Bulletin of Electrical Engineering and Informatics
Vol. 9, No. 2, April 2020, pp. 635~643
ISSN: 2302-9285, DOI: 10.11591/eei.v9i2.2090

A comparative study of different imputation methods for daily rainfall data in east-coast Peninsular Malaysia

Siti Mariana Che Mat Nor[1], Shazlyn Milleana Shaharudin[2], Shuhaida Ismail[3], Nurul Hila Zainuddin[4], Mou Leong Tan[5]
[1,2,4]Department of Mathematics, Faculty of Science and Mathematics, Universiti Pendidikan Sultan Idris, Malaysia
[3]Department of Mathematics and Statistics, Faculty of Applied Sciences and Technology, Universiti Tun Hussein Onn Malaysia, Malaysia
[5]Geography Section, School of Humanities, Universiti Sains Malaysia, Malaysia

Article Info

Article history:

Received Oct 14, 2019
Revised Dec 28, 2019
Accepted Feb 12, 2020

Keywords:

MCMC
Missing value
Nearest neighbor
NIPALS
Random forest
Replace by mean

ABSTRACT

Rainfall data are the most significant values in hydrology and climatology modelling. However, the datasets are prone to missing values due to various issues. This study aspires to impute the rainfall missing values by using various imputation method such as Replace by Mean, Nearest Neighbor, Random Forest, Non-linear Interactive Partial Least-Square (NIPALS) and Markov Chain Monte Carlo (MCMC). Daily rainfall datasets from 48 rainfall stations across east-coast Peninsular Malaysia were used in this study. The dataset were then fed into Multiple Linear Regression (MLR) model. The performance of abovementioned methods were evaluated using Root Mean Square Method (RMSE), Mean Absolute Error (MAE) and Nash-Sutcliffe Efficiency Coefficient (CE). The experimental results showed that RF coupled with MLR (RF-MLR) approach was attained as more fitting for satisfying the missing data in east-coast Peninsular Malaysia.

Corresponding Author:

Shazlyn Milleana Shaharudin,
Department of Mathematics,
Faculty of Science and Mathematics,
Universiti Pendidikan Sultan Idris, Malaysia.
Email: shazlyn@fsmt.upsi.edu.my

1. INTRODUCTION

In climatology and hydrological modeling, daily rainfall data is among the significant variables. Water resources management requires comprehensive hydrological variables datasets, including volumes, temperature and water level. Nevertheless, many hydrologists had commonly encountered the challenge of missing data in hydrological datasets. Normally, the missing data occurred for various reasons such as relocation of rainfall station, environmental changes, malfunctioning instruments and reorganization of network [1]. In hydrology, there are three type of missing data were taken into account, which are Missing Completely At Random (MCAR), Missing At Random (MAR) and Missing Not At Random (MNAR). As for hydrological data especially in the case of missing in rainfall datasets, it is classified as MCAR since the data in that area or any area does not affect the occurrence of missing in rainfall datasets of an area [1, 2]. It had been reported by [3] that the imputation for univariate time series hydrological data were also classified as MCAR and MAR. MCAR concerns the data where the chance of a particular missing values are independent of any dataset variables [2]. The most convenient practice to handle the missing data is by deleting the entire observations containing the missing data and analyzing the retained complete data.

However, ignoring or moving the data could be unsuitable as it could possibly make discontinuous data, leading to information loss. Consequently, there is possible outcome of unfitting conclusions. The consistency and continuity of rainfall data are highly crucial in statistical analyses like time series analysis [4]. The continuity and consistency might be instable as well because of the observational procedure modification and inadequate records. [2] Therefore, filling the data sets in daily rainfall data is critical.

In hydrologic modeling, using the most efficient method to acquire precise valuation of rainfall is highly important. The most fitting valuation should retain the key feature of the datasets and obey the rainfall characters in specific location [1, 5]. Hence, accomplishing the finest results in data analyses before proceeding to modeling, the data must be complete with good quality.

