Pertanika J. Sci. & Technol. 29 (4): 2625 - 2643 (2021)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

A Comparative Study of Several EOF Based Imputation Methods for Long Gap Missing Values in a Single-Site Temporal Time Dependent (SSTTD) Air Quality (PM10) Dataset

Shamihah Muhammad Ghazali[1]*, Norshahida Shaadan[1,2] and Zainura Idrus[1]

[1]Center for Statistical and Decision Science Studies, Faculty of Computer & Mathematical Sciences, Universiti Teknologi MARA, 40450, Shah Alam, Selangor, Malaysia

[2]Business Analytics Research Group, Faculty of Computer & Mathematical Sciences, Universiti Teknologi MARA, Bukit Ilmu, 18500, Machang. Kelantan.Malaysia

ABSTRACT

Missing values are often a major problem in many scientific fields of environmental research, leading to prediction inaccuracy and biased analysis results. This study compares the performance of existing Empirical Orthogonal Functions (EOF) based imputation methods. The EOF mean centred approach (EOF-mean) with several proposed EOF based methods, which include the EOF-median, EOF-trimmean and the newly applied Regularised Expectation-Maximisation Principal Component Analysis based method, namely R-EMPCA in estimating missing values for long gap sequence of missing values problem that exists in a Single Site Temporal Time-Dependent (SSTTD) multivariate structure air quality (PM10) data set. The study was conducted using real PM10 data set from the Klang air quality monitoring station. Performance assessment and evaluation of the methods were conducted via a simulation plan which was carried out according to four percentages (5, 10, 20 and 30) of missing values with respect to several long gap sequences (12, 24, 168 and 720) of missing points (hours). Based on several performance indicators such as RMSE, MAE, R-Square and AI, the results have shown that R-EMPCA outperformed the other methods. The results also conclude that the proposed EOF-median and EOF-trimmean have better performance than the existing EOF-mean based method in which EOF-trimmean is the best among the three. The methodology and findings of this study contribute as a solution to the problem of missing values with long gap sequences for the SSTTD data set.

ARTICLE INFO

Article history:
Received: 16 April 2021
Accepted: 05 July 2021
Published: 08 October 2021

DOI: https://doi.org/10.47836/pjst.29.4.21

E-mail addresses:
shamihah.ghazali@gmail.com (Shamihah Muhammad Ghazali)
shahida@tmsk.uitm.edu.my (Norshahida Shaadan)
zainura@tmsk.uitm.edu.my (Zainura Idrus)
* Corresponding author

Keywords: Air quality, empirical orthogonal functions, imputation, long gap missing values, PM10

ISSN: 0128-7680
e-ISSN: 2231-8526

INTRODUCTION

Usually, missing values in the air quality data set occurs when the values are unobserved, or the values were missing due to several reasons such as failure of monitoring instruments during some bad seasonal weather, computer system crashes, routine maintenance, human errors, calibration process and staying off-line for several days at the monitoring stations (Ghazali et al., 2020; Shaadan & Rahim, 2019). The impact of the missing data on the statistical analysis results depends on the mechanism that made the data to be missing and on the way the data analyst deals with them (Plaia & Bondi, 2006). In environmental studies, three types of missing data were taken into account, which is Missing Completely at Random (MCAR), Missing at Random (MAR) and Missing Not at Random (MNAR). For most air quality data sets, the mechanism of missing air quality data is MAR; the probability of a missing that a value is not dependent on the missing part themselves (Josse & Husson, 2016; Plaia & Bondi, 2006; Shaadan et al., 2015).

In the presence of missing data, treatment to replace the missing values is crucial in many fields, especially in air quality data sets where high percentages of data are being missed with long gap sequences (Ghazali et al., 2020). Many existing imputation methods that deal with missing values were proposed in the literature. The methods include a simple approach such as using mean or median substitution, a model-based approach including Regression-based imputation (REGEM), nearest neighbour (NN), K-nearest neighbour (KNN), expectation-maximisation (EM), maximum likelihood method and other hybrids methods (Junninen et al., 2004). Ruggieri et al. (2013) claimed that the Empirical Orthogonal Functions (EOF) based method is among the most promising method of imputing missing values to solve long gap sequences of missing data present in the air quality data set. Several applications of EOF methodology on the observed data by a Singular Value Decomposition (SVD) in handling the missing values have been discussed in several areas. However, the issue of long gap missingness in air quality data set and the experimentation is limited in the number. Moreover, long gaps of missingness often occur due to a longer sequence of hours, which is more than 6 hours and occur within several days or weeks and months (Bartzokas et al., 2003).

A study conducted by Beckers and Rixen (2003) was among the earliest investigation that used EOF calculations and procedures to fill in missing data. The study had used the imputation method in spatial-temporal data sets in the oceanographic field of studies. Next, another research was conducted by Sorjamaa et al. (2010) that proposed an improved EOF methodology for filling missing values in spatial-temporal climates data sets using EOF Pruning which was based on an original linear projection method. Among other closely related research that is continuously being explored was the paper by Beckers and Rixen (2003), Hannachi et al. (2007), Sorjamaa et al. (2010), Ruggieri et al. (2010) and Di Salvo et al. (2016). In Ruggieri et al. (2013), the authors proposed spatial-temporal Functional

Principal Component Analysis (FPCA) and used the EOF procedure to fill in long gap sequences of missing data to investigate the temporal variation of multiple pollutant datasets measured at multi-site and multivariate at the same time.

Even though the large proportion of missing values and long gap sequences of missing values have been considered in the above studies, noticeably, the scope was mostly focused on the imputation methods for Multi-site and Spatial-temporal Multivariate data structures. In a study by Bai et al. (2020), the authors propose a novel gap-filling method that used the EOF procedure; the method is known as diurnal cycle constrained empirical orthogonal function (DCCEOF) that used to fill in missing data gaps in hourly PM2.5 concentration of air quality data and the data existed the long gaps about 40% of days missing in the dataset.

However, the study data was focused on the time series of hourly PM2.5 datasets. Therefore, another kind of air quality data format identified as Single Site Temporal Time-Dependent (SSTTD) was rarely highlighted. Meanwhile, for Malaysia, the recorded format of air quality data set for an air quality monitoring station for a single pollutant is usually in the form of SSTTD. The pollutant observations were normally recorded and arranged into daily (row) by hourly (column) matrix format. In conclusion, the application of EOF based methods and their capacity has not yet been explored and compared to be used in the imputation analysis when long gap sequences are present in the SSTTD format air quality data sets. An example of missing values in SSTTD format is shown in Table 1.

