data) clarified by the x-axis (observed data). Due to the more significant link between x and y, the more closely R^2 approaches 1, the better the forecast. The greater the value of R^2, the more significant the correlation between the expected and observed data (Noor et al., 2008).

These figures resulted from all parameters in Pegoh for 15% simulated data. Overall, R^2 values for all Pegoh observations were quite near 1. This suggested that the expected and actual values in Pegoh were virtually identical. This study

determined that the EM approach is better than other methods for air pollution datasets with long-missing hours. This indicates that the EM method is the most stable among other methods. All R^2 values were very close to the value of 1. All the R^2 values were over 0.8 in all of the parameters, which can be considered high in the R^2 values. EM is a stable and superior method to impute long gaps in missing data in air pollution datasets. This is consistent with the study run by Zakaria (2018) which stating that EM is the best method to fill in long missing hours of air pollution dataset.

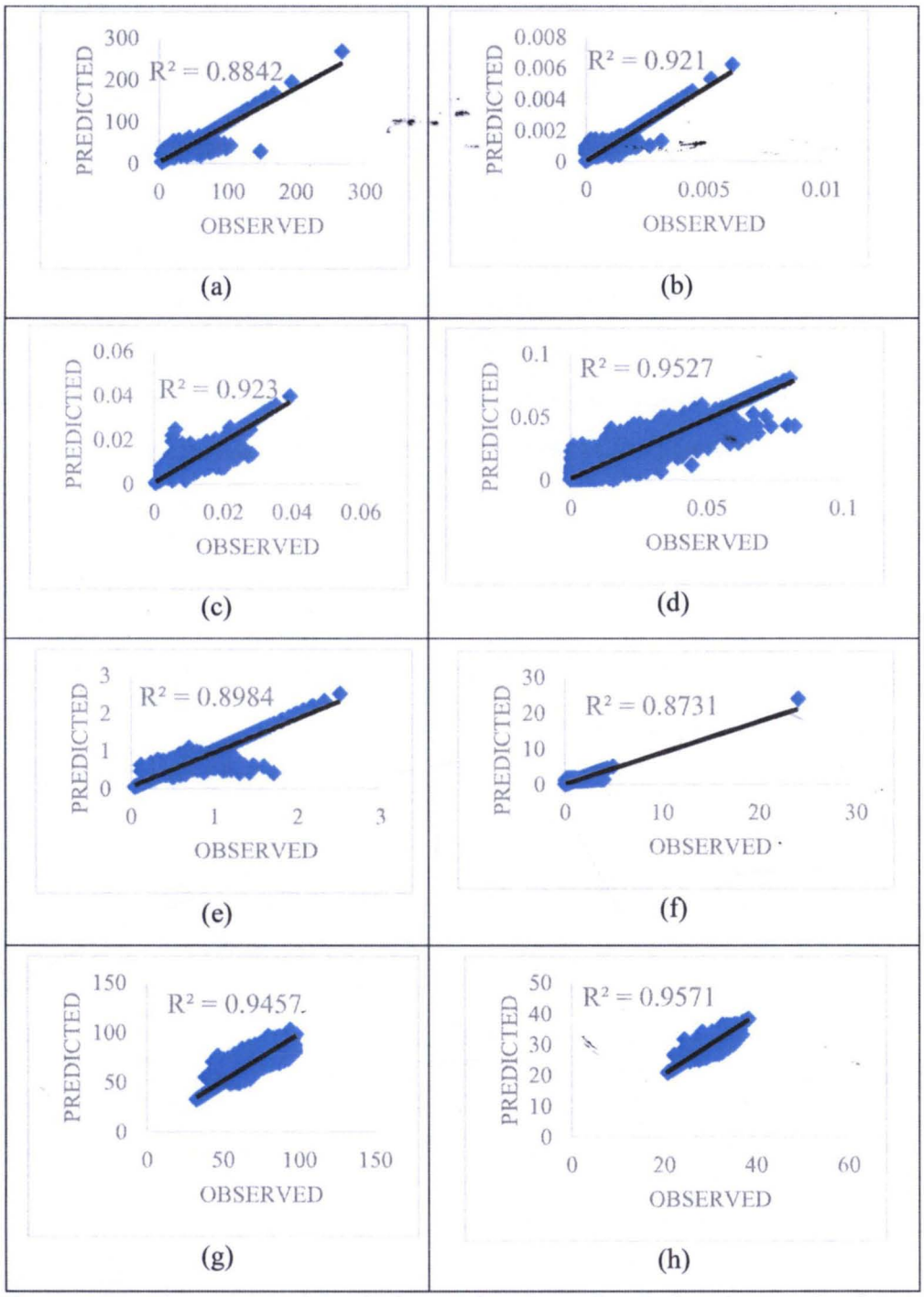


Figure 4.7: The scatter plot of observed and predicted data in Pegoh of 15% simulated missing data for (a) PM_{10}, (b) SO_2, (c) NO_2, (d) O_3, (e) CO, (f) wind speed, (g) relative humidity, and (h) ambient temperature.

Table 4.13 summarises each advantage and disadvantages of each imputation method. The performance of EM methods to estimate missing data in long-missing gaps was superior to that of other imputation techniques. MCMC methods were demonstrated to be the second-best imputation methods in competition with EM methods. MCMC was slightly behind the EM method in imputing the long gaps missing value. The EM and MCMC methods included all current, complete data in the air pollution dataset while estimating missing values. Therefore, extreme outliers may lower the accuracy of this method's predictions. ARIMA approaches shown poor performance to impute missing values and inconsistent performance to estimate missing data in long gaps of missing data in a dataset on air pollution.

Table 4.13: Summary of description of each imputation method

Tier	Methods	Description
Good	Expectation Maximize	• It is always assured that the likelihood will rise with each repetition. • The E-step and M-step are often pretty easy to implement for many problems. • Able to treat long gaps missing data.
