Revista Brasileira de Recursos Hídricos
Brazilian Journal of Water Resources
Versão On-line ISSN 2318-0331
RBRH, Porto Alegre, v. 24, e42, 2019
Scientific/Technical Article

https://doi.org/10.1590/2318-0331.241920180188

Analysis of the fluviometric network of Rio das Velhas using Entropy

Análise da rede fluviométrica do Rio das Velhas com emprego de Entropia

Luiz Henrique Resende de Pádua[1], Nilo de Oliveira Nascimento[1], Francisco Eustáquio Oliveira e Silva[1] and Leonardo Alfonso[2]

[1]Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brasil
[2]IHE Delft Institute for Water Education, Delft, Holanda
E-mails: lhr.padua@gmail.com (LHRP), niloon@ehr.ufmg.br (NON), fsilva@ehr.ufmg.br (FEOS), l.alfonso@un-ihe.org (LA)

Received: November 30, 2018 - Revised: July 22, 2019 - Accepted: August 18, 2019

ABSTRACT

In this work a comparative study was carried out, in which different methods were used in the literature that seek to evaluate the number of stations and the quality of the information generated by the hydrometric network of a watershed, using Information Theory concepts. The underlying idea is the so-called optimal network whose function, according to World Meteorological Organization (WMO) is to optimally and inexpensively meet the primary goal of hydrometry, which is to provide the necessary information with a minimum number of stations correctly positioned in the basin. Methodologies based on Information Theory ascend to fill the gap on a standard method for the design of hydrometric networks. The evaluated methods were applied to the subbasin of the Rio das Velhas belonging to the São Francisco River basin in Brazil. The results showed that the methods analyzed, which use the concept of entropy, are adequate and efficient for evaluation of existing fluviometric networks, since they allow the reduction of eventual redundancies and at the same time, seek to maximize the information generated. It was possible to compare them and indicate the most appropriate method for the application within the national context, as well as indicate new methods for use thereof.

Keywords: Theory of information; Entropy; Hydrometric network; Hydrological modeling; Rio das Velhas.

RESUMO

Neste trabalho foi realizado um estudo comparativo, onde foram empregados diferentes métodos disponíveis na literatura que buscam avaliar a quantidade de estações e a qualidade da informação gerada pela rede hidrométrica de uma bacia hidrográfica, utilizando conceitos da Teoria da Informação. A ideia subjacente é a da chamada rede ótima, cuja função, segundo a World Meteorological Organization (WMO), é a de atender, de forma otimizada e a baixo custo, o objetivo precípuo da hidrometria, qual seja, o de oferecer a informação necessária com um número mínimo de estações corretamente posicionadas na bacia. Metodologias baseadas na Teoria da Informação ascendem para ocupar a lacuna sobre um método padrão para projeto de redes hidrométricas. Os métodos avaliados foram aplicados à sub-bacia do Rio das Velhas pertencente à bacia do Rio São Francisco, no Brasil. Os resultados demonstraram que os métodos analisados, os quais se valem do conceito de entropia, são adequados e eficientes para avaliação de redes fluviométricas existentes, uma vez que permitem a redução de eventuais redundâncias e ao mesmo tempo, buscar a maximização da informação gerada. Foi possível compará-los e indicar o método mais adequado à aplicação dentro do contexto nacional, assim como indicar novas metodologias para utilização dos mesmos.

Palavras-chave: Teoria da informação; Entropia; Rede hidrométrica; Modelagem hidrológica; Rio das Velhas.

BACKGROUND

The hydrometric network, according to WMO (2008), can be defined as a set of facilities developed for the purpose of collecting data of the different components of the hydrological cycle. This is designed and operated to support decision making compatible with the broad context of hydrology and hydraulics.

A well-structured hydrometric network requires a certain amount of observation points duly located within the catchment area, and of course, an understanding of the information capacity that each station holds and can produce. However, building and maintaining a well-structured hydrometric network in most countries has been an arduous task, often due to the limited number of monitors installed and the amount of data available (MISHRA; COULIBALY, 2009; PYRCE, 2004; SAMUEL et al., 2013).

WMO (2008) recommends that a minimum hydrometric network be established before attempting to reach the network considered to be well structured. The minimum network should fulfill the role of providing good quality data to meet the most pressing requirements of regional water resources development, indicating the number of posts required within the basin in terms of size and geomorphology.

In this study, with particular reference to fluviometric monitors, we intend to identify the optimal network, as defined by WMO (2008) for river basins. In summary, it is proposed to identify among the already established monitors, those that best allow to characterize the seasonal flow regime (greater generation of information), and those whose generated information is redundant, and therefore can be relocated or even deactivated.

The main objective of this work is to evaluate the hydrometric network of existing fluviometric monitors of the Rio das Velhas basin, located in the central region of the State of Minas Gerais, Brazil. The watershed of Rio das Velhas, the largest tributary of the São Francisco River basin, is approximately 761 km long, which has an area of 29,173 km^2.

Different methods seek to evaluate and design hydrometric networks. Mishra and Coulibaly (2009) present a review of several methods, which can vary from those based on statistical analysis, geomorphological analysis, user surveys, Information Theory, among others. However, the most prominent or considered most promising method is certainly those based on Information Theory, which uses the concept of maximum Entropy and information transfer. Such relevance is pointed out by these authors, as well as in the complementary review by Keum et al. (2017).

For the analysis of said hydrometric network, the methodologies based on Information Theory (SHANNON, 1948) were used here. They use the concept of Entropy to assess the number of monitors and the quality of the information generated by the hydrometric network in hydrographic basins by measuring the information generated.

As stated by Alfonso et al. (2014), the Information Theory provides methods capable of quantifying the information contained in a single random variable and methods capable of quantifying the information contained and shared between two or more of these variables.

In this study, we compare the methods and recommendations already elaborated by similar studies carried out (e.g., ALFONSO et al. 2010a,b, 2014; LI et al., 2012; MISHRA; COULIBALY, 2010, 2014; WERSTUCK; COULIBALY, 2016, 2017; SAMUEL et al., 2013), applied to Rio das Velhas, in continuation of studies previously performed at the national level by Gontijo Junior and Koide (2012) and Pádua et al. (2018a,b).

This more comprehensive subsequent assessment will enable the analyses the relevance of the application of these methodologies in Brazilian basins of tropical climate, characterized by large geographic coverage, between urban and rural areas, and that show a great seasonal variation of flows. We attempted to mainly an analysis of sensitivity of input parameters of these methods, as its variation is influenced by data used in addition of the suitability of these to the spatial and temporal intervals indicated by the works available in the literature, mentioned above.

The choice of a methodology that employs the concept of Entropy for this study comes from its conceptual base, strongly linked to the examination of the variance of the flow series. Thus, it is possible to evaluate if, in these extreme cases, such as those observed in Brazilian basins, such methods can adequately evaluate and indicate an optimal minimum fluviometric network. The concept of a minimum network is re-emphasized, since such methods are based exclusively on the probabilistic analysis of the historical data of the monitors, being beyond the scope of this work more specific questions related to the hydrometric network design.

In short, we want to achieve the following objectives with this work: (i) to verify the fluviometric network of Rio das Velhas in relation to the recommendations of WMO (2008), by the amount of monitors required, compared to that obtained through analysis using the Theory of Information; (ii) propose a methodology of discretization of continuous flow data for the application of such concepts; (iii) evaluate two methods of ranking monitors using Entropy; (iv) identify the monitors that contribute the most and least in the amount of information to the network, as well as those that are contributing redundantly.

INFORMATION THEORY - ENTROPY

According to Shannon (1948), information theory provides mechanisms to measure the information contained in a discrete (or discretized) random variable X (SINGH 1997; ALFONSO et al., 2010a). This amount of information is known by Entropy, that within the Theory of Information, represents the measurement of information, and can sometimes also be referred to as Uncertainty or Marginal Entropy.

The meaning of this amount of information, according to these authors, is that, assuming a set of N events, uncertainty refers when it is not known which of these N events will occur. The magnitude of this uncertainty is related to the amount of information known about the events, that is, once one occurs and the same is observed, the uncertainty decreases as we receive the information, which leads to the assertion that: "information can be considered a reduction of uncertainty" (ALFONSO et al., 2013).

