Sains Malaysiana 44(3)(2015): 449–456

A Comparison of Various Imputation Methods for Missing Values in Air Quality Data

(Perbandingan Pelbagai Kaedah Imputasi bagi Data Lenyap untuk Data Kualiti Udara)

NURYAZMIN AHMAT ZAINURI*, ABDUL AZIZ JEMAIN & NORA MUDA

ABSTRACT

This paper presents various imputation methods for air quality data specifically in Malaysia. The main objective was to select the best method of imputation and to compare whether there was any difference in the methods used between stations in Peninsular Malaysia. Missing data for various cases are randomly simulated with 5, 10, 15, 20, 25 and 30% missing. Six methods used in this paper were mean and median substitution, expectation-maximization (EM) method, singular value decomposition (SVD), K-nearest neighbour (KNN) method and sequential K-nearest neighbour (SKNN) method. The performance of the imputations is compared using the performance indicator: The correlation coefficient (R), the index of agreement (d) and the mean absolute error (MAE). Based on the result obtained, it can be concluded that EM, KNN and SKNN are the three best methods. The same result are obtained for all the eight monitoring station used in this study.

Keywords: Imputation techniques; missing data; performance indicators

ABSTRAK

Kertas ini membincangkan pelbagai kaedah imputasi bagi rawatan data lenyap untuk data kualiti udara khususnya di Malaysia. Objektif utama kajian ini ialah memilih rawatan data lenyap yang terbaik dan juga perbandingan sama ada wujud perbezaan antara kaedah yang digunakan antara stesen di Semenanjung Malaysia. Pelbagai kes data lenyap telah dijana secara rawak iaitu dengan 5, 10, 15, 20, 25 dan 30% data lenyap. Enam kaedah rawatan data lenyap telah digunakan dalam kajian ini iaitu teknik berasaskan min, median, jangkaan pemaksimuman (EM), dekomposisi nilai tunggal (SVD), K-jiran terdekat (KNN) dan K-jujukan jiran terdekat (SKNN). Pemilihan teknik imputasi terbaik adalah berdasarkan kepada penunjuk prestasi yang menggunakan nilai pekali korelasi (R), indeks persetujuan (d) dan min ralat mutlak (MAE). Berdasarkan kepada keputusan yang diperoleh, dapat disimpulkan bahawa kaedah EM, KNN dan SKNN adalah tiga kaedah yang terbaik. Keputusan yang sama diperoleh bagi semua stesen yang digunakan dalam kajian ini.

Kata kunci: Data lenyap; penunjuk prestasi; teknik imputasi

INTRODUCTION

Missing data is a widespread problem in many fields such as longitudinal studies, experimental studies and also data obtain from surveys that may due to many reasons. This also includes environmental studies such as air quality data which were due to many reasons for example machine failure, human error and insufficient sampling. Complete data are required to perform statistical analysis such as in time series analysis, principal component analysis (PCA) and multivariate analysis. Data with missing value can cause a significant problem, for example in time series analysis; it requires continuous data in order to perform prediction. One approach to overcome this problem is the adoption of imputation techniques (Junninen et al. 2004).

However, an appropriate method for handling missing data depends on the pattern and on the missing data mechanism. This is important as it effects on how much the missing data will bias the results when performing statistical analysis. Generally, there are three types of missing data; namely missing complete at random (MCAR), missing at random (MAR) and not missing at random (NMAR) (Little & Rubin 2002). MCAR means that the missing data mechanism does not depend to the values of any variables in the data set, whether it is missing or observed. While for MAR, the cause of missing data is unrelated to the missing values, but may be related to the observed values of other variables. Based on Junninen et al. (2004), Plaia and Bondi (2006) and Pollice and Lasinio (2009), the missing data mechanism of air quality data is MAR.

In the last few decades a number of various imputation techniques to overcome the missing data have been proposed. One of the most popular approaches is by using listwise deletion method which simply removing the missing data and uses the remaining data set for analysis. Even though this method is easy to implement, Little and Rubin (2002) and Rubin (1987) have shown that by removing the missing values using listwise deletion method it can introduce a substantial biases in the study, especially when the mechanism of the missing data are not randomly distributed. Another common method is by using the average values for imputation of the missing values. This is the easiest imputation method where it imputes the mean value of each variable on the respective missing variables as an estimate of the missing value (Allison 2001). This method can lead to a problem of bias and large errors in the covariance matrix and this will affect the performance of the statistical modeling. Other approaches of imputation techniques that can be use are hot deck imputation which imply in the nearest neighbor method; proposed by Laaksonen (2000) and the imputation which based on the least squares and maximum likelihood estimation by Dempster et al. (1977) and Little and Rubin (2002).

