International Journal of GEOMATE, Aug, 2017, Vol.13, Issue 36, pp.131-137
Geotec., Const. Mat. & Env., ISSN:2186-2990, Japan, DOI: http://dx.doi.org/10.21660/2017.36.2760

NORMAL RATIO IN MULTIPLE IMPUTATION BASED ON BOOTSTRAPPED SAMPLE FOR RAINFALL DATA WITH MISSINGNESS

*Siti Nur Zahrah Amin Burhanuddin[1], Sayang Mohd Deni[2] and Norazan Mohamed Ramli[3]

[1,2,3] Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia

* Corresponding Author, Received: 01 July 2016, Revised: 09 Aug 2016, Accepted: 29 Nov 2016

ABSTRACT: The existence of missing values in rainfall data series is inevitably affects the quality of the data. This problem will influence the results of analysis and subsequently provide imprecise information to the hydrological and meteorological management. A practical and reliable approach is needed in developing estimation methods to impute the missing values. Single imputation is the most commonly used approach for missing values, but, it encounters with the limitation of not considering the uncertainty and natural variability in missing data imputation. Thus, this study has proposed multiple imputation approach based on bootstrap samples in order to overcome the limitation of single imputation approach. Three normal ratio estimation methods are implemented using the proposed approach. The performances of the estimation methods are evaluated at six different levels of missingness. Complete 40 years daily rainfall data from four meteorology stations were considered for the analysis purpose with Johor Bahru station was selected as the target station. The results of the proposed approach were compared to the results obtained from single imputation approach and the widely known built in software for multiple imputation, Amelia II package, in assessing the performance of proposed approach. The results showed that all estimation methods that implemented using proposed approach provided the most accurate estimation results at all percentages of missingness. This proves the advantage of adaption of variability and uncertainty element in the proposed approach in estimating the missing rainfall data at the area of the current study.

Keywords: Missing Rainfall Data, Normal Ratio, Multiple Imputation, Bootstrap

1. INTRODUCTION

Complete rainfall dataset is highly necessary to the effective hydrological and meteorological analyses. However, the presence of missing data in the rainfall dataset is inevitable [1]. This is due to several factors such as relocation of stations and faulty instruments. This problem will influence the accuracy of the analysis results and subsequently provide inaccurate information to the management and development of hydrology and meteorology.

Concerning this situation, various estimation methods have been explored in treating the missing values in rainfall time series, i.e. normal ratio, inverse distance weighting, multiple linear regression, and kriging. However, normal ratio (NR) has appeared to be the most commonly used method in estimating missing rainfall values as stated in literature [2]–[4] due to its simplicity and efficiency.

Reference [7] introduced the application of NR method in estimating missing rainfall records. The NR method was then modified by [8] and [9] through the adaption of correlation coefficients and the effect of distance in the original version of NR method to improve its performance. The development of the NR method is continually being explored and is recently being studied by [4], [5], and [6]. Reference [5] and [6] compared NR method to their more sophisticated proposed methods to estimate the missing values in monthly meteorological data. They have discovered that the proposed method produced more accurate results compared to the old NR method.

Due to its simplicity, NR method is considered as an evergreen method in imputing missing rainfall data. However, the limitations of the NR method are disclosed from its applications in missing data imputation. The implementation of NR method in the previous studies is only through the single imputation (SI) approach. This approach is commonly known with the limitation of not accurately represents the variability of missing data and the uncertainty of imputed values [10]. Therefore, in order to overcome the limitation of SI approach, multiple imputation approach is introduced by [11]. Multiple imputation (MI) is one of the advanced approaches in imputing the missing rainfall data.

MI is an approach that handles missing data in a way to produce a valid statistical inference instead of estimate the missing values as close as possible to the observed ones [12]. The approach has proven to

be a powerful tool in studies conducted by [13] and [5]. [5] were successful in developing such estimation methods based on MI approach called EM-MCMC. Although the method is computationally inefficient, it has produced more robust results compared to the SI based methods. Reference [13] applied four stages of imputation approach (based on MI approach) in estimating the missing rainfall values at Paya Kangsar, Malaysia, and proved that the MI performed better than the SI.