Generally, a few methods are used to manage the missing data. Normal Ratio method has become a generic method for estimating missing rainfall data [6]. It was initially recommended by [7], and afterwards altered by [8] and became a generic method for rainfall missing data valuation [9]. If any adjacent gauges had normal annual precipitation surpassing 10% of the measured gauge, this method was employed [10]. It was based on previous observations of rain gauge and the surroundings. Nonetheless, other significant factors include distances among rain gauges and aerial coverage of respective gauge which were considered when using this method and had become evident to significantly influence the rainfall valuation. Unlike the common methods listed, NIPALS and RF algorithm has the capability of preserving high dimensional data and constructing the efficient complete datasets of rainfall data. Furthermore, there is desirable properties of these algorithms which is its capability of handling diversified types of missing data. They could potentially scale to big data settings and adapt to nonlinearity and interactions [11].

On the contrary, there were alternative techniques which used other factors to evaluate the missing rainfall data. A number of researchers had presented means to manage the challenges in the missing data in their studies, specifically in the field of hydrology. One of the methods was inverse distance methods, which utilized the distances from the target station of two to five neighbor stations, giving more weight to the data from the nearest weather station [12]. Besides that, regression method was also a well-known method used for estimating missing values in hydrological data [13]. A different part in regression model was that it also considered other factors in order to impute the missing hydrological data such as elevation and topography [14]. Regression method employed step-wise regression to determine the coefficients for all the significant neighbor stations [15]. Regression-based method underestimated the amount of no rainfall days. However, regression methods suffered from the overestimation of rainy days total. Additionally, the probability distribution of rainfall was not well-preserved [16]. The regression method might misrepresent the amount freedom degrees and was challenging in noisy datasets [17].

The objective of this study are twofold, first to performed data imputation using Replace by Mean, Nearest Neighbor, Markov Chain Monte Carlo (MCMC), Nonlinear Iterative Partial Least Squares (NIPALS) and Random Forest (RF) methods for daily rainfall data in the East-Coast of Peninsular Malaysia. Second, to evaluate the performance of imputation methods coupled with Multiple Linear Regression (MLR) model in predicting the future daily rainfall values. The findings from this study is expected to contribute towards finding the best and finest method for data imputation technique that enables the reconstructions of complete rainfall datasets.

2. STUDY AREA AND DATA

This study centers on east-coast of Peninsular Malaysia that places in the latitude between 3.5° N and 6.5° N and longitudes 102° E and 104° E. The datasets used in this study is high dimensional data which were obtained from daily rainfall data of 1987-2018 from the Department of Irrigation and Drainage Malaysia (DID) for 32 years as represented in Figure 1. That 48181 data contained 8.59% missing values. According to [1], the datasets that contained of less than 10% of missing values are regarded as excellent data. A huge number of time series observations were needed to get a precise outline of the rainfall patterns [18]. Other than that, the reliability of frequency estimator of a long time series data is highly valuable since it strongly associates with sample size in data analysis. Table 1 shows the geographical coordinates of 48 rainfall stations chosen from east-coast of Peninsular Malaysia.

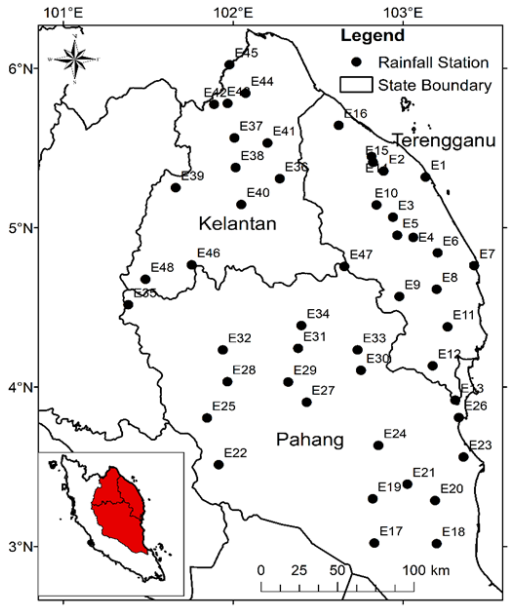

Figure 1. The location of 48 rainfall stations in east-coast peninsular Malaysia

Table 1. Geographical coordinates and percentage of missing values for east-coast of peninsular Malaysia

