ISSN 2278-3091
Volume 8, No.1.6, 2019
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse6981.62019.pdf
https://doi.org/10.30534/ijatcse/2019/6981.62019

Comparative Analysis of Missing Data Imputation Methods for Continuous Variables in Water Consumption Data

Norzanah Md Said[1], Zalhan Mohd Zin[2], Mohd Nazri Ismail[3], Termizi Abu Bakar[4]
[1,2] UniKL MFI, Malaysia, norzanah@unikl.edu.my
[3] UPNM, Malaysia, m.nazri@upnm.edu.my
[4] RAI Utility, Malaysia, termiziabubakar@gmail.com

ABSTRACT

Real-world data are often incomplete and/or lacking in certain behaviors or trends and usually contains many missing data. Missing data problem can be solved using missing values imputation technique in order to produce a good quality of data.

This paper presents a comparative analysis of three missing data imputation approaches based on the flow processes of handling missing data to address the issues of missing water consumption data in Sibu City. The aim is to determine the suitable imputation method for substituting missing values in the dataset. In this study, the research emphasizes the use of basic principles of dealing with missing data in choosing the use of missing data imputation methods. Thus, three existing imputation methods of advanced techniques for handling missing data are deployed; model-based K-Nearest Neighbor imputation (KNN), SVD Imputation (SI) and Random Forest (RF) respectively. The experiment was conducted using datasets with different missing entries (5% to 45% by 5). The missing data entries were created based on a complete water consumption dataset. The experimental comparative analysis was evaluated with respect of its standard deviation of residuals (Root Mean Square Error, RMSE), coefficient of determination (r-square, R^2) and execution time. The analysis shows that 0.26(5%, low missing rates) and 0.36(45%, high missing rates) of RMSE criteria and 0.93 of R^2 by RF method was demonstrated comparatively better to other imputation methods. In the term of execution time criteria, 0.29 seconds for low missing rates (5%) and 0.98 seconds for high missing rates (45%) by SI method performs the fastest among other two methods.

Key words: missing data; missing value imputation; root mean square error; r-square

1. INTRODUCTION

Nowadays, automatic meter reading (AMR) is used for monitoring, measurement and analysis of water consumption data [1]. This AMR is installed at customers' premises and monitored at various locations through various monitoring stations for utility control [2]. However, to conduct water consumption analysis which has large observations of missing data makes the task difficult to identify and extract potential useful information from datasets[3-4]. The occurrence of missing data has been assumed occurred during the process of integration [5], equipment failure, human error, routine maintenance, changes in sitting of monitors, measurement error, mishandling samples or due to some other factors [6]. This missing data or incomplete data set creates will introduced an element of ambiguity into data analysis that produce to low quality of data [7]. Besides that, most of researchers' claim that missing data are problematic which leads to problems in data handling, computation analysis and the results produce may be biased which has contributed to the significant reduction in efficiency [8].

In regards, missing data imputation forms a significant data preprocessing issue in order to provide an efficient and valid analysis. Potti et al., 2015 [9] have highlighted the issues of information quality examination which concerns a right handling and investigation of data towards producing data quality. With that, there are various imputation methods that be able to handle missing data form simple approaches to complex analysis. Many researchers are working on this problem to develop a new imputation method [10] or introduce more sophisticated methods. These researchers have successfully deployed to address missing values in various fields. Wai Yan Lai et. al [11] has utilized data mining algorithm using Sequential K-Nearest Neighbor (SKNN) imputation methods for treating missing rainfall data. The imputation method used by the author is based on the good performance of SKNN method proposed by [12]. Shahid Ali and Simon Dacey [13] has successfully applied machine learning algorithm (SVM ensemble) to replace missing values for Carbon monoxide(CO) concentrations for air pollution. The use of SVM ensemble due to computational air pollution data analysis is spatio-temporal in nature. Cronin-Fenton et.al [14] had used traditional data analysis such as complete-case analysis and multiple imputation methods in clinical epidemiological research. The study had shown good performance results on the small percentage of missing data (<5%) in the study field. Cheema and Jehanzeb R [15] have proposed some guidelines for choosing missing data handling methods based on factors such as sample size,

proportion of missing data, method of analysis and missing data handling method. Nevertheless, this study was discussed on statistical methods and analysis perspective. Based on above discussion, many methods are available, but analysts are facing difficulty in searching a suitable method due to lack of knowledge in term of flow processes of handling missing data in choosing the suitable imputation methods.

