



A Study on Bayesian Principal Component Analysis for Addressing Missing Rainfall Data

Wai Yan Lai¹  · K. K. Kuok¹

Received: 15 August 2018 / Accepted: 3 February 2019 /

Published online: 1 June 2019

© Springer Nature B.V. 2019

Abstract

This paper proposed the application of Bayesian Principal Component Analysis (BPCA) algorithm to address the issue of missing rainfall data in Kuching City. The experiment was conducted using six different combinations of rainfall data from different neighbouring rainfall stations at different missing data entries (1%, 5%, 10%, 15%, 20%, 25% and 30% of missing data entries). The performance of BPCA model in reconstructing the missing data was examined with respect to Bias (B_s), Efficiency (E) and Root Mean Square Error (RMSE). The reliability and robustness of BPCA was confirmed by comparing its performance with K-Nearest Neighbour (KNN) imputation model. The results support the addition of data from neighbouring rainfall stations to improve the imputation accuracy.

Keywords Bayesian principal component analysis (BPCA) · K-nearest neighbour (KNN) · Missing rainfall data · Imputation

1 Introduction

Rainfall is one of the most important hydrological parameters used in most hydrological and climatological studies (Kamaruzaman et al. 2017; Lee and Kang 2015). However, the occurrences of missing data are critical and unavoidable in various fields of research. Missing data may be contributed by human errors in managing the datasets, equipment failure and natural disasters that may damage the gauging equipment on site. The direct impact of having missing data will be the lack of input data or samples for performing any simulations. Consistent and complete rainfall datasets are required to obtain accurate hydrological simulation and prediction studies (Jajarmizadeh et al. 2015; Sattari et al.

✉ Wai Yan Lai
wlai@swinburne.edu.my

¹ Faculty of Engineering, Computing and Science, Swinburne University of Technology Sarawak Campus, Jalan Simpang Tiga, 93350 Kuching, Sarawak, Malaysia

2016). Thus, missing rainfall data need to be handled carefully to enhance the reliability of hydrological studies.

To address the missing rainfall observations, listwise deletion, pairwise deletion, zero imputation, and hot deck imputation are commonly adopted (Kamaruzaman et al. 2017; Pagano et al. 2014). However, these methods are yet proven to be reliable, accurate and scientifically approved. Listwise deletion and pairwise deletion eliminate the missing observations. Hence, using any data deletion methods will cause the loss of information and reduction in sample size. Zero imputation includes the substitution of missing observations with zeros. Replacing the missing entries with zeros will disrupt the nature of the data distribution. Thus, it may create bias and error in further studies. Zero imputation may be applicable in some of the hydrological parameters such as rainfall. However, it may not be suitable to be used in ground water level and studies with negative values (Gill et al. 2007). “Hot deck” imputation method is more reliable when compared to listwise deletion, pairwise deletion, and zero imputation. It is currently used in Malaysia to replace missing observations with available observations from other nearby gauging equipment or rainfall stations (Malek et al. 2010). However, this method is not reliable if the missing observations occurred simultaneously at the other gauging equipment and nearby rainfall stations. All these methods may create biases and result in unreliable and inaccurate studies.

Throughout the research on the impact of missing data, it is encouraged to implement data imputation during the data pre-processing process to boost the performance of the prediction studies. Ekeu-wei (2018) performed an experiment to estimate flood by adopting datasets with missing and imputed observations. The imputed observations were predicted using Monte Carlo Multiple imputation approach. The results show that using imputed observations can boost the accuracy of flood estimate consistently. The findings also suggest that using datasets with missing observations will cause underestimation and overestimation of flood estimate. Kuok and Bessaih (2007) used artificial neural network (ANN) to predict the daily rainfall runoff of Sungai Bedup Basin. The results indicate that ANNs performed better with the increased supply of input data. These findings highlight the importance of sustaining consistent and long term climatological and hydrological data. The literatures also emphasized on the implementation of data imputation to increase the data availability. By doing so, it can boost the performance and accuracy of simulation and prediction studies.

Statistical approaches, data mining approaches and machine learning approaches such as ANN and K-nearest neighbour (KNN), are some of the approaches that can be used to perform data imputation. Oba et al. (2003) created Bayesian Principal Component Analysis (BPCA) to address the missing values of gene expression profile data. The BPCA model outperformed the KNN impute and singular value decomposition method (SVD) in imputing the missing data. Bennett et al. (2007) used nearest neighbour by distance (ND) and correlation (NC), inverse distance weighted (IDW), average of gauges selected by correlation (A), and weighted average of gauges selected by correlation (WA) to impute the missing rainfall data. The results showed that WA method outperformed all the other proposed methods.

ANNs are also applied in the hydrological field for performing data prediction tasks. Luk et al. (2001) used multilayer feedforward network (MLFN), partial recurrent neural network (PRNN), and time delay neural network (TDNN) to predict the rainfall values.

