International Journal of Engineering & Technology, 7 (4.30) (2018) 5-10

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Identifying the Ideal Number Q −Components of the Bayesian Principal Component Analysis Model for Missing Daily Precipitation Data Treatment

Zun Liang Chuan[1*], Azlyna Senawi[1], Wan Nur Syahidah Wan Yusoff[1], Noriszura Ismail[2], Tan Lit Ken[3], Mu Wen Chuan[4*]

[1] Faculty of Industrial Sciences and Technology, Universiti Malaysia Pahang, Lebuhraya Tun Razak, 26300 Gambang Kuantan, Pahang DM, Malaysia

[2]School of Mathematical Sciences, Faculty Science and Technology, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor DE, Malaysia

[3] Malaysia-Japan International Institute of Technology (MJIIT), Universiti Teknologi Malaysia Kuala Lumpur, Jalan Sultan Yahya Petra (Jalan Semarak), 54100 Kuala Lumpur, Malaysia

[4] Faculty of Electrical Engineering, Universiti Teknologi Malaysia, 81310 UTM Skudai, Johor, Malaysia

**Corresponding author E-mail: chuanzunliang@ump.edu.my*

Abstract

The grassroots of the presence of missing precipitation data are due to the malfunction of instruments, error of recording and meteorological extremes. Consequently, an effective imputation algorithm is indeed much needed to provide a high quality complete time series in assessing the risk of occurrence of extreme precipitation tragedy. In order to overcome this issue, this study desired to investigate the effectiveness of various *Q*-components of the Bayesian Principal Component Analysis model associates with Variational Bayes algorithm (BPCA*Q*-VB) in missing daily precipitation data treatment, which the ideal number of *Q*-components is identified by using The Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) algorithm. The effectiveness of BPCA*Q*-VB algorithm in missing daily precipitation data treatment is evaluated by using four distinct precipitation time series, including two monitoring stations located in inland and coastal regions of Kuantan district, respectively. The analysis results rendered the BPCA5-VB is superior in missing daily precipitation data treatment for the coastal region time series compared to the single imputation algorithms proposed in previous studies. Contrarily, the single imputation algorithm is superior in missing daily precipitation data treatment for an inland region time series rather than the BPCA*Q*-VB algorithm.

Keywords: *Bayesian principal component analysis model; Data treatment; TOPSIS; Variational Bayes.*

1. Introduction

The East Coast Economic Region (ECER) of Malaysia is a unique mix industrial region, which plentiful endowment of nature and agricultural resources. These resources can be diversified Malaysia's economy with a vital manufacturing component of high technology industry, medical technology and pharmaceuticals. In addition, the natural and agricultural resources of ECER have been promoted and accelerated the development of the tourism sector, which is a fast tool to spread national economic development and poverty reduction. However, ECER is frequently encountered the risk of occurrence for extreme precipitation tragedies, such as drought, flood and landslide with associated massive damage and losses to the society and national economy when the monsoon season is prevailing. Although the risk assessment of extreme precipitation tragedies can be carried out by using historical precipitation time series, but the risk assessment is frequently flawed due to the lack of complete time series. Therefore, an effective imputation algorithm is indeed much needed in missing precipitation data treatment.

For the past decades, a number of imputation algorithms have been proposed, such as a normal ratio algorithm, inverse distance weighting algorithm and coefficient of weighting algorithm, where these conventional imputation algorithms are frequently applied in environmental sciences [1, 5, 9, 10]. Recently, Burhanuddin et al. [3] proposed a multiple imputation algorithm, which associate the conventional normal ratio algorithm and moving block bootstrapping. The analysis results rendered their proposed algorithm yielded a more reliable results compared to the single imputation algorithm in missing daily precipitation data treatment. Moreover, Burhanuddin et al. [4] also proposed an improved normal ratio algorithm by adapting geometric median as an alternative of the arithmetic mean, which invariant respect to outliers. The imputation algorithms proposed by Burhanuddin et al. [3, 4] are highly dependent on the homogeneous precipitation time series of neighbouring monitoring stations. Contrarily, Saeed et al. [9] proposed median algorithm, which the proposed single imputation algorithm is without depending on the homogeneous precipitation time series of neighbouring monitoring stations.

This study intended to develop another multiple imputation algorithm without depending on the homogeneous precipitation time series of neighbouring monitoring stations. The developed multi-

ple imputation algorithm associates the $Q-$components Bayesian Principal Component Analysis model and Variational Bayesian algorithm (BPCA*Q*-VB), where the ideal of *Q*-components of the Bayesian Principal Component Analysis (BPCA*Q*) model is identified by using the Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) algorithm. In order to pursue the main objective of this study, the rest of the paper is arranged as follows. Section 2 briefly the description of the four selected daily precipitation time series involved in this study. Section 3 presented the overview of the theoretical background, including BPCA*Q*-VB algorithm, performance indices and TOPSIS algorithm while Section 4 presented the analysis results. Finally, the concluding remark is rendered in Section 5.

