Geophysical Research Abstracts,
Vol. 11, EGU2009-12496, 2009
EGU General Assembly 2009

A regressive methodology for estimating missing data in rainfall daily time series

E. Barca and G. Passarella

Water Research Institute, National Research Council, Bari, Italy (emanuele.barca@ba.irsa.cnr.it / 00390805313365)

The "presence" of gaps in environmental data time series represents a very common, but extremely critical problem, since it can produce biased results (Rubin, 1976).
Missing data plagues almost all surveys. The problem is how to deal with missing data once it has been deemed impossible to recover the actual missing values. Apart from the amount of missing data, another issue which plays an important role in the choice of any recovery approach is the evaluation of "missingness" mechanisms.
When data missing is conditioned by some other variable observed in the data set (Schafer, 1997) the mechanism is called MAR (Missing at Random). Otherwise, when the missingness mechanism depends on the actual value of the missing data, it is called NCAR (Not Missing at Random). This last is the most difficult condition to model.
In the last decade interest arose in the estimation of missing data by using regression (single imputation). More recently multiple imputation has become also available, which returns a distribution of estimated values (Scheffer, 2002).
In this paper an automatic methodology for estimating missing data is presented. In practice, given a gauging station affected by missing data (target station), the methodology checks the randomness of the missing data and classifies the "similarity" between the target station and the other gauging stations spread over the study area. Among different methods useful for defining the similarity degree, whose effectiveness strongly depends on the data distribution, the Spearman correlation coefficient was chosen. Once defined the similarity matrix, a suitable, nonparametric, univariate, and regressive method was applied in order to estimate missing data in the target station: the Theil method (Theil, 1950).
Even though the methodology revealed to be rather reliable an improvement of the missing data estimation can be achieved by a generalization.
A first possible improvement consists in extending the univariate technique to the multivariate approach. Another approach follows the paradigm of the "multiple imputation" (Rubin, 1987; Rubin, 1988), which consists in using a set of "similar stations" instead than the most similar. This way, a sort of estimation range can be determined allowing the introduction of uncertainty.
Finally, time series can be grouped on the basis of monthly rainfall rates defining classes of wetness (i.e.: dry, moderately rainy and rainy), in order to achieve the estimation using homogeneous data subsets. We expect that integrating the methodology with these enhancements will certainly improve its reliability.
The methodology was applied to the daily rainfall time series data registered in the Candelaro River Basin (Apulia – South Italy) from 1970 to 2001.

REFERENCES

D.B., Rubin, 1976. Inference and Missing Data. Biometrika 63 581-592
D.B. Rubin, 1987. Multiple Imputation for Nonresponce in Surveys, New York: John Wiley & Sons, Inc.
D.B. Rubin, 1988. An overview of multiple imputation. In Survey Research Section, pp. 79-84, American Statistical Association, 1988.
J.L., Schafer, 1997. Analysis of Incomplete Multivariate Data, Chapman & Hall.
J., Scheffer, 2002. Dealing with Missing Data. Res. Lett. Inf. Math. Sci. 3, 153-160. Available online at http://www.massey.ac.nz/∼wwiims/research/letters/
H. Theil, 1950. A rank-invariant method of linear and polynomial regression analysis. Indicationes Mathematicae,

12, pp.85-91.