Estimation of Missing Rainfall Data in Pahang Using Modified Spatial Interpolation Weighting Methods

Muhammad Az-zuhri Azman [a], Roslinazairimah Zakaria [b] and Noor Fadhilah Ahmad Radi [c]

[a,b] *Faculty of Industrial Sciences & Technology, Universiti Malaysia Pahang, Lebuhraya Tun Razak, 26300 Gambang, Kuantan, Pahang, Malaysia*
[c] *Institute of Engineering Mathematics, Universiti Malaysia Perlis, Taman Bukit Kubu Jaya, Jalan Seraw, 02000 Kuala Perlis, Perlis, Malaysia*

Abstract. In meteorological and hydrological researches, missing rainfall data has always been one of the most challenging problems which need to be faced by the researchers. The problems of missing rainfall data are due to the wrong technique used when measuring the rainfall, relocation of the rain station and malfunctioned of instrument. Finding the suitable method to solve missing data problem is very critical before going to the next level of data analysis. Most researchers used the spatial interpolation method to estimate the missing rainfall data at a particular target station which is based on the available rainfall data at their neighboring stations. The spatial interpolation method is one of the traditional weighting factors which also consider the correlation between the stations. This study uses the modified of spatial interpolation weighting methods to estimate the missing rainfall data in Pahang and only assume that the particular target station has the missing value. A new modified method of normal ratio and inverse distance weighting with correlation is proposed by abbreviated by NRIDC. The performance of the modified spatial interpolation weighting methods used are assessed using the similarity index (S-index), mean absolute error (MAE) and coefficient of correlation (*R*) for different percentage of missing values (5%-30%).

Keywords: missing data; modified spatial interpolation; normal ratio; correlation; inverse distance weighting.
PACS: 92.40.eg

INTRODUCTION

. Rainfall plays a vital part in agriculture as well as in hydrological and meteorological studies. Hydrological modeling requires continuous rainfall data in identifying spatial and temporal rainfall patterns using the selected model. In the real world data set such as rainfall data, problem of missing data is common and unavoidable. Among the reasons of having missing value could be due to the wrong technique used by researchers when measuring the rainfall, malfunction of the instrument for a specific period of time especially during the extreme environmental events, unsystematic way when storing the data and relocation of meteorological station. Generally, the problem of missing data will reduce the representativeness of the sample taken from the population and will result in the wrong conclusion about the population. In some worst cases, this problem specifically can also prevent the vital analysis of variables considered from being executed and finally will create problems in many applications. Since this problem will highly affect the quality of the data, the step to handle missing data is very crucial before going to the next step of data analysis. Therefore, the proper method for estimating the missing rainfall data should be considered and discussed favourably.

The used of available point observations for interpolation can be optimized by employing the information from topography, data from satellite or radar information (Grimes, Pardo-Iguzquiza, and Bonifacio 1999; Haberlandt and Kite 1998; Seo, Krajewski, and Bowles 1990). Due to inconsistencies and lack of continuous data, estimating the missing rainfall data is very significant before performing the statistical analysis. The most common traditional spatial interpolation weighting method used in estimating the missing data is based on the normal ratio (NR) method (Young, 1992; Paulhus & Kohler, 1952). Paulhus and Kohler (1952) proposed the simple NR method based on the ratio means between the data from a target station and neighboring stations to estimate the missing rainfall data and it is named as old normal ratio (ONR) method. Young (1992) modified the ONR by including the correlation value which gives normal ratio with correlation (NRC). Correlation effect is considered since neighboring stations have similar characteristics with the target station and correlation is significant between those stations. To estimate

missing data in Malaysia, Tang *et al.* (1996) also proposed the new modification of NR by including the effect of square root of the distance (NRD).

Another simplest and widely used method for interpolation is the inverse distance weighting method (ID) (Hubbard 1994). This method estimates the missing data by calculating the weighted average of known data from neighboring station to the target station. The ID method has assumed the concepts of Tobler's first law (first law of geography) which states that "*everything is related to everything else, but near things are more related than distance things*" (Tobler 1970). The distance affects the strength of relatedness between the target station and neighbouring stations.