Moderate	Markov Chain Monte Carlo	• Implementation is fairly easy. • Needs more samples than simple importance sampling, which makes it slower.
Poor	Auto Regression Integrated Moving Average	• Implementation is quite challenging. • Has few ways to deal with a lot of variables and complicated relationships in long gaps missing data. • High accuracy and easy to understand if have strong past data. • Expensive computation cost.

CHAPTER 5

CONCLUSION

5.1 Conclusion

Various findings may be drawn from this study. This study utilized five air quality data hourly monitoring records and three meteorological data from two air quality monitoring stations in Pegoh, Perak, and Kota Kinabalu, Sabah, in 2018. The five air quality variables were particulate matter (PM_{10}), dioxide (SO_2), nitrogen dioxide (NO_2), ozone (O_3), and carbon monoxide (CO). In contrast, the three meteorological variables included wind speed (WS), relative humidity (H), and ambient temperature (AT). This dataset was used to investigate the properties of missing values in the air pollution dataset. Overall, the Kota Kinabalu air quality dataset has more missing data than the Pegoh air quality dataset. Kota Kinabalu had the most significant percentage of missing data at 4.40%, while Pegoh had the lowest percentage at 3.40%. Before selecting the observed data, the dataset must be treated with linear interpolation. The overall proportion of missing data in both areas is consisted of short gaps missing data, and linear regression is considered the most effective imputation method for short gaps of missing data in air pollution data.

This study focused on the long gap of missing data, which ranged from 24 to 120 hours. In research by Noor (2014), more than % of air pollution data gaps in Malaysia were shorter than 12 hours long (Noor, 2014). This work should assist in treating long hours of data gaps in air pollution datasets. For Pegoh and Kota Kinabalu, the largest number of lost hours in this research is 120, equivalent to 5 days. Even if the CAQM is faulty, large gaps in missing data should not be as extensive in practice. This equipment may be repaired immediately or replaced with

a brand-new one. The percentages of simulated missing data were simulated to be 5%, 10%, and 15%, respectively.

