Department of Civil and Environmental Engineering
University of Catania

QUALITY CONTROL OF DAILY PRECIPITATION DATA THROUGH NEURAL NETWORKS

European Geosciences Union
EGU General Assembly 2005
Vienna, Austria, 19 - 24 April 2009

G. Sciuto, B. Bonaccorso, A. Cancelliere and G. Rossi
Tel. +39 095 7382718 Fax +39 095 7382748 Email: guidosciuto@geographie.uni-bonn.de

Introduction and objectives

Hydrometeorological data acquired by a monitoring network are potentially affected by errors, originating from various causes, that can compromise their applicability in describing hydrological phenomena (Madsen, 1989). A traditional control of the data, based on a ‘‘manual’ inspection, is not suitable for real-time applications, where the information on the observed phenomenon has to be promptly available. Therefore, it appears necessary to develop methods for automatic quality control of the data that make it possible to perform a preliminary analysis of the acquired information aimed at identifying potential anomalies in the observations (Abbott, 1986; Reek et al., 1992).

In this paper, an automatic quality control procedure of daily rainfall data is presented (Sciuto et al., 2008), which extends a method previously developed for monthly rainfall data (Campisano et al., 2002). In particular, two types of controls are proposed. The first control is oriented to verify the null/not null nature of observed precipitation, without regard to its amount and therefore aims to detect mainly errors due to the total blockage of the rain gauge.

Once a not null observation is labeled as correct, the second control aims at verifying its value through a comparison with confidence intervals of fixed probability. Both controls are based on neural networks, appropriately trained by making use of historical observations already subject to manual inspection for correctness.

Quality control procedure

Due to the intermittent nature and high variability of daily precipitation, confidence intervals cannot be estimated from historical data of the considered station (Sciuto et al., 2008). The proposed procedure consists of two different neural networks: the first neural network, defined binary, is used to estimate the presence or not of rainfall value in the series of the target station, on the basis of contemporary observations in reference stations. The second neural network is adopted to detect errors in the amount of rainfall, on the basis of confidence intervals, only when the first neural network has confirmed that for the examined day a not-null rainfall value is expected.

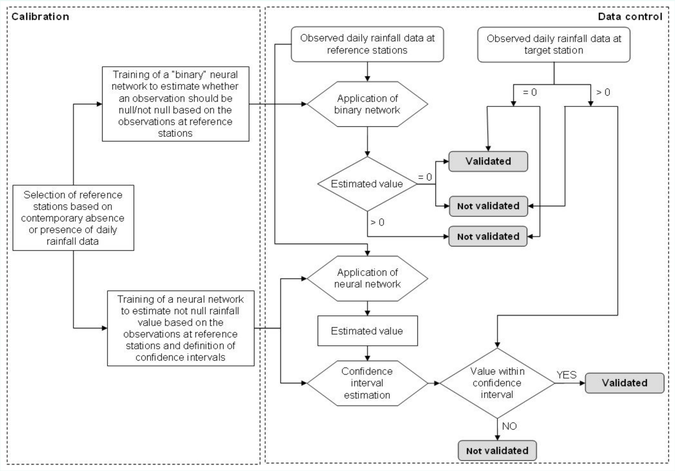

Figure 1. Flow chart of the automatic quality control procedure of daily rainfall data.

Selection of reference stations

Reference stations have been first selected among the neighboring ones within a maximum distance Δl=30 km and altitude difference ΔH=± 200 m. Then, a criteria based on the frequency of contemporary presence/absence of rain has been adopted, by computing the following conditional probability:

$$P_{ij} = P\left[Y^{(c)} = j \mid Y^{(r)} = i\right] \cong \frac{\#\left[Y^{(c)} = j, Y^{(r)} = i\right]}{\#\left[Y^{(r)} = i\right]} \qquad i, j = 0,1$$

being i,j equal to 1 in case of daily rainfall observed at the generic station, and equal to 0 in case of rainfall absence (c indicates the considered station and r indicates the reference station).

