INTERNATIONAL JOURNAL OF CONSERVATION SCIENCE
ISSN: 2067-533X Volume 10, Issue 4, October-December: 791-804 www.ijcs.uaic.ro

ESTIMATION OF MISSING VALUES IN AIR POLLUTION DATASET BY USING VARIOUS IMPUTATION METHODS

Fahren Fazzer SUKATIS[1], Norazian Mohamed NOOR[1,*], Nur Afiqah ZAKARIA[1], Ahmad Zia UL-SAUFIE[2], Annas SUWARDI[3]

[1] Sustainable Environment Research Group (SERG), Centre of Excellence Geopolymer and Green Technology (CEGeoGTech), School of Environmental Engineering, Universiti Malaysia Perlis (UniMAP), Kompleks Pusat Pengajian Jejawi 3,02600 Arau, Perlis, Malaysia.
[2] Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Cawangan Pulau Pinang, Kampus Permatang Pauh, Jalan Permatang Pauh 13500 Permatang Pauh, Pulau Pinang, Malaysia.
[3]Universitas Negeri Makasar, Faculty of Mathematics and Natural Sciences, Kampus UNM Parangtambung, Jalan Daeng Tata Makassar, Indonesia.

Abstract

The aim of this study is to determine the best imputation method to fill in the various gaps of missing values in air pollution dataset. Ten imputation methods such as Series Mean, Linear Interpolation, Mean Nearest Neighbour, Expectation Maximization, Markov Chain Monte Carlo, 12-hours Moving Average, 24-hours Moving Average, and Exponential Smoothing (α = 0.2, 0.5, and 0.8) were applied to fill in the missing values. Annual hourly monitoring data for ambient temperature, wind speed, humidity, SO2, NO2, O3, CO, and PM10 from Petaling Jaya and Shah Alam were used from 2012 to 2016. These datasets were simulated into three types of missing data patterns that vary in length gaps of missing patterns, i.e. simple, medium and complex patterns. Each patterns was simulated into two percentages of missing, i.e. 10% and 20%. The performance of these imputation methods was evaluated using four performance indicator: mean absolute error, root mean squared error, prediction accuracy, and index of agreement. Overall, the Expectation Maximization method was selected as the best method of imputation to fill in the simple, medium and complex patterns of simulated missing data, while the Series Mean method was shown as the worst method of imputation.

Keywords*: Air pollution; Estimation; Missing data; Imputation methods; Simulation; Performance indicators.*

Introduction

The purpose of air quality monitoring is to measure the ambient air quality, if there is a significant change in air quality level, the public should be told to be prepared for the situation. However, continuous air quality monitoring station (CAAQM) requires a frequent maintenance to ensure that the air pollution data obtained from this station is accurate. However, maintenance process will cause air pollution data from the station to be incomplete [1, 2]. According to [3], air pollution data can be missing because of too many uncontrollable conditions such as malfunctioning of the instruments, maintenance or repairing, and calibration.

In environmental studies, missing data is a problem repeatedly encountered by researchers [4-6]. Discontinuities of data pose a significant obstacle for time-series forecast schemes, which for most of the parts require continuous information as a condition for their application. Missing data hinder the ability to make exact conclusion or interpretations about the observation [7]. Therefore, the missing data need to be treated, because complete data are required to perform statistical analysis, for example in time series analysis, principal component

* Corresponding author: norazian@unimap.edu.my

analysis (PCA) and multivariate analysis; it requires continuous data in order to perform prediction [8]. Discontinuities of data pose a significant difficulty for time-series prediction schemes, this is because such analysis require a continuous data as a condition for their use [8].

The most popular method for handling the missing observation in the dataset is by deleting those observations [10]. By removing missing values using the deletion method, it can introduce substantial biases in the study [9, 11]. Other than deletion method, mean substitution is also one the popular imputation method because it is easy to use. Mean imputation method replaces the missing value with the mean value of each variable on the respective missing variables as an estimate of the missing value [9,10]. Referring to [3], the mean imputation underestimates the variance in the dataset and can alter any other derived chemometric study. Moreover, this method can lead to a problem of bias and large errors [9].

This study focuses on applying various imputation methods and selecting the best methods for several complexity of gaps in the simulated missing observation of air quality dataset. There are many imputation methods can be used to fill in the missing data in air pollution dataset [9]. The length of the missing gaps and the type of study conducted must be considered in determining the best method of imputation [1]. Hence, this study focuses on applying various imputation methods and selecting the best methods for several complexity gaps of the simulated missing observation in air quality dataset.

Experimental Part

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

Raw Air Pollution Dataset

Air pollution hourly dataset was used in Petaling Jaya and Shah Alam from 2012 to 2016. These locations were selected because Petaling Jaya is the industrial area, while Shah Alam is the urban area in Klang Valley, Selangor. These dataset was obtained from Department of Environment, Malaysia. There are five air pollution data which is carbon monoxide (CO) in ppm, nitrogen dioxide (NO_2) in ppm, sulphur dioxide (SO_2) in ppm, particulate matter (PM_{10}) in $\mu g/m^3$, ozone (O_3) in ppm and three meteorological data such as ambient temperature (AT) in ^oC, wind speed (WS) in km/h, and relative humidity (%HR).

Simulation of Missing Data

To test the effectiveness of different imputation methods, three random simulated missing patterns that are simple, medium and complex would be used (Table 1).