Three imputation methods replaced the missing observations in the simulated data. Expectation-Maximization (EM), Markov Chain Monte Carlo (MCMC), and Auto-Regression Integrated Moving Average (ARIMA) were used as imputation techniques. The effectiveness of suggested imputation methods was evaluated using four performance indicators: Prediction Accuracy (PA), Index of Agreement (d2), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). For both Pegoh and Kota Kinabalu, EM was chosen as the best imputation method to replace the long gaps of missing values in air pollution data. This is because Expectation Maximize provides superior performance in datasets with long hours of missing data gaps and high percentages of the accuracy of simulated missing data. EM also displays a low error in the error measurement. The ranking model validated the selection of the EM method as the superior option for filling in missing data on air pollution. This is shown by the fact that this method first appeared in almost all of the simulated percentages of missing data. The performance of the MCMC method is slightly lower than the Expectation-Maximization method but still, a preferable method to impute long gaps in missing data in the air pollution dataset. According to Junninen et al. (2004), MCMC is the best imputation method for filling in missing data in air pollution data. This is due to the procedure's complexity, which might represent the risk involved with missing data. However, in this study, the EM method imputes better than the MCMC method, even though the EM method is a simple, robust, and easy-to-use single-imputation method (Gómez-Carracedo et al., 2014). The suggested imputation method, ARIMA, is relatively poor in imputation performance due to the sizeable systematic error between the observed and the predicted observations.

The ARIMA method requires just the past time-series data to generalize a forecast or impute missing data. Therefore, the ARIMA approach can improve prediction accuracy while minimizing the number of factors. The classic model identification methods for finding the proper model from the class of alternative models are typically challenging to comprehend and computationally costly. This method is also subjective, and the predictor's skill and experience might influence

the model's accuracy. Second, the underlying theoretical model and structural links are not as different as they are in particular straightforward for imputation methods, such as EM and MCMC. Moreover, ARIMA models, like other imputation methods, are fundamental "retrospective," which let the past predict the future (Guarnaccia, et al., 2018). So that, in the long run, the forecast becomes a straight line and is not very good at predicting series with turning points.

5.2 Recommendations for Future Study

This study aimed to investigate the effectiveness of the time series method known as ARIMA in solving the issues of missing data in air quality monitoring datasets with large gaps. This research has led to the identification of the imputation method, which can be used to treat missing data in datasets related to air pollution. This study has the potential to be improved in the future by:

I. Model identification: at this stage, particularly for nonstationary models, the data must undergo differencing to become stationary and look for opportunities to expand the sample size (use at least four years of data).

II. Estimation is the process of choosing models that are as simple as possible. Here, fully utilize the Correlogram of the different data. Do not overlook the ACF and PACF in the model. Consider any lag of the ACF and PACF that falls outside of the 95% confidence bound to determine the AR and MA values in the model.

III. Diagnostics: To check to see if the models are still beneficial. The best fit model is most likely the one with the lowest Akaike Information Criterion (AIC) and the highest Log-Likelihood values. Also, the residuals correlation chart must be stable.

IV. If forecasts cannot be compared to actual data, there is a problem with the modelling procedure. To make a non-stationary series into stationary data, it must be differentiated. For the ARIMA model to determine the order, a

correlogram of the difference must be used. The predictions should then be consistent with the observed data in some manner.

V. Before initiating the simulation and imputation process, maybe consider replacing or removing the outliers and extreme outliers with suitable values to improve the performance of the EM and MCMC methods. This is necessary because any abnormality present in the dataset will impact the performance of the EM and MCMC methods.

REFERENCES

Md Yusof, N. F., Ramli, N. A., Yahaya, A. S., Sansuddin, N., Ghazali, N. A., & Madhoun, W. A. (2010). Monsoonal differences and probability distribution of PM10 concentration. *Environ Monit Assess*, 655–667.

Ahn, H., Sun, K., & Kim, K. P. (2021). Comparison of missing data imputation methods in time series forecasting. *Computers,Materials & Continua*, 767-779.

Allison, P. D. (2002). *Missing Data.* Pennsylvania: SAGE Publications.

Bauer, J., Angelini, O., & Denev, A. (2017). Imputation of multivariate time series data - Performance benchmarks for multiple imputation and spectral techniques. *SSRN Electronic Journal*, 1-5.

Bing, L., Jin, Y., & Li, C. (2021). Analysis and prediction of air quality in Nanjing from autumn 2018 to summer 2019 using PCR–SVR–ARMA combined model. *Scientific Reports (11)*, 1-14.

Bowker, G. E., Schwede, D. B., Lear, G. G., & Warren-Hicks, W. (2011). Quality assurance decisions with air models: A case study of imputation of missing input data using EPA's multi-layer model. *Water, Air and Soil Pollution*, 1-4.

