Controlled Sampling Approach in Improving Multiple Imputation for Missing Seasonal Rainfall Data

Siti Nur Zahrah Amin Burhanuddin (misssnzbab@gmail.com)
Universiti Teknologi MARA https://orcid.org/0000-0002-7136-494X
Sayang Mohd Deni
Universiti Teknologi MARA
Norshahida Shaadan
Universiti Teknologi MARA

Research Article

Keywords: Imputation, Missing data, Seasonal rainfall, Controlled sampling, Block bootstrap

Posted Date: July 6th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-679692/v1

Controlled sampling approach in improving multiple imputation for missing seasonal rainfall data

*Siti Nur Zahrah Amin Burhanuddin[a], Sayang Mohd Deni[b], Norshahida Shaadan[c]

[a,b,c]Centre for Statistics and Decision Science Studies, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia.

[a]misssnzbab@gmail.com, [b]sayang@tmsk.uitm.edu.my, [c]shahida@tmsk.uitm.edu.my

ORCID: [a]0000-0002-7136-494X

Abstract

Missingness in rainfall data is one of the well-known and challenging issues faced by meteorologists and researchers from all over the world. The problem would affect the quality of the data which is very important in representing the actual meteorological characteristics of a particular location. Therefore, the missing data should be properly treated in order to provide good quality dataset for the public domain. In furtherance of ensuring the accuracy of imputed missing data, the original structure of the rainfall data series must be specifically preserved when the data are having seasonal patterns. Most of the environmental datasets are generally characterized by outliers and seasonal patterns. These characteristics have certainly affected the performance of missing data imputation methods. The problem of missing data can be treated, but a specific structured approach must be employed when involving dataset that contains outliers and seasonal patterns. This study has highlighted and discussed the structured and comprehensive procedures on how to tackle the problem of missing data by emphasizing on controlled sampling approach for their implementation. The missing values were estimated by using multiple imputation based on block bootstrap approach associated with normal ratio methods compared to the conventional sampling (i.e. general bootstrap approach). The analysis and experimentation are illustrated using several datasets obtained for several locations in Peninsular Malaysia. The block bootstrap approach has revealed its advantage of preserving time series structure in its process and successfully improved the estimates of missing rainfall data imputation.

Keywords: Imputation; Missing data; Seasonal rainfall; Controlled sampling; Block bootstrap

Declarations

Funding (information that explains whether and by whom the research was supported):

Malaysian Fundamental Research Grant (FRGS/1/2014/ST06/UITM/02/6)

Conflicts of interest/Competing interests (include appropriate disclosures):

Not applicable

Availability of data and material (data transparency):

Not applicable

Code availability (software application or custom code):

Not applicable

Authors' contributions (optional: please review the submission guidelines from the journal whether statements are mandatory)

1. Introduction

Quality environmental data are highly necessary for performing efficient and effective hydrological and meteorological analyses. Completeness in the data is one of the most important factors to be considered in ensuring the quality of any data. However, the occurrence of missing values in environmental data, especially rainfall data is inevitable (Di Piazza et al., 2011; Miró et al., 2017) and has been a long-standing problem that needed to be addressed (Ramos-Calzado et al., 2008). There are various reasons that contribute to the missingness of rainfall data. Factors such as meteorological extremes, observational recording errors, relocation of stations, malfunctioning of instruments, and human errors during the measurement of rainfall data may affect the consistency in the rainfall records.

Missing data is considered one of the barriers in the production of accurate results in rainfall analysis. Missing data in the process of analysis may result in biasness and produce misleading results (Ramos-Calzado et al., 2008; Kalteh and Hjorth, 2009). As a result, inaccurate information will be passed on to the hydrological management and development activities such as water resource management, irrigation scheduling, and flood forecasting and prediction. Therefore, this shortcoming has motivated researchers to identify suitable imputation approaches in treating missing data.

A common practice when dealing with the missing values is to simply discard them from the rainfall records in the data preprocessing stage (Acuña and Rodriguez, 2004). However, it is not recommended because significant information could be lost when incomplete data is neglected. This will subsequently cause a loss of efficiency and produce severely biased estimates (Donders et al., 2006; Aydilek and Arslan, 2013). Therefore, to ensure that an effective rainfall analysis is performed, it is essential to impute the incomplete rainfall dataset by using appropriate methods.

It is important to understand the reasons and mechanisms behind the missing data for the observed dataset before introducing the missing values in the dataset. Generally, the missing data is introduced based on three fundamental mechanisms, i.e. i. missing completely at random (MCAR), ii. missing at random (MAR), and iii. missing not at random (MNAR) (Twumasi-Ankrah, Odoi, Pels, & Gyamfi, 2019). The missing mechanism is categorized as MAR if the probability of having missing data in a variable depends on observed data, however, it is known as MNAR if the missing data are related to unobserved data. On the other hand, the mechanism is called MCAR if the missing data is independent of both unobserved and observed data. These mechanisms are essentially related to the underlying reason why the data is missing and explain the relationships between measured variables and the probability of missing data (Jahan et al., 2018).

Meanwhile, it is necessary to identify appropriate treatment approaches for missing values based on the mechanism (or pattern) of missing data (Hanaish et al., 2013; Jahan et al., 2018). According to Kalteh and Hjorth (2009) and Yozgatligil et al. (2013), MCAR and MAR are called ignorable response mechanisms because the reasons for the missing data can be ignored during the analysis. Model-based methods require this assumption, and it is reasonable to assume that missing hydrological data are MCAR or MAR (Kalteh and Hjorth, 2009). Meanwhile, the NMAR is a nonignorable pattern of missingness.

Ingsrisawang and Potawee (2012) and Jahan et al. (2018) assumed that the existence of missing rainfall values in their studies were under the MAR mechanism. On the other hand, Hanaish et al. (2013) claimed that the missing values in Malaysian rainfall data is assumed to be in MCAR category, which means that the cause of missing data is either not related to the values that are missing or is not dependent on the observed data, totally random missingness (Fielding et al., 2009). In a similar study by Farhangfar et al. (2004), the missing values under MCAR mechanism was introduced in their study. From the reviewed literature, it can be concluded that most of the missing values in hydrological data are categorised under MAR mechanism because the presence of missing data only depends on the observed data rather than the other variables (e.g. the latitude, the longitude, or the elevation of the rainfall stations) (Chuan et al., 2019). Since several missing data estimation methods are to be used in this study, including multiple imputation approach, the missing data mechanism can be assumed to be missing at random (MAR).

Various approaches have been suggested and developed by previous studies for estimating the missing rainfall values. However, multiple imputation (MI) is the most popular treatment among the other

approaches used in the last few decades due to its high capability in handling the missing data. Several studied have found this approach to be a powerful tool in most of the situations (Yendra and Jemain, 2013; Yozgatligil et al., 2013; De Carvalho et al., 2017; Ekeu-Wei et al., 2018; Chen et al., 2019). Multiple imputation approach was introduced by Rubin (1987) to overcome the limitations of single imputation (SI) approach. This approach provides a more accurate and robust estimation results by considering the uncertainty of missing values and variability of the imputed values in the imputation process (Lo Presti et al., 2010). Multiple imputation approach treats the missing data in a way to produce a valid statistical inference instead of estimating the missing values as close as possible to the observed ones (Enders, 2010).

Enders (2010) organized the MI approach into three basic phases, i.e. the imputation phase, analysis phase, and pooling phase. Generally, imputation is the most important phase in the MI approach. It involved the imputation of the missing values as in the process of SI approach. The process of imputation is implemented multiple times (*m* times); in many cases, three to five times (Rubin, 1987) on different datasets. The imputed values are different for each dataset due to the inherent randomness in the algorithm itself. The complete imputed datasets are then analysed by standard methods which result in *m* sets of parameter estimates. Finally, these estimates are pooled to produce a single point estimate.

Yozgatligil et al. (2013) and Jahan et al. (2018) adopted the Monte Carlo Markov Chain based on the expectation–maximization (EM-MCMC) method in the application of MI approach in their studies. The method considers missingness as a proportional information of the sample to estimate the parameter of interest through conditional expectations (Yozgatligil et al., 2013). The MCMC estimates the imputation and parameters iteratively with the initial estimates provided by the EM algorithm. The EM-MCMC method has been found to produce robust estimation results, nevertheless, it was unfortunately proven by Yozgatligil et al. (2013) as inefficient in the application of MI approach. It takes longer run times with more intensive computations compared to simple conventional methods such as the arithmetic average and normal ratio.

In the last few decades, MI approach was frequently applied by several researchers (Khalifeloo et al. (2015), De Carvalho et al. (2017), Miró et al. (2017), Sattari et al. (2017), Jakhar et al. (2018), and Milo et al. (2019)) in the imputation of missing rainfall data. Due to its efficiency, there are various MI approach packages that have been introduced for missing data problems. R is one of the programming languages that offer MI packages. The packages comprise of “MICE”, “Amelia II package”, “missForest”, “Hmisc”, and “mi” however, the Amelia II package (Amelia) is the one that is mostly used in imputing the missing rainfall data.

Amelia is a bootstrapping-based algorithm with the adoption of the expectation maximization (EM) method that is applied in the estimation process (Honaker et al., 2018). Bootstrap is a resampling technique that is used to estimate statistics on a population by sampling a dataset with replacement (Bland and Altman, 2015). The bootstrap creates multiple datasets from the observed dataset without the need - to make any assumptions. The created sample is a set of randomly chosen observations that has the same size and equally representative of the original observed dataset. This was the main reason for considering bootstrap approach in multiple imputation. The bootstrap approach allows for the occurrence of uncertainties in producing the estimated values. The general bootstrap is the conventional approach that performs generalized random sampling. General bootstrap was considered in Amelia to simulate the estimation uncertainties in implementing multiple imputation. Cardenas et al. (2009) revealed that the weakness of the general bootstrap is that, it does not preserve the time series original structure in its process, and this makes the conventional sampling approach unsuitable to be used for data that has seasonal patterns, such as rainfall data.

According to Yunus et al. (2011), rainfall distribution in Peninsular Malaysia is commonly governed by two rainy seasons, i.e. southwest monsoon which usually occurs in mid of May and ends in August and northeast monsoon which initiates in early November and ends in February. There are transitional periods between the monsoon seasons which usually occur in April and October. The northeast monsoon season brings heavy rain to the east coast areas (involving the southern and eastern parts) of Peninsular Malaysia (Jamaludin et al., 2010) while the west coast region experienced heavy rainfall during the southwest monsoon season (Jamaludin and Jemain, 2009). Information about the seasonal element is

very important to be considered in dealing with the rainfall dataset (Deni et al., 2009). The original structure of the rainfall time series must be preserved so that all the significant information can be used to ensure accurate results of the analysis being conducted (De Carvalho et al., 2016).

Therefore, block bootstrap with controlled sampling technique was introduced in the MI approach to improving its performance, especially when dealing with rainfall time series data. This study is aimed to introduce the block bootstrap for data sampling purposes in the MI approach. This is an effort to improve the performance of the currently existing approach which uses generalized random sampling (general bootstrap) in order to produce accurate imputed values in providing a good quality dataset to be used for the public domain.

2. Material and Methods

2.1. Data description and study area

The daily rainfall data was collected from the Malaysian Drainage and Irrigation Department and the Malaysian Meteorology Department. It comprises of data from 13 rainfall stations throughout the West and Southern regions of Peninsular Malaysia for the 40-year period between 1975 to 2014. Most of the rainfall stations recorded data using automatic tipping bucket rain gauge, which has a sensitivity of 0.5 mm per tip. However, for some stations, the data collection is still executed using manual methods of measurement. The amount of rainfall for a particular day is the amount collected over the 24-hour period starting at 8:00 a.m.

