MJFAS MALAYSIAN JOURNAL OF FUNDAMENTAL AND APPLIED SCIENCES
PRINT ISSN: 2289-5981 | ONLINE ISSN: 2289-599X

RESEARCH ARTICLE

A comparison of method for treating missing daily rainfall data in Peninsular Malaysia

Izzat Fakhruddin Kamaruzaman [a, b,*], Wan Zawiah Wan Zin [a], Noratiqah Mohd Ariff [a]

[a] School of Mathematical Sciences, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia
[b] Faculty of Business, Multimedia University, Jalan Ayer Keroh Lama, 75450 Bukit Beruang, Melaka, Malaysia

* Corresponding author: izzat.kamaruzaman@mmu.edu.my

Article history
Received 3 October 2017
Accepted 7 November 2017

Abstract

This study modified a method for treating missing values in daily rainfall data from 104 selected rainfall stations. The daily rainfall data were obtained from the Department of Irrigation and Drainage Malaysia (DID) for the periods of 1965 to 2015. The missing values throughout the 51 years period were estimated using the various types of weighting methods. In determining the best imputation method, three test for evaluating model performance has been used. The findings of this study indicate that the proposed method is more efficient than the traditional method. The homogeneity of the data series was checked using the homogeneity tests recommended by the existing literatures. The results indicated that more than 40% of the rainfall stations were homogenous based on the proposed method.

Keywords: Daily rainfall, imputation, inverse distance, homogeneity

INTRODUCTION

Daily rainfall data are one of the most important variables in hydrological and environmental modelling and also in assessing the water quality. However, studies involving the use of long and continuous time series data are always faced with the problem of missing value especially in developing countries like Malaysia. Mostly the existing data series are too short to perform a good and meaningful analyses and often contain a large number of missing values [1-3]. Normally lack of data and inhomogeneity problem are due to rainfall station relocation, changes in the environment, instrument malfunctions and network reorganizations [4]. In hydrologic modelling, developing a method to get an accurate estimation of rainfall are very crucial. In order to get an accurate results in analyses, the rainfall data that is used must be complete, homogeneous and have a good quality.

Basically, there are two ways to resolve this problem which is by using removal methods such as listwise deletion and pairwise deletion and the other method is imputation which it is divided into several sections such as single imputation, multiple imputation and iterative imputation [5]. Ad hoc methods that are commonly applied by many researchers are listwise deletion, pairwise deletion and single imputation. The listwise deletion method eliminate all information contained in the sample even though only one data is missing. This will lead to a reduction in the number of samples. Pairwise deletion method do not eliminate all of the information available in a missing data sample, but the incomplete information are excluded from the analysis. As for the single imputation method, each missing value will be filled with an appropriate value such as an average value. This can maintain the original amount of data. However, by replacing the missing value with a single value will result in the reduction of variance and next will change the shape of the distribution. According to Peugh and Enders [6], listwise deletion method and pairwise deletion method are the common method used in the treatment of missing value. However in the study practiced using time series data like hydrological data, removal methods are not suitable because it may cause the data to become discontinuous. Meanwhile, a single imputation in which each gap is filled by a single value while in the real situation hydrological data are in the random form.

In general, there are a number of methods have been proposed to estimate missing value [7-10]. The best estimation should not change the important characteristics of the dataset and should follow the character of rainfall in a given area [11]. In hydrological study, spatial correlations also exist among rainfall occurrence and amounts of studied stations. The fact is reasonable as fairly negative relationship were observed between correlation of daily rainfall and distances among the neighboring stations. Therefore it is important to include neighboring stations in estimation process.

Within station methods for estimating missing values in climate series are the easiest and simplest approaches. Eischeid *et al.* [12] suggested that the treatment of missing value can be performed by using the data station itself or stations nearby however applications using data from nearby stations are more reliable [12,13]. Among the methods that used neighbouring station data as data generation is a method that is based on the weighted method. There are many studies related to statistical analysis that have been discussed about missing value by using weighted method [1,3,9,10,14-17].

