Jurnal Kejuruteraan 35(1) 2023: 191-201
https://doi.org/10.17576/jkukm-2023-35(1)-18

Comparisons of Various Imputation Methods for Incomplete Water Quality Data: A Case Study of The Langat River, Malaysia

Naeimah Mamat & Siti Fatin Mohd Razali*

Department of Civil Engineering, Faculty of Engineering and Built Environment, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor, Malaysia.

**Corresponding author: fatinrazali@ukm.edu.my*

Received 7 May 2022, Received in revised form 27 June 2022
Accepted 30 July 2022, Available online 30 January 2023

ABSTRACT

In this study, the ability of numerous statistical and machine learning models to impute water quality data was investigated at three monitoring stations along the Langat River in Malaysia. Inconsistencies in the percentage of missing data between monitoring stations (varying from 20 percent (moderate) to over 50 percent (high)) represent the greatest obstacle of the study. The main objective was to select the best method for imputation and compare whether there are differences between the methods used by the different stations. The paper focuses on different imputation methods such as Multiple Predictive Mean Matching (PMM), Multiple Random Forest Imputation (RF), Multiple Bayesian Linear Regression Imputation (BLR), Multiple Linear Regression (non-Bayesian) Imputation (LRNB), Multiple Classification and Regression Tree (CART), k-nearest neighbours (kNN) and Bootstrap-based Expectation Maximisation (EMB). Remarkably, among all seven imputation techniques, the kNN produces identically reliable results. The imputed data is all rated as 'very good' (NSE > 0.75). This was confirmed by the calculation of |PBIAS|<5.30 (all imputed data are'very good') and KGE≥0.87 (all imputations are rated as' good'). Imputation performance improves for all three monitoring stations with an index of agreement, WI ≥ 0.94, despite varying percentages of missing data. According to the findings, the kNN imputation approach outperforms the others and should be prioritised in actual use. Future research with the existing methods could benefit from the addition of geographical data.

Keywords: Imputation methods; missing data; multiple imputation; evaluation criteria; water quality

INTRODUCTION

Water is essential to all life and can be used for a variety of purposes, including drinking, irrigation, industry, fishing, boating, and swimming (Çadraku 2021; Khan et al. 2021; Kerich 2020). The problem of missing data arises frequently in environmental fields for a variety of reasons. In developing nations, where the proportion of monitoring stations with missing data varies widely and is high, this water-quality data issue is of particular concern (Aguilera, Guardiola-Albert & Serrano-Hidalgo 2020).

Therefore, dealing with incomplete data is crucial, especially in modelling, as it can negatively affect the interpretation of the data, or the models created from the data (Ratolojanahary et al. 2019). The lack of data can pose significant challenges to researchers as it can lead to incorrect conclusions being drawn from a research project.

When it comes to research, lack of data is a common problem for many researchers. Lack of data in environmental and ecological studies can be caused by a variety of factors, such as insufficient samples, loss of samples, or malfunction of measurement instruments (Cheliotis et al. 2019; Hadeed et al. 2020).

Little and Rubin (2019) distinguished three types of missing data: Completely random missing data means that there is no relationship between the known values and the number of cases in which a variable is missing; the probability that a case contains a missing value for a variable may depend on the known values but not on the value of the missing data itself if the data are missing at random; if the probability of a missing value for a variable depends on the value of that variable, the data are not missing at random.

The simplest method for dealing with missing data is to remove all incomplete cases from the data collection. This method, called complete case analysis, can exclude critical information, especially for small samples. Another approach to missing values is to calculate them using the information contained in the data set. This is called the imputation method. It is crucial to choose the optimal imputation approach for missing data, as the consequences of an error are reflected in both the quality of the estimators and the results.

Another commonly used technique is to impute the missing values using averages. This is the simplest imputation approach as it uses the mean of each variable to estimate the missing value for the corresponding missing variables (Hamzah et al. 2020; Little & Rubin 2019). This approach can lead to biases and large errors in the covariance matrix, which affects the performance of statistical modelling. Other imputation algorithms include hot deck imputation based on the nearest neighbour method (Andridge, Bechtel & Thompson 2021), least squares imputation and maximum likelihood estimation (Shin, Davison, & Long 2017).

Numerous studies show that multiple imputation (MI) outperforms deletion and single imputation approaches for dealing with missing data in the context of incomplete data reconstruction (Aleryani, Wang, & de la Iglesia 2020; Audigier et al. 2018; de Silva et al. 2017; Hamzah et al. 2021; Hayati Rezvan, Lee, & Simpson 2015; Mandel J 2015; Morita 2021; Ratolojanahary et al. 2019). If the imputation model at least approximates the underlying mechanism of missing data, the MI technique has shown promising results (Murray 2018).

Similar methods have been used in previous studies to assess the superiority of one imputation method over another. Examples of such methods include the kernel-based iterative missing data estimation method by Liu et al. (2020), which was evaluated against other conventional frequency estimators as well as non-parametric iterative signals with a radius basis function kernel and other conventional frequency estimators. The methods were compared by simulating different levels of missing data, using each approach to predict missing values, and then comparing the predictions with the removed data using RMSE.

Tak, Woo & Yeo (2016) proposed an imputation method based on a modified k-nearest neighbour approach that takes into account spatial and temporal correlation. Missing observations were simulated by eliminating values between 0.1 and 50% of the total data, and then imputed using the proposed approach, the nearest history method, bootstrapping-based expectation maximisation and maximum likelihood estimation. The RMSE, MAPE and percentage change in variance were used to compare the imputation approaches.

In addition, Schmitt, Mandel & Guedj (2015) examined six different imputation techniques: mean, k-nearest neighbours (kNN), Fuzzy k-means, Singular Value Decomposition, Bayesian Principal Component Analysis and multiple imputations by chained equations. The comparison was carried out on four real data sets ranging in size from four to sixty-five variables under the completely random missing data assumption and using four evaluation criteria: RMSE, Unsupervised Classification Error, Supervised Classification Error and processing time.

Therefore, the main objective of this study is to determine the most appropriate imputation method and whether there are inconsistencies between the methods used by different stations. In this study, different imputation techniques are investigated, including Predictive Mean Matching (PMM), Multiple Random Forest Imputation (RF), Multiple Bayesian Linear Regression Imputation (BLR), Multiple Linear Regression (Non-Bayesian) Imputation (LRNB), Multiple Classification and Regression Tree (CART), k-nearest neighbours (kNN) and bootstrap-based expectation maximisation (EMB).

