The Andhra Agric. J 70 (2): 254-262, 2023
10.61657/aaj.2023.38

An Investigation to Identify the Best Precipitation Missing Data Methods

D Tejasri, K N Sreenivasulu, V Srinivasa Rao and B Aparna
Department of Statistics and Computer Applications, Agricultural college, Bapatla, A.P.

ABSTRACT

Rainfall is very much important for the agricultural activities and also important part of the hydrological cycle. Studying about precipitation is important in identifying precipitation characteristics; occurrence is a spatial and temporal variability, statistical modeling and forecasting of precipitation and resolving the problems such as floods, droughts, landslides etc., Numerous methods have been introduced for estimating and reconstructing missing data. Several methods are available to predict the missing data. Those methods are Arithmetic average method, Normal ratio method, Inverse distance method, Aerial precipitation ratio method, Multiple linear regression method and UK Traditional method. These methods derive the missing values using observations from neighboring stations and to estimate the missing data of precipitation. Some of these methods are taken up in this study to identify the best method based on the method selection criteria *i.e.* R^2, RMSE and MAPE. The results needed that the multiple linear regression method, Normal ratio method, Decision tree method, UK Traditional method provide successful estimation of the missing precipitation data. It is found that multiple linear regression method performs well over other standard methods when missing rainfall is estimated for both dry and rainy months.

Keywords**:** *MAPE, Missing data, R^2, Rainfall and RMSE.*

Rainfall is needed as a source of fresh water, which is essential for the survival of mankind, plants and animals as well as stream flow availability. Rainfall and stream flow plays a significant role in hydrological, agricultural methods and in assessing water quality. Studying about rainfall is important in order to identify the rainfall characteristics, occurrence of spatial and temporal variability, forecasting extreme rainfall events and hence, the problems such as floods, droughts and landslides may be resolve. Meanwhile, stream flow cycle is the section where rainfall occurs and results in flow. Floods also can happen when the volume of water exceeds the capacity of the river.

The consistency and continuity of rainfall data are very important in statistical analyses such as time series analysis. Both consistency and continuity may be disturbed due to change in observational procedure and incomplete records (missing observations) which may vary in length from one or two days to decades of years. However, inconsistency in a rainfall record can be identified by graphical or statistical methods such as double mass curve analysis, the Von Neumann ratio test, cumulative deviations, likelihood ratio test, run test etc., Nevertheless, filling of the gaps generated by inconsistent data is essential, and different procedures and approaches are available to accomplish this task. The most common method used to estimate missing rainfall data is Normal Ratio method (Chow *et al.,* 1988). This method is based only on past observations of that rain gauge and surrounding gauges. However, there are other important factors such as distances among rain gauges, aerial coverage of each gauge etc. which are disregarded in this method but are proved to have significant influences on estimation of rainfall. There are other techniques which use different other factors to estimate missing rainfall data, few of them including Normal Ratio method, Inverse Distance method, and Arithmetic Mean method/local Mean method (Chow *et al.,* 1988). Regression and time series methods were used in the past for estimation of missing rainfall data. One of the major limitations of such methods is the necessity to define the functional form of the relationships a priori.

Most of these methods derive the missing value using observations from neighboring stations. To estimate the missing data of precipitation several methods is taken up in this study to identify the best

method based on the method selection criteria i.e. R square, MAPE, RMSE etc., Then, based on the identified best method the estimates can be done for the missing data.

MATERIALS AND METHODS

To identify the best method to estimate the precipitation missing data at five rain-gauge stations located in Andhra Pradesh, namely Nandigama, Vijayawada, Tiruvuru, Guntur and Jaggaiahpet were taken in this study by using Nandigama has a target station and remaining four are nearby stations for development of various methods of missing data techniques on rainfall. The study completely relied on secondary data collected from IMD (India Meteorological Department), Government of India. The monthly data was collected for the period of 1961-2000.