Xia *et al.* [18] using the nearest neighbour station to estimate the missing value based on a geometric weighting, Willmott et al. [19] using arithmetic averaging of data from neighbouring stations to treat the missing value and Teegavarapu and Chandramouli [16] using the inverse distance weighting method of neighbouring stations in the process of rainfall data imputation. Jemain *et al.* [11] argued that the inverse distance weighting method is the superior traditional methods

in the estimation of missing value. Improvements to inverse distance weighting method also can be traced, for example, in Teegavarapu and Chandramouli [16] and Suhaila *et al.* [20]. Young [21] and Filippini *et al.* [22] proposed the interpolation of the correlation of each station to ascertain the weighted value. The use of correlation coefficients between data series as weightage has been examined on a daily basis [16,23], and generally found to outperform distance based methods.

Prior to the imputation process, the type of mechanism missing data should be interpreted because the effectiveness of an imputation technique depends entirely on their assumptions. There are three features that missing data mechanism is often applied in previous studies, namely missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR) [24]. The term MCAR refers to data where the missingness mechanism does not depend on the variable under investigation or any other variable, which is observed in the dataset. The term MAR means data is missing, but conditioned by some other variable observed in the dataset that is other than the variable under investigation [25]. Finally, not missing at random (MNAR) occurs when the missingness mechanism depends on the actual value of the missing data. This is the most difficult condition to the development of a model. Based on the definition of Little and Rubin [26], missing value in the rainfall study is determined as MCAR because of the occurrence of missingness in the rainfall data of an area not affected by the data in that area or any area. There is also a study on rainfall data imputation using MAR assumption [27]. However, Moritz *et al.* [28] have stated that imputation MCAR and MAR for univariate time series study is similar. In this study, the mechanism of missing value has been classified as MCAR.

STUDY AREA AND DATA

The focus of this study is the state of Peninsular Malaysia which lies in the Equatorial zone of Northern latitude between 1 and 6º N and Eastern longitude from 100 to 103º E. Peninsular Malaysia experienced hot and humid weather all year round. Typically, the Malaysian climate influenced by winds blowing from the Indian Ocean which is known as Southwest Monsoon Wind occurs from May to September and the South China Sea which is the Northeast Monsoon Wind occurs from November to March. Whereas the transition period between the two monsoons is recognized as the intermonsoon periods occurring in March to April and September to October, bringing intense convective rain to many areas in the peninsula. Annual rainfall is 80% per annum between 2000mm to 2500mm.

The data used in this study can be considered good quality data with less than 10% missing values throughout the 51 years period. A large amount of time series observations are required in order to obtain an accurate overview pertaining to the pattern of the rainfall [11]. In addition, long time series data is valuable because the credibility of the frequency estimator is closely related to the size of the sample during the analysis process [29]. A 51-year record of data during the years 1965–2015 is obtained from the Department of Irrigation and Drainage Malaysia (DID) for further analysis. Table 1 and Fig. 1 show the geographical coordinates and percentages of missing observations of the 104 selected rainfall stations used to collect rainfall data.

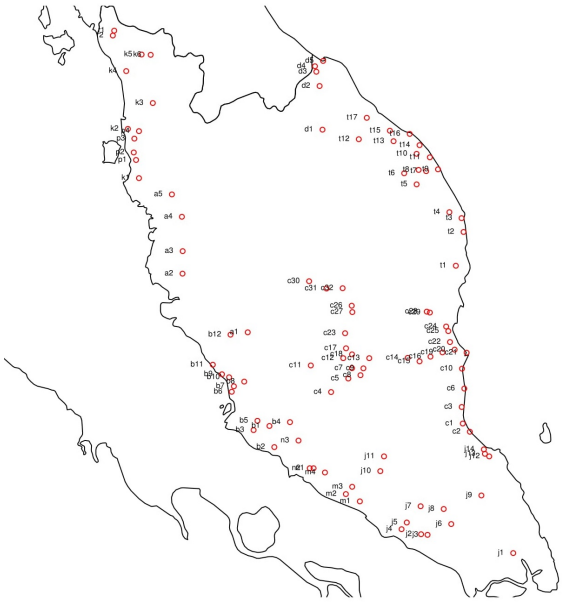

Fig. 1 The locations of the 104 rain gauged stations in Peninsular Malaysia.

Table 1 Geographic locations of the 104 selected stations in Peninsular Malaysia.

