Classification des hydrogrammes avec des outils de l'analyse de données fonctionnelles.

Camille Ternynck [1] & Mohammed Ali Ben Alaya [2] & Fateh Chebana [2] & Sophie Dabo-Niang [3] & Taha B.M.J. Ouarda [1]

[1] *iWater, Masdar Institute, Masdar City, Abu Dhabi, Emirats Arabes Unis; cternynck@masdar.ac.ae; touarda@masdar.ac.ae*
[2] *Institut National de la Recherche Scientifique, Québec (QC), Canada; Mohammed_Ali.Ben_Alaya@ete.inrs.ca; Fateh.Chebana@ete.inrs.ca*
[3] *Laboratoire EQUIPPE, Université de Lille, Villeneuve d'Ascq, France; sophie.dabo@univ-lille3.fr*

Résumé. La classification des hydrogrammes de débit joue un rôle important dans un grand nombre d'études hydrologiques et hydrauliques. Elle permet, par exemple, de prendre des décisions quant à l'implémentation de structures hydrauliques, de caractériser différents types de crues induisant une meilleure compréhension des comportements extrêmes des débits. Les méthodes employées pour classifier les hydrogrammes sont généralement basées sur un nombre fini de caractéristiques de l'hydrogramme, n'incluant pas toute l'information disponible contenue dans la série de données. Dans ce travail, nous adaptons et appliquons trois méthodes statistiques de classification pour données fonctionnelles pour l'analyse des hydrogrammes de débit. La classification fonctionnelle emploie directement toutes les données de la série étudiée et utilise ainsi toute l'information disponible sur la forme, le pic, la date, etc. Les méthodes sont appliquées aux données provenant de la province du Québec, Canada. Nous montrons que les classes obtenues en utilisant la méthodologie fonctionnelle présentent de l'intérêt et peuvent mener à une meilleure représentation que celles obtenues en utilisant une méthode multidimensionnelle hiérarchique usuelle. L'approche fonctionnelle présente l'avantage d'utiliser toute l'information contenue dans l'hydrogramme, réduisant ainsi la subjectivité inhérente à l'analyse multidimensionnelle sur le type et le nombre de caractéristiques à utiliser, et par conséquent diminuant l'incertitude associée.

Mots-clés. Données fonctionnelles, Classification non supervisée, Hydrogramme de débit, k-moyennes, Classification hiérarchique

Abstract. Classification of streamflow hydrographs plays an important role in a large number of hydrological and hydraulic studies. For instance, it allows to make decisions regarding the implementation of hydraulic structures and to characterize different flood types leading to a better understanding of extreme flow behavior. The employed hydrograph classification methods are generally based on a finite number of hydrograph characteristics, and do not include all the available information contained in a discharge

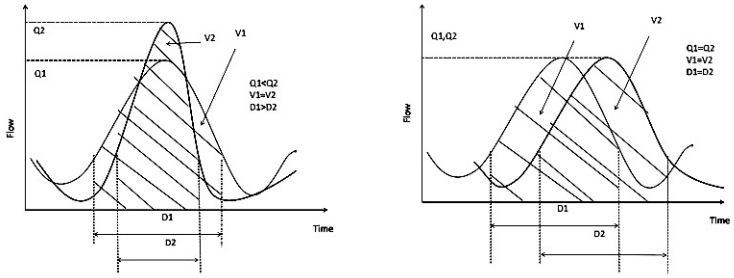
time series. In this work, we adapt and apply three statistical techniques from the theory of functional data classification for the analysis of flood hydrographs. Functional classification directly employs all data of a discharge time series and thus contains all available information on shape, peak, timing, etc. The proposed functional methodology is applied to streamflow datasets from the province of Quebec, Canada. We show that classes obtained using functional approaches have merit and can lead to better representation than those obtained using the usual multidimensional hierarchical classification method. The proposed methodology has the advantage of using the entire information contained in the hydrograph, reducing hence the subjectivity that is inherent in multidimensional analysis on the type and number of characteristics to be used, and diminishing consequently the associated uncertainty.

Keywords. Functional data, Clustering, Streamflow hydrograph, k-means, Hierarchical classification.

1 Résumé long

L'hydrogramme est la principale source d'information permettant d'étudier le comportement des débits. Il s'agit de la représentation graphique de la variation du débit au cours du temps. L'information qu'il fournit est essentielle pour déterminer la sévérité et la fréquence d'évènements hydrologiques extrêmes, comme les inondations et les sécheresses. Pour un bassin versant donné, les hydrogrammes annuels peuvent ne pas être similaires d'année en année. Classifier ces hydrogrammes en classes homogènes présente un intérêt pour identifier et comprendre les différents régimes, caractériser des groupes, séparer les évènements et détecter des changements possibles.

