2014

The Potential Of Kohonen Method For Classifying Rivers Focused On Regional Environmental Flow Determination With ELOHA

Federico Gonzalez Cuellar

Nelson Obregon Neira

11th International Conference on Hydroinformatics
HIC 2014, New York City, USA

THE POTENTIAL OF KOHONEN METHOD FOR CLASSIFYING RIVERS FOCUSED ON REGIONAL ENVIRONMENTAL FLOW DETERMINATION WITH ELOHA

FEDERICO GONZÁLEZ CUÉLLAR (1), NELSON OBREGÓN NEIRA (2)
(1): Master of Hydro-Systems, Pontifical Javeriana University, Career 7 # 42-27 7th floor, Bogotá, Colombia
(2): Geophysical Institute, Pontifical Javeriana University, Career 7 # 42-27 7th floor, Bogotá, Colombia

This article highlights the potential benefits that the Kohonen method has for the classification of rivers with similar characteristics by determining regional ecological flows using the ELOHA (Ecological Limits of Hydrologic Alteration) methodology. Currently, there are many methodologies for the classification of rivers, however none of them include the characteristics found in the Kohonen method such as (i) providing the number of groups that actually underlie the information presented, (ii) used to make variable importance analysis, (iii) which in any case can display two-dimensional classification process, and (iv) that regardless of the parameters used in the model, the clustering structure remains. In order to evaluate the potential benefits of the Kohonen method, 174 flow stations distributed along the great river basin "Magdalena-Cauca" (Colombia) were analyzed. 73 variables were obtained for the classification process in each case. Six trials were done using different combinations of variables and the results were validated against reference classification obtained by Ingfocol in 2010, whose results were also framed using ELOHA guidelines. In the process of validation it was found that two of the tested models reproduced a higher level than 80% of the reference classification with the first trial, meaning that more than 80% of the flow stations analyzed in both models formed invariant groups of streams.

1. ELOHA METHODOLOGY IN THE MAGDALENA-CAUCA BASIN

This article presents characteristics of the Kohonen method that can be classified as virtues in the classification of rivers, focusing on the determination of regional ecological flows, compared to other methods of classification.

From the eco-hydrologic point of view, the methodology which highlights the features of Kohonen is the methodology called Ecological Limits of Hydrologic Alteration (ELOHA) [1]. The most important feature of ELOHA is that it corresponds to a regional methodology because it encompasses large areas [2]. The classification of the rivers is the condition which gives it the regional character, since instead of assigning ecological flows for each particular river, it establishes eco-hydrological recommendations by type of river, for which it is necessary to gather together the rivers in families that maintain similar characteristics. In this regard, it is precisely in the process of classification of rivers where Kohonen takes relevance.

The case of the analyzed study corresponds to the Magdalena-Cauca basin, in Colombia. This basin is perhaps the most important basin of this country, hosting in its interior more than half of the population (approximately 77% of the country's population) and much of its economy. The Magdalena-Cauca basin has an approximate area of 273,000 km^2 [3] and presents a great variety of climates, which makes the eco-hydrologic sorting process more difficult.

In 2010, Ingfocol Ltda applied part of the ELOHA methodology in the Magdalena-Cauca basin with the objective of establishing the basis for allocating recommendations of ecological flows by type of river [4]. In this case, 174 flow stations were taken into account and 73 eco-hydrological variables, of which 68 correspond to the parameters supplied by the IHA software (Indicators of Hydrologic Alteration) [5] for each series of flow, 3 to the spatial coordinates of each station and the other two variables correspond to the average flow of each series and the corresponding coefficient of variation. In turn, of the variables supplied by the IHA software, 34 correspond to hydrological parameters of alteration called IHA parameters and the other 34 are related to the components of the ecological flow, which are called EFC parameters.

The classification process carried out by Ingfocol in 2010 was quite complex, since it had to resort to different techniques to obtain a proper classification of rivers. Initially it was necessary to implement principal components analysis, linear correlation and ecological significance of variables, to obtain 26 of the 73 variables for the classification process. With these 26 relevant variables, methods were used for the classification of hierarchical type and non-hierarchical as for example the method of K-means and the cluster with a probability of distribution. Finally, the streams classification process was guided and supplemented manually with the help of experts in the basin and the development of several workshops for this purpose. The Figure 1 shows the summary of the process of classification obtained by Ingfocol in 2010.