According to Shannon and Weaver (1949), the Entropy $H(X)$ of a discrete random variable X, with discrete values $x_1, x_2 \ldots x_n$ with probabilitites $p(x_1)$, $p(x_2)$,...$p(x_n)$, where n represents the number of elements, is given by:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i) bits. \quad (1)$$

In Equation (1), H refers to the standard notation for quantity of information, its unit being in information *bits*, according to Shannon (1948). In this work the logarithmic base can be omitted unless it is necessary to express it on a basis other than 2. In this same equation, X represents a discrete random variable and its entropy will be denoted by $H(X)$. This means that X is not an argument of a function, it represents only the indication of that variable, to differentiate it from another possible variable Y, which would have its entropy represented by $H(Y)$, for example.

As shown by Singh (1997), entropy represents a measure of the amount of chaos, or rather the lack of information about a system. What finally, according to this author, can be affirmed that the entropy represents the quantification of our ignorance on a system. Which then leads to its opposite, if there has all the information, consequently the entropy of the system will be zero.

As detailed by Shannon and Weaver (1949), the use of the amount of entropy H allows advantages over other possible quantifiers because of their properties, which make it interesting to measure quantity of information mainly when the study is aimed at more than one variable.

Among the properties, some stand out as the Joint Entropy (JH), that is, the amount of information contained between two random variables. For two random variables, X and Y, stochastically independent, the total entropy will be the sum of the individual entropies ($H(X) + H(Y)$). But if these variables are stochastically dependent, some of the contained information is shared, their joint entropy will be then (ALFONSO et al., 2013),

$$H(X,Y) = -\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i, y_j) \log p(x_i, y_j), \quad (2)$$

whre $p(x_i, y_j)$ represents the joint probability distribution between the variables X and Y.

According to (COVER; THOMAS, 2012), another important relationship is the one that refers to the Mutual Information between two variables, sometimes also called Transformation (I), that is, the relation that expresses the reduction of the uncertainty of X due to the knowledge of Y. This relation is expressed as follows,

$$I(X;Y) = \sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i, y_j) \log \frac{p(x_i, y_j)}{p(x_i)\, p(y_j)}, \quad (3)$$

The principle of application of the entropy measures for the planning of hydrometric networks is that the data coming from the monitors must have among themselves the least Transformation possible, that is, they must be minimally dependent, from the statistical point of view. If the Transformation between the observed data in two monitors is high, the same information is duplicated, which allows to dispense with one of the analyzed monitors and reallocate it to a new fluvial section, with corresponding Transformation close to zero.

Another way of directly and effectively accessing the dependency between multiple random variables can be performed using the Total Correlation (CT), C, by the following formulation (ALFONSO et al. 2010a; MCGILL, 1954):

$$C(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i) - H(X_1, X_2, \ldots, X_n). \quad (4)$$

The total correlation C will always be positive if the entropy sums of all variables are greater than their joint entropies; greater than zero if two variables have some dependence; and will be zero if all these same variables are independent of each other.

Following is the methodology used for network optimization using the concept of Entropy and multiobjective optimization, followed by results and discussion.

METHODOLOGY

Two different models for analysis of hydrometric networks using entropy were used. The first one developed by Alfonso et al. (2010b), called *Water Level Monitoring Design in Polders* (WMP) and the second one by Li et al. (2012), called *Maximum Information Minimum Redundancy* (MIMR).

The choice of these two methods, even if both based on the concept of Information Theory, comes from its different approaches to form of optimization and design of hydrometric networks. But specifically, this distinction falls on how algorithms work to find the optimal set of monitors, capable of maximizing information with as little redundancy as possible.

The first one, WMP, searches for this optimal set through a system of evaluation of the monitors that most aggregate information to the network, simultaneously analyzing the generation of information and its dependencies by minimizing Transformation, to a parameter that seeks the identification of independent variables. Thus, the primary objective of the method is to maximize the information provided by the set of independent monitors by minimizing Transformation and Total Correlation.

The second method, MIMR, in a different way, searches for maximum information by the joint analysis of Entropy and Transformation, maximizing both, contradicting the basic idea of searching for independent monitors. However, reducing redundancy is a third objective function, referring to Total Correlation. Both methods are best detailed in the following topics.

WMP method

Developed by Alfonso et al. (2010a) and Alfonso et al. (2010b), the method was initially designed with the objective of locating and evaluating monitors for monitoring water levels in polders. Polders are regions formed by portions of low, underwater, floodplain, protected by dikes that artificially disconnect them from the hydrological regime of neighboring systems, in order to keep water levels to be used for different purposes.

The chosen study region was Pijnacker, Delfland, in the Netherlands. This model was also applied later by Alfonso et al. (2013) in the Magdalena River basin, Colombia, for the analysis of fluviometric monitors networks. In all cases the results were very promising, becoming a model for the design and analysis of

hydrometric networks, using the concept of entropy, reference in the area.

The application of the WMP model consists of five basic parts as detailed in Alfonso et al. (2010a): (i) the generation of a dense time series of data, by a hydrodynamic model or regionalization, if potential sites are studied for the monitors; (ii) the quantization of the data to remove possible noises from the series; (iii) use of three different criteria in parallel to evaluate the dependency between the data series; (iv) a procedure to locate the potential points for the installation of the monitoring monitors; and (v) the use of Equation 4 to evaluate the dependence between multiple variables.

For step (iv), two parallel processes are used. The first evaluates the network by Equation (2), that is, from the points with the highest entropy values. In this process a multiobjective analysis is performed aiming at the maximization of the individual entropy and the minimization of Transinformation, according to Equation 5,

$$\begin{cases} Max.H(X_1)+\ldots H(X_i) \\ Min.\sum_{j=1}^{i-1} T(X_i;X_j) \end{cases} \quad (5)$$

The second uses the process introduced by Jakulin and Bratko (2004) and Fass (2006), adapted by Alfonso et al. (2010a) for use in water resources, called Directional Information Transfer Index (DIT). This refers to a normalization of the mutual information between two monitors, to obtain the fraction of information transferred from one station to another, receiving a value between 0 and 1, according to Equation 6,

$$DIT_{XY} = \frac{I(X;Y)}{H(X)} \quad (6)$$

where DIT_{XY} refers to the information received from *X* por *Y*. When *H(Y)* is used in the denominator of Equation 10, DIT_{YX} becomes the information sent from *X* to *Y*.

Thus, the algorithm of the proposed methodology follows the sequence of calculations (Alfonso et al., 2010a):

i. Read the data of all proposed monitors.

ii. Calculates the marginal entropy of each of the monitors.

iii. Calculates the mutual information between all the monitors, generating a symmetric matrix T,

$$T = \begin{bmatrix} I(X_1;X_1) & I(X_1;X_2) & I(X_1;X_n) \\ I(X_2;X_1) & I(X_2;X_2) & I(X_2;X_n) \\ I(X_n;X_1) & I(X_n;X_2) & I(X_n;X_n) \end{bmatrix} \quad (7)$$

iv. The system is separated into two so-called dependent and independent sets, where the values evaluated sequentially by row are considered independent, where each element is checked by the relation $I(X_i;X_n) < \varepsilon$, where ε is considered as the mean between the row array vectors.

Mark as chosen the station that has the largest entropy within the independent vector.

These procedures can be verified in more detail in (ALFONSO et al., 2010a).

MIMR method

The second method, MIMR, is a model based on multiobjective analysis, which seeks to lease the studied monitors by maximizing the amount of information each one generates (marginal entropy), maximizing the information transformation and minimizing information redundancy between monitors (total correlation),

$$\begin{cases} Max.H(X_1)+\ldots H(X_i) \\ Max.\sum_{j=1}^{i-1} T(X_i;X_j) \\ Min.\sum_{j=1}^{i-1} C(X_i;X_j) \end{cases} \quad (8)$$

As can be observed by Equation 12, this indicates that the algorithm will result as an optimal network, the one that reaches the best point of convergence between the maximum effective information, by minimizing the information redundancy measured by the total correlation. Due to the complexity of solving the three equations simultaneously, the algorithm adopts an integrated equation of resolution by maximizing the sum of the three,

$$Max.\lambda_1\left(H(X_1)+\ldots H(X_i)\right)+\sum_{j=1}^{i-1} T(X_i;X_j)-\lambda_2\sum_{j=1}^{i-1} C(X_i;X_j) \quad (9)$$

In Equation 9, according to Li et al. (2012), two coefficients, λ_1 and λ_2, were added. These coefficients function as an additional resource made available so that the analysis operator can, according to their perception of the system, adjust the portions of the equation according to their respective representations or weights, that is, delegate importance to the total information or if there is doubts, allocate greater weight to the redundancy of the information generated.

