Analele Universităţii din Oradea, Seria Geografie **XXIX**, no. 2/**2019**, pp.**158-173**
ISSN **1221-1273**, E-ISSN **2065-3409** DOI 10.30892/auog.292116-813

DATA CLASSIFICATION TECHNIQUES AND SYSTEM FOR PREDICTING DISCHARGES IN THE GAMBIA RIVER BASIN

Cheikh FAYE *
Assane Seck University of Ziguinchor, Sciences and Technology Faculty, Department of Geography, Ziguinchor, Senegal, e-mail: cheikh.faye@univ-zig.sn

Bouly SANÉ
Assane Seck University of Ziguinchor, Sciences and Technology Faculty, Department of Geography, Ziguinchor, Senegal, e-mail: B.SANE79@zig.univ.sn

Ibrahima THIAW
Cheikh Anta Diop University of Dakar, Arts and Human Sciences Faculty, Department of Geography, Dakar, Senegal, e-mail: ibrahima4.thiaw@ucad.edu.sn

Cheikh Tidiane WADE
Assane Seck University of Ziguinchor, Sciences and Technology Faculty, Department of Geography, Ziguinchor, Senegal, e-mail: cheikh-tidiane.wade@univ-zig.sn

Citation: Faye, C., Sané, B., Thiaw, I., & Wade, C.T., (2019). Data Classification Techniques and System for Predicting Discharges in the Gambia River Basin. *Analele Universităţii din Oradea, Seria Geografie*, 29(2), 158-173. https://doi.org/10.30892/auog.292116-813

Abstract: Within the framework of water resources management, numerous research works and methods were led in world. In this trail, we noted a fast development of time series data mining (TSDM) which supplies a new method for water resources management. This article examines the trend of discharge during the high water period (from July till November) in the basin of Gambia measured at the Mako station for 1970-2013 period. Methodology consisted at first in calculation and in standardization of data by the method of z-score of some statistical parameters (mean, maximum, minimum, range and standard deviation). Obtained series were afterward submitted to classifications techniques such as k-means clustering and Agglomerative Hierarchical Clustering (AHC) of TSDM to cluster and discover the discharge patterns in terms of the autoregressive model. Based on these methods, a discharge forecast model has been developed. For the validation of the indicated model, and with respect to the maximum discharge, the coefficients of discharge growth and decay, respectively on the phase of rise and the phases of rise and descent waters, were calculated. This study presents basin discharge dynamics in high water period based on TSDM.

Key words: data mining; discharge; forecast model; hydrological process; clustering; techniques

* * * * * *

* Corresponding Author

INTRODUCTION

In Senegal, the collection of climatological and hydrological data is managed respectively by the National Agency for Civil Aviation and Meteorology (ANACM) and the Directorate of Management and Planning of Water Resources (DGPRE). These data are very useful in research, analysis of historical trends and future forecasts. With the multiplication of databases, various techniques of data analysis and knowledge extraction are used worldwide (Mishra et al., 2014) and by researchers of various disciplines: hydrology, the environment, climatology, computer science, mathematics, etc. Today, the development of information technology has generated huge amounts of databases covering various fields of science and technology. Data mining is widely applied in the scientific research. Finding association rules, sequential patterns, classifying and grouping data are typical tasks involved in the data mining process. The various classification techniques all aim at distributing n individuals, characterized by p variables X1, X2, …, XP, into a certain number m of subgroups that are as homogeneous as possible, each group being well differentiated from the others (Larose, 2005). Two major classification techniques exist: partitioning and hierarchical classification.

Data mining refers to the extraction of knowledge from large amounts of data. The time series data mining (TSDM) methodology follows the delayed integration process to predict future occurrences of significant events. The tools of data mining boil down to neural networks and decision trees allowing the prediction of a qualitative variable (classification trees) or quantitative (regression tree) (Gupta and Chaturvedi, 2013). Nevertheless, the most innovative methods concern the search for association rules (Agrawal et al., 1993) which can lead to observations of the "composition of the consumer's shopping cart" type, and the study of frequent sequences allowing to understand customer behavior over time (Agrawal and Srikant, 1995). Two major types of methodologies preside over data mining techniques: the supervised mode which requires the definition of a dependent variable (thus some hypotheses) and the unsupervised mode where all the variables are considered on the same plane (detection of associations, classification, partition, etc.) (Crié, 2003). Hydrological databases are sets of various record values that diverge over time. Research based on the theory of data mining and hydrological techniques is needed to analyze hydrological, climatological and sedimentary databases for different types of study (Gupta and Chaturvedi, 2013).

The term "data mining" does not mean the generation of data or the data sets themselves, but only the practice of data analysis. Many of the methods used come from statistics: however, data mining is not a purely statistical process, but an interdisciplinary process that uses learning techniques from computer science and mathematics (especially learning unsupervised) and allied with artificial intelligence (Piatetsky-Shapiro and Frawley, 1991). These efficient methods are integrated into data mining software to enable the evaluation of large datasets (Agrawal and Srikant, 1995; Crié, 2003).

Classification has important role in science through usage of multidimensional statistic techniques. In biological sciences, such as botany, zoology, ecology, the term "taxonomy" is used to designate the art of classification. Also, the classification techniques are widely used in geosiences (geology, pedology, geography, study of the pollutions etc.). Over the past decade, a lot of effort has been conducted to extract knowledge from hidden historical data (Jayanthi, 2007). To this end, real-time hydrological forecasting is a major challenge for the scientific community (Mishra et al., 2014).

In the field of hydrological data mining, various techniques are used to extract knowledge from historical data (Mishra et al., 2013). Some of the topics of interest to this study are the discovery of models for the exploitation of hydrological data during the high water period in the Gambia River Basin.

Database extraction that combines the fields of time series analysis and data mining techniques (Aydin et al., 2009) therefore remains an essential technique for hydrological analyzes. The main objective of this study is to develop a data mining application using modern

information technology and to discover the hidden information or models behind the historical hydrological data during the high water period at Mako station in the Gambia River Basin. Data mining tools such as similarity search, k-means grouping and Agglomerative Hierarchical Clustering (AHC) model are used in this article.

STUDY AREA

For this work, the Gambia River was selected because of the high variability of water resources. Its basin, with an area of nearly 77,100 km², extends in latitude, from 11° 22' North (in the Fouta-Djalon) to 14° 40' North (in the Far-Eastern Ferlo) and, in longitude, from 11° 13' West (Fouta-Djalon) to 16° 42' West (Banjul, mouth) (Lamagat, 1989; Dione 1996; Sow 2007).