2. Study areas

Kuantan River Basin located at the ECER is one of the vital tributaries, which irrigates the majority of the rural, urban agriculture and industrial areas of Kuantan District. This district is frequently exposed the risk of occurrence for extreme precipitation tragedies when the monsoon season is prevailing. In this study, daily precipitation time series of monitoring stations located in the inland and coastal regions in the Kuantan district [6] as depicted in Fig. 1 are selected to evaluate the effectiveness of the proposed multiple imputation algorithm, which this daily precipitation time series data are acquired from the Department of Irrigation and Drainage (DID) Malaysia. The daily precipitation time series of monitoring stations located in the Kuantan River Basin are selected due to the quality of time series for this district is frequently degraded by the missing data. The topography and descriptive statistical characteristics for the monitoring stations and its complete daily precipitation time series are presented in Table 1, which the missing data are simulated and withdraw from the complete time series. The period of precipitation time series selected is range 2010 to 2013, which coverage the timeframe of monsoon season.

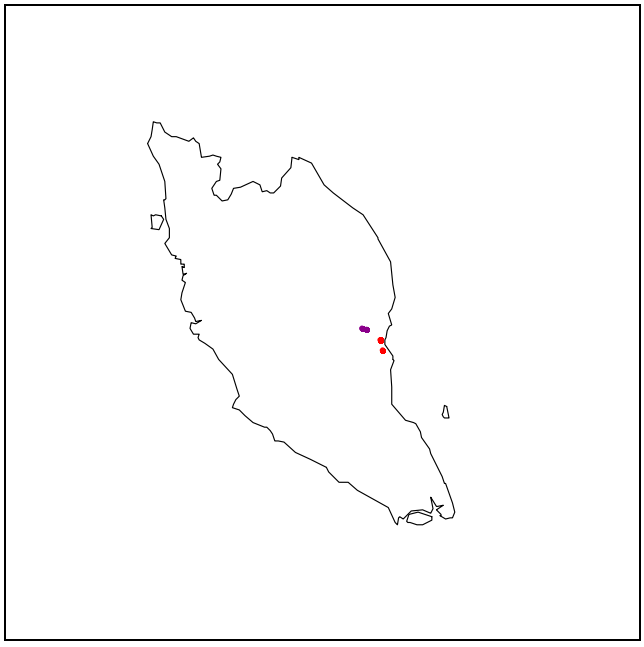


Fig.1: The location of the four selected monitoring precipitation stations in Kuantan District, which the magenta and red colour points represent the monitoring precipitation stations located in inland and coastal regions, respectively.

Table 1: The topography and descriptive statistical characteristics of the four selected monitoring precipitation stations in Kuantan District

Region (Station ID)	Inland (3930012)	Inland (3931013)	Coastal (3633104)	Coastal (3732020)
Station name	Sg. Lembing	Ldg. Nada	Kg. Bahru Penor	Paya Besar Kuantan
Latitude	03°55'00"	03°54'30"	03°37'50"	03°46'20"
Longitude	103°02'10"	103°06'20"	103°18'55"	103°16'50"
Period	11/06/2012-06/10/2013	06/09/2010-26/05/2013	06/06/2011-01/09/2013	08/10/2012-03/11/2013
Number of days	483	994	819	392
Mean (mm)	8.0683	8.9085	7.3425	7.1977
Coefficient of variation	1.9219	1.9991	2.9753	2.7827

3. Theoretical Background

3.1. Bayesian principal component analysis model

Tipping and Bishop [11] introduced the probabilistic principal component (PPCA) model, which associates the probabilistic formulation with regular principal component analysis model. The intention of PPCA takes into account the probabilistic formulation is to facilitate the estimation of latent variable by using expectation-maximization algorithm, where the missing daily precipitation data behave towards as an addition. In general, PPCA model which is relative to the factor analysis model can be expressed as

$$\mathbf{Y}_k = \boldsymbol{\omega}\mathbf{L}_k + \boldsymbol{\theta} + \boldsymbol{\xi}_k \tag{1}$$

where $\mathbf{Y}_k$ is the $P-$dimensional of precipitation variable vectors, $\mathbf{L}_k$ is the $Q-$dimensional of latent variable vectors, $\boldsymbol{\omega}$ is $P \times Q-$dimensional projection matrix with pre-defined Q, $\boldsymbol{\theta}$ is the mean vectors of $\mathbf{Y}_k$, $\boldsymbol{\xi}_k$ is the residual errors and $k = 1, 2, \ldots, n.$ In addition, $\mathbf{L}_k$ and $\boldsymbol{\xi}_k$ are assumed to adequate to a Gaussian distribution, namely $\mathbf{L}_k \sim N(0, \mathbf{I})$ and $\boldsymbol{\xi}_k \sim N(0, \gamma^2\mathbf{I})$, where γ^2 is the variance and $\mathbf{I}$ is the identity matrix. By integrating the prior probability distribution function of parameters, $\Pr(\boldsymbol{\omega}, \boldsymbol{\theta}, \gamma^2)$ and the conditional probability distribution function of PPCA model, $\Pr(\mathbf{Y}_k, \mathbf{L}_k | \boldsymbol{\omega}, \boldsymbol{\theta}, \gamma^2)$, therefore the posterior probability distribution of $\mathbf{L}_k$ and the parameters of $\boldsymbol{\omega}$, $\boldsymbol{\theta}$ and γ^2 is resulted, such that

$$\Pr(\mathbf{L}_k, \boldsymbol{\omega}, \boldsymbol{\theta}, \gamma^2 | \mathbf{Y}_k) \propto \Pr(\mathbf{Y}_k, \mathbf{L}_k | \boldsymbol{\omega}, \boldsymbol{\theta}, \gamma^2) \Pr(\boldsymbol{\omega}, \boldsymbol{\theta}, \gamma^2) \tag{2}$$

