Revue des sciences de l'eau
Journal of Water Science

Une approche floue pour la détermination de la région d'influence d'une station hydrométrique
A fuzzy approach to the delineation of region of influence for hydrometric stations

Z. K. Bargaoui, V. Fortin, B. Bobée and L. Duckstein

Volume 11, Number 2, 1998

URI: https://id.erudit.org/iderudit/705307ar
DOI: https://doi.org/10.7202/705307ar

See table of contents

Publisher(s)

Université du Québec - INRS-Eau, Terre et Environnement (INRS-ETE)

ISSN

0992-7158 (print)
1718-8598 (digital)

Explore this journal

Cite this article

Bargaoui, Z. K., Fortin, V., Bobée, B. & Duckstein, L. (1998). Une approche floue pour la détermination de la région d'influence d'une station hydrométrique. *Revue des sciences de l'eau / Journal of Water Science*, *11*(2), 255–282. https://doi.org/10.7202/705307ar

Article abstract

The concept of partial membership of a hydrometric station in a hydrologic region is modeled using fuzzy sets theory. Hydrometric stations are represented in spaces of hydrologic (coefficient of variation: *CV*, coefficient of skewness: *CS*, and their counterparts based on *L*- moments: *L-CV* and *L-CS*) and/or physiographic attributes (surface of watershed: *S*, specific flow: Q_s=*Qmoyen*/*S*, and a shape index: I_c). Two fuzzy clustering methods are considered.

First a clustering method by coherence (Iphigénie) is considered. It is based on the principle of transitivity: if two pairs of stations (*A*,*B*) and (*B*,*C*) are known to be "close" to one another, then it is incoherent to state that *A* is "far" from *C*. Using a Euclidean distance, all pairs of stations are sorted from the closest pairs to the farthest. Then, the pairs of stations starting and ending this list are removed and classified respectively as "close" and "far". The process is then continued until an incoherence is detected. Clusters of stations are then determined from the graph of "close" stations. A disadvantage of Iphigénie is that crisp (non fuzzy) membership functions are obtained.

A second method of clustering is considered (ISODATA), which consists of minimizing fuzziness of clusters as measured by an objective function, and which can assign any degree of membership between 0 to 1 to a station to reflect its partial membership in a hydrologic region. It is a generalization of the classical method of mobile centers, in which crisp clusters minimizing entropy are obtained. When using Iphigénie, the number of clusters is determined automatically by the method, but for ISODATA it must be determined beforehand.

An application of both methods of clustering to the Tunisian hydrometric network (which consists of 39 stations, see Figure 1) is considered, with the objective of obtaining regional estimates of the flood frequency curves. Four planes are considered: P1: (Q_s,*CV*), P2: (*CS*,*CV*), P3: (*L-CS*,*L-CV*), and P4: (S,I_c), based on a correlation study of the available variables (Table 1).

Figures 2, 3a, 4 and 5 show the clusters obtained using Iphigénie for planes P1 through P4. Estimates of skewness (*CS*) being quite biased and variable for small sample sizes, it was decided to determine the influence of sample size in the clusters obtained for P2. Figure 3b shows the clusters obtained when the network is restricted to the 20 stations of the network for which at least 20 observations of maximum annual flood are available. Fewer clusters are obtained than in Figure 3, but it can be observed that the structure is the same: additional clusters appearing in Figure 3 may be obtained by breaking up certain large clusters of Figure 3b. In Figure 3c, the sample size of each of the 39 stations of the network is plotted in the plane (*CS*,*CV*), to see if extreme estimated values of *CS* and *CV* were caused by small samples. This does not seem to be the case, since many of the most extreme points correspond to long series.

ISODATA was also applied to the network. Based on entropy criteria (Table 2, Figures 6a and 6b), the number of clusters for ISODATA was set to 4. It turns out that the groups obtained using ISODATA are not very fuzzy. The fuzzy groups determined by ISODATA are generally conditioned by only one variable, as shown by Figures 7a-7d, which respectively show the fuzzy clusters obtained for planes P1-P4. Only lines of iso-membership of level 0.9 were plotted to facilitate the analysis. For hydrologic spaces (P2 and P3), it is skewness (*CS* and *L-CS*) and for physiographic spaces (P1 and P4) it is surface (Q_s and *S*).

Regionalization of the 100-year return period flood is performed based on the homogeneous groups obtained (using an index-flood method), and compared to the well-known region of influence (ROI) approach, both under the hypothesis of a 2-parameter Gamma distribution and a 3-parameter Pareto distribution. For the ROI approach, the threshold corresponding to the size of the ROI of a station is taken to be the distance at which an incoherence first appeared when applying Iphigénie. Correlation of the regional estimate with a local estimation for space P1 is 0.91 for Iphigénie and 0.85 both for ISODATA and the ROI approach. Relative bias of regional estimates of the 100-year flood based on P1 is plotted on Figures 9 (Gamma distribution) and Figure 10 (Pareto distribution). The three methods considered give similar results for a Gamma distribution, but Iphigénie estimates are less biased when a Pareto distribution is used. Thus Iphigénie appears superior, in this case, to ISODATA and ROI. Values of bias and standard error for all four planes are given for Iphigénie in Table 3.

Application of an index-flood regionalization approach at ungauged sites requires the estimation of mean flow (also called the flood index) from physiographic attributes. A regression study shows that the best explanatory variables are watershed surface *S*, the shape index I_c and the average slope of the river. In Figure 8, the observed flood index is plotted against the flood index obtained by regression. The correlation coefficient is 0.93.

Iphigénie and ISODATA could also be used in conjunction with other regionalization methods. For example, when using the ROI approach, it is necessary to, quite arbitrarily, determine the ROI threshold. It has been shown that this is a byproduct of the use of Iphigénie. ISODATA is most useful for pattern identification when the data is very fuzzy, unlike the example considered in this paper. But even in the case of the Tunisian network, its application gives indications as to which variables (skewness and surface) are most useful for clustering.

REVUE DES SCIENCES DE L'EAU, Rev. Sci. Eau 11(1998) 255-282

Une approche floue pour la détermination de la région d'influence d'une station hydrométrique

A fuzzy approach to the delineation of region of influence for hydrometric stations

Z.-K. BARGAOUI[1*], V. FORTIN[2], B. BOBÉE[2] et L. DUCKSTEIN[3]

Reçu le 12 août 1996, accepté le 14 janvier 1998**.

SUMMARY

The concept of partial membership of a hydrometric station to a hydrologic region is modeled using fuzzy sets theory. Hydrometric stations are represented in spaces of hydrologic (coefficient of variation: *CV*, coefficient of skewness: *CS*, and their counterparts based on *L*-moments: *L-CV* and *L-CS*) and/or physiographic attributes (surface of watershed: *S*, specific flow: $Q_s = Q_s = \overline{Q}/S$, and a shape index: I_c). Two fuzzy clustering methods are considered. First a clustering method by coherence (Iphigénie) is considered. It is based on the principle of transitivity: if two pairs of stations (*A*, *B*) and (*B*, *C*) are known to be "close" to one another, then it is incoherent to state that *A* is "far" from *C*. Using a Euclidean distance, all pairs of stations are sorted from the closest pairs to the farthest. Then, the pairs of stations starting and ending this list are removed and classified respectively as "close" and "far". The process is then continued until an incoherence is detected. Clusters of stations are then determined from the graph of "close" stations. A disadvantage of Iphigénie is that crisp (non fuzzy) membership functions are obtained. A second method of clustering is considered (ISODATA), which consists of minimizing fuzziness of clusters as measured by an objective function, and which can assign any degree of membership between 0 to 1 to a station to reflect its partial membership to a hydrologic region. It is a generalization of the classical method of mobile centers, in which crisp clusters minimizing entropy are obtained. When using Iphigénie, the number of clusters is determined automatically by the method, but for ISODATA it must be determined beforehand. An application of both methods of clustering to the Tunisian hydrometric network (which consists of 39 stations, see figure 1) is considered, with the objective of obtaining regional estimates of the flood frequency curves. Four planes are considered: P1:

1. École Nationale d'Ingénieurs de Tunis BP 37 1002 Tunis, Tunisie.
2. Chaire CRSNG/Hydro-Québec en hydrologie statistique, Institut national de la recherche scientifique, 2800 rue Einstein, CP 7500, Sainte-Foy (Québec) Canada G1V 4C7.
3. ENGREF, 19 rue du Maine, 75732 Paris cedex 15.

