SNRU Journal of Science and Technology

Journal home page: snrujst.snru.ac.th

Evaluating Inverse Distance Weighting and Correlation Coefficient Weighting Infilling Methods on Daily Rainfall Time Series

Srisunee Wuthiwongyothin*, Chanyut Kalkan, Jantana Panyavaraporn
Facutly of Engineering, Burapha University, Chonburi, 20131 Thailand.
*Corresponding Author: srisunee.wu@eng.buu.ac.th

Received: 24 December 2020; **Revised:** 13 February 2021; **Accepted:** 19 February 2021; **Available online:** 1 May 2021

Abstract

Encountering missing daily rainfall records is inevitable and filling the gap is a challenging issue. Spatial interpolation is one of the most widely used methods to estimate missing daily rainfall. The method is easy to apply, less time consuming, and requires inexpensive computation than more complex methods. This study attempted to evaluate the inverse distance weighting (IDW) and correlation coefficient weighting (CCW) methods and compare each method's performance over the Upper Ping River Basin. Daily rainfall data from 92 stations over about 65 years (1953 – 2017) were obtained. After screening, 44 stations were used in this study. Before implementing infilling methods, cluster analysis using K-means was applied to group the station into three sub-regions. Three and four neighbor stations (source stations, SSs) were tested to find the optimal number of SSs. Six target stations from three sub-areas were chosen to test the infilling method with different percentages of missing values to represent various numbers of missing data. The study results revealed that CCW provided better performance than the IDW method. The optimal number of source stations to estimate missing data was four stations assessed by evaluating mean values, R correlation and similarity index. Moreover, CCW also yielded less error of mean absolute error (MAE) and root means square error (RMSE) compared to IDW. Varying the percentage missing values between 5%, 10%, 20%, 30%, 40%, and 50% revealed that each infilling method was not sensitive to the percentage of missing data.

Keywords: infilling method; daily rainfall; missing data; inverse distance weighting method; correlation coefficient weighting method

1. Introduction

Rainfall data is fundamental and crucial for studying water resources, hydrology, climatology, meteorology, environment, agriculture, and ecology. Accurate rainfall data can provide reliable and meaningful analyses in any related study. The completeness of the rainfall dataset could directly affect the goodness-of-fit of hydrological, climate, and environmental models because the data is used to set up, tune, and improve model performance via calibration and validation [1]. In contrast, an incomplete dataset can create difficulties in estimating model parameters or identifying processes in those studies, which leads to misinterpretation for both spatial and temporal variations of all involved indicators. Thus, an accurate dataset is helpful to many fields of study.

Encountering missing daily rainfall data is an inevitable and common issue due to various problems extending from human problems, natural conditions, and instrumental issues [2]. General instrumental issues are instrumental malfunction or interruption of telecommunication. Some examples of human problems are the absence of data collectors for manual rain-gauges, processing data issues, measuring

techniques, and relocation of the stations. Natural conditions include heavy storms or lighting that damages rain-gauge systems.

Typically, rainfall data varies in both spatial and temporal characteristics. Daily rainfall characteristics contain more noise than monthly rainfall. Daily rainfall depth is deterministic continuous information, while rainfall occurrence (rain day and no-rain day) behaves as discrete data [3]. Monthly rainfall exhibits continuous data based on a seasonal pattern. Seasonal and annual rainfall present a more homogeneous pattern because it is constructed by aggregating smaller time scale data. Normally, different time scales and different locations contain different behavior patterns of the rainfall information. In this study, daily rainfall is chosen for study because it is generally measured daily in Thailand and is the basic data for monthly and annual rainfall time series.