Kuok and Bessaih (2007) used multilayer perceptron (MLP) and recurrent (REC) network to estimate the daily rainfall runoff of Sungai Bedup Basin. Particle Swarm Optimisation Feedforward Neural Network (PSOENN) was also created by Kuok et al. (2010) to calibrate the water tank model and the relationship of the rainfall runoff model at Sungai Bedup Basin. Chai et al. (2017) estimated the rainfall data by using six daily meteorology data and two types of neural networks: Backpropagation Neural Network (BPN) and Radial Basis Function Network (RBFN).

In this study, the potential of using BPCA imputation model to treat missing rainfall data is investigated. The accuracy and robustness of BPCA imputation model is expected to be the major challenge in this study. Accurate prediction of hydrological data such as rainfall data is challenging due to their high degree of temporal and spatial variability, and non-linear characteristics (Bennett et al. 2007; Chai et al. 2017; Gill et al. 2007). Furthermore, different climate zones have different rainfall pattern and spatial distribution. This increases the challenge in reconstructing the missing rainfall data because different climate zones have different best imputation method (De Silva et al. 2007). To the knowledge of the authors, there is no published work that applies BPCA imputation model to treat missing rainfall data. The BPCA imputation model is known to have good imputation performance in the medical domain. Hence, it is motivated to study the application of BPCA algorithm in patching the missing rainfall data. Considering all the issues and challenges, there is a need to develop a novel approach to apply and boost the imputation performance of BPCA model in treating the missing rainfall data. As such, the objectives of this study are aligned as below:

- To predict the missing rainfall data using BPCA model and rainfall data
- To study the parameters that will affect the performance of BPCA model
- To evaluate the performance of BPCA model with the introduction of reference rainfall data from neighbouring rainfall station
- To compare the performance of BPCA model with existing imputation model, KNN within the study area

2 Study Area and Rainfall Stations

The Kuching City in Sarawak, Malaysia was chosen as the study area of this research study. The rainfall data within Sarawak River Basin were adopted in this study. The distance between the rainfall stations was set to be the benchmark for selecting the neighbouring stations. The rainfall data from further stations are expected to have large difference in terms of spatial and temporal distribution that may lower the imputation performance. Hence, only the rainfall data from neighbouring rainfall stations were selected in this study.

The rainfall stations at Kuching Saberkas (1), Kuching Third Mile (2), Ulu Maong (3), and Kuching Airport (4) were selected in this study. The location of the selected rainfall stations was illustrated in Fig. 1. They are relatively close to one another when compared to other available stations in Kuching. The daily rainfall data in the year 1991 were collected from Department of Irrigation and Drainage (DID) Sarawak. The rainfall data from the four stations were analysed to study the impact of distance and correlation of data between the neighbouring

