PAPER • **OPEN ACCESS**

Prediction of Missing Data in Rainfall Dataset by using Simple Statistical Method

To cite this article: Izzati Amani Mohd Jafri *et al* 2020 *IOP Conf. Ser.: Earth Environ. Sci.* **616** 012005

Prediction of Missing Data in Rainfall Dataset by using Simple Statistical Method

Izzati Amani Mohd Jafri[1], Norazian Mohamed Noor[1,2], Ahmad Zia Ul-Saufie[2,3] Annas Suwardi[4]

[1]School of Environmental Engineering, Universiti Malaysia Perlis, Kompleks Pengajian Jejawi 3, 02600 Arau, Perlis, Malaysia
[2]Sustainable Environment Research Group, Center of Excellence Geopolymer and Green Technology (CEGeoGTech), Universiti Malaysia Perlis, Kompleks Pengajian Jejawi 2, 02600 Arau, Perlis, Malaysia
[3]Faculty of Science Computer and Mathematics, Universiti Teknologi Mara, Cawangan Pulau Pinang, Malaysia
[4]Universitas Negeri Makasar, Faculty of Mathematics and Natural Sciences, Kampus UNM Parangtambung, Jalan Daeng Tata Makassar, Indonesia

E-mail: norazian@unimap.edu.my

Abstract. Almost all of the data obtained from hydrological station contains missing data. Usually, this problem occurs due to equipment failures, maintenance work and human error. Incomplete dataset will reduce the ability of a statistical analysis and can cause a bias estimation due to systematic differences between observed and unobserved data. In this study, four simple statistical method such as Series Mean, Average Mean Top Bottom, Linear Interpolation and Nearest Neighbour were applied to predict the missing values in a rainfall dataset. An annual daily data for rainfall from nine selected monitoring station (from 2009 until 2018) were described using descriptive statistic. Then, the dataset were randomly simulated into 4 percentages of missing (5%, 10%, 15% and 20%) by using statistical package for social sciences software. The performance of this imputation methods were evaluated by using four performance indicators namely Mean Absolute Error, Root Mean Squared Error, Prediction Accuracy, and Index of Agreement. Overall, Linear Interpolation method was selected as the best imputation method to predict the missing data in the rainfall dataset.

1. Introduction

Understanding and predicting on the climate change over a period of time is very important to prepare from any extreme event that might occur and to plan for a better environmental management. As climate is the average weather, the characteristics involved in defining weather includes surface temperature, rainfall, relative humidity, atmospheric pressure and wind speed. The two primary parameters in measuring climate change are temperature and precipitation [1]. In Malaysia, the standard instrumentation, installation and maintenance for rainfall stations in Malaysia is managed by the Department of Irrigation and Drainage (DID). As DID have been using the rain gauge for over

than 50 years, it is a common problem that the rainfall dataset often have a missing gap due to several factor includes human error and instrument error such as measurement error and damaged measuring instrument.

This hydrological data is very important in predicting and understanding the climate change. It is likely to brings a significant effect on Malaysia, includes increasing sea levels and rainfall, increasing flood risks and droughts. Although this data can not necessarily stop any extreme events from occurring, however it can help our society to prepare and plan for extreme events [2], provide timely warnings and identify alternate method to overcome the problems [3]. In view, a complete dataset is important to determine climate changes. This study will focus on making prediction of missing observation in rainfall dataset by using several statistical methods. The methods will be evaluate using some performance measures to determine the best prediction method that can be used to fill in the rainfall data.

2. Experimental

In this study, daily rainfall data from 9 rainfall stations at selected areas in Pulau Pinang were used. The selected data is within the range of 10 years which is from 2009 to 2018. The 9 monitoring stations were divided to the north (Simpang Ampat, Ladang Batu Kawan and Bagan Buaya), the centre (Kompleks Prai, Juru Dam, Permatang Rawa and Cherok To' Kun) and the south (Permatang Binjai and Bumbong Lima). The data of the rainfall is collected by days in millimeter (mm). The descriptive statistic of rainfall data such as mean, median, standard deviation, variance and range were studied by using SPSS. The characteristic of missing data pattern such as the percentages of missing and gaps length of missing were discussed and portrayed by using visual graph. Then, the data were simulated into four percentages of missing data i.e. 5%, 10%, 15% and 20%. After that, five imputation methods which is Series Mean (SM), Nearest Neighbour (NN), Linear Interpolation (LI) and Mean Top Bottom (MTP) were applied to fill in the simulated missing data. The proposed imputation methods were compared to each other by using four performances indicators to select the best imputation method.

