Spatial Interpolation for Missing Rainfall Data in Northern Region of Peninsular Malaysia

Mohd Khaidir Mohamed Salleh¹, Noor Fadhilah Ahmad Radib², and Nor Azrita Mohd Amin³

¹Institut Kemahiran MARA, KM 14 Jalan Kaki Bukit, 02400 Beseri Perlis.
²Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Jalan Ilmu 1/1, 40450 Shah Alam, Selangor.
³Institute of Engineering Mathematics, Universiti Malaysia Perlis, Kampus Pauh Putra, 02600 Arau, Perlis.

*E-mail: khaidir@mara.gov.my

Abstract. This study is aimed to estimate missing rainfall values for daily rainfall data from 30 selected rainfall stations. The daily rainfall data were obtained from the Department of Irrigation and Drainage Malaysia (DID) for the periods of 1999 to 2019. The missing values throughout the 20 years period were estimated using spatial interpolation methods. These methods include arithmetic average (AA), normal ratio (NR), inverse distance (ID) and coefficient of correlation (CC) weighting methods. The methods consider the distance between the target and the neighbourhood stations as well as the correlation between them. In determining the best spatial interpolation method, three tests for evaluating model performance have been used namely similarity index (S-index), mean absolute error (MAE) and root mean square error (RMSE). The homogeneity test using Standard normal homogeneity (SNHT), Buishand range (BR), Pettitt and Von Neumann (VNR) ratio are conducted to test the homogeneity of the rainfall data. The results show that the ID method is more efficient than the others method and 85% of the rainfall stations were homogenous based on this method. This study is important as it can be used to fill in the missing value rainfall data so that the conclusions that can be drawn from the data is valid.

1. Introduction

Daily rainfall data are one of the most important measurement variables in hydrological and environmental modelling. Rainfall data is important in assessing the water quality. However, studies involving the use of continuous series data are always faced with the problem of missing value. Mostly the existing data series are not enough to perform a good and meaningful analyses and often contain a large number of missing values [1]. Normally lack of data and inhomogeneity problem are due to rainfall station relocation, instrument malfunction and network reorganization [2]. In hydrologic modelling, developing a method to get an accurate estimation of rainfall are very crucial. In order to get an accurate result in analyses, the rainfall data must be complete, homogeneous and have a good quality.

In general, there are a number of methods have been proposed to estimate missing value. The best estimation should not change the important characteristics of the dataset and should follow the character of rainfall in a given area [3]. There are two ways to solve a missing value in a dataset which is by using removal methods such as listwise deletion and pairwise deletion and the other methods is imputation which it is divided into several section such as single imputation, multiple imputation and interative
imputation [4]. Spatial Interpolation is another method that can be used to estimate missing value where spatial factor take into consideration as a main factor.

Basically, the most common spatial interpolation method used in estimating rainfall data is normal ratio weighting method [5]. This method is based on the past observations of the target station and neighbourhood stations as proposed by [6]. A simpler method of spatial interpolation used in estimating missing data is arithmetic average weighting method. This method considers the average annual rainfall amount at the target station as well as the neighbourhood stations and able to be used if the average annual rainfall amount at the target station is within 10% of the difference of the average annual rainfall amount from the neighbourhood stations [5]. The inverse distance weighting method is another simpler method which is based on the assumption that the rainfall amount at the target station could be influenced most by the nearest stations and less by the more distant stations [7]. The weight of inverse distance method is made by finding the inverse distance of the neighbourhood station to some power of the distances from the target station. [8] proposed the correlation coefficient weighting method by replacing the weighting value of the inverse distance with the correlation coefficient. Thus, the weighting value of inverse distance method is $1/d_i^b$ whereas for coefficient correlation is $r_i$. The study show that the inverse distance method is far more above superior than the other methods.

Hence, the main objective of this study is to examine the best spatial interpolation methods for treating daily rainfall at 30 stations in northern region of Peninsular Malaysia. The performance of methods used are compared and assessed using the similarity index, mean absolute error and root mean square error. Then, the data series will be test to detect any inhomogeneous rainfall series data by adopting approach introduced by [9]. Two testing variables which are annual rainfall amount and annual maximum amount are used in this test.