DATA AND STUDY AREA

The Langat River catchment is located in the western part of Peninsular Malaysia, more specifically between latitudes 2^o 40’ 152” N and 3^o 16’ 15” N and longitudes 101^o 19’ 20” E to 102^o 1’ 10” E (Hamzah et al. 2021). The catchment covers an area of about 2,394.38 km^2, with the main river channel being about 141 kilometres long. The river flows south into the Lower Mainland and west to the coast of Selangor State, with its mouth in the Strait of Malacca (Ebrahimian et al. 2018). This river basin, which is the most densely populated in Malaysia, is believed to offset the benefits of overdevelopment in the Klang Valley (Wan Mohtar, Bassa Nawang & Rahman 2017; Ahmed et al. 2016). It is an important raw water resource for drinking, recreational, industrial and agricultural purposes (Ahmed, Mokhtar, and Majid 2021)

Within the Langat River, there are four sub-basins (Kajang, Dengkil, Lui & Semenyih). The Langat River in Kajang was selected for water quality assessment. The Department of Environment (DOE) Malaysia, Ministry of Natural Resources and Environment, provided monthly water quality time series for the Langat River in Kajang. Three stations were selected for water quality monitoring: S01, S02 and S03. The details for the selected sampling water quality monitoring stations are depicted in Table 1 and Figure 1. At stations S01, S02, and S03, the average percentage of missing values is 25%, 22%, and 52%, respectively.

Department of Environment (DOE) had implemented WQI to measure the quality of water in Malaysia for over 25 years. DOE use six water parameters quality to define the status of surface water quality based on national water quality status (NWQS) for Malaysia, which are dissolved oxygen (DO), biological oxygen demand (BOD), chemical oxygen demand (COD), pH value, ammoniacal-nitrogen (AN) and total suspended solid (TSS).

For this study, the parameters for six water quality variables were selected for the period 2000-2019.

TABLE 1. Coordinates of selected sampling water quality monitoring station

Station	DOE station code	Coordinate		Location	Sampling data	Percentage missing values
		Latitude	Longitude			
S01	L15	03º02’46.0” N	101º46’38.8” E	Pekan Batu 11	Once in a month 2000 - 2019	25%
S02	L05	02º59’52.2” N	101º47’14.8” E	Kajang Bridge		22%
S03	L04	02º57’51.4” N	101º47’01.1” E	Near west country estate		52%

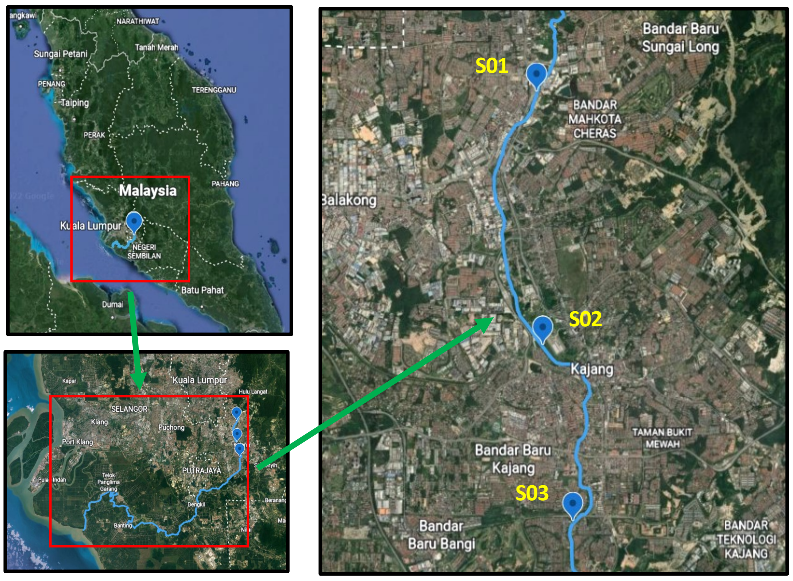


FIGURE 1. Selected water quality monitoring stations of the Langat River in Selangor, Malaysia.

METHODOLOGY

Two primary subsections constitute this section. In the first subsection, approaches for imputed missing data are presented. While the second subsection explains how the performance of the methods used is evaluated. This study used water quality data from 2000 to 2019 to evaluate the effectiveness of infilling techniques. The missing water quality data were restored after simulating the entire time series data. The technique for incorporating missing data into the complete time series is illustrated in the following phases.

1. Data were summarised to learn about the variables’ frequency distributions, the percentages of missing data, and the overall quality of the dataset.
2. Evaluations and computations of the chosen imputation models were performed.
3. The model that achieved the highest levels of performance across the range was determined to be the superior option for use with each variable at each monitoring location.

IMPUTATION METHODS

Numerous imputation techniques are described in the literature. Since there is not an optimal model for imputation for each type of variable (Rodríguez et al. 2021), several statistical and machine learning approaches (single and multiple imputation) were evaluated to achieve the aim of this study. This study found that there are seven imputation options that can be used to fill data gaps in water quality measurements.

MULTIPLE IMPUTATION (MI)

Multiple imputation (MI) combines the maximum likelihood technique with the ability to generate five to ten data sets, including raw data, that can be used to replace the missing data (Ser, Keskin & Yilmaz 2016). After merging the data from the imputed dataset, the parameters are estimated. MI generates a covariance matrix and a vector of means using maximum likelihood estimation.

MI goes one step further by incorporating statistical uncertainty into the model and using this uncertainty to simulate the natural variability found in an entire dataset of cases. MI then imputes actual data values to fill in the missing data points in the data matrix (Little & Rubin 2019). Next, the data analyst analyses the individual data collections, compiles the analysis results, and summarises them into a single set of summary results.

In multivariate analysis, MI seems to be one of the most appropriate methods to deal with missing data. Linear and non-linear models benefit greatly from the flexibility and adaptability of MI. The various multiple imputation methods examined in this study are as follows:

PREDICTIVE MEAN MATCH (PMM)

The closest possible match between predicted and observed values is used to replace missing values for a continuous variable in PMM.

