

## On the use of informational entropy in GIS

**CINZIA CALORE, PAOLO LA BARBERA & GIORGIO ROTH**

*Istituto di Idraulica, Università di Genova, I Montallegro, 16145 Genova, Italy*

**Abstract** In recent years much attention has been devoted to the possible use of the informational entropy concept as a tool to explain hydrological phenomena and to understand the hydrologic response at the basin scale. In the present work, it is argued that the informational entropy measure introduced by Shannon (1948) includes a shape as well as a scale component. Whilst many authors hypothesize a relationship between entropy and homogeneity on one side, and the space-time scales of hydrologic processes on the other side, suggesting an entropy increase — associated to expected heterogeneity increase — with increasing spatial resolutions, we found that in drainage network analysis, over a certain threshold, this increase is linked to the scale factor alone and not to an increase of effective information. This result can represent a new criterion in the choice of the most appropriate hydrological modelling approach at the basin scale.

### INTRODUCTION

The entropy concept is used in water resources to cope with uncertainties associated with hydrologic variables, hydrologic systems and their models, and probability distribution function parameters. Different definitions of entropy are available for the different uses. Entropy is a mathematical quantity describing disorder or uncertainty. Boltzmann (1872) introduced the concept of entropy to measure the disorder in a thermodynamic system, while Shannon (1948) used the concept of informational entropy to measure the uncertainty associated with a given information. According to Wilson (1970) four ways to view the concept of entropy can be defined: entropy as a measure of system properties (e.g. order and disorder, reversibility and irreversibility, complexity and simplicity); entropy as a measure of information, uncertainty, or probability; entropy as statistics of a probability distribution for measure of information or uncertainty; entropy as the negative of a Bayesian log-likelihood function for measure of information.

In hydrology, as in any other science, information is an important tool to deal with in order to understand various phenomena, to characterize and to forecast them. Describing the environment of hydrological phenomena under investigation and to understand their development, a way has been found to be fruitful, i.e. the application of the rule of minimal energy expenditure. The pioneers of this idea were Leopold & Langbein (1962): by analogy with the thermodynamic application, they applied concepts of energy dissipation and entropy to river channels analysis, founding some general rules affecting their hydraulic geometry. Yang (1971) introduced the law of least rate of energy expenditure which states that, during the evolution toward its equilibrium condition, a natural stream chooses its flow course in such a manner that the rate of potential energy expenditure per unit mass of water is minimum. Entropy concepts can be used to investigate channel networks and

plano-altimetric variations in the river configuration. Much of the work employing entropy concepts in hydrology have been done with reference to the informational entropy, in particular on applications of geomorphology to the understanding of drainage basin response (for a review see Singh & Fiorentino, 1992).

Informational entropy concepts are here applied to investigate the informational content associated with geomorphologic data: in particular data used in hydrologically oriented Geographical Information Systems (GIS) are analysed to reproduce channel networks for hydrological mathematical simulations in order to find a scale discretization criterion.

## INFORMATIONAL ENTROPY OF A DRAINAGE BASIN

The concept of river network informational entropy can be associated with the hypothesis that water discharge is the main source of information used by the drainage system to adjust its configuration: according to the principles of maximum entropy and minimum entropy production rate (Singh & Fiorentino, 1992), the basin tends to shape its form in order to spend the least amount of energy to produce and transfer runoff. A drainage network can be described as a treelike structure in which external (sources) and internal (segment junction) nodes can be identified. Links are segment between nodes. The topological distance  $d$  of a node from the outlet — or topological level of a link which originates from it — is defined as the number of links forming the path between that node and the outlet. The maximum of the topological distances within the network is called the topological diameter  $D$ . Another useful descriptor of the basin is the width function  $W(x)$ , giving the number of links at a given distance  $x$  from the outlet. Following Shannon (1948), the informational entropy  $S$  can be defined as

$$S = -\sum_{i=1}^N p_i \ln(p_i) \quad (1)$$

in which  $N$  is the number of possible states of the system and  $p_i$  the probability to be in state  $i$ . When applied to a river network, the informational entropy can be measured through the width function of the network. In this case  $N$  is equal to the topological diameter of the network, and  $p_i$  is the ratio between the number of links at the topological distance  $i$  from the outlet, and the total number of links within the whole network. In this case the informational entropy can be written as

$$S_D = -\sum_{i=1}^D p_i \ln(p_i) \quad (2)$$

The maximum value of equation (2) is obtained in absence of natural constraints as

$$S_{\max} = \ln(D) \quad (3)$$

and is referred to as primary entropy of the network (Kapur, 1990). If the river network is rescaled, i.e. each link is subdivided into  $k$  elements, each of them characterized by a probability  $p_i/k$ , the informational entropy of the rescaled network