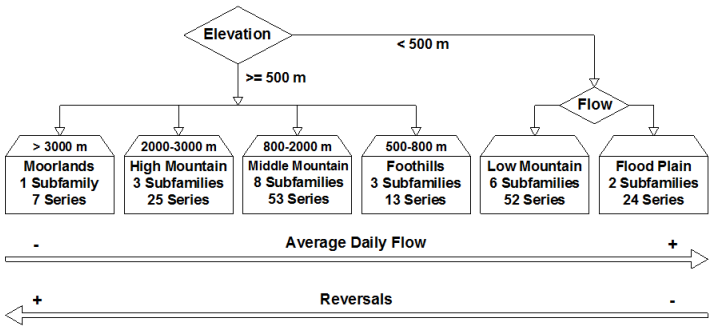

Figure 1. Summary of the classification process carried out by Ingfocol Ltda in 2010.

As can be seen in the figure above, in 2010 Ingfocol obtained a total of 23 groups of rivers, which were taken as a reference for the validation process for the Kohonen method described below.

2. KOHONEN METHOD FOR THE CLASSIFICATION OF INFORMATION

As mentioned above, the Kohonen method was used in this article to classify the eco-hydrologic information, whose most important aspects are described below.

The Kohonen method, also known as Self-Organized Maps (SOM) of Kohonen, is part of the called Artificial Neural Networks and is mainly used for analysis of classification of information. This method has two major advantages over other methods of classification: (i) It

doesn't need the number of groups by default (unsupervised learning) [6] and (ii) the display of the results, regardless of the number of variables, can always be done in two dimensions.

In summary, the learning of the SOM, in addition to the fact that it is not supervised, it is also competitive, that is to say that the neurons of the model compete with each other to see which one can most closely resemble the pattern of training presented, with which the weights of the winning neuron are updated in a greater proportion than the weights of neighboring neurons. The proportion of the update of the neurons belonging to the vicinity of the winning neuron decreases as a function of distance in connection with it. The greater the similarity between two patterns of training, will be smaller the distance between their winning neurons. This gives it the sense of self-organization to the model, since as it is training the network, the winning neurons of similar patterns form independent neighborhoods that finally reflect groups of similar patterns.

Finally, of the multiple maps that can be obtained with this method, the map of distances between neurons is one of interest in this case, which in a flat two-dimensional shows the borders of the groups of neurons that represent groups of rivers, and the map of weights of the input variables, which displays in a color scale the magnitude of the weight of each neuron to a particular variable. In the next section of this article this type of maps are shown.

3. DISCUSSION AND RESULTS

There are several advantages of the Kohonen method that postulate it as a suitable method to classify information, in this case, eco-hydrologic information. Before submitting such benefits, it is important to describe briefly the methodological process that was undertaken to achieve the results that will be provided later.

Based on the same information used by Ingfocol Ltda in 2010 for the classification process and taking into account the results obtained in that case as validation patterns, were raised 6 typologies with different combinations of variables, which are described below:

- Typology 1- all 73 variables.
- Typology 2 - variables IHA (34 variables).
- Typology 3 - variables EFC (34 variables).
- Typology 4 - Ingfocol variables 2010 (26 variables).
- Typology 5 - 12 monthly flows and station elevation (13 variables).
- Typology 6 – The quarter where presents the highest flow, medium flow extremely low, average flow of large floods, medium flow and station elevation (5 variables).

Of the 6 previous typologies, It is important to stress the fact that typologies 5 and 6 were raised in order to establish how effective was the Kohonen method with very few variables, which were chosen trying to represent the average hydrograph of each hydrological station with variables easy to obtain. The Kohonen method was applied for each of the previous typologies of variables and their results were compared with the results obtained by Ingfocol in 2010. At this stage of the process, models with a different number of neurons were tested until optimal results were obtained to identify the number of groups in each case, activity that was performed manually.