* Correspondance. e-mail : chaire_hydro@inrs-eau.uquebec.ca.
** Les commentaires seront reçus jusqu'au 31 décembre 1998.

(Q_s,*CV*), P2: (*CS*,*CV*), P3: (*L-CS*,*L-CV*), and P4: (S,I_c), based on a correlation study of the available variables *(Table 1)*. Figures 2, 3a, 4 and 5 show the clusters obtained using Iphigénie for planes P1 through P4. Estimates of skewness (*CS*) being quite biased and variable for small sample sizes, it was decided to determine the influence of sample size in the clusters obtained for P2. Figure 3b shows the clusters obtained when the network is restricted to the 20 stations of the network for which at least 20 observations of maximum annual flood are available. Fewer clusters are obtained than in Figure 3, but it can be observed that the structure is the same: additional clusters appearing in Figure 3 may be obtained by breaking up certain large clusters of Figure 3b. In Figure 3c, the sample size of each of the 39 stations of the network is plotted in the plane (*CS*,*CV*), to see if extreme estimated values of *CS* and *CV* were caused by small samples. This does not seem to be the case, since many of the most extreme points correspond to long series. ISODATA was also applied to the network. Based on entropy criteria *(Table 2, Figures 6a and 6b)*, the number of clusters for ISODATA was set to 4. It turns out that the groups obtained using ISODATA are not very fuzzy. The fuzzy groups determined by ISODATA are generally conditioned by only one variable, as shown by Figures 7a-7d, which respectively show the fuzzy clusters obtained for planes P1-P4. Only lines of isomembership of level 0.9 were plotted to facilitate the analysis. For hydrologic spaces (P2 and P3), it is skewness (*CS* and *L-CS*) and for physiographic spaces (P1 and P4) it is surface (Q_s and S). Regionalization of the 100-year return period flood is performed based on the homogeneous groups obtained (using an index-flood method), and compared to the well-known region of influence (ROI) approach, both under the hypothesis of a 2-parameter Gamma distribution and a 3-parameter Pareto distribution. For the ROI approach, the threshold corresponding to the size of the ROI of a station is taken to be the distance at which an incoherence first appeared when applying Iphigénie. Correlation of the regional estimate with a local estimation for space P1 is 0.91 for Iphigénie and 0.85 both for ISODATA and the ROI approach. Relative bias of regional estimates of the 100-year flood based on P1 is plotted on Figures 9 (Gamma distribution) and Figure 10 (Pareto distribution). The three methods considered give similar results for a Gamma distribution, but Iphigénie estimates are less biased when a Pareto distribution is used. Thus Iphigénie appears superior, in this case, to ISODATA and ROI. Values of bias and standard error for all four planes are given for Iphigénie in Table 3. Application of an index-flood regionalization approach at ungauged sites requires the estimation of mean flow (also called the flood index) from physiographic attributes. A regression study shows that the best explanatory variables are watershed surface S, the shape index I_c and the average slope of the river. In Figure 8, the observed flood index is plotted against the flood index obtained by regression. The correlation coefficient is 0.93. Iphigénie and ISODATA could also be used in conjunction with other regionalization methods. For example, when using the ROI approach, it is necessary to, quite arbitrarily, determine the ROI threshold. It has been shown that this is a byproduct of the use of Iphigénie. ISODATA is most useful for pattern identification when the data is very fuzzy, unlike the example considered in this paper. But even in the case of the Tunisian network, its application gives indications as to which variables (skewness and surface) are most useful for clustering.

Key-words: *hydrology, regionalization, 100-year flood, fuzzy sets, gamma, pareto, iphigénie, isodata.*

RÉSUMÉ

La notion d'appartenance partielle d'une station hydrométrique à une région hydrologique est modélisée par une fonction d'appartenance obtenue en appliquant les concepts de l'analyse floue. Les stations hydrométriques sont représentées dans des plans dont les axes sont des attributs hydrologiques et/ou phy-

siographiques. Les régions hydrologiques sont considérées comme des sous-ensembles flous. Une méthode d'agrégation par cohérence (Iphigénie) permet d'établir des classes d'équivalence pour la relation floue « il n'y a pas d'incohérence entre les éléments d'une même classe » : ce sont des classes d'équivalence qui représentent les régions floues. La fonction d'appartenance dans ce cas est stricte. Par opposition, la seconde méthode de type centres mobiles flous (ISODATA) permet d'attribuer un degré d'appartenance d'une station à une région floue dans l'intervalle [0,1]. Celle-ci reflète le degré d'appartenance de la station à un groupe donné (le nombre de groupes étant préalablement choisi de façon heuristique). Pour le cas traité (réseau hydrométrique tunisien, débits maximums annuels de crue), il s'avère cependant que le caractère flou des stations n'est pas très prononcé. Sur la base des agrégats obtenus par la méthode Iphigénie et des régions floues obtenues par ISODATA, est effectuée une estimation régionale des débits maximums de crue de période de retour 100 ans. Celle-ci est ensuite comparée à l'estimation régionale obtenue par la méthode de la région d'influence ainsi qu'à l'estimation utilisant les seules données du site, sous l'hypothèse que les populations parentes sont des lois Gamma à deux paramètres et Pareto à trois paramètres.

Mots clés : *hydrologie, régionalisation, crues, flou, gamma, pareto, iphigénie, isodata.*

INTRODUCTION

La détermination des régions hydrologiques homogènes par les méthodes d'analyse multivariée a fait l'objet de plusieurs études. Dans la panoplie des méthodes de partitionnement (nuées dynamiques, centres mobiles, agrégations de relations binaires et classification hiérarchique), WILTSHIRE (1986) et BURN (1989) ont choisi la méthode des nuées dynamiques alors que MOSLEY (1981) ainsi qu'ACREMAN et SINCLAIR (1986) ont utilisé une procédure de classification hiérarchique. BURN (1988) a traité le problème en ayant recours à l'analyse en composantes principales. Les attributs avec lesquels ces méthodes ont été mises en œuvre sont hydrologiques (coefficient de variation des débits, débit moyen spécifique) et/ou physiographiques (surface du bassin versant, pourcentage de lac et marais, coordonnées en latitude et longitude des stations...). Le point commun entre ces méthodes est qu'elles affectent chaque site observé à une région unique. Ce sont ACREMAN et WILTSHIRE (1989) qui ont entrevu la possibilité pour un site d'appartenir à plusieurs régions à la fois. En fait, c'est en utilisant l'analyse discriminante sur les caractéristiques physiographiques et en vue de la comparaison avec les résultats de la méthode des centres mobiles, que WILTSHIRE (1986) a calculé la probabilité *a posteriori* d'un site d'appartenir à une région. Cette probabilité a ensuite été interprétée comme une appartenance partielle.

BURN (1990), en systématisant cette approche, a rattaché à chaque site d'intérêt, une « région d'influence » et a quantifié la similarité entre un site et sa région d'influence par une fonction poids qui peut être interprétée comme un degré d'appartenance du site à la région. Plus récemment FORTIN *et al.* (1995) ont proposé de formaliser cette approche d'appartenance partielle d'un site à une région à l'aide de la théorie des sous-ensembles flous. La notion de sous-ensemble flou permet en effet de rendre compte du fait qu'une station hydrométrique, au lieu d'appartenir ou de ne pas appartenir de façon stricte à une région, peut y

appartenir à un certain degré, défini par un scalaire dans l'intervalle [0,1]. En conséquence, la région hydrologique prend le caractère de sous-ensemble flou, c'est-à-dire d'un sous-ensemble dont les éléments (les sites) sont affectés d'un degré d'appartenance.