Station Name	Code	Latitude	Longitude	Missing Values (%)
Kelantan				
Ladang Kuala Nal	d1	5.5708	102.1639	2.71
Tok Uban	d2	5.9694	102.1375	0.75
Meranti	d3	6.1000	102.1083	2.93
Kuala Jambu	d4	6.1500	102.0958	2.02
Stesen Keretapi Tumpat	d5	6.1986	102.1694	1.34
Terengganu				
Kampung Ibok	t1	4.3278	103.3681	0.17
Paka	t2	4.6361	103.4375	1.10
Dungun	t3	4.7625	103.4194	0.82
Ulu Dungun	t4	4.8167	103.3097	3.56
Kuala Brang	t5	5.0708	103.0139	3.66
Paya Rapat	t6	5.1722	102.9014	3.94
Bukit Sawa	t7	5.1917	103.1000	0.35
Kuala Telemong	t8	5.2028	103.0319	0.79
Marang	t9	5.2083	103.2069	1.35
Kampong Gemuroh	t10	5.3500	103.0139	1.71
Kuala Terengganu	t11	5.3181	103.1333	4.85
Ulu Besut	t12	5.4833	102.4917	3.72
Kampung Rahmat	t13	5.4653	102.8056	4.12
Batu Rakit	t14	5.4292	103.0403	0.71
Banggol	t15	5.5611	102.7722	3.71
Setiu	t16	5.5319	102.9514	4.34
Kampung Jabi	t17	5.6792	102.5639	3.78
Pahang				
Leban Chondong	c1	2.8861	103.4306	5.99
Rompin	c2	2.8111	103.4944	4.33
Kampung Bebar	c3	3.0375	103.4208	5.57
Pelangi	c4	3.1736	102.2417	1.31
Kampung Jawi 2				
Mengkarak	c5	3.2972	102.3972	3.71
Tanjung Batu	c6	3.2056	103.4444	4.89
Lebak	c7	3.3917	102.4306	1.65
Kampung Batu Che Mek	c8	3.3278	102.5069	1.87
Kampung Kuala Bera	c9	3.3903	102.5333	3.11