BAYESIAN LINEAR REGRESSION (BLR)

Univariate missing data are inferred using Bayesian linear regression, a type of statistical inference.

LINEAR REGRESSION (NON-BAYESIAN) (LRNB)

A linear regression line from ‘*y*’ at ‘*x*’ is fitted to the observed data, ignoring model errors.

REGRESSION AND CLASSIFICATION TREES (CART)

CART is a classification and regression algorithm that uses binary decision trees to classify new data.

RANDOM FOREST (RF)

An ensemble approach that uses fully evolved regression trees. The goal is to generate a strong regressor from numerous weak learners (regression trees).

k-NEAREST NEIGHBOUR (kNN)

It is widely accepted that one of the top ten data mining techniques is the k-nearest neighbour (kNN) method, where the mean of the relevant column of the nearest neighbour of the corresponding row, that has no missing values is used to fill in the gaps when a value is missing. The distance between two points in Euclidean space can be used to define the nearest neighbour (Santos et al. 2020).

BOOTSTRAP-BASED EXPECTATION MAXIMIZATION (EMB)

This approach is based on the bootstrap sampling procedure and the expectation maximisation algorithm. A sampling method called bootstrap sampling is used to estimate the sampling distribution of statistics, and the expectation maximisation method is a well-known tool for statistical imputation of missing data in various disciplines (Cara 2019; Gunn et al. 2019; Tak, Woo, & Yeo 2016). The EMB method generates a random sample for the bootstrap sample. Next, the expectation maximisation algorithm calculates the maximum likelihood estimate if missing data are available before regressing the data.

EVALUATION INDICATORS

Several key metrics were utilized to assess the study’s imputation techniques. Comparing theoretical and real data helped identify the optimal missing value estimation method. RMSE, MAE, MAPE, NSE, d, KGE, and PBIAS were used to compare the accuracy of the deployed techniques in reconstructing missing water-quality data. The objective function was chosen as NSE since it is the most constraining (Narbondo et al. 2020). RMSE, MAE, and MAPE were utilized for estimation, whereas d, KGE, and PBIAS were employed for validation. These metrics are presented in Equations (1) – (7).

ROOT MEAN SQUARE ERROR (RMSE)

Root mean square error (RMSE) is used in most studies to quantify the difference between imputed and observed values. It essentially represents the sample standard deviation of the difference.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i^{observed} - x_i^{imputed}\right)^2}{n}} \qquad (1)$$

The root mean square error (RMSE) is dimensionally equivalent to the actual and imputed values. The lower the root mean square error, the more accurate the performance of the model.

MEAN ABSOLUTE ERROR (MAE)

Mean absolute error (MAE) is defined as the average difference between imputed and observed data points and is calculated as follows (Avila et al. 2018).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|x_i{}^{observed} - x_i{}^{imputed}\right| \qquad (2)$$

The MAE ranges from 0 to infinity, with 0 being the best fit. Some researchers suggest using MAE instead of RMSE. MAE is more interpretable than RMSE. In mathematics, MAE is the average absolute difference between two variables. MAE is easier to understand than the average of squared errors. Moreover, unlike RMSE, each error affects MAE proportionally to its absolute value.

MEAN ABSOLUTE PERCENTAGE ERROR (MAPE)

In statistics, mean absolute percentage error (MAPE), also known as mean absolute percentage deviation, is a measure of the accuracy of an imputation procedure.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{x_i{}^{observed} - x_i{}^{imputed}}{x_i{}^{observed}}\right| \times 100\% \qquad (3)$$

NASH SUTCLIFFE EFFICIENCY (NSE)

The Nash-Sutcliffe efficiency (NSE) (Nash and Sutcliffe 1970) was used to evaluate the performance of the model.

As a normalised statistic, the NSE determines how much 'noise' is present compared to how much 'information' is present in the data of an experiment (Moriasi et al. 2015). In terms of NSE, the NSE is a measure of how well the observed and estimated data plots match the 1:1 line.

$$NSE = 1 - \frac{\sum_{i=1}^{n}\left(x_i^{observed} - x_i^{imputed}\right)^2}{\sum_{i=1}^{n}\left(x_i^{observed} - \bar{x}\right)^2} \tag{4}$$

NSE is between -∞ and 1.0 (including 1), where NSE = 1 is the optimal value. Performance levels between 0.0 and 1.0 are generally considered acceptable, but values below 0.0 indicate poorer correlation between observed and imputed values, indicating poor performance.

WILLMOTT'S INDEX OF AGREEMENT (WI)

Willmott (1984) introduced the agreement index (WI) as a standard method for assessing the extent of model prediction error. It is calculated by dividing the 'potential error' by the 'mean square error'. It is able to incorporate measurement uncertainty (Martín, Reyes & Taguas 2017).

$$WI = 1 - \frac{\sum_{i=1}^{n}\left(x_i^{observed} - x_i^{imputed}\right)^2}{\sum_{i=1}^{n}\left(\left|x_i^{imputed} - \bar{x}^{observed}\right| + \left|x_i^{observed} - \bar{x}^{observed}\right|\right)^2} \tag{5}$$

KLING-GUPTA EFFICIENCY (KGE)

Unlike NSE, there are no clearly defined criteria for KGE to define a 'good' model. Therefore, in current research, KGE scores are interpreted similarly to NSE: negative scores represent 'poor' model performance, while positive scores represent 'good' model performance (Andersson et al. 2017; Knoben, Woods & Freer 2018). However, a recent study (Knoben, Freer & Woods 2019) suggests that all model results −0.41≤KGE≤1 can be considered good efficiency.

$$KGE = 1 - \sqrt{(r-1)^2 + (\alpha-1)^2 + (\beta-1)^2} \tag{6}$$

PERCENT BIAS (PBIAS)

The percentage bias metric (PBIAS) quantifies the average probability that the simulated data is greater or less than the observed data. PBIAS is ideally 0.0, with low values indicating efficient model simulation. Positive numbers indicate an overestimation of the model, while negative values indicate an underestimation of the model (Moriasi et al. 2015). PBIAS is determined using equation 7, where PBIAS is the deviation of the data analysed, represented as a percentage of the mean.

$$PBIAS = \left[\frac{\sum_{i=1}^{n}\left(x_i^{observed} - x_i^{imputed}\right) * (100)}{\sum_{i=1}^{n}\left(x_i^{observed}\right)}\right] \tag{7}$$

The following table contains the performance values and ratings for NSE, WI, KGE, and PBIAS used in this work, as indicated in Table 2 (Chen et al. 2017; Knoben, Woods & Freer 2018; Moriasi et al. 2015).

