Similarity indices of meteo-climatic gauging stations for missing data handling: definition and comparison with the MICE method

E. Barca[1,*], G. Passarella[1]

[1] *Water Research Institute of the National Research Council, Department of Bari, Viale F. De Blasio, 5 70123 Bari, Italy; emanuele.barca@ba.irsa.cnr.it; giuseppe.passarella@ba.irsa.cnr.it*

[*] *Corresponding author*

Abstract. *The meteo-climatic datasets are at the basis of a great deal of studies on environmental state and its consequent management. In this frame, the completeness of meteo-climatic datasets is required for accurate and reliable analysis. Unfortunately, completeness is a rare in practice and, consequently, a preliminary treatment for filling in all gaps is needed. In this work, two intuitive and easy procedures for handling missing data are presented based on the "similarity station" concept. Finally, a comparison between the proposed methods and the Multiple Imputation Chained Equations, which is the state of the art in the field of missing data handling, has been carried out.*

Keywords. *Missing data; Time series; Multiple Imputation Chained Equations; Similarity methods.*

1 Introduction

Cimatic series are rarely complete, usually because of malfunctioning, effects of extreme events on the probes, etc. Consequently, a preliminary formal treatment of the time series is needed in order to fill all the gaps in. Such a treatment is very critical mostly because it is (i) inherently time consuming, particularly for long time series and large amount of missing data; (ii) affected by a high level of uncertainty, particularly for variables irregularly distributed in space and time; (iii) strongly dependent on the missing data mechanism ([2]); (iv) a blind estimation and only a global reliability can be assessed by means of population statistics. At present, a number of robust and powerful methods exist for missing data handling such as the Multiple Imputation Chained Equations (MICE) ([3]) and the Expectation-Maximization (EM) ([1]), which have been designed so that the estimation takes into account the available numerical and distributional information. Such methods revealed their efficacy also in cases where the missing data percentage is particularly severe, overcoming the critical threshold of 15/20%; nevertheless, some authors still claim the need of further investigations to definitively state their reliability ([4]). Furthermore, these methods are practically difficult to be implemented and not very intuitive. In a previous work ([2]) a methodological proposal was presented for a quick and reliable estimation of climatic missing data based on the concept of twin gauging stations. The proposed method is based on the intuitive concept of persistence, in time, of the spatial continuity of the climatic processes. On this basis, a refined and improved methodology is presented for determining similar gauging stations through which estimating missing values. Statistical and topographic properties are combined in order to determine a "similarity matrix". Given a gauging station whose time-series is affected by missing values, these are assessed "combining" the corresponding values of the n most similar stations. The proposed method and MICE were both applied to the rainfall gauging network of the Apulia Region (South-Eastern Italy). Statistical tests on both the estimated time series confirmed a substantial identity between the results of both the methods.

2 Materials and Methods

Usually, matrices of meteo-climatic time-series, measured at different locations of a regional monitoring network, are affected by different rates of missing data. To address this issue, various missing data handling methods have been propose in literature, usually based on the time autocorrelation. Nevertheless, meteo-climatic events, at the considered scale, are notoriously characterized by a strong spatial structure. In practice, we expect similar events in time in "similar stations in space". In this frame, a methodology is proposed in order to assess a degree of similarity of the monitoring network stations. Once a ranked list of "similar stations" has been determined for any considered station, missing values in it can be computed as the average of univariate regression estimations. In this paper two similarity metrics are proposed (equations (1) and (2)): the first approach requires the correlation matrix to be computed. In this case a simple index of similarity $I_R(i,j)$is provided consisting in the determination coefficient (equation (1)). A refinement of $I_R(i,j)$ is also provided ($I_S(i,j)$ in equation (2)) which involves the effective pairs number, $\frac{n_{i,j}}{N}$ and the relative distance and elevation differences, $\hat{d}_{i,j}$and $\widehat{\Delta h}_{i,j}$.

$$I_R(i,j) = R_{i,j}^2 \tag{1}$$

$$I_S(i,j) = \frac{1}{3}\left(\frac{n_{i,j}}{N} \cdot R_{i,j}^2 + \hat{d}_{i,j} + \widehat{\Delta h}_{i,j}\right) \tag{2}$$

$$\hat{d}_{i,j} = \frac{d_{i,j} - d_{k,j}^{max}}{d_{k,j}^{min} - d_{k,j}^{max}} \qquad \begin{matrix} given\ k \\ j=1,\ 2,\ \ldots,\ n \end{matrix} \tag{3}$$

$$\widehat{\Delta h}_{i,j} = \frac{\Delta h_{i,j} - \Delta h_{k,j}^{max}}{\Delta h_{k,j}^{min} - \Delta h_{k,j}^{max}} \qquad \begin{matrix} given\ k \\ j=1,\ 2,\ \ldots,\ n \end{matrix} \tag{4}$$

