Journal of Fundamental and Applied Sciences

ISSN 1112-9867

Available online at http://www.jfas.info

Research Article

SYSTEM FOR PREDICTING DISCHEARGES OVER THE HIGH WATER PERIOD THROUGH THE CLASSIFICATION TECHNIQUES DATA: CASE OF THE GAMBIA RIVER BASIN OF MAKO

C. Faye

Département de Géographie, U.F.R. Sciences et Technologies, UASZ, Laboratoire de Géomatique et d'Environnement, BP 523 Ziguinchor (Sénégal)

Received: 27 January 2019 / Accepted: 24 April 2019 / Published online: 01 May 2019

ABSTRACT

This article examines the trend of flow during the high water period (from July till November) in the basin of Gambia measured at the Mako station of over 2004-2013 period. Methodology consisted at first in calculation and in standardization of data by the method of z-score of some statistical parameters (average, 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 Time Series Data Mining to cluster and discover the discharge patterns in terms of the autoregressive model.. From these methods, a forecast model has been developed for the discharge process on average over these years. This study presents basin flow dynamics in high water period from Time Series Data Mining technique.

Keywords: *data Mining, flow, forecast model, hydrological process, clustering; technics*

Author Correspondence, e-mail: cheikh.faye@univ-zig.sn

doi: http://dx.doi.org/10.4314/jfas.v11i2.22

1. INTRODUCTION

Au Sénégal, la collecte des données climatologiques et hydrologiques est gérée respectivement par l'Agence nationale de l'aviation civile et de la météorologie (ANACIM) et la Direction de la Gestion et de la Planification des ressources en Eau (DGPRE). Ces données sont très utiles dans la recherche, l'analyse des tendances historiques et les prévisions futures. Avec le développement de la technologie des bases de données, diverses techniques d'analyse de données et d'extraction des connaissances sont utilisées pour découvrir les connaissances issues de ces données collectées dans diverses organisations telles que l'hydrologie, l'environnement, la météorologie, etc [1]. Aujourd'hui, le développement de la technologie de l'information a généré d'énormes quantités de bases de données couvrant divers domaines la science et la technologie. L'exploration de données est largement appliquée dans les domaines de la recherche. Trouver des règles d'association, des modèles séquentiels, la classification et le regroupement de données sont des tâches typiques impliquées dans le processus d'exploration de données. Les diverses techniques de classification visent toutes à répartir *n individus, caractérisés par p variables X1, X2, ..., Xp en un* certain nombre m de sous-groupes aussi homogènes que possible, chaque groupe étant bien différencié des autres [2]. Deux grandes techniques de classification existent : le partitionnement et la classification hiérarchique.

L'exploration de données désigne l'extraction ou l'extraction de connaissances à partir de grandes quantités de données. La méthodologie de l'exploitation des données de série temporelle suit le processus d'intégration retardé pour prédire les occurrences futures d'événements importants [3]. Ce cadre combine les méthodes de reconstruction d'espace de phase et d'exploration de données pour révéler des modèles cachés prédictifs d'événements futurs dans des séries temporelles non linéaires et non stationnaires. L'exploration de données de série temporelle est consacrée au développement et à l'application de nouvelles techniques et modèles informatiques pour l'analyse de bases de données temporelles de grande envergure. Le développement rapide de l'exploration de données offre une nouvelle méthode pour la recherche en l'hydrologie et la gestion des ressources en eau. En hydrologie, les données exploitées sont en grande partie des données hydroclimatiques qui prennent généralement la

structure des séries chronologiques. Les bases de données hydrologiques sont des ensembles de diverses valeurs record qui divergent avec le temps. Des recherches basées sur la théorie de l'exploration de données et sur les techniques hydrologiques sont nécessaires pour analyser les bases de données hydrologiques, climatologiques et sédimentaires pour différents types d'étude.

La capacité de construire un modèle de prédiction réussi dépend des données passées. En s'appuyant sur des succès et échecs passés, l'exploration de données peut prédire ce qui va se passer ensuite (prédiction future) [4]. L'exploration de données est un domaine multidisciplinaire qui permet de travailler dans des domaines tels que l'apprentissage automatique, la statistique, la reconnaissance de formes, la récupération d'informations, les réseaux neuronaux, la visualisation de données… [5]. Au cours de la dernière décennie, de nombreuses recherches sont menées pour extraire des connaissances à partir de données historiques cachées. La prévision hydrologique en temps réel constitue un grand défi pour la communauté scientifique car elle permet de sauver la vie, l'infrastructure et l'économie de n'importe quel pays [1]. Dans le domaine de l'exploration des données hydrologiques, diverses recherches et techniques sont réalisées pour l'extraction de connaissances à partir de données historiques [6]. Certains d'entre eux qui sont pertinents pour cette étude sont la découverte de modèles dans l'exploitation de données hydrologiques pendant la période de hautes eaux dans le bassin de la Gambie.

L'exploration de données par séries chronologiques combine les champs de l'analyse des séries chronologiques et des techniques d'exploration de données [7]. Cette méthode crée un ensemble de processus qui révèlent des modèles temporels cachés qui sont caractéristiques et prédictifs des conséquences de séries chronologiques. L'objectif principal de cette étude est de développer une application d'exploration de données en utilisant la technique de l'information moderne et de découvrir les informations cachées ou les modèles derrière les données hydrologiques historiques pendant la période de hautes eaux dans le bassin de la Gambie à Mako. Les outils d'exploration de données comme la recherche de similarité, le regroupement des k-moyens et le modèle de Classification Ascendante Hiérarchique (CAH) sont utilisés sur cet article.