In summary, the proposed algorithm follows the following steps to select the optimal network (LI et al., 2012): (i) First read the data of the monitors used; (ii) performs the discretization of the data of the monitors. This step will be further explained in the next topic; (iii) identifies the central station, delegated by the station with the highest marginal entropy value; (iv) create the first set of optimal monitors; (v) selects the second set of optimal networks by the resolution of Equation 9. These steps are then repeated countless times until the set that best satisfies Equation 08 is obtained.

Discretization / quantization

The quantization or discretization of the data series of the monitors consists basically of a "cleaning" of the existing noises. Its concept derives from the systems of Information Theory, and its main objective is to convert a continuous signal into a signal of discrete pulses, to enable its digital transmission by applying a mathematical leveling function (ALFONSO et al., 2010a).

The data discretization procedure was successfully applied in similar cases by authors (RUDDELL; KUMAR, 2009; ALFONSO et al., 2010a,b; LI et al., 2012). Its necessity comes from

the subjectivity of methods hitherto applied for discretization of data by histogram, mainly related to the size of the intervals used.

Techniques that use discretization by histograms are questionable due to the arbitrary step in the number of histogram intervals, even after some already well-known studies (SHIMAZAKI; SHINOMOTO, 2007) to optimize their number, since the marginal entropy is very sensitive or influenced by the choice of this number of intervals (LI et al., 2012). Thus, using this type of technique, depending on the number of intervals adopted, different entropy values will be obtained.

In this work, we chose to follow the methodology proposed by Alfonso et al. (2010a) and Li et al., (2012), using the discretization of the data by the quantization of the data by a mathematical leveling function, where a continuous value of the data series *x*, is converted to a nearest integer multiplied by a constant *a*, according to Equation 10,

$$x_q = a\left(\frac{2x+a}{2a}\right) \tag{10}$$

where the term in parentheses represents the leveling function. As shown by (Li et al., 2012), the use of the leveling function provides two important advantages: (i) it eliminates the need to use a parametric function that fits the continuous data; (ii) it incorporates a physical parameter, represented by constant *a* in its resolution, directly involved in the nature of the data series.

The value for parameter a can be obtained in different ways according to the context of the study carried out. According to Alfonso et al. (2010a), this value can be obtained by analyzing the data series, taking as reference the minimum dimensional value of the series or the observed average value of the lower tributary to the studied channel, which as placed by those authors, is crucial for the calculation of the marginal entropy, since high values for the data series can generate high values of entropy, but do not necessarily provide the information for the management of the system.

In contrast to the form proposed by Alfonso et al. (2010a), an important study by Li et al., (2012) on the choice of parameter *a*, given its empirical nature, and not always explicit, is the use of a trial and error. This methodology to obtain this parameter, starting from a median value, so as to guarantee the maintenance of some basic concepts, such as: (i) ensure that all candidate monitors have significant and distinguishable information contents; (ii) the spatial and temporal variability of the station series must be preserved as much as possible before and after discretization; (iii) the selected monitors should be stable as much as possible, since the value of a fluctuates within a central range close to its optimal value.

In this way, as shown by Li et al. (2012), after the application of this mathematical leveling function, Equation 10, the marginal entropy of the data series of the station no longer represents its generated information, but its information leveled around of the constant *a*. It also adds that, in relation to hydrometric network designs, it is not necessary to quantify in a very precise way the information contained in each station, an approximation is enough since the transformation of a series continues in discrete pulses is subject in any way to noises or errors that can be caused by the measuring instruments themselves.

In the present study both techniques were evaluated. In addition, a new method is proposed to obtain this parameter. Their choice was guided by obtaining the value in which the maximum difference between the Joint Entropy of the series and its respective Total Correlation was observed. The motivation for the formulation of this third methodology comes from the relative subjectivity presented by the previous methodologies, where in the first applications carried out, the difference in the results after the application by different analysts was perceived to be relevant.

In this method it is proposed to make the most objective and effective choice of parameter a as follows: (i) the values of all series of data are first computed together; (ii) a range of integer data around this median value is created, approximately 50% for less and more; (iii) the Entropy of each of the individual series, the Joint Entropy (EC) and the Total Correlation (CT), is calculated for each of these values of a of the stipulated interval; (iv) a new series of data obtained by the subtraction of EC by the CT, (EC-CT) is generated; (v) the smallest value of a corresponding to the highest value of the EC-CT series is verified analytically and graphically, thus obtaining the optimal value for this parameter, which leads to a greater amount of information and less redundancy.

Case study

The chosen area is the Rio das Velhas basin, located in the central region of the State of Minas Gerais, Brazil. The Rio das Velhas is the largest tributary of the São Francisco River Basin, approximately 761 km long, with Barra do Guaicuí in the municipality of Várzea da Palma as an end. Its basin has an area of 29,173 km^2, covering 51 municipalities of Minas Gerais, with a population of close to 5 million inhabitants, according to the 2010 Brazilian Institute of Geography and Statistics (IBGE).

Throughout the Rio das Velhas basin 61 monitors were identified, between active and inactive, according to data from the Agência Nacional de Águas (ANA), available through the Hidroweb web portal.

For the application of the models, a selection of these monitors was made, more propitious to use, since the objective here refers to the application and validation of the proposed methodology, and not specifically a work of designing a new network for this basin. Thus, the basin chosen becomes very propitious because, besides being a basin with monitoring, there are no structures such as reservoirs or large hydroelectric plants, which may cause some interference in the variance of the flow series.

Thus, first were eliminated monitors having a time series below 5 years. The application of this criterion was sufficient to reduce the number of candidates monitors to only 16.

In a second stage the information produced in these monitors was analyzed. In addition to an analysis of the consistency regarding the temporal variability of the data series, we evaluated the flows present in the historical series. Monitors with long periods no record were discarded. In the end they were selected 9 monitors currently active for the development of the study, which had records of 20 years in common periods, using the interval between years 1994 and 2014.

Considering that in the present work these techniques will be evaluated in hydrological conditions significantly different from those analyzed in related studies, it was chosen to use, as Alfonso et al. (2010a) and Li et al. (2012), series of monthly average

flows. The series of flows used have a much higher seasonal variability than those observed in previous studies. Therefore, because it is a first application in basins with these characteristics, it was decided to restrict the evaluation of the behavior of these methods to the conditions described in the literature.

The results obtained for the proposed methodology for evaluation of fluviometric networks are detailed below.

RESULTS

Figure 1 shows the monitors selected for the application of the models.

Information related to the basic statistics of the monitors are shown in Table 1. It should be noted the data on the variance. In general, it is noticed that the variance of the monitors is increasing to downstream. Following the same finding found by Li et al. (2012), but according to these authors, even though an