The assumed conjugate priors for the parameters of $\boldsymbol{\theta}$, γ^2 and the hierarchical prior of $\boldsymbol{\omega}$ in (2), which corresponds to a non-informative prior is given as

$$\Pr(\boldsymbol{\omega}, \boldsymbol{\theta}, \gamma^2 | \boldsymbol{\lambda}) = \Pr(\boldsymbol{\theta} | \gamma^2) \Pr(\gamma^2) \prod_{m=1}^{Q} \Pr(\boldsymbol{\omega}_m | \boldsymbol{\lambda}_m) \tag{3}$$

where $\boldsymbol{\theta} | \gamma^2 \sim N\left(\mathbf{0}, \left(10^{-10}\gamma^2\right)^{-1}\mathbf{I}\right)$, $\gamma^2 \sim \text{Gamma}\left(1, 10^{-10}\right)$ and $\boldsymbol{\omega}_m | \gamma^2, \lambda_m \sim N\left(\mathbf{0}, \left(\gamma^2\boldsymbol{\lambda}_m\right)^{-1}\mathbf{I}\right)$, while the hierarchical prior $\boldsymbol{\omega}$ is parameterized by a hyper-parameter, $\boldsymbol{\lambda}_m \in \square^{Q}$. The hyper-parameter $\boldsymbol{\lambda}_m$ plays a crucial role in dominating the effectiveness of dimensionality of the latent variables [2].

3.2. Variational Bayes algorithm

In general, the posterior of missing daily precipitation data, $\mathbf{Y}^{miss}$, with the presence of true parameters information, including $\boldsymbol{\omega}_T$, $\boldsymbol{\theta}_T$ and γ^2 can be expressed as

$$\pi\left(\mathbf{Y}^{miss}\right)=\Pr\left(\mathbf{Y}^{miss}\middle|\mathbf{Y}^{obs},\ \boldsymbol{\omega}_T,\ \boldsymbol{\theta}_T,\ \gamma_T^2\right) \tag{4}$$

which acquire with marginalizing (2) respect to $\mathbf{Y}^{obs}$. In practice, there is a mere presence of the information of parameter posterior, $\pi\left(\boldsymbol{\omega},\ \boldsymbol{\theta},\ \gamma^2\right)$, rather than the true parameters as rendered in Section 3.1, therefore (4) is required to re-written as

$$\pi\left(\mathbf{Y}^{miss}\right)=\iiint\left\{\pi\left(\boldsymbol{\omega},\ \boldsymbol{\theta},\ \gamma^2\right)\Pr\left(\mathbf{Y}^{miss}\middle|\mathbf{Y}^{obs},\ \boldsymbol{\omega}_T,\ \boldsymbol{\theta}_T,\ \gamma_T^2\right)\right\}\,d\boldsymbol{\omega}\,d\boldsymbol{\theta}\,d\gamma^2 \tag{5}$$

which corresponds to the Bayesian Principal Component regression. Based on (5), it is essential to acquire $\pi\left(\mathbf{Y}^{miss}\right)$ and $\pi\left(\boldsymbol{\omega},\ \boldsymbol{\theta},\ \gamma^2\right)$ by using Variational Bayesian algorithm, where the posterior of $\pi\left(\mathbf{Y}^{miss}\right)$ is initiated by imputing each of the missing daily precipitation data to instance-wise average. Afterwards, a sequence of process, including estimates $\pi\left(\boldsymbol{\omega},\ \boldsymbol{\theta},\ \gamma^2\right)$ using $\mathbf{Y}^{obs}$ and current $\pi\left(\mathbf{Y}^{miss}\right)$, estimates $\pi\left(\mathbf{Y}^{miss}\right)$ using the current $\pi\left(\boldsymbol{\omega},\ \boldsymbol{\theta},\ \gamma^2\right)$ and updates the $\boldsymbol{\lambda}$ using the current $\pi\left(\mathbf{Y}^{miss}\right)$ and $\pi\left(\boldsymbol{\omega},\ \boldsymbol{\theta},\ \gamma^2\right)$ is iterated to reach the predefined convergence criteria. Subsequently, the missing daily precipitation data is imputed to expectation with respect to the estimated posterior distribution [8], such that

$$\hat{\mathbf{Y}}^{miss}=\int\mathbf{Y}^{miss}\pi\left(\mathbf{Y}^{miss}\right)d\mathbf{Y}^{miss} \tag{6}$$