2. Study area and data

The methods are applied for the state of northern region of Peninsular Malaysia which are Penang, Kedah and Perlis. These region lies in the Equatorial zone of Northern latitude between 5° N and 6° N and Eastern longitude from 99° E to 100° E. Peninsular Malaysia experienced hot and humid weather all year round. Typically, the Malaysian climate influenced by wind 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 occur from November to March. Whereas the transition period between the two monsoons is recognized as the inter-monsoon period occurring in March to April and September to October, bringing intense convective rain to many areas in Peninsular Malaysia. 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 20% missing values throughout the 20 years period. A large amount of time series observations is required in order to obtain an accurate overview pertaining to the pattern of the rainfall [3]. In addition, long time series data is valuable because the credibility of the frequency estimator is close related to the size of the sample during the analysis process [10]. A 20-year record of data during the years 1999-2019 is obtained from the Department of Irrigation and Drainage Malaysia (DID) for further analysis. In this study, 30 rainfall stations where scattered in Penang, Kedah and Perlis are considered with Jeniang Klinik, T as the target station. In general, the target station has a complete set of data. In spatial interpolation procedure, the information about distance and correlation between target station and neighbourhood station are needed. Table 1 shows the description of the 30 rainfall stations in northern region Peninsular Malaysia with Jeniang Klinik, T as the target station.
Table 1. Description of the 30 rainfall stations in northern region peninsular Malaysia with Jeniang Klinik, T as the target station.

| Station | Station Name                        | Latitude (S°) | Longitude (E°) | Euclidean Distance (km) | Correlation | Missing Value (%) |
|---------|-------------------------------------|---------------|----------------|-------------------------|-------------|-------------------|
| T       | Jeniang Klinik                      | 5.8139        | 100.6319       | 4.8                     | 0.4062      | 8.6               |
| 1       | Bt. 27 Jln. Baling                  | 5.5833        | 100.7361       | 0.25305 (28)            | 0.4783      | 4.2               |
| 2       | Kedah Peak                           | 5.7958        | 100.4389       | 0.19385 (22)            | 0.4404      | 4.3               |
| 3       | Bt. 61 Jln. Baling                  | 5.8806        | 100.8944       | 0.37787 (42)            | 0.3361      | 8.3               |
| 4       | Stor JPS Alor Setar                 | 6.1056        | 100.3917       | 0.36255 (40)            | 0.3804      | 6.1               |
| 5       | Komplek Rumah Muda                  | 6.1056        | 100.8472       | 0.44073 (49)            | 0.3386      | 5.1               |
| 6       | Kuala Nerang                        | 6.2542        | 100.6125       | 0.53241 (59)            | 0.2918      | 10.0              |
| 7       | Padang Sanai                        | 6.3431        | 100.6903       | 0.4062                  | 0.4062      | 8.6               |
| 8       | Rumah Jps. Padang Mat Sirat         | 6.3558        | 99.7317        | 1.05072 (116)           | 0.2650      | 8.3               |
| 9       | Padang Katong di Kangar             | 6.4458        | 100.1875       | 0.77252 (85)            | 0.2812      | 6.0               |
| 10      | Ngolang                             | 6.4750        | 100.2472       | 0.76488 (84)            | 0.2992      | 6.8               |
| 11      | Padang Besar di Titi Keretapi       | 6.6569        | 100.3097       | 0.90248 (100)           | 0.2431      | 7.2               |
| 12      | Kolam Takongan Bkt. Panchor         | 5.1611        | 100.5361       | 0.65979 (73)            | 0.3108      | 10.7              |
| 13      | Sg. Simpang Ampat                   | 5.2939        | 100.4806       | 0.54156 (60)            | 0.3531      | 6.7               |
| 14      | Ldg. Batu Kawan                     | 5.2569        | 100.4306       | 0.59226 (66)            | 0.3286      | 10.4              |
| 15      | Talair Besar Sg. Pinang             | 5.3917        | 100.2125       | 0.59510 (66)            | 0.3322      | 5.2               |
| 16      | Pintu Air Bagan di Air Itam         | 5.3542        | 100.2000       | 0.63076 (70)            | 0.3325      | 8.1               |
| 17      | Kolam Takongan Air Itam             | 5.3958        | 100.2653       | 0.55606 (62)            | 0.3646      | 6.7               |
| 18      | Komplek Prai                        | 5.3819        | 100.3917       | 0.49429 (55)            | 0.3627      | 9.1               |
| 19      | Pusat Kesihatan Bukit Berapit       | 5.3756        | 100.4756       | 0.46533 (51)            | 0.3850      | 9.5               |
| 20      | Permatang Rawa                      | 5.3625        | 100.4597       | 0.48313 (53)            | 0.3629      | 8.9               |
| 21      | Pusat Pertanian Cherok To’Kun       | 5.3497        | 100.4742       | 0.49026 (54)            | 0.3662      | 10.5              |
| 22      | Klinik Bkt. Bendera                 | 5.4236        | 100.2708       | 0.53172 (59)            | 0.3438      | 5.3               |
| 23      | Kolam Bersih Pulau Pinang           | 5.4403        | 100.2861       | 0.50907 (56)            | 0.3398      | 4.6               |
| 24      | Lorong Batu Lanchang                | 5.4025        | 100.2994       | 0.52897 (59)            | 0.3453      | 5.4               |
| 25      | Ldg. Malakoff                       | 5.4889        | 100.4653       | 0.36521 (40)            | 0.4183      | 10.4              |
| 26      | Permatang Binjai                    | 5.5028        | 100.3944       | 0.39139 (43)            | 0.4063      | 9.8               |
| 27      | Rumah Pam Bumbung Lima              | 5.5589        | 100.4364       | 0.32132 (36)            | 0.4138      | 9.2               |
| 28      | Lahar Ikan Mati di Kepala Batas     | 5.5347        | 100.4306       | 0.34420 (38)            | 0.4083      | 5.3               |
| 29      | Rumah Pam Pinang Tunggal            | 5.5572        | 100.5069       | 0.28552 (32)            | 0.4225      | 10.3              |
3. Research methodology