Nevertheless, other complicated methods have also been applied such as geostatistical methods (simple kriging, ordinary kriging, block kriging, directional kriging, universal kriging and co-kriging) and regression based method which gives better estimation compared to ID (Di Piazza et al. 2011; Vicente-Serrano, Saz-Sánchez, and Cuadrat 2003). Instead of using complicated methods, many researchers improved the ID method to maintain its simplicity (Nejad and Ghahraman 2012). However, Teegavarapu and Chandramouli (2005) stated that the Tobler's first law is only valid when positive spatial autocorrelation exists. Their results have proved that ID method failed when negative and zero spatial autocorrelation exists. They also found that the correlation coefficient weighting (CC) method is superior in providing estimation of missing data compared to IDW method.

The purpose of this paper is to estimate the missing rainfall data in Pahang using the modified spatial interpolation weighting method used by Suhaila *et al.* (2008) and we proposed a new modified of normal ratio and inverse distance weighting by including correlation coefficient (NRIDC). All the modified spatial interpolation weighting method proposed by Suhaila *et. al* (2008) are compared to NRIDC. Then, the performance of all the modified methods is assessed using similarity-index (SI), mean absolute error (MAE) and correlation coefficient (*R*).

DATA AND AREA OF STUDY

In this study, daily rainfall data from ten meteorological stations located in Pahang state are considered and Kg. Bahru, Pekan (3.38ºN, 103.4236ºE) becomes the target station. Data ranges from 1st of January 1960 to 31st of December 2012 are used in this study. Summarization of distance and correlation of the target station and their neighboring stations are shown in **TABLE (1).**

TABLE (1). Distance (km) and correlation between the target station (TS) and nine neighboring stations (within the radius of 200 km) in Pahang.

Station Points	Station Names	Latitude	Longitude	Euclidean Distance (km)	Correlation
TS	**Kg. Bahru, Bt.9, Jln. Nenasi**	**3.388**	**103.427**		
1	Kastam Kuala Pahang	3.533	103.465	0.151 (16.88)	0.23
2	Kg. Tering di Tanjung Batu	3.206	103.444	0.183 (20.38)	0.23
3	Rumah Pam Pahang Tua di Pekan	3.561	103.357	0.185 (20.69)	0.23
4	Kg. Serambi	3.497	103.139	0.305 (33.91)	0.18
5	Sg. Jerik	3.781	102.643	0.873 (97.13)	0.15
6	Pelangi Kg. Jawi 2	3.174	102.242	1.201 (133.49)	0.13
7	Ldg. Bukit Dinding di Bentong	3.417	102.058	1.365 (151.75)	0.14
8	Kuala Marong di Bentong	3.513	101.915	1.513 (168.18)	0.12
9	Janda Baik	3.326	101.863	1.562 (173.62)	0.09

Estimation Method of Missing Data

In this section, five estimation methods of missing data are presented. The first four methods are the existing methods used by Suhaila et al. (2008), namely modified coefficient correlation weighting method (CCM), modified correlation coefficient with inverse distance weighting method (CID), modified normal ratio with inverse distance method (NRID), modified old normal ratio with Inverse distance method (ONRID). In this study, the fifth method of the normal ratio inverse distance weighting with correlation (NRIDC) is proposed.

(a) *Modified Coefficient Correlation Weighting Method*

In general, the spatial interpolating weighting method is based on the successful of ID method which is highly depends on the positive spatial correlation between the neighbouring station and target station. Thus, the distance of ID method is interchangeable with correlation coefficient(Teegavarapu and Chandramouli, 2005). Nevertheless, Suhaila et al. (2008) modified this method by considering different power of correlation coefficient to give more weight to the existing method of CCM. The modified of CCM weighting is expressed as follows:

$$W_i = \frac{r_{it}^{p}}{\sum_{\substack{i=1 \\ i \neq t}}^{N} r_{it}^{p}}$$

where r_{it}^{p} is the correlation coefficient which indicates the strength of the relationship between the target station, t and the i^{th} neighboring stations with p value ranging from 2 to 6. As stated by Suhaila et al. (2008), the result of CCM is superior than the ID (Hubbard 1994), modified NR method by Young (1992) and CC method by Teegavarapu and Chandramouli (2005) especially when the value of $p \geq 4$.