3.3. Performance indices

Based on previous studies, several performance indices such as the correlation coefficient, root index of agreement, mean square error, mean relative error and normalized root mean square error are widely used to evaluate the imputation algorithms [9]. In this study, three dimensionless performance indices, including Modified Willmott's Index of Agreement (MWIA), Modified Agreement Coefficient (MAC) and Normalized Root Mean Square Error (NRMSE) are used as the attributes for TOPSIS algorithm in identifying the ideal number of $Q-$components of the BPCA model. The values of these three performance indices close in zero indicate better fit, given as

$$\text{MWAI}=\frac{\sum_{l=1}^{P}\sum_{k=1}^{n}\left(\mathbf{Y}_{kl}^{\text{imp}}-\mathbf{Y}_{kl}^{\text{act}}\right)^2}{\sum_{l=1}^{P}\sum_{k=1}^{n}\left(\delta_1+\delta_2\right)^2} \tag{7}$$

$$\text{MAC}=\frac{\sum_{l=1}^{P}\sum_{k=1}^{n}\left(\mathbf{Y}_{kl}^{\text{imp}}-\mathbf{Y}_{kl}^{\text{act}}\right)^2}{\sum_{l=1}^{P}\sum_{k=1}^{n}\left[\left(\delta_3+\delta_2\right)\left(\delta_3+\delta_4\right)\right]} \tag{8}$$

$$\text{NRMSE}=\sqrt{\left(1-\frac{1}{nP}\right)\frac{\sum_{l=1}^{P}\sum_{k=1}^{n}\left(\mathbf{Y}_{kl}^{\text{imp}}-\mathbf{Y}_{kl}^{\text{act}}\right)^2}{\sum_{l=1}^{P}\sum_{j=1}^{n}\left(\mathbf{Y}_{kl}^{\text{act}}-\bar{\mathbf{Y}}^{\text{act}}\right)^2}} \tag{9}$$

Where $\delta_1=\left|\mathbf{Y}_{kl}^{\text{imp}}-\bar{\mathbf{Y}}^{\text{act}}\right|$, $\delta_2=\left|\mathbf{Y}_{kl}^{\text{act}}-\bar{\mathbf{Y}}^{\text{act}}\right|$, $\delta_3=\left|\bar{\mathbf{Y}}^{\text{act}}-\bar{\mathbf{Y}}^{\text{imp}}\right|$, $\delta_4=\left|\mathbf{Y}_{kl}^{\text{imp}}-\bar{\mathbf{Y}}^{\text{imp}}\right|$, and $\mathbf{Y}_{kl}^{\text{imp}}$ and $\mathbf{Y}_{kl}^{\text{act}}$ represents imputed and observed daily precipitation, respectively. Meanwhile, $\bar{\mathbf{Y}}^{\text{imp}}$ and $\bar{\mathbf{Y}}^{\text{act}}$ represents the average of imputed and observed daily precipitation, respectively.

3.4. TOPSIS algorithm

TOPSIS introduced by Hwang and Yoon [7] is suggested in this study as an alternative algorithm to identify the ideal number of $Q-$components of the BPCA. Suppose that $\mathbf{A}=\left[\alpha_{gh}\right]_{GH}$; $g,(h)=1,\ 2,\ \ldots,\ G,(H)$ represents a normalized matrix with G attributes and H alternatives. An alternative is designated as the best alternative, which rendered the highest value of relative closeness (κ), given as

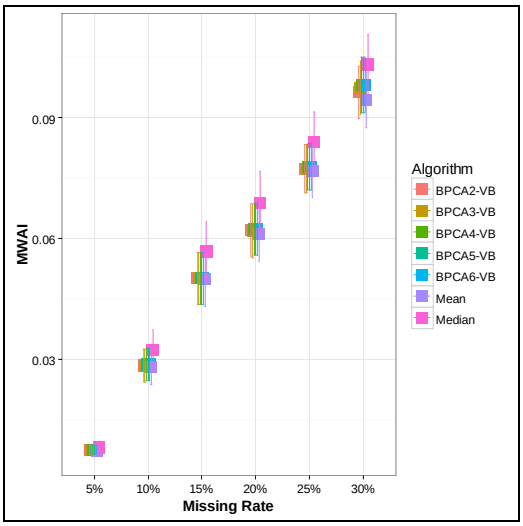
$$\kappa=\arg\left(\max_h\left(\frac{\sqrt{\sum_{g=1}^{G}\left(\tau_{gh}-\tau_h^-\right)^2}}{\sqrt{\sum_{g=1}^{G}\left(\tau_{gh}-\tau_h^-\right)^2}+\sqrt{\sum_{g=1}^{G}\left(\tau_{gh}-\tau_h^+\right)^2}}\right)\right) \tag{10}$$

Where $\tau_h^-=\min_h\left(\tau_{gh}\right)$, $\tau_h^+=\max_h\left(\tau_{gh}\right)$, $\tau_{gh}=\phi_h\alpha_{gh}$ represents the weighted normalized observations with weighted function, $\phi_h=\dfrac{\sum_{g=1}^{G}\frac{\left(\alpha_{gh}-\bar{\alpha}_h\right)^2}{G-1}}{\sum_{h=1}^{H}\left(\sum_{g=1}^{G}\frac{\left(\alpha_{gh}-\bar{\alpha}_h\right)^2}{G-1}\right)}$ on condition $\sum_{h=1}^{H}\phi_h=1$ and $0\le\kappa\le1$.