Table 1
Example of daily by hourly recorded PM_{10} data within 24 hours

Day	Hour							
	Hour 1	Hour 2	Hour 3	Hour 4	Hour 5	.	.	Hour 24
Day 1	30	40	60	70	90	.	.	140
Day 2	20	NA	NA	NA	NA		.	120
Day 3	50	70	90	NA	NA	.	.	NA
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
Day n	70	80	NA	.	.	.	.	NA

Thus, to fill the gap, in this paper, several EOF based methods are employed to find the most appropriate method for a good reconstruction of long sequences of missing values in SSTTD multivariate air quality data format with the application for PM10 air pollutant data set.

MATERIAL AND METHODS

Data and Study Area

The data used in this study is a real secondary air quality data of particulate matter with a size 10 micrometre and smaller called PM10 measured in ugm^{-3} of Klang air quality

monitoring station. The data was obtained from the Air Quality Division of the Department of Environmental (DOE) Malaysia. The data was recorded using a Continuous Ambient Air Quality Monitoring (CAQM) system by Alam Sekitar Malaysia Sdn. Bhd. (ASMA), a private sector authorised by the DOE. For experimentation analysis in this study, a complete data set was identified for Klang station, consisting of 479 days observations with 11,496 hourly cell records from 1st June 2014 up to 24th April 2015 and will be treated as reference data.

Figure 1 shows the location of Klang station in the Malaysia map with longitude and latitude (N03°00.620', E101° 24.484'). Klang station is located at Sekolah Menengah Perempuan Raja Zarina, Klang, Selangor. The area is surrounded by the crowded industries, residential and commercial areas. Klang station was chosen in this study because this station is among the popular stations with high PM_{10} levels recorded by DOE.

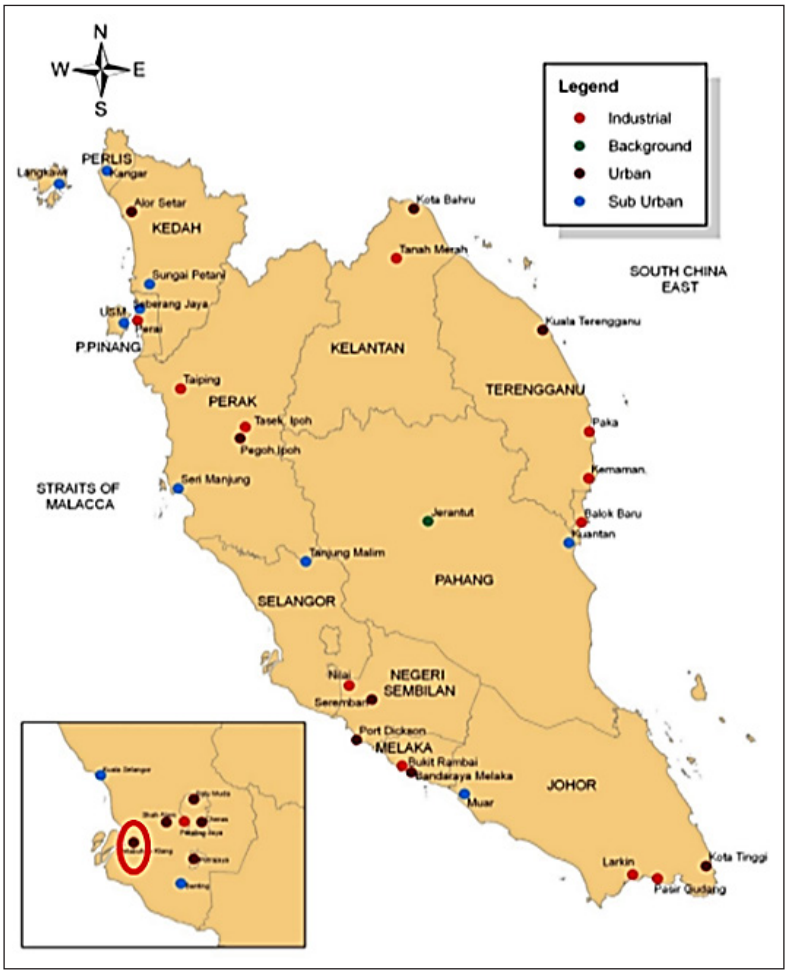


Figure 1. Location maps of air quality monitoring stations, including Klang station (i.e. highlighted in red circle). (Source: Malaysia Environmental Quality Report, 2013)

Methodology Framework

In order to achieve the research objective, the following steps of procedure as depicted in Figure 2 was employed in this study. The research methodology consists of four phases of stages; phase 1: Obtaining reference data, phase 2: Generating missing data pattern, phase 3: Imputing missing values and phase 4: Performance comparison.

Phase 1 step aims to prepare a reference data set for experimentation purposes to compare imputation methods. The reference data must be such a compulsory procedure to validate the performance of the imputation method with better accuracy (Shaadan et al., 2015). This study used a whole PM10 data set of hourly-recorded observations at the Klang air quality monitoring station as a reference data set. Selecting the reference data for this study begins with understanding and viewing the whole structure of the SSTTD data and only selecting the data with a complete case. As mentioned earlier, the selected reference

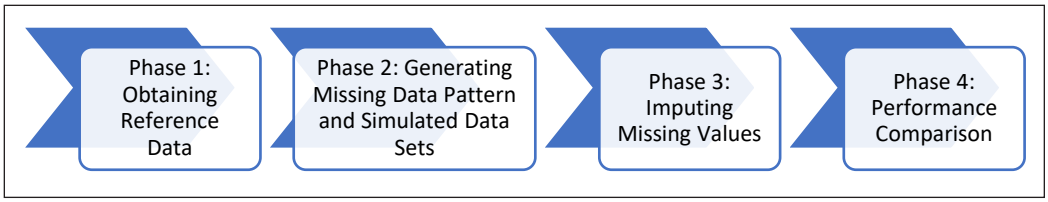


Figure 2. The methodology framework for research analysis

data for Klang station consists of 479 days observations with 11,496 hourly cell records from 1st June 2014 up to 24th April 2015. In phase 2, the simulations of artificial missing data set with several designed patterns were conducted. The patterns were generated according to several percentages of missing values of 5%, 10% and 30% with a different gap size of a sequence of missing points (i.e. hourly points) within 12 (half day), 24 (1 day), 168 (1 week) and 720 (1 month) period. Therefore, 16 missing values patterns need to be designed in this study, as shown in Table 2. Each design pattern will be generated into 100 simulated or artificial datasets.