2.1. Simulation of Missing Data

Four percentages of simulated missing data were used for evaluating the accuracy of imputation techniques that were 5%, 10%, 15% and 20%. In order to simulate the missing data, there must be a complete dataset as performances of the evaluated methods were compared between the predicted and observed value [4]. Therefore, the mode value was used to complete the dataset. Then, the data was randomly simulated by the software into 4 percentages of missing which is 5%, 10%, 15% and 20%. Generally, the pattern of missing observations for all percentages of the simulated missing data do not vary much from one to another. This is one important aspect in simulating the missing measurement as it will not interrupt the main structure of dataset [5].

2.2. Imputation Method

There are several imputation methods used to fill the missing values of four percentages of simulated missing data. The imputation methods used were Series Mean (SM), Mean Top Bottom (MTB), Nearest Neighbor (NN), Linear Interpolation (LI).

2.2.1. Series Mean (SM). It is a simple method which used the average mean values of the data series to be imputed in the missing observation. In this study, the average rainfall values of each station were used to fill in missing values.

$$\tilde{x} = \frac{\sum f_x}{n} \quad (1)$$

where $\sum f_x$ represents total daily rainfall of a stations, n is the total number of days.

2.2.2. Mean Top Bottom (MTB). A simple method which used the average mean values of the data above and below the missing value. Then, the mean values will fill in the missing data.

2.2.3. Nearest Neighbor (NN). Nearest Neighbour (NN) imputation methods was the method to replace the missing data with the nearest value to the missing datum [6]. In this study, the missing values will be fill in by the nearest above values of the rainfall data. Nearest Neighbour imputation was the simplest method available, in that the end points of the gaps were used as estimates for all the missing values. The equation is [6];

$$y = y_1 \text{if } x \leq x_1 + (x_2 - x_1)/2 \quad (2)$$
$$y = y_2 \text{ if } x > x_1 + (x_2 - x_1)/2$$

2.2.4. Linear Interpolation (LI). Linear Interpolation (LI) method fill the gaps of missing data by replacing the missing value with average value of the before and after data in sequential pattern [4].This method performed better for a short gap of missing data [7]. The equation of LI is written as follow [6]:

$$y - y_1 = \frac{y_2 - y_1}{x_2 - x_1}(x - x_1) \quad (3)$$

Where y is the missing observation, x is the time of point of missing observation. x_1 and y_1 are the coordinates of the starting point of the gap, x_2 and y_2 are the coordinates of the end point of the gap.

2.3. Performances Indicator

Four performances indicator will be used to measure the effectiveness of the imputation methods used in this study such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Index of Agreement (d_2), and Prediction Accuracy (PA). The performances for each imputation method were displayed in the form of rank and the best imputation method for overall will be selected. Table 1 shows the performance indicator formulae:

Table 1. Performances Indicator [6].

Performance Indicator	Formula	Description
Normalized Absolute Error (NAE)	$NAE = \sum_{i=1}^{n} \frac{Abs(P_i - O_i)}{\sum_{i=1}^{n} O_i}$	NAE value closer to zero indicates better method
Mean Absolute Error (MAE)	$RMSE = \frac{1}{N}\sum_{i=1}^{N} \lvert P_i - O_i \rvert$	MAE value closer to zero indicates better method
Index of Agreement (d_2)	$d_2 = 1 - [\frac{\sum_{i=N}^{N} (P_i - O_i)^2}{\sum_{i=N}^{N} (\lvert P_i - \overline{O} \rvert + \lvert O_i - \overline{O} \rvert)^2}$	$d_2$value closer to 1 indicates better method
Prediction Accuracy (PA)	$PA = \sum_{i=1}^{N} \frac{[(P_i - \overline{P})(O_i - \overline{O})]}{(N-1)\sigma_P \sigma_O}$	PA value closer to 1 indicates better method

Where N is the number of imputations, O_i is the observed data points, P_i is the imputed sata points, $\overline{P}$ is the average of imputed data, $\overline{O}$ is the average of observed data, σ_P is the standard deviation of the imputed data, and σ_O is the standard deviation of observed data.

3. Results and Discussions

3.1. Characteristic of Rainfall Dataset

In this study, the data of daily rainfall dataset at the 9 stations in Pulau Pinang were analysed. Table 2 shows the summary of descriptive statistic for rainfall distribution at nine station at Penang within ten years.

Table 2. The descriptive statistic for rainfall distribution at nine station at Penang within ten years.