where i and j represent a pair of monitoring stations; $n_{i,j}$ and N represent the number of pairs shared by i and j and the total length of the time-series, respectively; $\hat{d}_{i,j}$and $\widehat{\Delta h}_{i,j}$ represent the distance and the difference of elevation between i and j standardized with respect to the related variable ranges (equations (3) and (4)). The time–series of the m most similar stations are used for providing m estimations of missing data in i, according to $I_R(i,j)$ and $I_S(i,j)$. Such estimations are finally combined in a single value through the arithmetic average or a weighted average using $I_R(i,j)$ and $I_S(i,j)$ as weights, respectively. An application of the proposed methodologies is presented related to the "Canosa di Puglia" rainfall gauging station time series, which was affected by a severe, fictitious amount of missing data (33%). Estimated values of missing data have been statistically compared with true data and those estimated by the MICE method.

2.1 Study area, monitoring network and rainfall time series

The proposed method has been applied to the monthly total rainfall time series originating from 81 stations irregularly positioned within the Apulia Region (South-Eastern Italy) (Figure 1). In general, rainfall over the Region is characterised by a twofold behaviour depending on the season. Concerning the rainfall regime, it is usually assumed as Mediterranean ([2]). The gauging stations all belong to the meteorological monitoring network of the Regional Hydrographic Services of Land Protection Department. The time series range from January 1931 to December, 2010. The elevation of each station ranges from 2.00 m a.s.l., (Manfredonia station) to 954.00 m a.s.l. (Pescopagano station) and the average distance between the monitoring stations is around 120 km with a standard deviation of 26 km.

3 Results and Discussion

The whole time-series length of Canosa station was made of $N = 960$ monthly total rainfall rates related to the period from January 1931 to December 2010. Values ranging from September 1957 to April 1984 were cut out from the series in order to simulate a long period with missing values (320 values). Table 1 reports the sets of $m = 5$ most correlated (MC) and most similar (MS) stations to Canosa resulting after the indices computation.

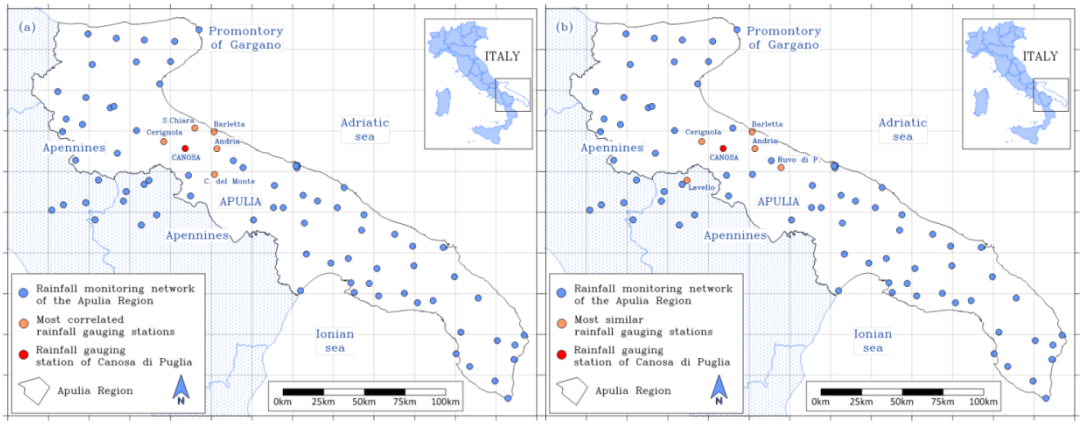

Figure 1: Study area, rainfall monitoring network, Canosa Station and most similar stations according to $I_R(i,j)$in (a) and $I_S(i,j)$ in (b).

Rain gauge Station	**MC/MS**	$R^2_{i,j}$	$d_{i,j}$	$\Delta h_{i,j}$	$n_{i,j}$	$I_S(i,j)$
Masseria Santa	MC	0.789	13.9	145.0	328	-
Cerignola	MC/MS	0.782	13.9	30.0	634	0.852
Andria	MC/MS	0.742	19.5	3.0	634	0.845
Barletta	MC/MS	0.731	20.4	124.0	597	0.793
Castel del Monte	MC	0.718	24.0	371.0	509	-
Lavello	MS	0.706	29.6	159.0	625	0.774
Ruvo di Puglia	MS	0.676	37.4	106.0	621	0.774

Table 1: Summary of parameters involved in the indices computation.