In the next step, in phase 3, missing values imputation with the existing EOF-based method and several proposed EOF-

Table 2
Sixteen different patterns of missing generated data

Pattern of missingness	Percentages of missingness (%)	Gap length of missingness (hours)
P05_G12	5	12 (half day)
P05_G24	5	24 (1 day)
P05_G168	5	168 (1 week)
P05_G720	5	720 (1 month)
P10_G12	10	12 (half day)
P10_G24	10	24 (1 day)
P10_G168	10	168 (1 week)
P10_G720	10	720 (1 month)
P20_G12	20	12 (half day)
P20_G24	20	24 (1 day)
P20_G168	20	168 (1 week)
P20_G720	20	720 (1 month)
P30_G12	30	12 (half day)
P30_G24	30	24 (1 day)
P30_G168	30	168 (1 week)
P30_G720	30	720 (1 month)

based methods were applied. The analysis was executed on the 100 simulated data sets for each missing pattern obtained in phase 2. The number of simulated missing data sets of 100 for each pattern is decided to be used, following the experimentation conducted by the study of Di Salvo et al. (2016) and Shaadan et al. (2015). Other researchers could also increase the size of the simulated data sets for further detailing the sensitivity analysis, but this sensitivity analysis is not the scope of this research. In this study, the objective of this experimentation based on these 100 simulated data sets for each different designed pattern is to evaluate the performance of the imputation methods towards the consistency of the performance results. Thus, further, validate the performance obtained based on the average value approach, which at the same time can be used to evaluate the performance of the methods towards the complexity of the missing patterns.

Several performance indicators will then be applied to assess the performance of the imputation at phase 4, which includes Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Coefficient of Determination (R-Square) and Agreement Index (AI).

Imputation Methods

Empirical Orthogonal Functions (EOF) Method. The Empirical Orthogonal Functions is a deterministic method for reconstructing the new data matrix through data reduction and identifying temporal variation relationships in the data. In solving the missing values, the EOF allows a linear, continuous projection to a high-dimensional space. The EOF method is performed using Singular Value Decomposition (SVD) by extracting the salient empirical modes of variation from the temporal dependent singular vectors of the data matrix and construct a new set of variables that capture most of the observed variance from the data through a linear combination of the original variables. In understanding the concept of EOF, let us consider a data matrix **X** containing the observations, which is arranged such that the element of SSTTD data field *t, h* of the matrix called $x(t,h)$ where *t* and *h* denote respectively time and hours position and M is the number of modes contained in the field, using an optimal set of basis functions of temporal dependent $U_k(s)$ and expansion functions of time $C_k(t)$, as below in Equation 1:

$$x(t,h)=\sum_{k=1}^{M} C_k(t)U_k(s) \qquad (1)$$

In this study, the EOF is performed on SVD. The SVD aims to extract the loading of the principal components, EOFs time (score) and EOFs temporal (loading). SVD is computed for the 2 dimensional of temporal dependent singular vectors of the n x p data matrix **X**, where n is the time-series (days), and p is the temporal-dependent (hours), U and V are a collection of eigendecomposition of vectors of **X**, as in Equation 2:

$$\hat{\mathbf{X}} = \mathbf{UDV}^{*} = \sum_{k=1}^{r} \rho_k a_{tk} u_k \qquad (2)$$

The column $\mathbf{a_k} = \left(a_{1k}, a_{2k}, \ldots, a_{nk}\right)$ of **U** which the score and $\mathbf{u_k} = \left(u_{k1}, u_{k2}, \ldots, u_{kp}\right)$ of **V** which the loading are respectively the left and right singular vectors of the data matrix **X**. **D** is the diagonal matrix with the singular values ρ in its diagonal, the diagonal elements are $\lambda_1 \geq \lambda_2, \ldots, \geq \lambda_r \geq 0$ of **D**, are singular values of **X**. Where r is the smaller dimension of **X** with $r \leq \min\left(n, p\right)$ is the rank of **X**. The singular values and the singular vectors have been sorted in decreasing order.

From the SVD, the EOF removes the noise from the data. Only the selected singular values and vectors are used for the reconstruction of a new data matrix. This study selects the optimal number of dimension *k'th* using Generalised Cross-Validation (GCV) methods. The GCV value can be interpreted as a classical model selection criterion where the residual sum of squares is penalised by the number of degrees of freedom. Equation 3 is as follows:

$$GCV\left(S\right) = \frac{np \sum_{i=1}^{n} \sum_{j=1}^{p} \left(x_{ij} - \left(\hat{x}_{ij}\right)^{S}\right)^{2}}{np - p - nS - pS + S^{2} + S} \qquad (3)$$

The EOF cannot be directly used with a database that contains missing values. In the common practice, in the existing EOF imputation method, the column mean was normally treated as the initial value for the missing cells (Beckers & Rixen, 2003; Ruggieri et al., 2010; Sorjamaa et al., 2010). However, the existing EOF imputation method has a drawback because it uses data matrix centralisation based on statistic mean for EOF computation. To be applied for the air quality dataset, the existing EOF need to be improved because the dataset often consists of extreme observations due to climatic variations and random processes. In this study, a robust statistic of statistic median and trimmed mean is employed in the matrix centralisation computation and proposed as initial values for the missing values. In this paper, four EOF-based imputation methods are introduced, and the capability of the methods for estimating missing values for long gap missingness problems in Malaysia air quality of SSTTD multivariate datasets is investigated. The existing of EOF method based on the mean (EOF-mean) is compared with the several proposed EOF based on median (EOF-median), EOF based on the trimmed mean (EOF-trimmean) and the newly applied Regularised Expectation-Maximisation Principal Component Analysis (R-EMPCA), which an iterative-based imputation method that is iteratively performing the EOF analysis by means of EM algorithm on the incomplete data sets. Up until now, the performance of R-EMPCA is not yet being explored for solving long gap missingness problems in air quality with SSTTD datasets.

The Existing method: EOF based on the Mean (EOF-mean) Method. Given that a few data points are missing with a long gap of missingness, the aim was to replace the missing value x_{miss} at time t with a value on the estimated data point from EOF based imputation, which $x_{ij}(t,h)$ at the same time t is missing, where $\hat{x}_{miss} = x_{ij}(t,h)$. This procedure of EOF-mean starts from allows the generated missing values at time *t* to be initially replaced with the mean of the observed values column. Next, the completed data are centralised using the mean centralisation. Then, the EOF procedure is applied to the centralised data matrix, and a new reconstructed matrix of the EOF-mean is built. Finally, the formula for the initial values of the mean column is written as Equation 4:

$$Mean, \overline{x} = \frac{\sum_{k=1}^{n} x_{ki.}}{n} \tag{4}$$

The Proposed Methods: EOF based on Median (EOF-median) and EOF based on Trimmed Mean (EOF-trimmean) Methods. These methods propose an enhancement approach on the existing EOF-mean by using a different strategy in the initialisation step. The EOF-median and EOF-trimmean initially replace the missing values using robust median statistics trimmed mean, respectively, before performing any EOF procedure. The enhancement focuses on using the values from the median and trimmed mean of the observed values as the initial values to replace missing values in the data matrix **X**. These initial values are calculated from the available data in the dataset to form a completed data matrix. Finally, the completed data are centralised using the median and trimmed mean centralisation, respectively. The formula for the initial values of the median column is written as Equation 5:

$$Median, m = l + \left(\frac{\frac{N}{2} - F_l}{f_m} \right) \times C \tag{5}$$

where *l* is the lower-class boundary of the median class, N is the total frequency, F_l is the cumulative frequency before the median class, f_m is the class width of the median class and is the frequency of the median class. Another proposed method is EOF-trimmean. The formula for the initial values of the trimmed mean column is written as Equation 6:

$$Trimmedmean,\ T = \frac{1}{|R|} * \sum_{|k|+1}^{n-|k|} X_{ij} \tag{6}$$

where n is the number of observations, k is an integer of the trim proportion with the calculation of $k = n\alpha$ with α the percentages to trim, and R is the denominator of trimmed

mean where $R = n - 2k$. In particular, the algorithm of EOF-based methods to impute the missing values using the different initialisation approach by applying three different initial values, which are column mean, the proposed median and trimmed mean, can be stated as follows in Table 3.

Table 3
Algorithm of EOF-mean, EOF-median and EOF-trimmean methods

Start
Step 1: Identify the missing values, x_{miss} in the data matrix $\mathbf{X}_M$.
Step 2: Initial values are filled into missing using the column mean/median/trimmed mean (hours), of the data matrix **X** and turned into a new completed matrix $\mathbf{X}_M$.
Step 3: After the initial value replacement, centralised the data matrix $\mathbf{X}_M$ by subtracting the new completed matrix $\mathbf{X}_M$ using the column mean/median/trimmed mean of $\mathbf{X}_M$.
Step 4: Computed a Singular Value Decomposition (SVD) in Equation 2 on the centralised matrix $\mathbf{X}_M$.
a) The loading and scores EOFs **U** and **V** are extracted from the SVD
b) An optimal number of EOFs are selected using cross validation method of GCV formula in Equation 3.
The selected EOF loadings and EOF scores are used to make the reconstruction by multiply the loadings and scores and adding the subtracted column mean/median/trimmean to form a new reconstructed data matrix $\mathbf{X}_M$.
Step 5: Replace the missing values, x_{miss} in the data matrix **X** by their new estimated value obtained from the reconstructed data matrix $\mathbf{X}_M$ of EOF-mean/ EOF-median/ EOF-trimmean.
End

The Newly Applied: Regularised Expectation-Maximisation Principal Component Analysis (R-EMPCA) Method. Another application of the EOF-based imputation method proposed in this study to solve the long gap of missingness is the Regularised Expectation Maximisation Principal Component Analysis (R-EMPCA), with the regularised iterative EOF approach that was previously introduced in Josse & Husson (2016). However, this method is not yet explored for solving the long gap missing data problem.

Generally, the regularised iterative EOF is very similar to iterative EOF. Both methods use an iterative approach and are based on the EOF model, extracting the EOF score and loading. This regularised iterative method starts with the initialisation step of the missing values by mean values. Then, an estimation step of the parameters where the appropriate optimal number of EOFs modes is predefined from the temporal covariance matrix. The third step is the imputation step of missing values. The estimation of the mean matrix for the missing values is to be updated after each iteration during the imputation process. Finally, the last step is to reconstruct the new matrix of completed data using the updated or converge mean matrix from the imputation step. The R-EMPCA has extended the normal EOF analysis by replacing the centralisation process by using a weighted least squares criterion as in Equation 7:

$$\mathbf{W}_{n\times p} * (\mathbf{X}_{n\times p} - \hat{\mathbf{X}}) \,||_2^2 : \mathbf{rank(r)} \leq \mathbf{S} \tag{7}$$

Then the missing values are imputed in an iterative loop with the fitted matrix with noise variance as Equation 8:

$$\hat{x}_{ij} = \sum_{s=1}^{S} \left(\sqrt{\lambda_s^{\ell}} - \frac{(\hat{\sigma}^2)^{\ell}}{\sqrt{\lambda_s^{\ell}}} \right) \times u_{is}^{\ell} v_{js}^{\ell}, \tag{8}$$

The noise variance estimated as Equation 9:

$$(\hat{\boldsymbol{\sigma}}^2)^{\ell} = \frac{|| \mathbf{X}^{\ell-1} - \mathbf{U}^{\ell}\mathbf{D}^{\ell}(\mathbf{V}^{\ell})^{*} ||^2}{np - n\mathbf{S} - p\mathbf{S} + \mathbf{S}^2}, \tag{9}$$

$\mathbf{1}_{n\times p}$ is a matrix filled with values one. In addition, only the missing values are replaced with estimated values, and the mean matrix is re-centred and updated after each iteration to give the same weight to each variable. Then, the estimation and imputation steps are repeated until the difference between two successive estimated matrices is less than the threshold or the iteration number exceeds the maximum fixed iterations. The algorithm for R-EMPCA is shown in Table 4. Thus, the iteration will make full use of useful information in the process of missing values imputation.

Table 4
Summary of the R-EMPCA algorithm of the EOF based method by applying an iterative approach for filling the missing values

Start

Step 1: An optimal number of EOFs are selected using cross validation method of GCV formula in Equation 3

Step 2: Identify the missing values, x_{miss} in the data matrix **X**.

Step 3: Initialisation $\ell = 0$. Initial values are substituted into missing values using the mean of the column (hours), $\bar{x}_{mean}$ of the original data matrix **X**.

Step 4: The step $\ell \geq 1$ are:

- a) Performing the reconstruction of PCA analysis to estimates EOF scores and loadings using the SVD formula in Equation 2.
- b) Missing values are imputed with the fitted values with the noise variance estimated. Then, update and replace the missing values in the data matrix **X** by their new estimated values from the reconstruction.

Step 5: Step at 4(a) of estimation of parameters by SVD and 4(b) the imputation step are repeated until the convergence criterion is fulfilled.