Handling missing time-series data such as daily rainfall is a challenging issue. A variety of infilling methods ranging from simple to sophisticated approaches have been studied. The infilling methods can be grouped as spatial interpolation such as simple arithmetic averaging, normal ratio, inverse distance weighting (IDW) and its modification, ordinary kriging, co-kriging, correlation coefficient weighting (CCW), and geographic coordination (GC). Generally, spatial interpolation techniques are the most widely used in the hydrology field. Statistical approaches such as simple linear regression, multiple regression [4], and multivariate regression [5] are common approaches. More complex statistical methods are probabilistic based or stochastic methods [6, 7]. Over the past few decades, machine learning techniques have been developed, such as artificial neural networks (ANN) [8] and fuzzy neural networks (CFNN).

A significant number of missing daily rainfall data could be a single day to several months or years. Apart from the number of missing data points, the nature of missing data occurrence is a concern. Rubin [9] classified missingness mechanisms of data into three categories by looking at an entire sample dataset separated into observed and missing data. Missing completely at random (MCAR) refers to missing data, and its probability is unrelated to both the observed and missing data response. If the missing data and its probability are related to the available observed data but unrelated to the missing values, this missing data mechanism is missing at random (MAR). The last mechanism is missing not at random (MNAR), which means missing data is not random because its probability depends on both observed and missing data [10]. Presti et al. [3] tested a statistical missing mechanism on missing daily rainfall of Italy rain stations. The study confirmed that missing daily rainfall data follows the MAR missingness mechanism.

The spatial interpolation technique is the most widely used to estimate missing daily rainfall because it is easy to apply, less time consuming, and requires inexpensive computation. Teegavarapu and Chandramouli [11] suggested more complicated infilling technique may not significantly improve estimation accuracy. Recently, Barrios et al. [12] mentioned that ANN and MLR were not significantly different from the IDW method. However, popular machine learning methods mostly need higher computational resources and are more time-consuming. Therefore, traditional spatial estimation methods such as IDW and normal ratio still exist and are commonly used today. These approaches yield enough accuracy to fill missing values with economical computation effort. However, few studies focusing on spatial interpolation conventional methods have tested, evaluated and compared these different infilling methods performance [11, 13 – 15] up to now. Although the development of a variety of infilling methods has been studied, there is no consensus of scientific publications about the best infilling method. Rainfall information is highly variable and produces no specific spatial- and temporal characteristics for a large region. The study used missing daily rainfall information in the study area, the upper Ping River Basin in Thailand.

This study attempted to investigate traditional IDW and conventional CCW spatial interpolation techniques with clustering analysis using K-means. IDW has been one of the most widely used methods to fill the gap in daily rainfall records for many decades. CCW is like IDW but replaces distances with correlation coefficients as the weight. Teegavarapu and his colleges [11, 16] found that CCW would

improve missing rainfall data estimates. In Thailand, CCW has not yet been tested, investigated, and compared to the IDW method specifically for this selected watershed area.

Cluster analysis has been applied before estimating missing values. J. Kim and J.H. Ryu [2] proved that using K-means can improve estimates of missing data regardless of infilling methods, especially for a large region like Idaho in the U.S., which has a highly spatial variation of rainfall. Thus, the objective of this study is to 1) investigate IDW and CCW infilling daily rainfall based on three and four neighbor stations and 2) evaluate CCW and IDW performance with different percentages of missing values of 5%, 10%, 20%, 30%, 40% and 50%. Results from this study could reveal a better spatial infilling method to fill missing daily rainfall time series.

2. Materials and Methods

Data Collection and Case Study Area

Daily rainfall data from ground-based rain gauges of 92 stations located in the upper Ping River Basin were collected from 1 January 1953 to 31 December 2017 (65 years). The data were obtained from the Thai Meteorology Department (TMD), the Royal Irrigation Department (RID), and the Department of Water Resources (DWR). Upper Ping River Basin is one of the main watersheds in northern Thailand. It is a major drainage area of about 26,111 km^2, contributing about 5.63 billion cubic meters (bm^3) of water annually to the Bhumibol reservoir. The dam can also generate hydroelectricity with the highest capacity of 713 MW per year. The study area topography is mostly hilly and mountainous, covered mainly by subtropical forest and vegetation of about 80%. The altitude of the catchment varies between +300 m a.s.l. (meter above mean sea level) near the dam (at the south) up to +2,595 m a.s.l. at the mountain hilltop (in the north at Doi Inthanon). The study area and rain stations are presented in Fig. 1.