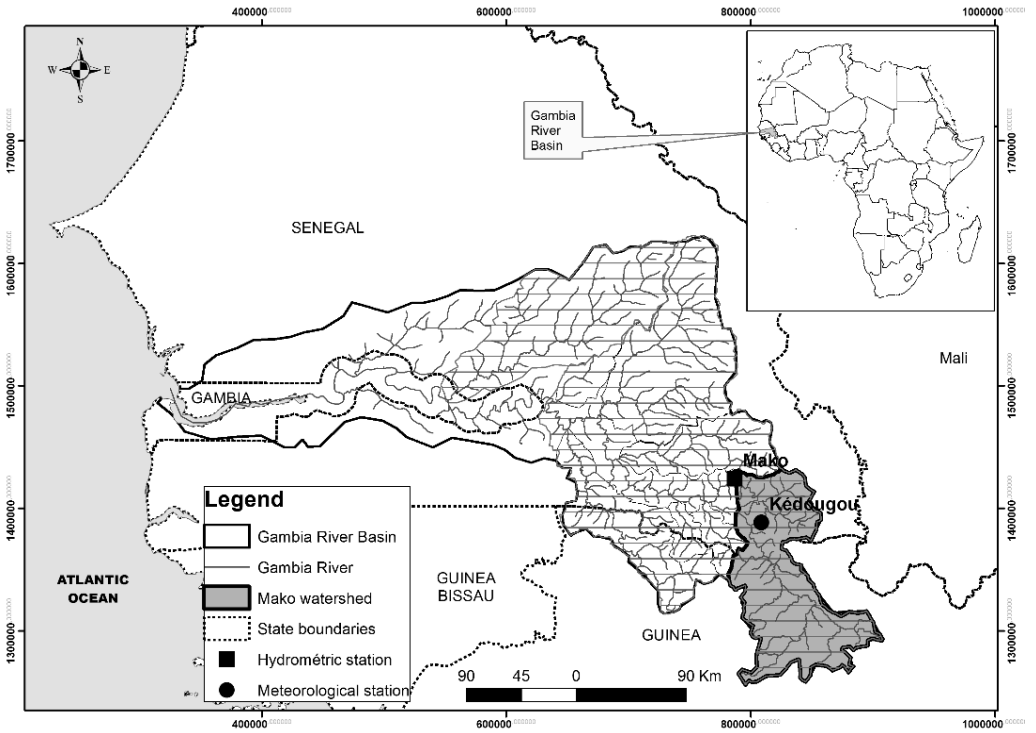


Figure 1. Gambia River Basin
(Source: CSE)

The length of the river at Mako Station is 328 km. The Mako Hydrometric Station is located on the main stream of the river. For this work, daily discharge data were taken over a period of 44 years (1970-2013). At the Kedougou meteorological station, the average maximum temperature is 30° C, the minimum temperature is 25° C and the precipitation is 1000 mm (1970-2013) (figure 1).

DATA AND METHODS

Select a data set

For this work, the daily data of discharge of the Mako station and climatological data (precipitation and temperature) of Kédougou were collected at the DGPRE. The daily data, in accordance with the requirements of the methods used, were converted into monthly means. On the monthly values of discharge (Q), a series 44 years is chosen (1970-2013) and the tests were carried out on a period of high water (July-November).

Statistical analysis, data standardization

The five statistical parameters (Q_{mean}, $Q_{maximum}$, $Q_{minimum}$, Q_{range} and $Q_{deviation}$) were calculated on each month with the discharge data. In order to have an efficient analysis of the data on the series and the period considered, the calculated parameters were normalized using the z-score technique through the following formula:

$$z = \frac{(x_i - x_m)}{\sigma}$$

With x_i which is the value of the month, x_m the mean of the series and σ the standard deviation of the series.

Standardization was necessary to prevent the results of the study from being affected by large variations in the data.

Data segmentation

For data segmentation, analysis of the basin hydrograph and Monthly Discharge Ratio (MDR, as the ratio of the monthly discharge to the annualy discharge) (Table 1) at the Mako station over the period 1970-2013 divides it into 3 segments: low water (May-July), high water (August-October) and low water (November-April). For this study, although the months of July and November are months of lowwater (MDR <1), they are used in the period (so-called high water) on which the tests are applied. This choice can be explained by the importance of their past discharges.

Table 1. Monthy Q values and MRD at Mako station (1970-2013)
(Data source: DGPRE)

	M	June	July	A	S	O	N	D	J	F	M	A	Year
Q (m^3/s)	0.23	7.42	78.8	310	432	201	64.1	23.6	11.3	5.30	2.09	0.37	95
MDR	0.002	0.08	0.83	3.26	4.54	2.12	0.67	0.25	0.12	0.06	0.02	0.00	1
	Low waters			High waters			Low waters						

The discharge of watercourses changes gradually with the changes in rainfall. The different climates in the basin thus cause different discharge processes for a better study of the discharge processes, the study of the climatic framework is fundamental as indicated in the works of Faye (2018) and Faye and Mendy (2018). This studies highlight the great climatic variability in the Gambia Basin with the presence of two periods: a wet period marked by pluviometric abundance during the 1950s and 1960s and a dry period marked by drought in the 1970s and 1980s (Faye, 2018; Faye and Mendy, 2018). On the other hand, during the 2000s, it was noted in the Gambia River Basin that an increase in rainfall predicted the improvement of rainfall patterns in the basin compared with the drought period of previous decades. However the persistence and sustainability of the increase are still to be proven, knowing that the long enough climatological scale is thirty years (Faye et al., 2017). In this article, the period of high water is chosen and the statistical data obtained here have been subjected to classification tests.

Apply K-means clustering and find out number of clusters

The k-means classification is a very efficient iterative method for finding spherical groups in small and medium sized databases. Its application requires several times the calculations in order to retain only the most optimal solution for the chosen criterion. For the first iteration we choose a starting point which consists in associating the center of the k classes with k objects (taken randomly or not). The distance between the objects and the k centers is then calculated and the objects are assigned to the centers of which they are nearest. Then we redefine the centers from the objects that have been assigned to the different classes. Then the objects are reassigned according to their distance to the new centers, until the convergence is reached (Gupta and Chaturvedi, 2013).

Apply DTW and find out similarities and dissimilarities

The search for similarity in time series analysis is one of the fastest and most demanding areas of development in data mining. Unlike normal database queries, which find the data corresponding to the given query exactly, a similarity search finds data sequences that differ only slightly from the given query sequence. It can be classified into two categories (Gupta and Chaturvedi, 2013):

The "all correspondence" category: In this type of time series data matching must be of equal length.

The category "subsequent correspondence": In this category, a sequence of requests X and a longer sequence Y were taken. The goal is to identify the sequence in Y, starting with Yi, which has the best matches of X, and to account for its shift within Y. The main difficulty is to define a measure of similarity (Rakthanmanon et al., 2012). For the analysis of similarity of time series data, Euclidean distance is generally used as a measure of similarity. Given two sequences, X = (x_1, x_2,…., xn) and Y = (y_1, y_2,…., y_n) with n = m, the Euclidean distance is defined as follows (Mishra et al., 2014):

$$D(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

The DTW (Dynamic Time Warping) Algorithm: The DTW is an algorithm for measuring the optimal similarity between two data sequences. The series data vary not only over time amplitudes, but also with the progression of time that hydrological processes can reveal different velocities in response to different environmental conditions. A nonlinear alignment produces a similar measure, allowing similar shapes to match even if they are out of phase in the time axis (Ding et al., 2008). The sequences are nonlinearly "deformed" in the time dimension to determine a measure of their similarity independent of certain nonlinear variations in the time dimension. To find the best alignment between X and Y time sequences one needs to find the way through the grid.

