Classification of pluviometric networks located in the region of Bogotá, Colombia using artificial neural networks

A R Garrido-Arévalo[1], L M Agudelo[1], N Obregon[2], and V M Garrido[3]
[1] Grupo de Investigación Human Centered Design, Universidad de La Sabana, Bogotá, Colombia
[2] Ciencia e Ingeniería del Agua y del Ambiente, Pontificia Universidad Javeriana, Bogotá, Colombia
[3] Ingeniería Eléctrica y Electrónica, Universidad Tecnológica de Bolívar, Cartagena de Indias, Colombia

E-mail: augustoga@unisabana.edu.co

Abstract. This work presents a methodology for the classification of pluviometric networks using artificial neural networks. For this, the network of stations registered in the Corporación Autónoma Regional de Cundinamarca, Colombia, was analyzed. The network studied consists of 182 stations for the measurement of precipitation and it has a historical series that goes, in some cases, from 1931 to the present. For the classification, three scenarios called types were proposed, in which the number of neurons in the output layer was varied. It was significant that when comparing the results of the different types, the permanence of certain features in the classification was found, indicating the validity of the classification.

1. Introduction

The behavior of the climate depends on many variables, so its prediction is a complex process that requires detailed information. The pluviometric networks provide information for the knowledge and determination of the state of the climate and the flow conditions in a given basin. With the phenomenon of climate change, such knowledge becomes increasingly important because its effectiveness can depend on the effectiveness of the measures taken to prevent the population from being affected by the extreme conditions that arise. Although, it is true that, for the design of the networks there are various methods, in practice they are located with criteria that do not respond to technical requirements [1], due to their complexity. This fact is reflected in the fact that the monitoring stations are located in redundant sites, neglecting other areas.

The concept of neural networks has been applied in fields as varied as electricity [2,3], chemistry [4], recognition of visual patterns [5], geotechnics [6], among others [7-9]. Similarly, this technique has shown positive results in the field of hydraulics and hydrology [10-12]. In this project, the design of a pluviometric network was evaluated based on the available information and artificial neural networks. For this, the network of stations registered in the Corporación Autónoma Regional de Cundinamarca was analyzed. As a result of this project, some general recommendations for the redesign of the pluviometric network in the study region are offered.

The pluviometric networks are fundamental for the planning of river basin management strategies since they provide the basic information for the design, construction and operation of hydraulic structures such as the urban stormwater drainage system. The design of these networks consists in determining the number and location of stations in a region to obtain a historical record of data that can characterize the phenomenon of precipitation in space and time.

Different methods are proposed for the design of an appropriate rain network, including Kriging, the Ward method, self-organized maps (SOM) and the K-means method [13]. Kriging is one of the most used methods for network optimization. However, it has been found that the use of a non-linear pattern learning method, such as an artificial neural network (ANN), produces 15% more reliable results under the same restrictions compared to the conventional kriging method [14].

The aim of this work is to evaluate the distribution of the stations of the pluviometric network of the Bogotá, Cundinamarca region through the application of artificial neural networks.

2. Methodology

2.1. Information collection
The information of the climatological network was provided by the Corporación Autónoma Regional de Cundinamarca, Colombia. The data of interest at this stage were the location of the station (coordinates), its status (in operation or out of operation) and its historical record (monthly precipitation).

2.2. Information processing
Taking into account that the stations of the analyzed network have come into operation at different times, a period was selected in which the largest possible number of stations possesses information and from these the ones that have at least 80% of the data of the data were selected. monthly precipitation in the established period [15].

2.3. Clustering
To perform the classification process, some of the most common methods were analyzed, such as, K-means, Ward and SOM. According to what is stated in [16], the SOM method is selected because it allows to identify the homogeneous regions with more precision. For this, Matlab 2017b licensed by Universidad Tecnológica de Bolívar was used.

2.3.1. Normalization of the variables. The input variables for the classification were: latitude (m), longitude (m), elevation (m), annual average precipitation (mm), standard deviation of the annual average precipitation and average monthly precipitation (mm) of each of the seasons. To prevent the difference in the scale of the variables from affecting the classification, these were transformed so that their ranges were comparable [16].