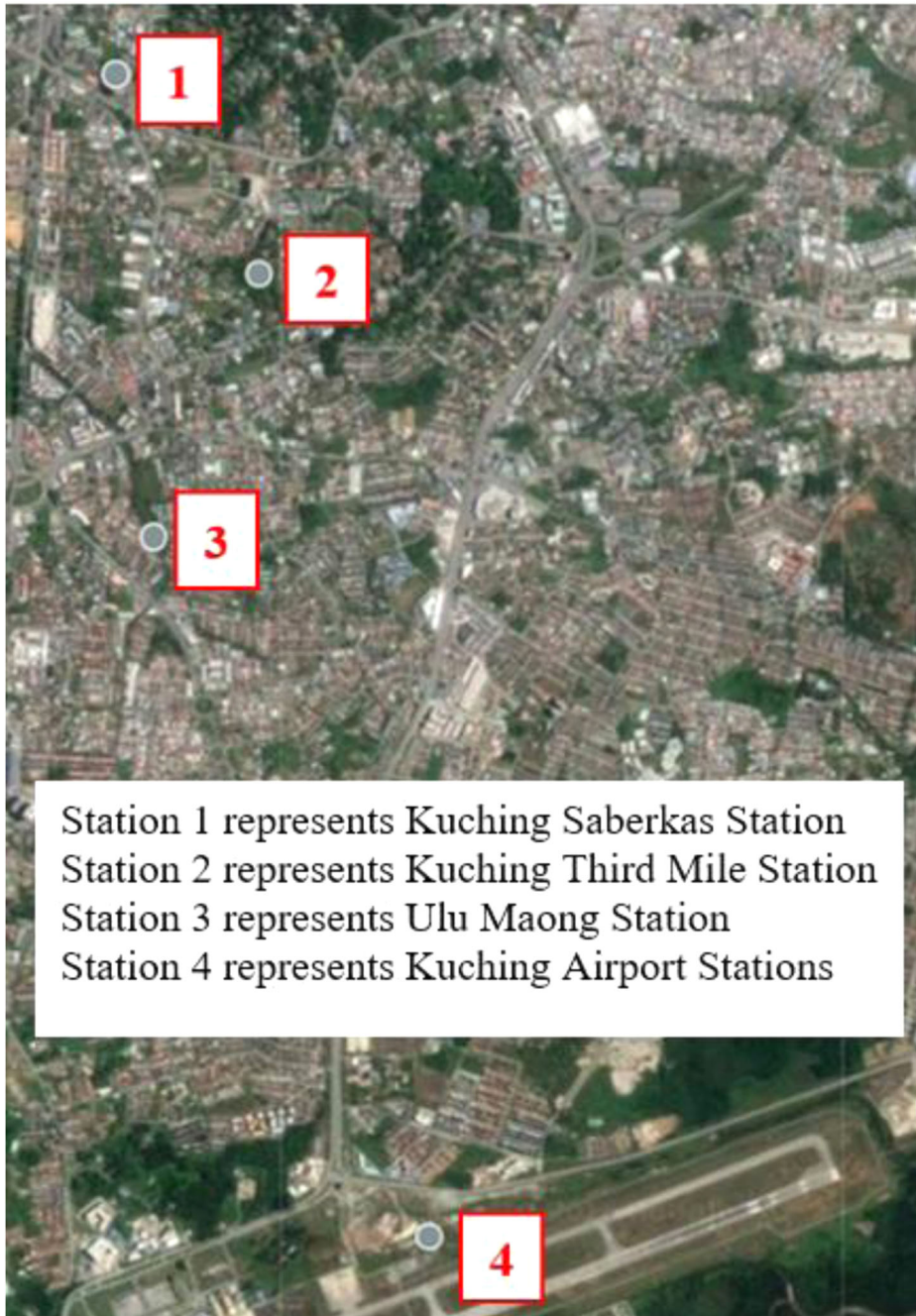


Fig. 1 Selected rainfall station

rainfall stations on the imputation performance of BPCA model. The correlation coefficient (r) of the rainfall data between stations were calculated using Eq. (1).

$$r = \frac{\sum(A-\bar{A})(B-\bar{B})}{\sqrt{\sum(A-\bar{A})^2 \sum(B-\bar{B})^2}} \tag{1}$$

where,

- A data from Station A
- B data from Station B
- \bar{A} mean of the data from Station A
- \bar{B} mean of the data from Station B

2.1 Data Correlation between the Selected Rainfall Stations

The correlations of $r \geq 0.7$, $0.4 \leq r < 0.7$, and $r < 0.4$ are defined as a high, medium, and low correlation relationship, respectively. The coefficient of correlations between the stations are tabulated in Table 1. From Table 1, it shows that the rainfall datasets are considered as highly correlated because all the r fall between the range of 0.75–0.97. It is observed that r decreases as the distance between the rainfall stations increases. The pairing of stations with Kuching Airport Station result in lower r than the pairs without Kuching Airport Station. This may be due to the small difference in rainfall received between the three rainfall stations other than Kuching Airport Station. The collected data shows that the rainfall amount received at Kuching Airport Station is significantly lower than the other rainfall stations. Another reason may be due to the geographical location of the rainfall stations. Ulu Maong Station is closer to Kuching Airport Station when compared to Kuching Saberkas Station and Kuching Third Mile Station. Thus, the correlation between Kuching Airport Station and Ulu Maong Station is higher than the other two stations.

3 Imputation Models

3.1 Bayesian Principal Component Analysis (BPCA)

The BPCA imputation model that is created by Oba et al. (2003) considers the whole dataset of gene expression profiles by a matrix, Y . Y is arranged in the order of $(D \times N)$.

Table 1 Calculated correlation between the rainfall stations

Station A	Station B	Coefficient of correlation, r
Kuching Third Mile	Kuching Airport	0.7911
	Kuching Saberkas	0.9711
	Ulu Maong	0.9452
Kuching Airport	Kuching Saberkas	0.7513
	Ulu Maong	0.8663
Kuching Saberkas	Ulu Maong	0.9083

N and D are known as the number of genes and the number of samples, respectively. The prediction of missing values is executed based on three elementary processes: principal component (PC) regression, followed by Bayesian estimation and the expectation-maximization (EM) like repetitive algorithm. The first two steps, PC regression and Bayesian estimation, are used for deriving, determining and setting up appropriate parameters. The missing values estimation will only be carried out at EM like repetitive algorithm represented by Eq. (2). The details of the derivations and assumptions had been outlined by Oba et al. (2003) and Oba (2013).

$$\hat{Y}^{miss} = \int Y^{miss} q(Y^{miss}) dY^{miss} \quad (2)$$

where,

\hat{Y}^{miss} imputed missing variables of matrix Y
 Y^{miss} missing variables of matrix Y
 $q(Y^{miss})$ posterior distribution of missing value

The BPCA imputation model had been applied widely in the field of biomedical for patching the missing microarray data. Shi et al. (2013) proposed a new hybrid imputation method that utilised both BPCA imputation and Local Least Square (LLS) imputation. The proposed method was named as Bayesian Principal Component Analysis and Iterative Local Least Square method (BPCA-iLLS). The BPCA-iLLS model outperformed the BPCA model and LLS model. However, the performance of BPCA and LLS models varied significantly when different datasets were used to perform the imputations. The literature also showed that LLS tends to outperform BPCA when dominant local similarity exists within the dataset. On the other hand, BPCA works better when the datasets have lower complexity. Another similar approach had been done by Severson et al. (2017). Several Principal Component Analysis (PCA) based methods were introduced and evaluated for imputing the missing microarray data. The methods that had been used in their studies were mean imputation, alternating least squares (ALS), singular value decomposition method (SVDImpute), probabilistic principal component analysis (PPCA), PCA-data augmentation (PCADA), PPCA-M (another variation of PPCA), BPCA, singular value thresholding (SVT), another variation of alternating least square (Alternating), and Lagrange multiplier method (ALM). It was mentioned that the SVDImpute and the probabilistic methods (PPCA, PPCA-M, and BPCA) performed the best overall. However, it was suggested that the suitability of the methods chosen for performing the imputation may vary. The missingness mechanism is found to be the main factor that affects the suitability of the imputation methods.