Lalang Sg Lui (West region) and Johor Bahru (Southern region) were chosen as the target stations due to the following reasons: (i) provide supply to water resource (i) existence of main economic and administrative centers and (ii) serves as the central industrial and commercial hub. The neighbouring stations are chosen by considering the radius distance to the target station as followed by Jamaludin et al. (2008) and Jahan et al. (2018). The stations located within the radius of 100 km from the target station were selected as the neighbouring stations. This is the optimal distance as it considers a suitable number of stations and gives reasonable estimation results (Jamaludin et al., 2008). A greater radius could possibly slow down the computation time as the number of stations increases while a smaller radius will result in no neighbouring station available within the range.

Therefore, 100 km was chosen as the best distance for selecting the neighbouring stations. This is due to the same study area (Peninsular Malaysia) considered in both studies which are involved with almost the same distribution of rainfall stations. Three neighbours were assigned to Johor Bahru while eight neighbours were chosen for Lalang Sg Lui station. All the considered stations with their respective geographical coordinates and spatial information are listed in Table 1. Their locations in the geographical map of Peninsular Malaysia are presented in Fig. 1.

Table 1
The target stations (in bolded) and their neighbouring stations within the respective region with geographical coordinates and distances.

Region	Station No	Station Name	Geographical Coordinate			Euclidean Distance	Great-Circle Distance (km)
			Longitude	Latitude	Elevation (m)		
West	**4**	**Lalang Sg Lui**	**101.91**	**3.14**	**118**	**0**	**0**
	1	Ibu Bekalan Sg Bernam	101.35	3.70	7	0.79	88
	2	JPS Jaya Setia	101.41	3.37	17	0.55	61
	3	JPS Ampang	101.75	3.16	52	0.16	18
	5	Kg. Chennah	102.07	3.09	127	0.17	19
	6	Telok Gong	101.39	2.93	3	0.56	62
	7	JPS Sikamat	101.96	2.74	93	0.40	45
	8	Kg. Sawah Lebar	102.26	2.76	97	0.52	57
	9	Ldg. New Rompin	102.51	2.72	187	0.73	81
Southern	**13**	**Johor Bahru**	**103.75**	**1.47**	**40**	**0**	**0**
	10	Bt. 42 Jln Kluang/ Mersing	103.74	2.26	16.6	0.79	88
	11	Simpang Mawai - Kuala Sedili	103.97	1.85	26	0.44	49
	12	Sek. Men. Bkt. Besar	103.72	1.76	45	0.29	32

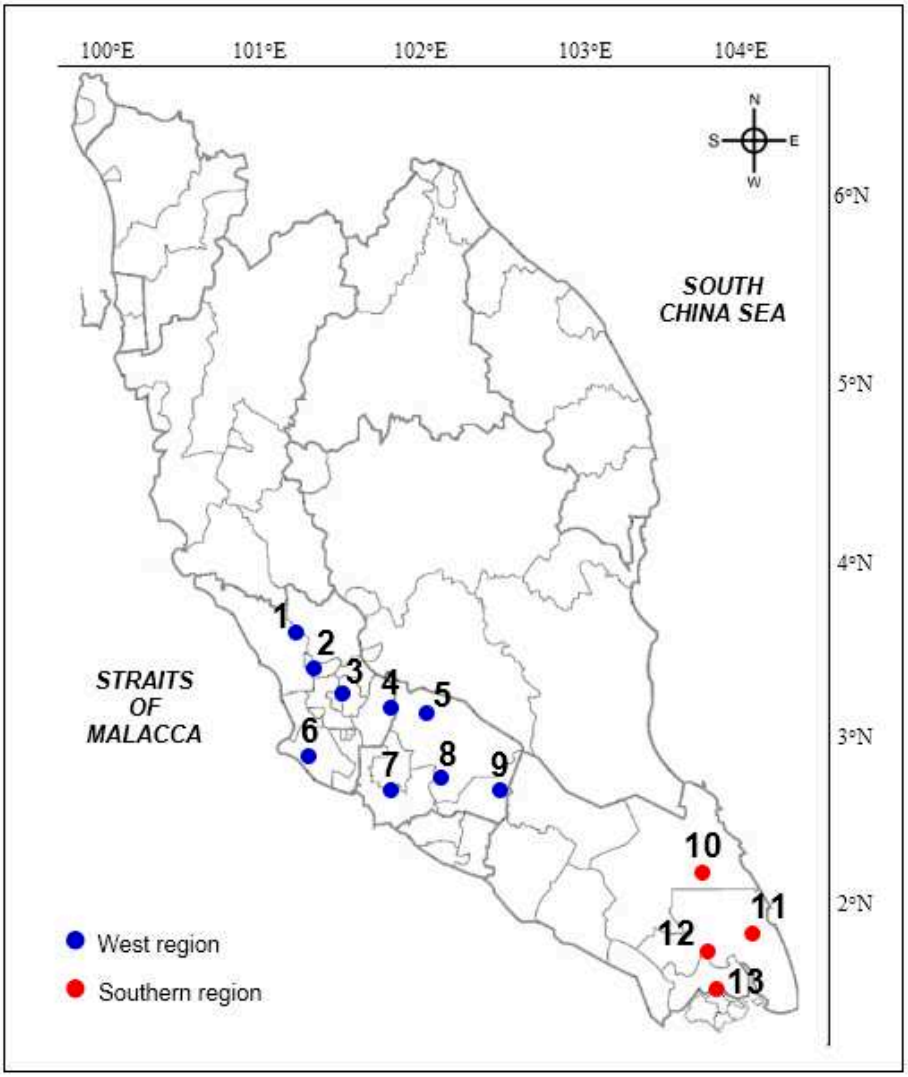


Fig. 1. Location of the selected rainfall stations.

2.2. Research design

Fig. 2 illustrates the stages involved in missing data imputation analysis. The multiple imputation procedure starts with the process of resampling for the observed dataset. Two sampling approaches are considered, i.e. general bootstrap which represents the conventional approach and block bootstrap denotes the controlled sampling approach. Then, the missing values are estimated using several normal ratio methods such as Old normal ratio (ONR), Modified normal ratio based on the trimmed mean (NRTR), Modified normal ratio based on the median (NRMED), and Modified normal ratio based on the geometric median (NRGMED). Finally, the performance of the estimation methods for both of the sampling approaches are compared based on several criteria, namely, MSE, RMSE, MAE, and SIndex. Their performances are evaluated based on various percentages of missing data and the outliers that will be created in the dataset. This way, the accuracy and consistency of the results is spontaneously assessed. The detailed explanation for each stage will be discussed in the next section.

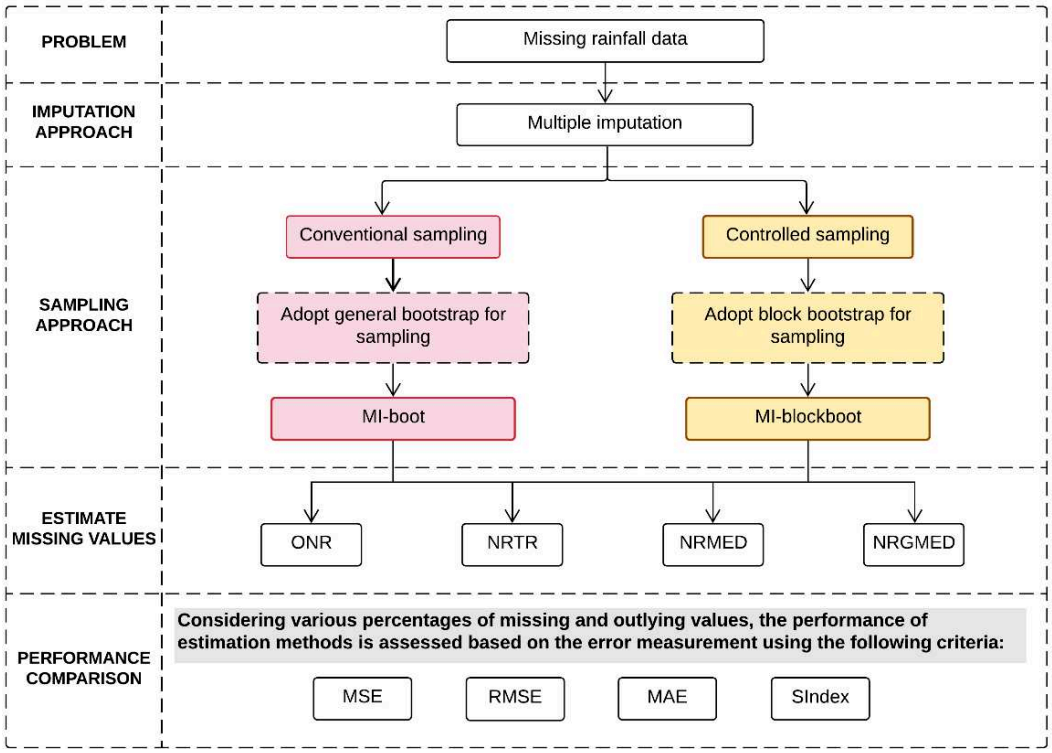


Fig. 2. The stages of missing data imputation analysis

2.3. Experimentation part

The experimental design executed in this study for the imputation of missing values in the daily rainfall dataset is discussed below. The procedure for the experimentation is conducted based on several stages, i.e. data handling, sampling approach, missing values imputation, and performance evaluation.

2.3.1. Data handling

There are two different types of procedures employed in the study. The first procedure utilizes the observed dataset whereas the second procedure uses the same dataset with few numbers of created outliers. The datasets used in the first and second procedures are named Dataset 1 and Dataset 2, respectively. Fig. 3 presents the data handling procedure implemented in this study.

In order to assess the robustness of the imputation approaches in this study, outliers were created in the observed dataset of the target station, which is named Dataset 2. The outliers were created based on three different levels, in terms of outliers' percent (namely 5%, 10%, and 15%). The procedure for the creation of outliers to be used in the rainfall dataset is explained below:

i. Set the number of outliers to be included in the dataset for the experimentation (i.e. 5%, 10%, and 15% of the dataset).
ii. Identify the existing outliers in the observed dataset of the target station using the boxplot method as applied by Laurikkala et al. (2000). In the boxplot, the boundaries are set to determine the outlying observations. The observations that exceed the boundaries are considered as outliers.
iii. Determine the percentage of existing outliers in the observed dataset.
iv. Create outliers that have been set in (i) by considering the percentage of available outliers in (iii). The artificial outliers are generated from the uniform distribution as suggested by (Tax and Duin, 2001).
v. The generated outliers are randomly replaced with the observed values in the dataset.