(b) Modified Correlation Coefficient with Inverse Distance Weighting Method

This method is a combination between the ID and CCM methods in estimating the missing rainfall data (Suhaila, Sayang, and Jemain 2008). The ID method is based on distance weighting which highly depends on the minimum distance between the target station and neighboring station for interpolating the spatial data. Inevitably, the correlation coefficient should not be neglected. Hence, the correlation value is incorporated into ID method. The CID weighting is given by

$$W_i = \frac{r_{it}^{p} d_{it}^{-p}}{\sum_{\substack{i=1 \\ i \neq t}}^{N} r_{it}^{p} d_{it}^{-p}}$$

where W_i is the respective weight, r_{it}^{p} is the correlation coefficient between the target station t and the i^{th} neighboring station with the best exponent value of $p \geq 4$ and d_{it} is the distance between the target station t and the i^{th} neighboring. According to Xia, Fabian, Stohl and Winterhalter (1999), the most commonly used value for p in ID is 2 with p ranges from 1.0 to 6.0. Based on Suhaila et al. (2008), only $p = 2$ is chosen in this study.

(c) Modified Normal Ratio with Inverse Distance Method

The combination of NR modified method (Young, 1992) and ID method (Hubbard 1994) is considered the best due to their simplicity. The NR modified method (Young 1992) depends on the positive spatial correlation while the ID method is based on the Tobler (1970) assumption. The weighting of NRID can be written as follows:

$$W_i = \frac{(n_i - 2) r_{it}^2 (1 - r_{it}^2)^{-1} d_{it}^{-2}}{\sum_{\substack{i=1 \\ i \neq t}}^{N} (n_i) r_{it}^p (1 - r_{it}^p)^{-1} d_{it}^{-2}}$$

where W_i is the respective weight, n_i is the number of points used to compute the correlation coefficient, r_{it}^2 is the square of correlation coefficient of daily rainfall data between the target station t and the i^{th} neighboring station and d_{it} is the distance between the target station t and the i^{th} neighboring station.

(d) Modified Old Normal Ratio with Inverse Distance Method

Normal ratio method modified by Tang et al. (1996) based on the square root distance (NRD) is found to be inferior compared to NRM proposed by Young (1992). The NRID method (Suhaila, Sayang, and Jemain 2008) is the combination between NRD and NRM methods which gives better results in performance compared to NRD and NRM modelled separately. The weighting of ONRID is given by

$$W_i = \frac{\frac{\mu_t}{\mu_i} d_{it}^{-2}}{\sum_{\substack{i=1 \\ i \neq t}}^{N} \frac{\mu_t}{\mu_i} d_{it}^{-2}}$$

where W_i is the respective weight, μ_t is the sample mean of available data at the target station, t while μ_i is the sample mean of available data at the i^{th} neighboring station and d_{it} is the distance between the target station, t and the i^{th} neighboring station.

(e) Normal Ratio Inverse Distance Weighting with Correlation Method

In this study, we proposed the combination between NR, ID and correlation value and it is named as normal ratio inverse distance weighting with correlation (NRIDC). This method is modified from Suhaila et al. (2008) by including the correlation value to NRID. Nevertheless, this slight modification still maintained the idea proposed by Suhaila et al. (2008) to combine between normal ratio, correlation and inverse distance in one weighting method in order to give more weight for better estimation of missing rainfall data. The weighting of NRIDC is given by

$$W_i = \frac{r_{it}^p \frac{\mu_t}{\mu_i} d_{it}^{-2}}{\sum_{\substack{i=1 \\ i \neq t}}^{N} r_{it}^p \frac{\mu_t}{\mu_i} d_{it}^{-2}}$$

where W_i is the respective weight of NRIDC method, r_{it}^p is the correlation coefficient between the target station t and the i^{th} neighboring station with the best exponent value of $p \geq 4$, μ_t is the sample mean of available data at the target station, t while μ_i is the sample mean of available data at the i^{th} neighboring station and d_{it} is the distance between the target station, t and the i^{th} neighboring station.