Thus, this paper presents a comparative analysis of three missing data imputation approaches based on the flow processes of handling missing data to address the issues of missing water consumption data in Sibu City. This paper identifies the suitable imputation method that can be applied for recovering the missing values for continuous variables in the dataset. In this study, the research emphasizes the use of basic principles of dealing with missing data in choosing the use of missing data imputation methods. The three of existing imputation methods of advanced techniques for handling missing data are deployed based on the analysis of proportion of missing data; model-based K-Nearest Neighbor imputation (KNN), SVD Imputation (SI) and Random Forest (RF) respectively. For each method, the performance evaluation criteria are based on Root Mean Squared Error (RMSE), R-square (R^2) and execution time represent a dependent (response) variable and various percentage of missing values represents independent (predictors) variables (5%, small missing rates to 45%, high missing rates; by 5).

2. STATE OF THE ART OF MISSING DATA HANDLING METHODS

In general, handling for missing data can be categorised into two phases; design and analytics [14]. The design phase can be described as a plan to minimize the extent of missing data (MD) especially during data collection towards improvement of data completeness. In this phase, the requirement of data quality monitoring is prioritized by organization in order to support a particular business need. For the analytics phase, it can be described as the processes that should be taken for dealing with missing data.

A recent review about missing data handling is provided in [5][10][16]. A wide variety of approaches has been reviewed and their application differs according to different datasets. Furthermore, some practical consideration are useful for initial stage for determining to the right imputation technique. There are two basic principles components to dealing with missing data that have been determined. From the observation, many of researchers are implementing the same handling technique as well [17][18].

2.1 Basic Principles of Dealing with Missing Data

There are two basic steps when dealing with missing data. It consists of the type of missingness mechanism and missing data classification technique as shown in Figure 1.

Figure 1: Basic principles for Dealing With Missing Data (derived from [10] [17])

2.1.1Type of Missing Data Mechanism

There are three possible types of missing data mechanism as described in [19][20][21]. Basic principle to identify missing data mechanism is depending whether a relationship is exist between the missing data with the other data in the dataset. Through identifying the category of missingness mechanism, it might assists the pattern of missingness and the reasons of missing data occurs in the dataset [22].

Suppose (Y, M) is a data matrix with complete data; M being the missing data indicator matrix is expressed as :

$$M = \begin{cases} 1, & if\ Ymis \\ 0, & Otherwise,\ Yobs \end{cases}$$

Y^{obs} and Y^{mis} are respective observed and missing parts of Y.

An observation is assume to be:

- Data that is missing completely at random (MCAR), if the missingness independent of all observed and unobserved values (missing data). The reasons of missing values are unrelated to the data in the data set.
  p(M | Y) = p(M) for all Y; M is independent of both Y^{obs} and Y^{mis}
- Data that is missing at random (MAR); if the missingness is independent of unobserved values but dependent on observed values
  p(M | Y) = p(M | Y^{obs}) for all Y; M is independent of Y^{mis}
- Data that is missing not at random (MNAR); if the missingness depends on missing (as well as perhaps on observed) values of Y
  p(M | Y) depends on (Y^{mis})

The assumption of MCAR mechanism can be tested using null-hypothesis tests. It can be denoted as:

The null hypothesis is: The missing data is missing completely at random (MCAR)

Alternative hypothesis is: The missing data is not MCAR (either MAR or MNAR)

Formally, this is written as:

H_0: p-value > 0.05

H_1: p-value <= 0.05

The p-value is a statistical measure to compare a chosen significance level (alpha) such as 0.05 to determine if the null hypothesis can be rejected. Bear in mind that, there is still no proper method to assume that missingness mechanism either MAR or MNAR can be determined from the observed data directly [23].

2.1.2 Missing Data Classification Techniques

There are two group of missing data classification techniques; traditional and modern data analysis. The classification techniques are determined based on the analysis of missing data proportions. With regards, traditional data analysis techniques is an appropriate to apply when a small percentage of data (less than 5%) is missing. Due to that, the deletion and single value data imputation methods are the most selection to handle missing data. There have been a large number of traditional data analysis reviewed in [24][15].

For modern data analysis technique is the right selection when deals with more than 5% of missing data in a dataset. Sometimes, it is called as advanced technique. Multiple imputation, model-based procedures and machine-based learning have been determined as the right selection to impute missing values in advanced techniques. A variety of advance techniques have been reviewed in [25]–[29][30].