	Valid N	Missing N	Mean	Median	Mode	Std Dvtn	Var	Range	Min	Max	Percentile 25	Percentile 75
SA	1737	53	13.33	6.5	0.5	17.18	295.22	172.5	0.5	173.0	2.0	6.5
LBK	1672	118	13.05	6.5	0.5	17.38	302.00	139.5	0.5	140.0	1.5	6.5
BB	1683	107	13.08	6	0.5	17.74	314.66	155.0	0.5	155.5	1.5	6.0
KP	1685	105	12.68	6	0.5	18.97	359.89	364.0	0.5	364.5	1.5	6.0
JD	1329	461	13.42	6	0.5	24.69	609.68	498.4	0.1	498.5	1.5	6.0
PR	1708	82	14.16	6.75	0.5	19.41	376.83	310.5	0.5	311.0	2.0	6.8
CTK	1790	0	14.89	7.5	0.5	20.76	430.96	367.4	0.1	367.5	2.0	7.5
PB	1548	206	14.51	6.5	0.5	22.20	492.77	393.0	0.5	393.5	2.0	6.5
BL	1692	98	13.54	6.5	0.5	19.55	382.04	271.0	0.5	271.5	2.0	6.5

Where: SA – Simpang Ampat, LBK – Ladang Batu Kawan, BB – Bagan Buaya, KP – Kompleks Prai, JD – Juru Dam, PR – Permatang Rawa, CTK – Cherok To' Kun, PB – Permatang Binjai, BL – Bumbong Lima.

From the table, Juru Dam station recorded the most highest missing rainfall measurements with almost 35% of the data were missing. The mean showed that the station toward the south of Pulau Pinang (from station Permatang Rawa downward) recieved higher amount of rainfall compared to the north. Highest variation was found out at Juru Dam station due to high range of dataset between the maximum and minimum measurement.

3.2. The Performances of Imputation Method

Figure 1 shows the overall performance and error measures for 5%, 10%, 15% and 20% simulated missing data. Referring to the figure, Linear Interpolation (LI) and Mean Top Bottom (MTB) method show the best performances for all percentages of simulated missing data. This is because almost all of the station showed small error values on mean absolute error (MAE) and normalized absolute error (NAE), while the performance value in prediction of accuracy (PA) and index of agreement (d_2) were high for both LI and MTB method. Series Mean (SM) method was shown to be the worst imputation method for estimation of all 5%, 10%, 15% and 20% simulated data. These method contribute large error and small performance compared to other methods.

