Replacing missing data

Pulak Guhathakurta

Climate Data Management & Services

भारत मौसम विज्ञान विभाग

INDIA METEOROLOGICAL DEPARTMENT

ESTIMATING MISSING DATA

- One of the main applications of statistics to climatology is the estimation of values of elements when few or no observed data are available or when expected data are missing.

- In many cases, the planning and execution of user projects for the need of enough meteorological or climatological observations; estimation is used to extend a dataset.

- Estimation also has a role in quality control by allowing an observed value to be compared to its neighbours in both time and space.

- Techniques for estimating data are essentially applications of statistics, but should also rely on the physical properties of the system being considered.

- In all cases, it is essential that values statistically estimated be realistic and consistent with physical considerations.

Estimation of missing values in monthly/seasonal/annual data

Difference and ratio adjustments: (WMO Guide to Climatological Practices)

When concurrent values for station pairs are compared, it is found that in some cases their difference (e.g. temperature) or their ratio (e.g. precipitation, wind speed, duration of bright sunshine) tends to be constant. However this may not be always be apparent for single observations, daily sums or mean values but is usually rather striking for monthly and annual sums and means.

Difference d or ratio q between values of a given element observed at stations A and B can be established from corresponding sums or mean values (from the simultaneous observations)

$$\mathbf{d} = \frac{\sum(\boldsymbol{b_i} - \mathbf{a}_{\boldsymbol{i}})}{\boldsymbol{n}} \quad \text{and} \quad q = \frac{\sum b_i}{\sum a_i}$$

If the value b_j is missing at B, it is reconstructed by using the corresponding value a_j at A and adding the established difference, or multiplying by the established ratio, depending on the element in question :

$$b_j = a_j + d$$

or,

$$b_j = a_j . q$$

INTERPOLATION /EXTRAPOLATION

Time interpolation

Interpolation uses data that are available both before and after a missing value,

Space interpolation

Interpolation using the data surrounding the missing value, to estimate the missing value.

Extrapolation extends the range of available data values. There are more possibilities for error of extrapolated values because relations are used outside the domain of the values from which the relationships were derived. Even if empirical relations found for a given place or period of time seem reasonable, care must be taken when applying them to another place or time because the underlying physics at one place and time may not be the same as at another place and time.

Estimation of missing point rainfall with spatial values :

- **Arithmetic Average Method:**

$$P_x = \frac{1}{n} \sum_{i=1}^{n} P_i$$

- **Normal Ratio Method:**

$$\frac{P_x}{N_x} = \frac{1}{n} \sum_{i=1}^{n} \frac{P_i}{N_i}$$

or

$$P_x = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{N_x}{N_i} \right) P_i$$

where N_i = Average annual total rainfall at station i.

Different methods for estimating daily missing values

1.	Average value nearer stations by distance	AVNSd
2.	Average value nearer stations by correlation coefficient	R AVNSR
3.	Climatological mean of the day	CMD
4.	Nearest neighbour value by distance	NNVd
5.	Nearest neighbour value by correlation coefficient	R NNVR
6.	Inverse distance weight power 1	IDW+1
7.	Inverse distance weight power 2	IDW+2
8.	Inverse distance weight power ½	IDW+1/2
9.	Linear regression with the nearest station by distance	LRNd
10.	Linear regression with the nearest station by R	R LRNR
11.	Linear regression with the nearest station by R/d ratio	LRNR/d
12.	Multiple linear regression weighted by distance power -1	MLRWd-1
13.	Multiple linear regression weighted by distance power -2	MLRWd-2
14.	Multiple linear regression weighted by R power 1	MLRWR+1
15.	Multiple linear regression weighted by R power 2	MLRWR+2
16.	Multiple linear regression weighted by R/d ratio power 1	MLRWR/d+1
17.	Multiple linear regression weighted by R/d ratio power 2	MLRWR/d+2

Average value nearer stations by distance d (AVNSd)

Missing data are obtained by arithmetically averaging data of the closest weather stations around the station of interest applying the following equation:

$$V_{est} = \frac{\sum_{i=1}^{n} V_i}{n}$$

Where V_{est} is the estimated value of the missing data, V_i is the value of the same parameter at the ith nearest weather station and n is the number of the nearest stations, from which information was used for the estimation of the missing value.

Average value nearer stations by the Pearson coefficient R (AVNSR)

The Pearson correlation coefficient, R, is a measure of the linear relationship between two random variables. R is an index that can be used to measure the degree of relationship of two variables. The R values are converted into weights by using the weighting formula for the ith R; where n is the total number of stations. Estimation of the missing variable is computed as the weighted sum of the available variable-values (n) and their respective weights (w)

$$w_i = \frac{R_i}{\sum_{j=1}^{n} R_j}$$

$$V_{est} = \sum_{j=1}^{n} V_j * w_i$$

Climatological mean of the day (CMD).