TABLE 2. Evaluation indicators and associated rating of performance.

Indicator	Rating of Performance	Physical Water Quality Variables	Chemical Water Quality Variables
NSE	Very good	$NSE > 0.80$	$NSE > 0.65$
	Good	$0.70 < NSE \leq 0.80$	$0.50 < NSE \leq 0.65$
	Satisfactory	$0.45 < NSE \leq 0.70$	$0.35 < NSE \leq 0.50$
	Unsatisfactory	$NSE \leq 0.45$	$NSE \leq 0.35$
PBIAS	Very good	$\lvert PBIAS\rvert < 10$	$\lvert PBIAS\rvert < 15$
	Good	$10 \leq \lvert PBIAS\rvert < 15$	$15 \leq \lvert PBIAS\rvert < 20$
	Satisfactory	$15 \leq \lvert PBIAS\rvert < 20$	$20 \leq \lvert PBIAS\rvert < 30$
	Unsatisfactory	$\lvert PBIAS\rvert \geq 20$	$\lvert PBIAS\rvert \geq 30$
WI	Very good	$0.75 < WI \leq 1.00$	$0.75 < WI \leq 1.00$
	Good	$0.65 \leq WI \leq 0.75$	$0.65 \leq WI \leq 0.75$
	Satisfactory	$0.50 < WI < 0.65$	$0.50 < WI < 0.65$
	Unsatisfactory	$WI \leq 0.5$	$WI \leq 0.5$
KGE	Satisfactory/Good	$KGE \geq -0.41$	$KGE \geq -0.41$
	Unsatisfactory	$KGE < -0.41$	$KGE < -0.41$

RESULTS AND DISCUSSION

Table 3 lists six water quality parameters with the percentage of missing values at each site. The percentage of missing values for ammoniacal nitrogen (AN), biological oxygen demand (BOD), chemical oxygen demand (COD), dissolved oxygen (DO), potential hydrogen (pH) and suspended solids (SS) is greater than 20% at stations S01 and S02 and greater than 50% at station S03. The missingness maps and patterns for the three stations discussed in the previous section are shown in Figure 2.

TABLE 3. Percentage of missing data

Parameter		% Missing Data		
		S01	S02	S03
Chemical	AN	27.08	21.25	53.33
	BOD	23.75	20.42	52.08
	COD	27.50	21.25	52.92
	DO	22.92	22.50	50.42
	pH	23.33	21.67	50.00
Physical	SS	25.00	24.58	51.67

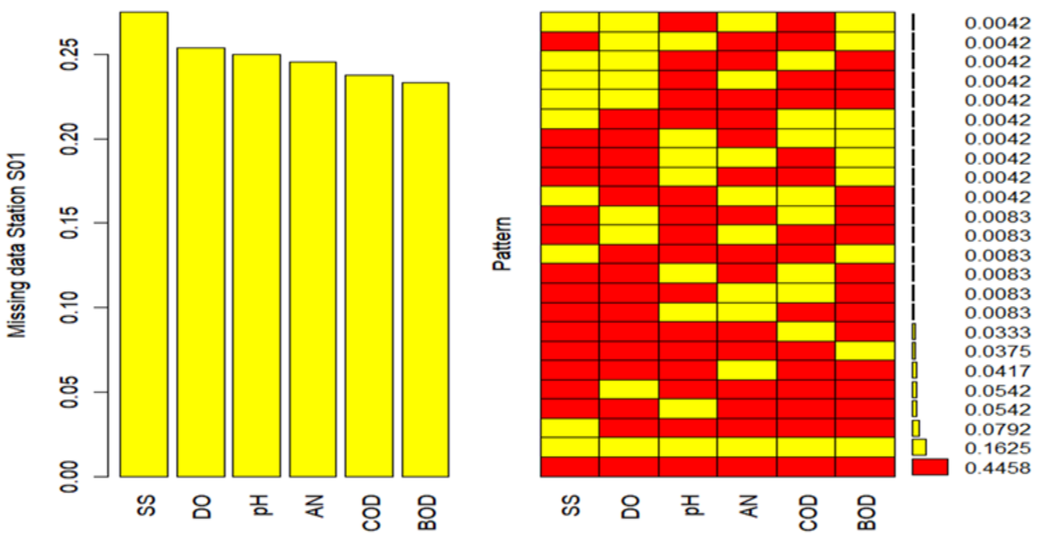
The observed data (available data) is represented by the red highlighted area, while the missing data is represented by the yellow highlighted area (missing data). The map of missing data clearly shows that the parameter of station S03 has a high percentage of missing values, more than 50%, while the parameters of stations S01 and S02 have an average missing value of 20%, which is considered a moderate percentage of missing values.

In this study, PMM, LRNB, BLR, RF, CART, EMB and kNN were compared to determine the optimal imputation method for calculating missing water quality data. The dataset for this study consisted of six water quality indicators and three monitoring stations. It was used to compare the accuracy of the different imputation methods and to select the optimal method for each parameter. Before each analysis, the dataset was min-max normalised to account for the different units and magnitudes.

The most efficient method had the highest NSE and the lowest RMSE, MAE and MAPE (Rodríguez et al. 2021; Moriasi et al. 2015). Consequently, the most accurate method for each parameter was selected and validated with the formulas WI, KGE and PBIAS. The results of this approach are expressed as time series of water quality with a one-month frequency. The leading model for each variable is shown in Tables 4 and 5 respectively, together with the values derived from the performance assessors and the corresponding score. For the NSE assessment, the imputation result is generally considered satisfactory. The water quality parameters measured at the three monitoring sites provided the best estimate with a 'very good' performance for all variables assessed. The validation of the imputation technique was exceptional and gave 'very good' results for the assessments WI, PBIAS and KGE.