In this paper, we investigated and make comparison on several imputation techniques in order to overcome the missing values in the air quality data sets. The techniques used are expectation-maximization (*EM*) method (Dempster et al. 1977), singular value decomposition (*SVD*) and *K*-nearest neighbour (*KNN*) method (Troyanskaya et al. 2001) and sequential *K*-nearest neighbour (*SKNN*) method (Kim et al. 2004). The techniques of mean and median substitution were also included for comparison. The main objectives of this study were to select the best imputation method for the air quality data specifically in Malaysia and to compare whether there was any difference in the methods used between the eight stations in Peninsular Malaysia. In order to compare the performance of the imputation methods, the correlation coefficient (*R*), the index agreement (*d*) and the mean absolute error (*MAE*) were used.

Data and Methods

DATA

In this study, the hourly ozone data with a length of three months which are obtained from the Department of Environment (DOE) were used. The data were taken from eight monitoring stations located at central, south and north region of Peninsular Malaysia. The list of stations considered is shown in Table 1.

The approach used for this study is a cross validation approach in order to investigate the efficiency of imputation methods (Gelman et al. 1998; Porter et al. 2009). In this approach, some values are removed at random from the original data set, replaced with the imputed values and are then compared for differences (Troyanskaya et al. 2001).

To implement the cross validation approach, the original data set from 8 monitoring stations are used which is referred as the observed data. Next, we select randomly some of the data say 5% data and the selected data are deleted. Here the deleted data was treated as missing and will be referred as test data set. This process will be repeated for different percentage such as 10, 15, 20, 25 and 30%, to get test data set with different percentage of missing. The same approach will be implemented to all 8 stations. Thus, there will be 6 test data sets for each station.

Then, the missing data will be imputed using various imputation techniques to recover the deleted values referred as imputed data. Finally, the imputed data obtained will then be compared to their corresponding observed values. Throughout the paper, the data was denoted as $\mathbf{X}$ of dimension $n \times p$, where the rows represent the number of days and the columns represent the number of hours.

IMPUTATION METHODS

There are various imputation methods available in the literature. In this study, six methods of imputation were compared to see which method is the best suit to impute missing values in the air quality data. Some of the selected methods are chosen because of the availability in the *R* package.

Mean and median imputation The mean and median imputation techniques are used in this study as a simple reference method. This techniques were used in the respective column in the data set where we find the mean or the median at a particular hour. For each of the column with missing values, the corresponding mean or median of the observed values are calculated and this values will be used to replace the missing value.

Expectation-maximization method The expectation-maximization (*EM*) method is a method with general approach for computing maximum likelihood estimates from incomplete data set. It consists of an iterative calculation which involved two steps; prediction and estimation. Firstly we discuss the basic concept of *EM*

TABLE 1. List of stations

Code	Region	Name of the stations	State
St1	Central	Gombak	Selangor
St2	Central	Kelang	Selangor
St3	Central	Petaling Jaya	Selangor
St4	Central	Kajang	Selangor
St5	South	Pasir Gudang	Johor
St6	South	Johor Bahru	Johor
St7	North	Perai	Pulau Pinang
St8	North	Universiti Sains Malaysia	Pulau Pinang

when data is complete; no missing value. Assume that there is n observations with p variables; $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$. If we partition $\mathbf{X}$ into two groups, $\mathbf{X}_{(p\times 1)} = \begin{bmatrix} \mathbf{X}^{1}_{(q\times 1)} \\ \cdots\cdots\cdots\cdots \\ \mathbf{X}^{(2)}_{((p-q)\times 1)} \end{bmatrix}$ then its mean vector and its covariance matrix is given as $\boldsymbol{\mu}_{(p\times 1)} = \begin{bmatrix} \boldsymbol{\mu}^{1}_{(q\times 1)} \\ \cdots\cdots\cdots\cdots \\ \boldsymbol{\mu}^{(2)}_{(p-q)\times 1)} \end{bmatrix}$ and $\sum_{(p\times p)} = \begin{bmatrix} \sum_{11}\limits_{(q\times q)} & & \sum_{12}\limits_{(q\times(p-q))} \\ \cdots\cdots\cdots & \vdots & \cdots\cdots\cdots \\ \sum_{12}\limits_{((p-q)\times q)} & & \sum_{22}\limits_{((p-q)\times(p-q))} \end{bmatrix}$ where $\mathbf{X}^{(1)}$ is distributed as $N_q(\mu^{(1)}, \Sigma_{11})$ and $\mathbf{X}^{(2)}$ is distributed as $N_{p-q}(\mu^{(2)}, \Sigma_{22})$. From this operation we can find $\mathbf{T}_1$ and $\mathbf{T}_2$ by equation: $\mathbf{T}_1 = \sum_{j=1}^{n} \mathbf{X}_j = n\bar{\mathbf{X}}$ and $\mathbf{T}_2 = \sum_{j=1}^{n} \mathbf{X}_j \mathbf{X}_j^{-1} = \sum_{j=1}^{n} \left(\mathbf{X}_j - \bar{\mathbf{X}}\right)\left(\mathbf{X}_j - \bar{\mathbf{X}}\right)' + n\bar{\mathbf{X}}\bar{\mathbf{X}}'$.