2. DONNÉES ET MÉTHODES

2.1. Zone d'étude

Pour ce travail, le fleuve Gambie a été sélectionnée en raison de la forte variabilité des ressources en eau. Son bassin, d'une superficie de près de 77 100 km, s'étend en latitude, du 11°22 Nord (dans le Fouta-Djalon) au 14°40 Nord (dans le Ferlo sud-oriental) et, en longitude, du 11°13 Ouest (Fouta-Djalon) au 16"42 Ouest (Banjul, embouchure) [8]; [9]; [10]. La longueur du fleuve à la station de Mako est de 2562 km. La station hydrométriquee de Mako est située sur le cours principal du fleuve. Pour ce travail, les données de débits journaliers ont été prises sur une période de 10 ans (2004-2013). A la station de Kédougou, la température maximale moyenne est de 30°C, la température minimale de 25°C et la pluviométrie de 1000 mm (figure 1).

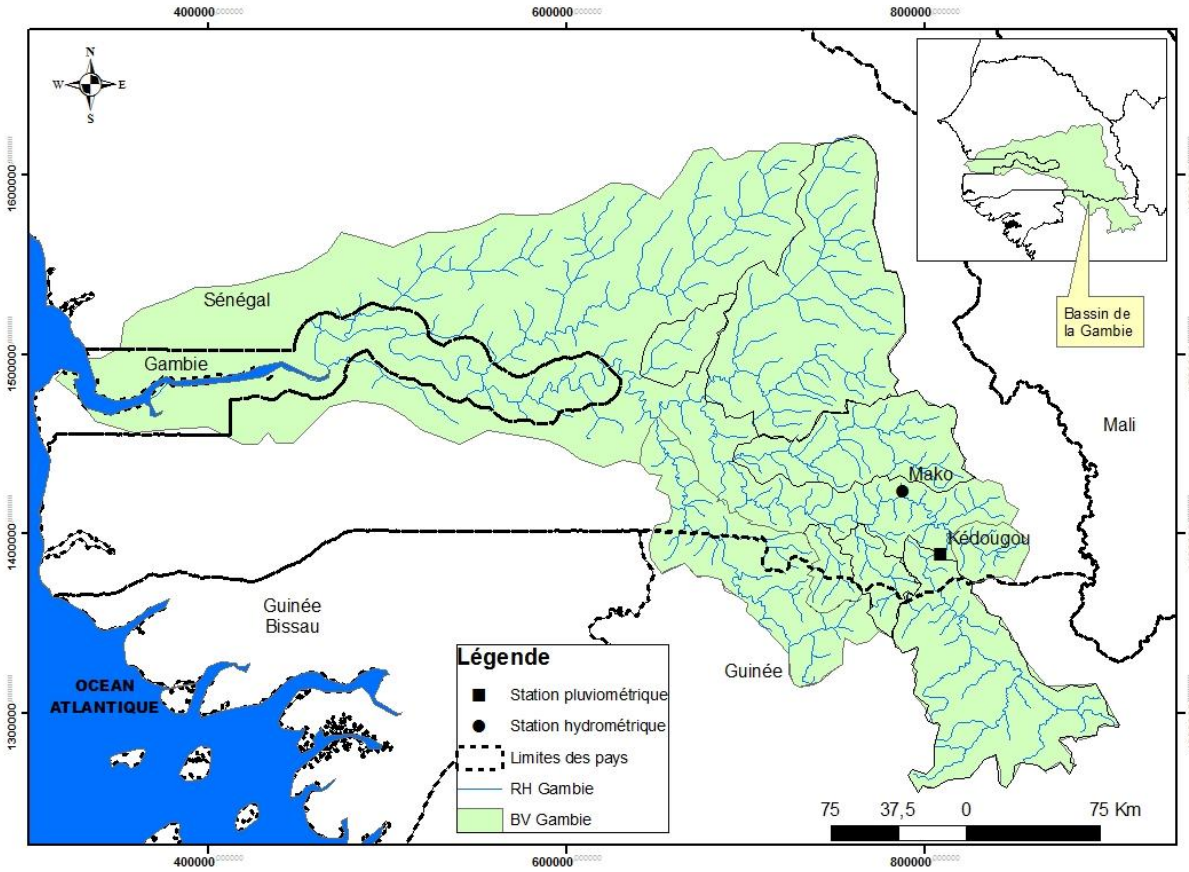

Fig.1. Situation du bassin versant de la Gambie

2.1. Données

Pour ce travail, les données journalière de débit écoulé de la station de Mako et des données climatologiques (pluviométrie et température) de Kédougou ont été recueillies au auprès de la

Direction de le Gestion et de la Planification des Ressources en Eau. Les données journalières, conformément aux exigences des méthodes utilisées, ont été converties sous forme de moyennes mensuelles. Sur les valeurs mensuelles de débit, une série 10 ans est choisie (2004-2013) et les tests effectués sur une période de hautes eaux (juillet-novembre)

2.2. Méthodes

2.2.1. Analyse statistique, normalisation des données

Les cinq paramètres statistiques (Qmoyenne, Qmaximum, Qminimum Qétendue et Q écartype) ont été calculés sur chaque mois avec les données de débit.

Afin d'avoir une analyse efficace des données sur la série et la période considérée, les paramètres calculés ont été normalisées en utilisant la technique de Z-score à travers la formule suivante

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

Avec x_i qui est la valeur du mois, x_m la moyenne de la série et σ l'écartype de la série.