Ce travail est une application à la détermination de régions hydrologiques homogènes de la théorie des sous-ensembles flous. Le cas d'étude est celui des 39 stations principales du réseau hydrométrique tunisien : 3 comportent plus de 50 années d'observation, 8 comportent entre 30 et 50 observations, 9 comportent entre 20 et 30 ans de données, et finalement 19 comportent entre 10 et 20 ans de données. Aucune station ne comporte moins de 10 observations. La figure 1 montre la répartition de ces stations sur le territoire tunisien. La grande majorité des stations se situent dans la partie nord du pays, avec seulement quelques stations situées dans le sud, en bordure de la mer. Notons que malgré la taille modeste de la Tunisie, son climat varie considérablement du nord au sud et d'est en ouest, à cause de la variation de latitude et de l'effet de la continentalité.

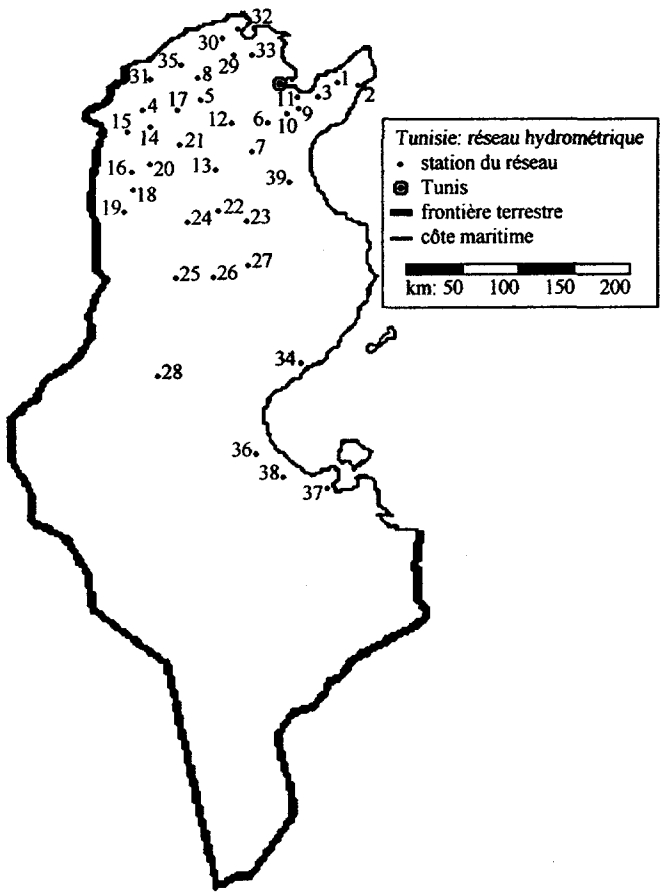

Figure 1 Distribution des 39 principales stations hydrométriques sur le territoire tunisien.

Areal distribution of the 39 hydrometric stations in Tunisia.

MÉTHODOLOGIE

Plusieurs concepts de la théorie des sous-ensembles flous (Kaufmann, 1975) sont utilisés dans ce travail :

(i) Le concept de sous-ensemble flou, c'est-à-dire un sous-ensemble regroupant des éléments dont le degré d'appartenance à cet ensemble est caractérisé par un scalaire dans l'intervalle [0,1]. On peut définir un sous-ensemble flou A par une fonction d'appartenance $\mu_A : x \rightarrow [0,1]$, qui associe à chaque élément x du référentiel son degré d'appartenance $\mu_A(x) \in [0,1]$ à l'ensemble A. La fonction d'appartenance floue généralise le concept de fonction indicatrice $I_A : x \rightarrow \{0,1\}$ d'un ensemble non flou, qui permet uniquement de spécifier si un élément x appartient complètement ou n'appartient pas du tout à l'ensemble. Chaque région hydrologique sera considérée comme un sous-ensemble flou, et chaque station sera caractérisée par son degré d'appartenance à chacune de ces régions floues.

(ii) Le concept de relation floue, c'est-à-dire une généralisation du concept de relation d'équivalence permettant de mesurer par un scalaire dans l'intervalle [0,1] le degré auquel une proposition logique est vérifiée. La relation « la station x est voisine de la station y », que nous considérerons être une relation floue, sera à la base d'une procédure d'agrégation des stations en régions. Cette relation peut être représentée sous forme matricielle, ou par un graphe dont les sommets sont les stations et qui matérialise la relation floue entre les sites.

(iii) Le concept de partition floue. L'ensemble des N stations sera décomposé en R sous-ensembles flous, ce qui revient à affecter à chaque station k un R-uplet $\{\mu_1(k), \mu_1(k)..., \mu_R(k)\}$ où $\mu_i(k)$ est le degré d'appartenance à la région i de la station k. Pour former ces R agrégats, il faut adopter une méthode de décomposition à la quelle sera associé un critère.

Une première méthode, pouvant être rattachée aux méthodes de classification basées sur l'analyse des graphes de relations (BEZDEK, 1981) est appliquée. Cette méthode d'agrégation, dénommée Iphigénie par KAUFMANN (1975), permet de subdiviser les sites en régions telles qu'au sein de chacune d'elles il n'y ait pas d'incohérence. Ainsi, toutes les stations d'une région sont voisines au sens d'une relation floue. Par construction, cette méthode conduit à des classes strictes. Le nombre de classes n'est toutefois pas fixé à l'avance.

Une seconde méthode de classification, basée sur la minimisation d'une fonction objectif, est considérée. C'est RUSPINI (1970) qui, le premier, a proposé ce type de méthodes dans le cadre de la théorie des sous-ensembles flous. Dans ce travail, la procédure adoptée n'est pas celle de Ruspini mais est une extension aux ensembles flous de la procédure de classification classique des centres mobiles décrite par DUDA et HART (1973). Il s'agit de la méthode ISODATA (BEZDEK, 1981) qui, par construction, conduit à des fonctions d'appartenance floues. Un désavantage de cette méthode est qu'il faut fixer à l'avance le nombre de classes, bien que certaines méthodes heuristiques existent pour fixer ce nombre.

Agrégation par cohérence (procédure Iphigénie)

Cette procédure a été initialement proposée pour l'agrégation de sous-ensembles ordinaires (BERNARD et BESSON, 1971). KAUFMANN (1975) l'a appliquée à l'agrégation de sous-ensembles flous prédéfinis. Dans notre cas, nous en

appliquons le principe afin de constituer, à partir des $N = 39$ stations du réseau, des sous-ensembles flous (les régions hydrologiques) au sein desquels tous les éléments (les stations) sont voisins au sens de la relation floue.

La procédure consiste d'abord à choisir une métrique pour mesurer la distance entre les stations. Cette métrique est construite dans notre cas à partir de certains attributs hydrologiques et/ou physiographiques. Ensuite, on détermine la distance entre chacun des $N(N-1)/2$ couples de stations que l'on peut former à partir des N stations. Ceci permet de construire deux listes : une liste de stations voisines, et une liste de stations éloignées. On ajoute tour à tour un élément dans chaque liste, en choisissant d'abord pour la liste des voisins les stations les plus proches, et pour la liste des éloignées les stations les plus distantes. L'arrêt de la procédure a lieu lorsqu'est rencontrée une incohérence de transitivité. Par exemple, on a obtenu précédemment « station *a* voisine de *b* » et « station *b* voisine de *c* » et l'étape actuelle donne « station *a* éloignée de station *c* ». La liste des stations voisines est ensuite utilisée pour construire un graphe. On forme des regroupements de stations en découpant le graphe de façon à ce que chaque groupe possède la propriété de transitivité, c'est-à-dire qu'à l'intérieur d'un même groupe, toutes les stations sont voisines les unes des autres.