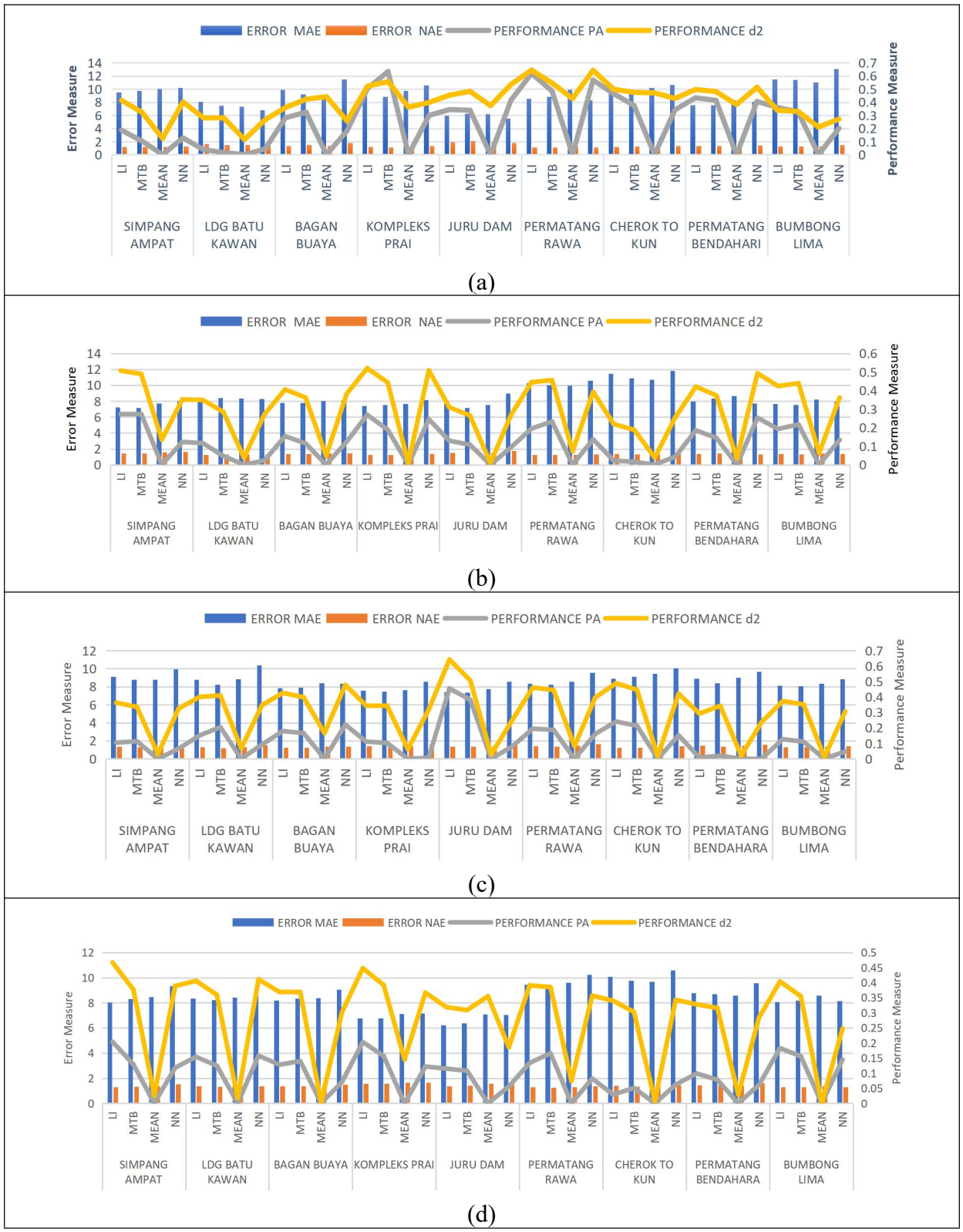


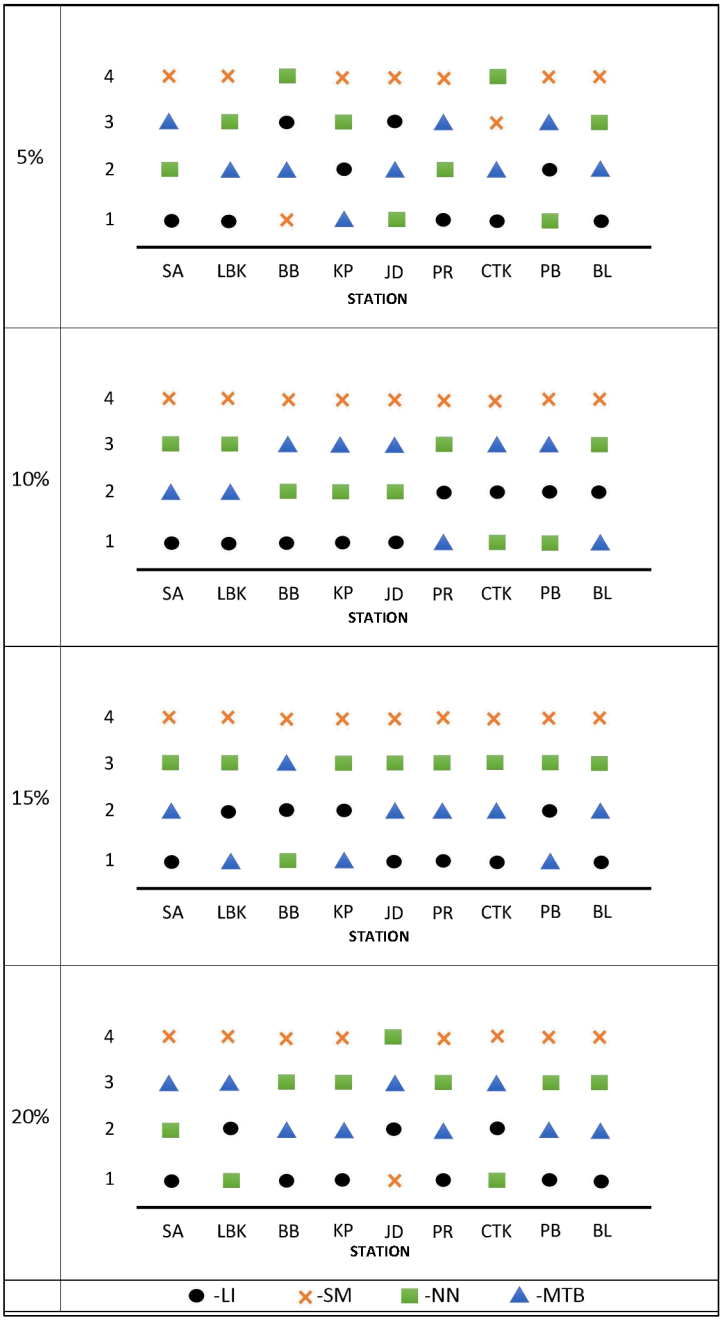
Figure 1. The overall performance and error measures for (a) 5% and (b) 10% (c)15% and (d) 20% simulated missing data.