La normalisation a été nécessaire pour éviter que les résultats de l'étude soient affectés par les grandes variations des données.

2.2.2. Segmentation des données

Pour la segmentation des données, l'analyse de l'hydrogramme du bassin et des coefficients mensuels de débits (CMD) (Tableau 1) à la station de Mako sur la période 2004-2013 permet de le divisée en 3 segments : basses eaux (Mai-juillet), hautes eaux (août-octobre) et basses eaux (novembre-décembre). Pour cette étude, bien que les mois de Juillet et Novembre soient des mois de basses eaux (CMD<1), ils sont utilisés dans la période (dite de hautes eaux) sur laquelle les tests sont appliqués. Ce choix peut s'expliquer par l'importance de leurs débits écoulés.

Tableau 1. Débits écoulés et CMD mensuels à Mako de 2004 à 2013

Date	M	J	J	A	S	O	N	D	J	F	M	A	
Débits	0,16	6,74	77,2	321	530	299	104	30,7	12,34	6,17	3,39	0,83	**116**
CMD	0,00	0,06	0,67	2,77	4,57	2,57	0,89	0,26	0,11	0,05	0,03	0,01	1
CMD	Basses eaux			Hautes eaux					Basses eaux				

Le débit des cours d'eau change progressivement avec la diminution de la pluviométrie. Les différents climats dans le bassin provoquent donc des processus d'écoulement différents. Ainsi, donc pour une meilleure étude des processus de l'écoulement, l'étude du cadre climatique est fondamentale. Sur cet article, la période de hautes eaux est choisie et les données statistiques qui y sont obtenues ont été soumises aux tests de classification.

2.2.3. Classification par la méthode k-moyennes (k-means clustering)

La classification k-means est une méthode itérative très efficace pour trouver des groupes sphériques dans les bases de données de taille petites et moyenne. Son application nécessite plusieurs fois les calculs dans le but de ne retenir que la solution la plus optimale pour le critère choisi. Pour la première itération on choisit un point de départ qui consiste à associer le centre des k classes à k objets (pris au hasard ou non). On calcule ensuite la distance entre les objets et les k centres et on affecte les objets aux centres dont ils sont les plus proches. Puis on redéfinit les centres à partir des objets qui ont été affectés aux différentes classes. Puis on réaffecte les objets en fonction de leur distance aux nouveaux centres, jusqu'à ce que la convergence soit atteinte.

2.2.4. Mesure de similarité

La recherche de similarité dans l'analyse des séries chronologiques est l'un des domaines de développement les plus rapides et exigeantes dans l'exploration de données. Contrairement à des requêtes de base de données normales, qui trouvent les données correspondant à la requête donnée exactement, une recherche de similarité trouve des séquences de données qui ne diffèrent que légèrement de la séquence de requête donnée. On peut le classer en deux catégories:

La catégorie « toute correspondance »: Dans ce type d'appariement des données de séries chronologiques doit être de longueur égale.

La catégorie « correspondance subséquente »: Dans cette catégorie, une séquence de requêtes X et une séquence plus longue Y ont été prises. L'objectif est d'identifier la séquence en Y, en commençant par Yi, qui a les meilleurs correspondances de X, et de rendre compte de son décalage au sein de Y. La principale difficulté consiste à définir une mesure de similarité [11]. Pour l'analyse de la similarité des données de séries chronologiques, la distance

euclidienne est généralement utilisée comme une mesure de similarité. Compte tenu de deux séquences, X=(x_1…, x_n) et Y=(y_1…, y_n) avec n = m, la distance euclidienne est défini comme suit :

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

Le DTW (Dynamic Time Warping) Algorithme : Le DTW est un algorithme de mesure de similitude optimale entre deux séquences de données. Les données des séries varient non seulement sur les amplitudes du temps mais aussi en fonction de la progression du temps que les processus hydrologiques peuvent révéler des vitesses différentes en réponse à différentes conditions environnementales. Un alignement non linéaire produit une mesure similaire, permettant des formes similaires pour correspondre même si elles sont hors de phase dans l'axe de temps [12]. Les séquences sont "déformés" non-linéaire dans la dimension du temps pour déterminer une mesure de leur similitude indépendante de certaines variations non linéaires dans la dimension temporelle. Pour trouver le meilleur alignement entre les séquences de temps X et Y on a besoin de trouver le chemin à travers la grille.

2.2.5. Classification Ascendante Hiérarchique (CAH)

Pour identifier le modèle de débit à partir des séries de données correspondant, chaque période hydrologique obtenue après la segmentation du k- means a été prise, puis l'analyse d'un modèle de débit dans chacune des périodes a été faite. Les analyses ont impliqué des techniques de classification hiérarchique sur les 5 ans avec l'observation des données moyennes de débit pour les mois choisis sur la période hydrologique. Maintenant, les données des séries chronologiques de débit du centre du groupe est le modèle parce que tous les autres objets dans un groupe particulier montrent après une similitude au centre seulement. Ainsi, le centre du groupe peut être considéré comme le modèle de débit.

Dans notre analyse, le modèle de débit (données de débit pour les mois de la période hydrologique) d'un an (entre 10 ans) est regroupé en plusieurs groupes. Mais l'année qui formait le centre du groupe formait également le modèle avec ses données de débit pour les mois de la période. Toutes les autres années dans le groupe atteignent l'appartenance du

groupe parce qu'il y avait similitude avec l'année représentant le centre afin qu'elles puissent suivre le modèle. Une fois le centre (l'année) dans le groupe est obtenu, la représentation graphique des données de débit de cette année correspondant au débit dans les mois, le long de l'axe x, donnerait le modèle.