Apply Agglomerative Hierarchical Clustering (AHC) algorithms and find out dendograms

To identify the discharge model from the corresponding data series, each hydrological period obtained after k-means clustering was taken, and then analysis of a discharge model in each of the periods was made. The analysis involved 5-year hierarchical classification techniques with observance of mean discharge data for selected months over the hydrological period. AHC is a subdivision of hierarchical clustering and a bottom-up approach, which proceeds by a series of fusions of the N objects into groups. If it is given a set of N items to be clustered and an N*N distance (or similarity) matrix then the basic process of agglomerative hierarchical clustering applied in this study is done iteratively following these four steps (Gupta and Chaturvedi, 2013) (figure 2): 1. Start with N clusters each containing a single entity, and an N × N symmetric matrix of distances (or similarities) Let dij = distance between item i and item j. 2. Search the distance matrix for the nearest pair clusters (i.e., the two clusters that are separated by the smallest distance). Denote the distance between these most similar clusters U and V by d_{UV}. 3. Merge clusters U and V into new clusters labeled T. Update the entries in the distance matrix by (a). Deleting the rows and columns corresponding to clusters U and V, and (b). Adding a row and column giving the distances between the new cluster T and all the remaining clusters. 4. Repeat steps (2) and (3) a total of N-1times.

In our analysis, the discharge pattern (time series data of discharge for the months in the hydrological period) of a year (among 44 years) is clustered into several clusters. But the year which formed the cluster center formed the pattern with its discharge data for the months in the period. All other members (years) in the cluster attained membership of the cluster because there was similarity to the year representing the center so they can be said to follow the pattern.

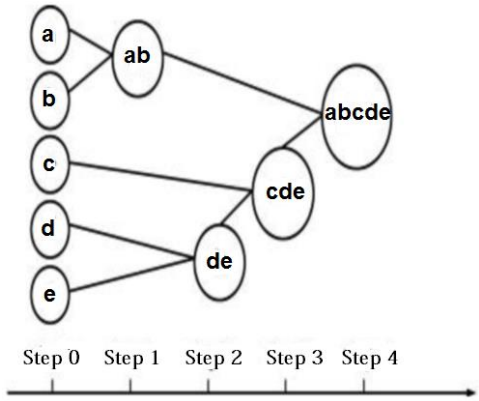


Figure 2. Cluster tree obtained by AHC
(Source: Gupta and Chaturvedi, 2013)

Now, the center (the year) in the cluster is obtained, plotting of the discharge data of that year corresponding to daily discharge in the months, along the x-axis would give the pattern.

Like the k-means classification, the techniques of AHC and the criterion Wards are applied over the period of high water (July to November) on the 44 years selected (1970-2013) for the analysis of the models. AHC is particularly useful for finding hidden models in multidimensional data. Since this is an unsupervised learning pattern, the number of classes can be large or small at times.

Calculate moving average in monthly discharge standardized data

A sliding mean, also known as moving average, is a type of finite impulse response filter used to analyze a set of data points by creating a series of means of the different subsets of the complete set of data. A moving average is commonly used with time series data to smooth out short-term fluctuations and highlight long-term trends or cycles. The threshold between the short term and the long term depends on the application, and the parameters of the moving average will be set accordingly. The simple moving average formula is given below (Gupta and Chaturvedi, 2013):

$$S_t = \frac{1}{k}\sum_{n=0}^{k-1} x_{t-n} = \frac{x_t + x_{t-1} + x_{t-2} + \cdots + x_{t-k+1}}{k} = x_{t-1} + \frac{x_t + x_{t-k}}{k}$$

With which is the simple moving average, k the observations, t the data point with respect to time and n the number of data points.

Calculate coefficient of growth and decay with respect to peak discharge

On either side of the month of the maximum discharge (September with 432 m^3/s) (figure 3), it seems necessary to calculate the growth coefficient and the decrease of the discharge to validate the stream discharge model before and after the peak. In hydrology, the coefficient of discharge growth or decay (k) is usually expressed in the following exponential decay (Mishra et al., 2014):

$$\frac{Q}{Q_0} = e^{-kt}; \qquad so \qquad \log\left(\frac{Q}{Q_0}\right) = \log(e^{-kt})$$

Applying the natural logarithm of the two sides and deducing k while knowing that t is here equal to 1 (one month), we obtain:

$$k = -l\,\mathrm{og}\left(\frac{Q}{Q_0}\right) \quad = \quad logQ_0 - logQ$$

Where k is the coefficient of discharge growth or decay, Q the monthly mean discharge, Q_0 the discharge in the previous month.

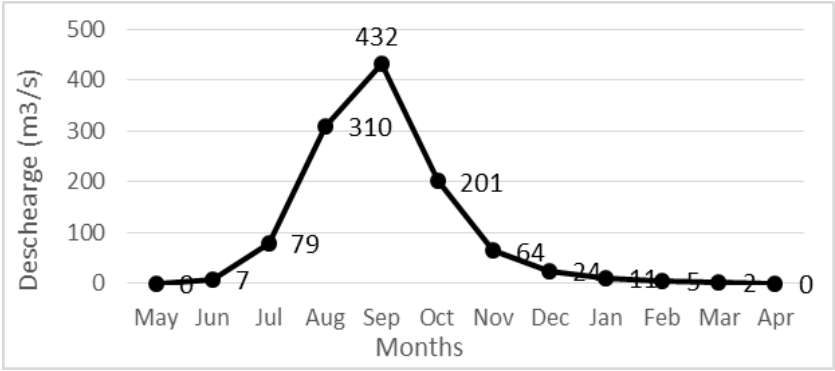


Figure 3. Evolution of the mean monthly discharges at Mako (1970-2013)
(Data source: DGPRE)

To base the methodology raised, the XLSTAT tool 2014 software version, which contains the implementation of various classification algorithms (such as K-means, AHC, etc., and other data mining techniques), is used. For this work, the daily discharge data for the period from July to November over 44 years (from 1970 to 2013) are used. XLSTAT is used to classify and find classes.

RESULTS AND DISCUSSION

The k-means method

After the division of the hydrological period into three segments and the choice of the period of high water (from July to November) for this study, the statistical parameters (mean, maximum, minimum, range and standard deviation) obtained and standardized over the period 1970 -2013 were subject to a grouping of k-means. Thus, a total of 220 (5 months in each of the 44 years) cases (months) based on the 5 parameters from the data of the chosen period was used and subject to this grouping. Table 2 indicates a classification of the different months according to obtained (four) classes.