The application of BPCA is not only limited within the biomedical field. It was also utilised for imputing the missing data of total electron content (TEC) Ionospheric satellite dataset. Under the work performed by Subashini and Krishnaveni (2011), the BPCA model was proven to be better than KNN imputation for imputing the missing TEC data. Other than imputation, BPCA was also applied for speech feature analysis. Oh-Wook et al. (2003) had proposed variational BPCA to estimate the speech feature dimensionality and the number of clusters used in Gaussian mixture model. The literatures imply that it is possible to implement BPCA in other fields of research. Hence, BPCA imputation model is introduced in this paper to impute the missing rainfall data.

3.2 K-Nearest Neighbour (KNN)

Lee and Kang (2015) patched the missing rainfall data using KNN regression with five different kernel estimation functions (Epanechnikov, Quartic, Triweight, Tricube and Cosine). The imputed rainfall datasets were then used to simulate the water runoff using Soil Water Assessment Tool (SWAT). The study showed that KNN can be applied to patch the missing hydrological data. It is also significant that utilising different kernel functions can improve the performance of KNN imputation in predicting the missing rainfall data. By doing so, it actually helps to enhance the accuracy of streamflow simulations.

As such, KNN imputation method is introduced in this paper to compare the performance of BPCA and KNN. The purpose of comparing the performance of BPCA and KNN is to observe the reliability and robustness of BPCA. The performance of KNN in missing data imputation had been proven to be reliable in both biomedical field and hydrological field. A built-in KNN imputation function in MATLAB, “*knnimpute*” was adopted in this research study. The KNN imputation function will impute the missing data by referring to the reference values from the nearest neighbour column with no missing values. The nearest-neighbour column is determined by identifying the Euclidean distance as shown in Eq. (3).

$$Euclidean \ distance = \sqrt{\sum_{i=1}^n (q_i - p_i)^2} \quad (3)$$

where p and q are the vectors of two different datasets.

4 Methodology

The missingness mechanism in this study was assumed to be Missing Completely at Random (MCAR). Malek (2008) stated that the cause of missing rainfall data in Malaysia is mainly due to errors and mistakes in data management, human resources, instrumentation, operation and maintenance. Hence, the missing rainfall data is not caused by the occurrences of random events. In order to evaluate the ability of the imputation models, the general experiment procedures were outlined as below:

Step 1: Collection of daily rainfall data from DID Sarawak

Step 2: Creation of six different input datasets without any missing values

Step 3: Introduction of artificial missing entries for all the datasets (1%, 5%, 10%, 15%, 20%, 25% and 30% of missing rainfall data entries)

Step 4: Import the rainfall data and source code into MATLAB

Step 5: Execution of the imputation under different parameters settings (different K values and percentage of missing data entries)

Step 6: Evaluation on the performance of BPCA model and KNN model using different evaluation methods

The selected rainfall data were arranged into six different input datasets. The datasets were created by combining the daily rainfall data of different neighbouring stations in a matrix form of $(X \times Y)$. X and Y represent daily rainfall amount and months,