Step 6: Replace the missing values, x_{miss} in the data matrix **X** with the new estimated value, $\hat{\mathbf{X}}_{R.EMPCA}$ obtained from the final reconstruction of data matrix

Performance Evaluation

The performance indicator involved in this study is Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Coefficient of Determination (R-Square) and Agreement

Index (AI) from (Junninen et al., 2004) are considered. The formulas for the performance indicator are given by the following Equations 10-13:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \hat{x}_i\right)^2}{n}} \tag{10}$$

$$\text{MAE} = \frac{\sum_{i=1}^{n}\left|x_i - \hat{x}_i\right|^2}{n} \tag{11}$$

$$\text{R-Square} = \left[\frac{1}{N}\frac{\sum_{i=1}^{n}\left(\hat{x}_i - \hat{\bar{x}}_i\right)\left(x_{i-}\bar{x}_i\right)}{\sigma_{\hat{y}}\sigma_y}\right]^2 \tag{12}$$

$$\text{AI} = 1 - \left[\frac{\sum_{i=1}^{n}\left(\hat{x}_{i-}x_i\right)^2}{\sum_{i=1}^{n}\left(\left|\hat{x}_i - \bar{x}_i\right| + \left|x_i - \bar{x}_i\right|\right)^2}\right] \tag{13}$$

where $\hat{x}_i$ is the observed value, $\hat{x}_i$ is the imputed value, $\bar{x}_i$ indicates the average of the actual data, and $\hat{\bar{x}}_i$ is the average of the imputed data with σ_y and $\sigma_{\hat{y}}$ are their standard deviations respectively. RMSE and MAE are used to assess the accuracy of the methods by looking at the residuals (i.e. the difference between the imputed and the observed values). At the same time, R-Square measures the imputation methods capability in predicting or estimating missing observation while AI measures the correlation between the imputed and the observed value. The ideal imputation method is the one that gives small error measures; the RMSE and MAE and high R-Square and AI.

RESULTS AND DISCUSSION

This section discusses the performance results of the imputation methods. The imputation methods were applied for the Klang station dataset, and the experiment was conducted for each missing data pattern using all four EOF based imputation methods. The results were then calculated as average results of the 100 simulated datasets for each missing pattern. The following Table 5 and Figure 3 represent the average score of each performance indicator based on error measures, the RMSE and MAE, while Figure 4 shows the performance based on R-Square and AI. The complexity evaluation was conducted according to the performance of the methods with respect to the different patterns of missing values with different levels of % and gap size–from low to higher percentage and from small long gap to larger long gap.

Table 5 indicates the average RMSE, MAE, R-Square and AI values for the four EOF methods computed from 100 artificial data sets for each missing pattern. Overall, the

R-EMPCA method has a very excellent performance indicated by the lowest RMSE and MSE values and the highest R-Square compared to the other three methods; the EOF-mean, EOF-median and the EOF-trimmean. Furthermore, it shows a stronger capability of the R-EMPCA method to estimate the missing values with higher accuracy and higher predictive power for each type of missing pattern except that for pattern P30_G720, whereby all methods are shown to have quite a similar performance for this pattern. Figures 3 and 4 also shows that among the EOF-mean, EOF-median and EOF-trimmean, the proposed EOF-trimmean method having a better performance, which are indicated by lower RMSE and MAE values and higher AI and R-Square values in comparison with EOF-mean and EOF-median methods. Therefore, this study has proven that the R-EMPCA method is the most suitable imputation method for a data set with a long missingness gap.

Noticeably, in this investigation, the value of R-square for R-EMPCA is not more than 0.7 for all the sixteen patterns of missingness. At the same time, the AI is relatively high with 0.78744 for the P05_G12 and become lower as this result depends on the pattern of missingness. It is also observed that there is a reduction in the performance of the methods when the proportion of missingness increases as the gap size increases. These findings are supported by Junger and Ponce de Leon (2015), who mentioned that the performance of the estimated value would be decreased when the missing values and the gap size increase in the data set. Even though the R-EMPCA method is found the best in terms of performance, it is believed that R-Square and the AI values recorded are due to the performance's ability when the methods experimented within the condition of long gap missingness situation. The results would be much better when the method is applied for a not so complex missing data set (i.e. data set with a small percentage and short gap sequence of missingness).

EOF-mean was the worst identified imputation method among the EOF-based imputation methods, while the proposed EOF-median has moderate performance. The following Figures 5-7 provide the analysis of the consistency of the results.

Table 5
Performances of four methods of imputation according to missing data pattern

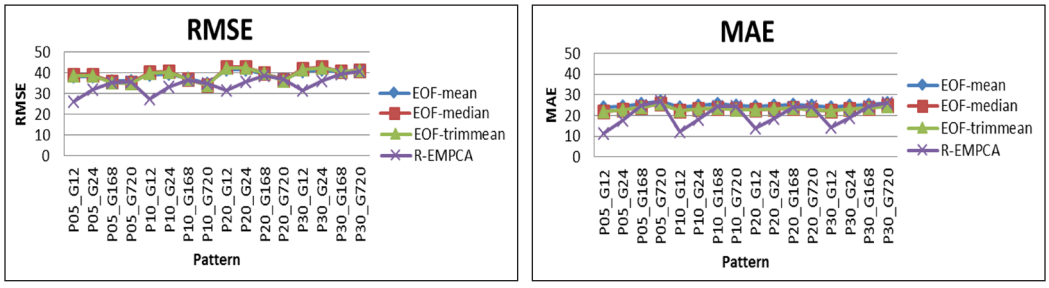
Patters	Performance Indicators	Klang Station			
		EOF-mean	EOF-median	EOF-trimmean	R-EMPCA
P05_G12	RMSE	38.11721	38.97415	38.70757	25.95430
	MAE	23.87387	22.07330	22.07434	11.46257
	R-Square	0.00835	0.00831	0.00904	0.59141
	AI	0.15670	0.23467	0.21502	0.78744
P05_G24	RMSE	38.17311	38.95518	38.68937	31.85556
	MAE	24.51937	22.82607	22.82319	17.80248
	R-Square	0.00359	0.00334	0.00383	0.33962
	AI	0.20054	0.23440	0.21749	0.61509
P05_G168	RMSE	35.96443	35.55519	35.37376	35.07408

Table 5 (*continue*)