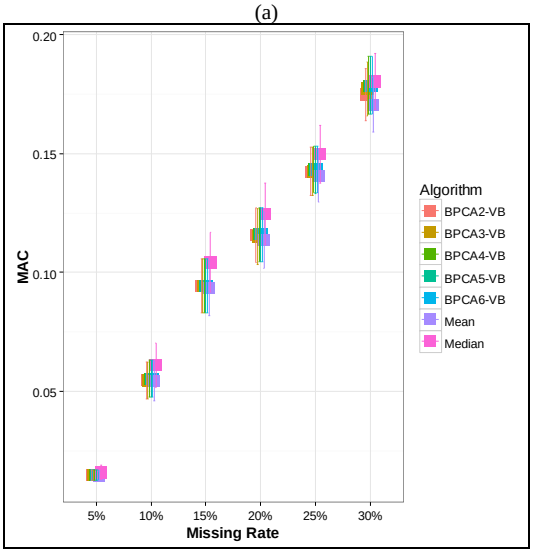
4. Analysis Results

In order to verify the effectiveness of the BPCA*Q*-VB algorithms, the comparison of efficiency between the BPCA*Q*-VB algorithms and the existing single imputation algorithms in missing daily precipitation data treatment such as mean and median algorithms [9] appraise at missing rates of 5%, 10%, 15%, 20%, 25% and 30%. Figs. 2 and 3 depicted the pictorial of performance comparison among the imputation algorithms in data treatment for the monitoring Stations 3930012 and 3931013 located in inland region corresponding to MWAI, MAC and NRMSE, while Figs. 4 and 5 depicted the pictorial of performance comparison for the monitoring Stations 3633104 and 3732020 located in coastal region. Based on Figs. 2-5, it can be observed that the performance uncertainty among the imputation algorithms are magnified respect to the amplification of missing rates, which the uncertainties summarized from 10 replications respect to the missing rates and imputation algorithms, respectively. Moreover, it also can find that the efficiency between the single and multiple imputation algorithms in missing daily precipitation data treatment are approximately equivalent irrespective to missing rates. However, the Multivariate Analysis of Variance (MANOVA) based on Roy' greatest root test rendered that there is a significant difference between the imputation algorithms in treating missing daily precipitation time series acquired from the both monitoring stations located in inland region, except for missing rate at 5%. Furthermore, this multivariate statistical test also rendered there is a significant difference between the imputation algorithms in treating missing daily precipitation time series acquired from the both monitoring stations located in coastal region. However, these significant results are invalid at missing rates as high as 10% and

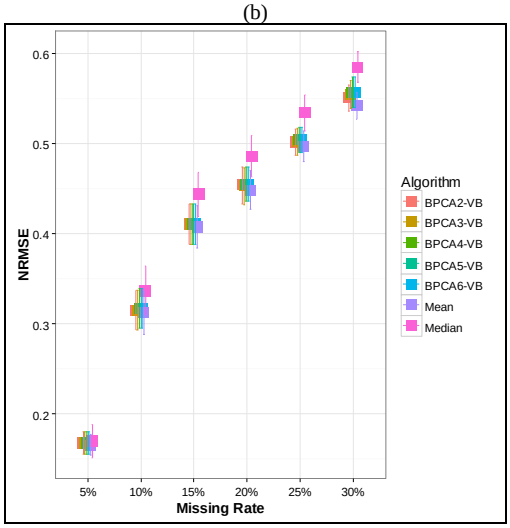
15% for the monitoring Stations 3633104 and 3732020, respectively.



(a)

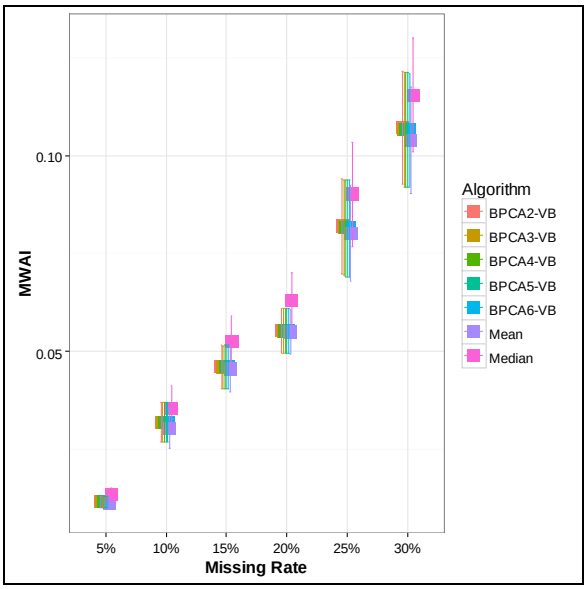


(b)

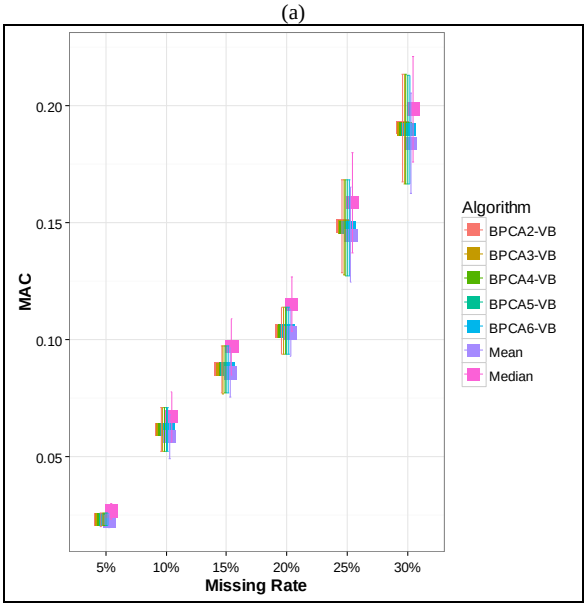


(c)