There are numerous methodologies proposed by prior studies as a remedy for the process of treating or estimating missing rainfall data. In this part, we will discuss shortly about the methods for estimating missing data and assessing the performance of the methods used. The analysis involved a target and some selected neighbourhood stations. Initially, the missing data are identified in the target stations. For the target station that has missing data, the average value between the available data from the neighbourhood stations is used. Then missing data are introduced in the target station. Using the spatial interpolation methods, the missing rainfall data in target stations are estimated and compared with the actual observations. The spatial interpolation methods used are as follows.

3.1. Spatial Interpolation and Single Imputation Methods

(i) Arithmetic Average Methods (AA)

The arithmetic average method is basically the average of rainfall amount of all the neighbourhood stations. The estimated missing value is given by

\[ p_i = \frac{1}{n} \sum_{i=1}^{n} y_i; \]

where \( p_i \) is the estimated value of the missing rainfall at the target station, \( y_i \) is the observed rainfall at neighbourhood station and \( n \) is the number of neighbourhood stations.

(ii) Normal Ratio Method (NR)

The missing rainfall at the target station is estimated as the weighted average of neighbourhood stations. The rainfall data at each of the neighbourhood stations is weighted by the ratio of the average annual rainfall at the target station and average annual rainfall of the neighbourhood station. The estimated missing value is given by

\[ p_i = \frac{1}{n} \sum_{i=1}^{n} \frac{N_i}{N_i} y_i; \]

where \( N_i \) is the annual rainfall amount at the target station and \( N_i \) is the annual rainfall amount at the \( i \)th neighbourhood station.

(iii) Inverse Distance Method (ID)

This method weights neighbourhood stations on the basis of their distance from the target station, on the assumption that closer stations are better correlated than those farther away. The estimated missing value is given by

\[ p_i = \frac{\sum_{i=1}^{n} y_i / d_{it}^b}{\sum_{i=1}^{n} 1 / d_{it}^b}; \]

where \( d_{it} \) is the distance between the target station and the \( i \)th neighbourhood station and \( b \) is the power of distance.

(iv) Coefficient Correlation Method (CC)
This method is used by replacing the distance with the correlation coefficient as the weighting value. The estimated missing value is given by

$$p_t = \frac{\sum_{i=1}^{n} y_i / r_{it}}{\sum_{i=1}^{n} r_{it}} ;$$

where \( r_{it} \) is the correlation coefficient of daily time series data between the target station and the \( i \)th neighbourhood station.