Nandigama is a town in NTR district of Andhra Pradesh. It is a Municipality and also the headquarters of Nandigama mandal in Nandigama revenue division. Nandigama is located 50 km West of Vijayawada and 46 km East of Kodada. It has an area of 25.90 km^2. Nandigama is located between 16.7833^o North Latitude and 80.3000^o East Longitude. Nandigama typically receives about 34.3 millimeters (1.35 inches) of precipitation and has 36.55 rainy days (10.01% of the time) annually.

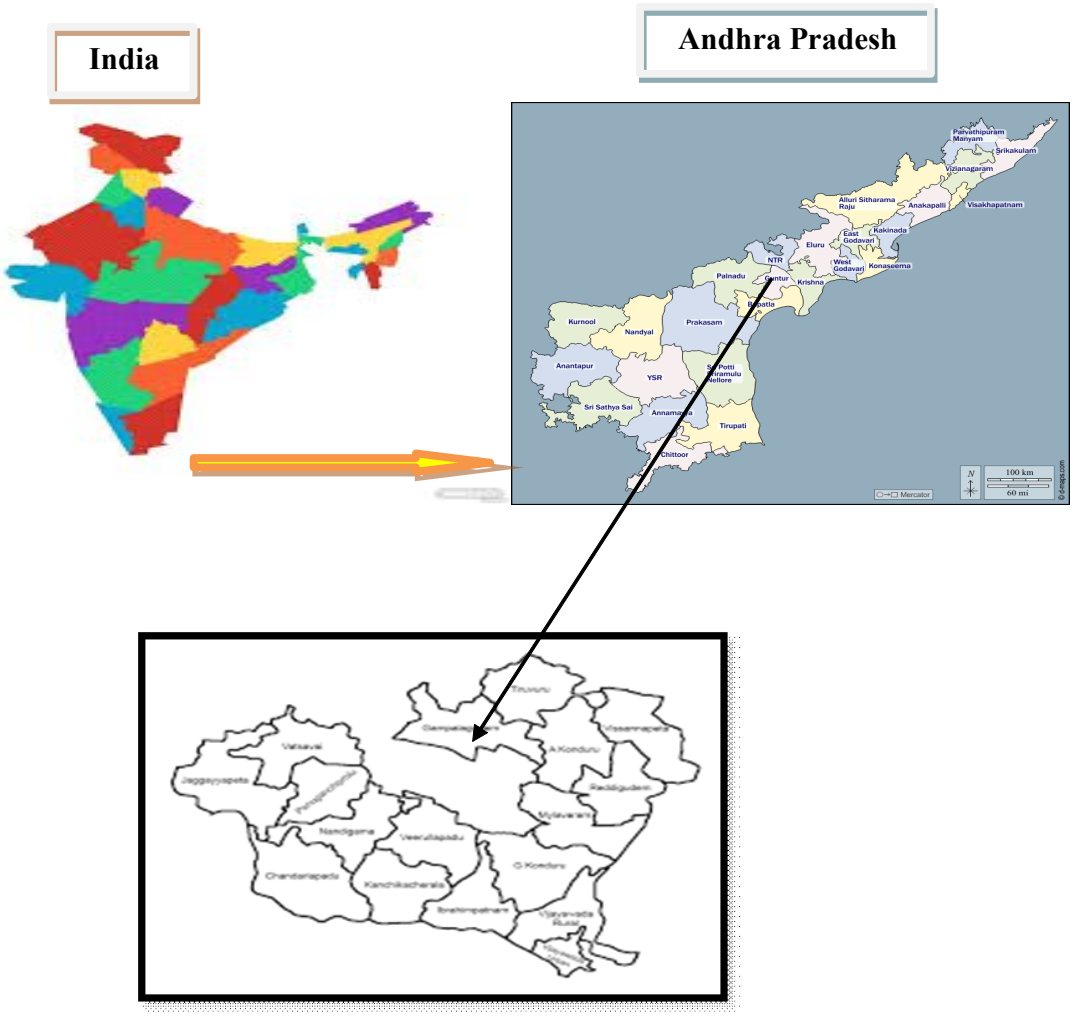

Fig.1 Location of Nandigama in NTR District (*source from Google website*)

Arithmetic Average Method (AAM)

Arithmetic Average method is the simplest method commonly used to fill in missing meteorological data in meteorology and climatology. Missing data can be obtained by computing the arithmetic average of the data corresponding to the nearest weather stations, as shown below

$$V_0 = \frac{\sum_{i=1}^{N} V_1}{N}$$

Where V_0 is the estimated value of the missing data, V_i is the value of same parameter at 1th nearest weather station and N is the number of the nearest stations.