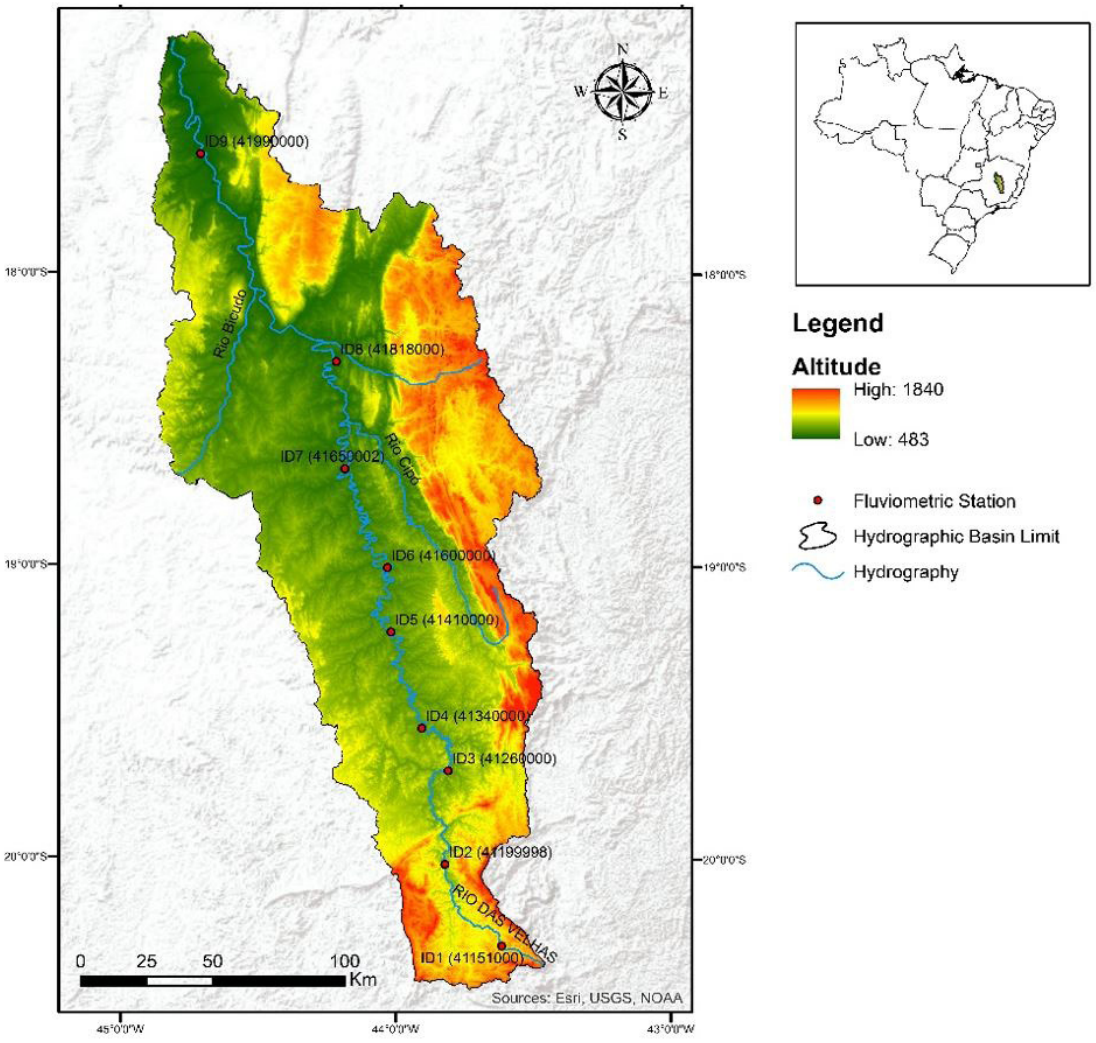

Figure 1. Rio das Velhas location map and location of selected fluviometric stations.

Table 1. Monitors data used in the study.

ID	Monitor	Area (km²)	Variance*	Mean*	Max*	Min*
1	41151000	175	3	3	13	1
2	41199998	1,550	315	31	114	10
3	41260000	3,730	1,719	57	295	7
4	41340000	4,860	2,492	70	409	20
5	41410000	7,080	5,472	94	461	21
6	41600000	8,050	5,013	99	388	22
7	41650002	10,700	8,651	120	600	25
8	41818000	16,600	25,738	183	904	31
9	41990000	26,500	50,782	252	1,165	37

*Streamflow is measured in m^3/s.

increasing variance toward the downstream, and consequently a greater amount of information contained in these monitors, one should not prematurely select the hydrometric network monitors only by your Entropy, this does not would ensure less redundancy between the selected monitors.

Discretization of flow series

The continuous series data of the monitors were discretized using Equation 14. Different values of *a* were used in order to evaluate the proposed methodologies.

The first value used (a1), according to (ALFONSO et al., 2010a), was obtained by analyzing the flow series of the most upstream station, aiming to obtain a minimum value that represents a level that, any change in the other series, higher to this value, can be measured by the Entropy. According to Table 1, the flow values for this station vary from 1.00 m^3/s to 13 m^3/s. Thus, as it will be better justified in the topic on the application of the methods, the value of 11 m^3/s was selected. According to these authors, the flows of the tributaries were also evaluated, as a way of verifying the magnitude of the difference between this minimum and these flows. However, since this flow value refers to the values measured at the station upstream of the basin (41151000), still upstream of tributaries, it was considered more appropriate, in view of the premises stipulated by this methodology, to maintain it as a value more representative.

The second value (*a2*), was obtained according to the methodology proposed by (LI et al., 2012), by the analysis of the relation of this parameter as a function of the variation of the entropies of the monitors and the respective standard deviations. As shown in Figure 2, the value of *a2* was varied until it reached an optimum range, around 210 m^3/s, not small enough to make it difficult to analyze the monitors by the similar entropy values generated, and not too large, that would nullify the entropy of the monitors closest to the head of the basin. For the formulation of Figure 2, the individual entropies of each of the monitors were also calculated.

Note that to reach this value by this methodology, it was not possible that it was small enough to be within the minimum and maximum of the station upstream in the basin.

As can see in Figure 2, the entropy value of the monitors increases from upstream to downstream, confirming previous statements of similarity with the variance of the data series.

The last applied methodology refers to the proposed here, by the joint analysis of Joint Entropy and Total Correlation. Figure 3 shows the graph of the series (JH-CT) along the variation of a

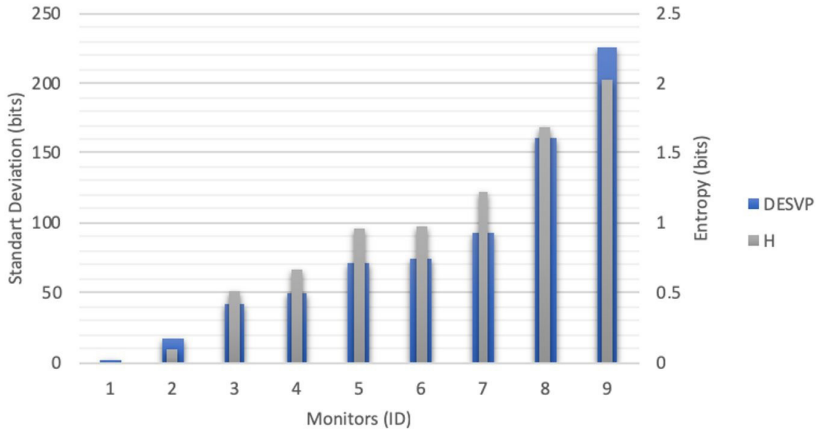

Figure 2. Individual entropy (H) after discretization and standard deviation of the flows of each station.

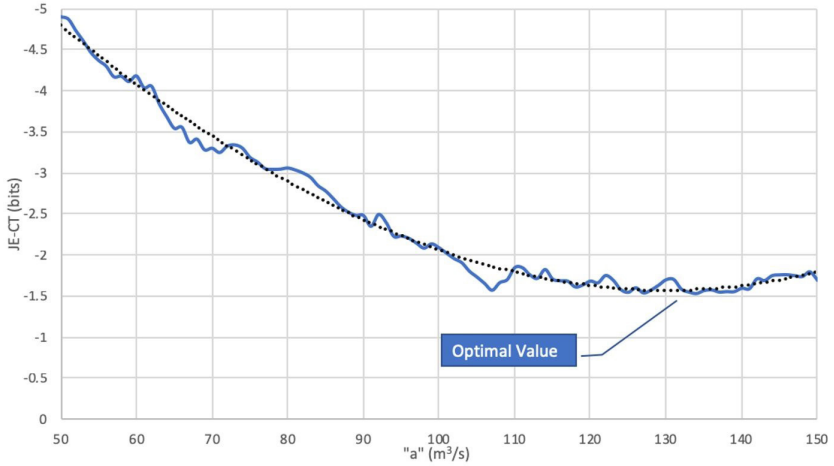

Figure 3. Comparison between JH and CT as a function of "*a*".

within the suggested range, around the mean value of the series of mean flows studied. The optimum value for this parameter (*a3*) can be clearly seen by the inflection region of the curve. Values lower than that designated as the optimal value oscillate, but in general tend to generate high JH but also high CT values. Conversely, higher values generate small values of CT, but low values for JH. For this study the value obtained for (a3) was 134 m^3/s.

Identification of the necessary number of fluviometric monitors

The first step for analysis of hydrometric network optimization refers to establishing the minimum density of monitors for particular river basin. Within this context, the ranges proposed by (WMO, 2008) are used as reference. The first step was to identify, according to national studies, the geographic characteristics of the Rio das Velhas basin.