Assessment of Estimation Methods

The performance of all the estimation methods of missing data applied are assessed using the similarity index (S-index), mean absolute error (MAE) and correlation coefficient (*R*). According to Wilmott (1981), S-index is a standardized measure of the degree of model prediction error and range between 0.0 (no agreement) and 1.0 (perfect match). This index of agreement detects the additive and proportional differences in the observed and simulated means and variances, but due to the squared differences, it was too sensitive to extreme values (Legates and Mccabe 1999). The error measures between the estimated data and their actual observed data are then compared using three error indices as follows:

$$S-index = 1 - \frac{\sum_{i=1}^{n}\left(\hat{x}_i - x_i\right)^2}{\sum\left(\left|\hat{x}_i - \overline{x}\right| + \left|x_i - \overline{x}\right|\right)^2}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{x}_i - x_i\right|$$

$$R = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(\hat{x}_i - \overline{x})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(\hat{x}_i - \overline{x})^2}}$$

where n is the total number of observations, $\hat{x}_i$ is the estimated value, x_i is the actual of the observed value and $\overline{x}$ is the mean value .

Results and Discussion

Missing rainfall data was estimated using different modified spatial weighting methods of CCM, CID, NRID, ONRID and NRIDC. We proposed the last method. Three indices of S-Index, MAE and R are applied to assess the performance of the estimation method for different percentages of missing data. The performance results of estimation methods are shown in **TABLE (2).**

TABLE (2). Comparison of estimation methods based on S-Index, MAE and R for various percentages of missing data.

Methods	Percentage of Missing Values (%)					
	5%	10%	15%	20%	25%	30%
			Similarity Index (S-index)			
CCM	**0.9863**	**0.9807**	**0.9638**	**0.9508**	**0.9397**	**0.9252**
ONRID	0.9860	0.9807	0.9625	0.9497	0.9384	0.9239
*ONRIDC	0.9857	0.9800	0.9614	0.9482	0.9367	0.9215
NRID	0.9857	0.9798	0.9614	0.9481	0.9365	0.9211
CID	0.9857	0.9798	0.9614	0.9481	0.9365	0.9211
			Mean Absolute Error (MAE)			
CCM	**0.4633**	**0.8285**	**1.3523**	**1.7941**	**2.1489**	**2.6917**
ONRID	0.4669	0.8267	1.3553	1.7938	2.1573	2.6981
*ONRIDC	0.4693	0.8309	1.3670	1.8073	2.1747	2.7250
NRID	0.4689	0.8326	1.3695	1.8106	2.1786	2.7323
CID	0.4689	0.8327	1.3696	1.8107	2.1787	2.7325

	Correlation Coefficient *(R)*					
CCM	**0.9732**	**0.9624**	**0.9307**	**0.9070**	**0.8872**	**0.8616**
ONRID	0.9725	0.9622	0.9279	0.9042	0.8840	0.8582
*ONRIDC	0.9719	0.9608	0.9257	0.9011	0.8805	0.8533
NRID	0.9719	0.9604	0.9257	0.9009	0.8801	0.8525
CID	0.9719	0.9604	0.9256	0.9009	0.8801	0.8525

* New proposed method

A good performance of spatial weighting method indicates the high value for both values of S-index and correlation but low for MAE value. In this study, a very slight difference could be seen based on the performance result between all the modified methods (Suhaila, Sayang, and Jemain 2008) and the new proposed method (NRIDC). In overall, the performance results showed that the method of modified correlation coefficient (CCM) is superior compared to other methods such results probably because of geographical characteristics which are almost similar since the stations used are in the same state and indicates strong relationship between all the stations. The proposed method is found to give better estimation compared to NRID and CID methods at four decimal places. FIGURE 1 shows the graphical performance assessment of all estimation weighting methods used for different percentages of missing data. The performance of each estimation method tends to decrease slightly in S-index and correlation coefficient but to increase slightly for MAE when using different percentages of missing data. Despite the graph shows drastically decreased for S-index and *R* and drastically increased for MAE, the performance of all methods gives perfect match and indicates good correlation.