Stations	Code	Stations	Code
Setor JPS KT	E1	Stor JPS Raub	E25
Kg. Sg. Tong	E2	Pejabat JPS Pahang	E26
Kg. Dura	E3	Rumah Pam Paya Kangsar	E27
Kg. Menerong	E4	JKR Benta	E28
Kg. Embong Sekayu	E5	Kg. Sg. Yap	E29
Jambatan Jerangau	E6	Kawasan B Ulu Tekai	E30
SM Sultan Omar	E7	Kg. Merting	E31
Al-Muktafi	E8	Bkt. Betong	E32
Rumah Pam Paya Kempian	E9	Ulu Tekai (A)	E33
Sg. Gawi	E10	Kuala Tahan	E34
Jambatan Tebak	E11	Gunong Brinchang	E35
Kg. Ban Ho	E12	Kg. Laloh	E36
Hulu Jabor	E13	Ulu Sekor	E37
Kg. Batu Hampar	E14	Dabong	E38
Klinik Chalok Barat	E15	Gob	E39
Inst. Pertanian Besut	E16	Balai Polis Bertam	E40
Sg. Kepasing	E17	Sek. Men. Teknik Kuala Krai	E41
Temeris	E18	Air Lanas	E42
Sg. Cabang Kanan	E19	Kg. Durian Daun	E43
Kg. Unchang	E20	Bendang Nyior	E44
Kg. Batu Gong	E21	Rumah Kastam	E45
Kuala Marong	E22	Blau	E46
Rumah Pam Pahang Tua	E23	Gunung Gagau	E47
Pintu Kawalan Pulau Kertam	E24	Brook	E48

3. RESULTS METHOD

3.1. Replace by mean

The easiest imputation technique used when the missing data are less than 10% is mean substitution method [19]. It comprises changing every missing value in the series $X^{(k)}$, k=1,…,d by the corresponding mean of respective component. This method had been employed in multivariate hydrological frequency analysis studies [20]. It was concluded that replacing missing data with overall average of the whole observations of the data provided excellent results [15]. The formula can be written as:

$$P_x = \frac{\sum_{i=1}^{n} P_i}{n} \quad (1)$$

where P_x is the observed rainfall data while n is the number of rainfall days.

3.2. Nearest neighbor

Another effective methods to fulfil the missing data are nearest neighbor imputation algorithms. Every missing value on a number of records is substituted by a value acquired from interrelated cases in the overall records set [21]. This method is based on the k observed values of the most similar time series. Then every value is employed into a single value with approaches like kernel function or the average methods [22]. Nearest neighbor imputation approaches are donor-based methods. The imputed value is regarded a value that was essentially valued for other dataset record or the measured values average from k records [23]. The process of imputation by nearest neighbor can be briefly explained as – Let *n* observations on *p* covariates be gathered. The corresponding n × p data matrix is given by X=(x_{is}), where x_{is} denotes the *i*th observation of the sth variable. Let O=(o_{is}) denote the corresponding $n \times p$ matrix of dummies with entries;

$$o_{is} = \begin{cases} 1 & \text{if } x_{is} \text{ was observed} \\ 0 & \text{for missing value} \end{cases} \quad (2)$$

distances between two observations x_i and x_j, which are signified by rows in the data matrix, can be calculated by using the L_q-metric for the data observed. Then one uses the distances;

$$d_q(x_i, x_j) = \left[\frac{1}{m_{ij}} \sum_{s=1}^{p} |x_{is} - x_{js}|^q I(o_{is} = 1) I(o_{js} = 1)\right]^{1/q} \quad (3)$$

where $m_{ij} = \sum_{s=1}^{p} I\,(o_{is} = 1) I(o_{js} = 1)$ denotes the number of valid components in the computation of distances. Parallel view conceptualize the distances and hence nearest neighbors were used [24].

3.3. Markov chain monte carlo (MCMC)

Other than any other methods for imputation of missing data, some still cannot be calculated explicitly due to missing data or complex dependence. So, the researchers can also use MCMC. The imputation has been performed by Monte Carlo simulation of MCMC method. According to [2], the expectation-maximization (EM) is a technique which figures the maximum probable valuations for MCMC method to replace missing data. MCMC method was used for the multi imputation procedure because of assuming multivariate normality. MCMC method is based on Bayesian inference with missing data by obeying several steps [25]:

– Imputation step. Estimate mean and covariance matrix, then simulates the missing values for each observation.
– Posterior step (P-step). P-step simulates the mean vector and covariance matrix from the imputed step.