According to the studies and reports made by the Geological Survey of Brazil (CPRM, 2001), there is a predominance of plateau areas in the Rio das Velhas basin. The classification of the physiographic units proposed by (WMO, 2008) cover six different types, these are: mountains, interior plains, hilly/undulating, coastal, small islands, and polar/ arid. Among these six, it can be said that, for the Rio das Velhas basin, a predominance of two, plains and hilly/undulating.

For these two types, according to this same organization, an average density of 1,875 km^2 of monitors is indicated. Thus, only for the sub basin of the Rio das Velhas, excluding the sub-basins of its tributaries, which covers about 20,000 km^2, we reach an approximate 12 monitors to meet this minimum. The areas used to calculate this minimum number of monitors were measured by GIS tools and are shown in Table 2.

This value will be compared with the results obtained by the methods that use the Information Theory concept to identify the same necessary minimum. The objective is to be able to verify if the estimated value at least by the WMO, agrees with the analysis of the methods in identifying the maximum information and points of redundancy in the network. In this way, it is possible to quantify if the existing set is generating redundant information, even for a lower number of stations stipulated by that organization, thus indicating if this network should be optimized and expanded.

Rank of monitors and informative values

As proposed initially, two methods were used that are presented today in the literature as more distinct for design and optimization of hydrometric networks. Both methodologies are somewhat similar, have as basic objective the maximization of information and reduction of total correlation. First, the performance between the methods was obtained by obtaining the Joint Entropy and Total Correlation, as a new station is included in the network.

For the MIMR method the results were obtained using both coefficients $\lambda_1 e \lambda_2$ of Equation 13, the value of 0.80 and 0.20, respectively. This value was obtained by the same methodology proposed by (LI et al., 2012).

In a first analysis of these results, some characteristics were expected, such as the ranking always starting at the ID 9 station, located at the point of greatest entropy, downstream of the basin. This result refers to the basic configuration of the algorithm, derived from the methodology initially proposed by Krstanovic and Singh (1992a, 1992b), which starts the ranking by a central station, defined by the one with the highest entropy of the set.

A second and important point observed in the ranking is that the sequencing of the ranks did not prevail by the increasing value of the marginal entropy of the monitors or by their variances. This shows the existence of redundant information in the system, as observed by the calculation of the Total Correlation superior to zero. This shows that assigning the monitors of a network simply by their maximum entropy values will not lead to the best set and can generate more redundant information.

For the value of a = 11 m^3/s, the WMP method indicated the rank of the monitors following the sequence: [9 3 4 2 8 6 5 7 1]. Due to this ranking, Figure 4 shows an efficiency in indicating the monitors that best contribute to the network, a fact verified by the largest increase in the EC curve for the first 5 monitors, which reached the value of 7.60 bits for this variable, the maximum being 7.89 bits, or 96%.

For the MIMR method, the indicated station rank was: [9 4 1 2 3 5 6 7 8]. Regarding EC, this method indicated a ranking with less accentuated growth, however, this less accentuated growth is also perceived for the CT curve. For the same first five monitors the CT, by the rank indicated by the MIMR method was 7.61 bits, against 11.80 bits of the WMP method.

In relation to the value of a = 134 m^3/s, presented in Figure 5, no difference was detected for the ranking proposed by the WMP method in comparison to the previous value of 11 m^3/s. For the MIMR method there was a change in the ranking, changing to: [9 7 1 2 4 3 6 5 8]. A decrease in effectiveness in the proposed initial set was observed, reaching for EC, with the first 5 monitors, 74% of the maximum, compared to 93% previously. This result also worsened for the TC, which for the same set represents 47% of network redundancy, compared to 31%.

Finally, for the value of a = 210 m^3/s, shown in Figure 6, the MIMR method maintained the same previous ranking. For the WMP method the proposed classification was: [9 6 4 3 8 5 2 7 1]. In relation to the two other values of a this was the least effective result for the WMP method. There is less significant growth for CHD and a higher TC increase at the first proposed monitors. The same 5 first monitors for this value reached 87% of the total EC.

Table 2. Monitors by km^2.

WMO Classification	WMO (km^2/monitor)	Area (km^2)	Monitor
Montains	1,000	2,997	3
Plains	1,875	4,065	2
hilly/undulating	1,875	12,947	7

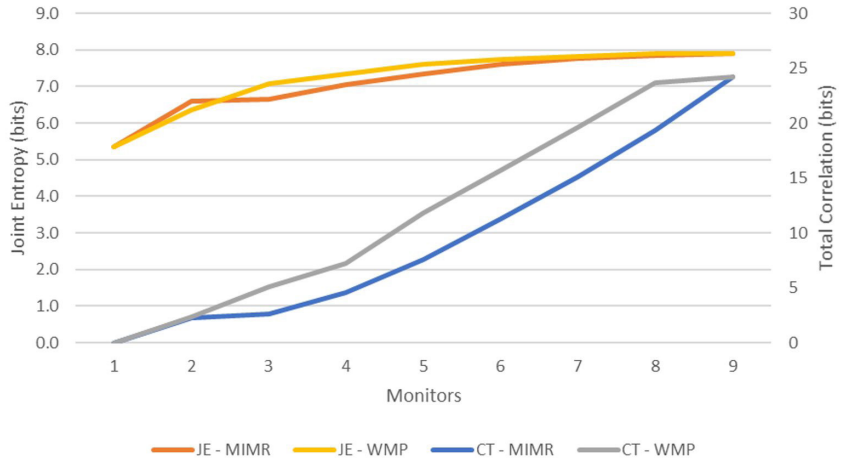

Figure 4. Progress of informative values as new monitors are added, a = 11 m^3/s.

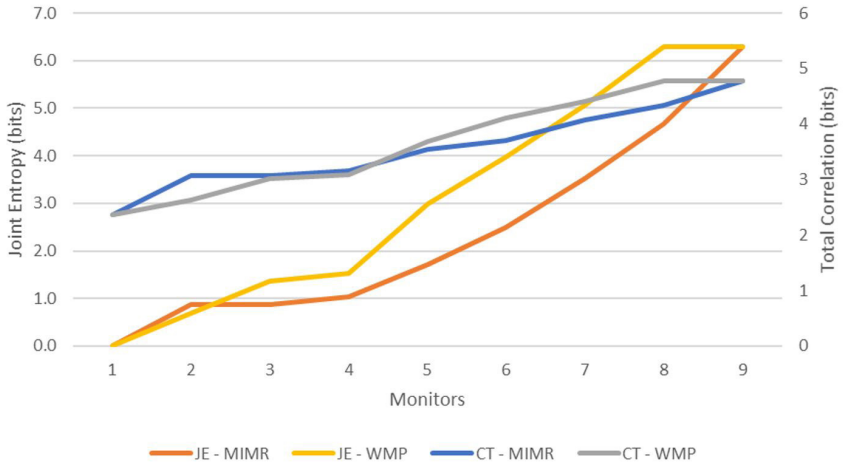

Figure 5. Progress of informative values as new monitors are added, a = 134 m^3/s.

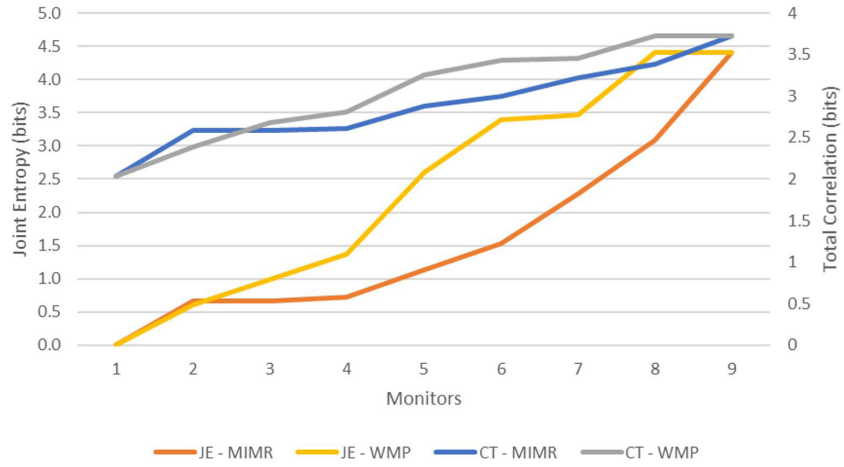

Figure 6. Progress of informative values as new monitors are added, a = 210 m^3/s.