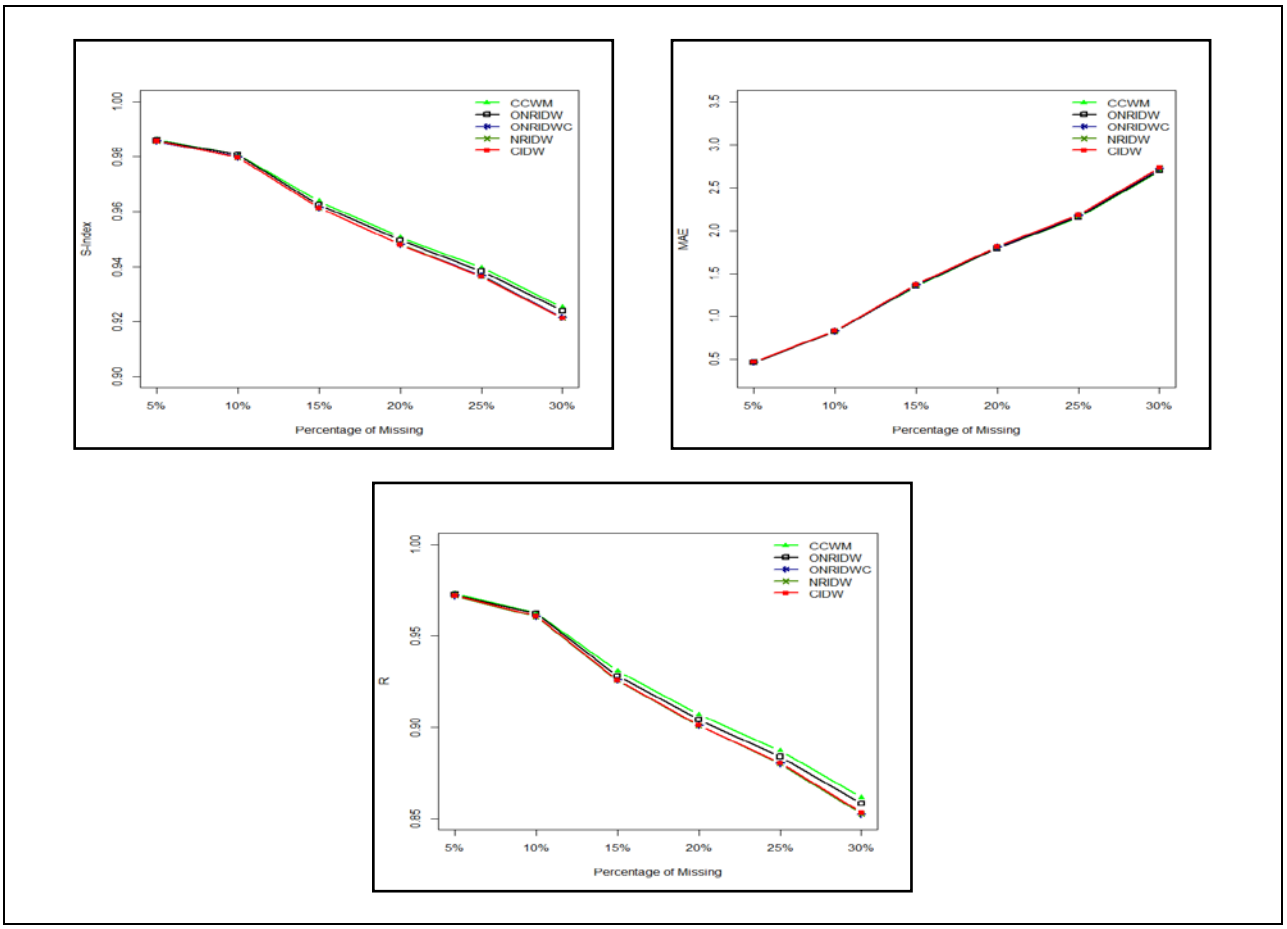

FIGURE 2. Graph of performance assessment using three indices (S-index, MAE and *R*) of various percentages of missing data.