Il est utile d'illustrer la méthode par un exemple. Soit **D** la matrice des distances entre les stations. Nous utiliserons dans cet exemple la distance euclidienne entre les stations dans le plan formé par le débit spécifique moyen et le coefficient de variation des débits. On obtient la plus petite des valeurs de $\mathbf{D}_{i,j}$ pour le couple de stations (22,31), et la plus grande pour le couple (2,5). On note alors le couple (22,31) dans la liste des stations voisines, et le couple (2,5) dans la liste des stations éloignées. On élimine ensuite de **D** les cases correspondant aux couples (2,5) et (22,31) :

	1	⋯	5	⋯	31	⋯	39
1	0	⋯	$\mathbf{D}_{1,5}$	⋯	$\mathbf{D}_{1,31}$	⋯	$\mathbf{D}_{1,39}$
2		⋯	×	⋯	$\mathbf{D}_{2,31}$	⋯	$\mathbf{D}_{2,39}$
⋮			⋮	⋮	⋮	⋮	⋮
22			$\mathbf{D}_{22,5}$	⋯	×	⋯	$\mathbf{D}_{22,39}$
⋮					⋮	⋮	⋮
39					$\mathbf{D}_{39,31}$	⋯	0

On répète ensuite l'opération sur les éléments restants de la matrice de distance. On remplit d'étape en étape un tableau à 3 colonnes où la première donne le numéro de l'étape de calcul, la seconde « les voisins » (couple de distance minimale à l'étape considérée) et où la troisième colonne rassemble les « éloignés » (couples de distance maximale à l'étape considérée). On s'arrête dès qu'une incohérence est rencontrée c'est-à-dire :

- on a obtenu (i,j) voisins et (j,k) voisins à des étapes antérieures, et on obtient (i,k) éloignés ;
- on a obtenu (i,j) voisins et (i,k) éloignés à des étapes antérieures, et (j,k) sont sélectionnés voisins.

Dans l'exemple du plan débit moyen spécifique/coefficient de variation, les itérations se sont arrêtées à l'étape 258. On a obtenu une incohérence du deuxième type :

• à l'étape 242, les stations 12 et 38 sont placées dans la liste des stations éloignées ;
• à l'étape 252, les stations 13 et 38 sont placées dans la liste des stations voisines ;
• à l'étape 258, les stations 12 et 13 sont placées dans la liste des stations voisines

Comment obtenir des classes une fois atteint l'arrêt de la procédure ? Nous adoptons une méthode heuristique consistant à élaborer ces classes de proche en proche : aux couples des stations les plus rapprochées (22,31), on agrège la première station apparaissant dans la colonne « voisins » en tant que voisine de l'un des constituants de ce couple. On continue ainsi à agglomérer les stations qui se trouvent en liaison avec tous les éléments déjà constitutifs d'un agrégat. Lorsqu'un nouvel élément ne peut s'ajouter à cet agrégat sans briser la propriété de transitivité commune à tous ses éléments, on obtient une classe d'équivalence et l'on forme de nouveaux groupes avec les stations restantes. Remarquons qu'en raison de l'utilisation de la matrice des distances, les deux autres propriétés des classes d'équivalence (réflexivité et symétrie) sont assurées. Ce procédé de partition permet d'obtenir des classes d'équivalence pour la relation floue « il n'y a pas d'incohérence de voisinage entre les éléments d'une même classe ». Ainsi, ces classes ont un caractère flou mais leur fonction d'appartenance est du type {0,1}.

L'interprétation du graphe des relations obtenu par la méthode Iphigénie n'est pas unique. En effet :
• une classe peut tout à fait être subdivisée en sous-classes, tant que celles-ci vérifient les trois propriétés des classes d'équivalence : réflexivité, symétrie et transitivité ;
• de même, un sommet peut être indifféremment rattaché à plus d'une classe lorsqu'il est en liaison avec chacun de leurs éléments (bien que tous ces éléments ne soient pas liés entre eux) ;
• enfin, des dipôles peuvent être formés de manière non unique avec le même graphe de relations.

Ainsi, en plus de la limite importante consistant à réduire la fonction d'appartenance au doublet {0,1}, la méthode Iphigénie peut conduire dans l'interprétation des graphes de relations, à des agrégats ayant des intersections non vides et qui perdent de ce fait leur caractère de classes d'équivalence strictes.

Classification floue (méthode ISODATA)

Largement utilisée dans les problèmes de reconnaissance des formes (TERANO *et al.*, 1992), la procédure ISODATA permet de répartir N stations dans un nombre (fixé à l'avance) de régions floues. Celles-ci sont définies dès lors que leurs fonctions d'appartenance sont établies. Soit U un ensemble de N sites à classer en R régions floues, et soit $\mu_i(k) \in [0,1]$ la fonction d'appartenance du site s_k à la région floue i avec, pour k fixé ($k = 1,..,N$) :

$$\sum_{i=1}^{R} \mu_i(k) = 1 \qquad (1)$$

Soit $\mathbf{D}_{i,k}$ la distance entre le site s_k et le centre de gravité v_i de la région i. On détermine les valeurs des fonctions d'appartenance en minimisant une fonction objectif $J(U)$ du type moindres carrés pondérés :

$$J(U) = \sum_{k=1}^{N} \sum_{i=1}^{R} \mu_i(k)^p D_{i,k}^2 \quad (2)$$

où l'exposant $p>1$ est réel. La distance étant considérée comme un indicateur de dissimilarité, en minimisant la somme des carrés des distances pondérées des éléments du groupe à son centre de gravité, on améliore l'homogénéité au sein du groupe. Le poids utilisé pour pondérer la mesure de distance est la puissance p de la fonction d'appartenance du site s_k à la région i, soit $\mu_i^p(k)$.

Évidemment, la valeur choisie pour l'exposant p se répercute sur le poids attribué à chaque élément dans le calcul de la fonction objectif. Pour $p = 1$, le poids se confond avec le degré d'appartenance. Dans le cas de la méthode des centres mobiles classiques, où l'on force $\mu_i^p(k)$ à prendre une des deux valeurs 0 ou 1, $J(U)$ représente l'inertie intra-classes. À valeur égale pour $\mu_i(k)$, le poids décroît quand p augmente. En élevant p, on affaiblit la contribution à la fonction objectif des sites ayant une grande distance au centre de gravité. En d'autres termes, on pénalise moins les stations de faible degré d'appartenance (celles dont la dissimilarité avec le centre de gravité est forte) en augmentant p. Alors, on accepte un caractère flou plus marqué dans le groupe. Nous expliquerons dans la suite ce qui nous a guidé pour le choix de p.

La position du centre de gravité est calculée à l'aide de (3) qui l'exprime comme la moyenne pondérée par le poids $\mu_i^p(k)$ des positions des sites de la région, soit :

$$v_i = \sum_{k=1}^{N} \mu_i^p(k) \cdot s_k \Big/ \sum_{k=1}^{N} \mu_i^p(k) \quad (3)$$

On peut montrer que minimiser (2) sous la contrainte (1) conduit à l'expression (4) pour $\mu_i(k)$. La procédure consiste d'abord à fixer $2 \le R \le N$ et $p > 1$, à choisir une métrique et à initialiser la matrice d'appartenance à des valeurs quelconques respectant les contraintes. On calcule ensuite la position des centres des classes à l'aide de (3) et on actualise la matrice d'appartenance en utilisant (4) :

$$\mu_i(k) = \begin{cases} \left\{ \sum_{j=1}^{R} \left(D_{k,i} / D_{k,j} \right)^{2/(p-1)} \right\}^{-1} & (\forall j) s_k \neq v_j : \\ & \text{il n'y a aucune station au centre de cette classe} \\ 0 & (\exists k \neq i) s_i = v_j : \\ & \text{il y a une station autre que } k \text{ au centre de cette classe} \\ 1 & s_k = v_j : \\ & \text{la station } k \text{ est au centre de la classe } i \end{cases} \quad (4)$$

Par la suite, on répète à chaque itération le calcul des équations (3) et (4), jusqu'à ce que les différences entre les valeurs de $\mu_i(k)$ d'une itération à la suivante soient suffisamment faibles. À titre de comparaison, rappelons que dans la procédure des centres mobiles stricte, $p = 1$ et les degrés d'appartenance, dans {0,1}, sont obtenus ainsi :

$$\mu_i(k) = \begin{cases} 1 & \text{si } D_{k,i} = \inf\limits_{j=1,2,\ldots,R} D_{k,j} \\ 0 & \text{sinon} \end{cases} \quad (5)$$

CHOIX DES ESPACES DE CLASSIFICATION

Les stations hydrométriques peuvent être décrites par deux ensembles de variables : des variables physiographiques et des caractéristiques statistiques des échantillons observés appelées variables hydrologiques. Afin de tenir compte de l'interdépendance entre ces deux ensembles de variables, CAVADIAS (1990) a choisi d'interpréter la représentation des stations dans l'espace des variables canoniques physiographiques d'une part et hydrologiques d'autre part. Dans ce travail, nous allons multiplier les espaces et considérerons trois espaces hydrologiques et un espace physiographique.