Table 1. The length of missing gaps for each missing pattern

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

l – The length of the gaps in hour.
[a] Simple and medium patterns are mixed in proportion of 1:1

The patterns were different in the length of missing gaps (hour). Simple pattern contained the missing data gaps that less than 24-hours, while medium was a gaps of missing data in between 24-hours to 168-hours, and complex was a combination of simple and medium patterns in proportion of ½ and ½ respectively. In this study two percentages of missing data were applied for each patterns i.e. as 10% and 20%. The purpose of the simulation is to evaluate

the proposed imputation method and compare its performance with the established methods, usually a simulation study was based on different missing data patterns [12].

Imputation Methods

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

Series Mean

The Series Mean (SM) method is the mean of all subjects related to a certain variable, and it is the default value in the program which is SPSS [13]. The equation of series mean as follows [14]:

$$\bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n \qquad (1)$$

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

Mean Nearest Neighbor

The Mean Nearest Neighbor (MNN) method is an arithmetical mean which is computed by using complete observation values under and above the missing data, then that value will be imputed instead of the missing data [13]. The equation of MNN as follows [8]:

$$y = y_1 \text{ if } x \le x_1 + \left[\frac{x_2 - x_1}{2}\right] \qquad (2)$$

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

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

Linear Interpolation

Linear interpolation (LI) method fill the gaps of missing data by replace the missing value with average value of the before and after data in sequential pattern [15]. This method performs better for short gap of missing data [14]. The equation of LI is written as follow [15]:

$$y^* = y_1 + \left(\frac{y_2 - y_1}{x_2 - x_1}\right)\left(x^* - x_1\right) \qquad (3)$$

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

Exponential Smoothing

Exponential Smoothing (ES) is methods that merge a linear trend with multiple seasonal components so that the seasonal effect is proportional to the current level of series [17]. The smoothing coefficient of 0.1 until 0.9 will be used. The equation of ES as follows [18]:

$$F_t = F_{t-1} + \alpha\left(A_{t-1} - F_{t-1}\right) \qquad (4)$$

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

Moving Average

Moving Average (MA) is a method by averaging a number of points from the input signal to produce each number in the output signal [19]. The equation of MA is written as [19]:

$$y_i = \left(\frac{1}{m}\right)\sum_{j=-k}^{k} x_{i+j} \qquad (5)$$

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

Expectation Maximization

The Expectation Maximization (EM) method involves two steps which is prediction and estimation by iterative calculation [9]. To undertake this method SPSS execute the several steps; (i) the mean, variance, and covariance are estimated from the individual complete data; (ii) the maximum likelihood procedures will be used to estimate a regression equations that relate each variable to each other variable which generate the formula, and; (iii) the formula are used to estimate the missing values [20].

Markov Chain Monte Carlo Method

The Markov Chain Monte Carlo (MCMC) method, the data are assuming from a multivariate normal distribution, then data augmentation will be applied to Bayesian inference with missing data by repeating the several steps such as the imputation I-step and posterior P-step, these two steps are iterated long enough to produce the results to be reliable for a multiply imputed data set [21]. This method objective is to have iterates converge to stationary distribution and then to simulate an approximately independent draw of the missing values. SPSS would execute this method.

Performance Indicators

Performance measure is used to describe the goodness of fit for each of the imputation methods [8]. There are four performance indicators to measure the goodness of fit of the imputation methods used in this study which are Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Index of Agreement (d_2), and Prediction Accuracy (PA). According to [14], there are two types of performance indicators to measure the goodness of fit for each imputation methods i.e. error measure (MAE and RMSE) and performance measure (d_2 and PA). The performance for each imputation methods were displayed in the form of rank and the best imputation methods for overall and for each pattern of simulated missing data were selected. Table 2 show the performance indicator formulae.

Table 2. The performance indicators formulae [15]

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

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

Results and Discussion

Characteristics of Missing Data (Raw Air Pollution Dataset)

Figure 1 show the percentage of total missing data (%) in Shah Alam and Petaling Jaya from 2012 to 2016. The highest missing data in Shah Alam was 29.68% (2015) and the lowest was 4.88% (2012), meanwhile the highest missing data in Petaling Jaya was 3.14% (2012) and the lowest in 2016 which is 2.11%. The percentage of missing data between Shah Alam and Petaling Jaya shows the large significant difference in all of 5 years except in 2012. This significant difference shows that the dataset in Shah Alam from 2013 to 2016 are not suitable for used as reference data.

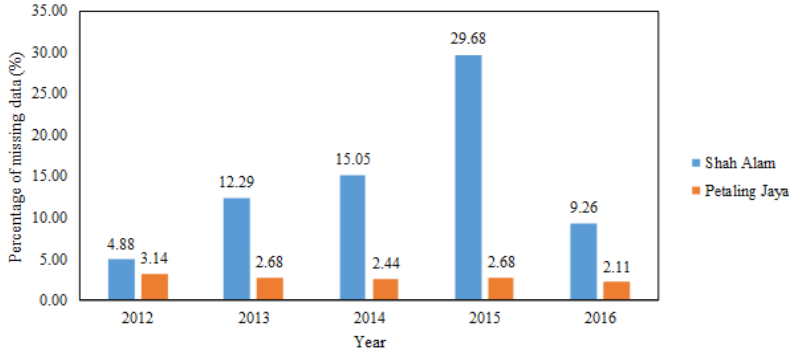

Fig. 1. The percentage of total missing data (%) at Shah Alam and Petaling Jaya from 2012 to 2016

Figure 2 and 3 show the longest gap of missing observation of air pollution (in hour) for Shah Alam and Petaling Jaya from 2012 to 2016. The longest gap of missing data in Shah Alam was 7963 hours (2014) for ambient temperature measurements records and those gaps of missing values are equal to about 332 days which was 90.7% of days in a year. The smallest gap of missing data observed was in humidity measurement at 46 hours (2012) and 46 hours (2014) of PM_{10} measurement. In Petaling Jaya (Fig. 3), NO_2 records contain the longest gaps of missing data which is 743 hours (2013) and it is equivalent to a month. Meanwhile, the smallest gap of missing data recorded was 3 hours (2016) in CO measurement records. However, in 2014 there was no missing data recorded in SO_2 concentration records.