Carracedo, M. P., Andrade, J., López-Mahía, P., & Lorenzo, S. M. (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*.

Cedar Lake Ventures, I. (31 December, 2018). *Weather Spark: The Weather Year Round Anywhere on Earth.* Retrieved from 2018 Weather History in Kota Kinabalu: https://weatherspark.com/h/y/130286/2018/Historical-Weather-during-2018-in-Kota-Kinabalu-Malaysia#Figures-Humidity

Chapra, S., & Canale, R. (1998). *Numerical methods for engineers.* Hoboken: McGraw-Hill Higher Education.

Dekker, R. (2006). The importance of having data-sets. *27th IATUL Conference*, 22-25.

Eekhout, I., Vet, H. C., Twisk, J. W., Brand, J. P., Boer, M. R., & Heymans, M. W. (2014). Missing data in a multi-item instrument were best handled by multiple imputation at the item score level. *Journal of Clinical Epidemiology*, 10-16.

Ghapor, A. A., Zubairi, Y. Z., & Imon, A. R. (2017). Missing Value Estimation Methods for Data in Linear. *Sains Malaysiana*, 317-326.

Guarnaccia, C., Griselda, C.-B., Breton, R., Tepedino, C., Quartieri, J., & Mastorakis, N. (2018). ARIMA models application to air pollution data in Monterrey, Mexico. *Mathematical Methods in Science and Engineering* (pp. 2-41). Mexico: AIP Conference Proceedings 1982.

Ibrahim, M. Z., Zailan, R., Ismail, M., & Lola, M. S. (2009). Forecasting and Time Series Analysis of Air Pollutants in Several Area of Malaysia. *American Journal of Environmental Sciences 5*, 625-632.

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)*, 2895-2907.

Kasyoki, A. (2013). Simple steps for fitting ARIMA model to time series data for forecasting using R. *International Journal of Science and Research (4)*, 318-321.

Kellermann, A. P. (2018). Missing data in complex sample surveys: Impact of deletion and imputation treatments on point and interval parameter estimates. *Mathematics*, 45-67.

Lee, M. H., Rahman, N. A., Suhartono, Nor, M. E., & Kamisan, N. B. (2012). Seasonal ARIMA for Forecasting Air Pollution Index: A Case Study. *American Journal of Applied Sciences 9 (4)*, 570-578.

Libasin, Z., Fauzi, W. W., Ul-Saufie, A., Idris, N. A., & Mazeni, N. A. (2021). Evaluation of Single Missing Value Imputation Techniques for Incomplete Air Particulates Matter (PM10) Data in Malaysia. *Pertanika Journals*, 3099-3112.

Little, R., & Rubin, D. (2002). *Statistical Analysis with Missing Data.* New York: John Wiley & Sons, Inc.

Myrtveit, I., Erik, S., & Olsson, U. H. (2001). Analyzing data sets with missing data: an empirical evaluation of imputation methods and likelihood-based methods. *IEEE Transactions on Software Engineering (27)*, 999-1013.

Noor, N. M., Abdullah, M. M., Yahaya, A. S., & Ramli, N. A. (2014). Comparison of linear interpolation method and mean method to replace the missing values in environmental data set. *Materials Science Forum*, 278-281.

Noor, N. M., Yahaya, A. S., & Abdullah, M. A. (2008). Estimation of missing values in air pollution data using single imputation techniques. *ScienceAsia (34)*, 341-345.

Penzer , J., & Shea, B. (1999). Finite Sample Prediction and Interpolation for ARIMA Models with Missing Data. *Journal of Forecasting*, 411-419.

Rahman, E. A., Hamzah, F. M., Latif, M. T., & Dominick, D. (2021). Assessment of PM2.5 patterns in Malaysia using the clustering method. *Aerosol and Air Quality Research*, 21-161.

Rubin, D. B. (1976). Inference and Missing Data. *Biometrika (63)*, 581-592.

Rumaling, M. I., Chee, F. p., Dayou, J., & Chang, J. (2020). Missing Value Imputation for PM10 Concentration in Sabah using Nearest Neighbour Method (NNM) and Expectation-Maximization (EM) Algorithm. *Asian Journal of Atmospheric Environment*, 62-72.