Batu 9 Jalan Nenasi	c10	3.3875	103.4236	8.74
Bentong	c11	3.4167	102.0583	2.25
Mentakab	c12	3.4833	102.3514	7.88
Kampung Chenor	c13	3.4833	102.5861	4.74
Kampung Salong	c14	3.4861	102.9333	0.76
Paya Membang	c15	3.4542	103.0403	1.65
Kampung Serambi	c16	3.4972	103.1389	1.13
Kerdau	c17	3.5736	102.3764	1.75
Sanggang	c18	3.5181	102.4306	1.01
Kampung Temai Hilir	c19	3.5361	103.2472	2.18
Pekan	c20	3.5611	103.3569	2.73
Kuala Pahang	c21	3.5333	103.4653	2.20
Penor	c22	3.6306	103.3153	1.80
Kuala Krau	c23	3.7111	102.3681	0.90
Kuantan	c24	3.7722	103.2806	3.63
Kampung Sungai Soi	c25	3.7306	103.3000	2.24
Jerantut	c26	3.9625	102.4278	1.43
Paya Kangsar	c27	3.9042	102.4333	3.54
Ladang Nada	c28	3.9083	103.1056	3.52
Ladang Kuala Reman	c29	3.9000	103.1333	2.01
Kuala Lipis	c30	4.1861	102.0431	1.74
Krambit	c31	4.1194	102.2000	1.17
Kampung Chebong	c32	4.1222	102.3458	1.47
Johor				
Kota Tinggi	j1	1.7028	103.8861	0.16
Sembrong	j2	1.8750	103.0542	1.80
Parit Raja	j3	1.8694	103.1125	1.70
Pesarai	j4	1.9208	102.8778	2.32
Batu Pahat	j5	1.9819	102.9250	3.49
Ladang Lambak	j6	1.9681	103.3264	2.16
Yong Peng	j7	2.1306	103.0500	0.99
Ladang Ulu Paloh	j8	2.1056	103.2583	3.66
Kluang	j9	2.2292	103.5986	1.00
Jementah	j10	2.4514	102.6861	1.82
Segamat	j11	2.5861	102.7194	1.21
Empangan Labong	j12	2.5861	103.6694	1.21
Pusat Pertanian Endau	j13	2.6097	103.6306	1.37
Stor JPS Endau	j14	2.6500	103.6208	0.54
Kedah				
Parit Nibong	k1	5.1278	100.5069	0.40
Rantau Panjang	k2	5.5778	100.4069	1.62
Jeniang	k3	5.8139	100.6319	0.40
Alor Setar	k4	6.1056	100.3917	1.01
Kampung Paya	k5	6.2569	100.5306	2.11
Kuala Nerang	k6	6.2542	100.6125	1.35
Melaka				
Telok Rimba	m1	2.1750	102.5014	0.16
Ladang Bukit Kajang	m2	2.2417	102.3750	0.05
Jasin	m3	2.3083	102.4319	0.09
Jalan Empat	m4	2.4389	102.1861	0.06
Negeri Sembilan				
Ladang Sungai Bahru	n1	2.4778	102.0833	2.11
Ladang Bukit Bertam	n2	2.4778	102.0486	2.77
Seremban	n3	2.7306	101.9472	2.94
Pulau Pinang				
Sungai Simpang Ampat	p1	5.2939	100.4806	4.66
Permatang Rawa	p2	5.3625	100.4597	3.69
Ladang Malakoff	p3	5.4889	100.4653	3.30
Pinang Tunggal	p4	5.5572	100.5069	4.47
Perak				
Tanjung Malim	a1	3.7194	101.4889	5.91
Telok Sena	a2	4.2556	100.9000	4.56
Kubang Haji	a3	4.4611	100.9014	5.04
Kuala Kangsar	a4	4.7750	100.8944	6.90
Batu Kurau	a5	4.9792	100.8042	3.39
Perlis				
Guar Nangka	r1	6.4750	100.2833	0.34
Arau	r2	6.4306	100.2708	0.22
Selangor				
Ladang Telok Merbau	b1	2.8639	101.6847	3.22
Ladang Sepang	b2	2.6708	101.7292	2.03
Sungai Mangg	b3	2.8264	101.5417	4.56
Semenyih	b4	2.8987	101.8704	5.69
Ladang Bukit Cheeding	b5	2.9111	101.5764	7.24
Ladang Bukit Kerayong	b6	3.1764	101.3444	2.99
Ladang Bukit Cherakah	b7	3.2264	101.3639	4.97
Ladang Tuan Mee	b8	3.2692	101.4571	6.41
Kuala Selangor	b9	3.3363	101.2562	3.27
Ladang Sungai Buloh	b10	3.3087	101.3210	2.14
Tanjung Karang	b11	3.4236	101.1733	4.63
Sungai Bernam	b12	3.6981	101.3333	3.79

RESEARCH METHODOLOGY

There are various methodologies proposed by prior studies as a remedy for the process of treating or estimating missing rainfall data. In this study, a method that uses data from the neighbours station will be presented such as inverse distance weighted (IDW), modified correlation weighted (MCW), combination correlation with inverse distance (CCID) and averaging correlation and inverse distance (ACID). Generally, a distance weighting technique appears to be one of the most accurate and frequently used for estimation process [27,30]. The main objective of this study is to examine the best imputation methods for treating daily rainfall at 104 stations in Peninsular Malaysia.

The process of imputation can be briefly explained as - suppose there are N neighbouring stations within a radius of 100 km, the rainfall amount for a station i is x_i with value i is equivalent to $1,...,N$. The rainfall amount in target station x_s is the value to be estimated, while the weightage for the neighbouring stations i is denoted as w_i. The formula for this method can be translated into

$$x_s = \sum_{\substack{i=1 \\ i \neq s}}^{N} w_i x_i \tag{1}$$

with constraints is given as $\sum_{i=1}^{N} w_i = 1$.

Inverse distance weighted (IDW)

IDW method is the traditional method that give the greatest weight to the nearest station and reduces weight proportionally as distance increases and minimizes the smoothing of the rainfall distribution. Weighting factor is written as follows:

$$w_i = \frac{d_{is}^{-p}}{\sum_{\substack{j=1 \\ j \neq s}}^{N} d_{js}^{-p}} \tag{2}$$

with d_{is} is euclidean distance between the target station with the neighbouring stations i. Weighting value w used will be less and less as the distance from the target station increase. The power value p also plays an important role in influencing the estimated value on target stations. The higher the p value is used, the greater its

influence in the estimation of data. The value of p that always used is greater than or equal to one.