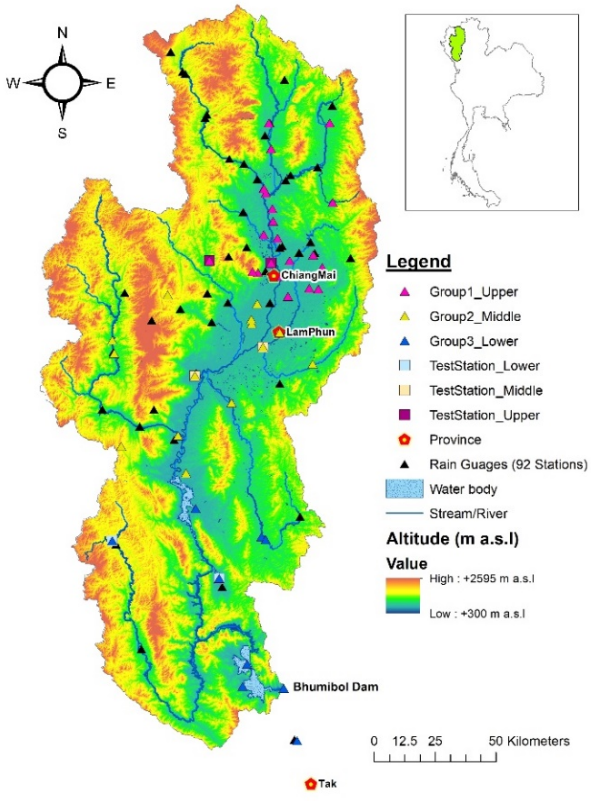


Fig. 1 Upper Ping River Basin topography and rain gauge stations

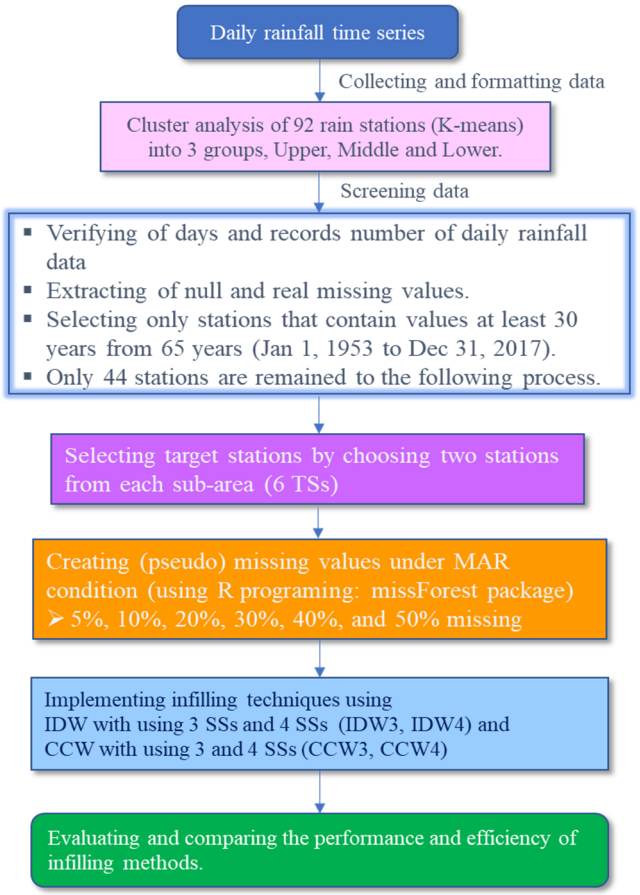


Fig. 2 Study Procedure

Study Method

In this study, IDW and CCW infilling methods were chosen to estimate and fill the gap of missing daily rainfall values. The estimation of missing values used data from the nearest (IDW) or highest correlation (CCW) neighbor stations that provide available data on the same days. These two methods can be inferred as the single best estimation method. The term "target stations (TS)" refers to the station that needs to fill the missing values gap. The term "source station (SS)" refers to neighbor stations used to estimate and fill missing values for the target station.