Table 2. Assignment Classes by Object After Application of K-means from 1970 to 2013
(Data source: DGPRE)

Observation	Class	Distance to centroid	Observation	Class	Distance to centroid	Observation	Class	Distance to centroid	Observation	Class	Distance to centroid
Jul-70	1	4.97	Jul-81	1	1.67	Jul-92	4	0.87	Jul-03	1	2.14
Aug-70	2	1.89	Aug-81	4	1.44	Aug-92	4	0.65	Aug-03	1	1.62
Sep-70	1	1.82	Sep-81	3	0.71	Sep-92	1	1.31	Sep-03	1	1.3
Oct-70	3	0.32	Oct-81	3	0.28	Oct-92	3	0.54	Oct-03	4	0.34
Nov-70	4	0.51	Nov-81	4	0.68	Nov-92	4	0.41	Nov-03	4	1.41
Jul-71	4	0.83	Jul-82	4	1.17	Jul-93	4	0.86	Jul-04	1	2.59
Aug-71	4	0.73	Aug-82	3	0.77	Aug-93	3	1.15	Aug-04	4	0.95
Sep-71	3	0.89	Sep-82	3	0.38	Sep-93	3	1.26	Sep-04	4	1.61
Oct-71	3	0.28	Oct-82	3	0.51	Oct-93	3	0.47	Oct-04	4	0.95
Nov-71	4	0.71	Nov-82	4	0.55	Nov-93	4	0.78	Nov-04	4	0.38
Jul-72	4	1.21	Jul-83	3	1.02	Jul-94	1	0.93	Jul-05	2	2.44
Aug-72	4	0.84	Aug-83	3	1.05	Aug-94	4	0.91	Aug-05	1	1.8
Sep-72	3	0.96	Sep-83	3	1.9	Sep-94	4	1.75	Sep-05	4	0.89
Oct-72	3	0.34	Oct-83	3	0.38	Oct-94	4	0.94	Oct-05	4	0.29
Nov-72	4	0.57	Nov-83	4	0.91	Nov-94	4	1.45	Nov-05	4	0.32
Jul-73	1	3.2	Jul-84	2	1.34	Jul-95	1	1.8	Jul-06	3	0.6
Aug-73	2	1.83	Aug-84	3	1.16	Aug-95	2	0.76	Aug-06	3	0.86
Sep-73	4	1.54	Sep-84	3	2.8	Sep-95	4	1.42	Sep-06	4	0.99
Oct-73	3	0.28	Oct-84	3	1.03	Oct-95	4	0.75	Oct-06	4	0.65
Nov-73	4	0.72	Nov-84	4	0.97	Nov-95	4	0.57	Nov-06	4	0.48

Jul-74	1	1.32	Jul-85	1	0.96	Jul-96	4	0.87	Jul-07	4	0.46
Aug-74	2	0.46	Aug-85	4	1.47	Aug-96	4	0.65	Aug-07	3	1.78
Sep-74	3	1.69	Sep-85	4	0.52	Sep-96	1	1.31	Sep-07	3	0.85
Oct-74	4	0.58	Oct-85	4	0.98	Oct-96	3	0.54	Oct-07	3	0.72
Nov-74	4	0.37	Nov-85	4	0.67	Nov-96	4	0.41	Nov-07	4	0.66
Jul-75	1	0.82	Jul-86	3	0.44	Jul-97	2	0.79	Jul-08	2	1.16
Aug-75	4	0.96	Aug-86	3	1.11	Aug-97	2	0.37	Aug-08	2	1.48
Sep-75	1	1.28	Sep-86	3	0.74	Sep-97	4	0.4	Sep-08	1	1.66
Oct-75	4	0.26	Oct-86	3	0.58	Oct-97	4	0.34	Oct-08	4	1.27
Nov-75	4	0.4	Nov-86	4	0.61	Nov-97	4	0.29	Nov-08	2	1.57
Jul-76	1	2.42	Jul-87	3	0.63	Jul-98	4	0.88	Jul-09	4	2.13
Aug-76	4	0.11	Aug-87	3	1.33	Aug-98	4	1.81	Aug-09	4	1.59
Sep-76	3	0.75	Sep-87	3	0.52	Sep-98	1	1.5	Sep-09	2	1.03
Oct-76	4	0.91	Oct-87	4	0.49	Oct-98	4	0.53	Oct-09	4	0.9
Nov-76	4	1.07	Nov-87	4	0.42	Nov-98	4	0.21	Nov-09	4	0.77
Jul-77	3	0.54	Jul-88	4	1.01	Jul-99	4	0.73	Jul-10	3	0.32
Aug-77	3	1.49	Aug-88	4	1.78	Aug-99	4	0.62	Aug-10	4	0.68
Sep-77	3	1.22	Sep-88	4	0.94	Sep-99	4	0.53	Sep-10	2	2.14
Oct-77	3	0.72	Oct-88	3	0.22	Oct-99	4	1.35	Oct-10	2	1.69
Nov-77	4	0.63	Nov-88	4	0.71	Nov-99	4	0.59	Nov-10	1	3.37
Jul-78	3	0.81	Jul-89	3	1.03	Jul-00	4	0.68	Jul-11	1	0.74
Aug-78	4	0.76	Aug-89	4	1.53	Aug-00	4	1.22	Aug-11	4	1.67
Sep-78	1	1.9	Sep-89	4	0.6	Sep-00	3	0.66	Sep-11	4	1.29
Oct-78	4	0.79	Oct-89	3	0.63	Oct-00	4	1.21	Oct-11	1	1.47
Nov-78	4	0.41	Nov-89	4	0.52	Nov-00	4	0.16	Nov-11	4	0.5
Jul-79	4	0.67	Jul-90	4	0.29	Jul-01	4	0.45	Jul-12	4	2.01
Aug-79	3	0.62	Aug-90	3	1.95	Aug-01	4	0.35	Aug-12	4	1.53
Sep-79	3	1.27	Sep-90	3	0.85	Sep-01	4	1.03	Sep-12	4	0.55
Oct-79	4	0.73	Oct-90	3	0.73	Oct-01	4	0.89	Oct-12	1	0.79
Nov-79	4	0.39	Nov-90	4	0.66	Nov-01	4	0.55	Nov-12	4	0.53
Jul-80	1	0.93	Jul-91	1	1.13	Jul-02	3	0.59	Jul-13	4	1.24
Aug-80	2	0.53	Aug-91	4	1.29	Aug-02	3	0.29	Aug-13	1	1.35
Sep-80	4	1.03	Sep-91	3	0.64	Sep-02	3	1.78	Sep-13	4	0.54
Oct-80	3	0.15	Oct-91	3	0.33	Oct-02	4	0.49	Oct-13	3	0.93
Nov-80	4	0.81	Nov-91	4	0.57	Nov-02	4	0.7	Nov-13	4	0.96

Table 3. Centroids of classes after the application of K-means from 1970 to 2013
(Data source: DGPRE)

Class	Mean	Max	Min	Range	Standard deviation	Sum of weights	Within-class variance	Minimum distance to centroid	Mean distance to centroid	Maximum distance to centroid
1	0,70	1,57	-0,10	1,73	1,72	28	3.91	0.74	1.72	4.97
2	1,31	0,46	1,92	-0,02	0,13	15	2.21	0.37	1.30	2.44
3	-1,05	-1,41	-0,73	-1,31	-1,25	59	0.97	0.15	0.84	2.80
4	-0,32	-0,34	-0,36	0,26	-0,31	118	0.86	0.11	0.82	2.13

From a typology (segmentation), this method of data analysis made it possible to obtain a simple schematic representation of the complex starting data table. This resulted in a partition of n individuals (months) into classes, defined by the observation of p variables (mean, maximum, minimum, range and deviation). Table 3 gives the class centers of gravity (these are the coordinates of the centroids of the classes for the different parameters) and the results by class (Sum of weights, Within-class variance, Minimum distance to centroid, Mean distance to centroid, Maximum distance to centroid).

Depending on the distribution of cases in the classes, the annual discharge process could be obtained as separate classes. For this work where only the high water period is used, the k-means algorithm also offers one of the graphical representations of the processed data. Thus figure 4 shows evolution curves of a set of statistical parameters analyzed by the k-means algorithm (mean, maximum, minimum, range and standard deviation).