Attributs hydrologiques

Historiquement, la régionalisation des débits maximaux a utilisé les attributs (Q_S,CV) où $Q_S = \overline{Q}/S$ est le débit spécifique moyen observé (en $m^3/s/km^2$) et $CV = \sigma/m$ le coefficient de variation des débits, qui correspond au rapport de la racine carré de la variance σ^2 à la moyenne μ. Ces deux attributs reflètent les moments d'ordre 1 (grâce à Q_S) et d'ordre 2 des débits (grâce à CV). On peut remarquer que la statistique Q_S fait intervenir la superficie du bassin versant qui est une caractéristique physiographique ; le plan P1 : (Q_S,CV) n'est donc pas un espace de représentation purement hydrologique. Afin que la métrique euclidienne donne un poids similaire aux deux attributs, la transformation logarithmique des débits spécifique est proposée : ainsi les deux variables ont approximativement un domaine de variation de même étendue. Pour modéliser l'asymétrie généralement caractéristique des distributions des débits, nous avons incorporé l'information véhiculée par l'estimateur du coefficient d'asymétrie CS.

$$CS = \mu_3/\sigma^3 = E\left[(X - E[X])^3\right] \Big/ E\left[(X - E[X])^2\right]^{3/2} \tag{6}$$

où $E[\cdot]$ désigne l'espérance mathématique. Le plan P2 : (CV, CS) est le second plan de travail. CV et CS ont l'avantage d'être adimensionnels et sont reliées pour certaines distributions statistiques. Par exemple, citons que $CS = 2 \cdot CV$ pour la distribution Gamma. BURN (1990) et WILTSHIRE (1986) avaient déjà envisagé la possibilité d'avoir recours au coefficient d'asymétrie en tant qu'attribut potentiel. Cependant, l'utilisation du coefficient d'asymétrie CS présente des inconvénients. Son estimateur est borné en fonction de la taille d'échantillon (KIRBY, 1974), biaisé (des corrections de biais ayant toutefois été envisagées, entre autres par BOBÉE et ASHKAR, 1991), sensible aux fluctuations d'échantillonnage et aux valeurs extrêmes. Nous avons, en raison des tailles peu élevées des échantillons de débits étudiés, examiné un troisième plan P3 : (L-CV,L-CS), constitué par des rapports de L-moments. Cela, parce que les estimateurs des L-moments proposés par HOSKING (1989), basés sur des statistiques d'ordre, sont moins biaisés et moins sensibles aux valeurs extrêmes de l'échantillon. Rappelons les définitions de L-CV et L-CS :

$$\begin{aligned} L-CV &= \lambda_2/\lambda_1 \\ L-CV &= \lambda_3/\lambda_2 \end{aligned} \tag{7}$$

où λ_r est le L-moment d'ordre r tel que :

$$\lambda_r = \frac{1}{r}\sum_{k=0}^{r-1}(-1)^k \binom{r-1}{k} E[X_{r-k:r}] \tag{8}$$

où $X_{r-k:r}$ est la valeur de rang $r-k$ dans un échantillon trié en ordre croissant de taille r.

Attributs physiographiques

La banque de données physiographiques disponible pour les bassins considérés dans l'étude comprend les variables suivantes :

- surface du bassin (km^2) S
- indice de compacité I_c, défini comme le rapport du périmètre du bassin au périmètre du cercle de même surface
- dénivelée spécifique (m/km) D_s, définie comme le rapport de la dénivelée à la longueur du rectangle équivalent. La dénivelée est la différence entre les altitudes ayant 5 % et 95 % de chances d'être dépassées dans le bassin. Le rectangle équivalent est le rectangle de même surface et même périmètre que le bassin versant.
- indice de pente globale (m/km^2) $I_g = D_s/\sqrt{S}$
- pente moyenne du cours d'eau principal (‰) I_m

La variable I_m est disponible pour 25 stations seulement parmi les 39, les autres étant disponibles pour 38 sites. Soulignons que seules les variables proprement physiographiques (obtenues à partir de la topographie par des procédés cartographiques) sont considérées dans ce travail. Les variables climatiques (pluie, température) et agro-climatiques (humidité du sol) n'ont pas été prises en compte. La matrice des corrélations entre les caractéristiques physiographiques est donnée par le tableau 1.

Tableau 1 Matrice des corrélations entre les caractéristiques physiographiques.

Table 1 *Correlation matrix between physiographic attributes.*

ρ	S	I_c	D_s	I_g	I_m
S	1	0,25	0,49	−0,37	−0,26
I_c		1	−0,27	−0,28	0,02
D_s			1	−0,07	−0,20
I_g				1	0,43
I_m					1

Ainsi, la surface S, l'indice de compacité I_c et la pente moyenne I_m du cours d'eau sont les variables les moins corrélées. Étant données les lacunes dans la base de donnée pour la pente moyenne, seules les deux premières seront retenues, formant ainsi le plan P4 : (log S, I_c). Si le diagramme log-log de la moyenne des débits maximums annuels Q_m (aussi nommé indice de crues) et de la surface du bassin S montre une assez bonne corrélation ($\rho = 0{,}72$), celle-ci n'est pas significative entre Q_m et I_c ($\rho = 0{,}08$) et reste faible entre Q_m et I_m ($\rho = -{,}25$). Il est intéressant de constater que la corrélation avec D_s est nettement plus significative ($\rho = 0{,}63$). Dans tous les cas traités, la métrique euclidienne a été utilisée.

RÉSULTATS

Agrégation par cohérence

Le graphe des relations dans le plan P1 : (log Q_s,*CV*) a permis d'obtenir les huit agrégats de la figure 2 dans laquelle cinq groupes se juxtaposent mais trois groupes sont assez distincts. Les régions dans le plan P2 : (*CV*,*CS*) sont au nombre de 7 *(figure 3a)* lorsque les 39 stations sont considérées. Dans le plan P3 : (*L-CV*,*L-CS*), 8 groupes sont obtenus *(figure 4)*. Sur cette figure où plusieurs intersections non vides sont envisageables apparaît la difficulté d'interprétation du graphe des relations. Remarquons que dans ce plan, les stations 36 et 37 sont très proches et constituent une entité tout à fait séparée des autres groupes, ce qui n'est pas le cas pour les autres plans où ces deux stations sont assez peu « voisines ».

Six régions ont été mises en évidence dans le plan P4 : (log *S*, I_c). On observe sur la figure 5 que les stations ont été essentiellement regroupées en fonction de la surface du bassin versant. Tenter d'incorporer la variable D_s s'est avéré infructueux. En effet, dans l'espace (log *S*, log D_s, I_c), les deux stations les plus proches ont des indices de crues Q_m fort différents. Il en est ainsi pour différentes combinaisons d'espaces où interviennent I_g ou D_s : les stations proches dans ces espaces s'avèrent assez distinctes du point de vue de leurs statistiques de crue. On peut conclure empiriquement pour ce cas particulier que ces deux variables n'expliquent pas la proximité hydrologique.