Bear in mind, the MCAR and MAR assumptions are eligible to handle missing data either in traditional or advanced technique. In summary, single data imputation gives biased coefficients if the data is not MCAR and ignores uncertainty and almost underestimates the variance [31]. In contrast to advanced technique which constitute better estimate of uncertainty and unbiased as well. Hence, the next sub-topic will discuss the advance technique in term of quantitative measurement for continuous variables which is in conjunction with research domain.

2.2 Missing Data Handling Methods for Continous Variables

The presence of missing data in quantitative measurement which usually measures in interval scales in the context of water consumption is investigated. In this research, the proportion of missing data is assumed to be more than 5% due to missing data can occurs for multiple reasons (i.e from data integration, error during data entry etc) and under MAR assumption. In fact, the nature of dataset characteristics in quantitative measurement can be applied for classification and regression tasks. In the context of water consumption, data is based on meter reading and estimated consumption which is identified to be in continuous variables. Due to that, it gives a direction to apply regression task to predict the missing data based on observed data. With that, existing imputation methods such as model-based K-Nearest Neighbor (KNN), Singular Value Decomposition (SI) and Random Forest (RF) are proposed as imputation methods for substituting missing data in the dataset.

2.2.1 K-Nearest Neighbor (KNN) Imputation

KNN imputation is referred as instance-based algorithm [32]. In the KNN, missing values are imputed using values calculated from the k-nearest neighbours. In particular, the nearest neighbours can be identified by minimizing the distance function, such as the Euclidean distance. Once the k-nearest neighbours have been found, a replacement value must be estimated to substitute for the missing attribute value. The KNN can predict both qualitative attributes (the most frequent value among the k-nearest neighbours) for nominal values and quantitative attributes (the mean among the k-nearest neighbours) in the case of numerical values. An important parameter for the kNN method is the value of k, which is typically set to 1 but is sensitive to outliers. However, KNNI has no theoretical criteria for selecting the best k-value and the k-value has to be determined empirically. The KNN technique is referred to be the most efficient when applied on microarray data situations [33].

2.2.2 Singular Value Decomposition (SI) Imputation

Singular Value Decomposition (SI) is used to obtain a set of mutually orthogonal expression patterns that can be linearly combined to approximate the values of all attributes in the data set [24] [34]. SVD is based on eigenvalues can be expressed as

$$A_{m\times m} = U_{m\times m}\Sigma_{m\times n}\, V^T_{n\times n}.$$

The most significant eigenvectors of V^T is used to linearly estimate missing values. Then, the exact fraction of eigenvectors best for estimation needs to be determined empirically. Once k most significant eigenvectors from V^T are selected, missing value j in row i by first regressing this row against the k eigenvectors need to be estimated. After that, the coefficients of the regression to reconstruct j from a linear combination of the k eigenvectors need to be used. The j th value of row i and the j th values of the k eigenvectors are not used in determining these regression coefficients. In SI, it has to originally fill all missing values by other methods in matrix A, obtaining A'. Then utilize an expectation maximization method to arrive at the final estimate. Each missing value in A is estimated using the above algorithm, and then the procedure is repeated on the newly obtained matrix, until the total change in the matrix falls below the empirically determined threshold (say 0.01).

2.2.3 Random Forest (RF) Imputation

Random Forest (RF) is referred to an iterative imputation method based on a random forest algorithm [35]. It's a non parametric imputation method applicable to various variable types. Non-parametric method does not make explicit assumptions about functional form of *f* (any arbitary function). Instead, it tries to estimate *f* such that it can be as close to the data points without seeming impractical. By averaging over many unpruned classification or regression trees, it builds a random forest model for each variable. Then it uses the model to predict missing values in the variable with the help of observed values [36].

2.3 Performance Error Metric

2.3.1 Root Mean Square Error (RMSE)

Error metric is important for the selection of reliable imputation method. The most widely adopted performance measure for regression performance is root mean square error (RMSE) [37][38]. RMSE measures the difference between imputed and true values. It is represented the sample standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are. RMSE can be expressed as.

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(X_i^{obs} - X_i^{imputed})^2}{n}}$$

where, X_i^{obs} is observed (true) values and $X_i^{imputed}$ represents predicted (imputed) values

A standard statistical metric, RMSE has been used to measure model performance in meteorology, air quality, and climate research studies [37].