Recently, various built-in packages are created for implementing the MI, hence, Amelia II package is the one that commonly used in estimating the missing rainfall data. The package is a bootstrapping-based algorithm that estimates the statistics by applying the expectation maximization method. The method imputes each missing value *m* times (*m*=5 is the program default), that created five completed datasets. The datasets can be straightly used for analysis purpose. General bootstrap used in this package is not suitable to be used for time series data since it does not preserve the original time series structure. Therefore, moving block bootstrapping (MBB) is considered in this study to improve the accuracy of bootstrap for time series data. MBB divides the data into several blocks and samples the whole blocks before concatenate them. The dependency structure of the time series was preserved within each block [14].

Accordingly, this study is aimed to propose multiple imputation approach onto the NR method, which will result in more accurate estimations of missing rainfall values by considering the variability and uncertainty of imputation at the same time. This effort is of the consideration in providing a good quality dataset to be used for public domain.

The remaining of this paper are organized as follows: Materials and Methods part describes the data preparation and the proposed multiple imputation approach with the performance criteria used in evaluating the performance of the estimation methods. Results and Discussions part presents the results of the study with a comparative evaluation of the method's performance followed by the conclusions.

2. MATERIALS AND METHODS

2.1 Data Preparation

This study was performed for the southern region of Peninsular Malaysia. The application of particle swarm optimization (PSO) approach to determine the optimal number and locations for the optimal rain gauge network in these areas had been proposed by [15]. However, four rainfall measuring stations including the target and neighboring stations were selected from the state of Johor for evaluation purposes. The name of the stations with their respective geographical coordinates and spatial and descriptive information are listed in Table 1.

Table 1 List of the selected stations with their geographical coordinates and spatial and descriptive information

Name of Station	Latitude	Longitude	Euclidean Distance (km)	Mean	Standard Deviation	Max Value
Johor Bahru	**103.75**	**1.47**	**0 (0)**	**6.635**	**10.43**	**285.4**
Sek. Men. Bkt. Besar	103.72	1.76	0.29 (32)	5.369	8.806	298
Kuala Sedili	103.97	1.85	0.44 (49)	7.067	10.369	397.5
Jln. Kluang/Mersing	103.74	2.26	0.79 (88)	6.431	10.265	305

The Johor Bahru station is considered as the target station (station in bold). In estimating the missing values of the target stations, rainfall data from their surrounding stations are also necessary. Data from stations that are closer to the target station tend to share similar characteristics with the target station, which are definitely valuable in providing more accurate estimation results [4]. Thus, the stations within the radius of 100 km to the target stations were all considered as neighboring stations (see Table 1).

The complete daily rainfall records from these stations were obtained from the Malaysian Drainage and Irrigation Department (DID) for the analysis of this study. The data consists of the daily rainfall amounts for the period of 40 years from January 1, 1975 to December 31, 2014.

2.2 Estimation Methods for Missing Rainfall Values

Three normal ratio methods, i.e. old normal ratio (ONR), normal ratio based on trimmed mean (NRTR), and normal ratio based on geometric median (NRGMED) are considered to be implemented using the multiple imputation (MI) approach proposed in this study. The application of these methods through the MI approach is rarely found in the previous studies. Generally, the NR methods are implemented using SI approach by most of studies, for example, [4], [5], and [6].

The old normal ratio (ONR) method was firstly introduced by [7] in estimating rainfall missing

values. It is based on the mean ratio of data between the target station and the neighboring stations. The NR method is further expressed as follows:

$$\hat{Y} = \frac{1}{N}\sum_{\substack{i=1 \\ i\neq t}}^{N}\left(\frac{\mu_t}{\mu_i}\right)Y_i \quad (1)$$

where; μ_t is the arithmetic mean of the available data at target station t; μ_i is the arithmetic mean of the available data at i^{th} neighboring station; $\hat{Y}$ is the estimated missing data at target station t; Y_i is the concurrently observed data at the i^{th} neighboring station; and N is the number of neighboring stations.