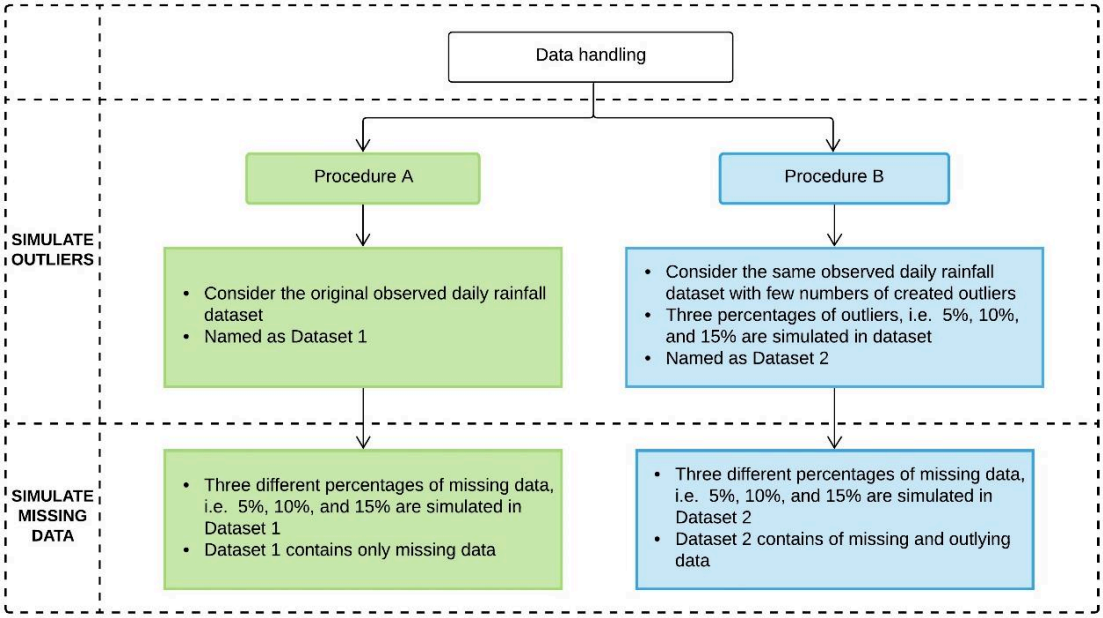


Fig. 3. Procedure for data handling

Since a complete observed rainfall dataset was used for this study, the missing data were artificially created in both datasets (Dataset 1 and Dataset 2) in order to evaluate the performance of the imputation approaches. Three different levels of missingness percentage (%) were considered to assess the consistency of the estimation results. The missing values are imputed by the considered estimation methods using these following procedures:

i. Identify a target station that has problems of missing values, however, for the purpose of analysis, the station with a complete dataset is used.
ii. Identify neighbouring stations based on a 100 km radius distance from the target station.
iii. Resample rainfall dataset of the identified target station (i) to produce bootstrapped dataset.
iv. Repeat step (iii) five times to produce five bootstrapped datasets.
v. Generate artificial missing values in the bootstrapped datasets in (iii) for the three different levels of missingness percentage (5%, 10%, and 15%) randomly.

    (Supposedly, if 10% of the data that is randomly chosen for testing methods are considered missing, the rest 90% of the data are used to calculate the estimation values for missing data).

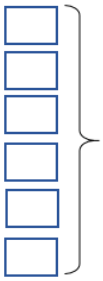
vi. Using the dataset of neighbouring stations obtained in (ii) to estimate the missing values in datasets in (iv) using the proposed estimation methods.
vii. The missing data is then imputed with the estimated values obtained in (vi) to produce the complete rainfall datasets.
viii. Do error analysis on each of the complete imputed dataset.
ix. Average the results of the analysis to produce a single result.
x. The results in (ix) are then compared between the estimation methods to evaluate their performance.

2.3.2. Sampling approach

In this study, two MI approaches were considered, i.e. MI based on general bootstrap (conventional sampling) and MI based on block bootstrap (controlled sampling). The former MI approach is named MI-boot while the latter is named MI-blockboot. Table 2 represents the comparison of both MI approaches.

The MI-blockboot is the enhancement of the conventional sampling MI approach. The block bootstrap was introduced to overcome the limitation of general bootstrap incorporated in the MI-boot wherein the original structure of the rainfall series is not preserved. Based on the comparison, the MI-blockboot shows its advantages over the MI-boot. It's ability to preserve the original structure of the time series dataset promises more accurate estimation results.

Table 2
Comparison of MI approaches considered in the study

Description	**Multiple Imputation Approach**	
	MI-boot	**MI-blockboot**
Approach Name	Multiple Imputation based on general bootstrapping	Multiple Imputation based on block bootstrapping
Sampling approach	Conventional sampling	Controlled sampling
Type of Bootstrapping	General Bootstrap	Block Bootstrap
Capability/ Limitation	Lose dependency structure of the dataset	Preserves dependency structure of the dataset (consider the seasonal monsoon aspect)
Bootstrapping Illustration	 Data are directly resampled	 Data are divided into several blocks before being resampled
Procedure	Dataset is directly resampled	i. Dataset is divided into *n* blocks ii. Resample each block iii. Concatenate all the blocks
Estimation Method	Normal Ratio (existing and modified)	Normal Ratio (existing and modified)

- Controlled sampling (block bootstrap)

Block bootstrap (also known as moving block bootstrapping (MBB)) was exclusively introduced for time series data (Inoue & Shintani, 2006). This is due to the characteristics of time series dataset wherein there is a need to preserve its original structure. Therefore, in order to make sure that the original structure of the studied dataset is not affected, block bootstrap was introduced in the current study. The adoption of block bootstrap in the MI was intended to improve the results of imputation, especially when dealing with time series data. The block bootstrap was pioneered by Carlstein (1986) and further developed by Kunsch (1989). The idea featured is to resample blocks (intervals of a time series) with replacement. The blocks are usually disjointed and cover the entire time series, but they can be overlapping (Chernick and LaBudde, 2011). Block bootstrap divides the rainfall time series into several blocks and samples all of the blocks before concatenating them. As a result, the dependency structure of the time series was preserved within each block (Li, 2006).

Block bootstrap was adopted by Burhanuddin et al. (2017b) in their MI that was specifically proposed for rainfall time series. The bootstrapping approach was applied to normal ratio methods to impute

missing values in the Malaysian daily rainfall dataset. They have proved the capability of the proposed MI in providing accurate estimation values, especially when dealing with the dataset that contains outliers. The procedure was a bit different from other studies in determining the size and number of blocks. Rainfall time series were divided into several blocks according to the characteristics of the rainfall data, such as the seasonal pattern and the rainfall amounts before going through the actual bootstrap process. The bootstrap was applied separately on each block of the data. The detailed procedure of block bootstrap that has been implemented in the current study is explained below (Burhanuddin et al., 2017b).

i. Divide rainfall time series into several blocks.

   Rainfall time series distribution is investigated based on the seasonal pattern. Monsoon seasons in Malaysia are reviewed to understand the pattern of rainfall amounts throughout a certain period of a year. Generally, the monsoon season is divided into three categories, i.e. southwest monsoon, northeast monsoon, and inter-monsoon. The southwest monsoon is the source of a high total amount of rainfall in the northwest, southwest, and west regions of Peninsular Malaysia which usually occurs from the middle of May and ends in August. On the other hand, the northeast monsoon which usually begins in early November and ends in February only caters to the eastern region of Peninsular Malaysia with a high total amount of rainfall. Inter monsoon period occurring in between the two monsoons, which are during the March/ April and September/ October is also usually associated with a heavy rainfall. Therefore, from the reviews, the rainfall time series is divided into four blocks, i.e. (1) the month of May to August, (2) March to April, (3) November to February, and (4) September to October.

Block 1	Block 2	Block 3	Block 4
(Mar - Apr)	(May - Aug)	(Sep - Oct)	(Nov - Feb)

ii. Resample each block of rainfall time series obtained in (i) for 1000 times (the size of the bootstrapped sample is the same size as the original data). Then merge the blocks of the bootstrapped sample to produce daily rainfall time series for a year.

Block 1	Block 2	Block 3	Block 4
x_i i = 1, ..., 61	x_j j = 1, ..., 123	x_k k = 1, ..., 61	x_l l = 1, ..., 120

iii. Repeat step (i) and (ii) for all 40 years' daily rainfall data and merge them to generate a set of bootstrapped samples representing a new sample for a period of 40 years, equal to 14610 daily data.

New Sample
x_n n = 1, 2, 3, ..., 14610

iv. Repeat step (i) to (iii) for 5 times to produce 5 new samples

The new samples obtained from the bootstrapping process were used in the next imputation procedure discussed in detail in the following section.

2.3.3. The imputation procedure

Several estimation methods were considered to estimate the imputed values of missing rainfall data. The normal ratio (NR) method has been chosen due to its formulation that can be implemented through the MI approach. Besides that, NR is selected due to its simplicity and efficiency in estimating missing rainfall values, as observed in several studies such as Mair and Fares (2010), Yozgatligil et al. (2013), and Radi et al. (2015). Several modified versions of NR methods proposed by Burhanuddin et al. (2017a) were considered in this study. They have robustified the existing ONR method by considering the trimmed mean and geometric median to improve the accuracy of the estimation results. One more modified NR was introduced in this study, i.e. NR based on the median. The estimation methods were implemented using both MI-boot and MI-blockboot approaches discussed earlier. The performance of the methods was compared to identify the best one. The different estimation methods are detailed as follows:

- Old normal ratio (ONR)

The existing old normal ratio (ONR) estimation method was modified to improve its performance in handling the missing rainfall data and, thereby treat the outliers to reduce their effect on the estimation results. Paulhus and Kohler (1952) were the pioneers to apply the ONR method in estimating the missing rainfall data. The method is given as follows:

$$Y_t = \sum_{i=1}^{N} \frac{\frac{\mu_t}{\mu_i}}{\sum_{\substack{i=1 \\ i \neq t}}^{N} \frac{\mu_t}{\mu_i}} Y_i \tag{1}$$

where μ_t and μ_i are the sample means of the available data at target station t and i^{th} neighbouring station respectively; Y_t is the missing data at target station t; Y_i is the concurrently observed data at the i^{th} neighbouring station; N is the number of surrounding stations.

- Modified normal ratio based on the trimmed mean (NRTR)

Trimmed mean is used to replace the arithmetic mean in the modified NR method. A 5% trimming percentage has been considered in this study. The modification was made to Eq. (1) and expressed as follows:

$$Y_t = \sum_{i=1}^{N} \frac{\frac{\mu_{trim_t}}{\mu_{trim_i}}}{\sum_{\substack{i=1 \\ i \neq t}}^{N} \frac{\mu_{trim_t}}{\mu_{trim_i}}} Y_i \tag{2}$$

where μ_{trim_t} and μ_{trim_i} are the sample trimmed means of the available data at target station t and i^{th} neighbouring station respectively

- Modified normal ratio based on the median (NRMED)

The mean or median is a simple way to impute the missing values. However, since the mean is very sensitive to the existence of outliers, the median is suggested to assure robustness (Fukuda and Rosta, 2005). Median imputation is preferable when the distribution of the underlying variable is not symmetric but rather skewed. Saeed et al. (2016) and Khamkong et al. (2017) applied the median to impute the missing values in rainfall dataset for their research studies. However, in this study, the median was considered as the weighting factor that replaces arithmetic mean in Eq. (1):

$$Y_t = \sum_{i=1}^{N} \frac{\frac{\mu_{med_t}}{\mu_{med_i}}}{\sum_{\substack{i=1 \\ i \neq t}}^{N} \frac{\mu_{med_t}}{\mu_{med_i}}} Y_i \tag{3}$$

where μ_{med_t} and μ_{med_i} are the sample medians of the available data at target station *t* and i^{th} neighbouring station respectively

- Modified normal ratio based on the geometric median (NRGMED)

The final modified NR method considered geometric mean (Gmed) in an effort to improve its performance. The Gmed is defined as the data minimizing the sum of distances to the sample dataset. The method is defined as follows (Das and Imon, 2014):