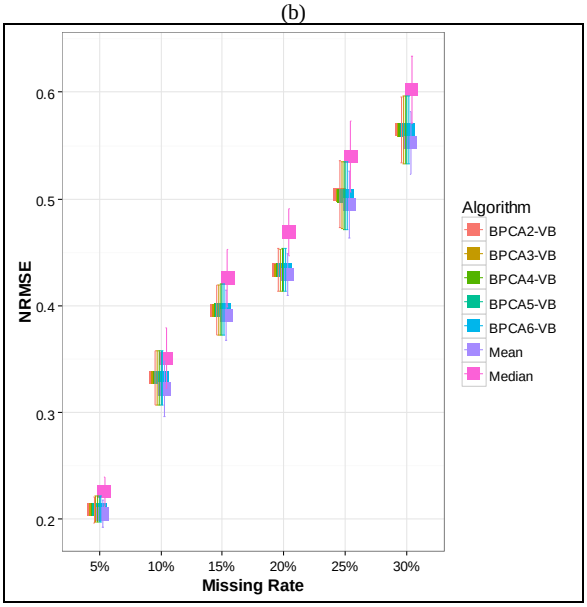
Fig 2: The pictorial of performance comparison of the single and multiple imputation algorithms in missing daily precipitation data treatment for the monitoring Station 3930012 located in inland region corresponding to (a) MWAI; (b) MAC; (c) NRMSE.



(a)

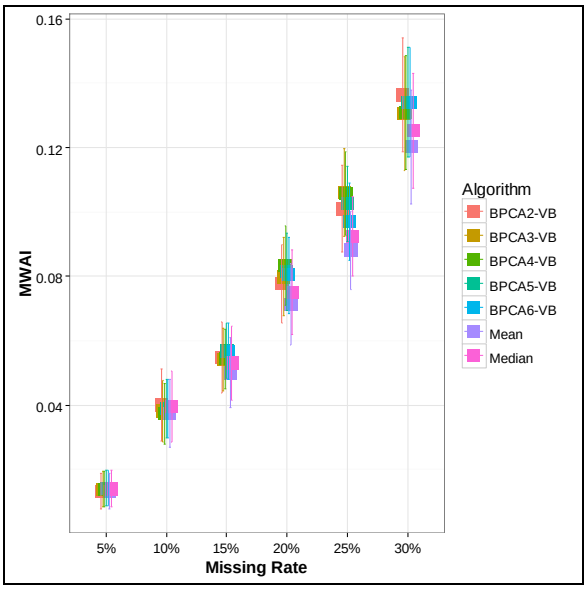


(b)

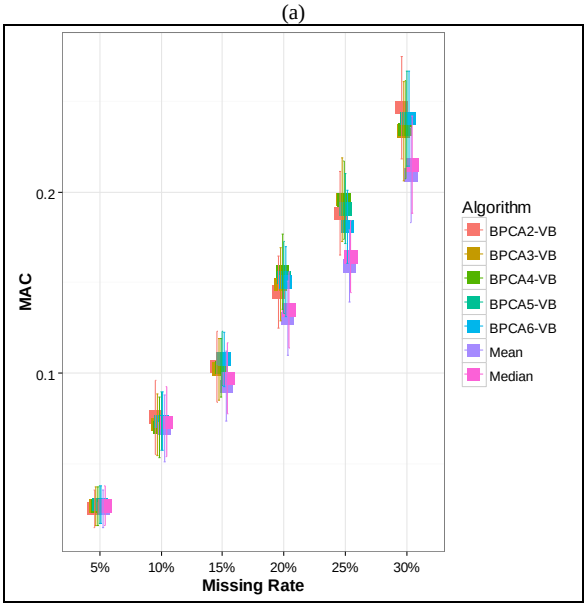


(c)

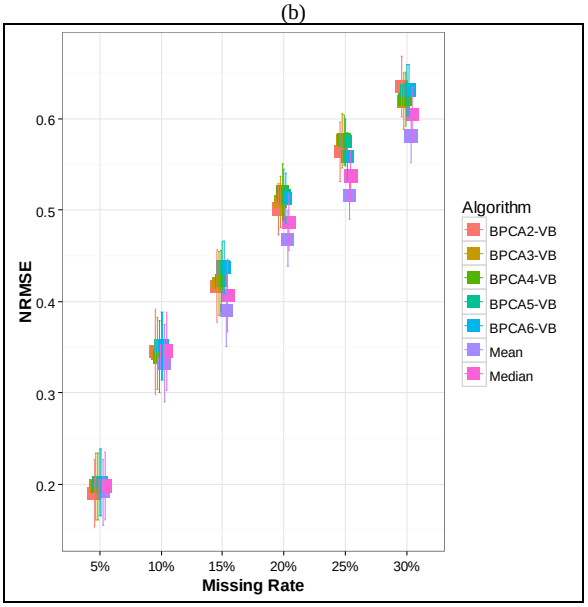
Fig 3: The pictorial of performance comparison of the single and multiple imputation algorithms in missing daily precipitation data treatment for the monitoring Station 3931013 located in inland region corresponding to (a) MWAI; (b) MAC; (c) NRMSE.