Figure 3 shows a boxplot representation of the performance of the technique (NSE, WI, PBIAS and KGE). NSE > 0.75 identifies 100% of imputed data as 'very good' and all imputed data have a positive NSE, indicating that the methodology outperforms the mean function used as an imputer for all imputations. The results of the validation were remarkable. In terms of WI-score and PBIAS scores, 100% of the imputed data are classified as 'very good'. With regard to KGE, all imputations are rated 'good'.

kNN outperformed the other methods in most cases (14 times), followed by CART (3 times) and BLR (1 time). One reason for this is that kNN is the only technique that, in addition to imputation, also takes temporal information into account by including neighbouring observations. The accuracy of the other imputation techniques used is quite similar.



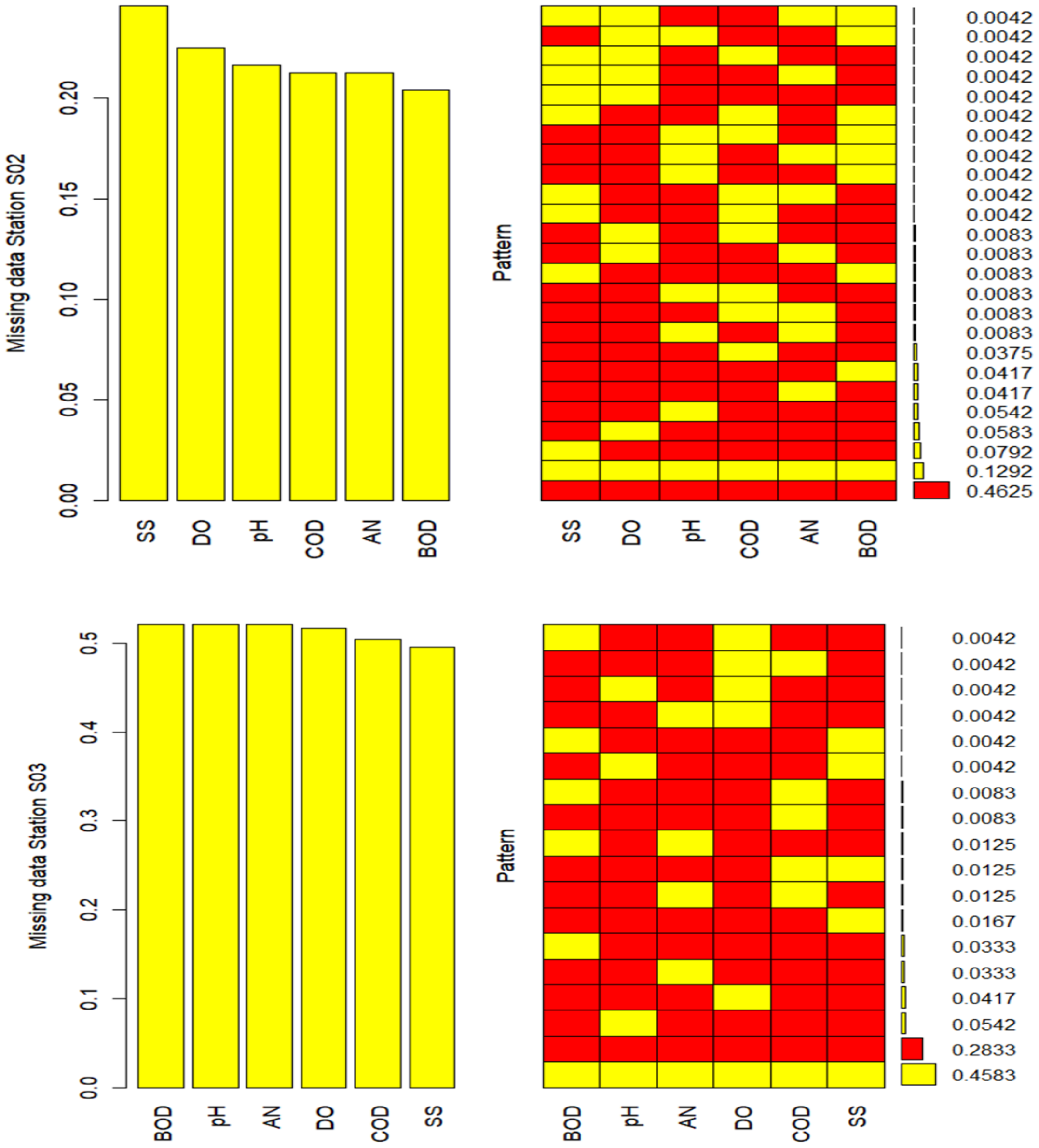


FIGURE 2. Missingness map and pattern at three monitoring stations

TABLE 4. Best imputation methods and corresponding performance evaluator values

PARAMETER	STATION	METHOD	NSE	NSE RATING	RMSE	MAE	MAPE
DO	S01	BLR	0.98	Very good	0.17	0.03	0.01
	S02	CART	0.97		0.44	0.07	0.01
	S03	kNN	0.99		0.16	0.04	0.01
BOD	S01	kNN	0.95	Very good	1.33	0.29	0.06
	S02	kNN	0.97		2.52	0.59	0.05
	S03	CART	0.98		0.75	0.20	0.04
COD	S01	CART	0.93	Very good	4.72	1.36	0.07
	S02	kNN	0.96		55.43	9.92	0.07
	S03	kNN	0.94		5.55	1.63	0.05
SS	S01	kNN	0.94	Very good	93.67	17.85	0.09
	S02	kNN	0.97		68.75	10.55	0.08
	S03	kNN	0.96		37.15	9.6	0.07
pH	S01	kNN	0.76	Very good	0.15	0.03	0.00
	S02	kNN	1.00		0.12	0.03	0.01
	S03	kNN	0.96		0.07	0.02	0
AN	S01	kNN	0.89	Very good	0.47	0.13	0.15
	S02	kNN	1.00		0.49	0.14	0.09
	S03	kNN	0.80		0.60	0.17	0.18