2.2.6. Calcul de la moyenne glissante sur les données de débits mensuels normalisés

Une moyenne glissante, également appelée moyenne mobile, est un type de filtre à réponse impulsionnelle finie utilisée pour analyser un ensemble de points de données en créant une série de moyennes des différents sous-ensembles de l'ensemble complet des données. Une moyenne mobile est couramment utilisé avec des données de séries chronologiques pour lisser les fluctuations à court terme et de mettre en évidence les tendances à long terme ou cycles. Le seuil entre le court terme et à long terme dépend de l'application, et les paramètres de la moyenne mobile seront fixés en conséquence. La formule de la moyenne mobile simple est donnée ci-dessous:

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

Avec x_t qui est la moyenne mobile simple, k les observations, t le point de données par rapport au temps et n le nombre des points de données.

2.2.7. Coefficient de croissance et de décroissance par rapport au pic de l'écoulement

De part et d'autre du mois du maximum de l'écoulement (septembre), il semble nécessaire de calculer le coefficient de croissance et celui de décroissance de l'écoulement pour valider le modèle de débit du cours d'eau avant et après le pic. En hydrologie, le coefficient de croissance ou de décroissance de l'écoulement (k) est habituellement exprimé dans la décroissance exponentielle suivante:

$$\frac{Q}{Q_0} = e^{-kt}\,; \qquad donc \qquad \log(\frac{Q}{Q_0}) = \log(e^{-kt})$$

En appliquant le logarithme naturel des deux côtés et en déduisant k tout en sachant que t est ici égal à 1 (un mois), on obtient:

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

Pour asseoir la méthodologie soulevée, l'outil Xlstat, qui contient la mise en œuvre de divers

algorithmes de classification (comme K-means, CAH, etc., et d'autres techniques d'exploration de données), est utilisé. Pour ce travail, les données de débits journaliers sur la période de juillet à novembre sur 10 ans (de 2004 à 2013), sont utilisées. Xlstat est donc utilisé pour faire une classification et trouver des classes.

3. RESULTS AND DISCUSSION

3.1. La méthode k-moyennes

Après la division de la période hydrologique en trois segments et le choix de la période de hautes eaux (de juillet à novembre) pour cette étude, les paramètres statistiques (moyennes, maximum, minimum, étendue et écartype) obtenus et standardisés sur la période 2004-2013 ont été soumis à un regroupement des K-moyennes. Ainsi, un total de 50 (5 mois sur chacune des 10 années) cas (mois) basés sur les 5 paramètres à partir des données de de la période choisie a été utilisé et soumis à ce regroupement. Le Tableau 2 indique une classification des différents objets (mois) suivant un nombre de 5 classes.

Tableau 2. Classes d'affectation par objet après application de K-means

Observations	Classes	Distance au barycentre	Observations	Classes	Distance au barycentre	Observations	Classes	Distance au barycentre
juil-04	1	1,3843	sept-07	2	1,5795	nov-10	3	0,3407
août-04	2	1,0043	oct-07	1	1,0198	juil-11	2	1,5554
sept-04	3	0,6764	nov-07	2	0,6214	août-11	1	0,6738
oct-04	3	0,9361	juil-08	2	1,1114	sept-11	3	1,1809
nov-04	3	0,4217	août-08	2	1,0565	oct-11	5	1,1538
juil-05	3	0,3262	sept-08	4	0,8171	nov-11	3	0,6249
août-05	3	1,1507	oct-08	2	0,8898	juil-12	2	0,7942
sept-05	3	0,7624	nov-08	3	0,3221	août-12	5	1,1535
oct-05	3	0,3304	juil-09	3	0,6014	sept-12	3	1,2332
nov-05	3	0,3101	août-09	3	0,4510	oct-12	3	0,4783
juil-06	2	0,9560	sept-09	4	1,6355	nov-12	2	0,6250
août-06	1	1,2687	oct-09	4	1,2620	juil-13	3	1,1115
sept-06	3	2,0021	nov-09	5	1,5214	août-13	3	0,9670
oct-06	3	0,4816	juil-10	5	0,8000	sept-13	3	0,4665
nov-06	3	0,2404	août-10	2	0,7753	oct-13	1	0,5307
juil-07	4	1,0682	sept-10	2	0,8801	nov-13	4	1,7255
août-07	4	1,3867	oct-10	2	1,1036			

A partir d'une typologie (segmentation), cette méthode d'analyse de données a permis d'obtenir une représentation schématique simple du tableau de données de départ complexe. Cela s'est traduit par une partition des n individus (mois) dans des classes, définies par l'observation de p variables (moyenne, maximum, minimum, étendue et écartype). Le Tableau 3 donne les barycentres des classes (il s'agit des coordonnées des barycentres des classes pour les différents paramètres) et les résultats par classe (il s'agit du nombre d'objets, de la variance intra-classe, de la distance minimale au barycentre, de la distance maximale au barycentre et de la distance moyenne au barycentre).