Inverse Distance Weighting Method (IDW)

The IDW method is based on the proximity of neighboring stations (or SSs) to the target station (TS). This method used distance between TS and SSs as a weighted factor by calculating as follows:

$$W_i = \frac{d_i^{-n}}{\sum_{i=1}^{m} d_i^{-n}} \tag{1}$$

Where d_i refers to the distance between TS and SS, and m is the total number of SSs being used. The value of power "n" usually ranges from 1 to 6. However, the most-used value of n is 2 [11, 12], which is also used in this study. The IDW weight decreases as the distance increases. The higher the value of power n is, the greater is the distance between the stations and the less weight any one station gives to a reading assigned to a neighboring station.

Correlation Coefficient Weighting Method (CCW)

The CCW method is like the IDW, but the CCW weighting factor is based on the correlation coefficient between TS and SS instead of distance. The CCW weighted factor is given as follows:

$$W_i = \frac{R_i}{\sum_{i=1}^{m} R_i} \tag{2}$$

where R_i is the correlation coefficient (R) between the TS and SS daily rainfall datasets.

Then the generated daily rainfall of both infilling methods can be estimated by

$$P_x = \sum_{i=1}^{m} W_i P_i \tag{3}$$

Cluster Analysis: K-means

K-means clustering is a technique used to classify data into K groups. The method is a simple and popular unsupervised algorithm to identifies the K number of centroids. Then, every data point is allocated to the nearest group by having minimum centroids, which is the minimum sum of the squared error (SSE). SSE is calculated by

$$SSE = \sum_{i=1}^{K} \sum_{n \in S_j} |x - m_i|^2 \tag{4}$$

Where x is a data point in group S_j, and m_i is the centroid of the dataset of the group. K is the designed number of groups.

Performance Measures Criteria

The performance of infilling methods was compared between the estimated values (replacing NA values) with their corresponding real observed values. Mean values and Pearson correlation (R) were evaluated. Common error measures used in this study were mean absolute error (MAE) and root means

square error (RMSE). The similarity index (S-index) was used to assess the agreement of the dataset. The S-index values range from 0 (disagreement) to 1 (perfect agreement). These measures indices are given as follows:

$$R = \frac{\sum_{i=1}^{N}(P_{obs,i}-\bar{P}_{obs,i})(P_{x,i}-\bar{P}_{x,i})}{\sqrt{\sum_{i=1}^{N}(P_{obs,i}-\bar{P}_{obs,i})^2 \sum_{i=1}^{N}(P_{x,i}-\bar{P}_{x,i})^2}} \quad (5)$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|P_{x,i} - P_{obs,i}| \quad (6)$$

$$RMSE = \sqrt{\sum_{i=1}^{N}\frac{(P_{x,i}-P_{obs,i})^2}{N}} \quad (7)$$

$$S-index = 1 - \frac{\sum_{i=1}^{N}(P_{obs,i}-\bar{P}_{obs,i})^2}{\sum_{i=1}^{N}(|P_{x,i}-\bar{P}_{x,i}|+|P_{obs,i}-\bar{P}_{obs,i}|)^2} \quad (8)$$

Where N is the total number of calculated data, and $P_{x,i}$ and $P_{obs,i}$ are the estimated and real observed values, respectively.

In addition, other statistical values such as maximum values, standard deviation, and 99th percentiles were also calculated to assess the efficiency of the infilling method in estimating heavy rainfall.

Study Procedure

Fig. 2 displays a summary of the overall steps implementing in this study. Data collected from January 1953 to December 2017 (65 years) at 92 rainfall stations managed by the TMD, RID, and DWR were collected. A brief explanation of each step is summarized herein.