TABLE 5. Best imputation methods and corresponding performance validation values

PARAMETER	STATION	METHOD	WI	WI RATING	PBIAS	PBIAS RATING	KGE	KGE RATING
DO	S01	BLR	0.99	Good	0.10	Very good	0.99	Good
	S02	CART	0.99	Good	-0.4		0.98	
	S03	KNN	1.00	Perfect	0.40		0.99	
BOD	S01	KNN	0.99	Good	1.10	Very good	0.97	Good
	S02	KNN	0.99	Good	1		0.97	
	S03	CART	0.99	Good	-0.60		0.99	
COD	S01	CART	0.98	Good	-0.60	Very good	0.96	Good
	S02	KNN	0.99	Good	3.80		0.95	
	S03	KNN	0.98	Good	1.50		0.96	
SS	S01	KNN	0.98	Good	2.20	Very good	0.96	Good
	S02	KNN	0.99	Good	3.8		0.95	
	S03	KNN	0.99	Good	0.5		0.98	
pH	S01	KNN	0.94	Satisfactory	-0.10	Very good	0.87	Good
	S02	KNN	1.00	Perfect	0.10		1.00	
	S03	KNN	0.99	Good	0		0.98	
AN	S01	KNN	0.97	Good	1.70	Very good	0.92	Good
	S02	KNN	1.00	Perfect	-0.40		1.00	
	S03	KNN	0.95	Good	5.30		0.87	

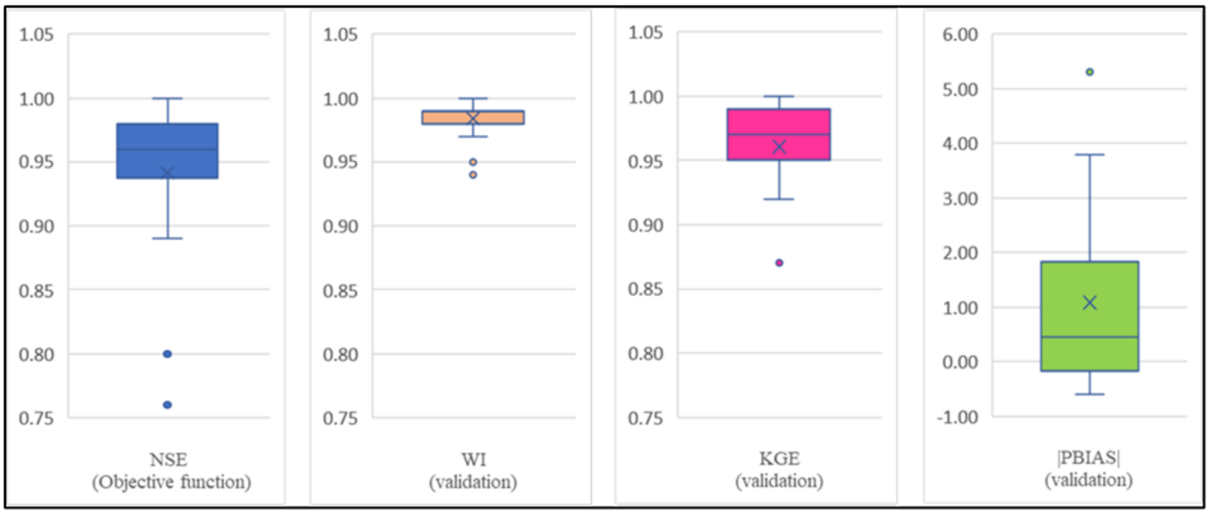


FIGURE 3. Box plots illustrating the performance of model imputation (NSE, WI, KGE, and PBIAS)

CONCLUSION

1. In Malaysia, the Department of Environment (DOE) had implemented WQI to measure the quality of water in for over 25 years.
2. DOE use six water quality parameters to define the status of surface water quality based on national water quality status (NWQS), which are dissolved oxygen (DO), biological oxygen demand (BOD), chemical oxygen demand (COD), pH value, ammoniacal-nitrogen (AN) and total suspended solid (TSS).
3. Hence, this study examined the variables DO, BOD, COD, pH, TSS and AN at three monitoring stations along the Langat River in Selangor, Malaysia.
4. The aim of this study to address the problem of imputing data in a water quality dataset.
5. Since there is no single best method for imputing water quality variables, it was very important to use a range of techniques.
6. The statistical and machine learning methods used in this study were PMM, LRNB, BLR, RF, CART, EM and kNN.
7. Among the implemented methods, kNN was selected as the best because it achieves remarkable accuracy with an optimised performance of 14 out of 18 for the six studied parameters at three monitoring stations.
8. The results show that the performance of kNN is ‘very good’ with an NSE > of 0.75 and the lowest values for RMSE, MAE and MAPE.
9. In addition, all imputed data with WI ≥ 0.94, -0.40 ≤ PBIAS ≤ 5.3 and KGE ≥ 0.87 were also rated as ‘very good’.
10. Consequently, this study provides the basis for future water quality studies in the study catchment, including developing the use of the data now available, so that the results of this study can help water managers and researchers around the world to improve water quality modelling and develop reliable modelling techniques, water quality predictions and sensitivity analyses.
11. It is believed that effective water quality data pollution control techniques can be improved by incorporating geographic information, which has already yielded promising results.

ACKNOWLEDGEMENT

The authors would like to sincerely thank the Earth Observation Centre, Universiti Kebangsaan Malaysia and the Department of Environment Malaysia, Ministry of Natural Resources and Environment for providing data for this study.

DECLARATION OF COMPETING INTEREST

None

NOMENCLATURE

$x_i^{observed}$	Observed data
$x_i^{imputed}$	Imputed data
n	Number of samples
$\bar{x}^{observed}$	Average of observed data
$\bar{x}^{imputed}$	Average of imputed data
$s^{observed}$	Standard deviation of observed data
$s^{imputed}$	Standard deviation of imputed data
r	Linear association between observed and imputed data
α	A measure of the flow variability error
β	A bias term
AN	Ammoniacal nitrogen
BLR	Bayesian linear regression
BOD	Biological oxygen demand

CART	Classification and regression tree
COD	Chemical oxygen demand
DO	Dissolved oxygen
DOE	Department of environment
EMB	Bootstrap-based expectation maximisation
KGE	Kling-Gupta efficiency
kNN	k-nearest neighbours
LRNB	Linear regression (non-Bayesian)
MAE	Mean absolute error
MAPE	Mean absolute percentage error
MI	Multiple imputation
NSE	Nash Sutcliffe efficiency
PBIAS	Percent bias
pH	Potential hydrogen
PMM	Predictive mean matching
RF	Random Forest
RMSE	Root mean square error
SS	Suspended solids
WI	Willmott's index of agreement

REFERENCES

Aguilera, Héctor, Carolina Guardiola-Albert & Carmen Serrano-Hidalgo. 2020. Estimating extremely large amounts of missing precipitation data. *Journal of Hydroinformatics* 22 (3). DOI:https://doi.org/10.2166/hydro.2020.127.