Modified correlation weighted (MCW)

Teegavarapu and Chandramouli [16] argues that the efficiency of the method of weighted is depend on the strength of correlation between the target stations with neighbouring stations. Thus, the formula of the inverse distance weighted method was modified as follows:

$$w_i = \frac{R_{is}^p}{\sum_{\substack{j=1 \\ j \neq s}}^{N} R_{js}^p} \tag{3}$$

with R_{is} is the correlation between the target stations with neighbouring stations and p is the power.

Combination correlation with inverse distance (CCID)

This method is a modification made to the traditional method. It involves a combination of the inverse distance weighting method and correlation weighted methods. Power value p also applies to the correlation coefficients and weighted as follows:

$$w_i = \frac{R_{is}^p d_{is}^{-2}}{\sum_{\substack{j=1 \\ j \neq s}}^{N} R_{is}^p d_{js}^{-2}} \tag{4}$$

Averaging correlation and inverse distance (ACID)

This method is the average of the two different methods, inverse distance and correlation. The number of station selected for inverse distance weighted still follow the previous assumptions which is within a radius of 100 km. However for correlation weighted part, the number of selected station must have correlation value greater than 0.4 based on the effects of moderate size [31]. Correlation value which is below the threshold will have less relevance to the target station thus lead to overestimate and underestimate rainfall values.

Goodness of fit test

The process of selecting the best estimator methods must be done carefully so that the results obtained did not contain any systematic errors. The first process is to eliminate some value in the target station. The percentage of the selected missing value is 5% because according to Jemain et al. [11], methods such as inverse distance weighted and weighted correlation is practically not sensitive to the percentage of missing value that was used. To determine the best imputation method, which is also known as the most frequent selected method, three model performance test will be considered. Three selected statistical tests to compare the effectiveness of the method in estimating are the root mean square error (RMSE), mean absolute error (MAE) and correlation coefficient (R) is given as follows:

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

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| \tag{6}$$

$$R = \frac{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)\left(\hat{y}_i - \bar{y}\right)}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 \sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2} \tag{7}$$

where n represents the number of data observations, $\hat{y}_i$ is the estimated value and y_i is the observed value.

Error is measured based on the difference between the estimated value and the observed values. For RMSE and MAE test, if the value obtained is small then it shows that the estimation method is the best. However, in R statistical tests, if the estimated value has many similarities with the observed values, then the R value will close to 1.

The selection of neighbouring stations of this study is based on a 100 km radius from the target station. This distance is considered the best and most optimal for areas in Peninsular Malaysia [11]. It is based on the suitability of the number of stations available for the analysis of the study. If the short distance is used, there might be a target stations that do not have neighbouring stations. Conversely, if the distance is large, then it will slow down the calculation process. Less than 10% of the number of neighbouring stations that have no more than 10 if the distance of 100 km is used as shown in Fig. 2.

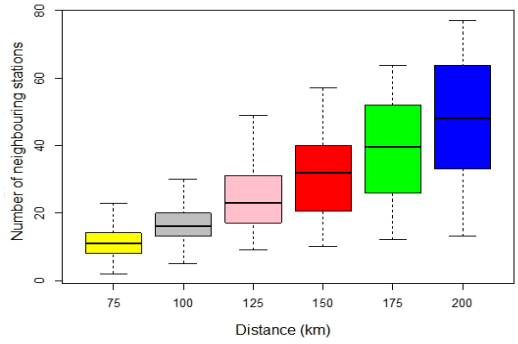

Fig. 2 Boxplots showing the relation between distance of the rainfall stations and the number of neighbouring stations.

The existence of a correlation that is not too high for each occurrence of the rainfall of neighbouring stations is the ultimate reason why correlation is to be considered. Fig. 3 shows a negative relationship between distance of the rainfall stations and their correlation.