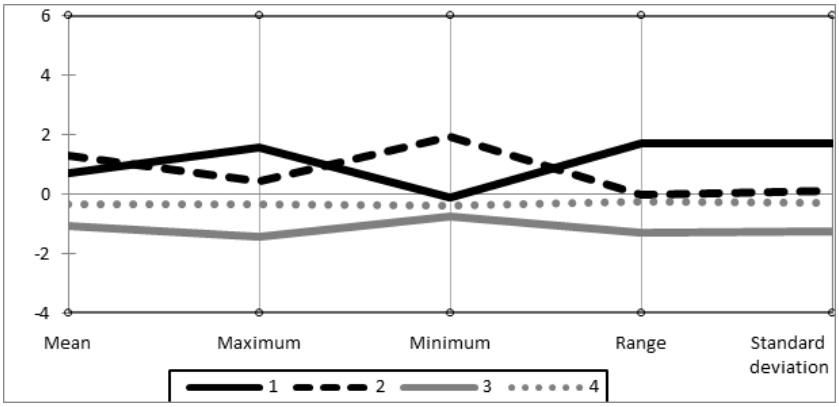


Figure 4. Graphical representation of the classes obtained by the application of K-means from 1970 to 2013 (Data source: DGPRE)

Agglomerative Hierarchical Clustering (AHC)

Like the k-means classification, the techniques of AHC are applied over the period of high water (July to November) on the 44 years selected (1970-2013) for the analysis of the models. The main role of AHC is to identify classes or groups of discharge series that are similar. By applying the AHC algorithm on the data set, four different classes were obtained (table 4).

Table 4. Classes of assignment by object after application of the AHC from 1970 to 2013 (Data source: DGPRE)

Observation	Class	Observation	Class	Observation	Class	Observation	Class	Observation	Class	Observation	Class
Jul-70	1	Sep-77	3	Nov-84	3	Aug-92	3	Oct-99	3	Jul-07	3
Aug-70	2	Oct-77	4	Jul-85	1	Sep-92	3	Nov-99	3	Aug-07	4
Sep-70	3	Nov-77	3	Aug-85	3	Oct-92	4	Jul-00	3	Sep-07	4
Oct-70	4	Jul-78	4	Sep-85	3	Nov-92	3	Aug-00	4	Oct-07	4
Nov-70	3	Aug-78	3	Oct-85	3	Jul-93	3	Sep-00	4	Nov-07	3
Jul-71	3	Sep-78	3	Nov-85	3	Aug-93	3	Oct-00	3	Jul-08	2
Aug-71	3	Oct-78	3	Jul-86	4	Sep-93	4	Nov-00	3	Aug-08	3
Sep-71	3	Nov-78	3	Aug-86	4	Oct-93	4	Jul-01	3	Sep-08	3
Oct-71	4	Jul-79	3	Sep-86	4	Nov-93	3	Aug-01	3	Oct-08	3
Nov-71	3	Aug-79	4	Oct-86	4	Jul-94	1	Sep-01	4	Nov-08	3
Jul-72	3	Sep-79	4	Nov-86	3	Aug-94	3	Oct-01	3	Jul-09	3
Aug-72	3	Oct-79	3	Jul-87	4	Sep-94	3	Nov-01	3	Aug-09	3
Sep-72	4	Nov-79	3	Aug-87	4	Oct-94	3	Jul-02	4	Sep-09	2
Oct-72	4	Jul-80	1	Sep-87	4	Nov-94	3	Aug-02	4	Oct-09	3
Nov-72	3	Aug-80	2	Oct-87	3	Jul-95	3	Sep-02	4	Nov-09	3
Jul-73	1	Sep-80	3	Nov-87	3	Aug-95	2	Oct-02	3	Jul-10	4
Aug-73	2	Oct-80	4	Jul-88	3	Sep-95	3	Nov-02	3	Aug-10	3
Sep-73	3	Nov-80	3	Aug-88	4	Oct-95	3	Jul-03	2	Sep-10	2
Oct-73	4	Jul-81	3	Sep-88	3	Nov-95	3	Aug-03	2	Oct-10	2
Nov-73	3	Aug-81	3	Oct-88	4	Jul-96	3	Sep-03	2	Nov-10	2
Jul-74	1	Sep-81	4	Nov-88	3	Aug-96	3	Oct-03	3	Jul-11	1
Aug-74	2	Oct-81	4	Jul-89	4	Sep-96	3	Nov-03	3	Aug-11	3
Sep-74	4	Nov-81	3	Aug-89	3	Oct-96	4	Jul-04	1	Sep-11	3
Oct-74	3	Jul-82	3	Sep-89	3	Nov-96	3	Aug-04	4	Oct-11	3

Nov-74	3	Aug-82	4	Oct-89	4	Jul-97	2	Sep-04	3	Nov-11	3
Jul-75	1	Sep-82	4	Nov-89	3	Aug-97	2	Oct-04	3	Jul-12	3
Aug-75	3	Oct-82	4	Jul-90	3	Sep-97	3	Nov-04	3	Aug-12	3
Sep-75	2	Nov-82	3	Aug-90	4	Oct-97	3	Jul-05	2	Sep-12	3
Oct-75	3	Jul-83	4	Sep-90	4	Nov-97	3	Aug-05	3	Oct-12	1
Nov-75	3	Aug-83	4	Oct-90	4	Jul-98	3	Sep-05	3	Nov-12	3
Jul-76	1	Sep-83	4	Nov-90	3	Aug-98	3	Oct-05	3	Jul-13	3
Aug-76	3	Oct-83	4	Jul-91	1	Sep-98	3	Nov-05	3	Aug-13	1
Sep-76	4	Nov-83	3	Aug-91	3	Oct-98	3	Jul-06	4	Sep-13	3
Oct-76	3	Jul-84	3	Sep-91	4	Nov-98	3	Aug-06	4	Oct-13	4
Nov-76	3	Aug-84	4	Oct-91	4	Jul-99	3	Sep-06	3	Nov-13	3
Jul-77	4	Sep-84	4	Nov-91	3	Aug-99	3	Oct-06	3		
Aug-77	4	Oct-84	3	Jul-92	3	Sep-99	3	Nov-06	3		

For class analysis of the AHC, the Sum of weights, the Within-class variance, the Minimum distance to centroid, the Mean distance to centroid, the Maximum distance to centroid are calculated (table 4) and represented in graph form (figure 5).