Comparison between MIMR and WMP methods

In relation to the main objective of maximization of the Joint Entropy, it is generally found that it increases as a new station is included in the network, even in a smaller amount for some steps, due to the redundancy and or little contribution of the station at that point. The reason that happens is related to the low number of monitors to a bowl as big as the Rio das Velhas, since, as minimum values for the recommendation d and WMO, about 12 monitors would be needed only in the basin of the main river.

If there was a saturation of the information values, due to the stagnation of the increase of the EC and increase of the CT, even with the addition of a new station, it could be affirmed that there is no need for more monitors for the basin.

What is perceived by the analysis of the existing network is that the Joint Entropy has an insignificant increase after 5 monitors added to the network, but maintains an increase of information, and in turn also generates a significant increase of Total Correlation. This shows that the network actually lacks optimization and scaling.

That a significant increase is related to that observed by Li et al. (2012), thus indicating that the simple addition of new monitors, can in certain cases little - interferon go on increasing total system information once again justified by the information generated redundancy.

Regarding the ranking of monitors, a distinction is made between the results of the WMP and MIMR methodology. The ranking differs as belonging to monitors including the optimal set, represented by the first five monitors, which account for 80% or more of the information generated throughout the network.

The ranking of these monitors by the WMP method was superior in all cases to the MIMR method. The proportion of the information generated by this optimal set against the total for the WMP and MIMR method in the three cases was (96%-93%), (87%-77%), (77%-74%) respectively, thus confirming the ability of the WMP method to indicate the best optimized set for the basin.

Another important point observed is Total Correlation. The optimum set of the five monitors obtained by the MIMR methodology cover about 30% of the information duplicated by the system in all situations, whereas by WMP this rate reaches about 50%. This shows the greater weight imposed by the MIMR method in a search for a less redundant optimal set.

Even a less interesting performance for the WMP method in relation to CT, we can see a good result for both, since only five monitors, representing 55% of the total monitors, were able to indicate a set that represents 80% of the total generated information.

These observations show an important finding that the MIMR method works on a prioritized focus of reducing redundancy, while the WMP prioritizes the significant increase of information as a primary goal.

Still on the ranking of the monitors, it is noticed that only the WMP method was able to identify a difficulty found in the discretization step of the data. As can be seen from Figures 4, 5 and 6, station ID 01, for the values of *a* used, does not aggregate information to the system, perceived by the zero increment after its insertion. This is because for these values of *a*, compared to the low values of flow in all other monitors, the Equation 1 for this station sums up to a single value, that's means, zero entropy.

The MIMR method, because it is a more analytical method, suggests the implementation of station ID 01, since even though it does not aggregate information, the station also does not generate redundancy, which causes this method to rank it better than the subsequent station, ID 02. The latter aggregates information to the network, but increases, albeit to a lesser extent, the redundancy, thus making it less "attractive" to the method during the ranking. This situation is apparently perceived by the WMP method, which in all cases ranked station ID 01 among the last three. This is certainly the most serious point in the points in disfavor to the MIMR method. By showing that in cases of seasons with large discrepancies between the data, common in Brazilian basins, this method may make a wrong choice of seasons. On the contrary, it counts as a merit to the WMP method, which by its more conceptual nature, based on the identification of independent monitors by the Entropy.

Regarding the variation of the results in relation to the variation of the value of a, it is possible to say that, at the beginning it is possible by the analysis of the JH and TC, to say with greater certainty which of the three methodologies is the most adequate. Method 1 was the only one to enable all monitors to aggregate information and make the analysis complete for all monitors. However, the JH value after step 5 undergoes minor variations. The use of a value of *a* so low may make it difficult to meet the premise of generating distinguishable values of Entropy for all monitors when variations between the magnitude between them are very high.

In this way, it can be concluded that, for these cases, with a great discrepancy of data, it is possible to indicate the discretization of the series focusing on an analysis directed to median values of *a*, allowing the obtainment of more distinct Entropy values for the monitors, even if it penalizes the monitors upstream of the basin.

Spatial location of fluviometric monitors

As a way of solving some questions still pending about the methods, when only evaluated by their capacities of indicating the monitors that add more information to the network, and at the same time reduce the redundancy in the same, a second stage of evaluation was executed. This step is aimed at analyzing their capacity to better spatially locate monitors in a fluviometric network.

This analysis was done by the spatial distribution of the monitors pointed out by the best ranked methods, which promote approximately 80% of the network information. Figures 7, 8 and 9 were assembled for this purpose. The locations of this main set of monitors pointed out by the methods are shown, according to the variation of the value of parameter *a* for discretization.

As can be seen in Figures 7 to 9, by this analysis, it is possible to perceive the best results when using the WMP method to analyze the networks. It is noticed that the MIMR method generated as an optimal set a more concentrated set in the upstream and central portion of the basin.

Also, on the location of these monitors it is noticed that only the WMP method indicated as an optimal set, beyond the central region of the basin, a station after the major confluences in the Rio das Velhas, Cipó and Bicudo rivers, maintaining for these points the respective ID 09 and ID 08.

Regarding the leasing capacity versus the variation of *a*, it can be seen that the methodologies 2 and 3 for discretization of the data obtained better results, more accentuated for the MIMR method.

This is because, as shown, the fluviometric monitors when thrown over the basin, by methodology 1, we notice a higher concentration of these monitors in the upstream portion of the basin. Differently from that found for the other methodologies,

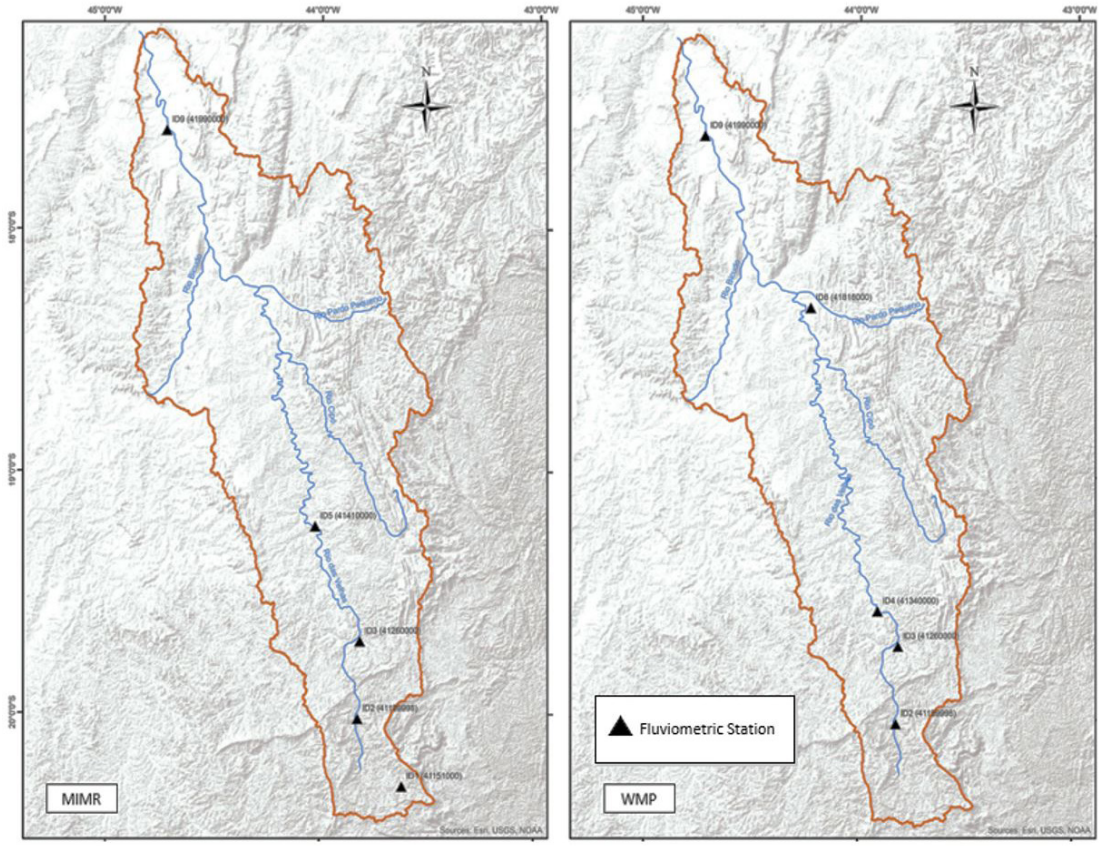

Figure 7. Spatial distribution of the main set of fluviometric monitors ranked by MIMR and WMPs for a = 11 m^3/s.