$S_{kD}$  can be written as

$$S_{kD} = -\sum_{i=1}^{kD} \frac{p_i}{k} \ln \frac{p_i}{k} = -\frac{1}{k} \sum_{i=1}^{kD} p_i \ln p_i + \sum_{i=1}^{kD} \frac{p_i}{k} \ln k \quad (4)$$

and therefore

$$S_{kD} = S_D + \ln(k) \quad (5)$$

From this, one can think of the informational entropy as the sum of two components, the first one linked to the structure of the drainage network, the second linked to the scale of description assumed for drainage representation. This approach can be viewed as limited by the assumption that the drainage must be fully described at level  $D$ . On the other side, equation (5) can be used to infer the optimal level of drainage description, starting from a network described at a finer scale. In fact, while reducing the informational content, one could expect that  $S$  follows the relation described by equation (5) down to a certain value, from which the measured value of  $S$  starts to deviate from the prediction given in equation (5). This value could be assumed to give, for the drainage under examination, the optimal scale of representation, that is the scale at which the natural complexity of the network is captured and no useless information is artificially included by a resolution finer than the optimal. In the following this concept is exploited with reference to drainage network analysis, even if we believe that it could be of relevant importance also in other fields.

## DRAINAGE NETWORKS FROM DIGITAL ELEVATION MODELS

Geographical Information Systems represent a fundamental tool for most applications in environmental and natural resources inventory and analysis. Land surveying and data collection of the spatial distribution of significant properties of Earth's surface has been an important part of activities of organized societies for long time. The demand of topographic maps and specific themes of the Earth's surface, such as natural resources, has accelerated. Since the 1960s the availability of computers and the development of techniques for remote data collection (aerial photographs, satellite images) have given new possibility to the science of data collection, management and elaboration.

In GIS the representation of data related to the description of relief over space is known as Digital Elevation Model (DEM). O'Challagan & Mark (1984) define a digital elevation model as any numerical representation of the elevation of all or part of a planetary surface, given as a function of geographic location. Following their example, we consider the most commonly used data structure for DEMs, i.e. the regular square grid. In this structure data are stored in an altitude matrix arranged on a grid with each value giving the elevation of a point. The location within the matrix supplies the spatial location of the point, so that no information about the value position has to be stored. Several algorithms for an automatic derivation of the channel network using DEM have been proposed. The method applied in this study is based on channel segment production following the direction of maximum slope

angle for each pixel (i.e. cell of DEM grid) within the boundary of the basin. As a consequence the drainage density is constant throughout the basin and dictated by the DEM resolution, i.e. the pixel size.

One of the primary questions dealing with automated extracted channel network is that of the appropriate drainage density. Some authors suggest criteria to find this appropriate scale. La Barbera & Roth (1994) proposed a filtering procedure based on the identification of a threshold value for the quantity  $AS^k$ , where  $A$  is the contributing area,  $S$  the stream slope and  $k \cong 2$ . This procedure consists in the progressive removal from the drainage network of the first order stream which presents the minimum  $AS^k$  value; the procedure is iterated up to a given target value for the area drained by first order streams. For the purposes of the present work the  $k$  value is set to zero and the drained area alone is used as a filter. Filtered drainage networks are studied in order to investigate the scaling properties of a proposed measure for appropriate drainage density identification: the informational entropy.

## INFORMATIONAL ENTROPY AND THE DISCRETIZATION SCALE

For a filtered drainage network, one could expect the diameter to be proportional to the mainstream length, and that this is proportional to the drainage area,  $l_\Omega \propto A^\alpha$  (Gray, 1961), over which streams are allowed to develop, that is

$$D_a \propto l_{\Omega,a} \propto (A^\alpha - a^\alpha) = A^\alpha \left[ 1 - \left( \frac{a}{A} \right)^\alpha \right] \quad (6)$$

in which  $A$  is the basin area,  $a$  is the area used to filter the network, i.e. the minimum contributing area, both measured in DEM pixels, and  $D_a$  and  $l_{\Omega,a}$  are the diameter and mainstream length of the filtered network. It follows that the primary entropy of the filtered network,  $S_{\max,a}$ , is a function of  $a$ , and can be written as

$$S_{\max,a} \propto \ln \left\{ A^\alpha \left[ 1 - \left( \frac{a}{A} \right)^\alpha \right] \right\} \quad (7)$$

Its maximum value,  $S_{\max,1}$ , is obtained when  $a$  reaches a minimum, that is, using a grid DEM as database for network identification,  $a = 1$  pixel.

$$S_{\max,1} \propto \ln(A^\alpha - 1) \quad (8)$$

For a given basin area, equation (7) predicts a linear relation (in a semi-log plot) between the minimum contributing area  $a$  and the entropy value. If this theoretical result is associated with the prediction given in equation (5) a critical value for  $a$  can be identified, below which the measured informational entropy is expected to deviate from the theoretical predictions.