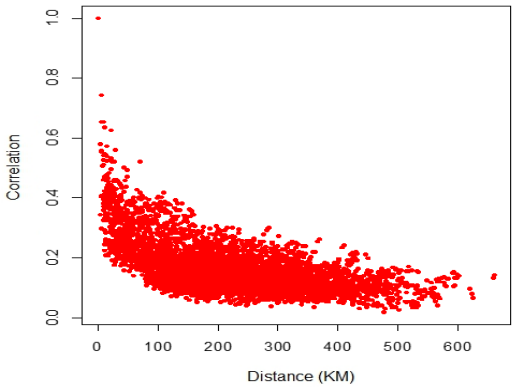

Fig. 3 Scatterplot showing the relation between distance of the rainfall stations and their correlation.

ANALYSIS AND FINDINGS

As mentioned in the earlier part of this paper, this study attempts to identify which method is considered as the best imputation method. To evaluate the performance of each method, the researchers tested three different model, namely RMSE, MAE and R . If the difference between estimated value and the observed value for each station are small, RMSE and MAE will show smallest value. If the estimated value has many similarities with the observed value, then the R value will close to 1. The most frequent method selected will be based

on the smallest RMSE and MAE value and largest R value. The three available methods and one proposed method are selected to estimate the missing value in the time series.

Overall, ACID method showed positive results when almost 41% of the total station prefer this method, followed by MCW methods, 31.73% and IDW method, 19.23%. The rest of the other stations choose CCID method as an estimating method. Table 2 indicates the number and percentage of rainfall stations based on the method of estimation.

Table 2 Number and percentage of rainfall station.

Estimation method	IDW	MCW	CCID	ACID
Number of station (%)	20 (19.23)	33 (31.73)	9 (8.65)	42 (40.38)

ACID method also recorded good results in the test statistic when it has the average RMSE and MAE lower than other methods as well as the highest average correlation value 0.6. While the IDW and MCW method respectively have the same average correlation value 0.55 and CCID method obtained 0.54. This clearly depicts that the selection of stations using an appropriate correlation is instrumental in getting the best estimation. Table 3 shows the average value of the test statistics for all four methods.

Table 3 Average value for test statistics.

Test statistics	RMSE	MAE	R
IDW	13.19	6.64	0.55
MCW	12.99	6.68	0.55
CCID	13.54	6.62	0.54
ACID	12.78	6.28	0.60

Once the missing values were estimated, the completed time series data have to go through homogeneity testing to ensure the quality of the data. This test is important because it can detect a change along a time series. There are four types of homogeneity tests namely standard normal homogeneity test (SNHT), Buishand range test, Von Neumann ratio and Pettitt. These tests have been applied by Wijngaard et al. [32] in their study of climate in Europe. In this study, two variables, namely the annual rainfall amount and annual maximum amount were tested.

Annual rainfall amount and annual maximum amount for each station were tested by using four types of homogeneity test. The critical value at 5% significance level selected is based on the Wijngaard et al. [32] which is valued at 8.45, 1.55, 235 and 1.54. For annual rainfall amount, from 104 stations that were tested only 56 station were consider homogenous. For annual maximum amount, 75 out of 104 were homogenous and finally only 48 stations were considered as homogenous for both variable. Table 4 shows homogenous stations available based on the method of estimation.

Table 4 Number and percentage of homogenous rainfall station.

Estimation method	IDW	MCW	CCID	ACID
Annual rainfall amount (%)	6 (10.71)	18 (32.14)	5 (8.93)	27 (48.21)
Annual maximum amount (%)	10 (13.33)	26 (34.67)	7 (9.33)	32 (42.67)
Annual rainfall and annual maximum amount (%)	5 (10.42)	17 (35.42)	5 (10.42)	21 (43.75)

DISCUSSIONS AND CONCLUSIONS

Ideally, rainfall data for an area and time normally show intrinsic spatial and temporal variation. Thus, the temporal and spatial analysis must be considered in this studies. In other words, the selection of methods should follow the appropriate criteria for any systematic errors can be reduced. Higher quality data are produced and the data mining outcomes also can be improved when estimation is performed in an appropriate way.