For a set of observations $x_i = \{x_1, x_2, x_3, ..., x_n\}$, take the natural logarithm of the data. Let it be defined as $a_i = \ln(x_i)$. Then, the median of a_i is computed and defined as follows:

$$\tilde{A} = \text{median}(a_i)$$

Finally, Gmed is obtained by the exponential of the median of logarithm values.

$$\text{Gmed} = \exp(\tilde{A})$$

For this study, the Gmed is introduced in NR replacing the arithmetic mean. The formulation of the NRGMED method is as follows:

$$Y_t = \sum_{i=1}^{N} \frac{\frac{\text{Gmed}_t}{\text{Gmed}_i}}{\sum_{\substack{i=1 \\ i \neq t}}^{N} \frac{\text{Gmed}_t}{\text{Gmed}_i}} Y_i \tag{4}$$

where Gmed_t and Gmed_i are the medians of the available data at target station *t* and i^{th} neighbouring station respectively

The application of the estimation methods through both MI-boot and MI-blockboot approaches is implemented based on the procedure given below. All the existing and modified NR methods were considered to be implemented and tested. The ONR method was chosen to be discussed as an example.

i. The new bootstrapped samples with created artificial missing values were utilized. The procedure of introducing missing data in the dataset will be explained in the following Application section.
ii. The ratio means of each sample of the target station and the dataset of its nearby stations are considered as the weighting factors for the estimation methods (an example of the ONR method - refer to Eq. (1)).

$$w_{L_i} = \frac{\mu_{L_i}}{\mu_i} \ , \qquad L = 1, 2, \ldots, 5$$

where w_{L_i} is the weight of the i^{th} neighbouring station for new sample L; μ_{L_i} and μ_i are the arithmetic means of the available data for new sample L and i^{th} neighbouring station respectively.

iii. Five weighting factors are obtained to produce five different estimated values for each individual missing value.
iv. Each estimated value will be used to impute the missing values in the original rainfall dataset, thus, producing five complete datasets with different imputed values and the same observed values.
v. The complete imputed datasets are analysed and the results of the analysis are pooled to produce a single result of the analysis.

2.3.4. Performance comparison

The performance of estimation methods in terms of error measurement is determined to evaluate the accuracy of the estimation results produced. Five elements of performance criteria were considered for this purpose, i.e. MSE, RMSE, MAE, and SIndex. The procedure of evaluation is executed as follows:

i. Find the difference between the estimated missing values and observed values.
ii. The error values obtained from (i) are then used to calculate different elements of the performance criteria, as mentioned in Eq. (5) to (8).
iii. The most appropriate methods are selected based on the least value of MSE, RMSE, MAE, and the highest value of SIndex.

$\hat{Y}$ referred to the estimated value, Y is the actual value of the observation, $\bar{Y}$ is the mean of the actual values, and n is the number of observations.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{Y}_i - Y_i\right)^2 \tag{5}$$

$$\text{RMSE} = \sqrt{\text{MSE}} \tag{6}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{Y}_i - Y_i\right| \tag{7}$$

$$\text{SIndex} = \frac{\sum_{i=1}^{n}\left(\hat{Y}_i - Y_i\right)^2}{\sum_{i=1}^{n}\left(\left|\hat{Y}_i - \bar{Y}\right| + \left|Y_i - \bar{Y}\right|\right)^2} \quad (8)$$

3. Results and discussion

3.1. Summary of rainfall pattern

Fig. 4 shows the matrix plot of the pattern of average monsoonal rainfall amount for all the rainfall stations considered in this study. The data on the amount of average rainfall for a period of 40 years for four monsoon periods i.e. southwest, inter-monsoon 1, northeast, and inter-monsoon 2 is comprehended from the figure. Southwest monsoon occurred on four consecutive months starting from May to August while northeast monsoon prevails from November to February. Inter-monsoon 1 represented the transition of southwest monsoon while inter-monsoon 2 is the transition of the northeast monsoon.

It can be seen that the rainfall event in the Southern region was exposed to the northeast monsoon. The amount of rainfall received during these monsoons was rather significantly high, especially for Bt. 42 Jln Kluang/ Mersing and Simpang Mawai - Kuala Sedili stations. This region received a rather high amount of rainfall during this monsoon i.e. up to 1050 mm. Meanwhile, for the West region, the rainfall event was not influenced by monsoon seasons. The pattern did not exhibit an obvious difference with the amount of rainfall received by each monsoon. The rainfall received by the stations in this region was distributed evenly throughout the year.

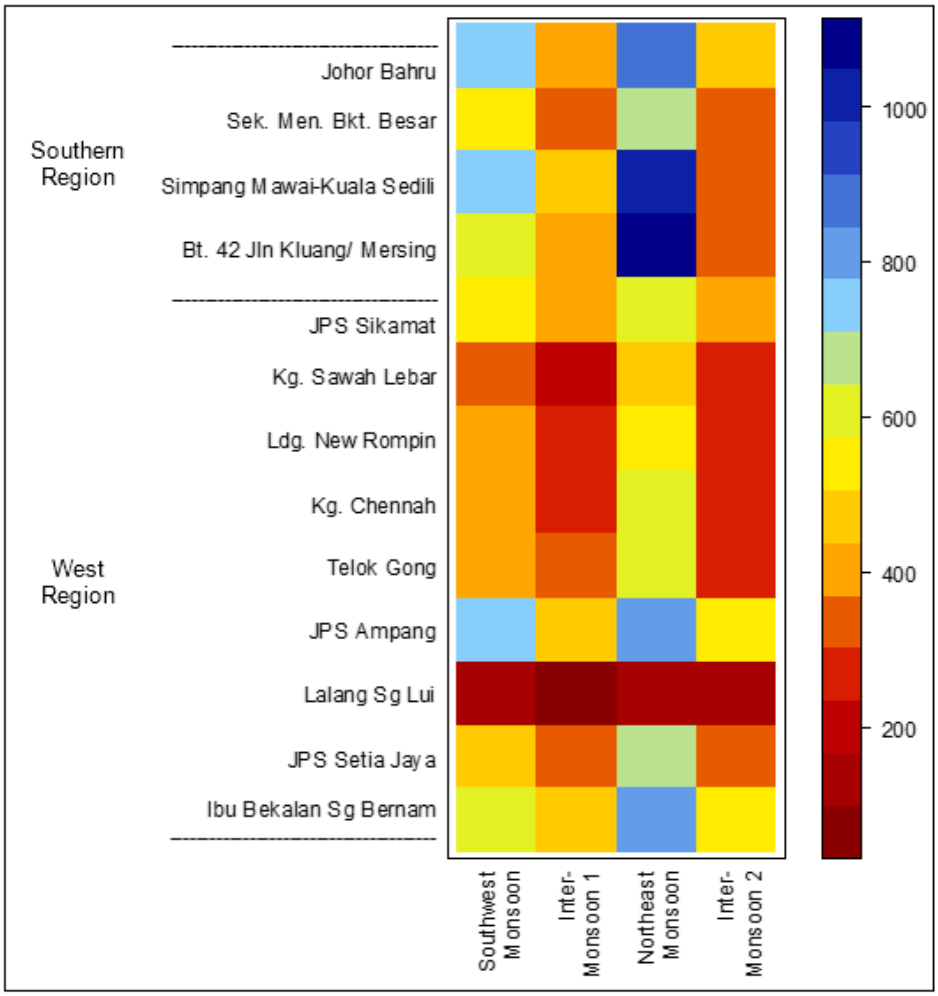


Fig. 4. The pattern of average monsoonal for a period of 40 years in all study stations.

Fig. 5 displays the matrix plot of the average monthly rainfall data pattern for all of the stations in both regions. It was observed that the level of average monthly rainfall received by the Southern region appears to be high at the end of the year (December). The amount of rainfall received by this region ranges between 97 mm to 435 mm with the highest at Bt. 42, Jln Kluang/ Mersing. The month of December is usually associated with the northeast monsoon season in which the stations will receive a higher amount of rainfall compared to the other months. For the West region, the average amount of monthly rainfall received by each station is observed to be evenly distributed throughout the year. The rainfall amount received by this region ranged between 24 mm and 310 mm. Therefore, it can be concluded that the rainfall event in the Western region was less exposed to the monsoon season.

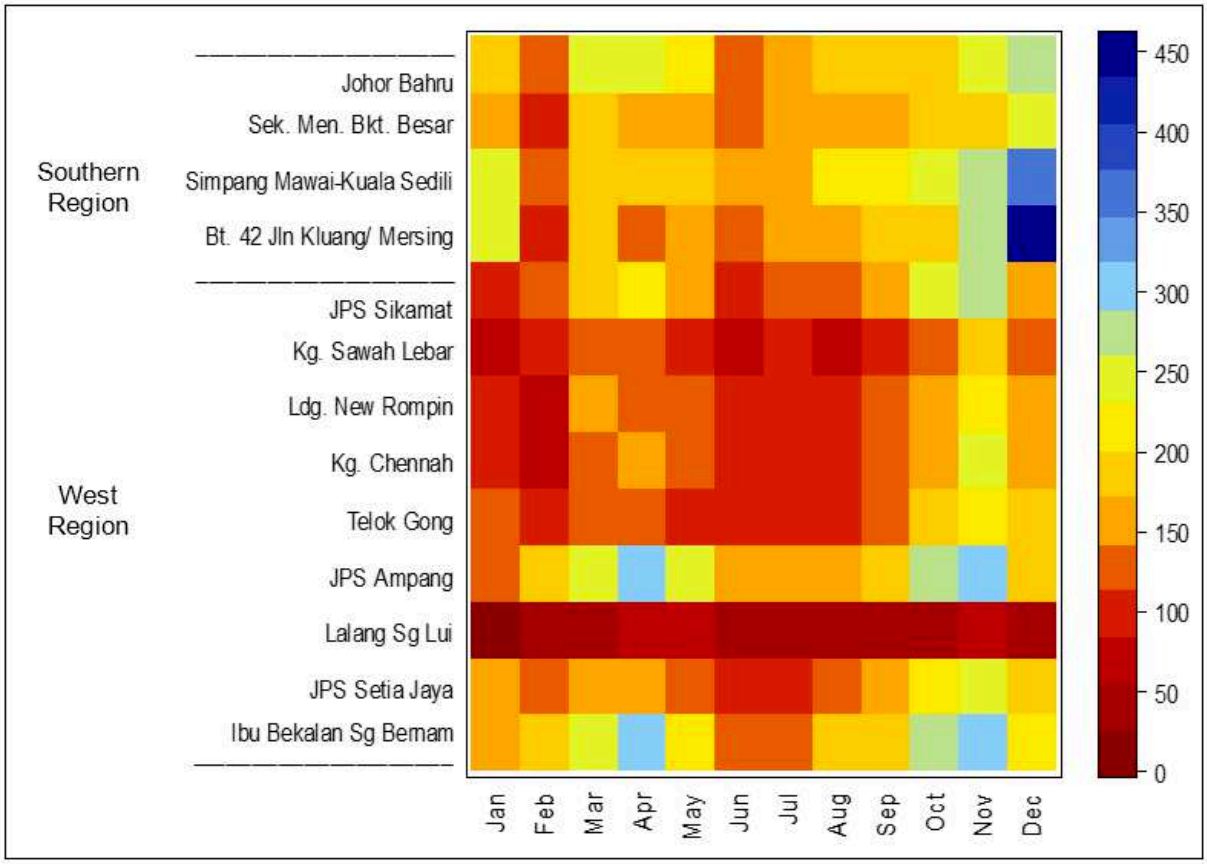


Fig. 5. The pattern of average monthly rainfall for a period of 40 years in all study stations.