Normal ratio based on trimmed mean (NRTR) method is the modified version of ONR which considering the trimmed mean as the weighting factor. Trimmed mean reduced the effect of outliers on the calculated average. Several levels of trimming (1%, 5%, and 10%) considered to assess the consistency of the estimation results. The application of the trimmed mean in the estimation methods can produce more accurate estimation results. The NRTR method is defined as follows:

$$\hat{Y} = \frac{1}{N}\sum_{\substack{i=1 \\ i\neq t}}^{N}\left(\frac{\mu_{trim_t}}{\mu_{trim_i}}\right)Y_i \quad (2)$$

where; μ_{trim_t} is the trimmed mean of the available data at target station t; μ_{trim_i} is the trimmed mean of the available data at i^{th} neighboring station.

The last method is normal ratio based on geometric median (NRGMED). Geometric median (Gmed) is a robust estimator of centrality in Euclidean spaces [16]. Gmed of a dataset is the data minimizing the sum of distances to the sample dataset. It is defined as a robust version of geometric mean that is more stable in the presence of outliers [15]. The method is expressed as follows:

$$\hat{Y} = \frac{1}{N}\sum_{\substack{i=1 \\ i\neq t}}^{N}\left(\frac{Gmed_t}{Gmed_i}\right)Y_i \quad (3)$$

Gmed is defined as: $Gmed_i = e^{(\text{median}(\ln(y_i)))}$

where; $Gmed_t$ is the geometric median of the available data at target station t; $Gmed_i$ is the geometric median of the available data at i^{th} neighboring station.

2.3 Proposed Multiple Imputation Using Bootstrapped Samples

Multiple imputation approach is proposed for the implementation of the estimation methods explained in the previous. Generally, MI consists of three consecutive phases; (i) imputation phase, (ii) analysis phase, and (iii) pooling phase [12]. Imputation phase is the most complicated phase which involves the process of estimating the missing values as in the SI approach. However, in MI, the imputation process is simulated m times (Reference [11] suggested 3 to 5 times) producing m estimated values for each individual missing values, consequently resulting in m different imputed datasets. The imputed datasets are then going through the process of analysis, which is the second phase. Several sets of parameter estimate and standard errors are produced before being pooled into single results through the last phase

The implementation of MI approach proposed in this study is quite different from the other studies. The execution has adapted the concept used in Amelia II package which involves bootstrapping. Moving block bootstrapping (MBB) is applied to rainfall time series in preserving the original time series structure. This effort is expected to produce more accurate missing values estimation results due to the variability and uncertainty inherent in the proposed approach algorithm. Figure 2 shows the diagram of the proposed MI approach.

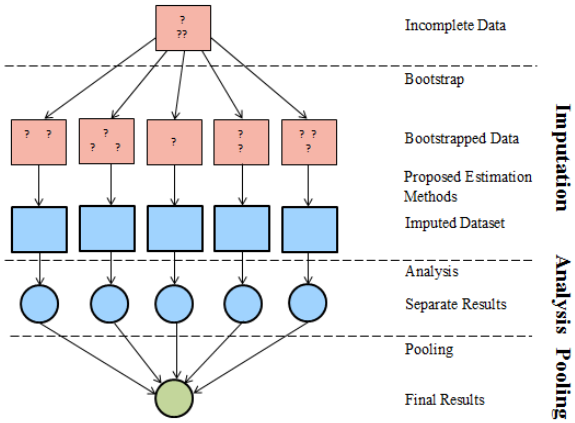

Fig. 1 A schematic of proposed approach to MI

Five bootstrapped samples are created to produce five sets of estimated values for each missing dataset. The bootstrapped sample are generated based on the procedure of MBB proposed in this study. The procedure of MBB in this study is a bit different from other studies. Rainfall time series of the target station are divided into several parts according to the