Table 5. Centroids of classes after the application of AHC from 1970 to 2013
(Data source: DGPRE)

Class	Mean	Max	Min	Range	Standard deviation	Sum of weights	Within-class variance	Minimum distance to centroid	Mean distance to centroid	Maximum distance to centroid
1	0,70	1,57	-0,10	1,73	1,72	13	3,37	0,50	1,56	3,96
2	1,31	0,46	1,92	-0,02	0,13	17	3,49	0,49	1,61	3,57
3	-1,05	-1,41	-0,73	-1,31	-1,25	131	1,27	0,22	0,99	2,58
4	-0,32	-0,34	-0,36	-0,26	-0,31	59	1,09	0,15	0,87	2,84

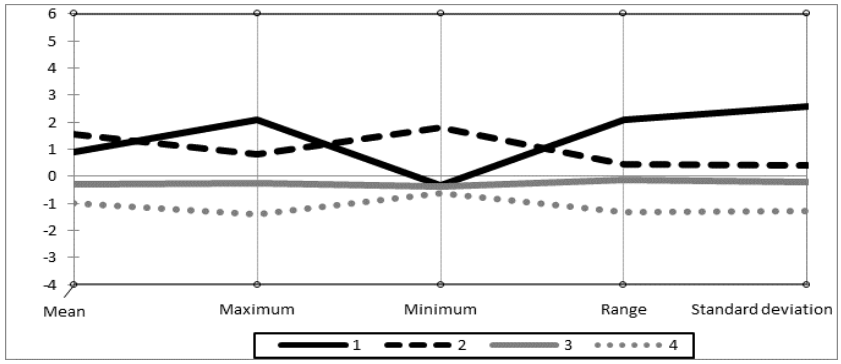


Figure 5. Graphical representation of the classes obtained by application of the AHC from 1970 to 2013
(Data source: DGPRE)

Similarity analysis, model detection and growth and decay coefficient with respect to peak discharge

The observation of similarities is also made on data from the high water period from 1970 to 2013. This technique is applied to indicate similar discharges between the months and years of the series. This is because the time series of discharge rates vary not only in terms of expression amplitudes, but also in terms of temporal progression, because the stream flows at different rates in time depending on different natural conditions or at different locations in the

basin at different times (Mishra et al., 2014). For this study, the distance between objects and k centers indicated by the k-means classification technique is used as the similarity matrix for the high-water period (table 6). These distances between the central objects represent the Euclidean distances between the central objects of the classes for the different descriptors. The similarity matrix made it possible to compare the monthly discharges of the 44 years of the series. For example, the distances between the central objects indicate strong similarities between the months of August 1976 and October 1980.

Table 6. Distances between the central objects for the period of high water according to the k-means from 1970 to 2013
(Data source: DGPRE)

	July-2011	**Aug-97**	**Oct-80**	**Aug-76**
July-2011	0	3.96	5.34	3.47
Aug-97	3.96	0	4.51	3.24
Oct-80	5.34	4.51	0	1.92
Aug-76	3.47	3.24	1.92	0

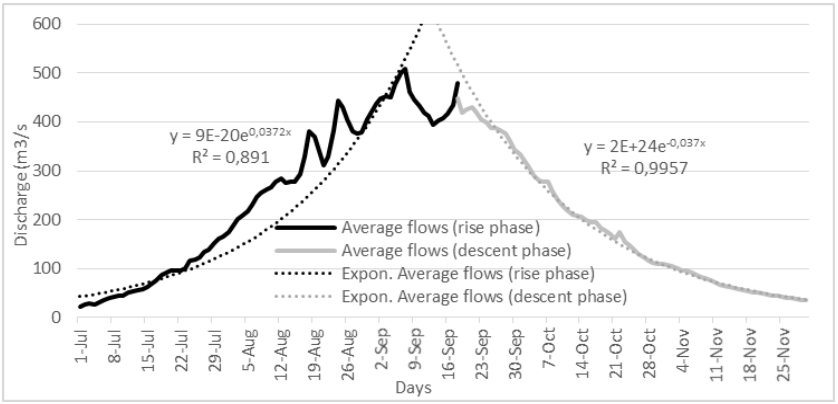
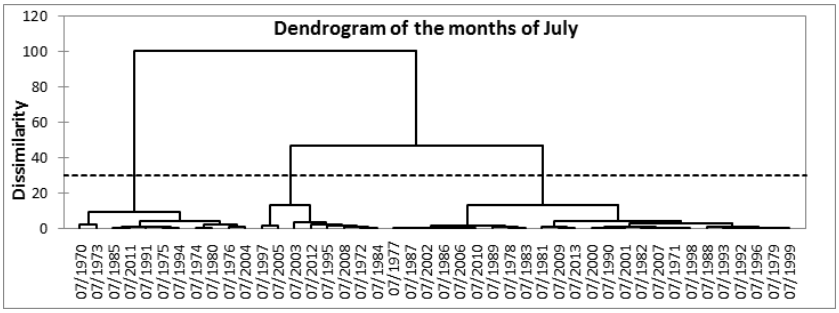


Figure 6. Growth models and the decrease in mean discharge from 1970 to 2013
(Data source: DGPRE)

The detection of the models is carried out from the means of the daily discharges of the period of high water of 1970 to 2013. The evolution of the mean daily discharges of the series makes it possible to indicate the model (figure 6), and taking into account the similarities, we can say all the years of the series follow it. In Figure 6, the model has been detected for both the rising up phase of the discharge and the falling down phase from the peak.



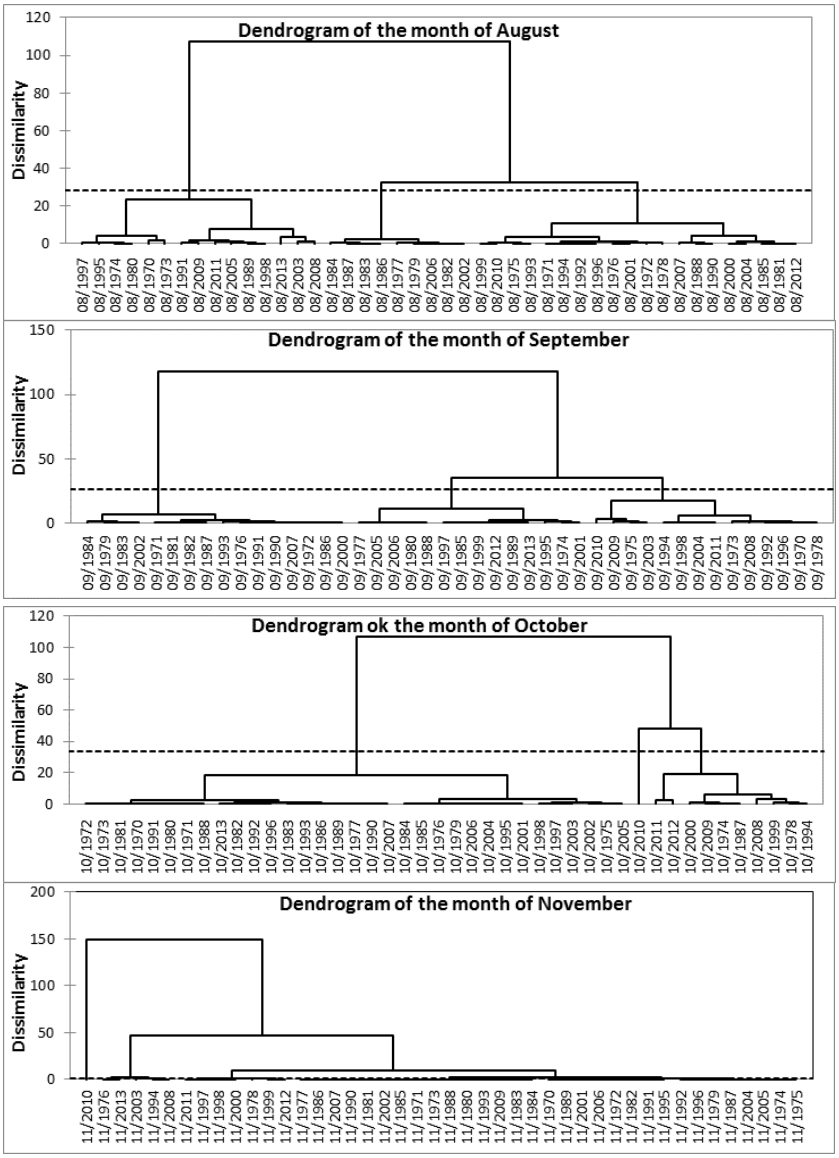


Figure 7. Dendrograms obtained after application of AHC from 1970 to 2013 (Data source: DGPRE)