Table 2 Evaluation summary for BPCA and KNN imputation

Imputation method	BPCA										KNN									
	Missing entries (%)	1	5	10	15	20	25	30	1	5	10	15	20	25	30					
Kuching Third Mile																				
B _s	1.0083	1.0223	1.0354	1.0177	1.0432	1.0140	0.9923	1.0094	1.0075	1.0172	1.0033	0.9993	1.0004	0.9743						
RMSE (mm)	1.1144	1.6915	2.6485	5.7350	6.3892	7.6825	9.0328	9.249	1.9839	2.9845	5.5958	6.2526	7.8557	9.2661						
E	0.9967	0.9924	0.9814	0.9127	0.8917	0.8434	0.7835	0.9977	0.9896	0.9764	0.9173	0.8968	0.8389	0.7758						
Kuching Third Mile & Ulu Maong																				
B _s	1.0046	1.0205	1.0354	1.0100	1.0307	1.0140	0.9923	0.9999	1.0004	1.0017	0.9996	0.9842	0.9804	0.9551						
RMSE (mm)	0.7525	1.6210	2.4690	4.7880	5.2720	6.2091	9.0328	0.1739	0.6849	1.0673	1.8368	3.2220	5.1610	6.5032						
E	0.9985	0.9930	0.9838	0.9392	0.9262	0.8977	0.7835	0.9999	0.9988	0.9970	0.9910	0.9724	0.9293	0.8878						
Kuching Third Mile & Kuching Saberkas																				
B _s	1.0021	1.0198	1.0354	1.0126	1.0239	1.0140	0.9974	0.9995	0.9990	0.9993	1.0008	1.0025	0.9998	0.9932						
RMSE (mm)	0.3526	1.3685	2.3242	4.0575	5.0257	5.6419	6.9788	0.0970	0.3897	0.8425	4.2121	5.2980	6.5802	7.8860						
E	0.9997	0.9950	0.9857	0.9563	0.9330	0.9155	0.8708	1.0000	0.9996	0.9981	0.9529	0.9255	0.8869	0.8376						
Kuching Third Mile & Kuching Airport																				
B _s	1.0050	1.0223	1.0325	1.0125	1.0368	1.0101	1.0021	0.9985	1.0025	1.0104	0.9984	0.9752	0.9798	0.9629						
RMSE (mm)	0.6309	1.5762	2.7474	5.7076	6.1439	7.6217	9.0328	0.2362	1.5289	2.5907	3.5788	5.4407	6.3311	8.0421						
E	0.9989	0.9934	0.9800	0.9135	0.8998	0.8458	0.7835	0.9999	0.9938	0.9822	0.9660	0.9214	0.8936	0.8284						
Kuching Third Mile, Kuching Saberkas & Ulu Maong																				
B _s	1.0011	1.0098	1.0152	1.0078	1.0211	1.0099	0.9987	1.0000	0.9994	1.0000	1.0003	0.9924	0.9986	0.9701						
RMSE (mm)	0.3047	0.9488	1.4934	2.5930	3.4675	4.1249	5.6238	0.0939	0.3715	0.7475	1.4883	3.2927	4.8407	6.2522						
E	0.9998	0.9976	0.9941	0.9822	0.9681	0.9548	0.9161	1.0000	0.9996	0.9985	0.9941	0.9712	0.9378	0.8963						
Kuching Third Mile, Kuching Saberkas, Ulu Maong & Kuching Airport																				
B _s	1.0019	1.0184	1.0246	1.0025	1.0153	1.0074	0.9986	0.9998	1.0005	0.9991	1.0004	1.0003	0.9986	0.9746						
RMSE (mm)	0.3553	1.3818	1.8929	3.0935	3.7875	4.4287	5.8424	0.0939	0.3715	0.7890	1.6219	3.6990	5.0848	6.4879						
E	0.9997	0.9949	0.9905	0.9746	0.9619	0.9479	0.9094	1.0000	0.9996	0.9983	0.9930	0.9637	0.9314	0.8883						

respectively. The motivation of creating different input datasets is to observe and evaluate the performance of imputation models under the increment of data availability. The relationship between the data correlation and imputation performance can also be observed with the utilisation of different input datasets. The input datasets were created by setting Kuching Third Mile rainfall station as the imputation and evaluation target. The missing data entries of 1%, 5%, 10%, 15%, 20%, 25% and 30% were artificially created and introduced in the rainfall data of Kuching Third Mile. The combination of datasets can be observed in the following list:

1. Kuching Third Mile
2. Kuching Third Mile & Kuching Saberkas
3. Kuching Third Mile & Ulu Maong
4. Kuching Third Mile & Kuching Airport
5. Kuching Third Mile, Kuching Saberkas & Ulu Maong
6. Kuching Third Mile, Kuching Saberkas, Ulu Maong & Kuching Airport

Similar to KNN imputation, BPCA also uses K as the selection parameter. The maximum adoptable K value will depend on the nature of the algorithms. The parameter, K, simply refers to the number of training samples that are needed to be referenced for performing the imputation. The performance of the imputation models was evaluated using Bias (B_s), Root Mean Square Error (RMSE) and Efficiency (E). The equations of the evaluation criteria are listed as in Eqs. (4) to (6). A perfect estimation of the missing observations will result in $B_s = 1$, $RMSE = 0$ and $E = 1$. The evaluation methods were selected based on relevant hydrological prediction studies as performed by Wang et al. (2016) (for B_s and RMSE) and Malek (2008) (for E). These evaluation methods account for the drastic and rapid behaviour change of convective precipitation field.

$$Bias, B_s = \frac{\sum_{i=1}^N F_i}{\sum_{i=1}^N O_i} \tag{4}$$

$$Root \ mean \ square \ error, RMSE = \sqrt{\frac{\sum_{i=1}^N (O_i - F_i)^2}{N}} \tag{5}$$

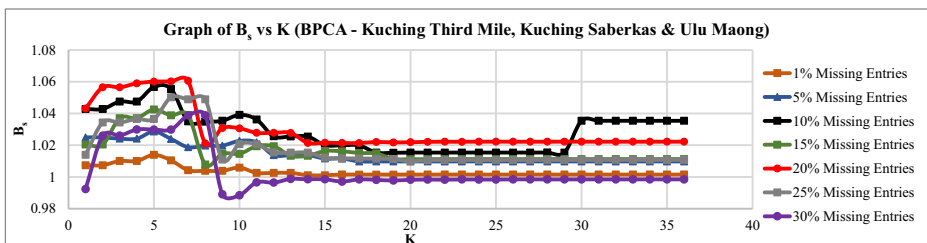


Fig. 2 Graph of B_s vs K (BPCA - Kuching Third Mile, Kuching Saberkas & Ulu Maong)