All air quality parameters in Shah Alam (Fig. 2) contained more than 1000 hours gaps of missing except in 2012.

Therefore, air quality dataset in 2012 contained the least missing data compared to other year in Shah Alam. Furthermore, the longest gap of missing data was observed in ambient temperature monitoring records of 2012 that was 824 hours. In Petaling Jaya, all air quality parameters contain less than 1000 hours of missing gaps. Referring to [8], the reference data was defined as the dataset that have the most complete data. As for Petaling Jaya the most complete data which containing the lowest total missing data was in 2016. However, in order to select the most suitable dataset as reference data for this research, dataset in 2012 was selected as the most suitable reference data for both location since that dataset only contained the lowest total missing data in Shah Alam and the total missing data in Petaling Jaya slightly below than Shah Alam in 2012. Table 3 show the descriptive statistics for all air pollutants in Petaling Jaya and Shah Alam (2012).

Table 3. The descriptive statistics for all air pollutants in petaling jaya and shah alam (2012)

Parameters	Location	Valid N	Missing	Mean	Median	Std Deviation	Outliers	Extreme Outliers
WS	PJ	8754	30	4.6	4.3	2.3	190	0
	SA	8737	47	4.9	4.2	3.3	12	0
AT	PJ	8754	30	28.5	27.9	2.8	0	0
	SA	7960	824	28	26.7	3.7	0	0
H	PJ	8754	30	70	72	15	0	0
	SA	8738	46	79	83	15	0	0
SO_2	PJ	8247	537	0.003	0.003	0.003	424	197
	SA	8081	703	0.003	0.002	0.005	403	211
NO_2	PJ	8318	466	0.03	0.028	0.012	77	3
	SA	8268	516	0.019	0.018	0.011	113	1
O_3	PJ	8283	501	0.013	0.006	0.015	219	11
	SA	8205	579	0.018	0.01	0.02	203	1
CO	PJ	8278	506	1.178	1.1	0.563	190	10
	SA	8302	482	0.75	0.66	0.46	188	8
PM_{10}	PJ	8677	107	49	45	25	259	63
	SA	8553	231	48	44	27	231	73

Where: PJ – Petaling Jaya, SA – Shah Alam

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

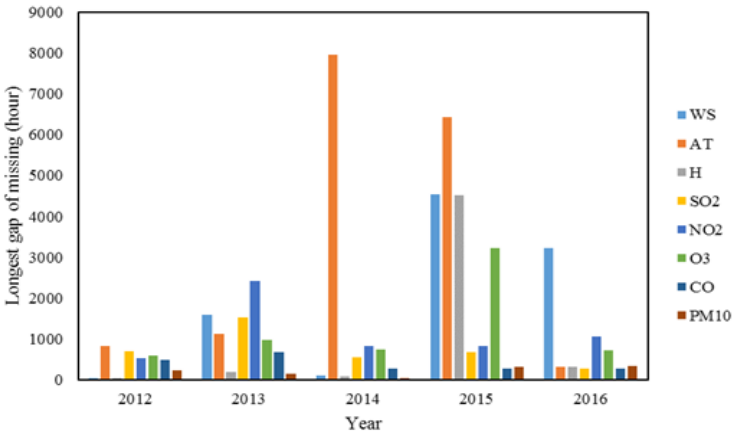

Fig. 2. The longest gap of missing observation (in hour) in Shah Alam from 2012 to 2016

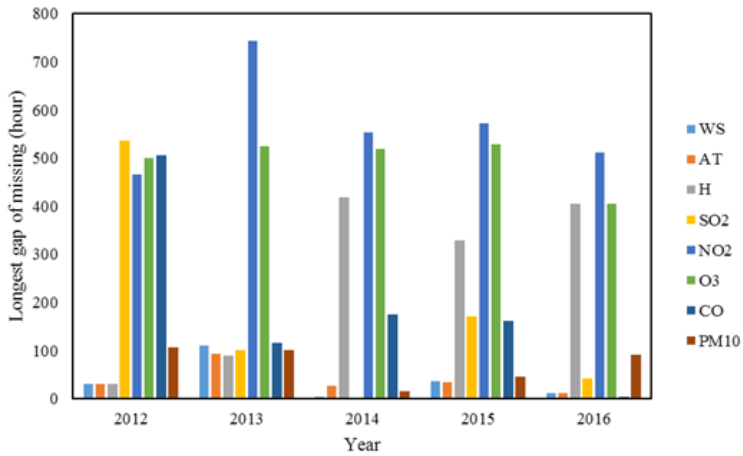

Fig. 3. The longest gap of missing observation (in hour) in Petaling Jaya from 2012 to 2016

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

concentration distribution varies more compared to the other parameters for both locations, hence, the range of the PM_{10} concentration was the highest. Based on the table, observation on PM_{10} and SO_2 data for both locations contained the highest reading of extreme outliers in 2012 which is expected due to haze episodes for PM_{10} observation.