Tableau 3. Barycentres des classes après l'application de K-means

Classes	Moy	Max	Min	Eten	Ecart	Nbre	Variance intra-classe	Distance min au barycentre	Distance moy au barycentre	Distance max au barycentre
1	0,3319	-0,5314	1,1898	-0,9389	-0,8204	5	1,3253	0,5307	0,9754	1,3843
2	-0,0540	0,5032	-0,3845	0,7191	0,4777	13	1,1651	0,6214	0,9963	1,5795
3	-0,6684	-0,7918	-0,5582	-0,6346	-0,6166	22	0,7016	0,2404	0,7007	2,0021
4	1,6078	1,1354	1,8487	0,3203	0,4018	6	2,1953	0,8171	1,3158	1,7255
5	1,0252	1,6803	0,0597	1,8463	2,2612	4	1,8721	0,8000	1,1572	1,5214

Selon la répartition des cas dans les classes, le processus annuel de débit pourrait être obtenu sous forme de classes séparées. Pour ce travail où seule la période de hautes eaux est utilisée, l'algorithme k-means propose une représentation graphique des classes (figure 2) est une courbe d'évolution en fonction des paramètres statistiques.

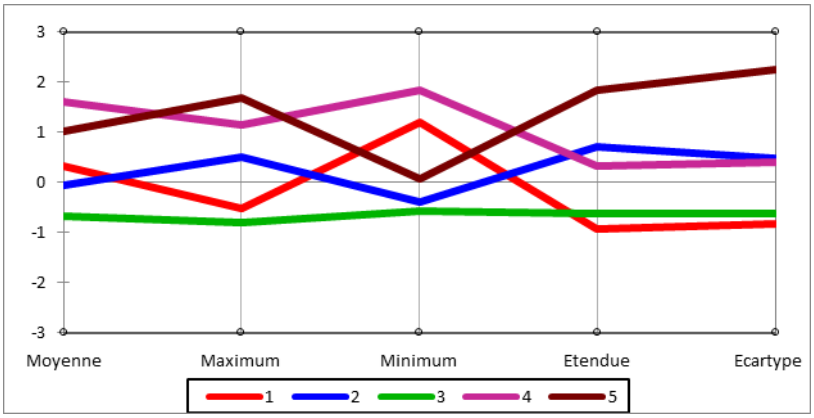

Fig.2. Représentation graphique des classes obtenues par l'application de K-means

3.2. Classification Ascendante Hiérarchique (CAH)

Tout comme la classification k-menans, les techniques de Classification Ascendante Hiérarchique (CAH) et le critère Wards sont appliqués sur la période de hautes eaux (juillet à novembre) sur les 10 années retenues (2004-2013) pour l'analyse des modèles. La CAH est particulièrement utile pour trouver des modèles cachés dans les données multidimensionnelles. Comme il s'agit d'un schéma d'apprentissage non supervisé, le nombre de classes peut être grand ou petit à certains moments. Le rôle principal de la CAH est d'identifier des classes ou des groupes de séries de débits qui sont semblables. En appliquant l'algorithme CAH sur l'ensemble de données, trois différentes classes ont été obtenues (Tableau 4).

Tableau 4. Classes d'affectation par objet après application de la CAH

Observations	Classes	Observations	Classes	Observations	Classes	Observations	Classes	Observations	Classes
juil-04	1	juil-06	2	juil-08	2	juil-10	2	juil-12	2
août-04	2	août-06	1	août-08	2	août-10	2	août-12	3
sept-04	1	sept-06	1	sept-08	3	sept-10	2	sept-12	1
oct-04	1	oct-06	1	oct-08	2	oct-10	2	oct-12	1
nov-04	1	nov-06	1	nov-08	1	nov-10	1	nov-12	2
juil-05	1	juil-07	3	juil-09	1	juil-11	2	juil-13	1
août-05	1	août-07	3	août-09	1	août-11	1	août-13	1
sept-05	1	sept-07	2	sept-09	3	sept-11	1	sept-13	1
oct-05	1	oct-07	1	oct-09	3	oct-11	2	oct-13	1
nov-05	1	nov-07	2	nov-09	3	nov-11	1	nov-13	3

Pour l'analyse des classes de la CAH, les barycentres des classes, la variance intra-classe et les distances (minimale, maximale et moyenne) au barycentre sont calculés (Tableau 4) et représentés sous forme de graphique (figure 3).

Tableau 5. Barycentre des classes après l'application de la CAH

Classes	Moy	Max	Min	Eten	Ecart	Nbre	Variance intra-classe	Distance min au barycentre	Distance moy au barycentre	Distance max au barycentre
1	-0,4832	-0,7436	-0,2345	-0,6910	-0,6543	27	1,4378	0,2795	0,9990	3,0291
2	-0,0198	0,6363	-0,3791	0,8539	0,7010	15	1,6165	0,8302	1,1727	2,0250
3	1,6678	1,3164	1,5024	0,7309	0,8940	8	3,7768	0,9167	1,6802	3,0402

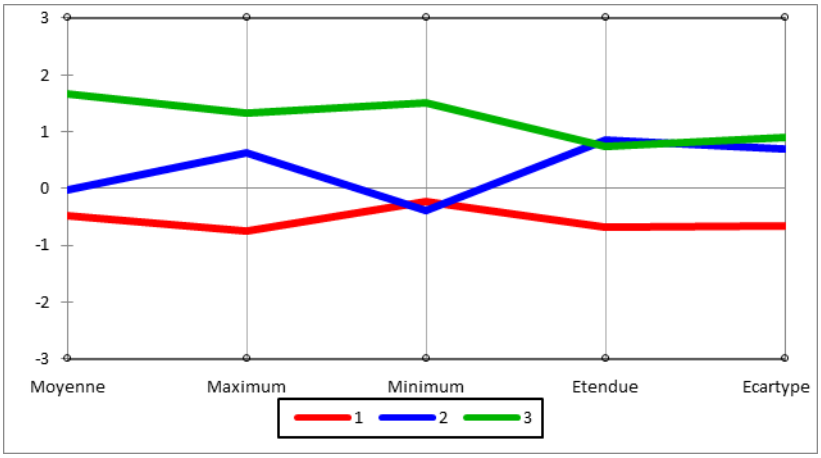

Fig.3. Représentation graphique des classes obtenues par application de la CAH

3.3. Analyse des similarités, détection de modèle et coefficient de croissance et de décroissance par rapport au pic de l'écoulement