The AA method is satisfactory if the gauges are uniformly distributed over the area and the individual gauge measurements do not vary greatly about the mean (Chow *et al.* 1998)

Inverse Distance Weighting Method (IDWM)

The inverse distance (reciprocal-distance) weighting method (IDWM) is the method most commonly used for estimating missing data. This method estimate the missing rainfall observation by considering the observed values at other stations, by using the formula

$$V_0 = \frac{\sum_{i=1}^{n} \left(\frac{v_i}{D_i} \right)}{\sum_{i=1}^{n} \frac{1}{D_i}}$$

Where D_i is the distance between the station with missing data and the nearest weather station, V_0 is the estimated value of the missing data and V_i is the value of same parameter at nearest weather station

Normal Ratio Method (NRM)

Normal Ratio (NR) method is weighted based on the ratio mean of the available data between the target station and the ith neighboring station. This method is used if any neighboring stations have the normal annual rainfall and stream flow data which exceeded more than 10% of the considered station. The estimated missing value is given by

$$P_X = \frac{1}{m} \sum_{i=1}^{m} \left[\frac{N_x}{N_i} \right] P_i$$

Where P_x = Estimate for the ungauged station, P_i = Rainfall values of rain gauges used for estimation, N_x = Normal annual precipitation of X station, N_i = Normal annual precipitation of surrounding stations and m = No. of surrounding stations

Aerial Precipitation Ratio Method (APR)

This method was developed based on spatial distribution of daily rainfall without accounting for the historical recurrence. The method leads the extension of point rainfall records to Thiessen Polygon areas. The APR method assumes the contribution of rainfall from surrounding stations is proportionate to the aerial contribution of each sub catchment (Thiessen polygon area claimed by each station without considering the missing gauge), when the station of missing values is excluded (De Silva, 1997). The formula of the method can be given as follows.

$$P_X = \frac{\sum_{i=1}^{N} \left[\left(A_j - A_i \right) P_i \right]}{\sum_{i=1}^{N} \left(A_j - A_i \right)}$$

Where A_j = Thiessen polygon area for the station with missing values, A_i = Thiessen polygon area when station with missing values is included, p_i = Annual precipitation of surroundings stations and p_X = Estimate for monthly rainfall for the station with missing observations.

Multiple Linear Regression Analysis (MLRM)

The MLRM using the least absolute deviation criteria (MLAD) is a robust version of a general linear least squares estimation. The method of least squares is an effective method when the errors are normally distributed and independent. However, for precipitation data especially, the assumption of normality over the wide range of situations can lead to poor estimations (Eischeid *et al.,* 1995). The main advantage of least absolute deviations is its resistance to outliers and to overemphasis of large tailed distributions. MLAD estimates the unknown parameters in a stochastic method so as to minimize the sum of absolute deviations of neighboring stations observations from the values predicted by the method. Kemp *et al.* (1983), Young (1992) and Eischeid *et al.* (1995) highlighted many advantages of the REG

in the data interpolation and estimation of missing data. Missing data (V0) were estimated as

$$V_0 = a_0 + \sum_{i=1}^{n} (a_i v_i)$$

Where a_i, a_1 and a_n are the regression coefficients and V_i is the value of same parameter at i^{th} nearest weather station

UK Traditional Method

The estimation of missing data involves assuming a constant difference between the long-term data from the target station and neighboring stations. For each month of the year, long-term data for each neighboring station is compared with data of the target station. The equation of estimation is as follows.

$$K_i = p_{i,j} + (\overline{q_j} > \overline{p_{i,j}}) \qquad if\ \overline{q_j} > \overline{p_{i,j}}$$

$$or$$

$$K_i = p_{i,j} - (\overline{p_{i,j}} > \overline{q_j}), \qquad if\ \overline{q_j} < \overline{p_{i,j}}$$