2.3.2 Coefficient of Determination (R Square, R^2)

The coefficient of determination is the ratio of explained variation in response (dependent) to the total variation in response. It is a statistical measure of how close the data are to the fitted regression line. The aim to measure for understanding how much of the variability in the key response can be explained.

2.3.3 Execution Time

Runtime of each imputation algorithms in term of CPU time (process time) is measured. The CPU time is measured in seconds.

3. METHODOLOGY

In this study, the analytics phase for handling missing data is proposed as shown in Figure 2. Then, follows by explanation for each process involve in handling missing data in the dataset. At the same time, the involvement of major task in data preprocessing are determined as described in [39][40] **.**

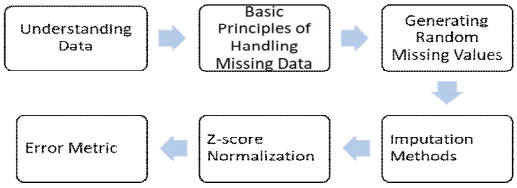


Figure 2: Proposed Flow Processes of Handling Missing Data

3.1 Understanding Water Consumption Data

The initial process when deals with missing data is to study the various attribute types, real values and discover pattern from water consumption data Dataset [41].

In this study, the residential water consumption data set was collected from one of the water utility in Sarawak. For finding information and discover patterns in a dataset, an exploratory data analysis (EDA) is applied. An EDA is an approach that perform an initial investigations on data which has its ability to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations [42]–[44]. The dataset contains 33046 instances (customers), 12 features (monthly usage) and 2 meta attributes (customer id and meter no). All features are holds numerical continuous.

3.1 Understanding Water Consumption Data

The initial process when deals with missing data is to study the various attribute types, real values and discover pattern from water consumption data [41].

In this study, the residential water consumption data set was collected from one of the water utility in Sarawak. For finding information and discover patterns in a dataset, an exploratory data analysis (EDA) is applied. An EDA is an approach that perform an initial investigations on data which has its ability to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations [42]–[44]. The dataset contains 33046 instances (customers), 12 features (monthly usage) and 2 meta attributes (customer id and meter no). All features are holds numerical continuous.

3.2 Basic Principles in Handling Missing Data

The next process is to identify the category of missingness mechanism and analysis the proportion of missing data as explained in [14][17] for water consumption data. In summary, the flow processes consists of two basic steps as shown in Figure 1.

3.3 Generating Random Missing Values

Generating random missing values is used to generate the percentage of missing values randomly. The missing rate is

usually taken from 5% until 45% from the complete data (Non-Nan Values) for simulation purposes [45]. The purpose of generating random missing values are to evaluate the performance of methods dealing with missing data.

3.4 Imputation Methods

Imputation can be described as the process of replacing missing data with substituted values using imputation methods (i.e advanced techniques). These imputatation methods are able to generate close to the true with unknown missing values. The idea is using complete data set and then artificially introducing missing values.Then, imputation methods are applied to the data and an estimation of how far is the estimation to the original value (known value). In the perspective of data preprocessing, it is the most important tasks f which invlove data cleaning in term of handle missing data in the dataset.

3.5 Z-Score Normalization

Z-score normalization is a strategy of normalizing data which allow to avoid outlier issue. Z-score normalization is applied to make every data point have the same scale so each feature is equally important.

3.3 Error Metric

Evaluation performance criteria is used to select a reliable imputation method. In this study, the RMSE is measured to quantify and compare the imputation methods over a set of datasets.

4. RESULTS AND DISCUSSION

4.1 Understanding Data

Figure 3 shows 91% (30072 instances) of the dataset is non-Nan values which is nearly to be completed data. Therefore, this study should investigate 9% related to the proportion of missing data (2974 instances). Studies need to be done to avoid losing information when dealing with missing data using simple techniques such as listwise deletion.