characteristics of rainfall data, such as the seasonal pattern and the rainfall amounts before going through the MBB process. The MBB is applied separately on each part of the data. The procedure is as follows:

i. Divide rainfall time series into four parts (blocks) according to monsoon seasons, 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 100 times (the size of bootstrap sample is the sample size of the original data). Then merge the blocks of bootstrapped data to produce a year daily rainfall time series.

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

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

New Sample
x_n
$n = 1, 2, 3, \ldots, 14610$

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

The new samples produced are used in the proposed MI approach (NRMI-boot). The NRMI-boot is implemented based on the following procedures:

a) The same procedures of introducing missing data in the target station used in the SI approach are applied in NRMI-boot.
b) The ratio means of each new sample of the target station and the data of its nearby stations are considered as the weighting factor for the estimation methods (an example of ONR method - refer to Eq. (1)).

$$w_{L_i} = \frac{\mu_{L_i}}{\mu_i}, \quad L = \mathbf{1}, \mathbf{2}, \ldots, \mathbf{5} \tag{4}$$

where, w_{L_i} is the weight of the i^{th} neighboring station for bootstrapped sample L; μ_{L_i} and μ_i the sample mean of the available data for bootstrapped sample L and i^{th} neighboring station respectively.

c) Five weighting factors are produced to yield five different estimated values for each individual missing value.
d) Each estimated values is then used to impute the missing values in the original rainfall dataset, in which producing five complete datasets with different imputed values and the same observed values.
e) The complete imputed datasets are analyzed and the results of analysis are pooled to produce a single result of analysis.

The results obtained from the proposed NRMI-boot were compared to the results of SI approach and the results obtained from Amelia II package. The performance of the considered NR methods is evaluated based on the implementation of these three approaches. Two error measurements are considered in evaluating the performance of the estimation methods, namely root mean square error (RMSE) and similarity index (SIndex). They are defined as follows:

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

$$\text{SIndex} = \frac{\sum_{i=1}^{n}\left(\hat{Y}_i - Y_i\right)^2}{\sum_{i=1}^{n}\left(\left|\hat{Y}_i - \bar{Y}\right| + \left|Y_i - \bar{Y}\right|\right)^2} \tag{6}$$

where; $\hat{Y}_i$ is the estimated value; Y_i is the actual value of the observation; $\bar{Y}$ is the mean of actual values; n is the number of observations.

3. RESULTS AND DISCUSSIONS

Six levels of missingness ranges from 5% to 30% are used in assessing the performance consistency of estimation methods. The performance of estimation methods is investigated through their implementation using SI and MI approaches. The results of the NRMI-boot are compared to the SI at

various level of missingness and displayed in Table 3. The results of estimation methods using NRMI-boot produced slight improvements compared to the SI approach based on the least RMSE and the highest SIndex (see the values in bold). It also produced better estimation results than SI at varying percentages of missingness. This shows that the variability and uncertainty accounts in the NRMI-boot produced more accurate results of estimation.

Of the five estimation methods, ONR appears as the best method in estimating the missing rainfall values, whereas NRGMED is the worst. The NRTR method was divided into three based on different levels of trimming, i.e. NRTR1 (1% trimming), NRTR5 (5% trimming), NRTR10 (10% trimming), in order to investigate the effects of the trimming levels of the mean on the accuracy of the estimation results. The accuracy of the estimation results using NRTR decrease with the increase of the trimming percentage.