(a)

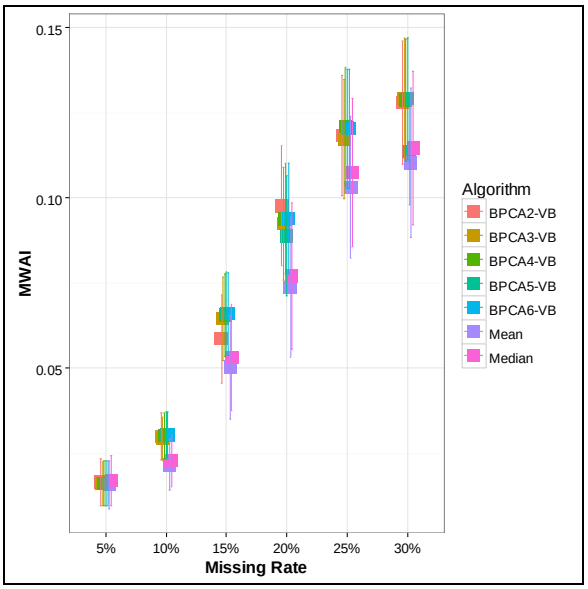


(b)

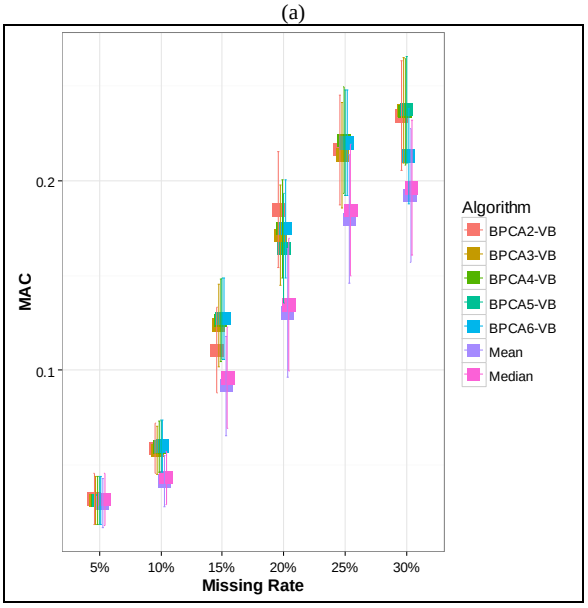


(c)

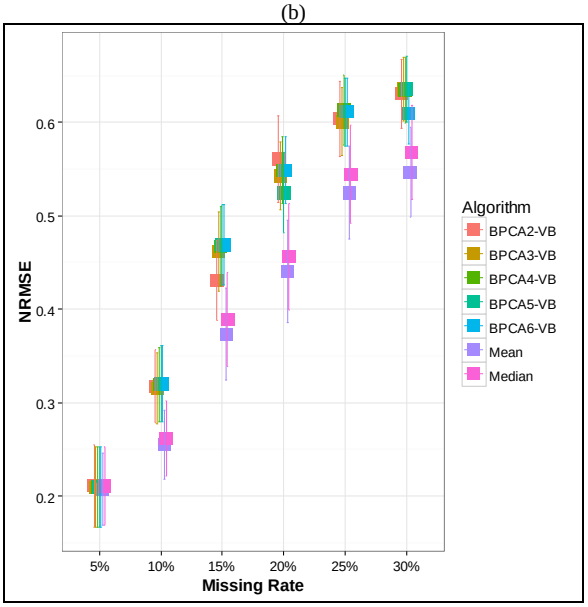
Fig. 4: The pictorial of performance comparison of the single and multiple imputation algorithms in missing daily precipitation data treatment for the monitoring Station 3633104 located in coastal region corresponding to (a) MWAI; (b) MAC; (c) NRMSE.



(a)



(b)



(c)

Fig. 5: The pictorial of performance comparison of the single and multiple imputation algorithms in missing daily precipitation data treatment for the monitoring Station 3732020 located in coastal region corresponding to (a) MWAI; (b) MAC; (c) NRMSE.

In identifying the superiority among the imputation algorithms for such complex condition, this study used TOPSIS algorithm to rank the superiority among the imputation algorithms as rendered in Table 2, which the average of three performance indices, namely MWAI, MAC and NRMSE are used as the attributes for TOPSIS algorithm. The main reason these three performance indices are used as the attributes for TOPSIS algorithm as these performance indices rendered a significant difference results for Roy's greatest root test. Table 2 rendered that the single imputation algorithm, namely median algorithm is invariably more superior compared to the mean algorithm and *BPCAQ*-VB algorithms for the inland region time series. In contrary, the BPCA5-VB algorithm is more superior compared to the single imputation data algorithms proposed in literature, which ranked in the best three irrespective of missing rates. Due to the spatial and temporal characteristics between the daily precipitation time series data acquired from the monitoring stations located in inland and coastal region are varied [6], therefore a vary imputation algorithm is indeed much needed in treating the missing daily precipitation time series.