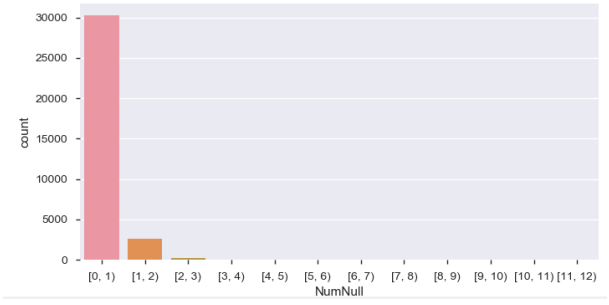


Figure 3: Plot represent Non-Nan Values

4.2 Basic Principles of Handling Missing Data

4.2.1 Type of Missingness Mechanism

Based on the null-hypothesis testing, the p-value is obtained closely to 0.00 which as indicator to reject the null hypothesis. The p-value is calculated using chi-square test in Python environment. Hence, the MAR assumption of missingness mechanism is assumed. Besides that, the percentage of 9% missing values assist in selecting an appropriate handling data technique.

4.2.2 Analyse the proportion of missing data

The presence of missing values in a dataset are recognized based on the symbols, NaN,NA,n/a, na,--,0. Table 1 shows the proportion of missing data for each month. To find meaningful data from the table, multivariate data analysis is applied. Multiple linear regression using scatter plot is selected for describing missing data and determining the relationship between the number of missing values and the month.

Table 1: Number of missing values per month

Month	Number of Missing Values
01	259
02	224
03	248
04	233
05	242
06	237
07	255
08	277
09	259
10	243
11	266
12	274

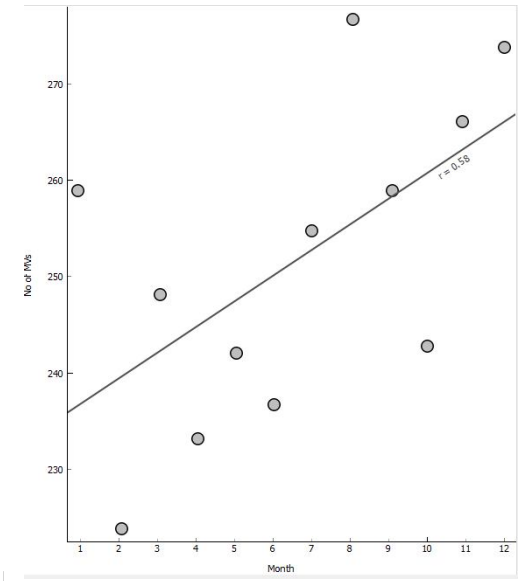
According to the Figure 4 the plot indicates linear positive with a moderate relationship. It is because the correlation coefficient denoted as r indicates 0.58 which explain in [46].

4.3 Experiment for comparison on Missing Values Imputation Methods

Table 2: Statistical Performance of three (3) different Imputations Techniques

Missing Rate	Singular Value Decomposition (SI)			K-Nearest Neighbor (KNN)			Random Forest (RF)		
	RMSE	E.T	R-Square	RMSE	E.T	R-Square	RMSE	E.T	R-Square
0.05	0.2137	0.299		0.2037	0.572		0.2572	4.156	
0.15	0.2389	0.409		0.3809	3.349		0.2896	6.873	
0.25	0.3021	0.646	0.8925	0.3831	9.082	0.4823	0.3524	12.188	0.9268
0.35	0.3231	0.837		0.4114	15.315		0.3592	12.120	
0.45	0.4651	0.984		0.3711	21.48		0.3813	15.431	

E.T = Execution Time



and month using scatter plot

4.3.1 Used Effect of the imputation methods on error metric

Results of experimental comparative analysis (ECA) was performed on three different imputation methods presented; SI, KNN and RF. The purpose of ECA is to evaluate the effect of using different imputation methods on missing values in the dataset in term of the standard deviation of the residuals (RMSE criteria), coefficient of determination (R^2) and time taken. Three comparative experiments were carried out based on the RMSE values according to the percentage of missing values (from 5% to 45%) to represent each method of imputation as shown in Table 2. Overall result indicates that the performance measures for three different imputation methods are in the range of 0.200 until 0.414 for their standard deviation of the residuals which was in acceptable error. The RF imputation method results was better performance which was indicated to 0.26 for small missing values,(5%) and 0.38 for higher missing values, (45%). In contrast to KKN, it was seen that their RMSE values is relatively high at missing rate between 5% (0.20) to 35% (0.41). Then it decreases slightly (0.37) when the missing rate is higher (45%). For SI method, the performance was expected to be incereased with increasing percentage of missing values in the dataset. The RMSE values was around 0.20 to 0.41.