Within the context of Peninsular Malaysia, the study revealed that the using of suitable correlation between target station with neighbouring stations plays a vital role in estimating process. Previous studies revealed that the selection of the neighbouring stations were totally depending on the distance. Thus, there is some neighbouring station that have low correlation with the target station. In this case, it may affect the estimation by overestimate or underestimate the missing value. The best way to gain an accurate result is by selecting the nearest station that have "good" correlation. This study consider 0.4 as a good correlation value based on moderate effect size [31]. Any correlation below than threshold will not be considered. As a result, ACID shows a good performance when nearly 40% from the total stations prefer this method compared to others. Thus, it is of utmost important to look at these good correlations in order to improve the estimation results.

Searching for the best method to estimate missing daily rainfall values has been a major interest in several studies. There are many methods that have been tested in order to find the best estimation technique. The existing methods such as inverse distance, correlation coefficient and modified of these two methods have been tested for estimation of missing rainfall values. The performance of these modified method has improved in terms of the RMSE, MAE and R. It is suggested that the selection of the neighbouring station with high correlation value must be considered in hydrological studies of the missing values due to the dependency of space and observations on a station always rely on other stations.

In this paper the researchers have made a comparison of some methods. The proposed method was found to be very useful in estimating missing daily rainfall data. Undeniably, determining the best imputation method is crucial not only for hydrological studies but also for other related studies.

ACKNOWLEDGMENT

The authors sincerely acknowledge the Department of Irrigation and Drainage Malaysia (DID), for providing the complete daily precipitation data that been used in this study. The work is financed by MyBrain15 Scholarship provided by the Ministry of Higher Education of Malaysia and Ministry of High Education (MOHE), STEM Grant with vote no. A. J091002.5600.07397.

REFERENCES

[1] Bennett, N. D., Newham, L. T. H., Croke, B. F. W. & Jakeman, A. J. 2007. Patching and Disaccumulation of Rainfall Data for Hydrological Modelling. *International Congress on Modelling and Simulation (MODSIM 2007)*. December 2007. University of Canterbury, Christchurch, New Zealand. 2520–2526.

[2] Elshorbagy, A. A., Panu, U. S. & Simonovic, S. P. 2000. Group-Based Estimation of Missing Hydrological Data: I. Approach and General Methodology. *Hydrological Sciences*. 45(6), 849–866.

[3] Kajornrit, J., Wong, K. W. & Fung, C. C. 2012. A Comparative Analysis of Soft Computing Techniques Used To Estimate Missing Precipitation Records. *19th ITS Biennial Conference 2012*. 18-21 November 2012. Bangkok, Thailand.

[4] Peterson, T. C., Easterling, D. R., Karl, T. R., Groisman, P., Nicholls, N., Plummer, N., Torok, S. et al. 1998. Homogeneity Adjustments of In Situ Atmospheric Climate Data: A Review. *International Journal of Climatology*. 18(13), 1493–1517.

[5] Zhang, S. 2012. Nearest Neighbor Selection For Iteratively kNN Imputation. *Journal of Systems and Software*. 85(11), 2541–2552.

[6] Peugh, J. L. & Enders, C. K. 2004. Missing Data in Educational Research : A Review of Reporting Practices and Suggestions for Improvement. *Review of Educational Research*. 74(4), 525–556.

[7] Di Piazza, A., Lo Conti, F., 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.