Data preparation

First, data from different organizations were arranged into a column format. Each row refers to date or time, and each column is rainfall data. Then, the data were checked for quality and missing values. Next, the real percentage of missing data over the study period was calculated.

Clustering analysis

Clustering analysis used geographical information (latitude and longitude) to classify and group all the 92 stations into three groups: upper, middle and lower parts, because the basin has an elongated shape.

Screening the data and deleting null and missing values

Data from 92 stations were checked and screened. In this study, the stations that have equivalently available data for at least 30 years were chosen. Only 44 stations remained in the study after applying this criterion.

Selecting target stations

This study selected six target stations from all sub-areas to test the infilling methods by choosing from the maximum available values among the rain stations within the group. Herein, the six target stations are 327009 (UTS1) and 327501 (UTS2) from the upper part, 327003 (MTS1) and 329003 (MTS2) from the middle part, and 327008 (LTS1) and 329006 (LTS2) from the lower area.

Creating missing data

The R-program with the missForest package tool generated missing data (NA) of the target stations under MAR by removing 5%, 10%, 20%, 30%, 40% and 50% of the data. Table 1 presented the total number of data points and the number of missing percentage values.

Table 1 Total number of data and number of missing values of the selected six target stations.

Target Station	% Missing >>	5%	10%	20%	30%	40%	50%
UTS1(327009)	Total number of NA	1,170	2,340	4,681	7,021	9,361	11,702
	Total data	23,406	23,406	23,406	23,406	23,406	23,406
UTS2(327501)	Total number of NA	1,159	2,318	4,637	6,954	9,274	11,591
	Total data	23,186	23,186	23,186	23,186	23,186	23,186
MTS1(327003)	Total number of NA	1,165	2,331	4,662	6,993	9,325	11,656
	Total data	23,314	23,314	23,314	23,314	23,314	23,314
MTS2(329003)	Total number of NA	1,106	2,213	4,426	6,640	8,853	11,067
	Total data	22,135	22,135	22,135	22,135	22,135	22,135
LTS1(327008)	Total number of NA	1,075	2,151	4,302	6,453	8,604	10,756
	Total data	21,513	21,513	21,513	21,513	21,513	21,513
LTS2(329006)	Total number of NA	1,044	2,088	4,176	6,264	8,352	10,440
	Total data	20,881	20,881	20,881	20,881	20,881	20,881

Implementing infilling methods

Then, IDW and CCW were applied using three and four SSs to test and find the optimal SSs for estimating missing values.