In the validation process, taking as reference the 23 groups obtained by Ingfocol Ltda in 2010, the groups obtained by Kohonen for each typology of variables were identified, and only the typologies that presented a number of groups similar to the baseline continued the process of analysis, i.e. close to 23. In this case, only the typologies 5 and 6 yielded a number of groups similar to Ingfocol with 20 and 25 groups respectively. The other 4 typologies obtained a much higher number than 23, and were therefore discarded. Figure 2 shows the map of distances obtained for Typology 5, where the groups obtained are also shown.

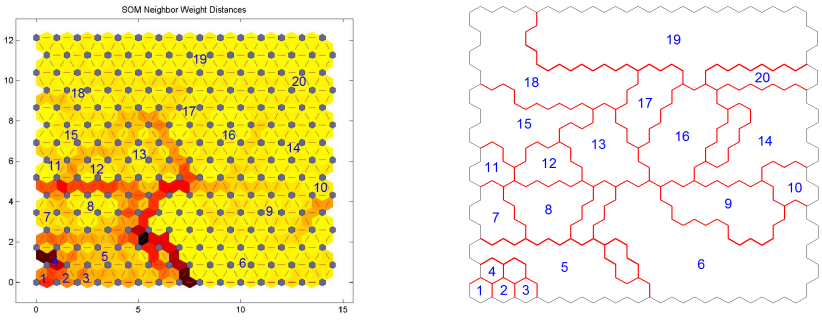

Figure 2. River Classification of the model of 225 neurons for the typology 5.

Finally, with typologies 5 and 6 the analysis of importance of variables was carried out within the process of classification and the analysis of invariant rivers groups in comparison with the classification obtained by Ingfocol in 2010. In this context, the following are the results and the benefits obtained by the Kohonen method to classify eco-hydrologic information.

In terms of results, as shown in Figure 3, with the Kohonen method doesn't matter the number of neurons chosen for the modeling, the classification structure is preserved, i.e. that is, the shape of the maps is similar.

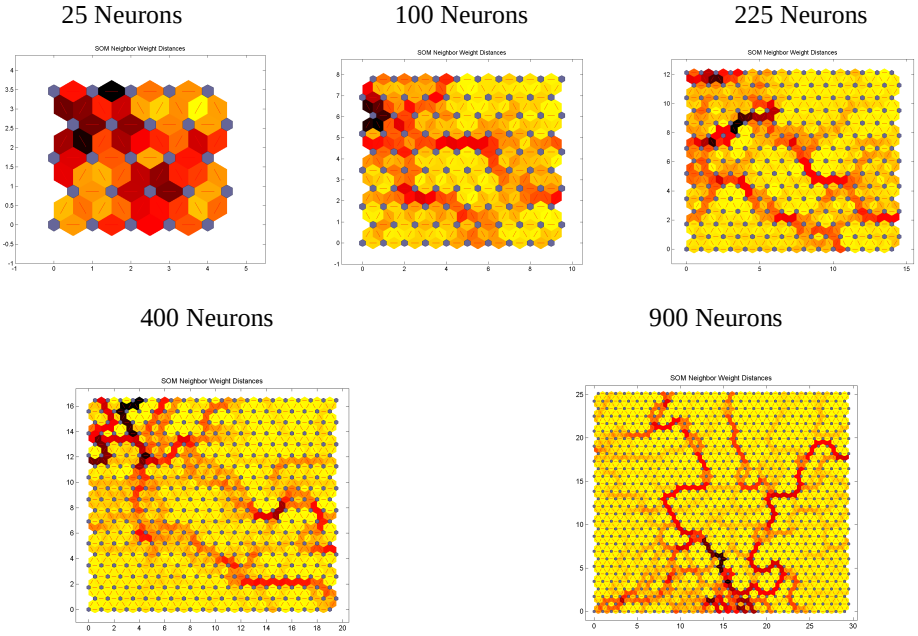

Figure 3. Map of distances for models with different numbers of neurons obtained for typology 6.

In this case, the fact that the Kohonen method does not require the number of groups by default is also highlighted, unlike other methods of classification, and in this sense, the number of neurons that is assigned to the model depends on the resolution to which the groups wish to be identified; among more neurons has the model, with higher resolution you will see the

classification of information. Under this perspective, the Kohonen method provides the number of groups that in fact underlies the information presented.