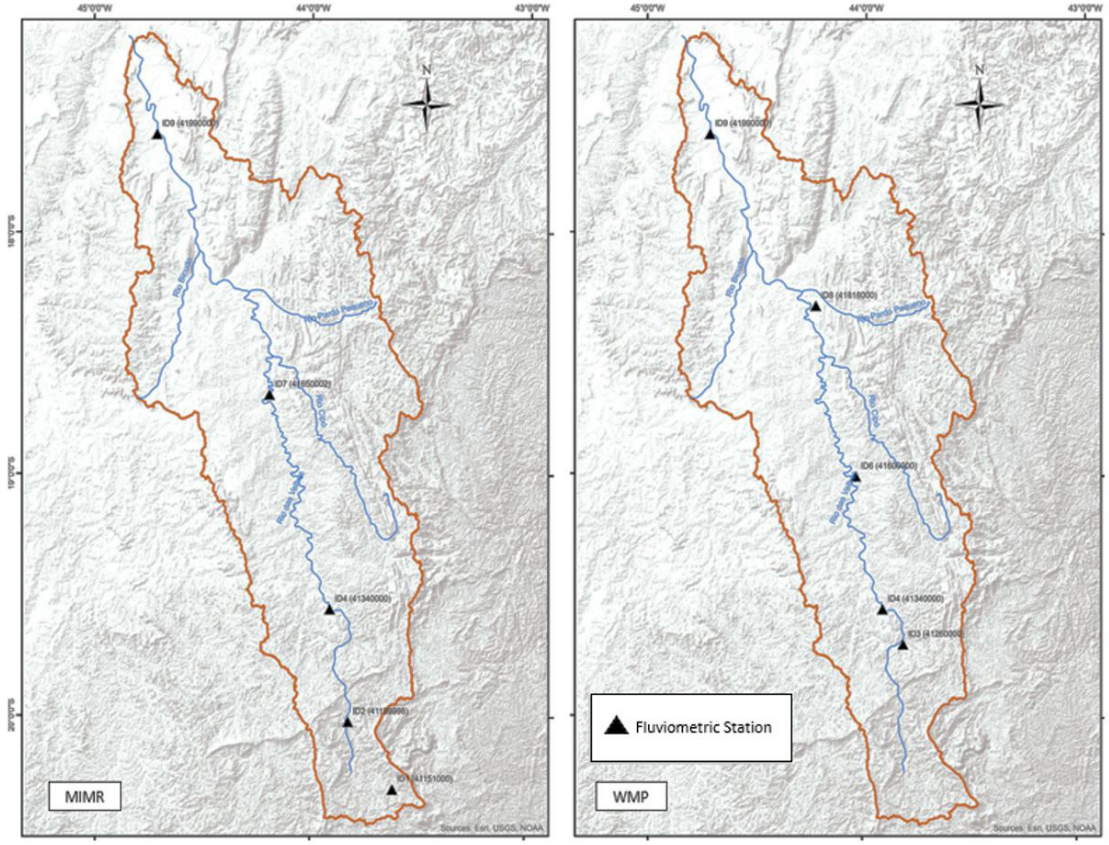

Figure 8. Spatial distribution of the main set of fluviometric monitors ranked by MIMR and WMPs for a = 210 m^3/s.

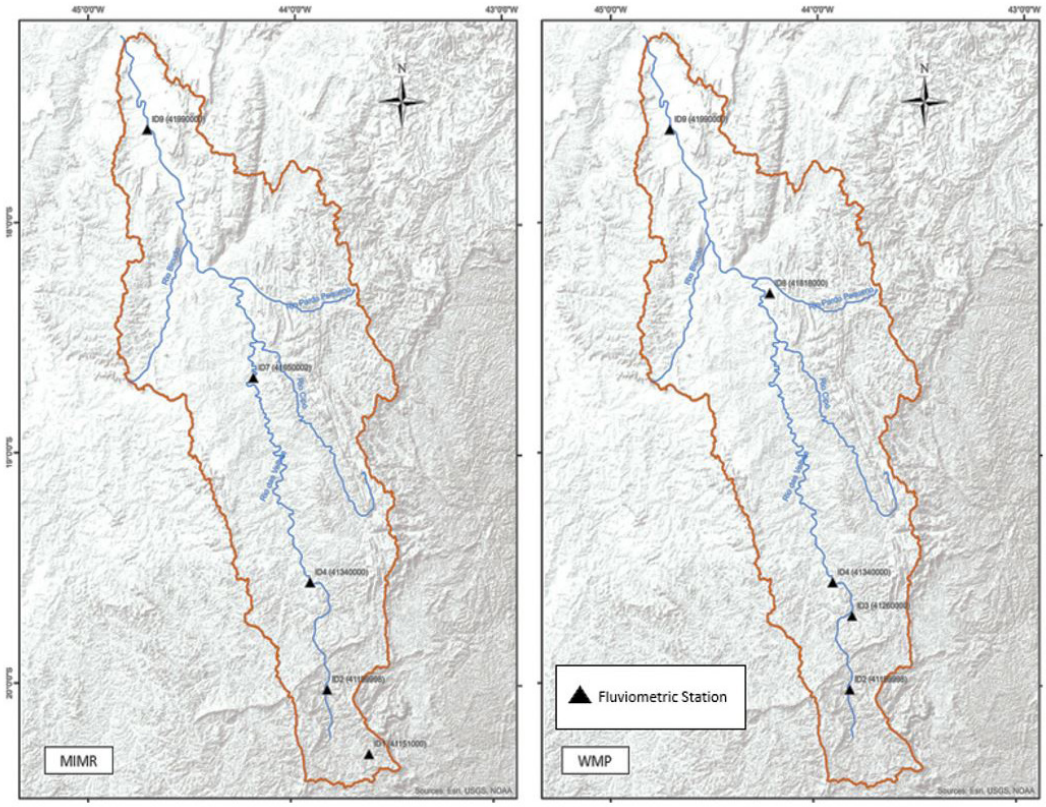

Figure 9. Spatial distribution of the main set of fluviometric monitors ranked by MIMR and WMPs for a = 134 m^3/s.

which achieved a more homogeneous distribution of the monitors along the Rio das Velhas. Thus, as noted in the previous item, the use of a median value for discretization of the data, as done for methodologies 2 and 3, better results are obtained when evaluating monitors with discrepant flow data.

CONCLUSIONS

The results found corroborate in large part those found by Alfonso et al. (2010a) and Li et al. (2012), showing the adequacy of these methodologies, now analyzed in hydrographic basins in tropical Brazil. However, a result different from that pointed out by Li et al. (2012), this shows that the method MIMR as superior to WMP. When analyzed s a s capacity s geographical location of the monitors, and adherence to data containing monitors with differing magnitudes of flows, it was realized that only the WMP method was able to adapt to this reality by its concept based on an analysis by monitors and not only by analytical mode by the analysis of EC and CT. It may be concluded here in this study that this is the most appropriate method within this context.

A new methodology for discretization of continuous series data was proposed. This methodology proved to be equivalent in the verification of the amount of information and redundancy of the system, and superior to obtain the best specialized set along the river channel.

It was verified that the set formed by 9 fluviometric monitors is undersized for the watershed of the Rio das Velhas. As recommended by WMO (2008), this number should be of the order of 12 monitors. In addition, even with a lower number of monitors, the analysis methods based on Information Theory indicated the monitors that contribute most in the amount of information to the network, and consequently those that are not adding information due to the redundancy of the information generated.

It should be noted that until this point the expansion of the fluviometric network is beyond the capacity of the evaluated methods. In the future this point can be solved by simulation. Hydrodynamic models can be generated to simulate a dense synthetic network of monitors. However, this task may still generate a great deal of uncertainty, given the low amount of information available or the empirical nature of such models.