En observant ces quatre plans de représentation, on réalise que certaines stations ne sont pas toujours voisines. Ceci justifie, *a posteriori*, l'utilisation

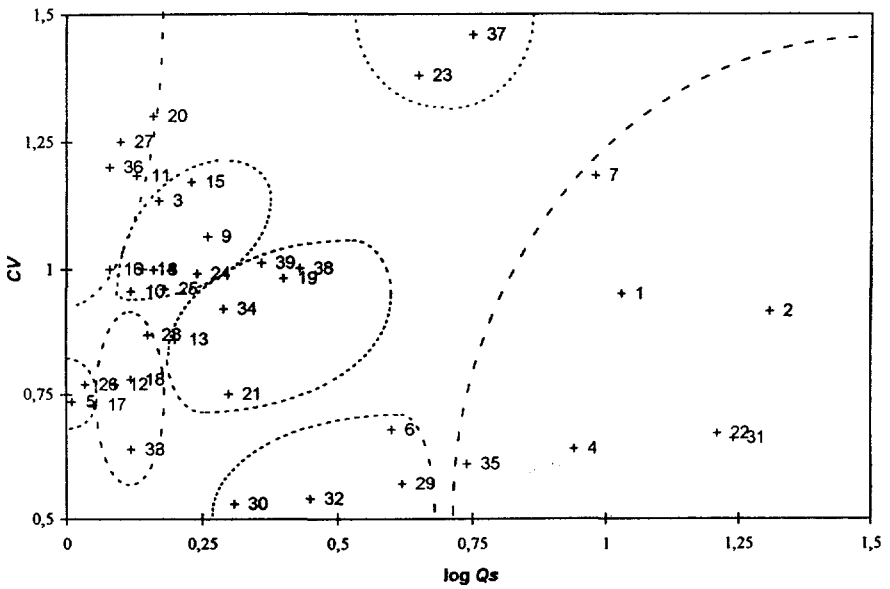

Figure 2 Groupes obtenus avec Iphigénie pour (logQs, CV).

Clusters obtained with Iphigénie in the (logQs, CV) plane.

conjointe de ces différents plans. Il serait possible, en pondérant les quatre matrices de relations issues de chaque application de la procédure Iphigénie, d'obtenir une matrice agrégée (KAUFMANN, 1975).

Sensibilité du résultat à la taille d'échantillon

La sensibilité des résultats à la taille d'échantillon a été étudiée dans le plan P2 : (*CV,CS*) puisque l'estimateur de coefficient d'asymétrie y est particulièrement sensible. La figure 3b montre les quatre groupes obtenus lorsqu'un réseau réduit aux 20 stations ayant des séries comportant au moins 20 années d'observation, que l'on peut comparer à la figure 3a, obtenue en considérant les 39 séries de plus de 10 observations. Pour le réseau restreint aux 20 stations les plus longues, le séparateur s'avère être le coefficient d'asymétrie, alors que pour le réseau complet le coefficient de variation intervient dans la séparation des groupes. Cependant, les 8 groupes de la figure 3a apparaissent être des sous-

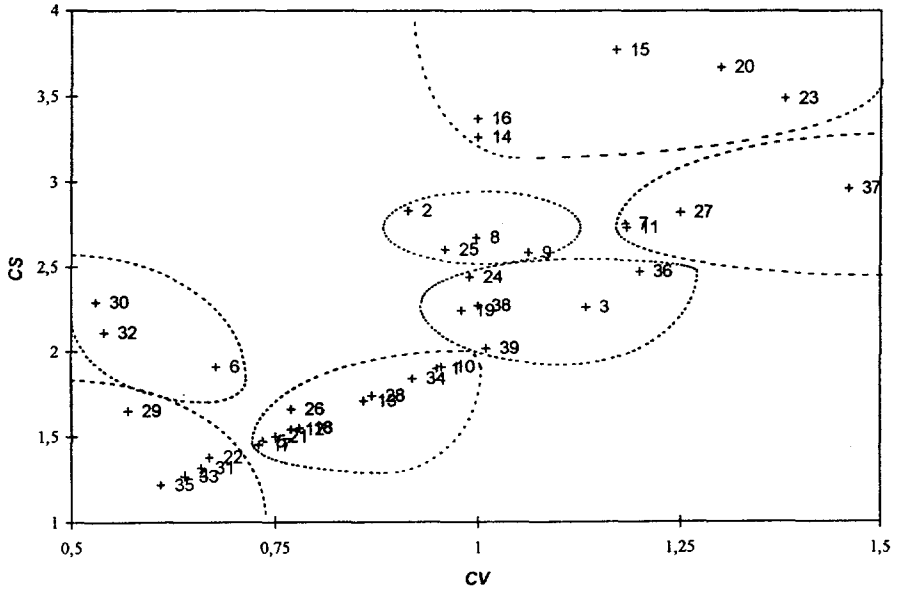

Figure 3a Groupes obtenus avec Iphigénie pour (CV, CS) avec les 39 séries du réseau.

Clusters obtained for Iphigénie in the (CV,CS) plane with the 39 stations in the network.

structures des 4 groupes de la figure 3b. L'organisation en groupes dans ce plan, ne semble donc pas dépendre de la taille d'échantillon. Il suffit pour s'en assurer d'examiner la figure 3c dans laquelle la répartition des stations dans le plan P2 : (*CV,CS*) est accompagnée de la taille d'échantillon. Les stations 6, 29, 30 et 32 dont le coefficient d'asymétrie varie de 1,65 à 2,30 mais qui ont un coefficient de variation relativement réduit (entre 0,53 et 0,74) ont des tailles d'échantillons respective de 13, 18, 18 et 21. Leur position dans ce plan ne semble donc pas être attribuée à leur faible taille.

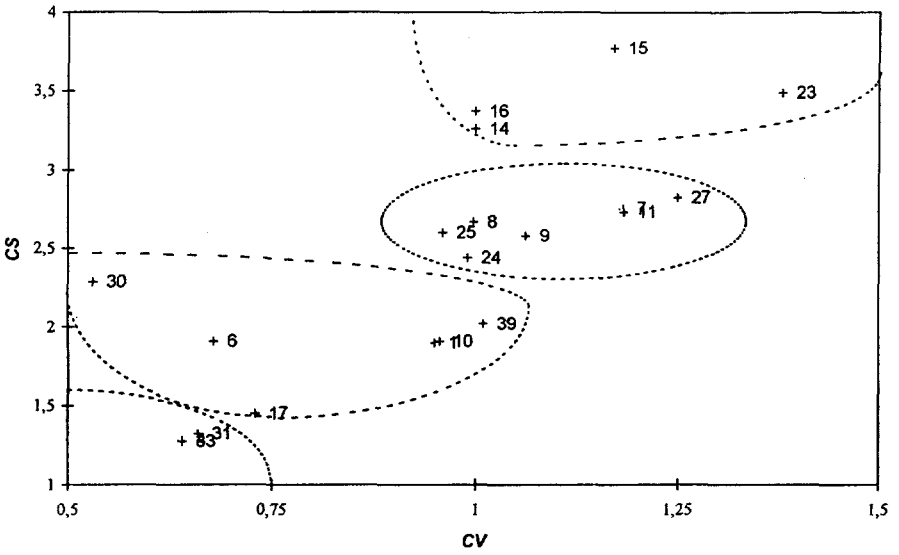

Figure 3b Groupes obtenus à l'aide d'Iphigénie dans le plan (CV,CS) à partir des séries comportant au moins 20 observations.

Clusters obtained with Iphigénie in the (CV,CS) plane for stations having at least 20 observations.