3.4. Non-linear interactive partial least-square (NIPALS)

Principal component analysis with missing values could be allowed by using NIPALS method is a method. It iteratively applied PCA to the datasets with missing values. The NIPALS algorithm is employed on the dataset and the attained PCA model is employed to expect the missing values [26]. The algorithms of NIPALS work as follows:

Given a rectangular data table of size $n \times p$, let us denote by $X = \{xij\}, 1 \le i \le n, 1 \le j \le p$, the matrix representing the observed values of the variables $x.j$ for n statistical units. Next, if X is of rank a, then the decomposition formula for principal component analysis of X is $X = \sum_{h=1}^{a} t_h p_h'$, where $t_h = (t_{h1}, \dots, t_{hi}, \dots, t_{hn})$' and $p_h = (p_{h1}, \dots, p_{hj}, \dots, p_{hp})$ are the principal factors and principal components, correspondingly. Hence, the NIPALS algorithm valuates a missing value conforming to the cell (i, j) as:

$$\hat{x}_{ji} = \sum_{l=1}^{k} t_{li} p_{lj} \quad (4)$$

where k ($k \le a$) is established by cross-validation. Execution of the NIPALS algorithm is straight forward, the base is simple linear regressions.

3.5. Random forest

Random Forest can manage mixed data type. It has been identified to be able to work efficiently under barren circumstances like non-linear data structures, complex interactions and high dimensions. According to [27], random forest is capable of dealing with mixed data type and as a non-parametric method, non-linear (regression) and interactive effects are probable. Let us assume $\mathbf{X}$=($\boldsymbol{X_1, X_2, \dots, X_P}$)

to be a $n \times p$-dimensional data matrix. For an arbitrary variable $\boldsymbol{X_s}$ including missing values at entries $\boldsymbol{i_{mis}^{(s)}} \subseteq \{1,\ldots,n\}$ the rainfall dataset in this study could be separated into two categories:

- The observed values of variable $\boldsymbol{X_s}$, denoted by $\boldsymbol{y_{obs}^{(s)}}$
- the missing values of variable $\boldsymbol{X_s}$, denoted by $\boldsymbol{y_{mis}^{(s)}}$;

For start, an initial guess for the missing values in $\mathbf{X}$ is made using mean or other imputation method. Next, the variables $\boldsymbol{X_s}$,$s=1,\ldots,p$ is sorted based on the overall missing values beginning with the smallest amount. For each variable $\boldsymbol{X_s}$, the missing values is imputed by first fitting an Random Forest with response $\boldsymbol{y_{obs}^{(s)}}$ and predictors $\boldsymbol{x_{obs}^{(s)}}$; and next, predicting the missing values $\boldsymbol{y_{mis}^{(s)}}$ by applying the trained Random Forest to $\boldsymbol{x_{mis}^{(s)}}$. The imputation procedure is repeated until a stopping criterion is satisfied.

3.6. Multiple linear regression

After all missing values are replaced by several approaches, the complete data set is analyzed using multiple linear regression to identify the best approaches of handling the missing data in east-coast Peninsular Malaysia. Regression analysis is a statistical method which uses the connection between at least two quantitative variables towards expected variables [28]. A common statistical method employed in a lot of disciplines including climate data is MLR model [29]. The Multiple linear regression model parameter can be stated as follows:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots \beta_k X_{ik} + \varepsilon_i(\beta), i = 1, \ldots, N \quad (5)$$

where $Y_i$is the value of response variable, $\beta_0, \beta_1, \beta_2$ and β_k are unknown constant, X_y is value of predictor variable ε_i is the random error.

3.7. Root mean square error (RMSE), mean absolute error (MAE) and nash-sutcliffe efficiency coefficient (CE)

The root mean square error (RMSE) has been used as a standard statistical metric to measure model performance in meteorology, air quality, and climate research studies. RMSE presents information on the short-term efficiency which is a benchmark of the difference of predicated values about the observed values. The lower the RMSE, the more accurate is the evaluation. Another statistical approach for model performance evaluation is Nash-Sutcliffe efficiency coefficient (CE). The CE is a popular index to assess the predictive power of hydrological models [30]. CE value of 1 are pursued in the best performance models. The mean absolute error (MAE) is another useful measure widely used in model evaluations. MAE (mean absolute error) is an indication of the average deviation of the predicted values from the corresponding observed values and can present information on long term performance of the models; the lower value of MAE represents the better results for the long term model [1]. The RMSE, CE and MAE are given by the following formula;