On the other hand, in order to establish how effective is the Kohonen method in the classification of eco-hydrologic information, taking into account the complexity of the classification process carried out by Ingfocol in 2010, a procedure to compare each families of rivers obtained by Kohonen for the typologies 5 and 6 with the families of rivers of reference (23 groups of rivers obtained by Ingfocol in 2010) was undertaken, to subsequently identify how many rivers remained in the same group. This analysis is called analysis of invariance. Table 1 shows the results of this analysis for the Typology 5.

Table 1. Array of invariance rivers of Typology 5 with regard to the classification obtained by Ingfocol in 2010.

Clusters Ingfocol 2010	Clusters T5																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	# patr.
1						7															7
2						5															5
3									2												2
4						16			2												18
5														5		2	2		5	1	15
6														3					2		5
7																			4	3	7
8										1									3		4
9										1				2		1			1		5
10														2					2		4
11									1	1				1					5		8
12														1			1		1	2	5
13																			6		6
14																			4		4
15																			3		3
16																		17	9		26
17																		3			3
18																		6	2		8
19																		1	5		6
20																		2			2
21					1			2					3		1						7
22					2		1	1			1	2			1			1			9
23	2	1	2	1	4		1	3							1						15
# patr.	2	1	2	1	7	28	2	6	5	3	1	2	3	14	3	3	3	30	52	6	174

By analyzing the table above, it can be shown that 151 rivers formed groups of at least two matching individuals, which represents approximately 87% of invariance with respect to the total number of rivers, which in this case are 174 along the Magdalena-Cauca basin. This same analysis was performed for the Typology 6, finding 141 rivers that formed invariant groups of at least two individuals, thus obtaining in this case a percentage of invariance of 81 %.

For its part, another goodness of the Kohonen method consists in that with it analysis of importance of variables can also be done, since it is possible to identify, based on the weights maps the input variables, as these determine the map of distances. The foregoing refers to how influential are the weights of each variable in the delimitation of the borders between groups of the map of distances. Figure 4 shows the 5 maps of weights of the input variables of the Typology 6, together with the map of distances.

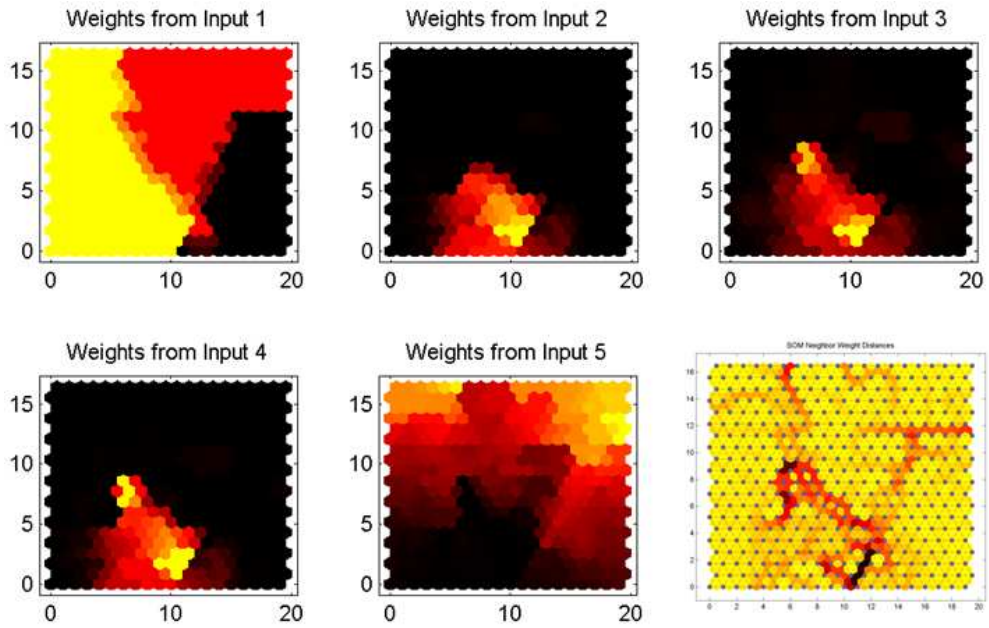

Figure 4. Weights maps of the input variables and map of distances for Typology No. 6.