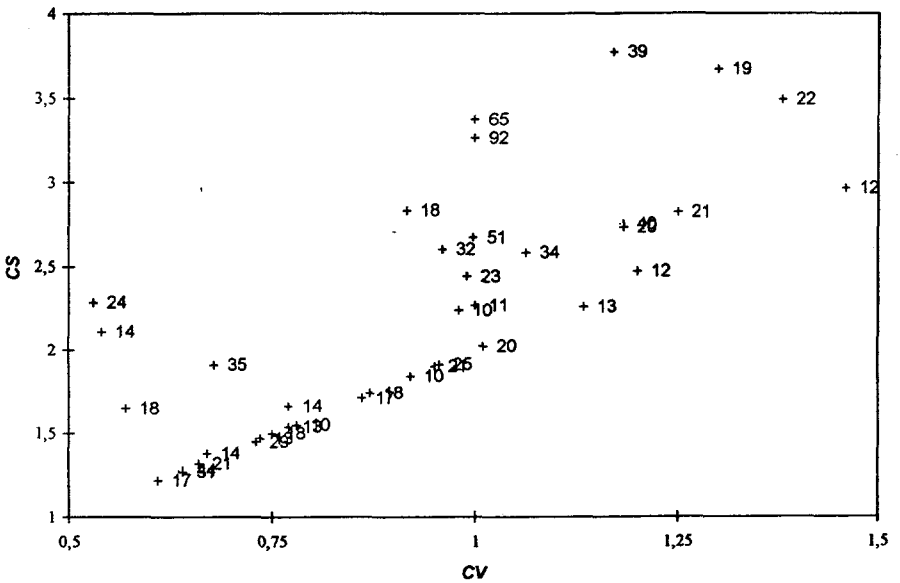

Figure 3c Taille d'échantillon des séries dans le plan (CV,CS).

Sample size available at each station in the (CV,CS) plane .

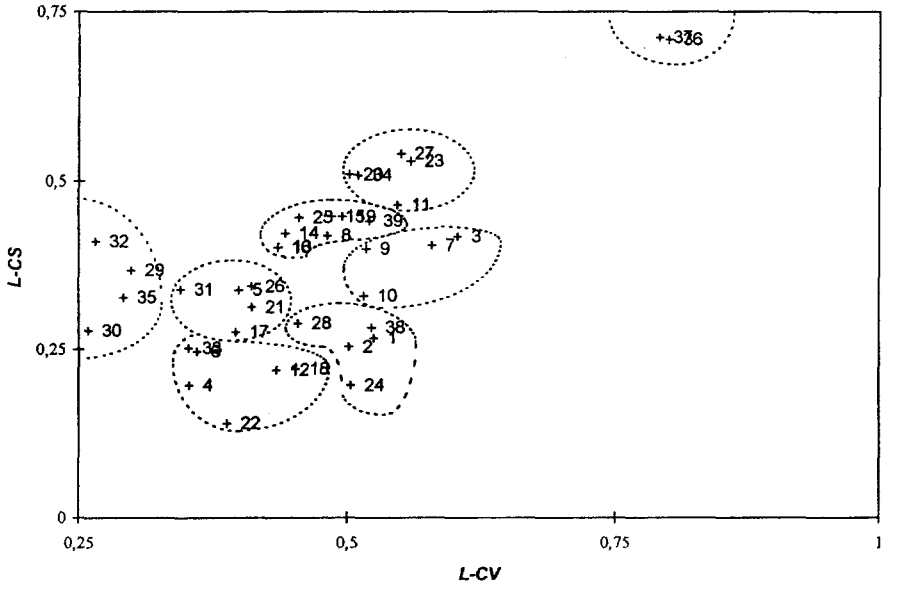

Figure 4 Groupes obtenus à l'aide de Iphigénie dans le plan (L-CV,L-CS).
Clusters obtained with Iphigénie in the (L-CV, L-CS) plane.

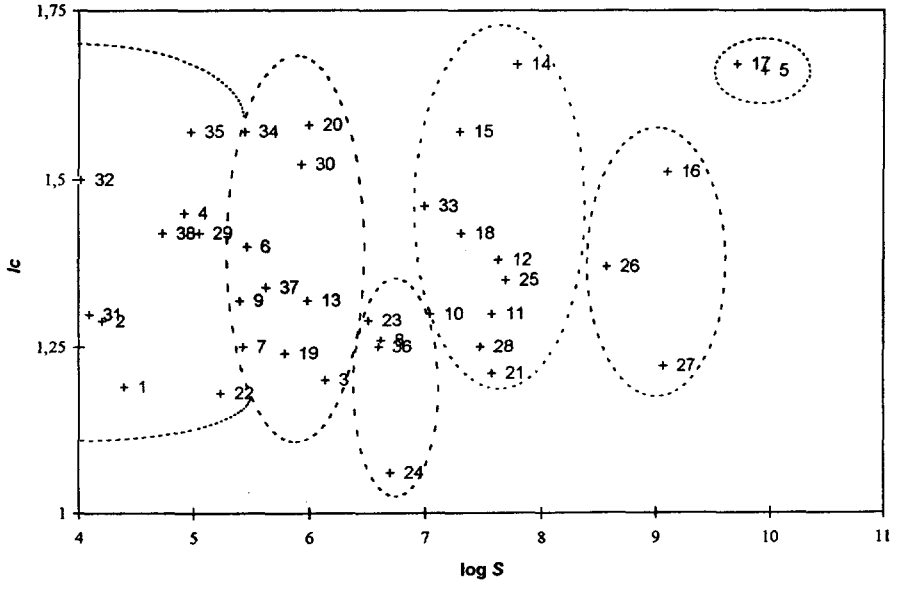

Figure 5 Groupes obtenus à l'aide de Iphigénie dans le plan (log S, Ic).
Clusters obtained with Iphigénie in the (log S, Ic) plane.

Comparaison avec la méthode de Burn (région d'influence) et des centres mobiles

La procédure Iphigénie permet de déterminer la distance limite à partir de laquelle les stations ne sont plus voisines : c'est la distance correspondant à l'arrêt de la procédure. Il s'agit de la distance critique qui constitue la base de la méthode de Burn, et qui, au lieu d'être fixée par l'opérateur (BURN, 1990) est ici un résultat de la procédure d'agrégation. Ainsi, la méthode d'agrégation par cohérence permet-elle de définir objectivement une distance critique pour l'ensemble des sites et de retrouver la région d'influence pour chacun d'eux. Elle indique également la région d'influence de chaque site. Pour cela, il suffit de recenser pour un site cible les stations dont la distance est inférieure à la distance critique. Ce recensement conduit à la construction d'une matrice de la relation entre sites, matrice dont les éléments prennent soit la valeur zéro soit l'unité. Un des résultats de la procédure Iphigénie est le nombre de classes d'équivalence, ce qui constitue un avantage sur des méthodes telles que celles des centres mobiles et ISODATA, pour lesquelles le nombre de classes doit être défini *a priori* (bien que des méthodes heuristiques que nous étudierons permettent de l'estimer).

Procédure ISODATA

Choix du nombre de classes floues et du paramètre p

Quelle est la validité des agrégats obtenus par l'algorithme de classification ? Combien de sous-ensembles flous sont réellement présents dans l'ensemble des stations ? Cela revient à valider le choix de R. On trouve dans BEZDEK (1981) une revue des méthodes de validation des agrégats. Parmi celles-ci, nous en avons choisi deux, basées respectivement sur le coefficient de partition et l'entropie de la partition.

1) le coefficient de partition F, est un scalaire compris entre $1/R$ et 1, défini par :

$$F(U,R) = \sum_{k=1}^{N} \sum_{i=1}^{R} \mu_i^2(k) / N \qquad (9)$$

C'est un indicateur qui dépend des NxR éléments de la matrice d'appartenance. *F est relié à la matrice de similarité S entre les régions floues, définie par :*

$$S_{ij} = \sum_{k=1}^{N} \mu_i(k) \cdot \mu_j(k) / N \qquad (10)$$

$F = 1$ équivaut à $S_{ij} = 0$ pour tout $i \neq j$.