Patters	Performance Indicators	Klang Station			
		EOF-mean	EOF-median	EOF-trimmean	R-EMPCA
	MAE	25.80377	24.20227	24.17143	24.81692
	R-Square	0.00033	0.00037	0.00050	0.05371
	AI	0.32522	0.30319	0.29982	0.39352
P05_G720	RMSE	35.92638	35.08161	34.91106	35.77932
	MAE	27.10393	25.70282	25.60557	26.94752
	R-Square	0.00038	0.00012	0.00023	0.01623
	AI	0.31892	0.33314	0.32900	0.33561
P10_G12	RMSE	39.01976	40.12672	39.85329	27.39822
	MAE	24.05786	22.29393	22.29608	12.15135
	R-Square	0.01373	0.01258	0.01378	0.55831
	AI	0.13102	0.23574	0.21656	0.77262
P10_G24	RMSE	39.29177	40.49170	40.19496	33.11227
	MAE	24.74710	23.15876	23.14616	18.23082
	R-Square	0.00616	0.00505	0.00567	0.32431
	AI	0.14871	0.24123	0.22060	0.59886
P10_G168	RMSE	37.30235	37.02084	36.84009	36.42588
	MAE	25.52515	23.44938	23.47681	24.51017
	R-Square	0.00049	0.00043	0.00077	0.05335
	AI	0.27288	0.24343	0.23635	0.34863
P10_G720	RMSE	34.89105	34.16551	33.98925	34.73723
	MAE	24.72334	23.01038	22.95128	24.55034
	R-Square	0.00014	0.00007	0.00017	0.01111
	AI	0.31434	0.29831	0.29199	0.33000
P20_G12	RMSE	41.37989	42.57698	42.29561	31.44141
	MAE	24.34981	22.69193	22.68838	13.96214
	R-Square	0.02221	0.02054	0.02184	0.46850
	AI	0.13507	0.24032	0.22357	0.71544
P20_G24	RMSE	41.47086	42.63278	42.35331	35.52967
	MAE	24.96108	23.25652	23.25844	18.62116
	R-Square	0.00953	0.00754	0.00844	0.29630
	AI	0.12609	0.22980	0.21189	0.56698
P20_G168	RMSE	39.36883	39.98125	39.72529	38.47219
	MAE	25.08108	23.43837	23.41567	24.09854
	R-Square	0.00049	0.00042	0.00060	0.05056
	AI	0.19992	0.24278	0.22545	0.28785
P20_G720	RMSE	37.01653	36.37736	36.20709	36.84366
	MAE	24.82146	22.53450	22.54400	24.62451
	R-Square	0.00014	0.00005	0.00018	0.00971

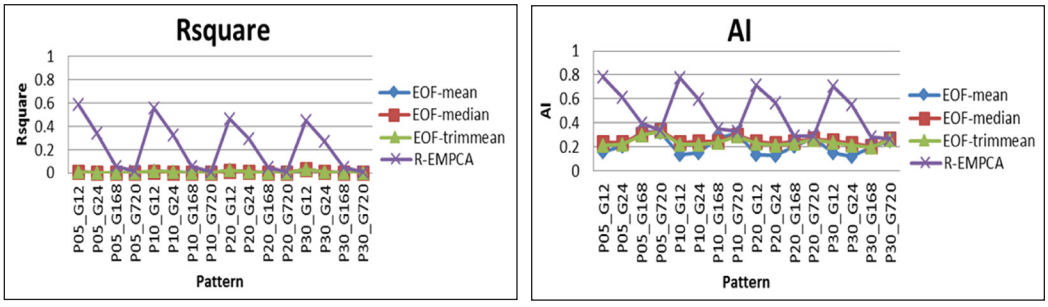
Table 5 (*continue*)

Patters	Performance Indicators	Klang Station			
		EOF-mean	EOF-median	EOF-trimmean	R-EMPCA
	AI	0.27432	0.26486	0.25474	0.29133
P30_G12	RMSE	40.66486	41.86543	41.58401	31.40857
	MAE	24.02346	22.40867	22.40350	14.25771
	R-Square	0.03081	0.02896	0.03038	0.44710
	AI	0.14768	0.25175	0.23628	0.70575
P30_G24	RMSE	41.52809	42.65726	42.37999	36.08452
	MAE	24.76144	23.02723	23.03267	18.84405
	R-Square	0.01278	0.01076	0.01168	0.27366
	AI	0.11905	0.22828	0.20970	0.54812
P30_G168	RMSE	40.38344	40.76528	40.54024	39.51843
	MAE	25.45146	23.41158	23.43521	24.44615
	R-Square	0.00070	0.00044	0.00078	0.04593
	AI	0.19293	0.20585	0.19629	0.27885
P30_G720	RMSE	40.96781	41.25684	41.00958	40.79544
	MAE	26.15329	24.69441	24.62649	25.94850
	R-Square	0.00012	0.00005	0.00015	0.01002
	AI	0.24747	0.26452	0.25627	0.26770



(a) (b)

Figure 3. Imputation methods performance based on average error measures: (a) RMSE; and (b) MAE



(a) (b)

Figure 4. Imputation methods performance based on correlational measures: (a) R-Square; and (b) AI

As depicted in Figure 5, the distribution of RMSE for each method shows an increasing pattern when the proportion and the gap size increase. Among the four methods, R-EMPCA clearly shows the lowest RMSE median for the data set with gap sizes 12 and 24 at levels of missing percentage (5, 10, 20, and 30). However, for the data set with a larger gap size, 168 and 720, R-EMPCA has shown a slight reduction in the performance indicated by a slight increase in RMSE score. The RMSE distribution also has a consistently decreasing performance pattern for EOF-mean, EOF-median and EOF-trimmean methods when the gap size increases. Figure 5 also shows that the distribution pattern of RMSE for all methods are similar for the largest missing gap size with 720 consecutive missing points (i.e. within a month duration of missingness).

In contrast, R-EMPCA keeps a better performance at different percentages of missingness and gap sizes of missingness when the RMSE produce the lowest error amongst other methods. It can be seen from the boxplots that EOF-mean, EOF-median and EOF-trimmean at all levels of missingness patterns have quite a similar pattern of boxplot where the lower and the upper whisker show a similar range. In addition, EOF-trimmean shows a lower median and variance compared to the EOF-mean and EOF-median. On the other hand, the median of EOF-mean shows slightly higher compared to the EOF-median ad EOF-trimmean. Therefore, EOF-trimmean outperforms the EOF-median in estimating missing values for long gap missingness, whereas EOF-mean demonstrates the worst performances.