The mean and the covariance of the conditional distribution of $\mathbf{X}^{(1)}$ given $\mathbf{X}^{(2)}$ are then given as $\boldsymbol{\mu} = \mathrm{E}(\mathbf{X}^{(1)}|\mathbf{X}^{(2)}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \boldsymbol{\mu}^{(1)} + \sum_{12}\sum_{12}^{-1} (\mathbf{X}^{(2)} - \boldsymbol{\mu}(^{2)})$ and $\boldsymbol{\Sigma} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}\Sigma_{22}^{-1}\Sigma_{21}$ where $|\Sigma_{22}| > 0$. The next step in *EM* is when data containing missing values. An initial average $\tilde{\boldsymbol{\mu}}$ are calculated where these average will be used to substitute the missing values in the incomplete data. Then from this values, the initial covariance estimates, $\tilde{\boldsymbol{\Sigma}}$ are obtained. By using the initial estimates $\tilde{\boldsymbol{\mu}}$ and $\tilde{\boldsymbol{\Sigma}}$, the contributions of the missing values to $\mathbf{T}_1$ and $\mathbf{T}_2$ can be predicted. Here in the prediction step, for each column $\mathbf{x}_j$ with missing values, let $\mathbf{x}_j^{(1)}$ denote the missing values and $\mathbf{x}_j^{(2)}$ denote the observed values; $\mathbf{x}'_j = \left[\mathbf{x}_j^{(1)\prime}, \mathbf{x}_j^{(2)\prime}\right]$. From the initial estimate of $\tilde{\boldsymbol{\mu}}$ and $\tilde{\boldsymbol{\Sigma}}$ which were obtained from the observed values, the missing values are estimate by using the mean of the conditional normal distribution of $\mathbf{x}^{(1)}$ given $\mathbf{x}^{(2)}$. The equation is given by

$$\tilde{\mathbf{x}}_j^{(1)} = E\left(\mathbf{X}_j^{(1)} \middle| \mathbf{x}_j^{(2)}; \tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}\right) = \tilde{\boldsymbol{\mu}}^{(1)} + \tilde{\boldsymbol{\Sigma}}_{12}\tilde{\boldsymbol{\Sigma}}_{22}^{-1}\left(\mathbf{x}_j^{(2)} - \tilde{\boldsymbol{\mu}}^{(2)}\right), \tag{1}$$

where it estimates the contribution of $\mathbf{x}_j^{(1)}$ to $\mathbf{T}_1$. Next, the predicted contribution of $\mathbf{x}_j^{(1)}$ to $\mathbf{T}_2$ is

$$\mathbf{x}_j^{(1)}\mathbf{x}_j^{(1)\prime} = E\left(\mathbf{X}_j^{(1)}\mathbf{X}_j^{(1)\prime} \middle| \mathbf{x}_j^{(2)}; \tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}\right) = \tilde{\boldsymbol{\Sigma}}_{11} - \tilde{\boldsymbol{\Sigma}}_{12}\tilde{\boldsymbol{\Sigma}}_{22}^{-1}\tilde{\boldsymbol{\Sigma}}_{21} + \tilde{\mathbf{x}}_j^{(1)}\tilde{\mathbf{x}}_j^{(1)\prime},$$

and

$$\mathbf{x}_j^{(1)}\mathbf{x}_j^{(2)\prime} = E\left(\mathbf{X}_j^{(1)}\mathbf{X}_j^{(2)\prime} \middle| \mathbf{x}_j^{(2)}; \tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}\right) = \tilde{\mathbf{x}}_j^{(1)}\tilde{\mathbf{x}}_j^{(2)\prime}. \tag{2}$$

The contribution in (1) and (2) are summed over all $\mathbf{x}_j$ with missing components. The results are combined with the sample data to yield $\tilde{\mathbf{T}}_1$ and $\tilde{\mathbf{T}}_2$ where this procedure completes one prediction step.