## APPLICATION AND RESULTS

The above reported theoretical derivations are compared with field results, as obtained from a DEM. For the proposed analysis two different areas of study have

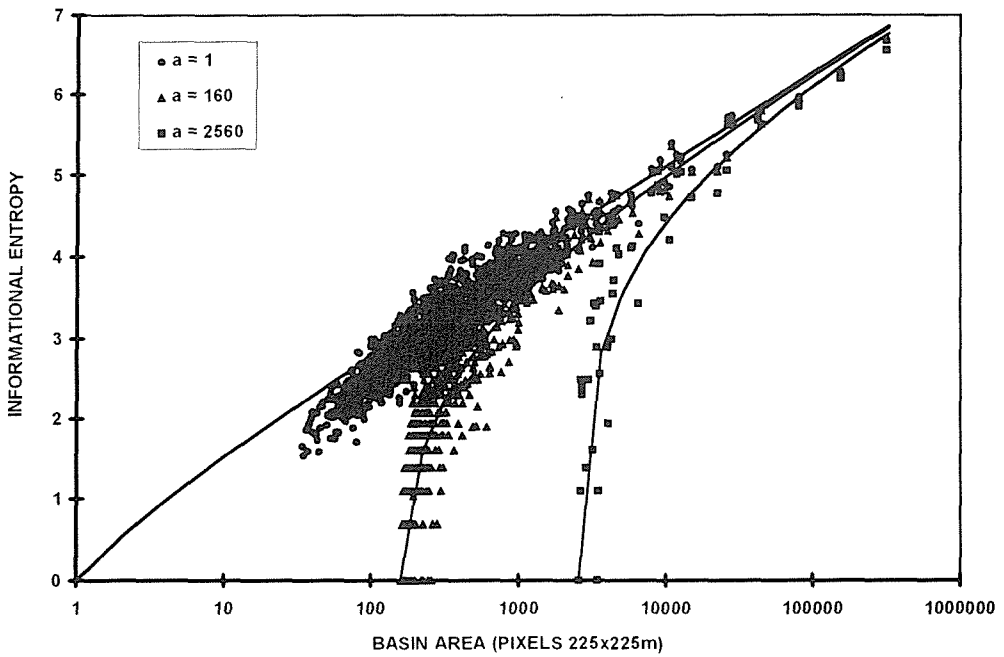
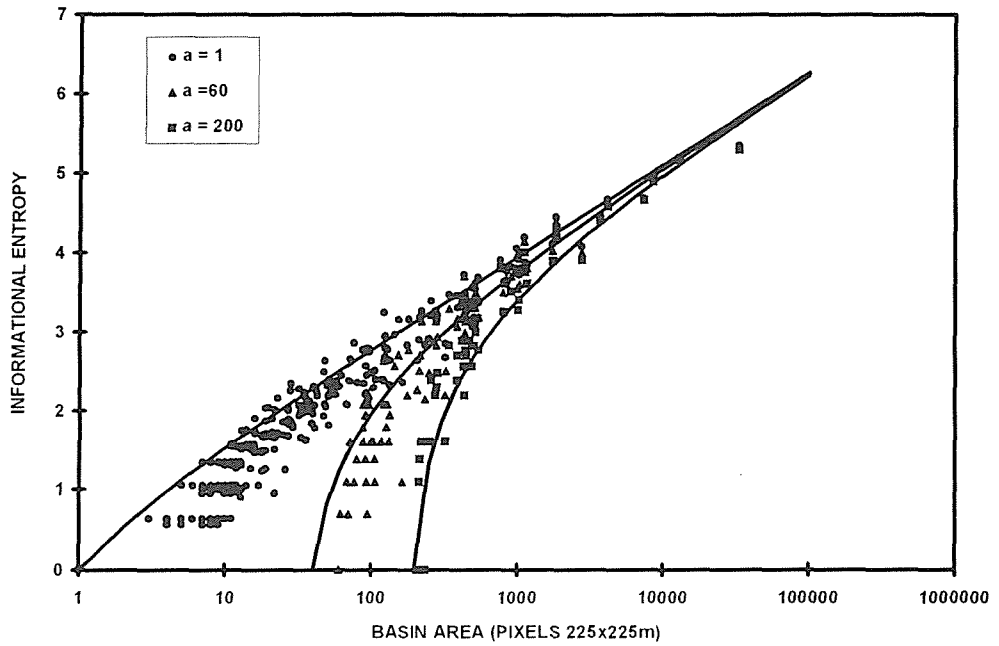