3.2. Performance evaluation of imputation approaches

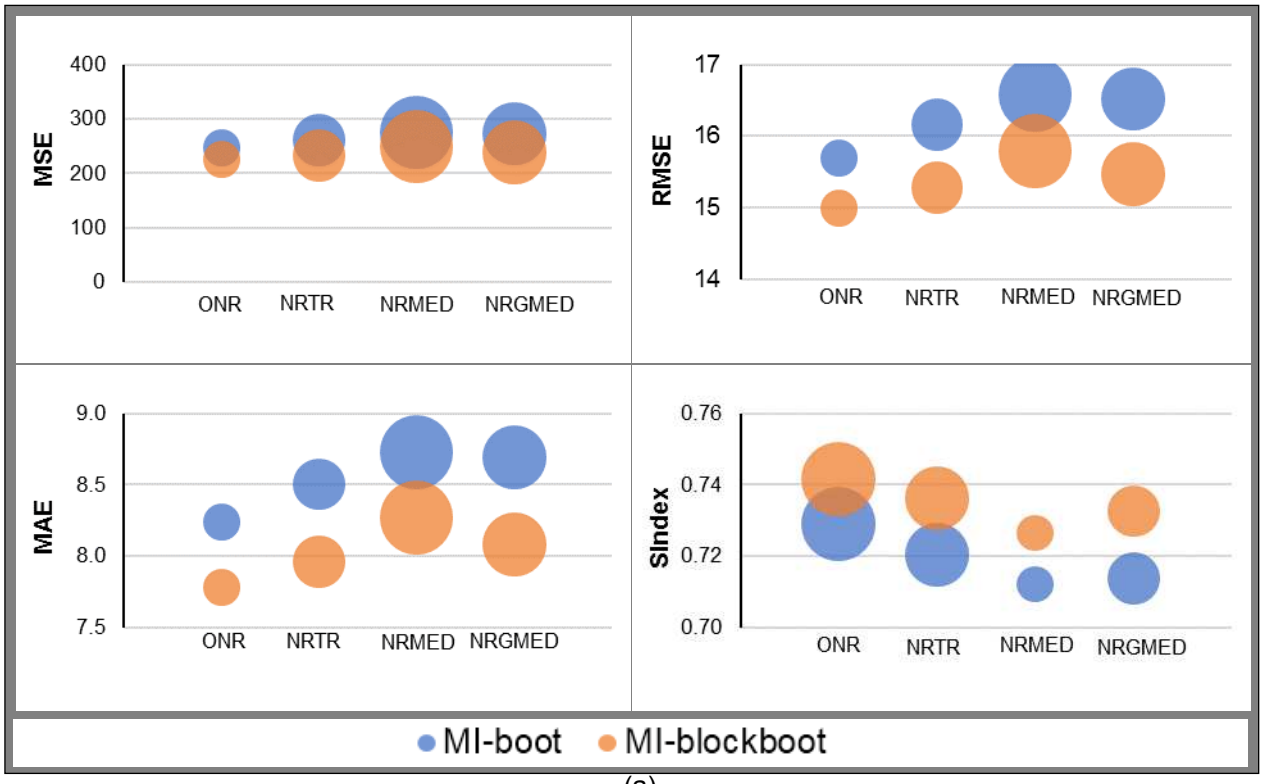
The errors between the estimation results and the observed values were obtained to assess the performance of the NR methods through their implementation of both MI-boot and MI-blockboot. The most appropriate estimation methods are determined based on the least value of MSE, RMSE, and MAE and the highest value of SIndex. The comparison of the performance of the association of NR methods and MI approaches for Dataset 1 and Dataset 2 to assess their capability in both situations, i.e. without and with the presence of created outliers in the dataset, respectively was carried out. The performance evaluations have been comprehensively explained below for each estimation method and rainfall station.

Fig. 6 shows the performance comparison of NR methods implemented through both MI-boot and MI-blockboot on Dataset 1 for Johor Bahru (a) and Lalang Sg Lui stations (b). The results illustrated correspond to the performance with a data missingness of 5% based on MSE, RMSE, MAE, and SIndex. The ONR, NRTR, and NRMED methods implemented through MI-blockboot are observed to provide more accurate estimation results compared to the MI-boot for both target stations. Block bootstrap has presented its benefits when dealing with time series data by providing more precise information on the data characteristics. The advantage of the combination of robustness and blocking elements has successfully improved the accuracy of the estimation results. The adoption of trimmed mean and median as a robust estimator can be clearly witnessed for the application at both stations.

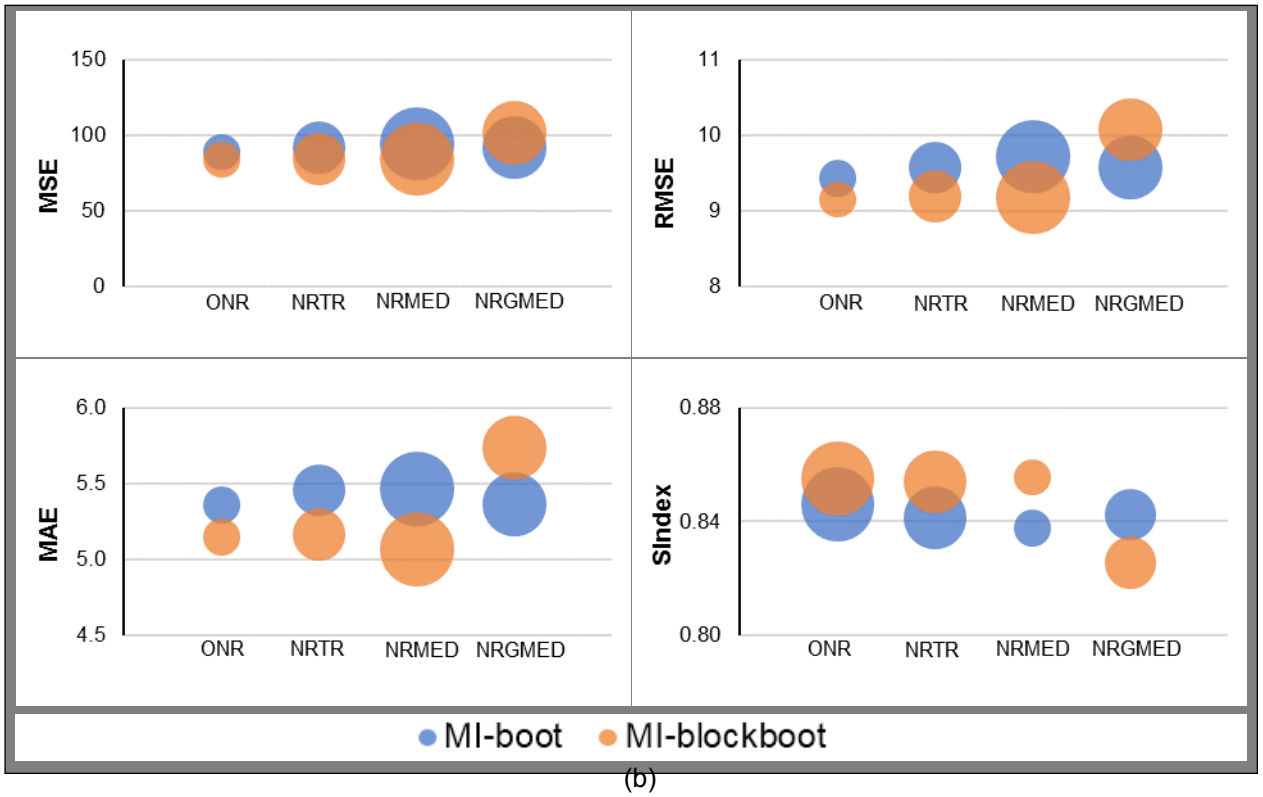
NRGMED method seemed to perform even better when implemented using MI-boot compared to MI-blockboot at Lalang Sg Lui station. This method was expected to perform better when incorporated with block bootstrap similar to the other methods. Therefore, in order to identify the causes of this situation, this study was going to execute all NR methods on the dataset that was affected by the presence of outliers (i.e. Dataset 2). The purpose of employing all the methods was to prove the capability of robust methods in estimating more accurate results. The efficiency of the methods was expected to be revealed when dealing with outliers. The strength of this method in producing more accurate estimation results can be more highlighted when dealing with dataset 2.

The consistency of the performance (MSE, RMSE, MAE, and SIndex) of each method for different levels of missingness (%) was evaluated and plotted in Fig. 7 for Johor Bahru and and Fig. 8 for Lalang Sg Lui stations. The performance of all methods for both MI-boot and MI-blockboot for Johor Bahru station slightly decreased with an increase in the level of missing data percentages. The comparison between the MI approaches indicate more accurate estimation results were produced by the controlled sampling MI approach for all levels of missingness (%).

Meanwhile, it can be noted that the performance of the NR methods for Lalang Sg Lui station was rather consistent as they were not really affected by the increasing number of missing values in the dataset. The MI-blockboot was found to perform consistently better than the other MI approach, except for NRGMED method. The NRGMED incorporated MI-boot approach provided more accurate estimation results irrespective of the level of missingness (%).



(a)



(b)

Fig. 6. Performance of estimation methods implemented using both of the MI approaches on Dataset 1 based on all performance indicators for Johor Bahru (a) and Lalang Sg Lui station (b)