2.3.2. Creation of the Network model. To correctly differentiate the groups of patterns, it is recommended to use arrays of neurons in the output layer as large as possible. However, if the number of neurons in the output layer is too large, the model may overtrain and highlight the differences between each of the patterns, throwing the same number of groups as the patterns. In accordance with the above, three scenarios were defined with different numbers of neurons in the output layer. The number of iterations is given, according to the recommendations of the literature [17], by the number of neurons in the output layer multiplied by 500. Table 1 details the characteristics of each of these scenarios.

Table 1. Characteristics of the types.

	Type-1	Type-2	Type-3
Neurons in the output layer	100	400	900
Iterations	50000	200000	450000

Journal of Physics: Conference Series **1448** (2020) 012008 doi:10.1088/1742-6596/1448/1/012008

For all cases, a hexagonal topology (hexagonal shaped neurons as well) was chosen, so that the neurons that are not at the edges of the Kohonen layer have 6 neighboring neurons with which they connect virtually.

2.4. Results display
The classification method used offers the advantage of displaying the results on two-dimensional maps regardless of the number of variables included. In this sense, for the visualization and interpretation of the results, two stand out: The Hits map and the distance map between neurons (U-matrix map). In the Hits map, the winning neurons are identified. In it, the number within each neuron indicates the number of stations it represents, that is, the number of victories of each neuron. The neurons with zero value correspond to those that do not represent any pattern (station). An example of Hits Map is shown in Figure 1.

U-matrix allows to visualize how different a neuron is from another. Dark colors indicate a large difference while light colors indicate similarity between neurons and therefore between seasons. This chart allows to identify the groups in which the information is classified. An example of a U-matrix is shown in Figure 2.

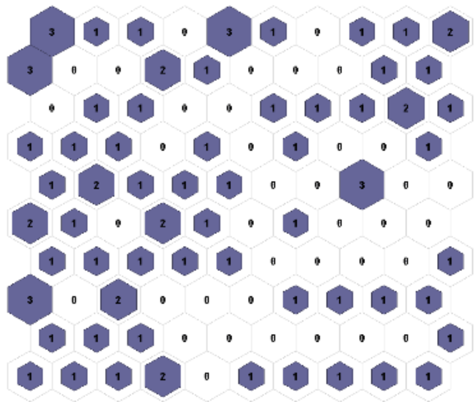

Figure 1. Hits map.

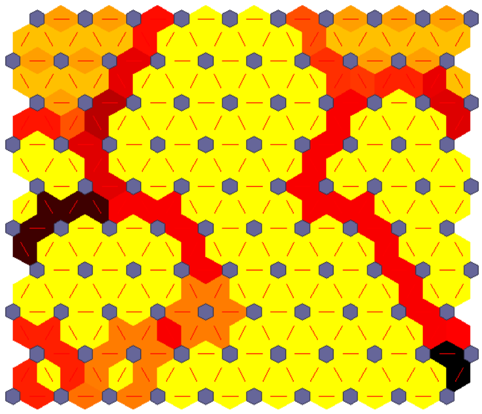

Figure 2. Self-organizing map.

When the number of groups that are formed has been identified, each neuron is associated with the station it represents, for this purpose, there is the Hits matrix, in the rows are the stations and in the columns are the neurons. Each station corresponds to a winning neuron that is indicated with the number 1. In this way, the groups of stations are established.

3. Results
The results obtained in the indicated scenarios are presented below.

3.1. Type-1
The distance map between beurons shows how different a neuron is from another, from which you can identify the groups into which the information is divided. In Figure 3, this map is observed for type-1, here the presence of darker areas that mark the division between the sets is noted. A total of 13 groups with different numbers of neurons are observed. For this classification the lines of greater intensity have been taken into account.

3.2. Type-2
Figure 4 shows the result for type-2, you can see the dark areas that divide the sets. In this case the demarcation is more noticeable than in type-1. 50 groups are distinguished, some of these will be made up of non-winning neurons, so they are not important for the study.

3.3. Type-3

Figure 5 shows the map for type-3. In this case the demarcation is more noticeable than in type-1. 65 groups are distinguished, some of these will be made up of non-winning neurons, so they are not important for the study.