[8] Kim, J. W. & Pachepsky, Y. A. 2010. Reconstructing Missing Daily Precipitation Data using Regression Trees and Artificial Neural Networks For SWAT Streamflow Simulation. *Journal of Hydrology*. 394(3–4), 305–314.
[9] Lee, H. & Kang, K. 2015. Interpolation of Missing Precipitation Data Using Kernel Estimations for Hydrologic Modeling. *Advances in Meteorology*. 2015, 1–12.
[10] Simolo, C., Brunetti, M., Maugeri, M. & Nanni, T. 2010. Improving Estimation of Missing Values In Daily Precipitation Series by a Probability Density Function-Preserving Approach. *International Journal of Climatology*. 30(10), 1564–1576.
[11] Jemain, A. A., Mohd Deni, S., Syed Jamaludin, S. S. & Wan Zin, W. Z. 2015. *Penyurihan Ikhtisar Data Hujan*. Kuala Lumpur: Dewan Bahasa dan Pustaka.
[12] Eischeid, J. K., Pasteris, P. A., Diaz, H. F., Plantico, M. S. & Lott, N. J. 2000. Creating a Serially Complete, National Daily Time Series of Temperature and Precipitation for The Western United States. *Journal of Applied Meteorology*. 39(9), 1580–1591.
[13] Paulhus, J. L. H. & Kohler, M. A. 1952. Interpolation of Missing Precipitation Records. *Monthly Weather Review*. 80(8), 129–133.
[14] Hasana, M. M. & Crokea, B. F. W. 2013. Filling Gaps in Daily Rainfall Data: A Statistical Approach. *20th International Congress on Modelling and Simulation*. 1-6 December 2013. Adelaide, South Australia380–386.
[15] Ramos-Calzado, P., Gomez-Camacho, J., Perez-Bernal, F. & Pita-Lopez, M. F. 2008. A Novel Approach to Precipitation Series Completion In Climatological Datasets: Application to Andalusia. *International Journal of Climatology*. 1525–1534.
[16] Teegavarapu, R. S. V. & Chandramouli, V. 2005. Improved Weighting Methods, Deterministic and Stochastic Data-Driven Models for Estimation Of Missing Precipitation Records. *Journal of Hydrology*. 312(1–4), 191–206.
[17] Zhang, S. 2008. Parimputation : From Imputation and Null-Imputation to Partially Imputation. *IEEE Intelligent Informatics Bulletin*, *9*(1), 32–38.
[18] Xia, Y., Fabian, P., Stohl, A. & Winterhalter, M. 1999. Forest Climatology: Estimation of Missing Values for Bavaria, Germany. *Agricultural and Forest Meteorology*. 96(1–3), 131–144.
[19] Willmott, C. J., Robeson, S. M. & Feddema, J. J. 1994. Estimating Continental and Terrestrial Precipitation Averages From Rain-Gauge Networks. *International Journal of Climatology*. 14(4), 403–414.
[20] Suhaila, J., Deni, S. M. & Jemain, A. A. 2008. Detecting inhomogeneity of rainfall series in Peninsular Malaysia. *Asia-Pacific Journal of Atmospheric Sciences*. 44(4), 369–380.
[21] Young, K. C. 1992. A Three-Way Model for Interpolating for Monthly Precipitation Values. *Monthly Weather Review*. 120(11), 2561–2569.
[22] Filippini, F., Galliani, G. & Pomi, L. 1994. The Estimation of Missing Meteorological Data in a Network of Automatic Stations. *Transactions on Ecology and the Environment*. 4(1), 14328–14336.
[23] Ahrens, B. 2005. Distance in Spatial Interpolation of Daily Rain Gauge Data. *Hydrology and Earth System Sciences Discussions*. 2(5), 1893–1922.
[24] Little, R. J. A. & Rubin, D. B. 2002. *Statistical Analysis with Missing Data*. New Jersey: John Wiley and Sons, Inc.
[25] Schafer, J. L. 1997. *Analysis of Incomplete Multivariate Data*. New York: Chapman and Hall.
[26] Little, R. J. A. & Rubin, D. B. 1987. *Statistical Analysis With Missing Data*. New York: John Wiley and Sons, Inc. 1987.
[27] Presti, R. Lo, 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(1–4), 1–22.
[28] Moritz, S., Sardá, A., Bartz-Beielstein, T., Zaefferer, M. & Stork, J. 2015. Comparison of Different Methods for Univariate Time Series Imputation in R. *arXiv preprint arXiv:1510.03924*, 1–20.
[29] Porth, L. S., Boes, D. C., Davis, R. A., Troendle, C. A. & King, R. M. 2001. Development of a Technique to Determine Adequate Sample Size Using Subsampling and Return Interval Estimation. *Journal of Hydrology*. 251(1–2), 110–116.
[30] Kang, K. & Merwade, V. 2014. The Effect of Spatially Uniform and Non-Uniform Precipitation Bias Correction Methods on Improving NEXRAD Rainfall Accuracy for Distributed Hydrologic Modeling. *Hydrology Research*. 45(1), 23–42.
[31] Evans, J. D. 1996. *Straightforward Statistics for the Behavioral Sciences*. University of California: Brooks/Cole Pub. Co.
[32] Wijngaard, J. B., Klein Tank, A. M. G. & Können, G. P. 2003. Homogeneity of 20th Century European Daily Temperature and Precipitation Series. *International Journal of Climatology*. 23(6), 679–692.