$$\text{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2/n} \quad (6)$$

$$CE = 1 - \sum_{i=1}^{n}(y_i - \tilde{y}_i)^2 \,/\, \sum_{i=1}^{n}(y_i - \bar{y}_i)^2 \quad (7)$$

$$\text{MAE}=\frac{1}{n}\sum_{i=1}^{n}|(y_i - \tilde{y}_i)| \quad (8)$$

where y_i is the observed rainfall, $\tilde{y}_i$ is the predicted rainfall data, and $\bar{y}_i$ indicates the average rainfall data over rainfall station in east-coast Peninsular Malaysia.

4. RESULTS AND DISCUSSION

This section discusses the results of imputation methods for daily rainfall datasets. The imputation methods were applied for 48 stations in east-coast of Peninsular Malaysia. The experiments were conducted for each station, using all five imputation methods. The results were then calculated as an average results, representing each imputation method. Root Mean Square Error (RMSE) and Nash-Sutcliffe Efficiency Coefficient (CE) were used in evaluating the performance of each method. If the discrepancy between the estimated and observed values for each station were small, RMSE will display the smallest values. Meanwhile, CE values may varied from -∞ to 1 and deemed satisfactory when the values are higher than 0.5. The method with smallest RMSE and highest CE values were selected as the best technique in filling

the missing data of daily rainfall datasets. The experimental results for each imputation methods are showed in Table 2.

Table 2 indicates the average RMSE and CE values for five methods. The results showed that, the smallest RMSE with highest CE was obtained from Replace by Mean method. However, CE values showed that all imputation methods generate satisfactory results whereby the obtained values were approximate to 1. In view of the obtained results, Replace by Mean provided the most fitting performance. Meanwhile Random Forest (RF) was the worst imputation method for daily rainfall data in east-coast Peninsular Malaysia as the results showed RF has the lowest CE and highest RMSE amongst other methods.

Table 2. Average RMSE and CE values for five imputation methods

Method	RMSE	CE
Replace by Mean	**2.3100***	**0.9873***
Nearest Neighbor	4.2362	0.9597
MCMC	5.0208	0.9461
NIPALS	3.8386	0.9703
Random Forest	6.3337	0.9150

*** indicate the best results**

Once the missing values have been filling in, the next step in this study is to analyze the full dataset using Multiple Linear Regression (MLR) model. The MLR model was used to identify the best approaches of handling missing data when the imputation values coupled with modelling. To evaluate the performance of imputation methods coupled with MLR model, MAE and RMSE were used respectively.

Table 3 presents the RMSE and MAE values for each statistical approaches for imputing the missing values of daily rainfall data in east-coast Peninsular Malaysia coupled with multiple linear regression model. It can be observed that RF-MLR has the lowest RMSE and MAE of 16.4428 and 8.6229, respectively compared to other approaches. Thus, the final results suggested that Random Forest is the best statistical approach for imputing the missing values of daily rainfall data when it coupled with regression model. Table 3 also showed that imputation method of Replace by Mean-MLR has relatively smallest values for RMSE and MAE, making the model competitive to RF-MLR.

Table 3. The results for MLR coupled with imputation methods

Model	RMSE	MAE
Replace by Mean-MLR	16.9564	8.9195
NN-MLR	17.4154	9.2199
MCMC-MLR	17.2803	9.3224
NIPALS-MLR	17.2179	9.0531
RF-MLR	**16.4428**	**8.6229**

Finally, the observed and predicted values for Replace by Mean-MLR, NN-MLR, MCMC-MLR, NIPALS-MLR and RF-MLR models were plot for visual inspection. Figure 2 shows the results of five imputation methods for 175 missing daily rainfall data in the east-coast Peninsular Malaysia at station Setor JPS KT (E1). Based from Figure 2, it can observed that the imputed values of daily rainfall data by using RF-MLR and NN-MLR showed similar trends. For instance, both models were responsive to rainfall occurrences with similar magnitude peaks and times. However, RSME and MAE for RF-MLR was significantly lower compared to NN-MLR. In contrast, Replace by Mean-MLR generally underestimated the discharges since the plotted line of the values of estimated tend to flatten out which did not seem to show any trends. With these results, RF-MLR was considered the best model at filling the gaps in the missing values of daily rainfall data in east-coast Peninsular Malaysia.