Ahmed, Minhaz Farid, Lubna Alam, Goh Choo Ta, Che Abd Rahim Mohamed & Mazlin Mokhtar. 2016. A review on the environmental pollution of Langat River, Malaysia. *Asian Journal of Water, Environment and Pollution*. DOI:https://doi.org/10.3233/AJW-160035.

Ahmed, Minhaz Farid, Mazlin bin Mokhtar & Nuriah Abd Majid. 2021. Household water filtration technology to ensure safe drinking water supply in the Langat River Basin, Malaysia. *Water (Switzerland)* 13 (8). DOI:https://doi.org/10.3390/w13081032.

Aleryani, Aliya, Wenjia Wang & Beatriz de la Iglesia. 2020. Multiple imputation ensembles (MIE) for dealing with Missing Data. *SN Computer Science* 1 (3). DOI: https://doi.org/10.1007/s42979-020-00131-0.

Andersson, Jafet C.M., Berit Arheimer, Farid Traoré, David Gustafsson & Abdou Ali. 2017. Process refinements improve a Hydrological Model Concept Applied to the Niger River Basin." *Hydrological Processes* 31 (25). DOI:https://doi.org/10.1002/hyp.11376.

Andridge, Rebecca, Laura Bechtel & Katherine Jenny Thompson. 2021. Finding a flexible Hot-Deck imputation method for multinomial data. *Journal of Survey Statistics and Methodology* 9 (4). DOI:https://doi.org/10.1093/jssam/smaa005.

Audigier, Vincent, Ian R. White, Shahab Jolani, Thomas P.A. Debray, Matteo Quartagno, James Carpenter, Stef van Buuren & Matthieu Resche-Rigon. 2018. Multiple Imputation for Multilevel Data with Continuous and Binary Variables. *Statistical Science* 33 (2). DOI:https://doi.org/10.1214/18-STS646.

Avila, Rodelyn, Beverley Horn, Elaine Moriarty, Roger Hodson, & Elena Moltchanova. 2018. Evaluating Statistical Model Performance in Water Quality Prediction. *Journal of Environmental Management* 206 (January): 910–19. DOI:https://doi.org/10.1016/j.jenvman.2017.11.049.

Çadraku & Hazir, S. 2021. Groundwater quality assessment for irrigation: Case Study in the Blinaja River Basin, Kosovo. *Civil Engineering Journal (Iran)* 7 (9). DOI: https://doi.org/10.28991/cej-2021-03091740.

Cara, Javier. 2019. Modal identification of structures from Input/Output data using the expectation-maximization algorithm and uncertainty quantification by mean of the Bootstrap. *Structural Control and Health Monitoring* 26 (1): e2272. DOI:https://doi.org/10.1002/stc.2272.

Cheliotis, Michail, Christos Gkerekos, Iraklis Lazakis & Gerasimos Theotokatos. 2019. A novel data condition and performance hybrid imputation method for energy efficient operations of Marine Systems." *Ocean Engineering* 188 (September): 106220. DOI:https://doi.org/10.1016/j.oceaneng.2019.106220.

Chen, Huajin, Yuzhou Luo, Christopher Potter, Patrick J. Moran, Michael L. Grieneisen & Minghua Zhang. 2017. Modeling pesticide diuron loading from the San Joaquin watershed into the Sacramento-San Joaquin Delta Using SWAT. *Water Research* 121. DOI: https://doi.org/10.1016/j.watres.2017.05.032.

Ebrahimian, Mahboubeh, Ahmad Ainuddin Nuruddin, Mohd Amin Mohd Soom, Alias Mohd Sood, Liew Ju Neng & Hadi Galavi. 2018. Trend analysis of major hydroclimatic variables in the Langat River Basin, Malaysia."*Singapore Journal of Tropical Geography* 39 (2). DOI:https://doi.org/10.1111/sjtg.12234.

Gunn, Virginia, Carles Muntaner, Edwin Ng, Michael Villeneuve, Montserrat Gea-Sanchez & Haejoo Chung. 2019. Gender Equality Policies, Nursing Professionalization, and the Nursing Workforce: A Cross-Sectional, Time-Series Analysis of 22 Countries, 2000–2015. *International Journal of Nursing Studies* 99. DOI:https://doi.org/10.1016/j.ijnurstu.2019.103388.

Hadeed, Steven J., Mary Kay O'Rourke, Jefferey L. Burgess, Robin B. Harris & Robert A. Canales. 2020. Imputation methods for addressing missing data in short-term monitoring of air pollutants. *Science of The Total Environment* 730 (August): 139140. DOI: https://doi.org/10.1016/j.scitotenv.2020.139140.

Hamzah, Fatimah Bibi, Firdaus Mohd Hamzah, Siti Fatin Mohd Razali & Hafiza Samad. 2021. A Comparison of Multiple Imputation Methods for Recovering Missing Data in Hydrological Studies. *Civil Engineering Journal (Iran)* 7 (9). DOI:https://doi.org/10.28991/cej-2021-03091747.

Hamzah, Fatimah Bibi, Firdaus Mohd Hamzah, Siti Fatin Mohd Razali, Othman Jaafar & Norhayati Abdul Jamil. 2020. Imputation methods for recovering streamflow observation: A Methodological review. *Cogent Environmental Science* 6 (1). DOI:https://doi.org/10.1080/23311843.2020.1745133.

Hayati Rezvan, Panteha, Katherine J. Lee & Julie, A. Simpson. 2015. The rise of multiple imputation: A review of the reporting and implementation of the method in Medical Research. *BMC Medical Research Methodology* 15 (1): 30. DOI:https://doi.org/10.1186/s12874-015-0022-1.

Kerich, Emmy C. 2020. Households drinking water sources and treatment methods options in a regional irrigation scheme. *Journal of Human, Earth, and Future* 1 (1). DOI:https://doi.org/10.28991/hef-2020-01-01-02.