Characteristics of the Simulated Missing Data

Table 4 show the percentage of missing data gap (in hour) for each of the missing gap patterns of simulated missing data. The number gap of missing data was presented in percentages and the gap intervals for simple pattern was 6 hours and for both medium and complex pattern were 24 hours.

Table 4. Percentage of the number of simulated missing data in gap length (hour) for each of the missing gaps pattern of simulated missing data

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

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

Generally, the pattern of descriptive statistics for 10% and 20% simulated missing data were not much vary from one to other percentages of missing data although the pattern of gaps for each percentages of missing were varied. As example, Figurer 4 shows the percentile for complex patterns of simulated missing data (PM_{10}) from Petaling Jaya.

From the figure, it can be seen that there are not much differences for values of every percentiles even though the percentage of missing increases. When the structure of simulated missing data changes from its original, it would be considered as a different dataset which is not the same as its original. Disturbed structure of simulated missing data may affect the performance of imputation process. This is because this process depends on existing data to estimate the missing values. This occurrence was due to the random number generated in

producing the simulated missing values patterns and the availability of large number of observation with the same range [16]. This finding is consistent with the other the study by [2], who also found that the structure of the simulated of missing data not interrupted after simulation process.

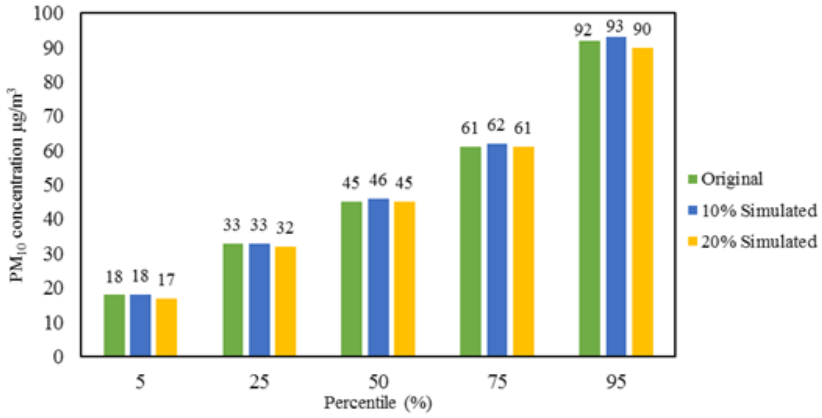

Fig. 4. The percentile for complex simulated of missing data (PM_{10}) in Petaling Jaya

The Performance of Imputation Method

Table 5 shows the average results of performance indicators of the simulated missing data for simple, medium, and complex pattern for Petaling Jaya and Shah Alam respectively.

Table 5. The average results of performance indicators of the simulated missing data for Petaling Jaya and Shah Alam

Patterns	Methods	10% simulated missing data				20% simulated missing data			
		Performance Indicators				Performance Indicators			
		MAE	RMSE	PA	d_2	MAE	RMSE	PA	d_2
Simple	**EM**	**2.7161**	**3.7195**	**0.7206**	**0.7988**	**2.8231**	**3.9357**	**0.7036**	**0.7788**
	MCMC	3.1705	4.3018	0.6429	0.7676	3.2791	4.5528	0.6230	0.7553
	LI	3.9319	5.3079	0.3873	0.6285	3.9971	5.4612	0.3712	0.6161
	MNN	4.7496	6.2644	0.2274	0.5246	4.8309	6.4419	0.2135	0.5158
	12MA	4.9040	6.1369	0.1081	0.4280	4.8134	6.0641	0.1209	0.4392
	24MA	4.2401	5.3014	0.2740	0.4806	4.1941	5.3123	0.2600	0.4742
	0.8ES	4.7578	6.2687	0.2237	0.5221	4.8376	6.4425	0.2101	0.5132
	0.5ES	4.7969	6.2850	0.2091	0.5118	4.8706	6.4475	0.1968	0.5037
	0.2ES	4.8794	6.2387	0.1523	0.4687	4.9060	6.3422	0.1473	0.4666
	SM	4.5586	5.7637	0.0000	0.1025	4.6899	5.9688	0.0000	0.0710
Medium	**EM**	**2.5774**	**3.2658**	**0.7098**	**0.7974**	**3.2020**	**4.8371**	**0.6923**	**0.7744**
	MCMC	3.0130	3.9090	0.6412	0.7705	3.6398	5.3480	0.6104	0.7478
	LI	5.0944	6.4377	0.1686	0.4714	5.1497	7.0362	0.1689	0.4805
	MNN	5.7321	7.1644	0.0898	0.4330	5.7637	7.8420	0.0965	0.4334
	12MA	5.3869	6.7704	0.0728	0.4089	5.4161	7.2905	0.1086	0.4305
	24MA	4.7776	5.9515	0.1397	0.4059	5.0083	6.6890	0.1697	0.4104
	0.8ES	5.7328	7.1645	0.0895	0.4328	5.7666	7.8435	0.0960	0.4330
	0.5ES	5.7350	7.1654	0.0878	0.4317	5.7746	7.8471	0.0937	0.4315
	0.2ES	5.7263	7.1421	0.0792	0.4265	5.7854	7.8389	0.0850	0.4295
	SM	4.2810	5.1570	0.0000	0.1053	5.0583	6.8406	0.0000	0.0951
Complex	**EM**	**2.6881**	**3.4816**	**0.6854**	**0.7832**	**2.9292**	**4.1235**	**0.6929**	**0.7735**
	MCMC	3.1225	4.1444	0.6096	0.7513	3.4066	4.7440	0.6131	0.7511
	LI	4.1946	5.3751	0.2098	0.5073	4.8906	6.3658	0.2169	0.5095
	MNN	5.1191	6.6129	0.1391	0.4740	5.7993	7.7586	0.1302	0.4559
	12MA	4.7922	5.9984	0.1294	0.4397	5.0164	6.4806	0.1133	0.4301
	24MA	4.4677	5.5336	0.1749	0.4269	4.6462	5.9436	0.1744	0.4165
	0.8ES	5.1205	6.6132	0.1393	0.4743	5.8023	7.7590	0.1291	0.4550
	0.5ES	5.1298	6.6172	0.1358	0.4720	5.8115	7.7607	0.1247	0.4518
	0.2ES	5.1605	6.6073	0.1186	0.4593	5.8116	7.7249	0.1078	0.4390
	SM	4.4132	5.3167	0.0000	0.1609	4.7603	6.1934	0.0000	0.1051