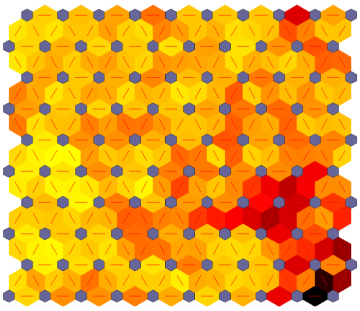

Figure 3. Self-organizing map of neighbor weight distances for type-1.

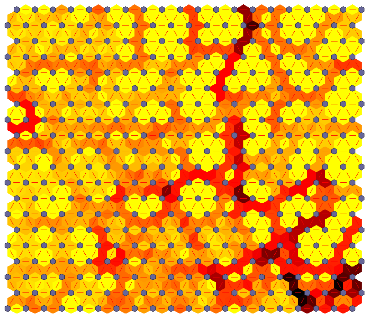

Figure 4. Self-organizing map of neighbor weight distances for type-2.

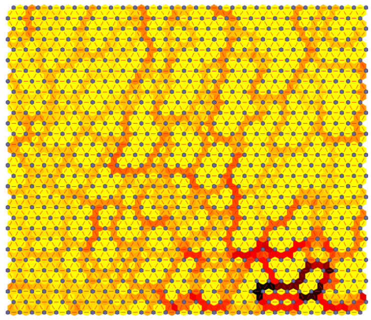

Figure 5. Self-organizing map of neighbor weight distances for type-3.

3.4. Analysis

With the help of the Hits matrix, a neuron and its corresponding group are associated with each station. A summary of the results obtained for the different types is presented in Table 2. It is observed that, by increasing the number of neurons, the number of groups formed increases, this is understood taking into account in a clustering process a high number of neurons in the output layer can cause the model to over-train, highlighting differences between each of the patterns, that is why in type-3, with 900 neurons, about 55% of the stations were classified individually.

Table 2. Classification by type.

Type	Number of neurons	Number of valid groups[a]
1	100	13
2	400	47
3	900	56

[a] Groups in which at least one neuron is winner.

It is significant to note that when conducting a comparative analysis between the different types, the permanence of certain features in the classification is found. A particular case is the E140 station that was always classified individually, as well as the E80 station, in type-1 and type-3. Table 3 shows the most frequent station combinations in the three classifications.

Table 3. Combination of common stations in types.

No.	Stations
1	E101-E61-E16-E65
2	E116-E154-E162-E163- E62-E67
3	E122-E29-E31-E36-E40-E60-E158
4	E142-E28
5	E144-E166-E86-E89-E138
6	E149-E27-E37-E52-E71
7	E150-E61-E102
8	E16-E101
9	E28-E142-E9
10	E42-E129
11	E76-E122

Journal of Physics: Conference Series **1448** (2020) 012008 doi:10.1088/1742-6596/1448/1/012008

4. Conclusions

The application of artificial neural networks for the classification of the stations of the pluviometric network of the area under study showed that such stations can be grouped into 13, 47 and 56 groups, depending on the number of neurons in the output layer. In type-1, with 100 neurons in the output layer, a total of 13 groups were obtained, in type-2, with 400 in the output layer, a total of 47 groups were obtained and, finally, in type-3, with 900 neurons in the output layer, a total of 56 groups were obtained. It is observed that, by increasing the number of these neurons, the number of groups formed increases, this is justified in that, in a process of clustering a high number of neurons in said layer, it can cause the model to over-train, highlighting differences between each of the patterns. The results obtained correspond with the stated in the literature.

Despite the change in the number of neurons in the output layer, while retaining the input variables, the permanence of certain traits in the classification was found, which ensures the existence of similar patterns in the groups obtained. Such is the case of stations E142 (La Casita) and E28 (Santa Teresa), located near the municipality of La Calera, which in the different types were classified in the same group. This fact points out some correspondence between them, so it is recommended, for future studies or for the reengineering of the network, to consider the possible reconstruction of the historical series of one of these from the other, which would allow the relocation of one of them in places where there are fewer stations. The same situation occurs with stations E42 (San Jorge) - E129 (Doña Juana), E76 (Monserrate) - E122 (Esclusa) and E16 (Tres esquinas) - E101 (El Hato No.2), which were classified into similar groups in the different types.