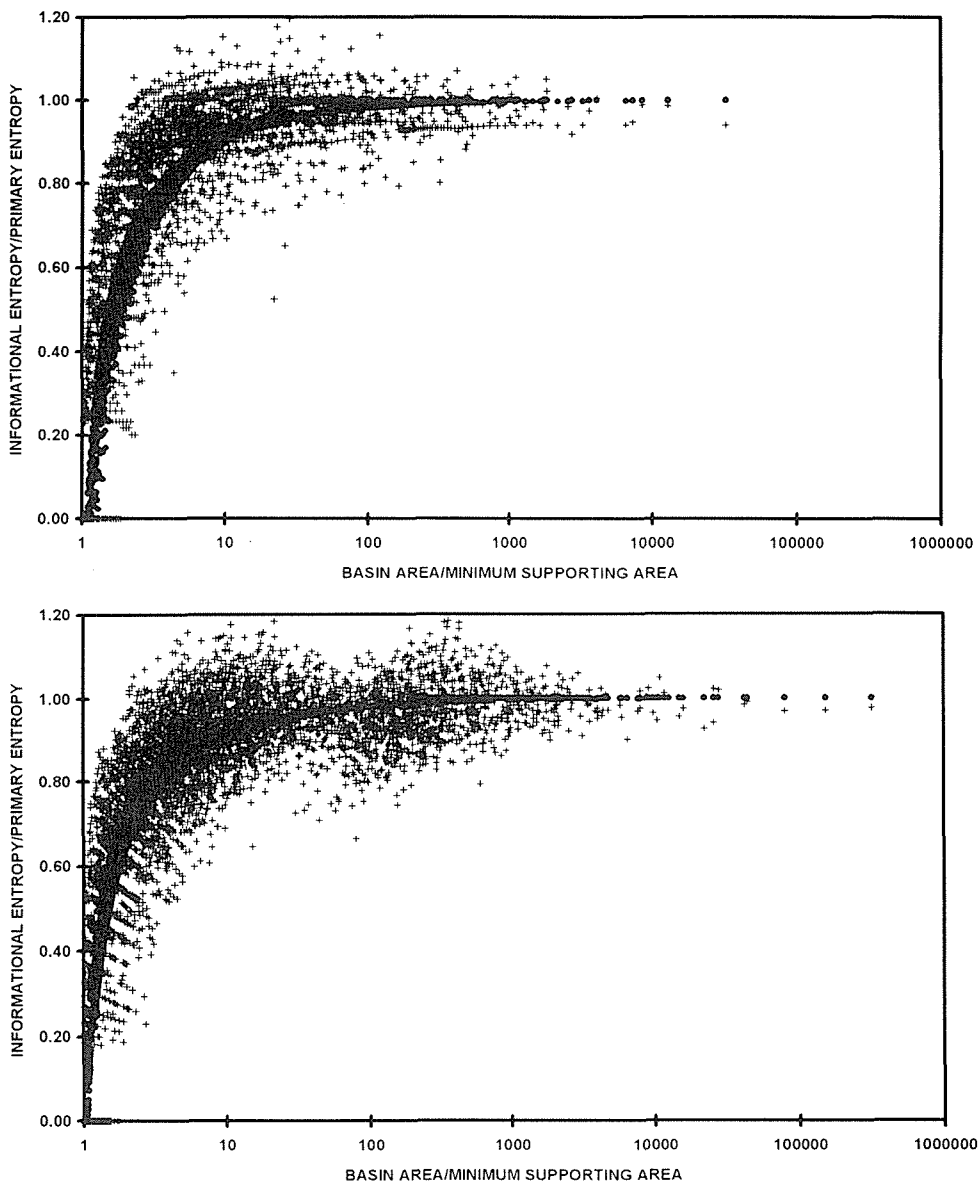


Fig. 1 Informational entropy as a function of basin area: theoretical predictions (continuous lines) and measured values at different minimum contributing areas (points). Liguria region (top) and Tevere River basin (bottom).