Table 2: The ranked of performance of the BPCA*Q*-VB algorithm in missing daily precipitation data treatment

Region	Station ID		Missing rates (%)					
		Algorithm	**5**	**10**	**15**	**20**	**25**	**30**
Inland	3930012	Mean	7	7	7	7	7	7
		Median	1	1	1	1	1	1
		BPCA2-VB	4	5	4	5	6	6
		BPCA3-VB	4	6	4	6	5	5
		BPCA4-VB	4	3.5	4	2	3	3
		BPCA5-VB	4	2	4	3	4	4
		BPCA6-VB	4	3.5	4	4	2	2
	3931013	Mean	7	7	7	7	7	7
		Median	1	1	1	1	1	1
		BPCA2-VB	6	2	5	2	2	2
		BPCA3-VB	3.5	6	6	6	3	4
		BPCA4-VB	3.5	5	4	3	6	3
		BPCA5-VB	3.5	3	2.5	4.5	4.5	5
		BPCA6-VB	3.5	4	2.5	4.5	4.5	6
Coastal	3633104	Mean	6	7	7	7	7	7
		Median	3	2	6	6	6	6
		BPCA2-VB	7	1	3	5	4	1
		BPCA3-VB	4	5	5	4	1	5
		BPCA4-VB	5	6	4	1	2	4
		BPCA5-VB	2	4	1	2	3	3
		BPCA6-VB	1	3	2	3	5	2
	3732020	Mean	7	7	7	7	7	7
		Median	1	6	6	6	6	6
		BPCA2-VB	2	4	5	1	4	4
		BPCA3-VB	6	5	4	4	5	2
		BPCA4-VB	4	3	3	3	1	3
		BPCA5-VB	3	2	1	5	2	1
		BPCA6-VB	5	1	2	2	3	5

5. Conclusion

This study rendered a comparison of the BPCA*Q*-VB algorithms in missing daily precipitation data treatment. These BPCA*Q*-VB algorithms are evaluated using four selected daily precipitation time series monitoring stations respectively located in inland and coastal regions of Kuantan district, which the ideal number of Q – components BPCA model is determined using TOPSIS algorithm. The analysis results rendered that the BPCA5-VB algorithm is more superior in missing daily precipitation data treatment for the coastal regions time series compared to the single imputation algorithms in the literature. In the inverse, the median algorithm is more superior in missing daily precipitation data treatment for inland regions time series rather than the BPCA*Q*-VB algorithms. In other words, a vary imputation algorithm are required in treating the missing daily precipitation time series acquired from the monitoring stations located in inland and coastal regions as the spatial and temporal characteristics of these regions is varied.

Acknowledgement

The authors would like to express unreserved appreciation to the Department of Irrigation and Drainage (DID) Malaysia for providing the daily precipitation data of this study. A word of appreciating also acknowledges to the Universiti Malaysia Pahang (UMP) for providing the flagship research grant RDU150393 and the internal research grant RDU1703184.

References

[1] Ahrens B (2006), Distance in spatial interpolation of daily rain gauge data. Hydrology and Earth System Sciences 10, 197-208.

[2] Bishop CM (1999), Bayesian PCA. *Proceedings of the Conference on Advances in Neural Information Processing Systems* 11, 382-388.

[3] Burhanuddin SNZA, Deni SM & Ramli NM (2017), Normal ratio in multiple based on bootstrapped sample for rainfall data with missingness. International Journal of GEOMATE 13(36), 131-137.

[4] Burhanuddin SNZA, Deni SM & Ramli NM (2017), Imputation of missing rainfall data using revised normal ratio method. *Advanced Science Letters* 23(11), 10981-10985.

[5] Chen FW & Liu CW (2012), Estimation of the spatial rainfall distribution using inverse distance weighting (IDW) in the middle of Taiwan. *Paddy and Water Environment* 10(3), 209-222.

[6] Chuan ZL, Ismail N, Shinyie WL, Ken TL, Fam SF, Senawi A & Yusoff WNSW (2018), The efficiency of average linkage hierarchical clustering algorithm associated multi-scale bootstrap resampling in identifying homogeneous precipitation catchments. *IOP Conference Series: Materials Science and Engineering* 342, 012070, doi:10.1088/1757-899X/342/1/012070.

[7] Hwang CL & Yoon K (1981), Multiple attribute decision making methods and applications a state art-of-the-art survey. Springer-Verlag, Heidelberg.

[8] Oba S, Sato M, Takemasa I, Monden M, Matsubara K & Ishii S (2003), A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics* 19(16), 2088-2096.

[9] Saeed GAA, Chuan ZL, Zakaria R, Yusoff WNSW & Salleh MZ (2016), Determination of the best single imputation algorithm for missing rainfall data treatment. *Journal of Quality Measurement and Analysis* 12(1-2), 79-87.

[10] Teegavarapu RSV & Chandramouli V (2005), Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records. *Journal of Hydrology* 312(1-4), 191-206.

[11] Tipping ME & Bishop CM (1997), Mixtures of principal component analysers. *Proceedings of the 5th International Conference on Artificial Neural Networks*, 13-18, doi: 10/1049/cp:19970694.