Besides, the following affinity index is also considered:

$$I_1 = \frac{N_{sp}}{N_t}; \qquad I_2 = \frac{N_p}{N_t}; \qquad I = I_1 + I_2$$

where N_p and N_{sp} are the number of days respectively with or without rainfall contemporary observed in two stations and Nt is the total number of observations.

The proposed methodology has been applied to daily rainfall observed from 1950 to 2004 in selected Sicilian stations, which belong to the real-time monitoring network of the Water Observatory of the Regional Agency for Waste and Water in Sicily (formerly, the Sicilian Regional Hydrographic Office).

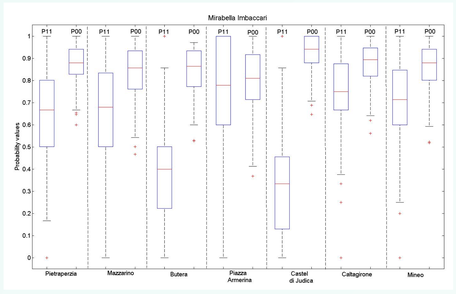

Figure 2. Box plots of P00 and P11 values between contemporary daily data observed at Mirabella Imbaccari station and corresponding reference stations.

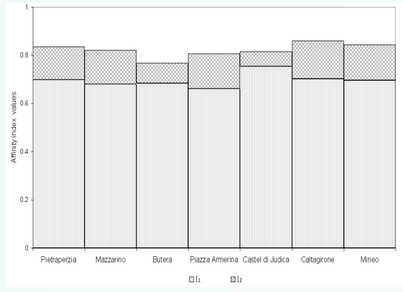

Figure 3. Affinity index values between contemporary daily data observed at Mirabella Imbaccari station and corresponding reference stations.

Applications

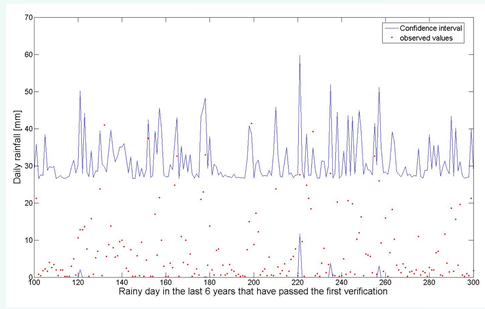

Figure 4. Daily rainfall series observed in the station of Mirabella Imbaccari and validated by the binary network and respective confidence interval

In Figure 4 not null values observed at the station of Mirabella Imbaccari, that have already passed the first quality control through the binary network, and relative confidence intervals (P= 99%) based on the reference stations' data computed by neural network, are shown.

In Table I, the percentage of the validated null and not null data are reported, with reference to a few considered stations, together with some statistics (r2 and RMSE) describing the goodness of the model estimates. From the table, it is possible to observe that such percentage is generally about 85%.

Table I. Statistics of neural network and validation results

Station	Calibration		Validation				
	r^2	RMSE [mm]	r^2	RMSE [mm]	Validated null values [%]	Validate not null values [%]	Validated null and not null [%]
Mirabella Imbaccari	0,52	4,27	0,49	5,22	85,14	84,19	85,00
Sommatino	0,64	6,57	0,47	5,80	85,88	88,74	86,58
Pioppo	0,67	5,49	0,70	3,91	92,98	92,63	92,88
Messina Ist. Geofisico	0,67	3,52	0,67	4,20	88,75	86,24	88,03

Accuracy of the quality control procedure

The accuracy of the proposed quality control procedure has been evaluated by quantifying the model's ability to validate correct data and, conversely, to detect errors in the data. In particular, the accuracy of the procedure has been tested by introducing known errors in the historical dataset and by estimating in terms of frequencies, the following probabilities:

$$P\left[not\text{-}\ correct \mid not\text{-}\ validated\right] \cong \frac{\#\left[not\text{-}\ correct, not\text{-}\ validated\right]}{\#\left[not\text{-}\ validated\right]} \qquad P\left[correct \mid validated\right] \cong \frac{\#\left[correct, validated\right]}{\#\left[validated\right]}$$