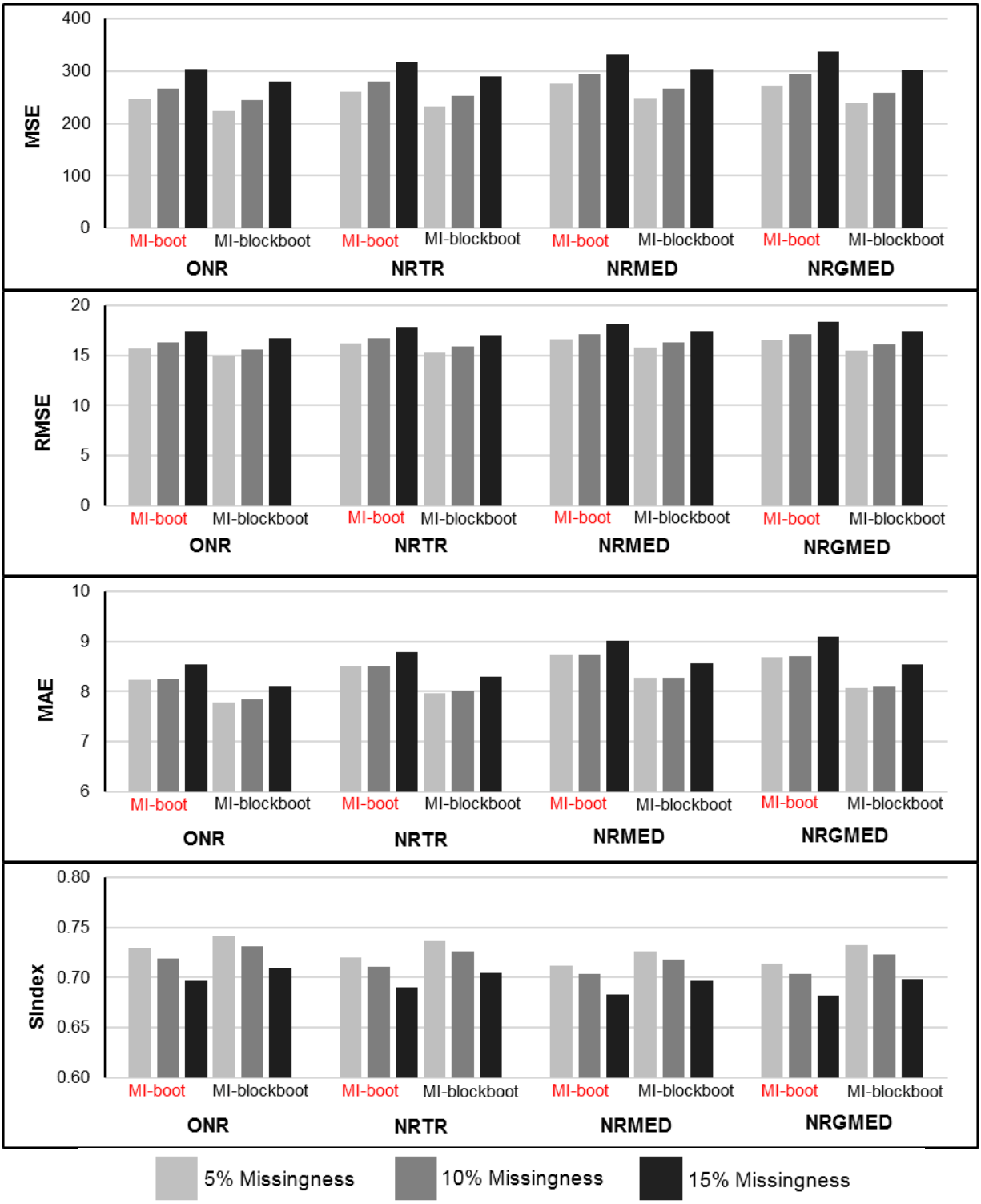


Fig. 7. Performance of estimation methods implemented using two different MI approaches on Dataset 1 based on all performance indicators for Johor Bahru for different percent of missingness.

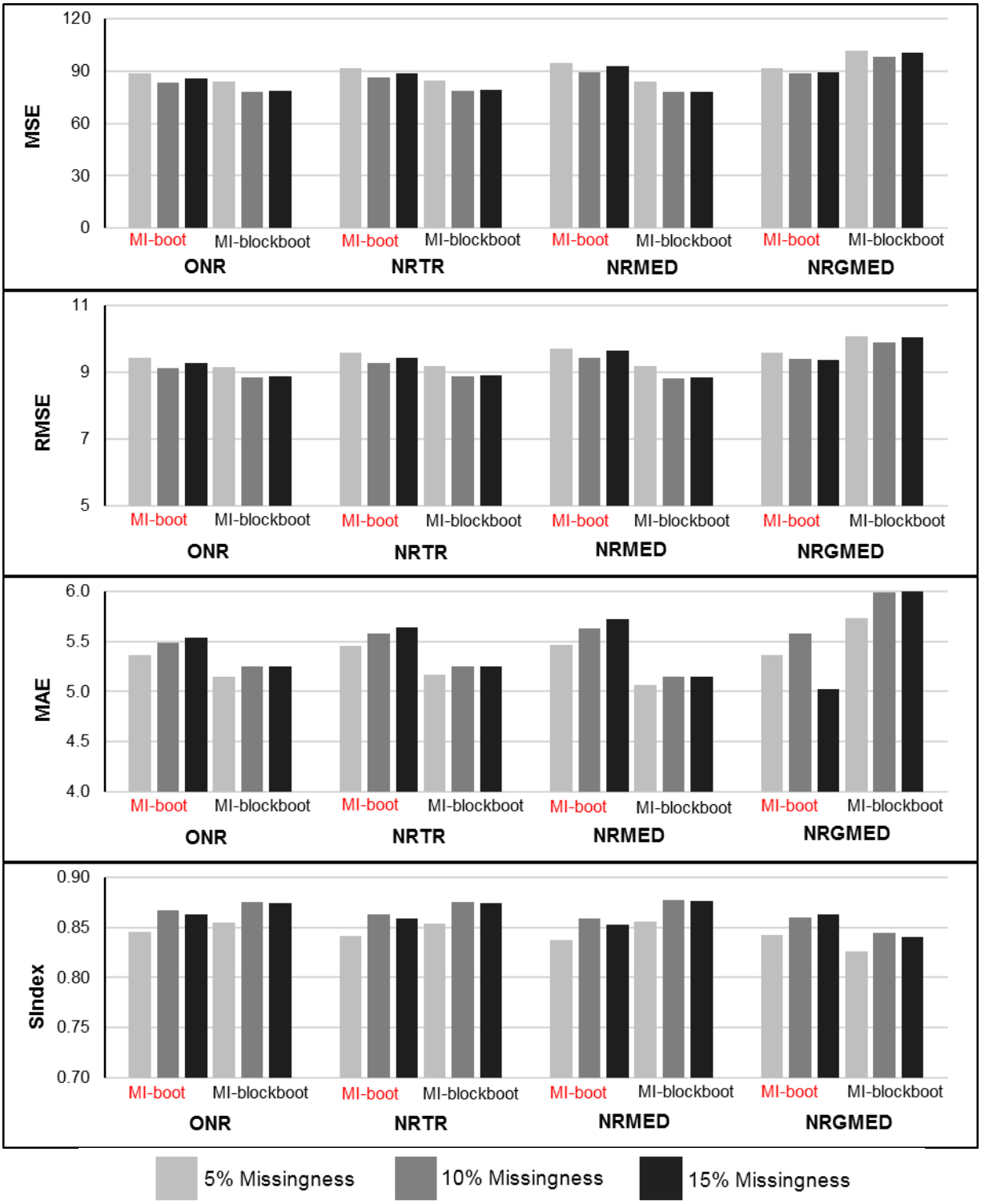


Fig. 8. Performance of estimation methods implemented using two different MI approaches on Dataset 1 based on all performance indicators for Lalang Sg Lui station for different percent of missingness

The robustness of an estimation method and the imputation approach is measured by its capability to be sturdy and insensitive to the presence of outliers in a dataset and maintain its performance by producing accurate results (Khamkong et al., 2017). The performance of NR methods implemented using both of the MI approaches was tested for different levels of outliers (%). Three levels of created outliers, i.e. 5%, 10%, and 15% were considered in Dataset 2 (as discussed in Section 4).

Table 3 and Table 4 present the performance comparison of different NR methods implemented through MI-boot and MI-blockboot on Dataset 2 for Johor Bahru and Lalang Sg Lui stations, respectively. In order to detail examine the effects of outliers, the discussion will be firstly focused on the results of imputation at 5% missingness. Overall, it is shown that the consideration of the various number of outliers in datasets has tremendously affected the estimation results. Their performance declined with increasing percentage of outliers and the effect was significant on MI-boot's method results, especially the NRGMED. The method exhibited highly sensitive to the presence of outliers; thus, its performance was seriously influenced by the increasing number of outliers, for both Johor Bahru and Lalang Sg Lui stations. Based on the four elements of criteria, MI-blockboot's methods produced a more accurate estimation result compared to the MI-boot, in spite of increasing outliers' percentage (refer Table 3 and Table 4, the values in bold represent more accurate results produced).

Table 3 demonstrates that all NR methods implemented using MI-blockboot performed better compared to the MI-boot for all levels of outliers' percentage, particularly the NRMED and NRGMED methods. Those methods were revealed to remain stable and powerful even under in the influence of outliers. The strength of the robust NRGMED associated with the controlled sampling (block bootstrap) was very obvious compared to its combination with conventional sampling (general bootstrap) as it continued to produce robust estimation results with increasing outliers in dataset. Meanwhile, the ONR and NRTR methods were seen as a sensitive and less stable option owing to their vulnerability to be affected with an increase in the outliers for both MI approaches, specifically at 15%. Therefore, the MI-blockboot's NRMED and NRGMED were found to be the most appropriate methods for Johor Bahru station.

Meanwhile, based on Table 4, through the MI-blockboot implementation, all methods produced more accurate estimation results for all levels of outliers' percent compared to the MI-boot. The NRGMED was identified to perform better than other methods through MI-blockboot. The method produced robust estimation results even when dealing with the dataset that was affected by a large number of outliers. Besides that, through the implementation of both MI approaches, ONR method has performed well over the NRTR and NRMED for higher level of outliers (10% and 15%). The capability of the methods was rather affected by the presence of large number of outliers, especially the NRMED. Thus, the most appropriate methods for imputing the missing data in Lalang Sg Lui station were the NRGMED and the ONR, as they exhibited a higher tendency to be robust.

The consistency of the performance for each method for various levels of missingness (%) was also evaluated (refer to Table 3 and Table 4). The evaluation was also made for different levels of outliers. As seen in the nature of performance for Dataset 1, all the methods performed with an acceptable level of consistency with slight increment and decrement for increasing levels of missingness (%). The performance of all the estimation methods implemented using both MI-boot and MI-blockboot were considered consistent as they were not really affected by the increasing number of missing values in the dataset.

Table 3
Performance results of imputation approaches on Dataset 2 based on performance indicators for Johor Bahru station for all different levels of missingness (%) and outliers (%)

Indicator	Outliers %	Missingness %	MI-boot				MI-blockboot			
			ONR	NRTR	NRMED	NRGMED	ONR	NRTR	NRMED	NRGMED
MSE	5%	5%	673.4	513.0	332.4	333.3	**556.7**	**405.7**	**316.0**	**287.3**
		10%	688.9	529.2	350.7	350.7	**569.9**	**420.9**	**332.3**	**303.5**
		15%	708.8	555.8	384.8	385.9	**598.4**	**455.5**	**371.5**	**342.5**
	10%	5%	1652.9	1673.8	515.8	53120.1	**1421.8**	**1360.8**	**479.9**	**389.3**
		10%	1644.4	1658.4	527.3	56851.7	**1398.7**	**1327.4**	**475.4**	**390.7**
		15%	1643.5	1663.0	560.5	30177.5	**1407.0**	**1344.0**	**515.1**	**424.5**
	15%	5%	3098.5	3801.6	924.7	13770.4	**2846.7**	**3414.5**	**823.8**	**701.5**
		10%	3084.1	3769.9	928.9	15048.3	**2812.2**	**3354.8**	**813.4**	**691.8**
		15%	3044.1	3725.6	967.3	16404.2	**2750.5**	**3275.1**	**839.6**	**696.9**
RMSE	5%	5%	25.95	22.64	18.23	18.24	**23.59**	**20.14**	**17.78**	**16.95**
		10%	26.24	23.00	18.72	18.71	**23.87**	**20.51**	**18.23**	**17.42**
		15%	26.62	23.57	19.61	19.63	**24.46**	**21.34**	**19.28**	**18.51**
	10%	5%	40.64	40.90	22.70	161.22	**37.70**	**36.88**	**21.90**	**19.73**
		10%	40.54	40.71	22.96	166.90	**37.39**	**36.42**	**21.80**	**19.76**
		15%	40.53	40.77	23.67	98.10	**37.50**	**36.65**	**22.69**	**20.60**
	15%	5%	55.66	61.64	30.4	87.14	**53.34**	**58.41**	**28.69**	**26.45**
		10%	55.52	61.38	30.46	90.42	**53.01**	**57.89**	**28.51**	**26.27**
		15%	55.16	61.02	31.08	94.26	**52.43**	**57.19**	**28.97**	**26.37**
MAE	5%	5%	13.31	11.74	9.57	9.58	**12.21**	**10.52**	**9.35**	**8.93**
		10%	12.96	11.49	9.50	9.49	**11.88**	**10.35**	**9.26**	**8.86**
		15%	13.09	11.67	9.76	9.76	**12.09**	**10.61**	**9.59**	**9.19**
	10%	5%	19.96	20.05	11.74	74.5	**18.65**	**18.26**	**11.36**	**10.30**
		10%	19.19	19.26	11.47	75.06	**17.83**	**17.41**	**10.94**	**9.99**
		15%	19.28	19.38	11.72	44.71	**17.95**	**17.57**	**11.26**	**10.25**
	15%	5%	26.60	29.24	15.31	40.78	**25.57**	**27.80**	**14.53**	**13.50**
		10%	25.63	28.16	14.82	41.07	**24.55**	**26.65**	**13.96**	**12.96**
		15%	25.69	28.26	15.10	43.23	**24.49**	**26.58**	**14.15**	**12.96**
SIndex	5%	5%	0.563	0.611	0.683	0.683	**0.597**	**0.651**	**0.691**	**0.706**
		10%	0.560	0.606	0.676	0.677	**0.594**	**0.646**	**0.684**	**0.699**
		15%	0.553	0.596	0.659	0.660	**0.583**	**0.631**	**0.665**	**0.679**
	10%	5%	0.407	0.405	0.609	0.357	**0.432**	**0.439**	**0.622**	**0.658**
		10%	0.409	0.408	0.607	0.351	**0.436**	**0.445**	**0.625**	**0.659**
		15%	0.405	0.403	0.594	0.468	**0.431**	**0.439**	**0.609**	**0.643**
	15%	5%	0.313	0.286	0.505	0.361	**0.325**	**0.300**	**0.526**	**0.556**
		10%	0.314	0.288	0.506	0.362	**0.327**	**0.303**	**0.530**	**0.560**
		15%	0.312	0.285	0.496	0.353	**0.326**	**0.302**	**0.521**	**0.556**