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

Referring to the table the Expectation Maximization (EM) methods shows selected as the best imputation method followed by Markov Chain Monte Carlo (MCMC) method as the second best imputation method for filling air pollution dataset. This is because all of the performance measures had shown the smallest error values in root mean squared error (RMSE) and mean absolute error (MAE), and as expected, high performance values in index of agreement (d_2) and prediction accuracy (PA) for both EM and MCMC methods respectively.

Figure 5 and 6 shows the overall performance and error measures for 10% and 20% - simulated missing data. The high value performance measures and low value of error measure indicates that the imputation was the most suitable for estimation of missing values. The Expectation Maximization (EM) methods show the best performance followed by Markov Chain Monte Carlo (MCMC) method. This is due to small error values in root mean squared error (RMSE) and means absolute error (MAE), while the performance values in index of agreement (d_2) and prediction accuracy (PA) were high for both EM and MCMC methods almost for all of the parameters had been observed. Meanwhile, Linear Interpolation (LI), Mean Nearest Neighbour (MNN), Moving Average (MA), and Exponential Smoothing (ES) show the moderate performance that competes with each other but not good as EM and MCMC methods. Series Mean (SM) method was shown to be the worst imputation method for estimation of all patterns in 20% of simulated missing data. This is because this method contributed to large error and small performance compared to the EM and MCMC methods. Overall, the performance of all imputation methods used for 20% simulated missing data from the good to worst was in the order of EM; MCMC; LI; MNN; 0.8ES; 0.5ES; 0.2ES; 12MA; 24MA; and SM.

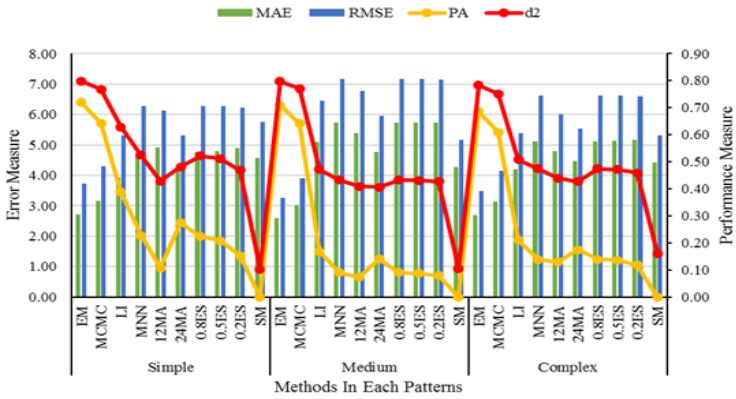

Fig. 5. The overall performance and error measures for10% - simulated missing data

Fig. 7 and 8 show the the ranking of all imputation methods for 10% and 20% simulated missing data in Petaling Jaya and Shah Alam respectively. Most of the air quality parameters agree that Expectation Maximization (EM) method is the best imputation method. This is because EM method was listed as the first in ranking for most of the air quality parameters except for PM_{10}. This result is consistent with the others researchers such [10,12, 14] who also found that EM imputation was the suitable method to impute the missing values in air pollution dataset. It seems that the performance of EM method was very good although the variety of the missing gaps in air pollution dataset were different. EM method was shown to have great performance and low error to impute the 10% and 20% of simple, medium, and complex simulated missing data. According to [20], in the first step of EM method, the mean, variance and covariance are estimated from the individual completed data. Therefore, any abnormality in the current complete data such outlier may affect the estimation of mean, variance and covariance. As consequent, the performance of estimation of the missing values by EM method reduced. Clearly, the performance of EM method was dropped when the number of extreme outliers was high. However, when the value of standard deviation was too small, EM method performance slightly better compared to the performance of EM method in dataset which contain high standard deviation.