According to the Figure 5, R-squared (R^2) measure for each imputation methods were obtained as well. The R-squared (R^2) measure is calculated to inform that the model can be explained all the variability of the response data (RMSE) around the mean. From above, 93%, 84% and 48% of R^2 were obtained representing for each model; RF, SI and KNN. It means that 93% of the variability in the standard deviation of residuals measure for RF model is accounted for by the percentage of missing data. So, the rate of missing data is the factor that influence the standard deviation of residuals for RF model since over 93% of the variabilities is left explained. Due to that, RF model was assumed to be better model fits to the data. While, the KNN method was assumed less good model fits to the data because the low R^2 was obtained (48%). Consequently, The model of RF which obtained higher R^2 is useful for modelling related to prediction issues for further analysis in future development.

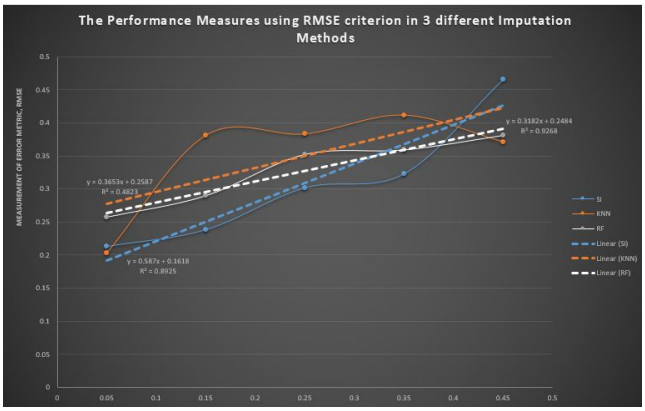


Figure 5: The effect use in 3 different imputation methods along the rate of missing values

4.3.2 Time taken

Execution time for each method is given in Figure 6. SI method was all very fast 0.2 sec to 1.0 sec duration following the missing value rate. For two methods of imputation; RF and KNN, were indicated that the time taken increased with the rate of missing values. But, RF imputation was look better than KNN which requires 15 sec for the highest rate of missing values (45%) whereas KNN imputation is the most time consuming.

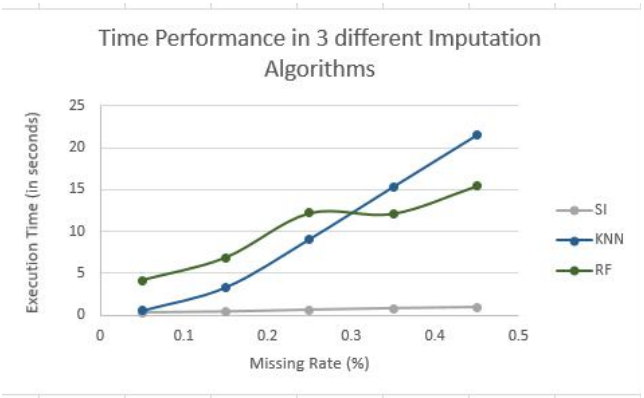


Figure 6: Time performance in 3 different imputation algorithms along the rate of missing values

5. CONCLUSION

This study described the concepts of missing value imputation that can be deployed at water utilities related to the water consumption data. The overall goal of the study was to determine the best imputation method for imputing missing values in the dataset according to the proposed flow processes of handling missing data. Error measures were calculated such as RMSE, R^2 and execution time in order to determine the performance error.

As a conclusion, this study have performed an experimental comparison among the imputation methods presented. Overall, the model produced results that were within the acceptable error. The results obtained in this study suggest that RF imputation method is likely the best imputation methods for filling missing data in water consumption in term of its standard deviation of residuals and coefficient of determination. In term of execution time, SI method is the fastest when comparing with other two methods presented.

ACKNOWLEDGEMENT

The authors are grateful to the RAI Utility Sdn Bhd for providing the necessary support to carry out the work presented in this manuscript.

REFERENCES

[1] A. Candelieri, D. Soldi, and F. Archetti, "Short-term forecasting of hourly water consumption by using automatic metering readers data," *Procedia Eng.*, vol. 119, no. 1, pp. 844–853, 2015. https://doi.org/10.1016/j.proeng.2015.08.948

[2] D. A. L. Owen, "The Technologies and Techniques Driving Smart Water From Innovation to Application – The Necessity of Integration," vol. 1, pp. 57–78, 2018.

[3] F. Km *et al.*, "Simplified Processing Method for Meter Data Analysis," no. November, 2015.