Dans la littérature hydrologique, un hydrogramme est généralement caractérisé par un nombre fini de caractéristiques (e.g. pic, volume, durée). Cependant, puisqu'il représente la variation du débit sur une certaine période, il ne peut être caractérisé seulement par un nombre fini, même grand, de caractéristiques mais plutôt par l'ensemble de l'hydrogramme comme une courbe. Les approches multidimensionnelles dépendent des indices utilisés pour caractériser le phénomène, et le fait de ne pas tenir compte de certains indices peut influencer les résultats. Par exemple, la figure suivante illustre deux situations pouvant justifier l'utilisation d'outils de l'analyse de données fonctionnelles pour étudier les hydrogrammes. A gauche, il s'agit de deux hydrogrammes de même volume mais dont les pics et les durées diffèrent. Ainsi une méthode basée sur ces caractéristiques pourrait facilement détecter les différences entres ces deux phénomènes. A droite, les deux hydrogrammes ont les mêmes volume, pic et durée. Par conséquent, une approche multivariée basée sur ces trois caractéristiques ne parviendrait pas à distinguer ces deux évènements. Il faudrait rajouter une caractéristique supplémentaire comme la date de début de l'évènement.



Par ailleurs, lorsqu'un très grand nombre de caractéristiques est utilisé, une importante quantité d'information peut être extraite et ainsi l'hydrogramme pourrait presque être représenté. Cependant, des inconvénients apparaissent, tels que l'augmentation de la dimension, la redondance et la subjectivité. De plus, quand le nombre de variables à inclure croît, le nombre de choix et possibilités de sous-ensembles de variables augmente aussi. Des techniques de sélection de variables peuvent être utilisées mais sont souvent coûteuses en temps de calculs. Par ailleurs, certaines variables ne sont pas directement disponibles et requièrent l'extraction à partir des données brutes, ce qui peut causer une augmentation de l'incertitude liée au manque de précisions dans les calculs.

Récemment, Chebana *et al.* (2012) montrent que le cadre de l'analyse de données fonctionnelles est adapté au contexte hydrologique avec un certain nombre d'avantages. La cadre fonctionnel y apparaît comme plus général, flexible et représentatif du phénomène hydrologique que celui de l'analyse multidimensionnelle. Ainsi, notre objectif est d'introduire le cadre de l'analyse de données fonctionnelles pour classifier des hydrogrammes de débits, en considérant les séries de décharges comme des courbes continues. Nous considérons donc un problème de classification non supervisée pour données fonctionnelles.

Dans cette présentation, nous rappelons les principes de trois méthodes de classification non supervisée rencontrées dans la littérature fonctionnelle et considérées dans notre étude. Parmi ces méthodes, on retrouve la méthode de classification descendante hiérarchique basée sur la distance entre les courbes moyenne et modale d'un ensemble de courbes (voir Dabo-Niang *et al.* (2007)). Nous avons également appliqué deux méthodes de k-moyennes adaptées au cadre des données fonctionnelles. D'une part, il s'agit du cas où la mesure de distortion classique entre les données est remplacée par une divergence de Bregman fonctionnelle (Fischer (2010)). D'autre part, nous avons considéré la méthode basée sur les coefficients des projections des courbes (Auder et Fischer (2012)).

Nous donnons ensuite les résultats issus de l'application de ces méthodes sur les hydrogrammes de plusieurs stations hydrologiques de la province de Québec (Canada). Les données traitées concernent les débits journaliers de certaines rivières Québécoises. Nous avons étudié, plus particulièrement, les débits printaniers de la rivière Romaine, observés de 1961 à 2000. Une étude comparative des résultats obtenus avec ces différentes techniques est donnée ainsi qu'une interprétation environnementale des résultats. Par soucis de comparaison, nous appliquons également une méthode de statistique multidimension-

nelle qui est habituellement utilisée dans la littérature hydrologique. Nous expliquons les avantages et inconvénients à considérer une méthode pour données fonctionnelles plutôt que multidimensionnelles.

Bibliographie

[1] Auder, B. et Fischer A. (2012). Projection-based curve clustering. *Journal of Statistical Computation and Simulation*, 82(8):1145–1168.
[2] Chebana, F., Dabo-Niang, S. et Ouarda, T.B.M.J. (2012). Exploratory functional flood frequency analysis and outlier detection *Water Resources Research*, 48(4): W04514
[3] Dabo-Niang, S., Ferraty, F. et Vieu, P. (2007). On the using of modal curves for radar waveforms classification. *Computational Statistics & Data Analysis*, 51(10):4878–4890.
[3] Fischer, A. (2010). Quantization and clustering with Bregman divergence. *Journal of Multivariate Analysis*, 101(9):2207–2221.