Where K_i is the UK coefficient value of the i^{th} neighboring station, $p_{i,j}$ is the observed temperature value of the i^{th} neighboring station of j^{th} month, $\overline{p_{i,j}}$ is the long-term average of the observed temperature of the i^{th} neighboring station of j^{th} month and $\overline{q_j}$ is the long-term average of observed temperature of the target station of the j^{th} month

The method can either be approached using correlation or distances between the target and neighboring stations. The following are the number of ways of estimating missing value under the UK method:

Averaging the Best Correlated Stations (UK_AA_C)

Calculating an arithmetic averaging of the individual estimations of the best correlated stations. The equation of estimation is as follows:

$$P_X = \frac{1}{n} \sum_{i=1}^{i=n} K_i$$

Where P_x is the estimated value, K_i is the UK coefficient value of the i^{th} neighboring station and n is the number of neighboring stations.

To identify best method for predicting the rainfall

The root mean square errors (RMSE), the mean absolute percent errors (MAPE) and the coefficient of determination (R^2) values were used.

Root mean square error (RMSE): It is the standard deviation of the residuals (prediction errors). The RMSE tells you how concentrated the data is around the line of best fit. It has also been used for method selection. The lesser value of RMSE the better method fit is mathematically

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_t - y_t)^2}{n}}$$

Mean absolute percent error (MAPE): It is a measure of prediction accuracy of a forecasting method in statistics. It expresses the mean of the percentage errors over different points of time, and is mathematically expressed as

$$MAPE = \frac{100}{n} \sum_{t=1}^{n} \left| \frac{y_t - \hat{y}_t}{y_t} \right|$$

R-Square: R^2 is a statistic that gives some information on overall variation explained by the method out of total variation. It is also known as coefficient of multiple determinations and it has been used by many researches for method selection having the highest value.

Mathematically,

$$R^2 = 1 - \frac{\text{Error sum of square}}{\text{Total sum of square}} = 1 - \frac{\sum_{t=1}^{n} (y_t - \hat{y}_t)^2}{\sum_{t=1}^{n} (y_t - \bar{y})^2}$$

Where,

y_t is the observed rainfall and stream flow at nearby station,

$\hat{y}_t$ is the estimated value and n is the number of nearby station.

RESULTS AND DISCUSSION

It deals with development of various statistical methods to estimate the missing rainfall data in Nandigama raingauge station. In this paper an attempt was made to

estimate the monthly missing rainfall data in Nandigama.

Estimation of monthly missing rainfall data in Nandigama

Table 1. Performance of standard methods during the month of January

S.NO	METHOD	R^2	RMSE	MAPE
1	Arithmetic average method	0.76	2.817	0.609
2	Normal ratio method	0.87	2.301	0.462
3	Inverse distance method	0.77	2.825	0.584
4	Multiple linear regression method	0.62	2.508	1.088
5	Aerial precipitation ratio method	0.76	2.856	0.609
6	UK Traditional method	0.8	2.728	0.956

From the Table.(1) it can be seen that the Normal ratio method performed better than other methods having low RMSE (2.301), MAPE (0.462) and with highest R^2 (0.866) value for the month of January.

Table 2. Performance of Standard methods during the month of February

S.NO	METHOD	R^2	RMSE	MAPE
1	Arithmetic average method	0.699	2.928	0.579
2	Normal ratio method	0.701	2.416	1.674
3	Inverse distance method	0.725	2.52	0.524
4	Multiple linear regression method	0.937	0.706	0.703
5	Aerial precipitation ratio method	0.698	2.921	0.566
6	UK Traditional method	0.647	2.802	0.589

From Table.(2) it can be observed that the performance of MLR is better with highest R^2 (0.937) and lowest RMSE (0.706) and MAPE (0.703) values as compared to the other five methods during the month of February.