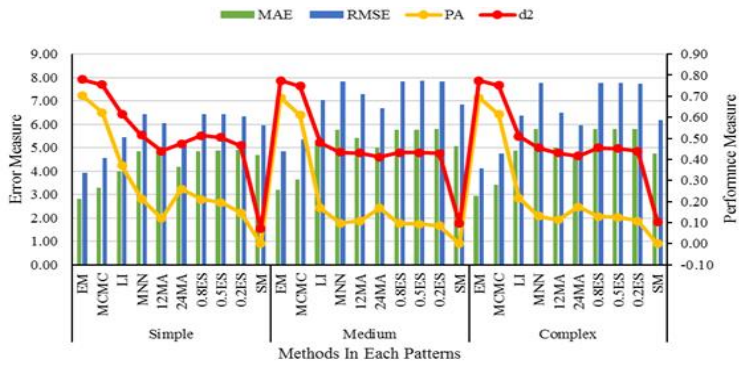

Fig. 6. The overall performance and error measures for 20% - simulated missing data

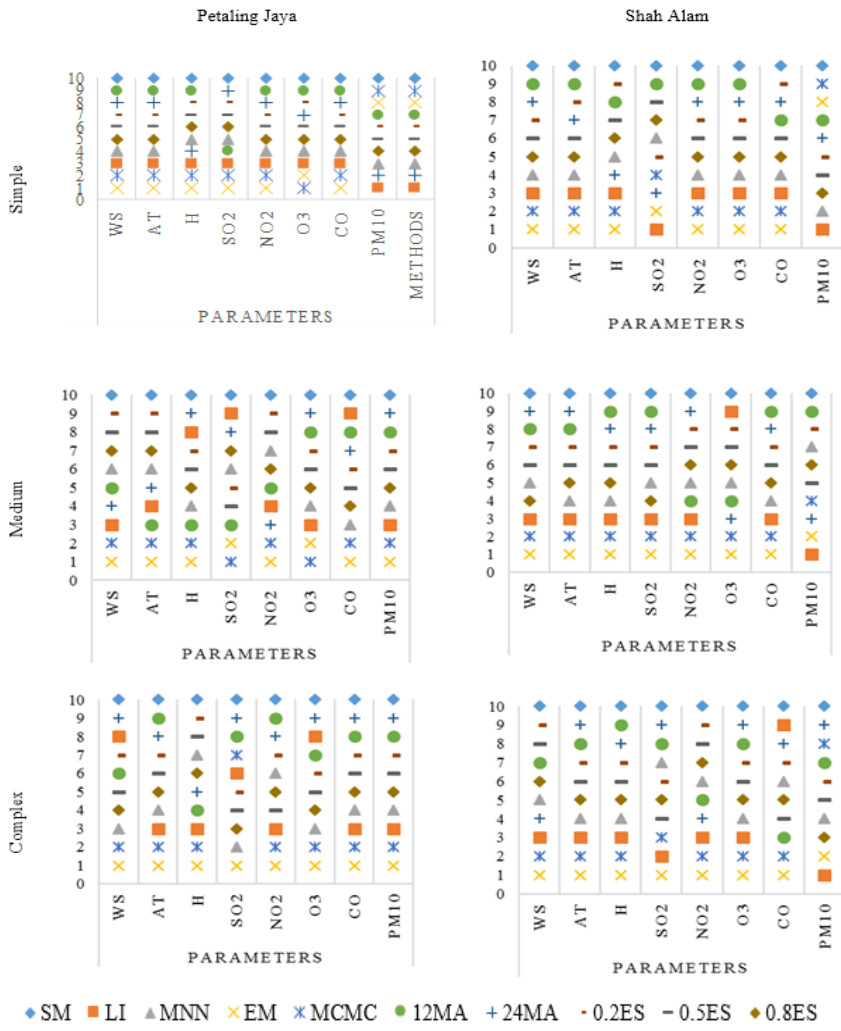

Fig. 7. The ranking of all imputation methods for 10% simulated missing data in Petaling Jaya and Shah Alam. Where: MAE – mean absolute error, RMSE – root mean squared error, PA – prediction accuracy, d_2 – index of agreement, EM – expectation maximization, MCMC – markov chain monte carlo, LI – linear interpolation, MNN – mean nearest neighbour, MA – moving average, ES – exponential smoothing

Markov Chain Monte Carlo (MCMC) method was the second most appropriate imputation method. In many cases, the EM and MCMC methods were competing to be the best imputation methods to fill in the values in 10% and 20% simulated missing data. This method performance was considered good but slightly lower compared to the EM method. This method performance was considered as good but slightly lower compared to the EM method. According to [8], MCMC method fills in each of missing data by averaging or pooling multiple simulated values. This process was done by complicated procedures such applying Bayesian inference and repeating several steps such as the imputation I-step and posterior P-step [21]. These complicated procedures would consume time but produce excellent estimation of missing data. Found that MCMC methods [14] were the second best method to impute the missing values especially in long gaps of missing.