For the second step that is the estimation step, the predicted values are used to compute a revised maximum likelihood estimate of the parameters using the formula:

$$\tilde{\mu} = \frac{\tilde{\mathbf{T}}_1}{n}, \qquad \tilde{\Sigma} = \frac{1}{n}\tilde{\mathbf{T}}_2 - \tilde{\mu}\tilde{\mu}'. \tag{3}$$

This process is repeated until the revised estimates, $\tilde{\boldsymbol{\mu}}$ and $\tilde{\boldsymbol{\Sigma}}$ dot not differ much from the previous iteration.

Singular value decomposition (SVD) has a powerful property where it can compress the information contained in $\mathbf{X}$ into the first few singular vectors which are mutually orthogonal and their importance rapidly decreases after the first columns/rows. In general, *SVD* decomposes a $n \times p$ matrix $\mathbf{X}$ into three matrices; $\mathbf{X}_{n\times p} = \mathbf{U}_{n\times p}\,\Lambda_{n\times p}\,\mathbf{V}'_{p\times p}$ where $\mathbf{U}_{n\times n}$ and $\mathbf{V}'_{p\times p}$ are orthogonal and normalised matrices, i.e., $\mathbf{U}'\mathbf{U} = \mathbf{I}$ and $\mathbf{V}'\mathbf{V} = \mathbf{I}$, $\Lambda_{n\times n}$ is a diagonal matrix with singular values in decreasing order, $\mathbf{U}$ columns are the left singular vectors and $\mathbf{V}'$ rows are the right singular vectors. The eigenvalues and eigenvectors of $\mathbf{XX}'$ and $\mathbf{X}'\mathbf{X}$ are computed, respectively. The eigenvectors $\mathbf{XX}'$ will be used to form the $\mathbf{U}$ columns while the eigenvectors of $\mathbf{X}'\mathbf{X}$ will form the $\mathbf{V}$ columns.

As *SVD* can only be performed on complete matrices, so for all the missing values in matrix $\mathbf{X}$ are filled with the row average to obtain a complete matrix. From the completed matrix, we identify k most significant eigenvectors by sorting the eigenvectors based on their corresponding eigenvalue. Once k most significant eigenvectors from $\mathbf{V}^T$ are selected, a missing value j in row i are estimate by first regressing this row against the k eigenvectors and then use the coefficients of the regression to reconstruct j from a linear combination of the k eigenvectors. The jth value of row i and the jth values of the k eigenvectors are not used in determining these regression coefficients. An expectation maximization method are used to get the final estimate. By using the above algorithm, each missing value in $\mathbf{X}$ is estimated to obtain a new matrix and then this procedure is repeated until the total change in the matrix are below the empirically determined threshold of 0.01.

K-nearest neighbour (KNN) method Missing value is filled by the mean value of corresponding column of the nearest neighbour of corresponding row that have no missing values. The nearest neighbour can be defined in terms of Euclidean distance.

Sequential K-nearest neighbour (SKNN) method This method separates the dataset into incomplete and complete set with missing or without missing values, respectively. The data in incomplete set are imputed by the order of missing rate that is the missing values are ranked from the fewest number of missing. Starting with the fewest number of missing value, this missing value is filled by the weighted mean value of corresponding column of the nearest neighbour of corresponding row in complete set. Then by taking into account the first imputed value, the process is repeated until all missing values are imputed.

PERFORMANCE INDICATORS

In order to evaluate the imputation techniques, three performance indicators are used namely the correlation coefficient (R), the index of agreement (d) and the mean absolute error (*MAE*) (Junninen et al. 2004). These three methods are built based on taking consideration between the predicted and their corresponding observed values to select the best method for estimating missing data.