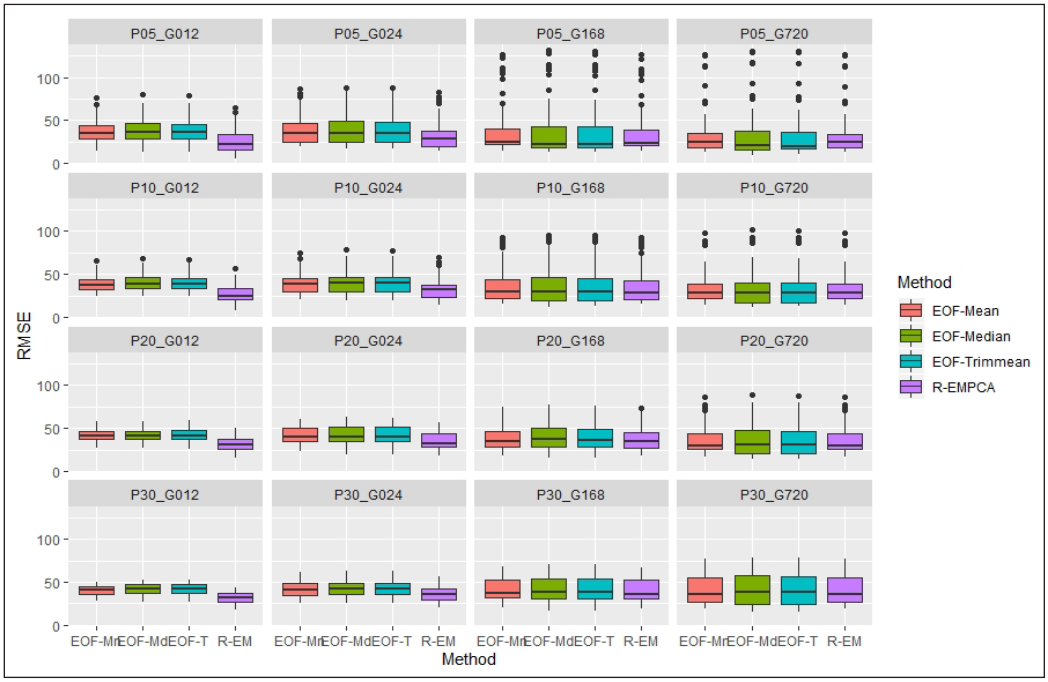


Figure 5. Box plot on RMSE distribution

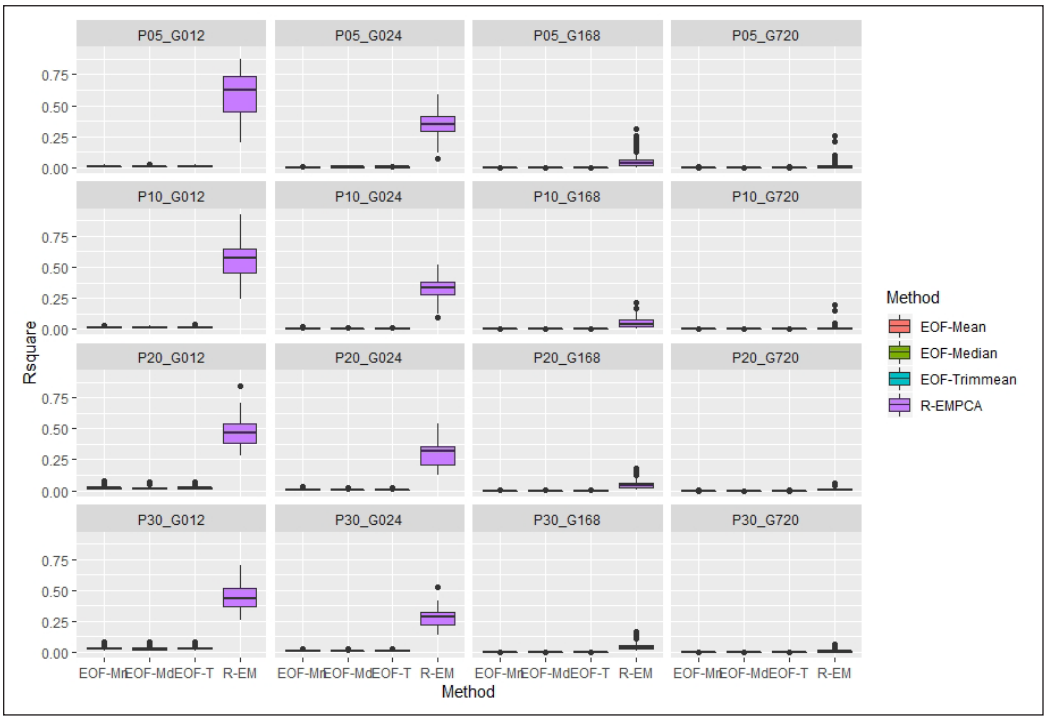


Figure 6. Box plot on R-Square distribution

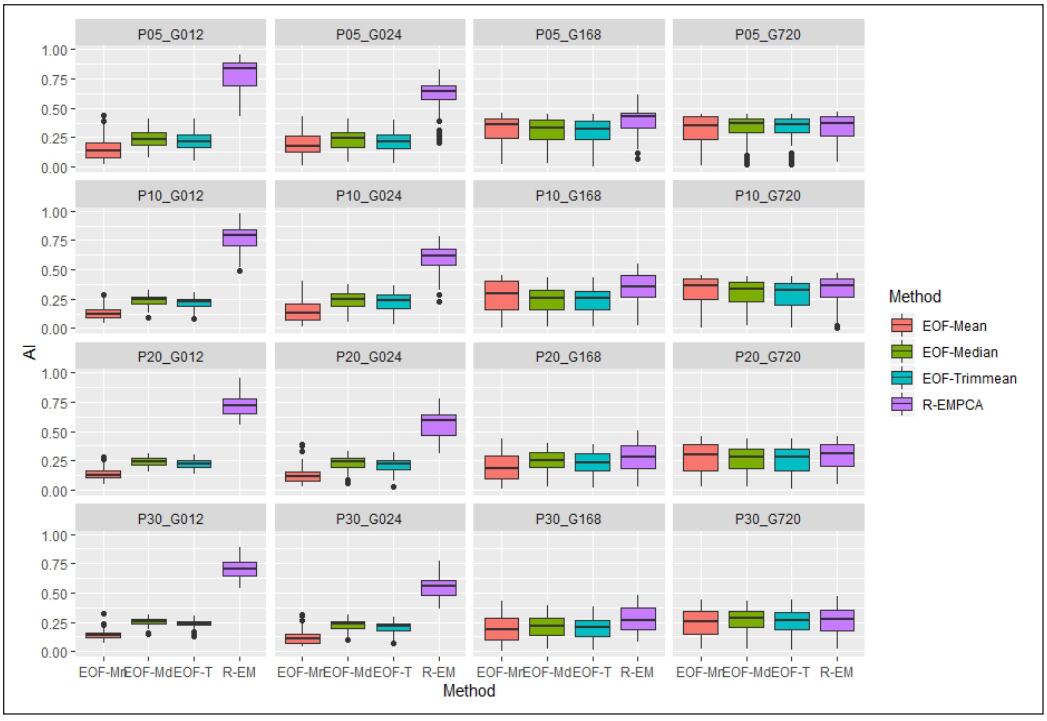


Figure 7. Box plot on AI distribution

The summary of the boxplot of R-Square score distribution for four EOF based imputation methods is shown in Figure 6. A higher value of R-Square that close to 1 indicates that the estimated values were almost close to the observed values, the boxplot of R-Square showed that R-EMPCA outperforms EOF-mean, EOF-median and EOF-trimmean with the highest value of R-Square for all the patterns of missingness from 12 gaps to 720 gaps of missingness. This plot directly explained that R-EMPCA has a good performance in terms of accuracy measure of R-Square for the generated long gap of missing data. However, the other three EOF-mean, EOF-median and EOF-trimmean, show a very short and thin boxplot, indicates that these three methods have a very low median and variance below that 0.01.