It is important to highlight the great contribution that such methodologies add to this line of research focused on the study of hydrometric networks, a problem faced in all countries that seek to optimize their networks and reduce costs.

ACKNOWLEDGEMENTS

Thanks to the Coordination of Improvement of Higher Education Personnel (CAPES) and Federal University of Minas Gerais (UFMG) for the economic support provided.

REFERENCES

ALFONSO, L.; HE, L.; LOBBRECHT, A.; PRICE, R. Information theory applied to evaluate the discharge monitoring network of the Magdalena River. *Journal of Hydroinformatics*, v. 15, n. 1, p. 211-228, 2013. http://dx.doi.org/10.2166/hydro.2012.066.

ALFONSO, L.; LOBBRECHT, A.; PRICE, R. Information theory–based approach for location of monitoring water level gauges in polders. *Water Resources Research*, v. 46, n. 3, p. W03528, 2010a. http://dx.doi.org/10.1029/2009WR008101.

ALFONSO, L.; LOBBRECHT, A.; PRICE, R. Optimization of water level monitoring network in polder systems using information theory. *Water Resources Research*, v. 46, n. 12, p. W12553, 2010b. http://dx.doi.org/10.1029/2009WR008953.

ALFONSO, L.; RIDOLFI, E.; GAYTAN-AGUILAR, S.; NAPOLITANO, F.; RUSSO, F. Ensemble entropy for monitoring network design. *Entropy*, v. 16, n. 3, p. 1365-1375, 2014. http://dx.doi.org/10.3390/e16031365.

COVER, T. M.; THOMAS, J. A. *Information theory*. New York: John Wiley, 2012.

CPRM – COMPANHIA DE PESQUISA DE RECURSOS MINERAIS. *Regionalização de vazões sub-bacias 40 e 41 – caracterização física e análise dos dados básicos*. Belo Horizonte: DNAEE/CPRM, 2001. (Relatório Final Volume I).

FASS, D. M. *Human sensitivity to mutual information*. 2006. Ph.D. (Dissertation) - Rutgers University The State University of New Jersey, New Brunswick, 2006.

GONTIJO JUNIOR, W. C.; KOIDE, S. Avaliação de redes de monitoramento fluviométrico utilizando o conceito de entropia. *Revista Brasileira de Recursos Hídricos*, v. 17, n. 1, p. 97-109, 2012. http://dx.doi.org/10.21168/rbrh.v17n1.p97-109.

JAKULIN, A.; BRATKO, I. Testing the significance of attribute interactions. In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING, 21., 2004 July 4-8, Banff, Alberta. *Proceedings*… New York: Association for Computing Machinery (ACM), 2004. v. 69, p. 52.

KEUM, J.; KORNELSEN, K. C.; LEACH, J. M.; COULIBALY, P. Entropy applications to water monitoring network design: a review. *Entropy*, v. 19, n. 11, p. 613, 2017. http://dx.doi.org/10.3390/e19110613.

KRSTANOVIC, P. F.; SINGH, V. P. Evaluation of rainfall networks using entropy: II. applications. *Water Resources Management*, v. 6, n. 4, p. 295-314, 1992a. http://dx.doi.org/10.1007/BF00872282.

KRSTANOVIC, P. F.; SINGH, V. P. Evaluation of rainfall networks using entropy: I. theoretical development. *Water Resources Management*, v. 6, n. 4, p. 279-293, 1992b. http://dx.doi.org/10.1007/BF00872281.

LI, C.; SINGH, V. P.; MISHRA, A. K. Entropy theory-based criterion for hydrometric network evaluation and design: maximum information minimum redundancy. *Water Resources Research*, v. 48, n. 5, p. 1-11, 2012. http://dx.doi.org/10.1029/2011WR011251.

MCGILL, W. J. Multivariate information transmission. *Psychometrika*, v. 19, n. 2, p. 97-116, 1954. http://dx.doi.org/10.1007/BF02289159.

MISHRA, A. K.; COULIBALY, E. P. Developments in hydrometric network design: a review. *Reviews of Geophysics*, v. 47, n. 2, p. RG2001, 2009. http://dx.doi.org/10.1029/2007RG000243.

MISHRA, A. K.; COULIBALY, P. Hydrometric network evaluation for canadian watersheds. *Journal of Hydrology*, v. 380, n. 3-4, p. 420-437, 2010. http://dx.doi.org/10.1016/j.jhydrol.2009.11.015.

MISHRA, A. K.; COULIBALY, P. Variability in Canadian seasonal streamflow information and its implication for hydrometric network design. *Journal of Hydrologic Engineering*, v. 19, n. 8, p. 1-11, 2014. http://dx.doi.org/10.1061/(ASCE)HE.1943-5584.0000971.

PÁDUA, L. H. R.; NAGHETTINI, M. C.; SILVA, F. E. O. E. Análise de redes hidrométricas com emprego de entropia. In: CONGRESSO DA ÁGUA, 14., 2018 Março 7-9, Évora, Portugal. *Anais*... Lisboa: Associação Portuguesa dos Recursos Hídricos – APRH, 2018a.

PÁDUA, L. H. R.; NAGHETTINI, M. C.; SILVA, F. E. O. E. Análise de redes fluviométricas com emprego de entropia. In: XIV SIMPÓSIO DE RECURSOS HÍDRICOS DO NORDESTE, 14., 2018 novembro 20-24, Maceió, Alagoas. *Anais*... Porto Alegre: Associação Brasileira de Recursos Hídricos – ABRH, 2018b. p. 1-10.

PYRCE, R. S. *Review and analysis of stream gauging networks for the Ontario stream gauge rehabilitation project – WSC Rep. 01-2004*. Peterborough: Institute for Watershed Science, Trent University, 2004.

RUDDELL, B. L.; KUMAR, P. Ecohydrologic process networks: 1. identification. *Water Resources Research*, v. 45, n. 3, p. W03419, 2009. http://dx.doi.org/10.1029/2008WR007279.

SAMUEL, J.; COULIBALY, P.; KOLLAT, J. CRDEMO: Combined regionalization and dual entropy-multiobjective optimization for hydrometric network design. *Water Resources Research*, v. 49, n. 12, p. 8070-8089, 2013. http://dx.doi.org/10.1002/2013WR014058.

SHANNON, C. E. A mathematical theory of communication. *The Bell System Technical Journal*, v. 27, n. 3, p. 379-423, 1948. http://dx.doi.org/10.1002/j.1538-7305.1948.tb01338.x.

SHANNON, C. E.; WEAVER, W. *The matematical theory of communication*. Champaign: University of Illinois Press, 1949.

SHIMAZAKI, H.; SHINOMOTO, S. A method for selecting the bin size of a time histogram. *Neural Computation*, v. 19, n. 6, p. 1503-1527, 2007. http://dx.doi.org/10.1162/neco.2007.19.6.1503. PMid:17444758.

SINGH, V. P. The use of entropy in hydrology and water resources. *Hydrological Processes*, v. 11, n. 6, p. 587-626, 1997. http://dx.doi.org/10.1002/(SICI)1099-1085(199705)11:6<587::AID-HYP479>3.0.CO;2-P.

WERSTUCK, C.; COULIBALY, P. Hydrometric network design using dual entropy multi-objective optimization in the Ottawa River Basin. *Hydrology Research*, v. 48, p. 1-13, 2016.

WERSTUCK, C.; COULIBALY, P. Assessing spatial scale effects on hydrometric network design using entropy and multi-objective methods. *Journal of the American Water Resources Association*, v. 54, n. 1, p. 275-286, 2017. http://dx.doi.org/10.1111/1752-1688.12611.

WMO – WORLD METEOROLOGICAL ORGANIZATION. *Guide to hydrological practices*. 16th ed. Geneva: WMO, 2008. (Volume I: Practices hydrology — From measurement to hydrological information, WMO 168).

Authors contributions

Luiz Henrique Resende de Pádua: primary author, manuscript structure, conceptualization, literature review, methods, analysis of the results, writing, original draft preparation.

Nilo de Oliveira Nascimento: study orientation, review, manuscript structure, analysis of the methods and results.

Francisco Eustáquio Oliveira e Silva: study orientation, review, manuscript structure, analysis of the methods and results.

Leonardo Alfonso: study orientation, review, manuscript structure, analysis of the methods and results.