The correlation coefficient (R) This indicator explains the variability in the imputed data and how much they are related to the observed values. It takes on values between 0 and 1, with values closer to 1 implying a better fit. The equation can be expressed as:

$$R = \left[\frac{1}{N} \frac{\sum_{i=1}^{N} \left(P_i - \bar{P}\right)\left(O_i - \bar{O}\right)}{\sigma_P \sigma_O} \right], \tag{4}$$

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

Index of agreement (d) It is a measure of relative error between imputed and observed data point and the formula is given by,

$$d = 1 - \left[\frac{\sum_{i=1}^{N} \left(P_i - O_i\right)^2}{\sum_{i=1}^{N} \left(\left|P_i - \bar{O}\right| + \left|O_i - \bar{O}\right|\right)^2} \right]. \tag{5}$$

Mean absolute error (MAE) The mean absolute error is the average difference between imputed and observed data points and is given by:

$$MAE = \frac{1}{N} \sum_{i=1}^{N} \left|P_i - O_i\right|. \tag{6}$$

RESULTS AND DISCUSSION

The analysis in this study has been carried out using the free software of R. Figure 1 shows the performance indicator based on the correlation coefficient for the six imputation methods with different percentage of missing values such as 5, 10, 15, 20, 25 and 30% missing. The result for all the eight stations can be divided into 2 groups with three methods; *EM*, *KNN* and *SKNN* performs consistently superior compared to the other three methods; mean, median and *SVD*. The *EM*, *KNN* and *SKNN* have a higher correlation coefficient compared to mean, median and *SVD*. Among the best three methods, *EM* has the highest correlation compared to *KNN* and *SKNN* as can be seen in the graph for St2, St3, St6, St7 and St8 for almost all percentage of missing values. While for St1, St4 and St5, *EM* performs equally good with the correlation coefficient are about the same for *KNN* and *SKNN*. Among mean, median and *SVD*, the worst method is median with the lowest correlation coefficient as can be clearly seen in almost all the stations regardless any percentage of missing values.

Figure 2 shows the performance indicator based on the index of agreement for the six imputation methods with different percentage of missing values such as 5, 10, 15, 20, 25 and 30% missing. The behaviour of the six imputation techniques is similar to the one shown in Figure 1. The best performances with the highest index agreement are *EM*, *KNN* and *SKNN* compared to mean, median and *SVD* with lower index of agreement. According to index of agreement for 5% missing values, all methods are almost equally good for St2, St3, St7 and St8.

Figure 3 shows the performance indicator based on the mean absolute error for the six imputation methods with different percentage of missing values such as 5, 10, 15, 20, 25 and 30% missing. The behaviour of the six imputation techniques is also similar to the one shown in Figures 1 and 2. The three methods that performs consistently better with the lowest *MAE* are *EM*, *KNN* and *SKNN* and three equally bad method are mean, median and *SVD*. For St2, St3, St4, St6 and St8, the best method among the three are *EM*.

A comparison of each plot in Figures 1, 2 and 3 show that all the performance indicator agrees that *EM*, *KNN* and *SKNN* are among the best method of imputation for all the percentage of missing values. The *EM*, *KNN* and *SKNN* method are consistently superior to the other three methods irrespective of percentage of missing values.

CONCLUSION

In this paper we aimed to investigate and compare the best method to overcome missing values for air quality data sets in Malaysia. Six methods of imputation are used: mean and median substitution, expectation-maximization (*EM*) method, singular value decomposition (*SVD*), K-nearest neighbour (*KNN*) method and sequential K-nearest neighbour (*SKNN*) method. The performance of imputations is compared using three performance indicators namely the correlation coefficient (R), the index of agreement (d) the mean absolute error (*MAE*). The consistent of each method was tested using different set of data from eight monitoring station located at central, south and north region of Peninsular Malaysia with 5, 10, 15, 20, 25 and 30% missing. From the result obtained, 3 method: *EM*, *KNN* and *SKNN* were consistently superior irrespective of the station and percentage of missing. All the three performance indicator agrees that this three methods are among the best method with a higher correlation coefficient and index