The pattern is also similar for AI score distribution. As shown in Figure 7, R-EMPCA possessed a very excellent AI score (high score) for missing data set with 12 and 168 gap sizes. However, the performance gradually decreased as gap size increase at all levels of missing percentage. Among the EOF-mean, EOF-median and EOF-trimmean, the EOF-mean has the lowest AI median for gap size 12 and 168 but having equal performance with EOF-median and EOF-trimmean when gap size increase to 168 and 720, indicated by quite similar median values. The results shown in Figures 5 to 7 have provided evidence and justification for the consistency of the imputation results obtained to validate and support the findings as summarised from Figures 3 and 4. These results indicate that R-EMPCA has consistent results as the best imputation method. EOF-trimmean is the second-best imputation method for the long gap missingness problem in the air quality dataset in Klang station.

CONCLUSION

In this study, four EOF-based imputation methods, which are EOF-mean, EOF-median, EOF-trimmean and R-EMPCA, were used and compared for the treatment of long gap missing values problem for a Single-Site Temporal Time-Dependent (SSTTD) type of multivariate air quality (PM10) data using real data set of Klang air quality monitoring station. The results have found that R-EMPCA outperformed the other EOF based methods, including the existing method, EOF-mean and the two proposed methods, EOF-median and EOF-trimmean, in some long gap sizes. The performance of R-EMPCA has proven its superiority as the method having the best result for not so large missing gap but gradually decrease and become at par with the other methods for enormous gap size; such as for one month (720 hourly consecutive missing points). However, the results also lead to a conclusion that the proposed EOF-median and EOF-trimmean give better performance as compared to the existing EOF-mean based method. Overall, the R-EMPCA provides a realistic and promising way to handle the long gap missingness presented in multivariate hourly air quality (PM10) of SSTTD data sets. To conclude, the use of various applications

on the imputation techniques based on the characteristics of the air quality dataset is recommended. A more general comparison of this method with many other different methodologies such as smoothing techniques or functional data analysis approach will be conducted in the future to evaluate further the performance and accuracy of the R-EMPCA method in handling long gaps of missingness. The data set was used generally in the experimentation without considering the seasonal influence on the imputed values. However, for the application in practice, it is suggested to apply the method according to season.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the Department of Environment Malaysia (DOE) for providing the information and data.

REFERENCES

Bai, K., Li, K., Guo, J., Yang, Y., & Chang, N. B. (2020). Filling the gaps of in situ hourly PM2.5 concentration data with the aid of empirical orthogonal function analysis constrained by diurnal cycles. *Atmospheric Measurement Techniques*, *13*(3), 1213-1226. https://doi.org/10.5194/amt-13-1213-2020

Bartzokas, A., Darula, S., Kambezidis, H. D., & Kittler, R. (2003). Sky luminance distribution in Central Europe and the Mediterranean area during the winter period. *Journal of Atmospheric and Solar-Terrestrial Physics*, *65*(1), 113-119. https://doi.org/10.1016/S1364-6826(02)00283-3

Beckers, J. M., & Rixen, M. (2003). EOF calculations and data filling from incomplete oceanographic datasets. *Journal of Atmospheric and Oceanic Technology*, *20*(12), 1839-1856. https://doi.org/10.1175/1520-0426(2003)020<1839:ECADFF>2.0.CO;2

Di Salvo, F., Plaia, A., Ruggieri, M., & Agro, G. (2016). Empirical orthogonal function and functional data analysis procedures to impute long gaps in environmental data. In *Studies in Theoretical and Applied Statistics, Selected Papers of the Statistical Societies* (pp. 3-13). Springer. https://doi.org/10.1007/978-3-319-27274-0_1

Ghazali, S. M., Shaadan, N., & Idrus, Z. (2020). Missing data exploration in air quality data set using R-package data visualisation tools. *Bulletin of Electrical Engineering and Informatics*, *9*(2), 755-763. https://doi.org/10.11591/eei.v9i2.2088

Hannachi, A., Jolliffe, I. T., & Stephenson, D. B. (2007). Empirical orthogonal functions and related techniques in atmospheric science: A review. *International Journal of Climatology: A Journal of the Royal Meteorological Society, 27*(9), 1119-1152. https://doi.org/10.1002/joc.1499

Josse, J., & Husson, F. (2016). missMDA: A package for handling missing values in multivariate data analysis. *Journal of Statistical Software*, *70*(1), 1-31. https://doi.org/10.18637/jss.v070.i01

Junger, W. L., & Ponce de Leon, A. (2015). Imputation of missing data in time series for air pollutants. *Atmospheric Environment*, *102*, 96-104. https://doi.org/10.1016/j.atmosenv.2014.11.049

Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J., & Kolehmainen, M. (2004). Methods for imputation of missing values in air quality data sets. *Atmospheric Environment*, *38*(18), 2895-2907. https://doi.org/10.1016/j.atmosenv.2004.02.026

Malaysia Environmental Quality Report. (2013). *Air Quality*. Department of Environment Malaysia.

Plaia, A., & Bondı, A. L. (2006). Imputation of missing values in air quality data sets. In *XLIII Riunione Scientifica Della Società Italiana Di Statistica* (pp. 667-670). CLEUP Publishing.

Ruggieri, M., Plaia, A., Di Salvo, F., & Agró, G. (2013). Functional principal component analysis for the explorative analysis of multisite-multivariate air pollution time series with long gaps. *Journal of Applied Statistics*, *40*(4), 795-807. https://doi.org/10.1080/02664763.2012.754852

Ruggieri, M., Di Salvo, F., Plaia, A., & Agró, G. (2010). EOFs for gap filling in multivariate air quality data: a FDA approach. In *Compstat 2010* (pp. 1557-1564). Physica-Verlag.

Shaadan, N., Deni, S. M., & Jemain, A. A. (2015). Application of functional data analysis for the treatment of missing air quality data. *Sains Malaysiana*, *44*(10), 1531-1540. https://doi.org/10.17576/jsm-2015-4410-19

Shaadan, N., & Rahim, N. A. (2019). Imputation analysis for time series air quality (PM10) data set: A comparison of several methods. *In Journal of Physics: Conference Series* (Vol. 1366, No. 1, p. 012107). IOP Publishing. https://doi.org/10.1088/1742-6596/1366/1/012107

Sorjamaa, A., Lendasse, A., Cornet, Y., & Deleersnijder, E. (2010). An improved methodology for filling missing values in spatiotemporal climate data set. *Computational Geosciences*, *14*(1), 55-64. https://doi.org/10.1007/s10596-009-9132-3