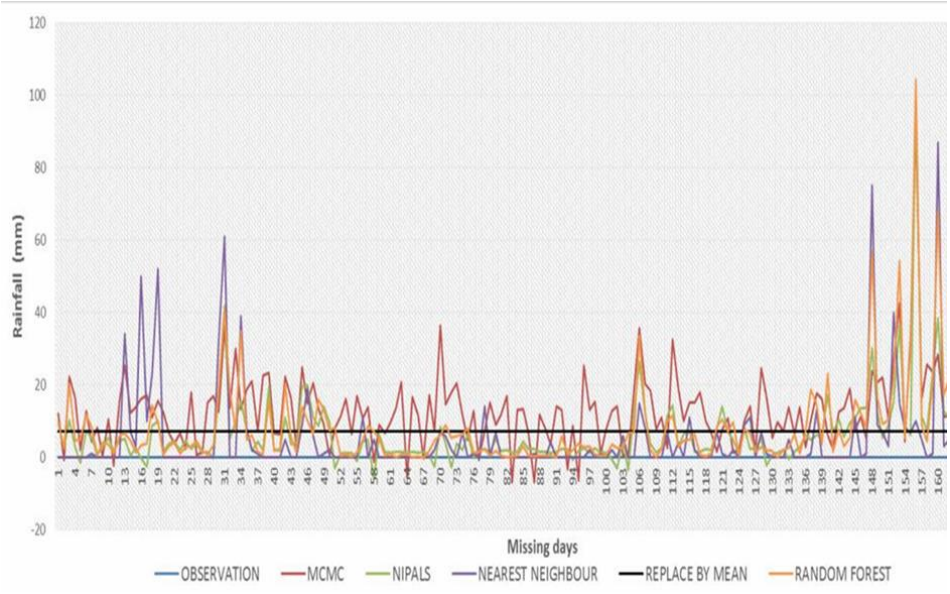

Figure 2. Data imputation results of 175 missing rainfall data for the MCMC, NIPALS, nearest neighbour and replace by mean models. Five line graphs represent the observation (blue), MCMC (red), NIPALS (green), NEAREST NEIGHBOR (PURPLE) AND REPLACE BY MEan (black)

5. CONCLUSION

The search of the most efficient method for imputation the missing values for rainfall data has continuously received a huge attention in many studies. In this study, five imputation methods which are Replace by Mean, NN, MCMC, NIPALS and RF were used and compared to obtain the most appropriate technique in filling the missing data for daily rainfall data in East-Coast Peninsular Malaysia. The results showed that Replace by Mean is the best method for single imputation. However, RF has proven it superiority as the method having the best result when coupled with MLR. The study has found out that the by using Replace by Mean, the dataset are prone to the risk of changing the standard deviation and the skewness of the data might change as well. Furthermore, it is also confirmed that, performance of predictive modelling coupled with imputation method may differ from single imputation alone. Therefore, in finding the best method for data imputation, it is crucial to test the dataset after imputation with any predictive modelling. To conclude, the use of various imputation techniques based on the characteristics of rainfall is endorsed and further studies with different methodologies and datasets should be explored.

ACKNOWLEDGEMENTS

This research has been carried out under Fundamental Research Grants Scheme (FRGS/1/2019/STG06/UPSI/02/4) provided by Ministry of Education of Malaysia