3.2. Performance of the estimation methods
In this trial, the performance of the estimation methods used are compared and assessed using the similarity index (S-index), mean absolute error (MAE) and root mean square error (RMSE). The error measures the difference between the estimation values and their corresponding observed values. For MAE and RMSE test, if the value obtained is small, then it shows that the estimation method is the best. While for S-index is the index of agreement for assessing model performance which implies the percentage of agreement between the observed and estimated values. The values of S-index range from 0.0 for complete disagreement to 1.0 for perfect agreement [11]. The three error indices are given by

$$S\text{-Index} = 1 - \frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n} (|\hat{y}_i - \bar{y}| + |y_i - \bar{y}|)^2}$$  \hspace{1cm} (5)$$

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |\hat{y}_i - y_i|$$  \hspace{1cm} (6)$$

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2}$$  \hspace{1cm} (7)$$

where \( y_i \) is the observed rainfall at target station, \( \hat{y}_i \) is the estimated value and \( n \) is the number of observed rainfall.

3.3. Homogeneity test
In this study, four homogeneity test namely Standard normal homogeneity (SNHT), Buishand range (BR), Pettitt and Von Neumann (VNR) ratio are used to test the homogeneity of the rainfall data. Null hypothesis \( Y_i (i = 1, 2, ..., n) \) is independent and identically distributed, where \( Y_i \) is the testing variable which could represent either the annual rainfall amount or annual maximum amount that have the same mean is said to be homogenous while alternative hypothesis SNHT, BR and Pettitt test assume the series consisted of break in the mean and is considered as inhomogeneous. These three tests are known as location-specific-tests in Wijngaard et al., 2003 able to detect the year where break occurs. On the other hand, VNR not able to detect the year of break because the series is not randomly distributed under alternative hypothesis.

There are some differences between the first three test mention earlier in this section. SNHT is sensitive in detecting the breaks near the beginning and the end of the series while BR and Pettitt test are used to identify the break in the middle of the series. Kang and Fadhilah, 2012 seek to identify that
SNHT and BR tests assume \( Y_i \) to be normally distributed. In contrary, Pettitt test does not need this assumption because it is a non-parametric rank test.

The next steps in homogeneity test is the evaluation of the four tests. The outcomes of the four tests for annual rainfall amount and annual maximum amount are grouped together. A classification is made depending on the number of tests rejecting the null hypothesis.

Class 1: ‘useful’ — one or zero tests reject the null hypothesis at the 5% level.
Class 2: ‘doubtful’ — two tests reject the null hypothesis at the 5% level.
Class 3: ‘suspect’ — three or four tests reject the null hypothesis at the 5% level.

4. Analysis and findings

As mentioned in the early part of this paper, this study attempts to identify which method is considered as the best spatial interpolation method. To evaluate the performance of each method, we tested three different model, namely MAE, RMSE and S-Index. If the difference between estimated value and the observed value for each station are small, MAE and RMSE will show smallest value. If the estimated value has many similarities with the observed value, then S-Index value will close to 1. The most frequent method selected will be based on the smallest MAE and RMSE value and largest S-Index value. The four available methods and one proposed method are selected to estimate the missing value in the time series.

Overall, ID method showed positive results when almost 87% of the total station prefer this method. The rest of the other stations choose NR and ID as an estimating method and none for AA method. Table 2 indicates the number and percentage of rainfall stations based on the method of estimation.

| Estimation method | AA  | NR  | ID  | CC  |
|-------------------|-----|-----|-----|-----|
| Number of station | 0   | 2   | 26  | 2   |
| (%)               | 0.0 | 6.67| 86.67| 6.67|

ID method also shows a good results in the test statistics with the average of MAE and RMSE are lower than other methods as well as the highest average S-Index value which is 0.7439. While the CC method is the second best method with the average of MAE and RMSE are the second lower than other methods with average S-Index value is 0.6874. This clearly depicts that the selection of station using an appropriate S-Index as instrumental in getting the best estimation. Table 3 shows the average value of the test statistics for all four methods.

| Test Statistics | MAE  | RMSE | S-Index |
|-----------------|------|------|---------|
| AA              | 5.5721| 11.3110| 0.5864  |
| NR              | 5.5133| 11.1549| 0.6280  |
| ID              | 4.7767| 9.9671 | 0.7439  |
| CC              | 5.2920| 10.8122| 0.6874  |