Khan, Afed Ullah, Hafiz Ur Rahman, Liaqat Ali, Muhammad Ijaz Khan, Humayun Mehmood Khan, Afnan Ullah Khan & Fayaz Ahmad Khan, et al. 2021. Complex linkage between watershed attributes and surface water quality: Gaining insight via path analysis. *Civil Engineering Journal (Iran)* 7 (4). DOI: https://doi.org/10.28991/cej-2021-03091683.

Knoben, Wouter J.M., Jim E. Freer & Ross A. Woods. 2019. Technical Note: Inherent Benchmark or Not? Comparing Nash-Sutcliffe and Kling-Gupta Efficiency Scores. *Hydrology and Earth System Sciences* 23 (10). DOI:https://doi.org/10.5194/hess-23-4323-2019.

Knoben, Wouter J.M., Ross A. Woods & Jim E. Freer. 2018. A quantitative hydrological climate classification evaluated with independent streamflow data. *Water Resources Research* 54 (7). DOI: https://doi.org/10.1029/2018WR022913.

Little, R. J. & D. B. Rubin. 2019. *Statistical Analysis with Missing Data* . Edited by John Wiley & Sons. Vol. 793.

Liu, Xinwang, Lei Wang, Xinzhong Zhu, Miaomiao Li, En Zhu, Tongliang Liu, Li Liu, Yong Dou & Jianping Yin. 2020. Absent multiple kernel learning algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42 (6). DOI: https://doi.org/10.1109/TPAMI.2019.2895608.

Mandel J, Schmitt P. 2015. A comparison of six methods for missing data imputation. *Journal of Biometrics & Biostatistics* 06 (01). DOI:https://doi.org/10.4172/2155-6180.1000224.

Martín, Miguel Ángel, Miguel Reyes & F. Javier Taguas. 2017. Estimating Soil Bulk Density with Information metrics of soil texture. *Geoderma* 287. DOI: https://doi.org/10.1016/j.geoderma.2016.09.008.

Moriasi, D. N., M. W. Gitau, N. Pai & P. Daggupati. 2015. Hydrologic and water quality models: performance measures and evaluation criteria.*Transactions of the ASABE* 58 (6): 1763–85. DOI: https://doi.org/10.13031/trans.58.10715.

Morita, Kojiro. 2021. Introduction to multiple imputation. *Annals of Clinical Epidemiology* 3 (1). DOI:https://doi.org/10.37737/ace.3.1_1.

Murray, Jared S. 2018. Multiple imputation: A review of practical and theoretical findings. *Statistical Science* 33 (2). https://doi.org/10.1214/18-STS644.

Narbondo, Santiago, Angela Gorgoglione, Magdalena Crisci & Christian Chreties. 2020. Enhancing physical similarity approach to predict runoff in ungauged watersheds in sub-tropical regions. *Water (Switzerland)* 12 (2). DOI:https://doi.org/10.3390/w12020528.

Nash, J.E. & J.V. Sutcliffe. 1970. River flow forecasting through conceptual models Part I — A discussion of principles. *Journal of Hydrology* 10 (3): 282–90. DOI: https://doi.org/10.1016/0022-1694(70)90255-6.

Ratolojanahary, Romy, Raymond Houé Ngouna, Kamal Medjaher, Jean Junca-Bourié, Fabien Dauriac & Mathieu Sebilo. 2019. Model selection to improve multiple imputation for handling high rate missingness in a water quality dataset. *Expert Systems with Applications* 131 (October): 299–307. DOI: https://doi.org/10.1016/j.eswa.2019.04.049.

Rodríguez, Rafael, Marcos Pastorini, Lorena Etcheverry, Christian Chreties, Mónica Fossati, Alberto Castro, & Angela Gorgoglione. 2021. Water-quality data imputation with a high percentage of missing values: A machine learning approach. *Sustainability* 13 (11): 6318. DOI:https://doi.org/10.3390/su13116318.

Santos, Miriam Seoane, Pedro Henriques Abreu, Szymon Wilk & João Santos. 2020. How distance metrics influence missing data imputation with K-Nearest Neighbours. *Pattern Recognition Letters* 136. DOI:https://doi.org/10.1016/j.patrec.2020.05.032.

Schmitt, Peter, Jonas Mandel & Mickael Guedj. 2015. A comparison of six methods for missing data imputation. *Journal of Biometrics & Biostatistics* 06 (01). DOI: https://doi.org/10.4172/2155-6180.1000224.

Ser, Gazel, Siddik Keskin & M. Can Yilmaz. 2016. The performance of multiple imputations for different number of imputations. In *Sains Malaysiana*. Vol. 45.

Shin, Tacksoo, Mark L. Davison & Jeffrey D. Long. 2017. Maximum likelihood versus multiple imputation for missing data in small longitudinal samples with nonnormality. *Psychological Methods* 22 (3): 426–49. DOI:https://doi.org/10.1037/met0000094.

Silva, Anurika Priyanjali de, Margarita Moreno-Betancur, Alysha Madhu de Livera, Katherine Jane Lee & Julie Anne Simpson. 2017. A comparison of multiple imputation methods for handling missing values in longitudinal data in the presence of a time-varying covariate with a non-linear association with time: A simulation study. *BMC Medical Research Methodology* 17 (1). DOI:https://doi.org/10.1186/s12874-017-0372-y.

Tak, Sehyun, Soomin Woo & Hwasoo Yeo. 2016. Data-driven imputation method for Traffic Data in Sectional Units of Road Links. *IEEE Transactions on Intelligent Transportation Systems* 17 (6): 1762–71. DOI:https://doi.org/10.1109/TITS.2016.2530312.

Wan Mohtar, Wan Hanna Melini, Siti Aminah Bassa Nawang & Mohd Noor Shafique Rahman. 2017. Statistical analysis in fluvial sediments of Selangor Rivers: Downstream variation in grain size distribution. *Jurnal Kejuruteraan* S (1): 37–45. DOI: https://doi.org/10.17576/jkukm-s-01-06.

Willmott, Cort J. 1984. On the evaluation of model performance in physical geography. In *Spatial Statistics and Models*, 443–60. Dordrecht: Springer Netherlands. DOI: https://doi.org/10.1007/978-94-017-3048-8_23.