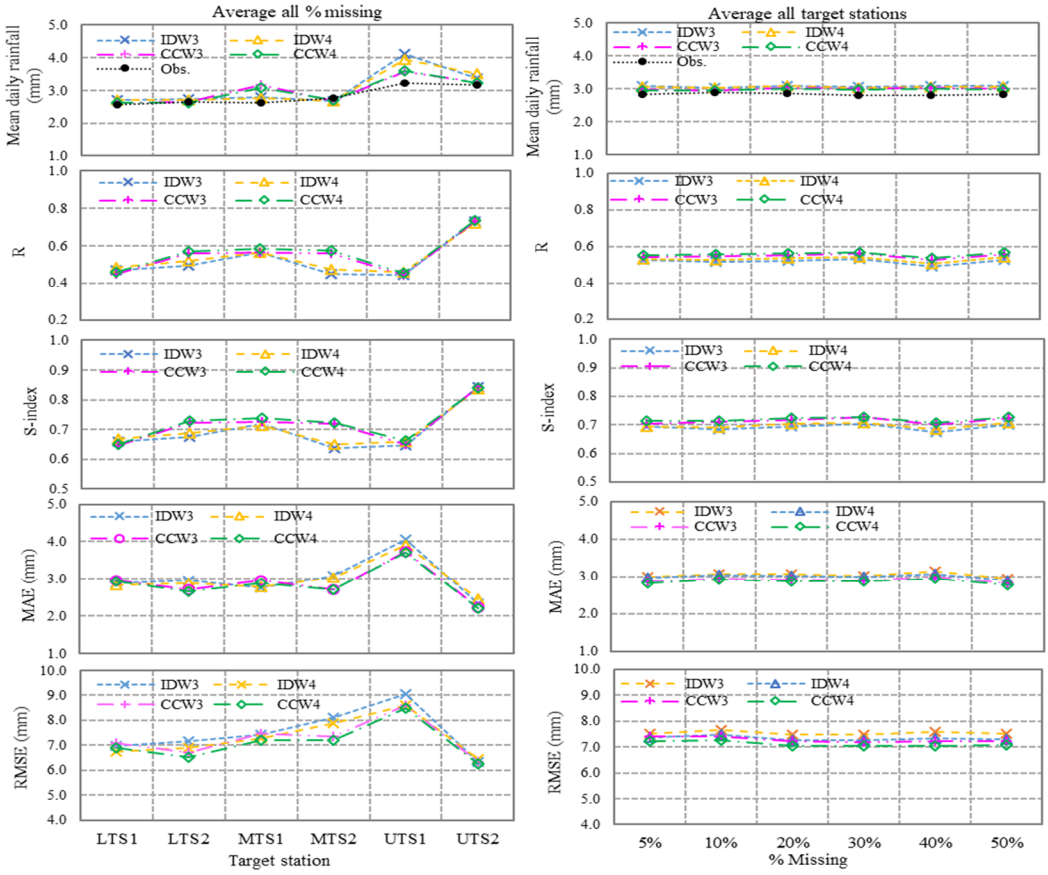
Identify infilling methods performance

Six TSs with four infilling tests (IDW3, IDW4, CCW3, and CCW4) and six different percentage missing ($6 \times 4 \times 6 = 144$ cases) were evaluated. The evaluating performances comprise mean, R correlation, S-index, RMSE, and MAE. Maximum, standard deviation, variance, and 99th percentile of the observed dataset and estimated values were compared to check infilling methods efficiency.

3. Results and Discussion

The performances of each infilling method: IDW3, IDW4, CCW3, and CCW4, are presented in Fig. 3, which shows mean values, correlation R, S-index, MAE, and RMSE. Fig. 3 a) shows the measurement values by averaging all percentage missing and comparing with each target station. In contrast, 3 b) compared the resulting performance of each percentage missing aspect. The plot of mean values, CCW4, performed closest to the observed data followed by CCW3, IDW4, and IDW3. CCW4 also yielded the highest R and S-index, followed by CCW3, IDW4 and IDW3. CCW4 gave the lowest measurement error values of MAE and RMSE, meaning that it produces less error than the other methods. Thus, among all methods tested in the current study, the CCW4 infilling method exhibited higher performance than CCW3, IDW4 and IDW3.