L'observation des similarités est aussi faite sur les données de la période de hautes eaux de 2004 à 2013. Cette technique est appliquée pour indiquer des débits similaires entre les mois et années de la série. Cela s'explique par le fait que les séries temporelles de débits varient non seulement en termes d'amplitudes d'expression, mais aussi en termes de progression temporelle, car le débit du cours d'eau peut s'écoule à des rythmes différents en fonction de différentes conditions naturelles ou à différents endroits dans le bassin à des moments différents [1]. Pour cette étude, la distance entre les objets et les k centres indiquée par la technique de la classification k-means est utilisée comme la matrice de similarité pour la période de hautes eaux (Tableau 6). Ces distances entre les objets centraux représentent les distances euclidiennes entre les objets centraux des classes pour les différents descripteurs. La matrice de similarité a permis de comparer les débits mensuels des dix années de la série. Par exemple, les distances entre les objets centraux indiquent de fortes similarités entre le mois d'octobre 2013 et le mois de juillet 2010.

Tableau 6. Distance entre les objets et les k centres pour la période de hautes eaux selon la k-means

	oct-13	nov-07	nov-06	sept-08	juil-10
oct-13	0	2,3519	2,2733	2,4944	4,629
nov-07	2,3519	0	1,7387	1,8399	2,5188
nov-06	2,2733	1,7387	0	3,3191	4,1046
sept-08	2,4944	1,8399	3,3191	0	2,6885
juil-10	4,629	2,5188	4,1046	2,6885	0

La détection des modèles est effectuée à partir des moyennes des débits journaliers de la période de hautes eaux de 2004 à 2013. L'évolution des débits moyens journaliers de la série permet d'indiquer le modèle standard (figure 4), et compte tenu des similarités, on peut dire toutes les années de la série la suivent. Sur la figure 4, le modèle standard a été détecté à la fois pour la phase de montée vers le pic de l'écoulement et la phase de descente à partir du pic. Le modèle standard montre les tendances futures de l'écoulement pendant la période de hautes eaux.

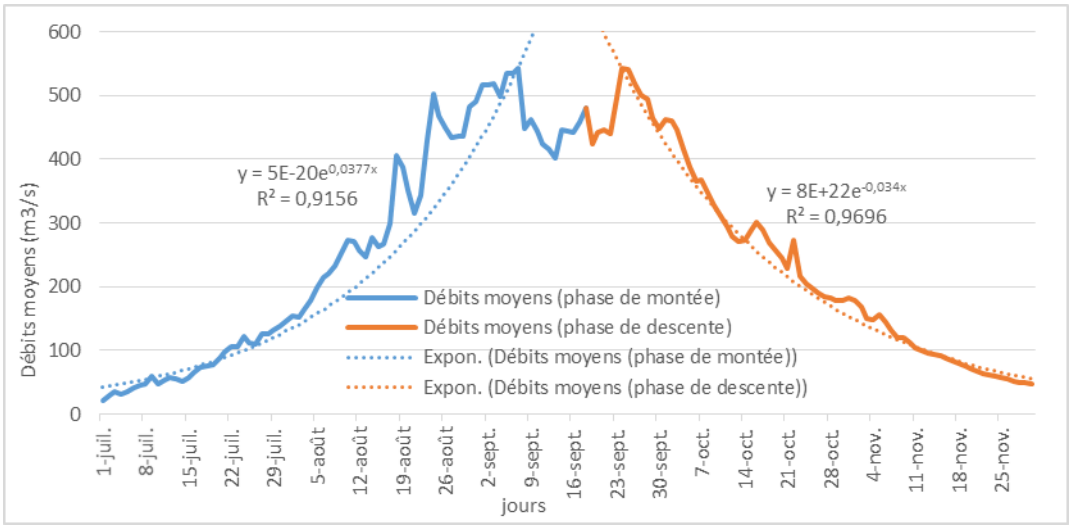

Fig.4. Modèles de croissance et celui de décroissance de l'écoulement moyen de 2004 à 2013

L'analyse de la classification ascendante hiérarchique est basée sur un diagramme en arbre : le dendogramme. Ce dernier, obtenu par l'application de la CAH, est une approche ascendante de classification hiérarchique, qui se déroule par séries de fusions des différents individus

(mois) en classes (figure 5). Dans ce diagramme en arbre, la hauteur de chaque ligne en forme de U indique la distance entre les différents individus (mois).