Table 4
Performance results of imputation approaches on Dataset 2 based on performance indicators for Lalang Sg Lui station for all different levels of missingness (%) and outliers (%)

Indicator	Outliers %	Missingness %	MI-boot				MI-blockboot			
			ONR	NRTR	NRMED	NRGMED	ONR	NRTR	NRMED	NRGMED
MSE	5%	5%	499.8	673.2	404.8	2428.2	**456.7**	**604.8**	**327.7**	**251.0**
		10%	484.3	654.7	386.9	2682.4	**448.6**	**598.1**	**323.3**	**208.9**
		15%	524.8	711.4	421.4	1838.8	**483.0**	**645.1**	**343.8**	**172.2**
	10%	5%	935.6	1383.2	6653.1	4723.1	**888.8**	**1305.1**	**4632.3**	**520.8**
		10%	920.4	1367.5	6638.3	4880.4	**869.0**	**1281.3**	**4299.3**	**390.4**
		15%	996.2	1481.4	7402.9	3836.2	**936.4**	**1381.4**	**4728.7**	**286.6**
	15%	5%	1268.6	1939.7	8279.1	5726.9	**1263.1**	**1931.0**	**8277.7**	**1333.5**
		10%	1256.0	1929.6	8305.4	5651.5	**1251.1**	**1922.2**	**8294.1**	**1049.2**
		15%	1353.0	2079.0	8961.8	4270.3	**1337.3**	**2053.1**	**8916.8**	**850.5**
RMSE	5%	5%	22.35	25.94	20.08	40.76	**21.37**	**24.59**	**18.09**	**15.82**
		10%	22.00	25.58	19.62	42.02	**21.17**	**24.44**	**17.96**	**14.44**
		15%	22.90	26.66	20.48	28.87	**21.97**	**25.38**	**18.52**	**12.82**
	10%	5%	30.59	37.19	81.53	59.93	**29.81**	**36.12**	**67.13**	**22.77**
		10%	30.34	36.98	81.42	60.69	**29.47**	**35.78**	**64.25**	**19.72**
		15%	31.56	38.49	86.04	47.04	**30.59**	**37.16**	**67.38**	**16.30**
	15%	5%	35.62	44.04	90.99	73.28	**35.54**	**43.94**	**90.98**	**36.22**
		10%	35.44	43.93	91.13	72.28	**35.37**	**43.84**	**91.07**	**32.02**
		15%	36.78	45.59	94.67	56.56	**36.57**	**45.31**	**94.43**	**28.89**
MAE	5%	5%	13.62	15.75	12.01	24.45	**13.02**	**14.94**	**10.80**	**9.41**
		10%	13.95	16.18	12.26	26.24	**13.43**	**15.47**	**11.22**	**8.99**
		15%	14.19	16.48	12.51	16.52	**13.61**	**15.69**	**11.31**	**7.26**
	10%	5%	18.58	22.56	48.96	35.42	**18.11**	**21.90**	**40.27**	**13.64**
		10%	19.20	23.35	50.78	37.28	**18.66**	**22.59**	**40.06**	**12.34**
		15%	19.53	23.78	52.62	27.62	**18.94**	**22.95**	**41.19**	**9.39**
	15%	5%	21.65	26.76	54.66	44.01	**21.60**	**26.70**	**54.66**	**21.66**
		10%	22.42	27.74	56.83	45.11	**22.38**	**27.68**	**56.79**	**20.00**
		15%	22.77	28.18	57.90	33.99	**22.63**	**28.00**	**57.76**	**17.63**
SIndex	5%	5%	0.448	0.374	0.506	0.369	**0.471**	**0.400**	**0.561**	**0.631**
		10%	0.482	0.407	0.545	0.402	**0.502**	**0.429**	**0.590**	**0.698**
		15%	0.458	0.383	0.521	0.611	**0.480**	**0.407**	**0.573**	**0.752**
	10%	5%	0.302	0.230	0.075	0.283	**0.312**	**0.240**	**0.104**	**0.442**
		10%	0.330	0.254	0.087	0.297	**0.342**	**0.266**	**0.128**	**0.543**
		15%	0.308	0.235	0.077	0.468	**0.321**	**0.247**	**0.116**	**0.645**
	15%	5%	0.245	0.181	0.065	0.101	**0.246**	**0.182**	**0.065**	**0.245**
		10%	0.269	0.202	0.075	0.121	**0.270**	**0.202**	**0.076**	**0.318**
		15%	0.251	0.186	0.068	0.292	**0.253**	**0.187**	**0.068**	**0.356**

From the previous discussion, it can be concluded that the association of the robust NR methods with the controlled sampling approach (block bootstrap) has provided a more reliable and efficient approach in dealing with time series datasets, especially the datasets of poor-quality (contains large number of outliers). The application of robust estimation methods has successfully reduced the effect of outliers on the estimated values to produced accurate estimation results. In comparison to the MI-blockboot approach, MI-boot has been found less efficient and more sensitive in dealing with time series dataset, especially for the poor-quality dataset. Therefore, it can be concluded that the controlled sampling approach was able to provide information with more accuracy on the original structure of rainfall time series.

Table 5 presents the most appropriate NR method to be implemented through the MI approach for each target station and type of dataset. The MI-blockboot's ONR was recommended as the best method to be executed for Dataset 1 (dataset that was not affected by outliers) while the MI-blockboot's NRGMED was suggested for Dataset 2 (dataset that was contaminated with outliers) for all levels of outliers. The suggested methods are proposed for imputation of missing rainfall data for both Johor Bahru and Lalang Sg Lui stations.

Table 5
Summary of the best estimation methods for Dataset 1 and Dataset 2

Target station	Best estimation method (MI approach)			
	Dataset 1	Dataset 2		
		5%	10%	15%
Johor Bahru	ONR (MI-blackboot)	NRGMED (MI-blackboot)	NRGMED (MI-blackboot)	NRGMED (MI-blackboot)
Lalang Sg Lui	ONR (MI-blackboot)	NRGMED (MI-blackboot)	NRGMED (MI-blackboot)	NRGMED (MI-blackboot)

4. Conclusions

The original structure of a dataset has to be inevitably considered as it provides crucial information on its characteristics, especially for time series data. Thus, the consideration of block bootstrap (controlled sampling) in the development of multiple imputation was a special algorithm created for rainfall time series. The block bootstrap ensured the original rainfall time series structure was preserved within each monsoon block and consequently improved the accuracy of the estimation results. These findings have successfully improved the limitation of the existing algorithm (i.e. general bootstrap) in Amelia which does not consider the structure of time series data, especially when dealing with the dataset that contains outliers (Dataset 2).

The association with the robust estimation methods has revealed its strength in this study. The advantage of MI-blockboot was evidenced over the MI-boot when dealing with the data that contains outliers. The involvement of robust NR methods, i.e. NRTR, NRMED, and NRGMED has increased its robustification to produce more accurate results. In conclusion, the introduction of controlled sampling in replacing the conventional sampling has succeeded towards improving the performance of the MI approach. Thus, the MI-blockboot was suggested for environmentalist as an alternative approach for better estimation of environmental datasets, particularly for rainfall data characterized by seasonal patterns.

Acknowledgement

The authors are indebted and thankful to the staff of the Drainage and Irrigation Department and the Malaysian Meteorological Department for providing the daily rainfall data for this study. This research would not have been possible without the sponsorship of the Ministry of Higher Education and also Universiti Teknologi MARA (UiTM), Malaysia. The authors also acknowledge their sincere appreciation to the reviewers for their valuable suggestions and remarks which improved the manuscript. This research was funded by Malaysian Fundamental Research Grant (FRGS/1/2014/ST06/UITM/02/6).

References

Acuña, E., & Rodriguez, C. (2004). The treatment of missing values and its effect on classifier accuracy. In D. Banks, F. R. McMorris, P. Arabie, & W. Gaul (Eds.), *Classification, clustering, and data mining applications* (pp. 639–647). Berlin, Heidelberg: Springer Berlin Heidelberg.

Aydilek, I. B., & Arslan, A. (2013). A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm. *Information Sciences*, *233*, 25–35. https://doi.org/10.1016/j.ins.2013.01.021

Bland, J. M., & Altman, D. G. (2015). Statistics notes: Bootstrap resampling methods. *BMJ (Online)*, *350*, *h2622*, 1–2. https://doi.org/10.1136/bmj.h2622

Burhanuddin, S. N. Z. A., Deni, S. M., & Ramli, N. M. (2017a). Imputation of missing rainfall data using revised normal ratio method. *Advanced Science Letters*, *23*(11), 10981–10985. https://doi.org/10.1166/asl.2017.10203

Burhanuddin, S. N. Z. A., Deni, S. M., & Ramli, N. M. (2017b). Normal Ratio in Multiple Imputation Based on Bootstrapped Sample for Rainfall Data with Missingness. *International Journal of GEOMATE*, *13*(36), 131–137. https://dx.doi.org/10.21660/2017.36.2760

Cardenas, R. R., Krainski, E. T., & Costa, M. A. (2009). Imports of climatic data and agricultural productivity: A comparison of approaches. In *1st PROCAD Project Workshop: Agricultural Insurance: Aesthetic Modeling and Precification*.