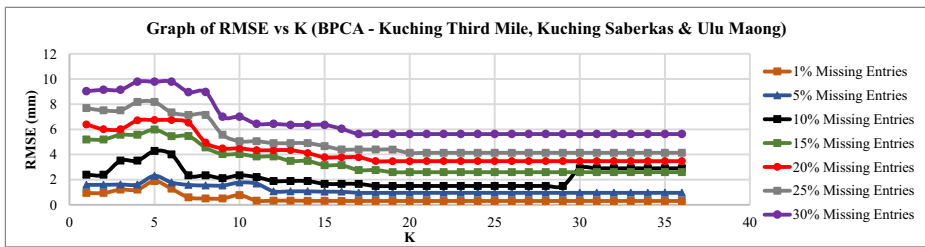


Fig. 3 Graph of RMSE vs K (BPCA - Kuching Third Mile, Kuching Saberkas & Ulu Maong)

$$Efficiency, E = \frac{\sum(O - \bar{O})^2 - \sum(O - F)^2}{\sum(O - \bar{O})^2} \tag{6}$$

where,

- F imputed value or predicted value
- O original value or observed value
- \bar{O} mean of original value or observed value
- \bar{F} mean of imputed value or observed value
- N number of data

5 Results and Discussion

The summary of the evaluation against the imputation models is tabulated in Table 2. To ease the difficulty of comparing the imputation performance for all the data combinations, Table 2 only tabulates the best imputation performance achieved by both of the imputation models at different experimental settings. Figures 2, 3, 4, 5, 6, and 7 illustrate the imputation performance of BPCA and KNN models at different K values and percentage of missing entries. Other graphs are not included in this paper as they show similar result patterns.

The performance of both KNN and BPCA do fluctuates as the K value increases. However, the range of K values utilised by BPCA is different from KNN. For BPCA, the range of K is defined to be equal to the number of column within the dataset (Oba 2013). Hence, the range of K values adopted for BPCA are different for each data combination. When only one rainfall

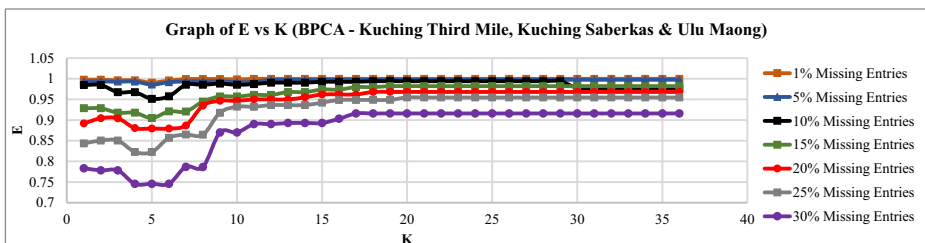


Fig. 4 Graph of E vs K (BPCA - Kuching Third Mile, Kuching Saberkas & Ulu Maong)

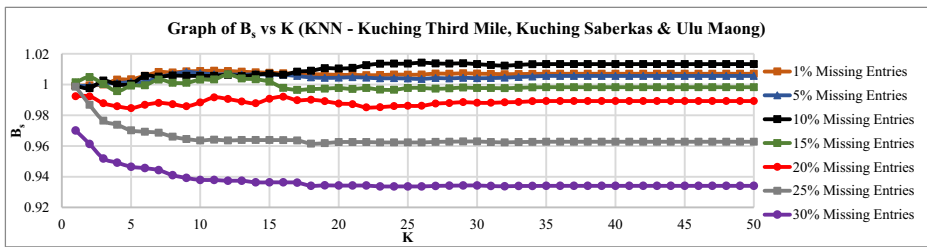


Fig. 5 Graph of B_s vs K (KNN – Kuching Third Mile, Kuching Saberkas & Ulu Maong)

dataset is used, the adoptable K values fall between the range of 1 to 12. The maximum adoptable K value increases by another 12 units when an additional rainfall data from one of the rainfall stations is added in. Figures 2, 3, and 4 show that the adopted K values fall within the range of 1 to 36. This is due to the utilisation of 3 rainfall datasets. Small fluctuation of performance is expected as the range of the K values obtained in this study is relatively small (maximum range of $1 \leq K \leq 48$). Unlike the experiment conducted by Oba et al. (2003), large performance difference is observed as the K values fall within the range of 1 to 200. Figures 2, 3, and 4 also show that the best and similar imputation performance can be achieved at different K values. This is different from the results obtained in the experiment performed by Oba et al. (2003). The results show that in gene profile data imputation, the BPCA model performed the best at $K = D - 1$. This might be due to the nature of the algorithm or rainfall data. The nature of rainfall data is much more random and has non-linear pattern. It is also observed that different occurrences of rain, rainfall amount and pattern were experienced by each of the rainfall station on the same timeline. They did not seem to be bounded or caused by any significant factor or reason.