The figure above shows that the maps of the first and the last variable (Input 1 and Input 5) are the most influential in the formation of the borders of the map of distances, which is shown in the bottom right of the figure. The names of variables of the 5 maps of weights in its order are:

- Input 1. The quarter where presents the highest flow
- Input 2. Average Flow extremely low
- Input 3. Average Flow of large floods
- Input 4. Average monthly multiannual flow
- Input 5. Elevation of the station

To determine if this analysis of importance of variables was reliable, was implemented the technique of decision trees with the "WEKA" software (Waikato Environment for Knowledge Analysis) [7], with which it was possible to establish that for the Typology 6, in the first three levels of importance of the tree was located the variables "The quarter where presents the highest flow" and "Elevation of the station", which confirmed the analysis performed with the Kohonen method results.

Finally, it highlights the feature of the Kohonen method to display the results always in two dimensions, no matter how many variables are analyzed for the classification process. As an example, Figures 2 and 3 present bidimensional maps of distances corresponding to the analysis of classification of variables 13 and 5 respectively (Typologies 5 and 6).

4. CONCLUSIONS

Although the Kohonen method for classification of eco-hydrologic information did not reproduce the exact results of the classification carried out by Ingfocol in 2010, the great potential of this tool is obvious, especially if taken into account that it provides the number of groups that in fact underlie the information presented without the user having to do so by

default and that it is also possible to understand and visualize the results easily in two dimensions regardless of the number of variables analyzed.

Within the typologies of variables analyzed, the typology 5 obtained a percentage of invariance of family groups of currents of 87% and typology 6 a percentage of 81 %, which suggests that these combinations of variables can be used as a first approximation of the classification of currents within ELOHA for the Magdalena-Cauca basin, especially taking into account that they possess only 13 and 5 variables respectively, which are relatively easy to obtain.

The results obtained suggest that despite the change in the number of neurons in the network model for the same typology of variables, the topological structure of the information is preserved. In this sense, the choice of the number of neurons of the model depends on the resolution to which one wishes to analyze the information. Under this premise it is clear that, in contrast to the methods of classification usually employed, with the Kohonen method is possible to obtain the number of groups that in fact underlie the analyzed information depending on the precision with which one wishes to analyze the information.

Based on the typologies of variables raised, it was observed that the typologies with fewer variables and which sought to synthesize the hydrograph for each station (typologies No. 5 and 6), were the typologies which best results provided, with which it is recommended to perform an analysis of importance of variables before entering the classification process of information in the ELOHA framework.

Finally, the analysis of importance of variables made based on the results of Kohonen suggests that, by comparing maps of input variables with the maps of distances between neurons, it is possible to determine the relevant variables within the process of classification.

REFERENCES

[1] N. Poff, B. Richter, A. Arthington, S. Bunn, R. Naiman, E. Kendy, M. Acreman, C. Apse, B. Bledsoe, M. Freeman, J. Henriksen, R. Jacobson, J. Kennen, D. Merritt, J. OKeeffe, J. Olden, K. Rogers, R. Tharme and A. Warner, "The ecological limits of hydrologic alteration (ELOHA): a new framework for developing regional enviromental flow standards," Freshwater Biology, vol. 55, (2010) pp. 147-170.

[2] A. Arthington, S. Bunn, N. Poff and R. Naiman, "The challenge of providing environmental flow rules to sustain river ecosystems," Ecological applications, vol. 16, (2006) pp. 377-395.

[3] CORMAGDALENA, "Plan de manejo de la cuenca del Río Magdalena" (2009).

[4] Ingfocol, Final report: “Implementación de la metodología ELOHA (Límites ecológicos de la alteración hidrológica) para la determinación de caudales ambientales regionales en la cuenca Magdalena-Cauca” (2010).

[5] R. Brian, J. Baumgartner, J. Powell and D. Braun, "A Method for Assessing Hydologic Alteration within Ecosystems.," Conservation Biology, (1996) pp. 1163-1174.

[6] Lin G-F. y Chen L-H., “Identification of homogeneous regions for regional frequency analysis using the self-organizing map”. Journal of Hydrology 324, (2006) pp. 1-9.

[7] H. Mark, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. Witten, "The WEKA Data Mining Software: An Update.," SIGKDD Explorations, vol. 11, no. 1, (2009) pp. 10-18.