Figure 2 shows the overall rank of all imputation method at 5%, 10%, 15% and 20% of simulated missing data. Generally, the best imputation method to fill in all percentages of simulated missing data were Linear Interpolation (LI) method. The second best method for 20% of simulated missing data were MTB method, followed by NN method and lastly SM method. This was evidenced when almost all performance indicators in each of the percentages agreed that LI method as the most appropriate for replacing the missing values. It seem that the performance of LI method was very good although the variety of the missing gaps in the rainfall data is different. LI method was shown to have great performance and low error to impute the 5%, 10%, 15% and 20% simulated missing data.

Table 3 shows the summary of each imputation method. The performances of LI method is the best compared to the other methods in estimating missing measurement in rainfall data. LI method show better performance because it fills the gaps of missing data by replacing this missing value with average value of data before and after missing data in sequential pattern [4]. MTB and NN method showed moderate performance to impute the missing values and they also showed inconsistent performance to estimate the missing data in different percentages of missing. Meanwhile, SM method was shown as the worst method to impute the missing data since the performance of this method in each percentages was lowest for almost all station.

Table 3. The summary of each imputation methods.

Tiers	Methods	Description
Good	Linear Interpolation (LI)	• Excellent performance in all percentages of simulated missing data. • The error in estimation tends to increase when the percentages of missing is higher.
Moderate	Mean Top Bottom (MTB) Nearest Neighbour (NN)	• Moderate performance in all percentages of simulated missing data. • The estimation tends to converge when the gaps become larger.
Bad	Series Mean (SM)	• Worst performance for all percentages of simulated missing data. • Lead to bias due to uncertainty are not covered.



Where: LI – Linear Interpolation, MTB – Mean Top Bottom, SM – Series Mean and NN – Nearest Neighbour, SA – Simpang Ampat, LBK – Ladang Batu Kawan, BB – Bagan Buaya, KP – Kompleks Prai, JD – Juru Dam, PR – Permatang Rawa, CTK – Cherok To' Kun, PB – Permatang Bendahari, BL – Bumbong Lima.

Figure 2. The rank of all imputation method at 5%, 10%, 15% and 20% of simulated missing data.

4. Conclusions

10 years of rainfall data at selected station in Pulau Pinang were used to simulate the dataset into 4 percentages of missing data i.e. 5%, 10%, 15% and 20%. Four simple statistical method were applied namely Linear Interpolation (LI), Mean Top Bottom (MTB), Nearest Neighbour (NN) and Series Mean (SM). Generally, the best imputation method selected to fill in all percentages of simulated missing data was Linear Interpolation (LI) method. The second best method for of simulated missing data were Mean Top Bottom (MTB) method, followed by Nearest Neighbour (NN) method and lastly Series Mean (SM) method. This was evidenced when almost all performance indicators in each of the percentages agreed that LI method as the most appropriate for replacing the missing values. It seem that the performance of LI method was good although the variety of the missing gaps in the rainfall data is different. LI method was shown to have great performance and low error to impute the 5%, 10%, 15% and 20% simulated missing data.

Acknowledgement

Author would like to thank Department of Irrigation and Drainage Malaysia for the rainfall dataset.

References

[1] Shako O 2015 Climate Measurement: *A review of rainfall and temperature measurement standards in Guyana* (Guyana: Ministry of Agriculture)
[2] Ben Aissia M A Chebana F and Ouarda T B M J 2017 *Adv. Water Resour.* **110** 299–309
[3] Junger W and Leon A P 2015 Atmospheric Environment **102** 96-104
[4] Sukatis F F Noor N M Zakaria N A Ul-Saufie A Z Suwardi A 2019 *International Journal of Conservation Science* **10**(4) 791–804
[5] Noor N M Yahaya A S Ramli N A Abdullah M M A 2006 Journal of Engineering Research & Education **3** 96-105
[6] Noor N M Abdullah M M A Yahaya A S and Ramli N A 2015 *Mater. Sci. Forum* **803** 278–281
[7] Norazian M N Mohd Mustafa A Ahmad Shukri Y and Nor Azam R 2006 Simulation **75** 94
[8] Kalteh A M and Hjorth P. 2009 *Hydrol. Res.* **40(**4) 420–432