The third and the last columns of Table 1 report the correspondent values of $I_R(i,j)$ and $I_S(i,j)$. Monthly missing values of Canosa have been estimated five times by means of linear univariate regression using each of the most correlated station. Finally, a single value has been computed by averaging the five estimated values. Figure 2 shows the plots of estimated versus true observed values, cut out previously. In particular plot a) refers to the results of the most correlated stations approach, plot b) to that of the most similar approach and finally, plot c) shows the results obtained estimating the missing values by means of the well-known MICE method. Figure 2 is not decisive in order to establish what of the three methods prevails over the others. In fact, the goodness of fit coefficient shows a slightly better value for the "most correlated" approach than the others, while the coefficient and the shape of the regression line seem to indicate better results from MICE. In any case the differences seem to be negligible. Even the error statistics, reported in Table 2, do not indicate the best approach, clearly. The values of the Mean Bias Error (MBE), of the Root Mean Squared Error (RMSE) and of the Relative Mean Absolute Error (RMAE) in Table 2 are very close each other.

4 Conclusions

Four methods have been proposed to estimate missing values in long time series of rainfall measures. The methods substantially propose a univariate regressive estimation of the missing values, in a given rainfall gauging station of a regional monitoring network, based on a set of "similar" stations located in the neighboring. Two indices of similarity have been proposed: the coefficient of determination

($I_R(i,j)$) between the dependent station i and the other stations j and the index $I_S(i,j)$ computed combining some statistical and topographic properties of the pairs (i,j). Both the proposed indices range from 0 to 1.

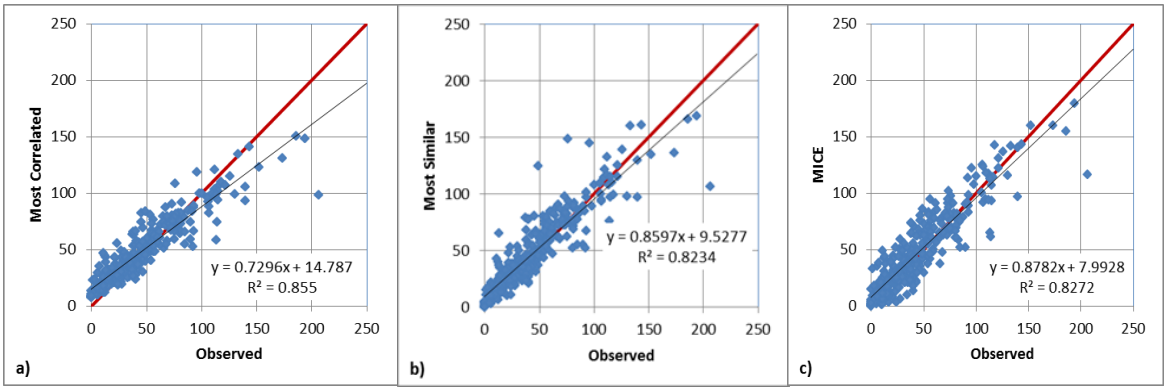

Figure 2: Observed vs. Estimation: a) Most Correlated; b) Most Similar; c) MICE method.

Once a ranked list of correlated/similar stations has been determined, missing values in i are computed as an average of univariate regression estimations. A case study related to the "Canosa di Puglia" gauging station, locate in Apulia Region (South-Eastern Italy), has been presented. In Canosa, monthly total rainfall rates have been measured continuously from 1931 to 2010. A set of 320 values (i.e. 33% of the whole dataset) has been cut out from the series and estimated with the proposed methods. Finally, provided that the MICE method is one of the most reliable for data missing estimation, available in literature, it has been used as benchmark to assess the performances of the proposed methods.

Method	**MBE**	**RMSE**	**RMAE**
Most Correlated	2.54	15.00	0.57
Weigthed Most Correlated	2.54	14.96	0.57
Most Similar	3.21	15.80	0.43
Weigthed Most Similar	3.17	15.72	0.43
MICE	2.47	15.50	0.54

Table 2: Summary statistics of models estimation error.

The results demonstrate a clear equivalence of all the methods, in terms of estimation error statistics and goodness of fit. In conclusion, considering meteo-climatic time-series missing data issue, the proposed methods seem to behave similarly to the most celebrated MICE; however, the procedural straightforwardness of the proposed methods can lead to prefer them instead of other methods well-known for their effectiveness but, undoubtedly, more complex in terms of computational efforts. Monthly total rainfall and monthly mean temperature time-series related to the 81 gauging stations of the regional meteo-climatic monitoring network have been filled in using the Weighted Most Correlated approach and the aforementioned aridity indices have been computed, spatialized and mapped for managerial uses.

References

[1] Dempster, A.P., Laird, N.M., Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*. Series B, **39**(1), 1–38.
[2] Lo Presti, R., Barca, E., Passarella, G. (2010). A methodology for treating missing data applied to daily rainfall data in the Candelaro River Basin (Italy). *Environmental monitoring and assessment*, **160**(1-4), 1-22.
[3] Rubin, D.B. (1987*). Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, New York.
[4] Schafer, J.L., Graham, J.W. (2002). Missing data: our view of the state of the art. *Psychological methods*, **7**(2), 147.