The analysis of the ascending hierarchical classification is based on a tree diagram: the dendrogram. The latter, obtained by the application of the AHC, is a bottom-up hierarchical classification approach, which takes place by series of mergers of different years (1970-2013) for every month (July to November) into classes (figure 7). In these tree diagrams, the height of each U-shaped line indicates the distance between the different years for every month. For the growth and decay coefficient with respect to the peak of the discharge, the results are shown in table 7. If for the rise phase, the coefficients k are all positive, which indicates a growth of the discharges, on the other hand on the descent phase, they are negative, which is synonymous with a decrease in discharge rates. The rate of growth of discharges (k) is higher between the months of June and July due to the low discharge. On the other hand between July-August and August-September, k is certainly positive (which indicates an increase of the discharge) but weaker

because of the importance of the discharge. On the descent phase, the decay rate (k) noted between September and October and between October and November indicates the gradual decline in the discharge in the basin. This decrease in discharge is more significant between October and November, in relation to the end of the rainy season in the basin. The coefficients k thus indicate the real image of the rise and fall of the discharges respectively on each side of the peak of the annual discharge. The absolute values of k indicate whether the discharges are low or high. Positive values mean the discharge is increasing and negative that it is decreasing.

Validation of results based on the causal effect

In this work, we try to analyze the nature of the variation of the discharge of the basin during the period of high water through graphs (figure 8) representing the data of past discharges. Hydrological processes have shown a cause-and-effect relationship with rain events, since it is the climatic setting that determines the modalities of river discharge. Thus, the evolution of rainfall and temperatures over the high-water period, on a monthly scale, is shown in Figure 8, which shows a fairly similar model consistent with the past-discharge model, which shows the importance of precipitated water slides and their contribution to the water slides that have passed. The evolution of the rain is accompanied by a second-order polynomial regression, whose equation is displayed at the top.This regression better reflects the monthly evolution of the rain in the basin at the Kédougou station) from 1970 to 2013.

Table 7. Mean values of coefficients (k) of growth (July, August, and September) and decay (September, October, and November) compared to discharge peak (1970-2013) per five years
(Data source: DGPRE)

Periods	Rise phase			Descent phase	
	July	**August**	**September**	**October**	**November**
1970-71 / 1974-75	0.87	0.64	-0.03	-0.47	-0.51
1975-76 / 1979-80	0.91	0.49	0.27	-0.27	-0.52
1980-81 / 1984-85	1.00	0.49	0.06	-0.42	-0.53
1985-86 / 1989-90	1.19	0.63	0.17	-0.40	-0.57
1990-91 / 1994-95	1.39	0.50	0.15	-0.36	-0.46
1995-96 / 1999-00	0.78	0.73	0.14	-0.36	-0.56
2000-01 / 2004-05	1.37	0.52	0.15	-0.37	-0.47
2005-06 / 2009-10	1.02	0.64	0.15	-0.26	-0.55
2010-11 / 2013-14	0.98	0.69	0.29	-0.16	-0.30

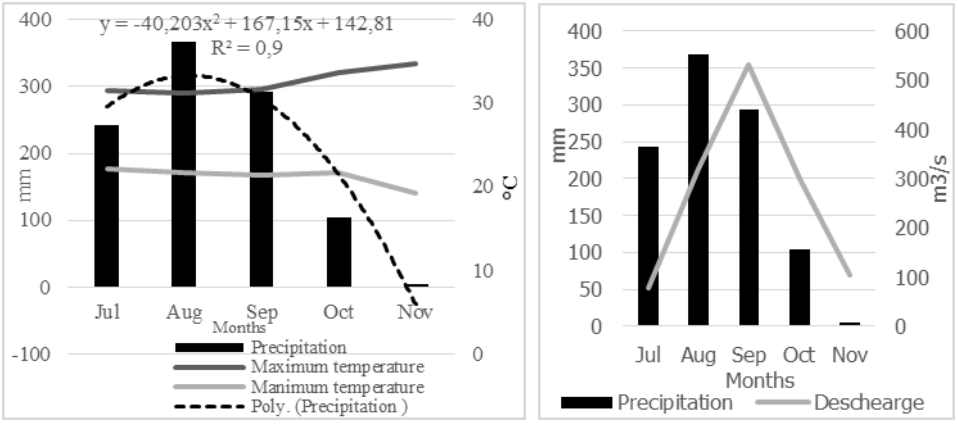


Figure 8. Mean monthly maximum and minimum temperatures and precipitation (a) and precipitation and discharge (b) at Mako (1970-2013)
(Data source: DGPRE)

Rainfall distribution could be observed throughout the high water period. In this paper, the following observations strongly validate the obtained models: 1) the evolution of past discharges is similar to hydrographs on each year during the high water period; 2) the evolution of discharge tends to increase during periods of high rainfall; 3) The rainfall - discharge pattern shows more similarities, despite the lag of one month between the peak of the rain and that of the discharge. However, the methods applied have disadvantages (Creusier and Biétry, 2014). For the k-means method, the disadvantage is that it does not make it possible to discover what can be a coherent number of classes, nor to visualize the proximity between classes or objects (Rakotomalala, undated; Lelu, 2008; Creusier and Biétry, 2014). For the AHC, the main disadvantage is that it requires the calculation of distances between individuals taken two by two. This is very quickly prohibitive as soon as the file size exceeds 1,000 individuals (Lerman, 2009; Creusier and Biétry, 2014). If several methods are proposed for the general problem of classification, each having its strengths and weaknesses, the ascending hierarchical methods (AHC technique) are used more than the K-means methods in case of small data as it was the case for our study period (44 years period) because the complexity is very high. On the other hand, if run time problems arise, then the K-means methods are used.

For the application of the k-means method and AHC techniques, it has been found that the AHC technique is more accurate and its main advantage over other classification methods lies in this representation in the form of a tree that evidence additional information: the increase of the dispersion in a group produced by an aggregation. The user can then have an idea of the adequate number of classes by choosing the partition corresponding to the highest jump in the increase of the dispersion within the classes (Lerman, 2009; Creusier and Biétry, 2014). This AHC technique makes it possible to highlight a "natural" grouping of a set of individuals described by characteristics (the variables). It offers a series of nested partitions represented in the form of trees called dendrograms. The algorithm proceeds by successive aggregations, starting from the most fragmentary partition, an individual is equal to a class, until the trivial partition, the grouping of all the individuals in one and only one class. In addition, the k-means and AHC methods are therefore complementary.