[4] A. David and O. Lloyd, "Smart Water Technologies and Techniques: Data Capture and Analysis for Sustainable Water Management, First Edition," pp. 229–230, 2018.

[5] D. Priya, R. Sivaraj, R. Assistant, and S. Gr, "a Review of Missing Data Handling Methods," *Int. J. Eng. Technol. Sci. – IJETS^{TM} ISSN*, vol. 2, no. 2, pp. 2349–3968, 2015.

[6] L. Hovany, "Error in Water Meter Measuring Due to Shorter Flow and Consumption Shorter Than the Time the Meter was Calibrated," *Water Supply Syst. Anal. - Sel. Top.*, 2012. https://doi.org/10.5772/51046

[7] S. A. N. Alexandropoulos, S. B. Kotsiantis, and M. N. Vrahatis, "Data preprocessing in predictive data mining," *Knowl. Eng. Rev.*, vol. 34, 2019.

[8] S. García, S. Ramírez-Gallego, J. Luengo, J. M. Benítez, and F. Herrera, "Big Data Analytics Big data preprocessing: methods and prospects," *Big Data Anal.*, vol. 1, no. 9, 2016.

[9] L. K. S. Potti and M. Madhavi, "An efficient and effective data quality management in health sector," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 8, no. 1, pp. 161–167, 2019.

[10] S. P. Mandel J, "A Comparison of Six Methods for Missing Data Imputation," *J. Biom. Biostat.*, vol. 06, no. 01, pp. 1–6, 2015. https://doi.org/10.4172/2155-6180.1000224

[11] W. Y. Lai, K. K. Kuok, S. Gato-trinidad, and K. X. Ling, "A Study on Sequential K-Nearest Neighbor (SKNN) Imputation for Treating Missing Rainfall Da," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 8, no. 3, pp. 363–368, 2019. https://doi.org/10.30534/ijatcse/2019/05832019

[12] and G.-S. Y. K.-Y. Kim, B.-J. Kim, "Reuse of imputed data in microarray analysis increases imputation efficiency," , journal article vol. 5, no. 1, p.," *BMC Bioinformatics*, vol. 5, no. 1, p. 160, 2004.

[13] S. Ali and S. Dacey, "Technical Review : Performance of Existing Imputation Methods for Missing Data in SVM Ensemble Creation," *Int. J. Data Min. Knowl. Manag. Process*, vol. 7, no. 5/6, pp. 75–91, 2017.

[14] D. Cronin-Fenton *et al.*, "Missing data and multiple imputation in clinical epidemiological research," *Clin. Epidemiol.*, vol. Volume 9, pp. 157–166, 2017.

[15] J. R. Cheema, "Some General Guidelines for Choosing Missing Data Handling Methods in Educational Research," *J. Mod. Appl. Stat. Methods*, vol. 13, no. 2, pp. 53–75, 2014.

[16] C. K. Enders, "Dealing With Missing Data in Developmental Research," *Child Dev. Perspect.*, 2013. https://doi.org/10.1111/cdep.12008

[17] M. S. Osman, A. M. Abu-Mahfouz, and P. R. Page, "A Survey on Data Imputation Techniques: Water Distribution System as a Use Case," *IEEE Access*, vol. 6, pp. 63279–63291, 2018.

[18] L. A. Belanche, V. Kobayashi, and T. Aluja, "Handling missing values in kernel methods with application to microbiology data," *Neurocomputing*, vol. 141, pp. 110–116, 2014.

[19] D. B. Rubin, "Cheracterizing the Estimation of Parameters in Incomplete-Data Problems.pdf," *J. Am. Stat. Assoc.*, no. 69, pp. 467–474, 1974.

[20] S. Tshering, T. Okazaki, and S. Endo, "A Method to Identify Missing Data Mechanism in Incomplete Dataset," *IJCSNS Int. J. Comput. Sci. Netw. Secur.*, vol. 13, no. 3, pp. 14–22, 2013.

[21] R. J. A. Little, "A test of missing completely at random for multivariate data with missing values," *J. Am. Stat. Assoc.*, vol. 83, no. 404, pp. 1198–1202, 1988.

[22] S. Fielding, P. M. Fayers, and C. R. Ramsay,

"Investigating the missing data mechanism in quality of life outcomes: A comparison of approaches," *Health Qual. Life Outcomes*, vol. 7, pp. 1–10, 2009.