For KNN, the maximum K value is not defined and definite. This requires the users to identify the convergence point to stop the increment of K value. To cope with this issue, KNN model is tested within the range of $1 \leq K \leq 50$. It is found that the performance of KNN model for all the data combinations remain unchanged when the K value exceeds a value of 40 (in Figs. 5, 6, and 7). This suggests that the range of K in this study should fall within the range of $1 \leq K \leq 40$ as further addition of K value is redundant. Similar to the performance of BPCA, the same imputation performance is achieved at different value of K. This is likely caused by the same issue as explained earlier on.

Generally, the performance of the BPCA and KNN models are quite similar in terms of all the evaluations performed in this study. This is because majority of the results tabulated in Table 2 are close with each other at their respective percentage of missing entries. The imputation performance is logical as it becomes worse when the percentage of missing entries

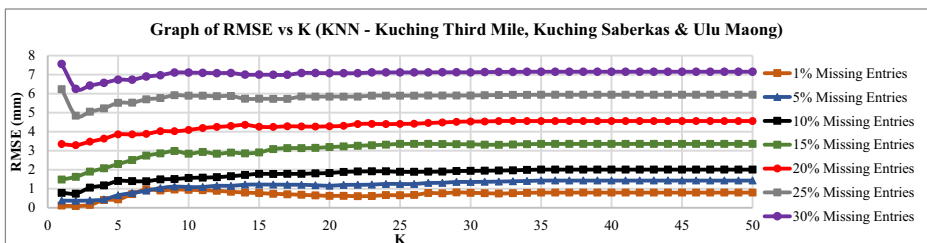


Fig. 6 Graph of RMSE vs K (KNN – Kuching Third Mile, Kuching Saberkas & Ulu Maong)

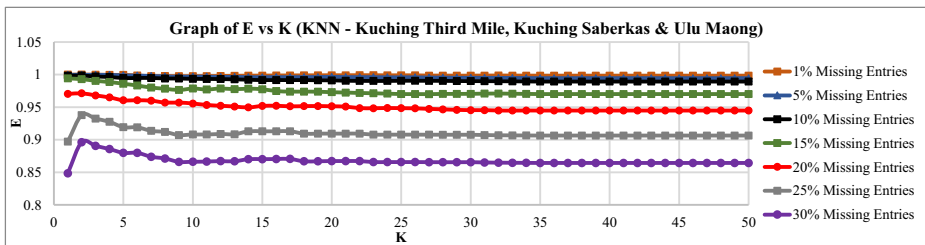


Fig. 7 Graph of E vs K (KNN – Kuching Third Mile, Kuching Saberkas & Ulu Maong)

increases. By referring to Table 2, the tabulated B_s values are not far from 1. This means that a slight overestimation ($B_s > 1$) and underestimation ($B_s < 1$) of data do occur on both KNN and BPCA. The accuracy of both models does improve when more data is provided for performing the imputation. The imputation performance achieves the lowest when only Kuching Third Mile's data is being utilised for executing the imputation. The combination of "Kuching Third Mile, Kuching Saberkas & Ulu Maong" performed the best for both KNN and BPCA model. This phenomenon suggests that further addition of data can be redundant as the performance dropped upon further addition of the rainfall data from Kuching Airport Station. For the data combination of two rainfall stations, the combination of "Kuching Third Mile & Kuching Saberkas" outperforms the rest for both imputation models. This might be due to the fact that the Kuching Saberkas Station is the nearest to Kuching Third Mile Station. The correlation between the rainfall data of the two stations are also the highest. Highest correlation between the datasets simply means that the possibilities of similar rainfall pattern are the highest. Thus, it will result in better imputation performance. This effect is also significant as the imputation performance drops when using rainfall data of the stations located further away from Kuching Third Mile.

The superiorities of both KNN and BPCA vary at different missing entries. From Table 2, it shows that 90% of the results from KNN at the missing entries of 1–20% are better than BPCA. On the other hand, 66% of the results from BPCA at the missing entries of 25–30% are better than KNN. This means that the general performance of BPCA is only superior to KNN at the missing entries of 25% and above. In terms of conveniences, BPCA is better as the range of K values is well defined. This reduce the time required to identify the adoptable range of K values. As for KNN, the suitable range of K values is identified via trial and error method. But, the performance of KNN is more consistent as shown in Figs. 2, 3, 4, 5, 6, and 7. These findings suggest that the suitability of KNN and BPCA imputation model may vary at different situations or settings.

6 Conclusion

In this study, the performance of BPCA imputation model is reliable as it exhibits similar results as the KNN imputation model. The missing data entries, K value and number of reference data are found to be the parameters that will affect the imputation performance of KNN and BPCA. The results support the idea of using correlation and distance to select the rainfall data from the neighbouring rainfall stations to be added into the input dataset. Improvement of the imputation performance for both BPCA and KNN is evident upon the addition of reference data. The findings also suggest that the suitability of the application of

BPCA and KNN imputation models is dependent on the situation. BPCA is found to be superior to KNN only at larger missing entries. The proposed method is recommended to be executed in other study area and other data mining or machine learning based imputation model. By doing so, it can help to determine if the proposed method is a viable alternative to boost their imputation performance.