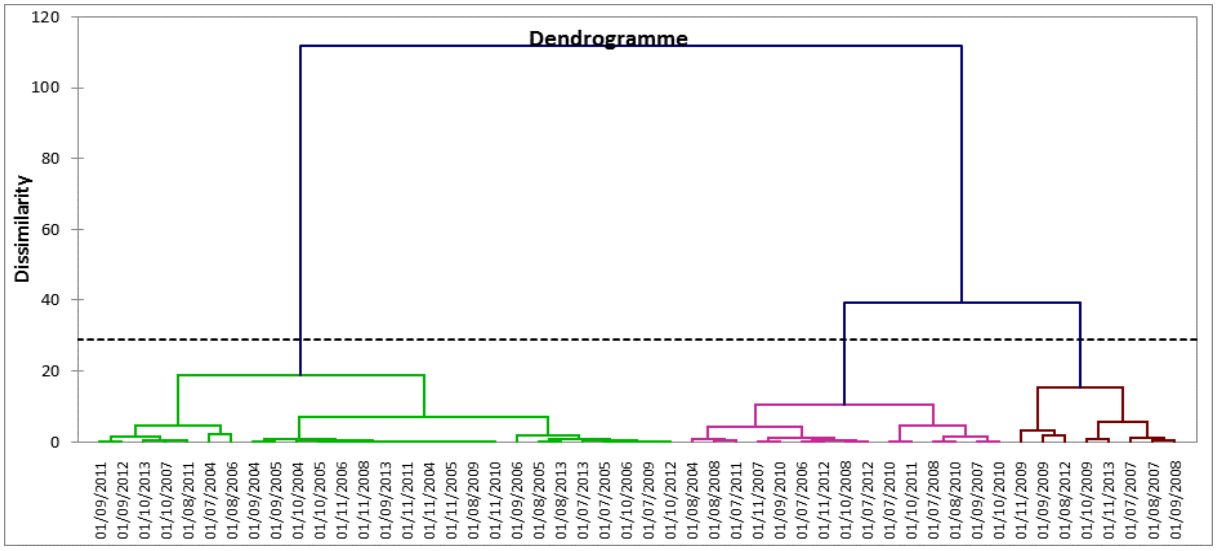

Fig.5. Dendrogrammes obtenus après l'application de la CAH

Pour le coefficient de croissance et de décroissance par rapport au pic de l'écoulement, les résultats sont indiqués dans le Tableau 7. Si pour la phase de montée, les coefficients k sont tous positifs, ce qui indique une croissance des débits, par contre sur la phase de descente, ils sont négatifs, ce qui est synonyme d'une décroissance des débits. Le coefficient de croissance des débits (k) est plus élevé entre les mois de juin et juillet compte tenu de la faiblesse de l'écoulement. Par contre entre juillet-août et août-septembre, k est certes positif (ce qui indique une hausse de l'écoulement) mais plus faible du fait de l'importance de l'écoulement. Sur la phase de descente, le coefficient de décroissance des débits (k) noté entre septembre et octobre et entre octobre et novembre indique la baisse progressive de l'écoulement dans le bassin. Cette baisse de l'écoulement est plus significative entre le mois d'octobre et celui de novembre, en rapport avec la fin de la saison des pluies dans le bassin. Les coefficients k indiquent donc l'image réelle de la hausse et de la baisse des débits respectivement de part et d'autres du pic de l'écoulement annuel. Les valeurs positives du coefficient k indique que les débits sont faibles et celles négatives que les débits sont élevés.

Tableau 7. Coefficient de croissance (Juillet, Août et Septembre) et de décroissance (Septembre, Octobre et Novembre) par rapport au pic de l'écoulement

	Phase de montée			Phase de descente	
	Juillet	**Août**	**Septembre**	**Octobre**	**Novembre**
2004-05	1,14	0,33	0,28	-0,52	-0,47
2005-06	0,80	0,61	0,03	-0,30	-0,59
2006-07	0,78	0,68	0,37	-0,31	-0,54
2007-08	1,07	0,60	0,06	-0,35	-0,50
2008-09	1,42	0,41	0,04	0,05	-0,45
2009-10	0,92	1,11	0,25	-0,48	-0,86
2010-11	1,00	0,86	0,62	-0,06	-0,32
2011-12	0,85	0,46	0,36	-0,33	-0,62
2012-13	1,78	0,47	0,14	-0,19	-0,59
2013-14	1,25	1,13	0,01	-0,44	-0,30

3.4. Validation des résultats sur la base de l'effet de causalité

Dans ce travail, on essaie d'analyser la nature de la variation des débits écoulés du bassin pendant la période de hautes eaux à travers de graphiques (figure 6) représentant les données de débits écoulés. Les processus hydrologiques ont montré une relation de cause à effet avec les événements pluvieux, du fait que c'est le cadre climatique qui détermine les modalités de l'écoulement fluvial. Ainsi, l'évolution de la pluviométrie et des températures sur la période de hautes eaux, à l'échelle mensuelle, est indiquée sur la figure 6 qui montre un modèle assez similaire et conforme au modèle de débits écoulés, ce qui montre l'importance des lames d'eau précipitées et leur contribution sur les lames d'eau écoulées. La répartition des précipitations pourrait être observée tout au long de la période de hautes eaux. Dans cet article, les observations suivantes valident fortement les modèles obtenus : *1)* l'évolution des débits écoulés est semblable aux hydrogrammes sur chaque année pendant la période de hautes eaux ; *2)* l'évolution des débits écoulés a tendance à augmenter pendant les périodes de forte pluviométrie ; *3)* la figure pluviométrique - écoulement présente plus de similitudes, malgré le décalage d'un mois entre le pic de la pluie et celui de l'écoulement.