**Fig. 2** Dimensionless informational entropy as a function of dimensionless basin area: theoretical predictions and measured values for the Liguria region (top) and the Tevere River basin (bottom).

been considered, both located in Italy. The first nearly overlaps the Liguria region while the second consists of the Tevere River basin. First of all, a space filling drainage network has been identified for all of the coastal basins of the Liguria region and for the Tevere River basin, using a DEM with a grid size of  $225 \text{ m} \times 225 \text{ m}$ . For each of the basins of the Liguria region and for the whole Tevere basin as well as for a number of its sub-basins, the width functions and the informational

entropy values related to their drainage network have been evaluated. The procedure has been repeated for different filtering levels, and the informational entropy for the filtered drainage networks identified.

Results of the proposed analysis are summarized in Fig. 1, where the entropy values are plotted against basin area, measured in a number of cells. Each point in the plot represents the entropy value related to a basin. Different symbols are introduced to identify different filtering levels. The curves are representing the primary entropy behaviour as function of the basin area for different minimum contributing areas, as presented in equation (7): one could observe that measured entropy values tend to approach the primary entropy component described by equation (8), with a dispersion that seems to indicate a shape factor. Most of all, it is important to note that the deviation from a linear behaviour (in the semi-log plot) is enhanced when the ratio between basin area and minimum contributing area tends to approach a value of the order of fifty. This behaviour is highlighted in Fig. 2, where entropy values are made dimensionless with reference to primary entropy values, while basin areas are nondimensional with reference to minimum contributing area. This trend seems to indicate a sharp decrease in the informational content when the system is described with less than about fifty lumped sub-systems.

## CONCLUSIONS

Theoretical derivations and experimental evidence show that, above a certain threshold, an increase of resolution in the spatial description of drainage networks obtained from digital elevation model interpretation, cannot be directly linked to an increase of information, at least if the informational entropy is assumed to measure the informational content of the network. In particular, informational entropy values obtained from filtered drainage networks are at first quite constant with respect to an increase of the area used to filter the network, then start to decrease sharply. This behaviour can be reasonably linked to the fact that, at large values of the minimum contributing area, the effective structure of the network is partially destroyed. When a DEM derived drainage network is introduced in a GIS structure to represent the effective drainage structure of the basin, this approach can be used to balance the conflict between the increase in the amount of data and the necessity to reproduce the effective drainage structure of the basin.

In this way a criterion can be identified for the choice of the appropriate modelling and scale approach for the simulation of hydrological phenomena at the basin scale. In fact, according to the optimization of the informational content, a maximum contributing area value is identified so as to define a threshold size basin which can be considered and described as a lumped system.

## REFERENCES

- Boltzmann, L. (1872) *Weitere Studien über das Wärmegleichgewicht unter Gas-molekulen* (in German), K. Acad. Wiss. (Wien) Sitzb., II Abt., 66, 275.
- Gray, D. M. (1961) Interrelationships of watershed characteristics. *J. Geophys. Res.* 66, 1215-1223.
- Kapur, J. N. (1990) *Maximum-Entropy Models in Science and Engineering*. John Wiley, New York.

- La Barbera, P. & Roth, G. (1994) Scale properties and scale problems: network morphology and network identification from digital elevation maps, In: *Advances in Distributed Hydrology* (ed. by R. Rosso et al.), 131-148. Water Resources Publ., Highlands Ranch, Colorado.
- Leopold, L. B. & Langbein, W. B. (1962) The concept of entropy in landscape evolution. *US Geol. Surv. Prof. Pap.* 550-A, A-20. Washington, DC.
- O'Challaghan, J. F. & Mark, D. M. (1984) The extraction of drainage networks from digital elevation data. *Computer Vision, Graphics and Image Processing* 28, 323-344.
- Shannon, C. E. (1948) A mathematical theory of communication. *Bell System Technol. J.* 27, 379-423, 623-659.
- Sing, W. P. & Fiorentino, M. (1992) A historical perspective of entropy applications in water resources. In: *Entropy and Energy Dissipation in Water Resources* (ed. by W. P. Sing & M. Fiorentino), 21-61. Kluwer, Boston, Massachusetts.
- Wilson, A. G. (1970) The use of the concept of entropy in system modelling. *Operational Res. Quart.* 21, 247-265.
- Yang, C. T. (1971) Potential energy and stream morphology. *Wat. Resour. Res.* 7, 311-322.