Schafer, J. L. (1999). Multiple imputation: a primer. *Statistical Methods in Medical Research*, 3-15.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 147–177.

Sharma, S. (2019). Descriptive statistics. *Horizon University*, 12-36.

Suhaimi, N., Ghazali, N. A., Nasir, M. Y., Mokhtar, M. Z., & Ramli, N. A. (2017). Markov chain monte carlo method for handling missing data in air quality datasets. *Malaysian Journal of Analytical Sciences (21)*, 552-559.

Sukatis, F. (2019). Filling the gaps of missing values in air pollution dataset by using various imputation methods. *International Journal of Conservation Science (10)*, 791-804.

Takahashi, M. (2017). Statistical inference in missing data by MCMC and non-MCMC multiple imputation algorithms: Assessing the effects of between-imputation iterations. *Data Science Journal*, 16-37.

Tang, W., Kassim, A., & Abubakar, S. (1996). Comparative studies of various missing data treatment methods. *Atmospheric Research 42*, 247-262.

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

Weerasinghe, S. (2010). A missing values imputation method for time series data: an efficient method to investigate the health effects of sulphur dioxide levels. *Environmetrics 2010 (21)*, 162-172.

Wui, J. C., Pien, C. F., Kai, S. K., & Sentian, J. (2018). Variability of the PM10 concentration in the urban atmosphere of Sabah and its responses to diurnal and weekly changes of CO, NO2, SO2 and Ozone. *Asian Journal of Atmospheric Environment (12)*, 109-126.

Wyzga, R. E. (2012). Note on a method to estimate missing air pollution data. *Journal of the Air Pollution Control Association (23)*, 207–208.

Ye, Z. (2019). Air pollutants prediction in Shenzhen based on ARIMA and prophet method. *E3S Web of Conferences (36)*, 5-10.

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*, 449-456.

Zakaria, N. A. (2018). Imputation methods for filling the long interval of missing data and meteorological dataset. *Urbanism*, 159-166.

Zakaria, N., & Noor, N. M. (2018). Imputation Methods For Filling Missing Data In Urban Air Pollution Data For Malaysia. *Urbanism. Architecture. Constructions.*, 159-166.

Zheng, Y. (2020). Predictive Study of Tuberculosis Incidence by ARMA Model Combined with Air Pollution Variables. *Complexity*, 1-11.

APPENDIX

The minimum and maximum range gap of 5% simulated missing data for Pegoh and Kota Kinabalu

Parameters	Range Gap for 5 % missing data (hour)			
	Minimum (Pegoh)	Maximum (Pegoh)	Minimum (KK)	Maximum (KK)
PM_{10} (g/m^3)	27	95	25	99
SO_2 (ppm)	51	108	44	96
NO_2 (ppm)	33	93	27	91
O_3 (ppm)	30	118	30	115
CO (ppm)	29	118	36	111
Wind Speed (m/s)	29	120	59	112
Relative Humidity (%)	26	74	25	115
Ambient Temperature (°c)	27	108	24	98

The minimum and maximum range gap of 10% simulated missing data for Pegoh and Kota Kinabalu

Parameters	Range Gap for 10 % missing data (hour)			
	Minimum (Pegoh)	Maximum (Pegoh)	Minimum (KK)	Maximum (KK)
PM_{10} (g/m^3)	26	108	24	115
SO_2 (ppm)	27	116	32	119
NO_2 (ppm)	25	116	36	114
O_3 (ppm)	27	113	28	117
CO (ppm)	26	120	36	117
Wind Speed (m/s)	55	118	40	115
Relative Humidity (%)	27	114	24	112
Ambient Temperature (°c)	39	113	27	102

The minimum and maximum range gap of 15% simulated missing data for Pegoh and Kota Kinabalu

Parameters	Range Gap for 15 % missing data (hour)			
	Minimum (Pegoh)	Maximum (Pegoh)	Minimum (KK)	Maximum (KK)
PM_{10} (g/m^3)	24	120	25	119
SO_2 (ppm)	28	117	32	120
NO_2 (ppm)	32	118	36	118
O_3 (ppm)	33	117	32	120
CO (ppm)	24	119	24	117
Wind Speed (m/s)	25	120	28	113
Relative Humidity (%)	35	113	29	118
Ambient Temperature (°c)	25	118	28	120