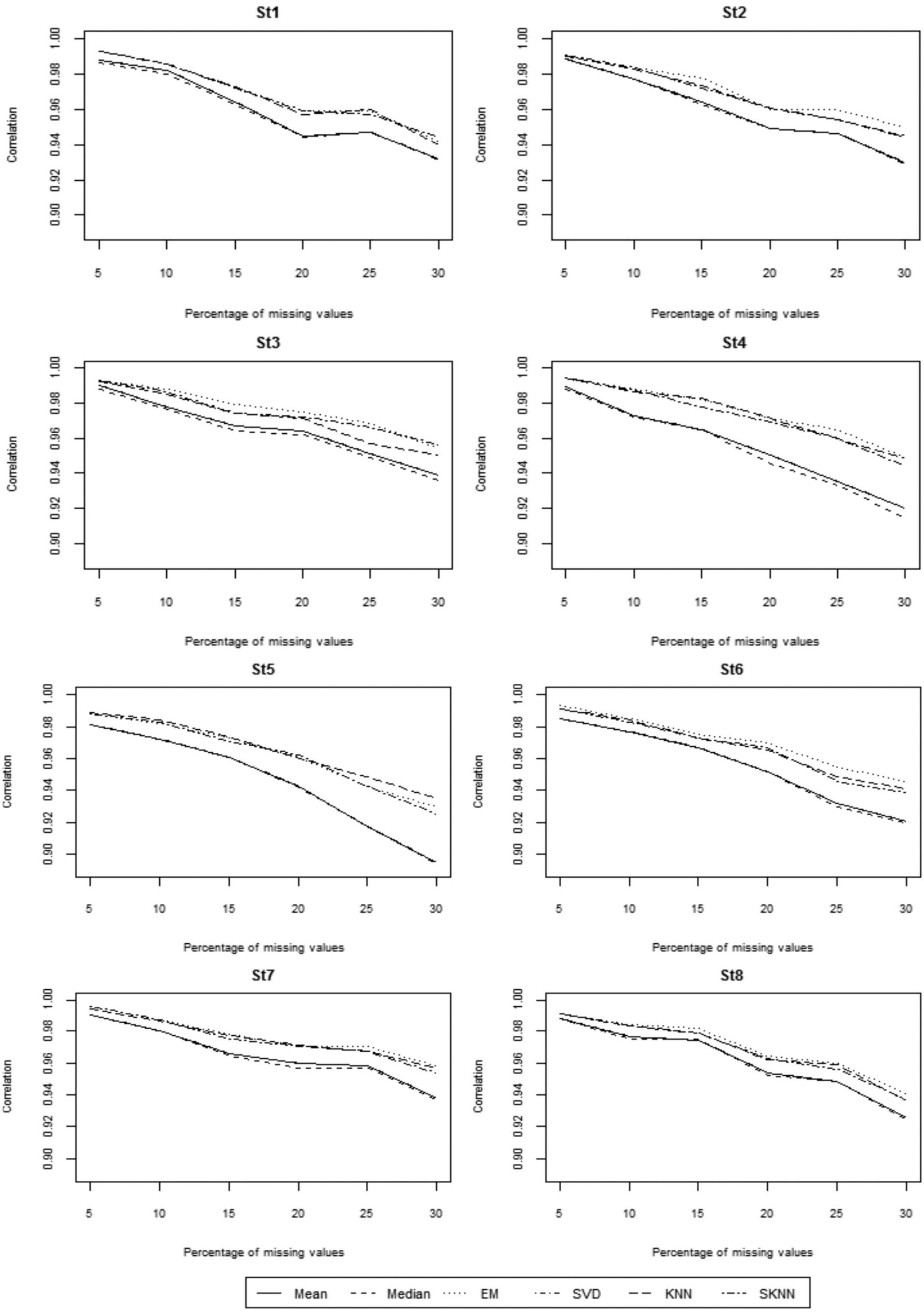

FIGURE 1. Performance indicator based on correlation coefficient for the six imputation methods