An opposite case is the E80 (Las Margaritas) and E140 (Central No. 2) stations, corresponding to the provincial offices of Sabana Occidente and Bogotá, La Calera, respectively. These stations were classified individually, this fact shows that the information they have could not be reconstructed from the other nearby stations, therefore it is recommended to the Corporación Autónoma Regional de Cundinamarca, in case of a possible reengineering of the network , do not relocate these stations since their information is unique, this added to the fact that they present a valuable historical record since 1959.

References

[1] Gontijo W 2007 *Avaliação e redimensionamento de redes para o monitoramento fluviométrico utilizando o método sharp e o conceito de entropía* (Brasilia: Universidad de Brasilia) p 52

[2] Asimakopoulou F, Tsekouras G, Gonos I and Stathopulos I 2013 Estimation of seasonal variation of ground resistance using artificial neural networks *Electric Power Systems Research* **94** 113

[3] Kheirkhah A, Azadeh A, Saberi M, Azaron M and Shakouri H 2013 Improved estimation of electricity demand function by using of artificial neural network, principal component analysis and data envelopment analysis *Computers & Industrial Engineering* **64** 425

[4] Adib H, Haghbakhsh R, Saidi M, Takassi M, Sharifi F, Koolivand M, Rahimpour M and Keshtkar S 2013 Modeling and optimization of FischereTropsch synthesis in the presence of Co (III)/Al2O3 catalyst using artificial neural networks and genetic algorithm *Journal of Natural Gas Science and Engineering* **10** 14

[5] Fukushima K 2013 Artificial vision by multi-layered neural networks: Neocognitron and its advances *Neural Networks* **37** 103

[6] Taormina R, Chau K and Sethi R 2012 Artificial neural network simulation of hourly groundwater levels in a coastal aquifer system of the Venice lagoon *Engineering Applications of Artificial Intelligence* **25** 1670

[7] Mishra N, Soni H, Sharma S and Upadhyay A 2018 Development and analysis of artificial neural network models for rainfall prediction by using time-series data *International Journal of Intelligent Systems and Applications* **11** 16

[8] Ainslie B, Reuten D, Le N and Zidek J 2009 Application of an entropy-based Bayesian optimization technique to the redesign of an existing monitoring network for single air pollutants *Journal of Environmental Management* **90** 2715

[9] Puangthongthub S, Wangwongwatana S, Kamens R and Serre M 2007 Modeling the space/time distribution of particulate matter in Thailand and optimizing its monitoring network *Atmospheric Environment* **41** 7788

[10] Turlapaty A, Anantharaj V, Younan N and Turk F 2010 Precipitation data fusion using vector space transformation and artificial neural networks *Pattern Recognition Letters* **31** 1184

Journal of Physics: Conference Series **1448** (2020) 012008 doi:10.1088/1742-6596/1448/1/012008

[11] Kim J and Pachepsky Y 2010 Reconstructing missing daily precipitation data using regression trees and artificial neural networks for SWAT streamflow simulation *Journal of Hydrology* **394** 305
[12] Liu Q, Shi Z, Fang N, Zhu H and Ai L 2013 Modeling the daily suspended sediment concentration in a hyperconcentrated river on the Loess Plateau, China, using the Wavelet–ANN approach *Geomorphology* **186** 181
[13] Lin G and Chen L 2006 Identification of homogeneous regions for regional frequency analysis using the self-organizing map *Journal of Hydrology* **324(4)** 1
[14] Chowdhury M, Alouani A and Hossain F 2010 Comparison of ordinary kriging and artificial neural network for spatial mapping of arsenic contamination of groundwater *Stochastic Environmental Research and Risk Assessment* **24(1)** 1
[15] Mishra A and Coulibaly P 2009 Hydrometric network evaluation for Canadian watersheds *Journal of Hydrology* **380(3-4)** 420
[16] Lin G and Chen L 2005 Identification of homogeneous regions for regional frequency analysis using the self-organizing map *Journal of Hydrology* **324(1)** 1
[17] González F 2012 *Agrupación ecohidrológica de corrientes en la cuenca Magdalena-Cauca dentro del marco de referencia ELOHA, empleando mapas autorganizados de Kohonen* (Bogotá: Pontificia Univerisidad Javeriana) p 29