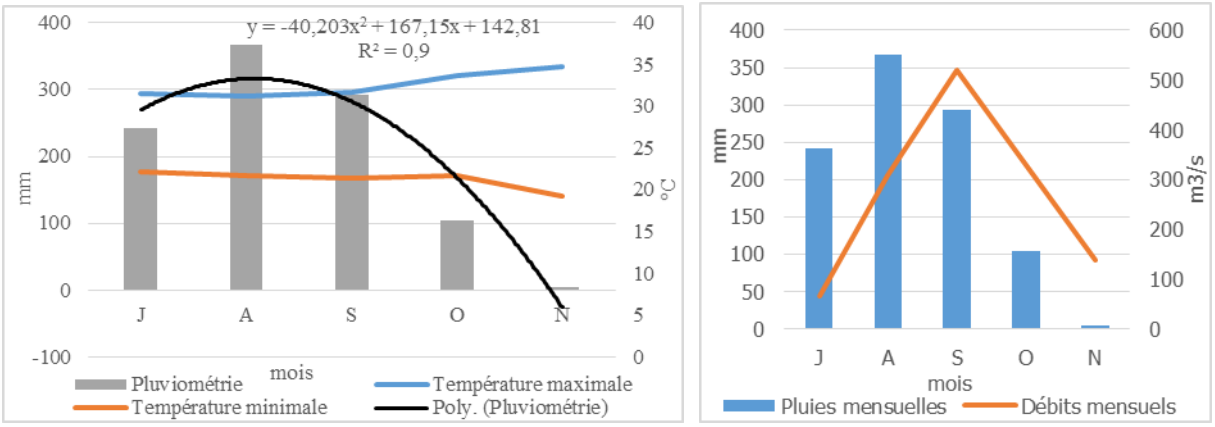

Fig.6. Pluies, températures (à Kédougou) et débits (à Mako) moyennes mensuelles

4. CONCLUSION

Sur cet article, les techniques d'exploration de données telles que les algorithmes de classifications comme la méthode des nuées dynamiques (k-means clustering), la Classification Ascendante Hiérarchique (CAH) et le critère de Ward, la recherche de similarité et la découverte de modèles sont utilisés dans les séries de données hydrologiques.

Les modèles découverts sont plus semblables aux modèles standards de débits. La comparaison des hydrogrammes et des précipitations au cours de la même période, a été faite et il est prouvé que les modèles de l'écoulement étaient plus semblables dans les mêmes périodes climatiques. Les modèles extraits des données hydrologiques peuvent être utilisés pour la prédiction de la valeur future de l'écoulement pendant la période de hautes eaux. Ces modèles sont valables pour les nouvelles données hydrologiques avec un certain degré de certitude.

Dans la zone tropicale ouest africaine, tout le système fluvial est divisé en 3 segments (périodes) : basses eaux (mai-juillet), hautes eaux (août-octobre) et basses eaux (novembre-décembre). Dans cette étude, nous n'avons utilisé que des données la période de hautes eaux sur un total de dix ans (2004-2013). Nos études futures pourraient se focaliser tantôt sur les autres périodes (celles de basses eaux), tantôt sur une série de données beaucoup plus longue (par exemple sur 20 ans) pour une étude plus complète du comportement hydrologique du bassin de la Gambie sur la station de Mako en particulier et même sur d'autres stations du bassin.

Toutefois, les méthodes appliquées renferment des inconvénients. Pour la méthode k-means, l'inconvénient est qu'elle ne permet pas de découvrir quel peut être un nombre cohérent de classes, ni de visualiser la proximité entre les classes ou les objets. Pour la CAH, le principal inconvénient est qu'elle nécessite le calcul des distances entre individus pris deux à deux. Ce qui est très rapidement prohibitif dès que la taille du fichier excède le millier d'individus. Après l'application de la méthode k-means et des techniques de CAH, on a constaté que la technique de CAH est plus précise et son principal avantage par rapport aux autres méthodes de classification réside dans cette représentation sous forme d'arbre qui met en évidence une information supplémentaire. De plus, les méthodes k-means et CAH sont donc complémentaires.

5. REFERENCES

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

[2] Larose D. T. Discovering Knowledge in Data: An Introduction to Data Mining, John Wiley & Sons, Inc., Hoboken, New Jersey. 2005.

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

[4] Jayanthi R. Application of data mining techniques in pharmaceutical industr. *JATIT*, 2007, 61-67.

[5] Piatetsky-Shapiro G. and Frawley W. J. Knowledge Discovery in Databases. *AAAI/MIT* Press: Boston, MA, 1991.

[6] Mishra S., Dwivedi V.K., Saravanan C. and Pathak K.K. Pattern discovery in hydrological time series data mining during the monsoon period of the high flood years in Brhamaputra river basin' *IJCA*, 2013, 67 (6), 7-14.

[7] Aydin I., Karakose and Akin A. The prediction Algorithm based on Fuzzy logic using time series data mining method, *World Academy of Science, Engg. and Technology*, 2009, 27, 91-98.

[8] Lamagat J.P. Monographie hydrologique du fleuve Gambie Collection M&m. ORSTOM-OMVG, 1989, 250 p.

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

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

[11] Rakthanmanon, T., Campana, B., Mueen, A., Batista, G., Westover, B., Zhu, Q., Zakaria, J., and Keogh, E., „Searching and mining trillions of time series sub sequences under dynamic time warping", ACM, 978-1-4503-1462, 2012, 340-356.

[12] Ding, H., Trajcevski, G., Scheuermann, P., Wang, X., and Keogh, E. J., „Querying and mining of time series data: experimental comparison of representations and distances measures", PVLDB, 1,2, 2008, 1542-52.

How to cite this article:
Faye C. System for predicting dischearges over the high water period through the classification techniques data: case of the Gambia river basin of Mako. J. Fundam. Appl. Sci., 2019, *11(2), 883-900.*