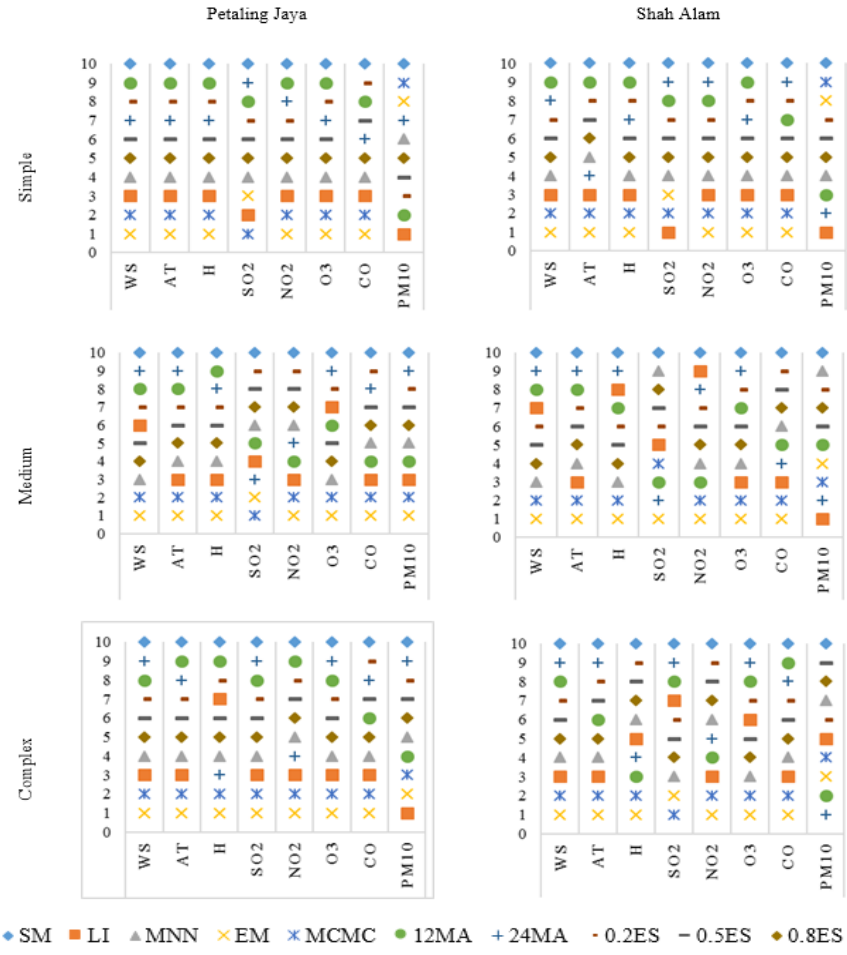

Fig. 8. The ranking of all imputation methods for 20% simulated missing data in Petaling Jaya and Shah Alam. Where: MAE – mean absolute error, RMSE – root mean squared error, PA – prediction accuracy, d_2 – index of agreement, EM – expectation maximization, MCMC – markov chain monte carlo, LI – linear interpolation, MNN – mean nearest neighbour, MA – moving average, ES – exponential smoothing

This method mechanism is quite similar to the EM method which also considers the existing values in dataset to make an estimation of missing values. Therefore, any abnormality or extreme values may reduce the prediction performance of this method.

On the other hand for PM_{10} dataset, Linear Interpolation (LI) method was ranked to have the highest performance in replacing the 10% simple patterns of simulated missing data for Petaling Jaya, and 10% medium and complex patterns for Shah Alam. This is due to the extreme outliers and standard deviation in PM_{10} dataset itself was the largest compared to the other parameters (Table 3). The performance of some of EM method in SO_2 was slightly better compared to PM_{10} parameter dataset, although SO_2 dataset contains the largest extreme outliers but the value of standard deviation (variability) in SO_2 was lower compared to PM_{10} data (Table 3). LI method show better performance compared to MNN, MA, and ES methods because the uncertainty was covered by this method. According to [2], LI method fill in the gaps of missing data by replacing the missing value with average value of data before and after missing data in sequential pattern.

Other methods such as Mean Nearest Neighbour (MNN), Moving Average (MA), and Exponential Smoothing (ES) methods show moderate performance. As they are always ranked between 3 and 9. This is due to the solving operation process of all this methods which only considers several current complete data to estimate the missing data. MNN method does not cover the uncertainty completely because this method only brings the previous complete data to fill in the missing gaps. Usually upper nearest the value would be selected to replace the missing values [15]. In this study, MA and ES methods show moderate performance because they were originally used for forecasting analysis which only considers several past values to predict the value in the future. This method converge the estimation values when the gaps of missing become larger which the uncertainty would not be covered especially in long gaps of missing data.

Meanwhile, the Series Mean (SM) method was shown as the worst imputation methods because this method was ranked at the last in rank model for all parameters in 10% and 20% - simple, medium, and complex simulated missing data. Reported by [3, 9], the mean imputation underestimates the variance and lead to the error. Overall, the ranking of all imputation methods used for 10% and 20% simple, medium, and complex simulated missing data from the good to worst performances for both location were EM; MCMC; LI; MNN; 0.8ES; 0.5ES; 0.2ES; 12MA; 24MA; and SM.

Table 6 shows the summary of each imputation methods. In this study, the performance of Expectation Maximization (EM) methods was superior compared to the other methods to estimate the missing data in various type of missing gaps.

Table 6. The summary of each imputation methods

Tiers	Methods	Description
Good	Expectation Maximization (EM)	✓ Excellence performances in all simulated of missing data gaps. ✓ Easy to implement and cost effective.
	Markov Chain Monte Carlo (MCMC)	✓ Complicated solving operations process. ✓ Sensitive to the abnormality in dataset i.e. outliers.
Moderate	Linear Interpolation (LI)	✓ Moderate performances in all simulated of missing data gaps estimation.
	Mean Nearest Neighbour (MNN)	✓ Easy to implement but time consuming. ✓ Easy solving operations process. ✓ Suitable for short gaps of missing data.
	Moving Average (MA)	✓ The estimation tends to converge when the gaps become large.
	Exponential Smoothing (ES)	
Bad	Series Mean (SM)	✓ Worst performances for all simulated of missing data gaps. ✓ Easy to implement and cost effectives. ✓ Easy solving operations process. ✓ Lead to bias due to uncertainty are not covered.

A Markov Chain Monte Carlo (MCMC) method was proved as the second best imputation methods which competing with EM methods. Observably, EM and MCMC methods consider all current complete data in air pollution dataset for estimation of missing values.