REFERENCES

[1] I. F. Kamaruzaman, *et al*., "A Comparison of Method for Treating Missing Daily Rainfall Data in Peninsular Malaysia," *Malaysian Journal of Fundamental and Applied Sciences*, pp. 375-380, 2017.
[2] R. J. Little and D. B. Rubin, "Statistical Analysis with Missing Data," *Wiley-Interscience*, ISBN 978-0471183860, 2002.
[3] S. Moritz, *et al*., "Comparison of Different Methods for Univariate Time Series Imputation in R," *Cologne University of Applied Sciences*, pp. 1-20, October 015.
[4] S. M. Shaharudin et al., "Modified Singular Spectrum Analysis in Identifynig Rainfall Trend over Peninsular Malaysia," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 15, no. 1, pp. 283-293, July 2019.
[5] A. A. Jemain *et al.,* "Penyurihan Ikhtisar Data Hujan," *Dewan Bahasa dan Pustaka*, p. 213, 2015.
[6] V. Chow et al., "Applied Hydrology", *Tata Mc Graw Hill Book Company*, ISBN 0-07-0101810-2, 1988.
[7] Paulhus, J. L. H., and M. A. Kohler, "Interpolation of missing precipitation records," *Monthly Weather Review*, vol. 80, no. 8, pp. 129-133, August 1952.
[8] N. F. A. Radi *et al.,* "Estimation of Missing Rainfall Data using Spatial Interpolation and Imputation Methods," *AIP Conference Proceedings*, vol. 1642, no. 1, pp. 1-11, February 2015.

[9] M.-T. Sattari *et al.,* "Assessment of Different Methods for Estimation of Missing Data in Precipitation Studies," *Hydrology Research*, vol. 48, no. 4, pp. 1032-1044, 2017.
[10] R. P. De Silva, *et al.,* "A Comparison of Methods used in Estimating Missing Rainfall Data," *The Journal of Agricultural Sciences*, vol. 3, pp. 101-108, January 2007.
[11] M. Kim et al., "Comparative Studies of Different imputation Methods for Recovering Streamflow Observation," in *Water*, vol. 7, no. 12, pp. 6847-6860, December 2015.
[12] Di Piazza et al., "Comparative Analysis of Different Techniques for Spatial Interpolation of Rainfall Data to Create a Serially Complete Monthly Time Series of Precipiation for Sicily, Italy," *International Journal of Applied Earth Observation and Geoinformation*, vol. 13, no. 3, pp. 396-408, June 2011.
[13] M. M. Hasan and B. F. W. Croke, "Filling Gaps in Daily Rainfall Data: A Statistical Approach," *20th International Congress on Modelling and Simulation*, Adelaide, Australia, pp. 380-386, December 2013.
[14] T. Makhuvha *et al.,* "Patching Rainfall Data using Regression Methods," *Journal of Hydrology*, vol. 198, no. 1-4, pp. 308-318, November 1997.
[15] Xia *et al.,* "Forest Climatology Estimation of Missing Values for Bavaria, Germany," A*gricultural and Forest Meteorology*, vol. 96, no. 1-3, pp. 131-144, August 1999.
[16] M. B. Aissia, *et al.,* "Multivariate Missing Data in Hydrology," *Advances in Water Resources*, vol. 110, pp. 299-309, December 2017.
[17] F. Tang and H. Ishwaran, "Random Forest Missing Data Algorithms," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 10, no. 6, pp. 363- 377, June 2017
[18] S. M. Shaharudin, *et al.,* "An Efficient Method to Improve the Clustering Performance using Hybrid Robust Principal Component Analysis-Spectral Biclustering in Rainfall Patterns Identification," *International Journal of Artificial Intelligence (IJ-AI)*," vol. 8, no. 3, pp. 237-243, September 2019.
[19] M. B. Aissia *et al.,* "Multivariate Missing Data in Hydrology," *Advances in Water Resources*, vol. 110, pp. 299-309, December 2017.
[20] L. R. Beard and A. J. Fredrich, "Hyrologic frequency analysis," *Hydrologic Engineering Methods for Water Resources Developmen*t, April 1975.
[21] L. Beretta and A. Santaniello, "Nearest Neighbor Imputation algorithms: a critical evaluation," *BMC Medical Informatics and Decision Making*, vol. 16, pp. 197-208, November 2015.
[22] M. Amiri and R. Jensen, "Missing Data Imputation using Fuzzy-rough Methods," *Neurocomputing*, vol. 205, pp. 152-164, September 2016.
[23] R. Pan *et al.,* "Missing Data Imputation by K Nearest Neighbours based on Grey Relational Structure and Mutual Information," *Applied Intelligence*, pp. 1-22, 2015.
[24] M. Aci *et al.,* "A Hybrid Classification Method of *k* Nearest Neighbor, Bayesian Methods and Genetic Algorithm," *Expert Systems with Applications*, vol. 37, no. 7, pp. 5061-5067, July 2010.
[25] N. Suhaimi *et al.,* "Markov Chain Monte Carlo Method for Handling Missing Data in Air Quality Datasets," *Malaysian Journal of Analytical Sciences*, vol. 21, no. 3, pp. 552-559, June 2017.
[26] S. M. Shaharudin, *et al.,* "Identification of Rainfall Patterns on Hydrological Simulation using Robust Principal Component Analysis," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*," vol. 11, no. 3, pp. 1162-1167, September 2018.
[27] L. Breiman, "Random Forest", *Machine Learning*, vol. 45, pp. 5–32, 2001.
[28] M. A. Shafi *et al.,* "A Hybrid of Multiple Linear Regression Clustering Model with Support Vector Machine for Colorectal Cancer Tumor Size Prediction," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 10, no. 4, pp. 323-328, January 2019.
[29] M. A. I. Navid and N. H. Niloy, "Multiple Linear Regression for Predicting Rainfall for Bangladesh," *Communications*, vol. 6, no. 1, pp. 1-4, 2018.
[30] T. Chai and R. R. Draxler, "Root Mean Square Error (RMSE) or Mean Absolute Error (MAE)," *Geoscientific Model Development*, vol. 7, no. 1, pp. 1247-1250, January 2014.