Table 3. Performance of Standard methods during the month of March

S.NO	METHOD	R^2	RMSE	MAPE
1	Arithmetic average method	0.6	1.543	1.059
2	Normal ratio method	0.77	1.48	1.654
3	Inverse distance method	0.62	1.674	1.115
4	Multiple linear regression method	0.89	1.237	0.752
5	Aerial precipitation ratio method	1	1.059	0.335
6	UK Traditional method	0.6	1.543	1.054

The above Table.(3) depicts that the RMSE (1.059), MAPE (0.335) are low and R^2 (0.999) is high for the Aerial precipitation method. It shows the superiority over the other Standard methods during the month of March.

Table 4. Performance of Standard methods during the month of April

S.NO	METHOD	R^2	RMSE	MAPE
1	Arithmetic average method	0.67	2.14	3.176
2	Normal ratio method	0.87	1.914	1.997
3	Inverse distance method	0.69	2.125	2.683
4	Multiple linear regression method	0.79	1.927	2.935
5	Aerial precipitation ratio method	0.7	2.175	3.603
6	UK Traditional method	0.67	2.107	3.201

From Table.(4) it is evident that the value of RMSE (1.914) and MAPE (1.997) are lowest and R^2 (0.871) is highest for Normal ratio method as compared with other methods. Therefore, it can be stated that the Normal ratio method performs better during April month.

Table 5. Performance of Standard methods during the month of May

S.NO	METHOD	R^2	RMSE	MAPE
1	Arithmetic average method	0.68	5.483	3.955
2	Normal ratio method	0.79	4.727	3.562
3	Inverse distance method	0.68	5.642	3.686
4	Multiple linear regression method	0.94	3.935	3.411
5	Aerial precipitation ratio method	0.65	5.455	3.825
6	UK Traditional method	0.68	5.391	3.04

The Table.(5) shows the values of RMSE (3.935) and MAPE (3.411) are lower and R^2 (0.935) is higher in Multiple linear regression method as compared with other methods. Hence Multiple regression method is superior over other methods for estimation of missing rainfall of Nandigama for the month of May.

Table 6. Performance of Standard methods during the month of June

S.NO	METHOD	R^2	RMSE	MAPE
1	Arithmetic average method	0.72	6.569	0.829
2	Normal ratio method	0.7	8.014	1.596
3	Inverse distance method	0.65	7.338	0.847
4	Multiple linear regression method	0.67	6.788	0.951
5	Aerial precipitation ratio method	0.67	7.429	0.87
6	UK Traditional method	0.62	7.338	1.038

From Table (6), Values of RMSE (6.569) and MAPE (0.829) are least and R^2 (0.721) is highest for Arithmetic average method over the other methods. Hence, it can be identified that Arithmetic average method performs better over the other Standard methods in the month of June.

Table 7. Performance of Standard methods during the month of July

S.NO	METHOD	R^2	RMSE	MAPE
1	Arithmetic average method	0.69	11.45	0.516
2	Normal ratio method	0.9	11.96	0.521
3	Inverse distance method	0.71	11.49	0.482
4	Multiple linear regression method	1	11.3	0.475
5	Aerial precipitation ratio method	0.69	11.35	0.517
6	UK Traditional method	0.66	11.45	0.519

Table.(7) shows that Multiple linear regression method has higher value of R^2 (0.997) and lower values of RMSE (11.30) and MAPE (0.475) than other methods. So it can be observed that MLRM is superior to other methods during the month of July.

Table 8. Performance of Standard methods during the month of August

S.NO	METHOD	R^2	RMSE	MAPE
1	Arithmetic average method	0.71	10.04	0.744
2	Normal ratio method	0.82	9.972	0.762
3	Inverse distance method	0.84	9.785	0.704
4	Multiple linear regression method	0.74	9.86	0.731
5	Aerial precipitation ratio method	0.75	10.161	0.749
6	UK Traditional method	0.71	10.034	0.764

As presented in Table.8, RMSE (9.785) and MAPE (0.704) are lower and R^2 (0.837) is higher in Inverse distance method over the other methods. It indicates the IDM performs better over the other methods for month of August.