[23] M. W. Huang, W. C. Lin, C. W. Chen, S. W. Ke, C. F. Tsai, and W. Eberle, "Data preprocessing issues for incomplete medical datasets," *Expert Syst.*, vol. 33, no. 5, pp. 432–438, 2016. https://doi.org/10.1111/exsy.12155

[24] S. García, J. Luengo, and F. Herrera, "Data Preprocessing in Data Mining," in *Data Preprocessing in Data Mining*, Intelligen., vol. 72, Springer International Publishing, 2015.

[25] A. N. Baraldi and C. K. Enders, "An introduction to modern missing data analyses," *J. Sch. Psychol.*, 2010.

[26] C. K. Enders, "A primer on the use of modern missing-data methods in psychosomatic medicine research," *Psychosom. Med.*, 2006.

[27] S. McPherson, C. Barbosa-Leiker, G. L. Burns, D. Howell, and J. Roll, "Missing data in substance abuse treatment research: Current methods and modern approaches," *Exp. Clin. Psychopharmacol.*, 2012.

[28] W. Leite and S. N. Beretvas, "The performance of multiple imputation for Likert-type items with missing data," *J. Mod. Appl. Stat. Methods*, 2010.

[29] M. Hopkins and M. Goeree, "Missing Data Methods," in *Health Technology Assessment*, 2015.

[30] M. B. Richman, T. B. Trafalis, and I. Adrianto, "Multiple imputation through machine learning algorithms," *87th AMS Annu. Meet.*, 2007.

[31] M. K. and J. Tian, "Machine Learning : Data Pre-processing," in *Prognostics and Health Management of Electronics: Fundamentals, Machine Learning, and the Internet of Things*, First., Michael G.Pecht and Myeongsu Kang, Ed. John Wiley & Sons Ltd, 2018, pp. 111–130. https://doi.org/10.1002/9781119515326.ch5

[32] S. García, J. Luengo, and F. Herrera, "Preface," in *Intelligent Systems Reference Library*, vol. 72, Springer International Publishing, 2015.

[33] D. V Nguyen, N. Wang, and R. J. Carroll, "Evaluation of Missing Value Estimation for Microarray Data," *J. Data Sci.*, vol. 2, pp. 347–370, 2004.

[34] O. Troyanskaya *et al.*, "Missing value estimation methods for DNA microarrays," *Bioinformatics*, vol. 17, no. 6, pp. 520–525, 2001.

[35] D. J. Stekhoven and P. Bühlmann, "Missforest-Non-parametric missing value imputation for mixed-type data," *Bioinformatics*, vol. 28, no. 1, pp. 112–118, 2012.

[36] D. J. Stekhoven and P. Bühlmann, "Missforest-Non-parametric missing value imputation for mixed-type data," *Bioinformatics*, 2012.

[37] T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)? -Arguments against avoiding RMSE in the literature," *Geosci. Model Dev.*, vol. 7, no. 3, pp. 1247–1250, 2014.

[38] R. C. Fair, "Evaluating predictive accuracy," *Specif. Estim. Anal. Macroecon. Model.*, vol. 1, no. 1, pp. 261–274, 1984.

[39] S. García, J. Luengo, and F. Herrera, "Tutorial on practical tips of the most influential data preprocessing algorithms in data mining," *Knowledge-Based Syst.*, vol. 98, pp. 1–29, 2016.

[40] I. Kononenko and M. Kukar, *Data Preprocessing*. 2013.

[41] P. M. Kellstedt and G. D. Whitten, *Getting to Know Your Data*. 2018.

[42] M. Abzalov, "Exploratory data analysis," in *Modern Approaches in Solid Earth Sciences*, 2016.

[43] W. L. Martinez and A. R. Martinez, *Exploratory data analysis with MATLAB®*. 2004. https://doi.org/10.1201/9780203483374

[44] J. T. Behrens, "Principles and Procedures of Exploratory Data Analysis," *Psychol. Methods*, 1997.

[45] R. M. Schouten, P. Lugtig, and G. Vink, "Generating missing values for simulation purposes: a multivariate amputation procedure," *J. Stat. Comput. Simul.*, vol. 88, no. 15, pp. 2909–2930, 2018.

[46] & F. Moore, D. S., Notz, W. I, *The basic practice of statistics*, 6th ed. New York, NY:, 2013.