BIOGRAPHIES OF AUTHORS

Siti Mariana is a graduate of Bachelor Degree in Education (Mathematics) from Universiti Pendidikan Sultan Idris (UPSI) in 2018. She is currently pursuing her studies in Masters of Science degree in Statistics while working on to publish papers in her scope of field. Her research focuses on the area of dimension reduction methods, specifically in climate informatics which involves analysis on huge climate-related datasets based on techniques in Data Mining.

Shazlyn Milleana was born in Johor Bahru, Malaysia, in 1988. She is a senior lecturer at the Department of Mathematics, Faculty of Science and Mathematics, Universiti Pendidikan Sultan Idris (UPSI). She graduated with a bachelor science degree in Industrial Mathematics from Universiti Teknologi Malaysia, in 2010. Upon graduation, she began her career as an Executive in banking institution. In the following year, she received an offer to continue her study as a fast-track PhD student at the same university. During her PhD journey, she developed an interest in multivariate analysis, specifically in finding patterns which deals with big data. Her research focuses on the area of dimension reduction methods specifically in climate informatics which involves analysis on huge climate-related datasets based on techniques in Data Mining. She had published her research in Scopus indexed journal and presented her work in various local and international conferences. She completed her PhD thesis at the end of 2016 and was conferred a doctorate degree in 2017.

Shuhaida Ismail is a lecturer at the Department of Mathematics and Staticstics, Faculty of Applied Sciences and Technology, Universiti Tun Hussein Onn Malaysia (UTHM). She obtained her first degree in Computer Sciences majoring from UTM. She also obtained a Master degree and PhD from the same university. Throughout she studies, she developed an interest in Machine Learning research area, specifically in predictive modelling, classification, and clustering. Her current research areas are in hydrological modelling, big data analytics and deep learning.

Nurul Hila was born in Kelantan, Malaysia, in 1987. She is a senior lecturer at the Department of Mathematics, Faculty of Science and Mathematics, Universiti Pendidikan Sultan Idris (UPSI). She graduated with a bachelor science d`egree in Financial Mathematics from Universiti Malaysia Terengganu, in 2010. In the following year, she continued her studies at the Universiti Malaysia Terengganu for her masters degree and in 2016 she managed to complete her Ph.D at the same university. During her PhD journey, she developed an interest in hybrid modelling, specifically double bootstrap on control chart. Her research focuses on the volatility point of sukuk (Islamic certificate) investment. She had published her research in Scopus indexed journal and presented her work in various local and international conferences.

Muo Leong Tan is a senior lecturer of Geography Section, School of Humanities at Universiti Sains Malaysia. Dr, Tan received his Ph.D in Remote Sensing from Universiti Teknologi Malaysia in 2016. He was awarded a postdoctoral fellowship at National University of Singapore and a visiting scholar at Fudan University, China. Dr Tan has demonstrated advanced research skill and foresight in the specific disciplines of remote sensing, geographic information system (GIS), climate change and hyrological modeling. He studied the impacts and uncertainties of climate change on streamflow in tropical river basins. Dr. Tan authored or co-authored over 35 scientific articles and proceedings, with an h-index of 11.