Table 9. Performance of Standard methods during the month of September

S.NO	METHOD	R^2	RMSE	MAPE
1	Arithmetic average method	0.65	10.009	0.712
2	Normal ratio method	0.76	10.395	0.78
3	Inverse distance method	0.66	10.212	0.725
4	Multiple linear regression method	0.78	8.521	0.709
5	Aerial precipitation ratio method	0.64	10.04	0.712
6	UK Traditional method	0.69	8.861	0.868

Table.(9) depicts that RMSE (8.521) and MAPE (0.709) are lower and R^2 (0.778) is higher in Multiple linear regression method than other methods. Hence, Multiple linear regression method performs better than all other methods during the September month.

Table 10. Performance of Standard methods during the month of October

S.NO	METHOD	R^2	RMSE	MAPE
1	Arithmetic average method	0.66	9.984	1.188
2	Normal ratio method	0.78	9.48	1.025
3	Inverse distance method	0.68	10.19	1.084
4	Multiple linear regression method	0.65	9.826	1.532
5	Aerial precipitation ratio method	0.67	10.064	1.127
6	UK Traditional method	0.65	9.846	1.54

Table.(10) shows that RMSE (9.480) and MAPE (1.025) are lower and R^2 (0.782) is higher in case of Normal Ratio method as compared to other methods. It shows the superiority of Normal Ratio method over the other methods for the month of October.

Table 11. Performance of Standard methods for during the month of November

S.NO	METHOD	R^2	RMSE	MAPE
1	Arithmetic average method	0.91	5.451	9.012
2	Normal ratio method	0.85	5.224	8.562
3	Inverse distance method	0.99	4.165	7.346
4	Multiple linear regression method	0.71	4.194	10.229
5	Aerial precipitation ratio method	0.97	5.086	8.205
6	UK Traditional method	0.96	5.413	7.47

Table.(11) shows that RMSE (4.165) and MAPE (7.346) are lower and R^2 (0.986) is higher in case of Inverse distance method. It can be observed that Inverse distance method performs better over the other methods for the month of November.

Table 12. Performance of Standard methods during the month of December

S.NO	METHOD	R^2	RMSE	MAPE
1	Arithmetic average method	0.94	0.805	0.836
2	Normal ratio method	0.9	1.066	0.987
3	Inverse distance method	0.9	0.697	0.697
4	Multiple linear regression method	0.93	0.617	0.759
5	Aerial precipitation ratio method	0.86	0.74	0.751
6	UK Traditional method	0.96	0.612	0.633

Table.(12) depicts that RMSE (0.612) and MAPE (0.633) are lower and R^2 (0.955) is higher for UK Traditional method so it can be identified that UK Traditional method performs better over other methods for the December month.

The investigation on Assessment of precipitation missing data methods and to estimate the missing rainfall data in Nandigama.

1) For the monthly rainfall it is found that Normal ratio method performed better for January, April, and October, Multiple linear regression method for the February, July and September, Inverse distance method performed better for August and November, UK Traditional method for the December and the Aerial precipitation method better for the March over other standard methods.
2) It is found that multiple linear regression method performs better over standard methods when missing rainfall is estimated for both dry and rainy months.
3) The results of the study showed that the target station has multiple linear regression method and Normal ratio method provided successful estimation of the missing precipitation data.

LITERATURE CITED

De Silva R P, Dayawansa N D K and Ratnasiri M D 2007. A comparison of methods used in estimating missing rainfall data.

Eischeid J K, Bruce Baker C, Karl T R and Diaz H F 1995. The quality control of long-term climatological data using objective data analysis. *Journal of applied meteorology and climatology*. 34(12): 2787-2795.

Kemp W P, Burnell D G, Everson D O and Thomson A J 1983. Estimating missing daily maximum and minimum temperatures. *Journal of Applied Meteorology and Climatology*. 22(9): 1587-1593.

Chow V, Maidment D R and Mays L W 1998. Applied Hydrology. McGraw-Hill, New York.

Young KC 1992. A three-way method for interpolating for monthly precipitation values. *Monthly Weather Review*. 120(11): 2561-2569.

Received on 17.03.2023 and Accepted on 24.06.2023