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 data. There are for 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., 2003. 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., 2003. For annual rainfall amount, from 30 stations that were tested only 22 stations were considered
homogenous. For annual maximum amount, all 30 stations were considered homogenous. Table 4 shows homogenous station available based on the method of estimation.

| Estimation method            | NR  | CC  | ID     |
|-----------------------------|-----|-----|--------|
| Annual rainfall amount (%)  | 0   | 1   | 21     |
| Annual Maximum Amount (%)   | 2   | 2   | 26     |

5. Discussions and conclusions
Ideally, rainfall data for an area and time normally show intrinsic spatial and temporal variation. Therefore, the temporal and spatial analysis errors can be reduced. Higher quality data can be obtained and the data mining outcomes also can be improved when estimation is performed in an appropriate way.

Within the context of northern region in peninsular Malaysia, the study discovered that the using of suitable correlation between target station with neighbourhood stations plays an important role in estimating process. Previous studies also discovered that the selection of the neighbourhood stations were totally depending on the distance. Therefore, there is some neighbourhood stations that have low correlation with the target station. In this case, it may affect the estimation either overestimate or underestimate the missing value. The best way to obtained an accurate result is by selecting the nearest station that have good in correlation. This study considers Pearson correlation, r > 0.4 as a good correlation value based on moderate effect size [13]. Any correlation below than threshold will not be considered. As a result, ID shows a good performance when nearly 87% from the total stations prefer this method compared to others. Therefore, it is of utmost important to look at these good correlation and distance 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. The existing spatial interpolation methods such as arithmetic average, normal ratio, inverse distance and coefficient correlation methods have been tested for estimation of missing rainfall values. The performance of these method has improved in term of S-Index, MAE and RMSE. It is proposed that the selection of the neighbourhood stations with high correlation and closest to target station must be considered in hydrological studies of the missing value due to the dependency of space and observations on a station always rely on other stations.

In this paper, we 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 spatial interpolation method is crucial not only for hydrological studies but also for other related studies.

References
[1] Kajornrit J, Wong K W and 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.
[2] Peterson T C, Easterling D R, Karl T R, Groisman P, Nicholls N, Plummer N, Torok S, Auer I, Boehm R, Gullett D and Vincent L. 1998. Homogeneity Adjustments of In Situ Atmospheric Climate Data: A Review. International Journal of Climatology. 18(13) 1493–1517.
[3] Jemain A A, Mohd Deni S, Syed Jamaludin S S and Wan Zin W Z 2015 Penyurihan Ikhtisar Data Hujan. (Kuala Lumpur Dewan Bahasa dan Pustaka).
[4] Zhang S 2012 Nearest Neighbor Selection For Iteratively kNN Imputation Journal of Systems and Software 85(11) 2541–2552.
[5] Chow V T, Maidment D R and Mays L W 1998, Applied Hydrology, Mc Graw Hill Book Company
[6] J. L. Paulhus and M. A. Kohler 1952 Mon. Wea. Rev, 80 129-133.
[7] Suhaila J, Deni S M and Jemain A A 2008. Detecting inhomogeneity of rainfall series in Peninsular Malaysia. Asia-Pacific Journal of Atmospheric Sciences. 44(4), 369–380.
[8] Teegavarapu R S V and 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.
[9] Wijngaard J B, Klein Tank A M G and Können G P 2003 Homogeneity of 20th Century European Daily Temperature and Precipitation Series. International Journal of Climatology 23(6) 679–692.
[10] Porth L S, Boes D C, Davis R A, Troendle C A and 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.
[11] Willmott C J 1981 Physical geography 2(2) 184-194.
[12] Kang H.M and Fadhilah Y 2012 Homogeneity test on daily rainfall series in Peninsular Malaysia, International Journal Contemporary Math. Sciences 7(1) 9-12.
[13] Evans J D 1996 Straightforwasrd Statistics for the Behavioral Sciences. University of California: Brooks/Cole Pub. Co.

Acknowledgements

The authors sincerely acknowledge the Department of Irrigation and Drainage Malaysia (DID), for providing the complete daily rainfall data that been used in this study. The work is financed by FRGS/1/2019/STG06/UNIMAP/02/4 provided by the Ministry of Higher Education (MOHE). Special thanks for two anonymous referees whose comments led to a clearer presentation.