Table 2 Performance of NR methods through single imputation (SI) and proposed NRMI-boot approaches at various level of missingness

Level of Missingness	Estimation Method	RMSE		SIndex	
		SI	NRMI-boot	SI	NRMI-boot
5%	ONR	14.869	**13.952**	0.648	**0.688**
	NRTR1	15.093	**14.174**	0.642	**0.682**
	NRTR5	15.340	**14.421**	0.635	**0.675**
	NRTR10	15.442	**14.540**	0.632	**0.672**
	NRGMED	15.633	**14.871**	0.627	**0.662**
10%	ONR	14.744	**13.672**	0.643	**0.682**
	NRTR1	14.985	**13.860**	0.637	**0.676**
	NRTR5	15.224	**14.060**	0.630	**0.671**
	NRTR10	15.322	**14.154**	0.627	**0.668**
	NRGMED	15.441	**14.400**	0.624	**0.661**
15%	ONR	14.182	**13.553**	0.646	**0.669**
	NRTR1	14.417	**13.752**	0.639	**0.664**
	NRTR5	14.642	**13.959**	0.632	**0.657**
	NRTR10	14.729	**14.053**	0.630	**0.655**
	NRGMED	14.870	**14.247**	0.626	**0.649**
20%	ONR	14.319	**14.296**	0.662	**0.656**
	NRTR1	14.538	**14.515**	0.656	**0.650**
	NRTR5	14.730	**14.732**	0.650	**0.644**
	NRTR10	14.806	**14.831**	0.648	**0.641**
	NRGMED	14.884	**15.051**	0.646	**0.635**
25%	ONR	14.610	**14.391**	0.660	**0.655**
	NRTR1	14.849	**14.621**	0.654	**0.648**
	NRTR5	15.055	**14.829**	0.649	**0.643**
	NRTR10	15.135	**14.920**	0.646	**0.640**
	NRGMED	15.333	**15.106**	0.641	**0.634**
30%	ONR	14.584	**14.157**	0.667	**0.664**
	NRTR1	14.821	**14.381**	0.660	**0.658**
	NRTR5	14.984	**14.571**	0.656	**0.652**
	NRTR10	15.044	**14.652**	0.654	**0.650**
	NRGMED	15.223	**14.892**	0.649	**0.643**

The next comparison involved the results of the best estimation method using the proposed NRMI-

boot, i.e. ONR method with the results obtained from the Amelia II package. Table 4 presents the results comparison of these MI approaches. The method in bold represented the most appropriate method in estimating the missing values. Based on the comparison, it can be seen that the results obtained from NRMI-boot gives more accurate results compared to the Amelia results. This may due to the moving block bootstrapping adapted in the NRMI-boot, in which it preserved the original time series structure in imputing the missing rainfall values.

Table 3 Performance of NR methods through the NRMI-boot and the Amelia II package results at various level of missingness.

Error Measures	Level of Missingness	Amelia	NRMI-boot
RMSE	5%	14.8312	**13.952**
	10%	15.3311	**13.672**
	15%	15.0944	**13.553**
	20%	15.8401	**14.296**
	25%	15.8237	**14.391**
	30%	15.9049	**14.157**
SIndex	5%	0.5795	**0.688**
	10%	0.6406	**0.682**
	15%	0.6612	**0.669**
	20%	0.6667	**0.656**
	25%	0.6691	**0.655**
	30%	0.6708	**0.664**

4. CONCLUSIONS

The implementation of estimation methods through the NRMI-boot produced the most accurate results among the SI approach and Amelia II package in estimating the missing rainfall values. This shows the advantage of the proposed approach that considering the variability in creating the missing values and uncertainty in estimating the imputed values. The involvement of these two elements has successfully improved the estimation results' accuracy for the area of the current study. Furthermore, the adaption of MBB gives advantage to the NRMI-boot and is more suitable when dealing with the time series data compare to the general bootstrapping adapted in Amelia II package.

Although the computation of the proposed approach is quite intensive compared to the SI approach, it is recommended to be applied in any studies related to missing values since it provided a robust approach and more accurate estimation results. For future research, the degree of suitability of the proposed MI approach (NRMI-boot) towards other climatic variables (e.g. temperature and wind speed) and time scales (e.g. monthly and yearly) needs to be investigated. Other suggestions are to consider other estimation methods such as inverse distance weighting and geographical coordinate methods in estimating the missing values by using the proposed approach.