This method uses the long-term average value of the same day of interest. V_{est} is the estimated value, V_i is the value of the variable for the ith day of year j, and T is the number of years data are available. The output of the following equation is simply a temporal average of the jth day value

$$V_{est,i} = \frac{\sum_{j=1}^{T} V_{ij}}{T}$$

Nearest neighbour value by d and R (NNV)

Observations from neighbouring stations are used in missing data reconstruction. It has been proposed to use geometrical distances to the stations and to apply the data from the closest station. However, the method gives poor results when the climate variable under analysis has a high spatial variability. For this reason, in this study a modified version of the criterion was used imposing conditions on the correlation coefficient

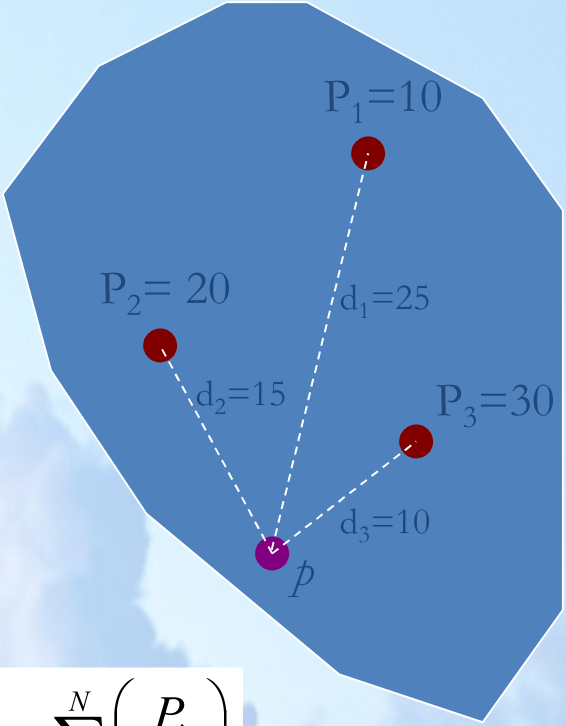
Inverse distance weight (IDW)

Computation of point rainfall at an ungauged/missing point : Inverse distance weighting

- Rainfall at a point is more influenced by nearby measurements than that by distant measurements
- The Rainfall at an ungauged point is inversely proportional to the distance to the measurement points
- Steps
  - Compute distance (d_i) from ungaged point to all measurement points.

$$d_{12} = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

  - Compute the precipitation at the ungaged point using the following formula



$$\hat{P} = \frac{\sum_{i=1}^{N}\left(\frac{P_i}{d_i^{\,2}}\right)}{\sum_{i=1}^{N}\left[\frac{1}{d_i^{\,2}}\right]}$$

$$\hat{P} = \frac{\frac{10}{25^2} + \frac{20}{15^2} + \frac{30}{10^2}}{\frac{1}{25^2} + \frac{1}{15^2} + \frac{1}{10^2}} = 25.24 \ mm$$

Linear regression with the nearest station by d, R and by the ratio R/d (LRN)

In this method the nearer station is selected by the nearest distance, the higher Pearson coefficient, or the higher ratio of the Pearson coefficient and the distance. The value in the nearer station V_i is defined . Then a linear fit between the target station V_{est} and the selected station is calculated to obtain the parameters a and b.

$$V_{est} = b * V_i + a$$

Multiple linear regression weighted by d, R and by the ratio R/d (MLRW)

In multiple linear regression, a linear combination of two or more predictor variables is used to explain the variation in a response. In essence, the additional predictors are used to explain the variation in the response not explained by a simple linear regression. Several powers are used to identify the best performing one. The weight used is the distance, the Pearson coefficient or the ratio of the Pearson coefficient to the distance, with power k. The coefficients a_i and b_i are calculated as a linear fit between V_{est} and $V_{i.}$

$$\boldsymbol{V_{est}} = \frac{\sum_1^n \{\boldsymbol{w_i^k} * (\boldsymbol{b_i} * \boldsymbol{V_i} + \boldsymbol{a_i})\}}{\sum_1^n \boldsymbol{w_i^k}}$$

For the methods using data of several stations generally four or more stations are to be used for the estimating missing value of precipitation, assuming that one station is representative of one cardinal direction. For temperature at least two stations (n = 2) are to be considered because of the scarcity of stations at diverse elevations, and taking into account that too distant stations could affect the prediction of closer stations.

THANK YOU

ANY QUESTION?