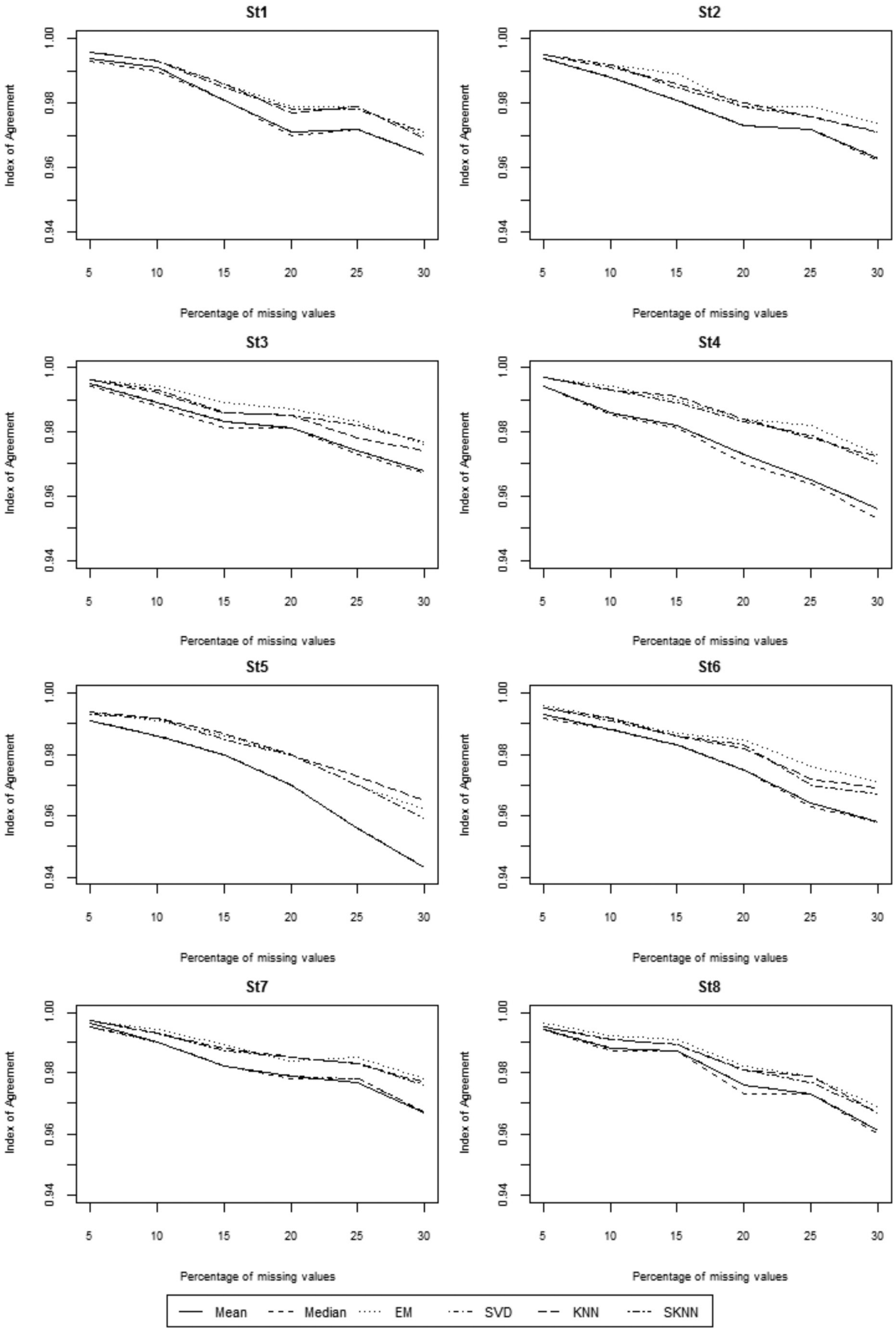

FIGURE 2. Performance indicator based on index of agreement for the six imputation methods