In many fields of the social sciences, we are led to form groups homogeneous within them and which differ sufficiently from each other. This is the purpose of the classification methods of which the k-means and AHC methods method is part (Creusier and Biétry, 2014). These techniques are currently one of the most used and most effective in data analysis. In fact, they make it possible to partition a finite population of elements into a number K (integer) of homogeneous classes. It is useful to note that their algorithms are very efficient in terms of execution time, but they suffer from the problem of dependence of the results on the choices made during the initialization. We can expand our work, trying to compare our results with other versions of K-means and AHC methods, working on other unsupervised classification algorithms, and even the supervised ones (Masmoudi, 2017).

CONCLUSION

The discovered models are more similar to discharge models. The comparison of hydrographs and precipitation during the same period was made and it is proved that the discharge models were more similar in this period. Our future studies could focus on other periods (low water periods), or a much longer data series (eg over 44 years) for a more complete study of the hydrological behavior of the Gambia Basin at the Mako hydrometric station in particular and even on other stations of the basin.

As we pointed out in the introduction, there are several classification methods. For better results and to overcome the disadvantages of using a single method, recent studies encourage the use of so-called mixed methods. A mixed method is a method that groups together several classification algorithms. In this study, the classification algorithm includes two different analyzes: the k-means method and AHC techniques. If several methods are proposed for the

general problem of classification, each having its strengths and weaknesses, the ascending hierarchical methods (AHC technique) are used more than the K-means methods in case of small data as it was the case for our study period (44 years period) because the complexity is very high. On the other hand, if run time problems arise, then the K-means methods are used.

The k-means method has allowed us not only to reduce the size of the data, but also to define a subspace in which our data will be easily represented. The AHC techniques made it possible to carry out the Hierarchical Ascendant classification proper. These techniques are currently one of the most used and most effective in data analysis. In fact, they make it possible to partition a finite population of elements into a number K (integer) of homogeneous classes.

From the hydrologic point of view, this classification can also be used as a tool to reduce the number of hydrological parameters to be used in the framework of design projects. We can identify the number of months in the high water period with the hydrological parameters close to the median of each class or, using the prediction model obtained by the discriminant analysis. It is however essential to evaluate the error related to the use of a set of parameters because the months on the period of high water belonging to a class do not have all the same hydrological parameters. Therefore, the uncertainty in the modeled temperature profile should be estimated when using the median as the metric that represents the class by varying the medians. It will also be possible to use the hydrological parameters of the other months over the high-water period, always with the same calibration, to evaluate the error by comparing the simulated temperature profiles.

We can expand our work, trying to compare our results with other versions of K-means and AHC methods, working on other unsupervised classification algorithms, and even the supervised ones.

REFERENCES

Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining Associations between Sets of Items in Massive Databases, *International Conference on Management of Data (ACM SIGMOD'93),* Washington D.C.

Agrawal, R., & Srikant, R. (1995). Mining Sequential Patterns. *11th International Conference on Data Engineering* (DE'95), Taipei, Taiwan.

Aydin, I., Karakose, M., & Akin, E. (2009). The prediction algorithm based on fuzzy logic using time series data mining method. *World Academy of Science, Engineering and Technology*, 51(27), 91-98.

Creusier, J., & Biétry, F. (2014). Analyse comparative des méthodes de classifications. *RIMHE: Revue Interdisciplinaire Management, Homme Entreprise,* (1), 105-123.

Crié, D. (2003). De l'extraction des connaissances au Knowledge Management. *Revue française de gestion*, (5), 59-79.

Ding, H., Trajcevski, G., Scheuermann, P., Wang, X., & Keogh, E. (2008). Querying and mining of time series data: experimental comparison of representations and distance measures. *Proceedings of the VLDB Endowment*, 1(2), 1542-1552.

Dione, O. (1996). *Evolution climatique récente et dynamique fluviale dans les hauts bassins des fleuves Sénégal et Gambie*. Thèse de doctorat, Université Lyon 3 Jean Moulin, 477 p.

Faye, C. (2018). Precipitation trends in the Gambia River basin (Senegal) for the period 1971-2010. *Bulletin of the Serbian Geographical Society*, 98 (2), 45-57.

Faye, C. (2018). Climatic Variability and Hydrological Impacts in West Africa: Case of the Gambia Watershed (Senegal). *Environmental and Water Sciences, public Health and Territorial Intelligence Journal,* 2(1), 54-66.

Faye, C., Ndiaye, A., & Mbaye, I. (2017). Une évaluation comparative des séquences de sècheresse météorologique par indices, par échelles de temps et par domaines climatiques au Sénégal. *Journal of Water and Environmental Sciences*, 1(1), 11-28.

Gupta, A., & Chaturvedi, S. K. (2013). Real Time Prediction System of Discharge of the Rivers using Clustering Technique of Data Mining. *International Journal of Engineering Research and Development*, 9 (4), 12-24

Jayanthi, R. (2007). Application of data mining techniques in pharmaceutical industr. *JATIT*, 61-67.

Lamagat, J.P. (1989). *Monographie hydrologique du fleuve Gambie Collection M&M*. ORSTOM-OMVG, 250 p.

Larose, D. T. (2005). *Discovering Knowledge in Data: An Introduction to Data Mining*, John Wiley & Sons, Inc., Hoboken, New Jersey.

Lelu, A. (2008). *La méthode de classification non-supervisée K-means axiales*. Rapport Technique, 12 p

Lerman, I.C. (2009). Analyse logique, combinatoire et statistique de la construction d'une hiérarchie implicative; niveaux et nœuds significatifs, *Revue Mathématiques et sciences Humaines*, 16, 265-286.

Masmoudi, N. (2017). *Modèle bio-inspiré pour le clustering de graphes: application à la fouille de données et à la distribution de simulations* (Doctoral dissertation, Normandie).

Mishra, S., Dwivedi, V. K., Sarvanan, C., & Pathak, K. K. (2013). Pattern discovery in hydrological time series data mining during the monsoon period of the high flood years in Brahmaputra River basin. *International Journal of Computer Applications*, 67(6), 7-14.

Mishra, S., Saravanan, C., Dwivedi, V.K., & Pathak, K.K. (2014). Discovering Flood Recession Pattern in Hydrological Time Series Data Mining during the Post Monsoon Period. *International Journal of Computer Applications*, 90(8), 35-44.

Piatetsky-Shapiro G., & Frawley W. J. (1991). *Knowledge Discovery in Databases*. AAAI/MIT Press: Boston, MA,

Rakotomalala R. (non daté). *Méthodes des centres mobiles. Classification par partition, Les méthodes de réallocation.* Université Lumière Lyon 2, 31 p.

Rakthanmanon, T., Campana, B., Mueen, A., Batista, G., Westover, B., Zhu, Q., ... & Keogh, E. (2012, August). Searching and mining trillions of time series subsequences under dynamic time warping. *In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining (*pp. 262-270). ACM.

Sow, A. A. (2007). *L'hydrologie du Sud-est du Sénégal et de ses Confins guinéo-maliens: les bassins de la Gambie et de la Falémé (Doctoral dissertation, Thèse (PhD).* Université Cheikh Anta Diop de Dakar), 1232 p.

Submitted: May 05, 2019 | Revised: November 17, 2019 | Accepted and published online December 13, 2019