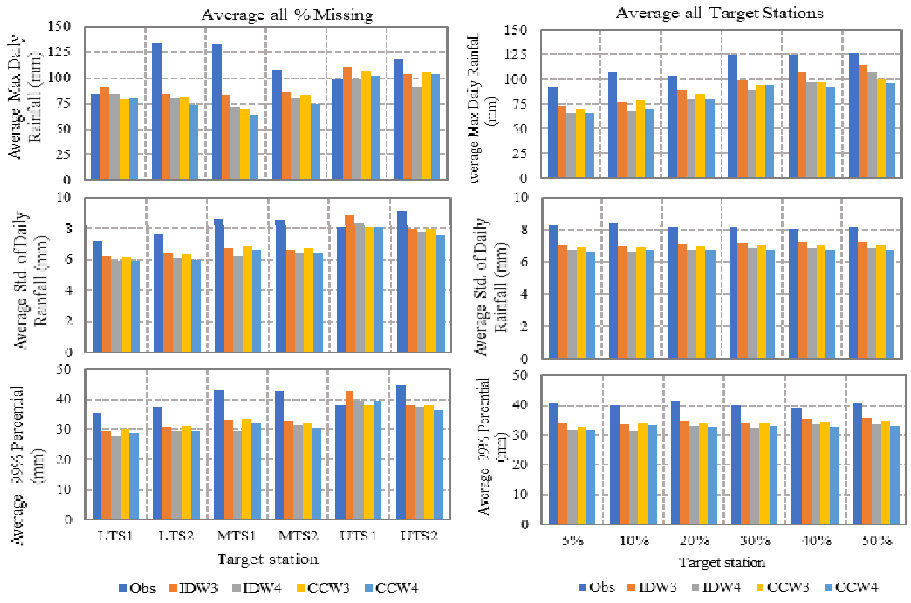
This study result was similar to a study from Teegavarapu [11, 16] that found that CCW is superior to IDW and some other traditional methods. In addition, by comparing the number of neighbor stations (SSs) used in estimation, utilizing data from four SSs could improve estimated results better than using three SSs. Varying different percentage missing values (5%, 10%, 20%, 30%, 40%, 50%) to represent various numbers of missing data as shown in Fig 4 b), it is clear that there is only a small variation of results depending on the different percentage missing amounts. Thus, the infilling methods are not sensitive to the percentage of missing values, which was similar to the work of Suhaila and colleagues [17].



a) Comparison of all stations

b) Comparing with different percentages of missing values

Fig. 3 Line plots comparing mean, R, S-index, MAE, and RMSE of each infilling method.



a) Comparison of all stations

b) Comparing with different percentages of missing values

Fig. 4 Comparison of maximum daily rainfall, standard deviation, and 99th percentile of observed rainfall for each infilling method. (Max refers to maximum, and Std is the standard deviation)

Although CCW4 performed best, Fig. 4 revealed that IDW3 mostly produced closer statistical values of maximum daily rainfall, standard deviation and 99th percentile of daily rainfall to the observed data, followed by CCW3. This result suggests that using three SSs to estimate missing values could produce more accurate values for heavy rainfall through the maximum and 99th percentile than using four SSs. Using three SSs yields higher variation (but closer to observed data), such as standard deviation or variance, of an estimated dataset compared to four SSs. Generally, the higher the number of stations used to estimate, the more average are the estimated values and the less variation is obtained.

4. Conclusion

This study examined methods of filling missing daily rainfall time series data based on spatial interpolation methods over the upper Ping River Basin. Cluster analysis using K-means was used to group all 92 stations that collected data from 1 January 1953 to 31 December 2017 into three sub-area based on coordinates. However, after screening and checking null and missing values, 44 stations remained in the study because they contained data for at least 30 years. Then, two representative target stations from each sub-area were selected based on the maximum available data. IDW is one of the most intensive methods used to estimate missing daily rainfall. IDW obtains available data from SSs to interpolate values with a weighting factor calculated by the distance between the TS and the SS. CCW is like the IDW method, but CCW used a weighting factor from the correlation coefficient between the TS and SSs.

This study found that CCW gave a better performance for R, S-index, MAE, and RMSE than IDW. Using four SSs (e.g., CCW4 and IDW4) was superior to using three SSs to estimate missing values. However, using three SSs (IDW3 and CCW3) would produce better heavy rainfall and standard deviation statistics. Lastly, by varying the different percentage missing values over the range 5%, 10%, 20%, 30%, 40%, and 50%, the results revealed no infilling method was sensitive to the percentage of missing data.

5. Suggestions

It is noted that spatial interpolation methods have some limitations. Primarily, they overestimate the number of rain days but underestimate extreme rainfall, as discussed in some studies [6,18]. Hence, any study needs to be concerned about extreme rainfall events; other techniques, such as probability-based or stochastic approaches, might produce better results than spatial interpolation infilling methods.

6. Acknowledgement

The author acknowledges to TMD, RID and DWR for supporting daily rainfall data. This work was supported by the Coordinating Center for Thai Government Science and Technology Scholarship Students (CSTS), National Science and Technology Development Agency (NSTDA) under funding number JRA-CO-2563-13103-TH. The author thankful to the anonymous reviewers who have comments which helped to improve the quality of this paper.

