The 3rd STAHY International Workshop on Statistical Methods for Hydrology and Water Resource Management
October 1 & 2 2012

Definition of homogeneous regions through entropy theory

M. Rianna[1], E. Ridolfi[1], L. Lorino[1], L. Alfonso[2], V. Montesarchio[1], G. Di Baldassarre[2], F. Russo[1] and F. Napolitano[1]

Email to: maura.rianna@uniroma1.it and elena.ridolfi@uniroma1.it

[1]Dipartimento di Ingegneria Civile, Edile e Ambientale – Università "Sapienza"

[2]UNESCO-IHE Institute for Water Education

Motivation

An important topic for regional flow frequency analysis is the definition of regions, within which the hydrological information can be transferred. In particular, a homogeneous region is a set of catchments, that can be considered homogeneous in terms of hydrologic response and statistical distribution of the hydrological extremes. In the scientific literature different techniques are presented and utilized to group catchments, such as the so called region of influence, the canonical correlation analysis, the correspondence analysis and the L-moments.

Objective

In this work the issue of identifying an homogeneous region is handled by a standard method (ROI defined by Euclidean distance) and by a new entropy based approach. Since basins are defined homogeneous if sharing redundant information, they are characterized by an high value of total correlation (i.e. a measure of redundancy). On the other hand the grouped basins must provide a relevant amount of information, that is measured through their joint entropy (i.e. measure of joint information). In this framework the cluster problem is faced through the definition of a new index (named SNI) that represents the rate between the redundant information and the joint information provided by couple of stations. The higher the value of SNI, the higher the relationship between the two considered stations. From the analysis of the SNI values it is possible to associate to each station the most related ones.
Statistical tests are then carried out to establish whether the grouped data arises from a common regional distribution.

1. ROI Methods through euclidean distance evaluation

Cluster analysis collects data into clusters by joining together objects with similar attribute data.
The ROI technique (Burn, 1990, a, b) is employed as the basis to cluster catchments. The ROI method permits to form a potentially unique collection of stations for each catchment, that can be used for the calculation of the flow quantiles at the site of interest.
The main point of ROI approach to regionalization is the selection of a method for evaluating the similarity between one station and others. In this work the closeness between N basins is evaluated by the Euclidean distance (Burn, 1990) in the M space, where M represents the variables used to define spatial similarity. It is possible to define a distance measure as:

$$D_{ij} = \left[\sum_{m=1}^{M} \left(X_m^i - X_m^j\right)^2\right]^{1/2}$$

where Dij is the distance between catchments i and j, and and are the values of attributes m for catchments i and j. The result of this method is a N x N symmetrical matrix.
The choice of the number of sites that falls into a region is done through homogeneity tests. Then for all the stations clustered the test is carried out from the maximum number of stations to the minimum number that satisfies the test.

2. Brief overview on entropy theory

H(x) is the marginal entropy of a discrete RV. $H(X_1,X_2)$ is the joint entropy, and p is the joint probability of a particular combination of the discharge values in two different sites:

$$H(X) = -\sum_i p_i \log_2 p_i \qquad H(X,Y) = -\sum_{i_1=1}^{N_1}\sum_{i_2=1}^{N_2} p_{i_1,i_2} \log_2 (p_{i_1,i_2})$$

Tot C is the total correlation:

$$TotC(X,Y) = \sum_{i=1}^{N} H(X_i) - H(X,Y)$$

The redundant information among couple of stations is measured through the coefficient of non transferred information:

$$T(X,Y) = H(X) - H(X \mid Y) = H(X) + H(Y) - H(X,Y)$$

It quantifies the amount of information of a variables (e.g. Y) contained in X. (Cover and Thomas, 1991) and it can be seen as the reduction of uncertainty of X, once that Y is known, (Alfonso et al., 2010a).

From the definition of T, several authors (e.g. Mogheir et al., 2004; Yang and Burn, 1994) employed a directional information transfer index (DIT) for measuring the relationship between each pair of stations:

$$DIT(X,Y) = \frac{T(X,Y)}{H(X)}$$

DIT(X,Y) is the fraction of information about station Y that can be inferred from X. It is not symmetrical, thus, DIT(X,Y) is not equal to DIT(Y,X).

In this work, the issue of identifying which stations are more related to another one is faced using a coefficient of shared information normalized by the joint information provided by the considered couple. This normalization allows to evaluate how much information is in common between stations in respect to the maximum information at their disposal, represented by their joint entropy.

$$SNI(X,Y) = \frac{TotC(X,Y)}{H(X,Y)}$$

DIT is the fraction of information transferred from one station to another (Yang and Burn, 1994),
SNI represents the fraction of shared information in respect to the total amount of information that the stations provide considered as a unique element.
The advantage of using ToTC instead of T is the possibility of assessing the total amount of information shared by N variables, without taking into account any partial dependencies among variables (i.e. conditional entropy).
In Figure 1, the concepts of marginal and joint entropy and ToTC are graphically explained.