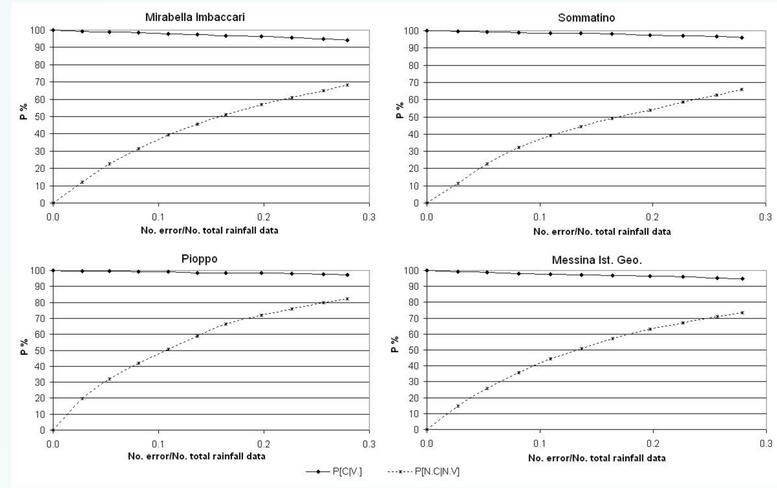

Figure 5. Performance of the model versus the percentage of the errors introduced in the series of daily rainfall data of Mirabella Imbaccari, Sommatino, Pioppo and Messina stations.

Only errors due to the total blockage of rain gauges have been considered by replacing in a random fashion observed values (supposedly correct) with null values, in order to simulate an obstruction in the rain gauge.

In Figure 5 probabilities of data being correct once it is validated, or not correct once it is not validated, are shown as a function of the percentage of erroneous data. In particular, it can be observed that, as the percentage of the introduced errors increases, the probability that validated data are correct remains always larger than 90%, whereas the probability that not validated data are actually not correct increases, but never exceeds 80%, with the exception of Pioppo station for which it goes up to 82%.

Conclusions

On the basis of the obtained results it can be concluded that, when the percentage of errors in the series increases, the probability that not validated data are not correct will increase too, while the probability that validated data are correct will decrease.

Ongoing research is oriented to test the performance of the proposed procedure by considering other kind of errors in the data, as well as to improve the procedure for the estimation of confidence intervals, making use of probability distributions conditioned on the observed values.

References

Abbott, P. F., (1986). *Guidelines on the quality control of surface climatological data.* WMO/TD-No.111, World Meteorological Organization, Geneva, Switzerland.

Campisano, A., Cancelliere, A., Rossi, G., (2002). Una procedura semi-automatica per il controllo di qualità dei dati pluviometrici e termometrici, *Atti del XXVIII Convegno di Idraulica e Costruzioni idrauliche*, Potenza 16-19 settembre, Vol. II, 415-424.

Madsen, H. (1989). Quality control of precipitation measurements in Denmark, *Proceedings of the fourth International Meeting on Statistical Climatology*, Rotura, New Zealand, pp. 13-15.

Reek, T., Doty, S.R., Owen, T.W.A., (1992). Deterministic approach to the validation of historical daily temperature and precipitation data from the cooperative network. *Bulletin American Meteorological Society*, 73(6), 753-762.

Sciuto, G., Bonaccorso, B., Cancelliere, A., and Rossi, G. (2008) Quality control of daily rainfall data through neural networks, *Journal of Hydrology*, Vol 364, Issues 1-2, 15 January 2009, p. 13-22 DOI:10.1016/j.jhydrol.2008.10.008.