7. References

[1] B. García, P. Sentelhas, G. Sparovek, L. Tapia, Filling in missing rainfall data in the Andes region of Venezuela, based on a cluster analysis approach, Rev. bras. meteorol. 14 (2006) 225 – 233.

[2] J. Kim, J.H. Ryu, A heuristic gap filling method for daily precipitation series, Water Resour Manag. 30 (2016) 2275 – 2294.

[3] L.R. Presti, E. Barca, G. Passarella, A methodology for treating missing data applied to daily rainfall data in the Candelaro River Basin (Italy), Environ Monit Assess. 160 (2010) 1 – 22.

[4] D. Mora, G. Wyseure, P. Willems, Gap filling based on a quantile perturbation factor technique, 11th International Conference on Hydroinformatics, New York City. 17 – 21 August 2014, 1 – 8.

[5] C. Simolo, M. Brunetti, M. Maugeri, T. Nanni, Improving estimation of missing values in daily precipitation series by a probability density function-preserving approach, Int J Climatol. 30 (2010) 1,564 – 1,576.
[6] M.M. Hasan, B.F.W. Crokea, Filling gaps in daily rainfall data: a statistical approach, 20th International Congress on Modelling and Simulation, Adelaide, Australia, 1-6 December 2013, 380 – 386.
[7] H. Aksoy, Use of gamma distribution in hydrological analysis, Turk J Engin Environ Sci. 24 (2000) 419 – 428.
[8] M.A. Malek, S. Harun, S.M. Shamsuddin, I. Mohamad, Reconstruction of missing daily rainfall data using unsupervised Artificial Neural Network, World Acad Sci Eng Technol. 44 (2008) 616 – 621.
[9] D.B. Rubin, Inference and missing data, Biometrika. 63 (1976) 581 – 592.
[10] Y. Xia, P. Fabian, A. Stohl, M. Winterhalter, Forest climatology: estimation of missing values for Bavaria, Germany, Agric for Meteorol. 96 (1999) 131 – 144.
[11] R.S.V. Teegavarapu, V. Chandramouli, Improved weighting methods, deterministic and stochastic data-driven, models for estimation of missing precipitation records, J. Hydrol. 312 (2005) 191 – 206.
[12] A. Barrios, G. Trincado, R. Garreaud, Alternative approaches for estimating missing climate data: application to monthly precipitation records in South-Central Chile, For. Ecosyst. 5 (2018) 1 – 10.
[13] Y. Tan, J.L. Ng, Y. Huang, Estimation of missing of missing daily rainfall during monsoon seasons for tropical region: A comparison between ANN and conventional methods, Carpathian J. Earth Environ. Sci. 15 (2020) 103 – 112.
[14] N.A. Rahman, S.M. Deni, N.M. Ramli, N. Shaadan, The improvement of missing rainfall data estimation during rainy season at Ampang station, Int. J. Eng. Technol. 7 (2018) 204 – 212.
[15] J.E. Kasri, A. Lahmili, O. Latifa, L. Bahi, S. Halima, M.A. Mitach, Comparison of the relevance and performance of filling in gaps methods in rainfall datasets, Int. J. Civ. Eng. Technol. 9 (2018) 992– 1000.
[16] R.S.V. Teegavarapu, Estimation of missing precipitation records integrating surface interpolation techniques and spatio-temporal association rules, J. Hydroinformatics. 11 (2019) 133 – 146.
[17] J. Suhaila, M.D. Sayang, A.A. Jemain, Revised Spatial Weighting Methods for Estimation of Missing Rainfall Data, APJAS. 44 (2008) 93 – 104
[18] R.S.V. Teegavarapu, Statistical corrections of spatially interpolated missing precipitation data estimates, Hydrol Process. 28 (2014) 3,789 – 3,808.