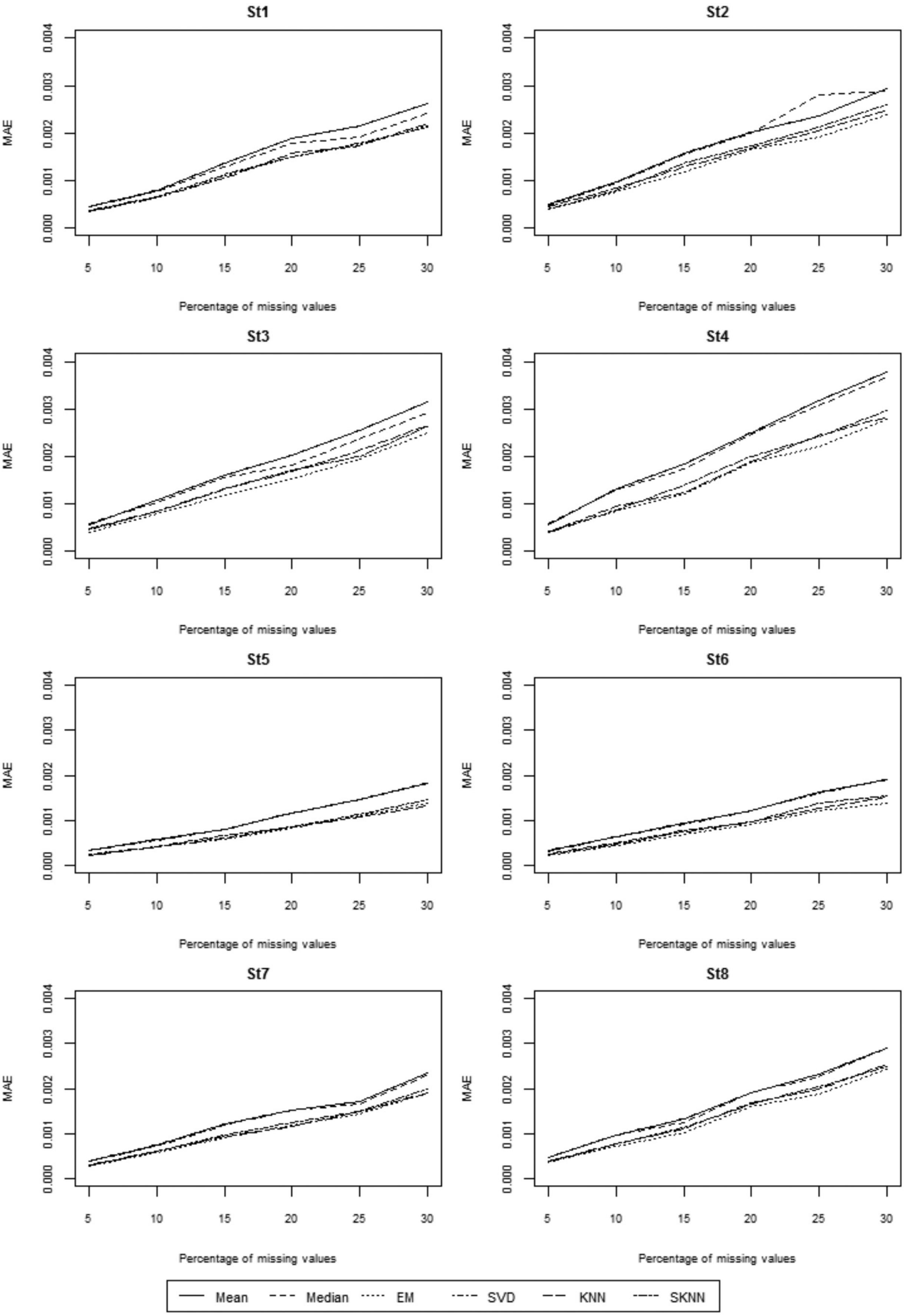

FIGURE 3. Performance indicator based on mean absolute error for the six imputation methods

agreement and with a lower min absolute error compared to mean, median and *SVD*. As a conclusion, we recommend that *EM*, *KNN* and *SKNN* are more preferable compared with the other three method.

ACKNOWLEDGEMENTS

The authors would like to thank Department of Environment Malaysia and Assoc. Prof. Dr. Mohd Talib Latif from Universiti Kebangsaan Malaysia for providing the information and data. This research is funded by Universiti Kebangsaan Malaysia research grant (ref no: UKM-AP-2011-19).

REFERENCES

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

Dempster, A.P., Laird, N.M. & Rubin, D.B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* 39(1): 1-38.

Gelman, A., King, G. & Liu, C. 1998. Not asked and not answered: Multiple imputation for multiple surveys. *Journal of the American Statistical Association* 93(443): 846-857.

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

Kim, K.Y., Kim, B.J. & Yi, G.S. 2004. Reuse of imputed data in microarray increases imputation efficiency. *BMC Bioinformatics* 5: 160.

Laaksonen, S. 2000. Regression-based nearest neighbor hot decking. *Computational Statistics*15(1): 65-71.

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

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

Pollice, A. & Lasinio, G.J. 2009. Two approaches to imputation and adjustment of air quality data from a composite monitoring network. *Journal of Data Science* 7: 43-59.

Porter, J., Cossman, R. & James, W. 2009. Research note: Imputing large group averages for missing data, using rural-urban continuum codes for density driven industry sectors. *Journal of Population Research* 26(3): 273-278.

Rubin, D.B. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., & Altman, R.B. 2001. Missing value estimation methods for DNA microarrays. *Bioinformatics* 17(6): 520-525.

Nuryazmin Ahmat Zainuri*
Fundamental Studies of Engineering Unit
Faculty of Engineering and Built Environment
Universiti Kebangsaan Malaysia
43600 Bangi, Selangor Darul Ehsan
Malaysia

Abdul Aziz Jemain & Nora Muda
School of Mathematical Sciences
Faculty of Science and Technology
Universiti Kebangsaan Malaysia
43600 Bangi, Selangor Darul Ehsan
Malaysia

*Corresponding author; email: yazmin@eng.ukm.my

Received: 30 May 2013
Accepted: 21 August 2014