5. ACKNOWLEDGEMENTS

The authors are indebted and thankful to the staff of Malaysian Meteorological Department, and Drainage and Irrigation Department for providing the daily rainfall data used in this study. This research would not have been possible without the sponsorships from Ministry of Higher Education and also Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Malaysia. This research is funded by the Malaysian Fundamental Research Grant, FRGS/1/2014/ST06/UITM/02/6.

6. REFERENCES

[1] Firat M, Dikbas F, Koc AC, and Gungor M, "Missing data analysis and homogeneity test for Turkish", Sadhana, Vol. 35, 2010, pp. 707–720.

[2] De Silva RP, Dayawansa NDK, and Ratnasiri MD, "A Comparison of methods used in estimating missing rainfall data", J. Agric. Sci., Vol. 3, 2007, pp. 101–108.

[3] Khorsandi Z, Mahdavi M, Salajeghe A, and Eslamian S, "Neural network application for monthly precipitation data reconstruction", J. Environ. Hydrol., Vol. 19, 2011, pp. 1–12.

[4] Suhaila J, Deni SM, and Jemain AA, "Revised spatial weighting methods for estimation of missing rainfall data", Asia-Pacific J. Atmos. Sc., Vol. 44, 2008, pp. 93–104.

[5] Yozgatligil C, Aslan S, Iyigun C, and Batmaz I, "Comparison of missing value imputation methods in time series: the case of Turkish meteorological data", Theor. Appl. Climatol., Vol. 112, Jul. 2013, pp. 143–167.

[6] Khosravi G, Nafarzadegan AR, Nohegar A, Fathizadeh H, and Malekian A, "A modified distance-weighted approach for filling annual precipitation gaps: application to different climates of Iran", Theor. Appl. Climatol., Jan. 2014, pp. 1–10.

[7] Paulhus JLH and Kohler MH, "Interpolation of missing precipitation records", Mon. Wea. Rev, Vol. 80, 1952, pp. 129–133.

[8] Young KC, "A three-way model for interpolating for monthly precipitation values", Mon. Weather Rev., Vol. 120, 1992, pp. 2561–2569.

[9] Tang WY, Kassim AHM, and Abubakar SH, "Comparative studies of various missing data treatment methods -Malaysian experience", Atmos. Res., Vol. 42, 1996, pp. 247–262.
[10] Donders ART, Van Der Heijden GJMG, Stijnen R, and Moons KGM, "Review: a gentle introduction to imputation of missing values", J. Clin. Epidemiol., Vol. 59, Oct. 2006, pp. 1087–91.
[11] Rubin DB, "Multiple Imputation After 18 + Years", J. Am. Stat. Assoc., Vol. 91, 1996, pp. 473–489.
[12] Enders CK, "The imputation phase of multiple imputation", Applied Missing Data Analysis, New York: The Guilford Press, 2010, pp. 187-216.
[13] Yendra R and Jemain AA, "Methods on handling missing rainfall data with Neyman-Scott rectangular pulse modeling", Proc. 2nd Natl. Symp. on Math. Sci., vol. 1213, 2013, pp.1213-1220.
[14] Li J, "The block bootstrap test of Hausman", Econ. Lett., Vol. 91, 2006, pp. 76–82.
[15] Aziz, M. K. B. M., Yusof, F., Daud, Z. M., Yusop, Z., & Kasno, M. A. (2016). "Optimal design of rain gauge network in Johor by using geostatistics and particle swarm optimization. Int. J. of GEOMATE, 11(3), 2422-2428.
[16] Das KR and Imon AHMR, "Geometric median and its application in the identification of multiple outliers", J. Appl. Stat., Vol. 41, 2014, pp. 817–831.