Carlstein, E. (1986). The use of subseries values for estimating the variance of a general statistic from a stationary sequence. *The Annals of Statistics*, *14*(3), 1171–1179. https://doi.org/10.1214/aos/1176350057

Chen, L., Xu, J., Wang, G., & Shen, Z. (2019). Comparison of the multiple imputation approaches for imputing rainfall data series and their applications to watershed models. *Journal of Hydrology*, *572*(March), 449–460. https://doi.org/10.1016/j.jhydrol.2019.03.025

Chernick, M., & LaBudde, R. (2011). *An introduction to bootstrap methods with applications to R*. United States of America: John Wiley & Sons, Inc. https://doi.org/10.1111/insr.12011_8

Chuan, Z. L., Deni, S. M., Fam, S.-F., & Ismail, N. (2019). The Effectiveness of a Probabilistic Principal Component Analysis Model and Expectation Maximisation Algorithm in Treating Missing Daily Rainfall Data. *Asia-Pacific Journal of Atmospheric Sciences*, *54*(S), 1–11. https://doi.org/10.1007/s13143-019-00135-8

De Carvalho, J. R. P., Almeida Monteiro, J. E. B., Nakai, A. M., & Assad, E. D. (2017). Modelo de imputação múltipla para estimar dados de precipitação diária e preenchimento de falhas. *Revista Brasileira de Meteorologia*, *32*(4), 575–583. https://doi.org/10.1590/0102-7786324006

De Carvalho, J. R. P., Nakai, A. M., & Monteiro, J. E. B. A. (2016). Spatio-Temporal Modeling of Data Imputation for Daily Rainfall Series in Homogeneous Zones. *Revista Brasileira de Meteorologia*, *31*(2), 196–201. https://doi.org/10.1590/0102-778631220150025

Deni, S., Jamaludin, S., Zin, W., & Jemain, A. (2009). Trends of wet spells over peninsular Malaysia during monsoon seasons. *Sains Malaysiana*, *38*(2), 133–142. Retrieved from https://eprints.utm.my/13159/

Di Piazza, A., Conti, F. Lo, Noto, L. V., Viola, F., & La Loggia, G. (2011). Comparative analysis of different techniques for spatial interpolation of rainfall data to create a serially complete monthly time series of precipitation for Sicily, Italy. *International Journal of Applied Earth Observation and Geoinformation*, *13*(3), 396–408. https://doi.org/10.1016/j.jag.2011.01.005

Donders, A. R., van der Heijden, G. J., Stijnen, T., & Moons, K. G. (2006). Review: a gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology*, *59*(10), 1087–1091. https://doi.org/10.1016/j.jclinepi.2006.01.014

Ekeu-Wei, I. T., Blackburn, G. A., & Pedruco, P. (2018). Infilling missing data in hydrology: Solutions using satellite radar altimetry and multiple imputation for data-sparse regions. *Water (Switzerland)*, *10*(1483), 1–22. https://doi.org/10.3390/w10101483

Enders, C. K. (2010). *Applied missing data analysis*. New York, London: The Guilford Press.

Farhangfar, A., Kurgan, L. a., & Pedrycz, W. (2004). Experimental analysis of methods for imputation of missing values in databases. In *Intelligent Computing: Theory and Applications II* (pp. 172–182). International Society for Optics and Photonics. https://doi.org/10.1117/12.542509

Fielding, S., Fayers, P. M., & Ramsay, C. R. (2009). Investigating the missing data mechanism in quality of life outcomes: a comparison of approaches. *Health and Quality of Life Outcomes*, *7*(57), 1–10. https://doi.org/10.1186/1477-7525-7-57

Fletcher, P. T., Venkatasubramanian, S., & Joshi, S. (2009). The geometric median on Riemannian manifolds with application to robust atlas estimation. *NeuroImage*, *45*, 143–152. https://doi.org/10.1016/j.neuroimage.2008.10.052

Fukuda, K., & Rosta, V. (2005). *Data depth and optimization*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.487.187&rep=rep1&type=pdf

Hanaish, I. S., Ibrahim, K., & Jemain, A. A. (2013). On the applicability of Bartlett Lewis Model: With reference to missing data. *Matematika*, *29*(1b), 53–65. Retrieved from https://matematika.utm.my/index.php/matematika/article/view/359

Honaker, J., King, G., & Blackwell, M. (2018). AMELIA II: A Program for Missing Data. Version 1.7.5, 54.

Ingsrisawang, L., & Potawee, D. (2012). Multiple imputation for missing data in repeated measurements using MCMC and copulas. In *International MultiConference of Engineers and Computer Scientist* (Vol. II, pp. 1–5). Retrieved from https://core.ac.uk/download/pdf/25758739.pdf

Inoue, A., & Shintani, M. (2006). Bootstrapping GMM estimators for time series. *Journal of Econometrics*, *133*(2), 531–555. https://dx.doi.org/10.1016/j.jeconom.2005.06.004

Jahan, F., Sinha, N. C., Rahman, M. M., Rahman, M. M., Mondal, M. S. H., & Islam, M. A. (2018). Comparison of missing value estimation techniques in rainfall data of Bangladesh. *Theoretical and Applied Climatology*, 1–17. https://doi.org/10.1007/s00704-018-2537-y

Jakhar, Y. K., Mishra, N., & Poonia, R. (2018). Predication accuracy analysis of data mining algorithms on meteorological data using R programming. In *3rd International Conference on Internet of Things and Connected Technologies (ICIoTCT)* (pp. 107–110). Elsevier-SSRN. https://doi.org/10.2139/ssrn.3166223

Jamaludin, S., Deni, S. M., & Jemain, A. A. (2008). Revised spatial weighting methods for estimation of missing rainfall data. *Asia-Pacific Journal of Atmospheric Sciences*, *44*(2), 93–104. Retrieved from https://people.utm.my/shariffah/files/2015/03/7451_APJASApril2008.pdf

Jamaludin, S., & Jemain, A. A. (2009). Investigating the impacts of adjoining wet days on the distribution

of daily rainfall amounts in Peninsular Malaysia. *Journal of Hydrology*, *368*(1–4), 17–25. https://doi.org/10.1016/j.jhydrol.2009.01.022

Jamaludin, Suhaila, Deni, S., Zin, W., & Jemain, A. (2010). Trends in peninsular Malaysia rainfall data during the Southwest Monsoon and Northeast Monsoon Seasons: 1975–2004. *Sains Malaysiana*, *39*(4), 533–542. Retrieved from http://www.ukm.my/jsm/pdf_files/SM-PDF-39-4-2010/03 Jamaludin Suhaila.pdf

Kalteh, A. M., & Hjorth, P. (2009). Imputation of missing values in a precipitation–runoff process database. *Hydrology Research*, *40*(4), 420. https://doi.org/10.2166/nh.2009.001

Khalifeloo, M. H., Mohammad, M., & Heydari, M. (2015). Multiple Imputation for Hydrological Missing Data By Using a Regression Method (Klang River Basin). *International Journal of Research in Engineering and Technology*, *04*(06), 519–524. https://doi.org/10.15623/ijret.2015.0406090

Khamkong, M., Bookkamana, P., Shin, Y., & Park, J. (2017). Modelling extreme rainfall in northern Thailand with estimated missing values. *Chiang Mai Journal of Sciences*, *44*(4), 1792–1804. Retrieved from http://www.thaiscience.info/journals/Article/CMJS/10987623.pdf

Kunsch, H. R. (1989). The jackknife and bootstrap for general stationary observations. *The Annals of Statistics*, *17*(3), 1217–1241. https://doi.org/10.1214/aos/1176348654

Laurikkala, J., Juhola, M., & Kentala, E. (2000). Informal Identification of Outliers in Medical Data. In *Fifth international workshop on intelligent data analysis in medicine and pharmacology* (pp. 20–24). Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.37.4330&rep=rep1&type=pdf#page=24

Li, J. (2006). The block bootstrap test of Hausman's exogeneity in the presence of serial correlation. *Economics Letters*, *91*(1), 76–82. https://dx.doi.org/10.1016/j.econlet.2005.11.001

Lo Presti, R., Barca, E., & Passarella, G. (2010). A methodology for treating missing data applied to daily rainfall data in the Candelaro River Basin (Italy). *Environmental Monitoring and Assessment*, *160*, 1–22. https://doi.org/10.1007/s10661-008-0653-3

Mair, A., & Fares, A. (2010). Assessing rainfall data homogeneity and estimating missing records in Mākaha valley, O 'ahu, Hawai 'i. *Journal of Hydrologic Engineering*, *15*, 61–66. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000145

Milo, E., Ekonomi, L., Margo, L., & Donefski, E. (2019). Seasonal means estimation and missing data in real data time series. *Applied Mathematical Sciences*, *13*(1), 25–32. https://doi.org/10.12988/ams.2019.812192

Miró, J. J., Caselles, V., & Estrela, M. J. (2017). Multiple imputation of rainfall missing data in the Iberian Mediterranean context. *Atmospheric Research*, *197*(February), 313–330. https://doi.org/10.1016/j.atmosres.2017.07.016

Paulhus, J. L. H., & Kohler, M. A. (1952). Interpolation of missing precipitation records. *Monthly Weather Review*, *80*(8), 129–133. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.394.6283&rep=rep1&type=pdf

Radi, N. F. A., Zakaria, R., & Azman, M. A. Z. (2015). *Estimation of missing rainfall data using spatial interpolation and imputation methods*. *The 2nd ISM International Statistical Conference 2014 (ISM-II): Empowering the Applications of Statistical and Mathematical Sciences*. Pahang: AIP Publishing. https://doi.org/10.1063/1.4907423

Ramos-Calzado, P., Gómez-Camacho, J., Pérez-Bernal, F., & Pita-López, M. F. (2008). A novel approach to precipitation series completion in climatological datasets: application to Andalusia. *International Journal of Climatology*, *28*, 1525–1534. https://doi.org/10.1002/joc.1657

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. *John Wiley & Sons*.

https://doi.org/10.1002/9780470316696

Saeed, G. A. A., Chuan, Z. L., Zakaria, R., Yusoff, W. N. S. W., & Salleh, M. Z. (2016). Determination of the best single imputation algorithm for missing rainfall data treatment. *Journal of Quality Measurement and Analysis*, *12*(1–2), 79–87. Retrieved from https://core.ac.uk/download/pdf/84306717.pdf

Sattari, M.-T., Rezazadeh-Joudi, A., & Kusiak, A. (2017). Assessment of different methods for estimation of missing data in precipitation studies. *Hydrology Research*, *48*(4), 1032–1044. https://doi.org/10.2166/nh.2016.364

Tax, D. M. J., & Duin, R. P. W. (2001). Uniform object generation for optimizing one-class classifiers. *Journal of Machine Learning Research*, *2*, 155–173. https://doi.org/10.1162/15324430260185583

Twumasi-Ankrah, S., Odoi, B., Pels, W. A., & Gyamfi, E. H. (2019). Efficiency of Imputation Techniques in Univatiate Time Series. *International Journal of Science, Environment, and Technology*, *8*(3), 430–453. https://www.researchgate.net/publication/333561806

Yendra, R., & Jemain, A. (2013). Methods on handling missing rainfall data with Neyman-Scott rectangular pulse modeling. In *Proceedings of the 20th National Symposium on Mathematical Sciences* (Vol. 1522, pp. 1213–1220). https://doi.org/10.1063/1.4801269

Yozgatligil, C., Aslan, S., Iyigun, C., & Batmaz, I. (2013). Comparison of missing value imputation methods in time series: the case of Turkish meteorological data. *Theoretical and Applied Climatology*, *112*(1–2), 143–167. https://doi.org/10.1007/s00704-012-0723-x

Yunus, F., Shafie, A., Jaafar, J., & Mahmud, Z. (2011). Homogeneous climate divisions for Peninsular Malaysia. *Geodinamica Acta*, *24*(2), 89–94. https://doi.org/10.3166/ga.24.89-94