Acknowledgements The authors wish to thank the reviewers for their feedback. The feedback contributes ideas and insights to improve this paper. This research did not receive any specific grant or fund from funding agencies in the public, commercial or not-for-profit sectors.

Compliance with Ethical Standards

Conflict of Interest No potential conflict of interest was reported by the authors.

References

- Bennett N, Newham L, Croke B, Jakeman A (2007) Patching and disaccumulation of rainfall data for hydrological modelling. In: Int. Congress on Modelling and Simulation (MODSIM 2007), Modelling and Simulation Society of Australia and New Zealand Inc., New Zealand, p 2520–2526
- Chai SS, Keat Wong W, Luong Goh K (2017) Rainfall classification for flood prediction using meteorology data of Kuching, Sarawak, Malaysia: backpropagation vs radial basis function neural network. *International Journal of Environmental Science and Development (IJESD)* 8:385–388. <https://doi.org/10.18178/ijesd.2017.8.5.982>
- De Silva R, Dayawansa N, Ratnasiri M (2007) A comparison of methods used in estimating missing rainfall data. *J Agric Sci* 3:101–108
- Ekeu-wei IT (2018) Evaluation of hydrological data collection challenges and flood estimation uncertainties in Nigeria. *Environment and Natural Resources Research* 8:44
- Gill MK, Asefa T, Kaheil Y, McKee M (2007) Effect of missing data on performance of learning algorithms for hydrologic predictions: implications to an imputation technique. *Water Resour Res* 43:1–12. <https://doi.org/10.1029/2006WR005298>
- Jajarmizadeh M, Harun S, Kuok KK, Sabari NS (2015) Contribution of climate forecast system meteorological data for flow prediction. In: Singapore. ISFRAM 2014. Springer Singapore, p 89–98
- Kamaruzaman IF, Zin WZW, Ariff NM (2017) A comparison of method for treating missing daily rainfall data in peninsular Malaysia. *Malaysian Journal of Fundamental and Applied Sciences (MJFAS)* 13:375–380
- Kuok KK, Bessaih N (2007) Artificial neural networks (ANNS) for daily rainfall runoff modelling. *Journal-The Institution of Engineers, Malaysia* 68:31–42
- Kuok KK, Harun S, Shamsuddin S (2010) Particle swarm optimization feedforward neural network for modeling runoff. *International Journal of Environmental Science & Technology (IJEST)* 7:67–78
- Lee H, Kang K (2015) Interpolation of missing precipitation data using kernel estimations for hydrologic modeling. *Adv Meteorol* 2015:1–12. <https://doi.org/10.1155/2015/935868>
- Luk KC, Ball JE, Sharma A (2001) An application of artificial neural networks for rainfall forecasting. *Math Comput Model* 33:683–693
- Malek MA (2008) Rainfall data in-filling model with expectation maximization and artificial neural network. PhD, Universiti Teknologi Malaysia
- Malek MA, Shamsuddin SM, Harun S (2010) Restoration of hydrological data in the presence of missing data via Kohonen self organizing maps. In: Ramov B (ed) *New trends in technologies. InTech, Rijeka*, pp 223–243
- Oba S (2013) [BPCAFill.m] BPCA Missing Value Estimator for MATLAB. Kyoto University. <http://ishiilab.jp/member/oba/tools/BPCAFill.html>. Accessed 4 May 2018
- Oba S, Sato M-a, Takemasa I, Monden M, Matsubara K-i, Ishii S (2003) A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics* 19:2088–2096
- Oh-Wook K, Kwokleung C, Te-Won L (2003) Speech feature analysis using variational Bayesian PCA. *IEEE Signal Processing Letters* 10:137–140. <https://doi.org/10.1109/LSP.2003.810017>
- Pagano TC et al (2014) Challenges of Operational River forecasting. *J Hydrometeorol* 15:1692–1707. <https://doi.org/10.1175/JHM-D-13-0188.1>

- Sattari M-T, Rezazadeh-Joudi A, Kusiak A (2016) Assessment of different methods for estimation of missing data in precipitation studies. *Hydrol Res*:1–13
- Severson AK, Molaro CM, Braatz DR (2017) Principal component analysis of process datasets with missing values processes 5:1–18. <https://doi.org/10.3390/pr5030038>
- Shi F, Zhang D, Chen J, Karimi HR (2013) Missing value estimation for microarray data by Bayesian principal component analysis and iterative local least squares. *Math Probl Eng* 2013:5. <https://doi.org/10.1155/2013/162938>
- Subashini P, Krishnaveni M (2011) Imputation of missing data using Bayesian Principal Component Analysis on TEC ionospheric satellite dataset. In: 2011 24th Canadian Conference on Electrical and Computer Engineering (CCECE), 8–11 May 2011. p 001540–001543. <https://doi.org/10.1109/CCECE.2011.6030724>
- Wang G, Wang D, Yang J, Liu L (2016) Evaluation and correction of quantitative precipitation forecast by storm-scale NWP model in Jiangsu, China. *Adv Meteorol* 2016:1–13. <https://doi.org/10.1155/2016/8476720>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.