Meanwhile, Series Mean (SM) Methods was shown as completely the worst methods to impute the missing data since almost all the performance of this method in each patterns gaps was lowest. For the Linear Interpolation (LI) method was considered as slightly well performance in this study. Other methods such as Mean Nearest Neighbour (MNN), Moving Average (MA), and Exponential Smoothing (ES) methods shown the moderate performance to impute the missing values and this methods also show an inconsistent performance to estimate the missing data in some different pattern of missing gaps.

Conclusion

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

Acknowledgments

The authors would like to thank Ministry of Education Malaysia for Fundamental Research Grant Scheme (FRGS 9003-00508) and Department of Environment, Malaysia for providing the air pollution dataset.

References

[1] S. Moshenberg, U. Lerner, B. Fishbain, *Spectral methods for imputation of missing air quality data*, **Environmental Systems Research**, **4**(26), 2015, pp. 1-13.
[2] N.A. Zakaria, N.M. Noor, *Imputation methods for filling missing data in urban air pollution data for Malaysia*, **Urbanism. Architecture. Constructions**, **9**(2), 2018, pp. 159-166.
[3] M. Gómez-Carracedo, J. Andrade., P. López-Mahía, S. Muniategui, D. Prada, *A practical comparison of single and multiple imputation methods to handle complex missing data in air quality*, **Chemometrics and Intelligent Laboratory Systems**, **134**, 2014, pp. 23-33.
[4] M. Zelenakova, P. Purcz, R.D. Pintilii, P. Blistan, P. Hlustik,M. Oravkova M. Abu Hashim, *Spatio-temporal variations in water quality parameter trends in river waters*, **Revista de Chimie**, **69**(10), 2018, pp. 2940-2947.
[5] C.R. Vintu, I.N. Alecu, A. Chiran, E. Leonte, A.F. Jitareanu, M. Stefan, *Researches on the Agrotouristic Offer of Guest Houses in Dornelor Bassin (Case Study)*, **International Journal of Conservation Science, 8**(3), 2017, pp. 419-430.
[6] A.A. Kadir, M.M. Al Bakri Abdullah, A.V. Sandu, N.M. Noor, A.L.A. Latif, K. Hussin, *Usage of palm shell activated carbon to treat landfill leachate*, **International Journal of Conservation Science, 5**(1), 2014, pp. 117-126.
[7] N.M. Noor, A.S. Yahaya, N.A. Ramli, M.M.A. Bakri, *Mean imputation techniques for filling the missing observations in air pollution dataset*, **Key Engineering Materials**, **594-595**, 2014, pp. 902-908.

[8] H. Junninnen, H. Niska, IK. Tuppurainen, J. Ruuskanen, M. Kolehmainen, *Methods for Imputation of missing values in air quality data sets*, **Atmospheric Environment**, **38**(18), 2004, pp. 2895-2907.
[9] N.A. Zainuri, A.A. Jemain, N.A. Muda, *A comparison of Various Imputation Methods for Missing Values in Air Quality Data,* **Sains Malaysiana**, **44**(3), 2015, pp. 449-456.
[10] N. Abd Razak, Y. Zulina Zubairi, R.M. Yunus, *Imputing Missing Values in Modelling the PM10 Concentrations*, **Sains Malaysiana, 43**(10), 2014, pp. 1599-1607.
[11] R.J.A. Little, D.B. Rubin, **Statistical Analysis with Missing Data II**, Wiley, New York, 2002.
[12] W. Junger, A.P. Leon, *Imputation of missing data in time series for air pollutants*. **Atmospheric Environment**, **102**, 2015, pp. 96-104.
[13] Ö. Çokluk, M. Kayri, *The Effects of Methods for Imputation for Missig Values on the Validity and Reliability of Scales*. **Educational Sciences: Theory & Practice**, **11**(1), 2011, pp. 303-309.
[14] N.A. Zakaria, *Imputation Methods for Filling the Long Interval of Missing Observations in Air Pollution Data and Meteorological Dataset*, **Thesis of Master of Science**. Universiti Malaysia Perlis, 2018.
[15] N.M. Noor, A.S. Yahaya, N.A. Ramli, M.M.A. Abdullah, *Filling the Missing Data of Air Pollutant Concentration Using Single Imputation Methods*, **Applied Mechanics and Materials**, **754-755**, 2015, pp. 923-932.
[16] M.N. Norazian, A.S. Shukri, R.N. Azam, A.M.M. Al Bakri, *Estimation of missing values in air pollution data using single imputation techniques*, **Science Asia, 34**(3), 2008, pp. 341-345.
[17] M. Akram, R.J. Hyndman, J.K. Ord, *Exponential Smoothing And Non-Negative Data*, **Australian & New Zealand Journal of Statistics**, **51***(*4), 2009, pp. 415-432.
[18] S. Glen, **Exponential Smoothing: Definition of Simple, Double and Triple,** Retrieved Sept.28.2018. https://www.statisticshowto.datasciencecentral.com/exponential-smoothing/
[19] P. Luo, M. Zhang, Y. Liu, D. Han, Q. Li, *A moving average filter based method of performance improvement for ultraviolet communication system*, **2012 8th International Symposium on Communication Systems, Networks & Digital Signal Processing** (CSNDSP), 18-20 July 2012, Poznan, Poland, 2012.
[20] S. Moss, **Expectation maximization to manage missing data**, Retrieved September 28, 2018, from https://www.sicotests.com/psyarticle.asp?id=267
[21] J.L. Schafer, **Analysis of incomplete multivariate data. Monographs on statistics and applied probability**, Chapman & Hall/CRC. London, 1997.

Received: January 18, 2018
Accepted: November, 20, 2019