Conclusion

This study compares the performance of modified estimation weighting method proposed by Suhaila et al. (2008) and the new proposed method using rainfall data in Pahang. The performance of all methods is assessed using three indices indicators (S-index, MAE and correlation coefficient) for different percentages of missing data (5% - 30%). The good performance results are shown by the S-index with value close to one and high correlation value with low MAE. The new proposed method gives better estimation compared to NRID and CID methods. For further research, the comparison between the geostatistical methods (simple kriging, ordinary kriging, block kriging, directional kriging, universal kriging and co-kriging) and regression based method to this modified method can be made. The use of different p values is also suggested.

ACKNOWLEDGMENTS

The authors wish to express their gratitude to all who have involved directly and indidrectly especially Mr. Abu Salim Abd. Aziz from Water Resources Management and Hydrology Division, Malaysia for providing the rainfall data. The authors also acknowledged Universiti Malaysia Pahang for the financial support (RDU120101).

REFERENCES

1. D.I.F. Grimes, E. Pardo-Iguzquiza, and R. Bonifacio, "Optimal Areal Rainfall Estimation Using Raingauges and Satellite Data." *Journal of Hydrology* 222(1-4): 93–108, 1999.
2. U. Haberlandt and G. W. Kite, "Estimation of Daily Space-Time Precipitation Series for Macroscale Hydrological Modelling." *Hydrological Processes* 12: 1419–32, 1998.
3. K.G. Hubbard, "Spatial Variability of Daily Weather Variables in the High Plains of the USA." 68(94): 29–41, 1994.
4. D.R. Legates and G. J. Mccabe, "Evaluating the Use of 'goodness-of-Fit' Measures in Hydrologic and Hydroclimatic Model Validation." *Water Resources Research* 35(1): 233–41, 1999.
5. S.S Nejad and B. Ghahraman, "Extended Modified Inverse Distance Method for Interpolation Rainfall." *International Journal of Engineering Inventions* 1(3): 57–65. (4 June 2014).
6. J. Paulhus and M.A. Kohler, "Interpolation of Missing Precipitation Records." *Mon. Wea. Rev*: 129–33, 1952.
7. A. Di Piazza, F. L. Conti, L.V. Noto, F.Viola and G. La Loggia "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, Jun 2011.
8. D-J. Seo, W. F. Krajewski and D. S. Bowles, "Stochastic Interpolation of Rainfall Data from Rain Gages and Radar Using Cokriging 2. Results." *Water Resources Research* 26: 915–24, 1990.
9. J. Suhaila, M. D. Sayang, and A.A Jemain, "Revised Spatial Weighting Methods for Estimation of Missing Rainfall Data." *Asia-Pacific Journal OF Atmospheric Sciences* 44(2): 93–104, 2008.
10. W.Y. Tang, A. H. M. Kassim and S.H Abubakar, "Comparative Studies of Various Missing Data Treatment Methods Malaysian Experience." *Atmospheric Research* 42: 247–62, 1996.
11. R. S. V. Teegavarapu and V. Chandramouli, "Improved Weighting Methods, Deterministic and Stochastic Data-Driven Models for Estimation of Missing Precipitation Records." *Journal of Hydrology* 312: 191–206, 2005.
12. A.W.R. Tobler, "Clark University.", vol. 46, 1970.
13. S. Vicente-Serrano, M. Saz-Sánchez and J. Cuadrat, "Comparative Analysis of Interpolation Methods in the Middle Ebro Valley (Spain): Application to Annual Precipitation and Temperature." *Climate Research* 24: 161–80, 2003.
14. C.J. Wilmott, "On the Validation of Models." 184–94, 1981.
15. Y. Xia, P. Fabian, A. Stohl, and M. Winterhalter, "Forest Climatology: Estimation of Missing Values for Bavaria, Germany." *Agricultural and Forest Meteorology* 96: 131–44, 1999.
16. K. C. Young, "A Three-Way Model for Interpolating for Monthly Precipitation Values." *Monthly Weather Review* 120: 2561–69, 1992.