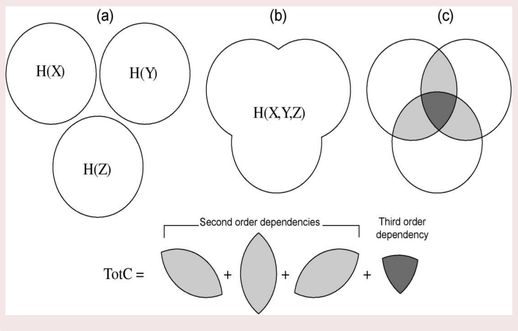
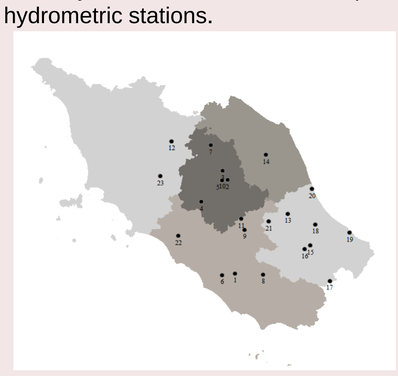


Figure 1 – (a) Three circles representing the marginal entropy of variables X, Y and Z; (b) joint entropy H(X,Y,Z) of the three variables; (c) areas of shared information. Total correlation is given by the summation of the three second-order dependencies and the third order dependency.

3. Case Study

The catchments considered in this study belong to the Central Italy area.
This area has large permeability, with a boundary of impervious rocks, that create a big natural reservoir that feeds the springs.
For this study 23 hydrometric stations are considered.
Data used, obtained by the VAPI project, are the maximum annual flows calculated by the mean daily flows.
For all the stations 31 contemporary years of data were available, starting from 1934 to 1940 and from 1950 to 1973.

Figure 2 – Case study area (Latium, Tuscany, Umbria, Marche, Abruzzo) and hydrometric stations.



Euclidean distance method: variables used are the basin area (km^2), the pervious percentage (%), the maximum and mean altitude (m), the latitude (m), longitude (m) and the mean of the maximum annual flows (m^3/s).

Entropy method: random vectors are the maximum annual discharge of each gauging station. Through the evaluation of the SNI index, stations are clustered forming ensembles from 2 to 7 elements.

For each site the similarity with other stations is evaluated in the parameters space.
The choice of the number of sites that fall into a region is done through homogeneity tests. The homogeneity is tested starting from the 7 more similar stations until the last 2.
Then for all the clustered stations, evaluated with both approaches, the test is done on the growth factor that is the maximum annual discharge standardized by the mean regional discharge.

4. Results

Figure 3 shows the percentage of regions that satisfies the homogeneity tests considering an homogeneous region with an increasing number of stations from 1 to 7 (expressed by the X axis).
Euclidean method: the percentage of stations that passes the test when the region includes only one station (i.e. the main one) is 17,4%, while 21,7%, passes the tests with maximum two stations in the considered region.
Entropy method: the 4,3% satisfies the tests with the main station only in the region, while the 8,0% overcomes the tests with two stations.
It is important to underline that 34,8% passes the tests with maximum 7 stations using the entropy approach, while only 13,0% passes the tests with the other method. Besides, with the standard method the big percentage of regions passes the tests with maximum 5 stations, while with the Entropy method the majority of stations satisfies it with 7 stations.

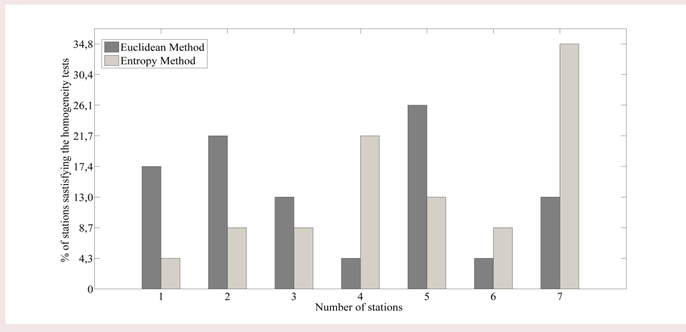


Figure 3 – Percentage of stations that pass the homogeneity tests with different number of stations.

5. Conclusions

In this work a new entropy based approach is compared to a standard procedure for finding the regions of influence.

Results show the big potentiality of the entropy method for regionalization purposes. Indeed this approach performs better in clustering homogeneous regions, because it creates bigger regions: even considering ensembles of 7 stations there is a large percentage of groups that satisfy the test.

The SNI based method is useful, because it allows a simple evaluation of similar sites, being based on the information content of recorded discharge data and it shows good performances compared to a traditional method.
Furthermore this approach can be applied only if the discharge values are known for the gauging stations, thus further studies are on going in order to use it in ungauged basins.
Future perspective is to compare results of the two methods in terms of quantiles.

6. References

Alfonso, L., Lobbrecht, A. and Price, R. (2010a). Information theory-based approach for location of monitoring water level gauges in polders. Water Resources Research, 46: W03528, doi:10.1029/2009WR008101.
Burn, D.H. (1990a). An appraisal of the region of influence approach to flood frequency analysis. Hydrological Sciences Journal, 35(2):149-165.
Burn, D.H. (1990b). Evaluation of regional flood frequency analysis with a region of influence approach. Water Resources Research, 26(10):2257-2265.
Cover, T.M. and Thomas, J.A. (1991). Elements of information theory. New York,Wiley.
Mogheir, Y., de Lima J. L. M. P. and Singh, V. P. (2004). Characterizing the spartial variability of groundwater quality using the entropy theory, I. Synthetic data. Hydrological Processes, 18:2165-2179.
Yang, Y., Burn, D.H. (1994). An entropy approach to data collection network design. Journal of Hydrology, 157:307-324.