Si F est à sa valeur maximale ($F = 1$), il n'y a pas d'appartenance partagée entre les groupes et les régions sont strictes. Si F est à sa valeur minimale ($F = 1/R$), $\mu_i(k) = 1/R \; \forall \; k = 1,N$ *et* $i = 1,R$, dénotant l'équi-appartenance des stations aux régions et conduisant à la partition la plus incertaine. Le principe de la recherche heuristique de R est le suivant : si l'ensemble des stations contient réellement des groupes distincts, la classification floue devrait conduire à des valeurs de F proches de 1. Aussi, une façon de déterminer R consiste à maximiser F, en faisant varier R. Nous avons considéré R = 2 à 7, avec $\varepsilon = 0{,}01$.

2) L'entropie associée à la partition

Comme le soulignent YAGER et FILEV (1994), l'essence du caractère flou d'un sous-ensemble A réside dans l'absence d'une distinction claire entre A et son

complément A_c. DE LUCCA et TERMINI (1972) ont proposé trois conditions à satisfaire par une fonction FUZ(A) de mesure du flou de A à savoir :

i) $FUZ(A) = 0 \leftrightarrow \mu_A(xi) = \{0,1\}\ \forall i = 1,N$ traduisant le caractère strict de A.

ii) FUZ(A) atteint son maximum pour $\mu_A(xi) = 0,5\ \forall i = 1,N$

iii) A* est un sous-ensemble flou dont le flou est moins marqué que celui de A : $FUZ(A) \geq FUZ(A^*)$ si
A* est tel que : $\mu_{A^*}(xi) \geq \mu_A(xi)$ si $\mu_A(xi) > 0,5$
et $\mu_{A^*}(xi) < \mu_A(xi)$ si $\mu_A(xi) < 0,5$

En s'inspirant du concept d'entropie statistique de Shannon, DE LUCCA et TERMINI (1972) ont proposé de mesurer le flou d'un sous-ensemble A (xi, $\mu_A(xi)$, i = 1,N) par la fonction « entropie normalisée » de A et de son complément noté A_c définie par :

$$h(\mu_A,\mu_{Ac}) = -\{\Sigma [\mu_A(xi) \log_a \mu_A(xi) + (1-\mu_A(xi)) \log_a \mu_A(xi)]\}/N \quad (11)$$

où $\log_a(\cdot)$ représente le logarithme à base $a \in (1,\infty)$. Rappelons que la fonction d'appartenance de A_c est $\mu_{Ac}(xi) = 1 - \mu_A(xi)$.

Ils ont démontré que la fonction $h(\mu_A,\mu_{Ac})$ satisfait les trois conditions précédentes. C'est BEZDEK (1981) qui a proposé le concept d'entropie de la partition notée H avec :

$$H(U,R) = -\sum_{k=1}^{N}\sum_{i=1}^{R}\mu_i(k)\cdot\log\mu_i(k)/N \quad (11)$$

où $\log(\cdot)$ représente le logarithme népérien et où $\mu_i(k)\cdot\log\mu_i(k) = 0$ si $\mu_i(k) = 0$. Dans le cas où R = 2, il montre qu'en raison de la condition (1), les expressions (11) et (12) sont équivalentes.

BEZDEK (1981) montre également que :

$$0 \leq H(U,R) \leq \log R \quad (13)$$

$H = 0$ témoigne d'une partition stricte alors que $H = \log R$ témoigne d'une partition incertaine car alors, les fonctions d'appartenance obtenues sont aussi partagées que possible. Minimiser H constitue ainsi une stratégie permettant de choisir le nombre de classes R lorsque des groupes bien définis sont supposés exister. BEZDEK (1981) montre que :

$$0 \leq 1-F(U,R) \leq H(U,R) \quad (14)$$

H et F, qui mesurent le flou d'une partition donnée, agissent en sens inverse : quand F augmente, H diminue. Toutefois, il a été remarqué que H est plus sensible que F à la variation de R (BEZDEK, 1981). Par ailleurs, il peut arriver que la meilleure valeur pour R obtenue en minimisant H ne coïncide pas avec celle obtenue en maximisant F. Dans plusieurs cas cependant, cette méthode heuristique de minimisation du flou ne donne pas le résultat escompté : le nombre de groupes sélectionné est souvent trop faible. Pour pallier à cette difficulté, il est en général préférable de maximiser le nombre de groupes tout en minimisant le flou, ce qui revient à faire un compromis puisque ces objectifs sont en général opposés. Une façon de faire consiste à choisir une valeur de R pour laquelle les indices de flou se stabilisent. C'est ainsi que nous procéderons, après avoir fixé la valeur du paramètre p. Ce dernier est usuellement fixé à 2 (BEZDEK, 1981), et sa valeur influence le degré de flou de la partition : plus il est élevé, plus la partition obtenue est floue. D'ailleurs, lorsqu'il tend vers 1, la partition floue tend vers une partition stricte. Le tableau 2 montre cet effet : pour trois plans de travail, les indices de flou H et F ont été calculés pour $R = 4$ et $R = 7$ en considérant $p = 1,5$ et $p = 2$.

Tableau 2 Critères F et H évalués pour trois plans avec R = 4 et R = 7, p = 1,5 et p = 2.

Table 2 *F and H criteria computed from three planes, with R = 4 and R = 7, p = 1,5 and p = 2.*

Plan	R = 4		R = 7		p
	F	H	F	H	
P1 : (log Q_S, CV)	0,90	0,17	0,88	0,24	
P2 : (CV,CS)	0,90	0,18	0,89	0,21	1,5
P3 : (L-CV, L-CS)	0,87	0,24	0,83	0,35	
P1 : (log Q_S, CV)	0,79	0,43	0,75	0,54	
P2 : (CV, CS)	0,74	0,51	0,68	0,68	2,0
P3 : (L-CV, L-CS)	0,67	0,67	0,60	0,86	

Comme prévu, quel que soit le plan de travail, $p = 1,5$ conduit à des valeurs nettement plus fortes pour F et plus faibles pour H, R étant fixé. Pour tenir compte du caractère flou des régions homogènes, p a été fixé à 2. D'après les figures 6a

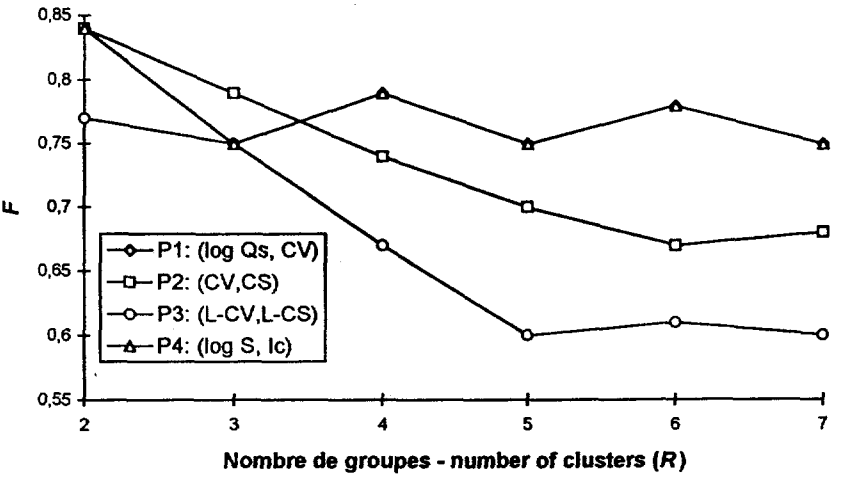

Figure 6a Variation du critère F en fonction de R.

Variation of the F criterion with R.

et 6b qui montrent la variation de F et H selon R, les trois autres espaces se distinguent de P3 du fait qu'ils présentent une classification au caractère plus strict. Les plans P1 et P4 sont très ressemblants. Quel que soit l'espace d'étude, $R = 2$ est retenu comme la meilleure valeur pour R. Cependant, la plus forte augmentation pour H s'observe à partir de $R = 3$ pour P2 et P3, H ayant tendance à se stabiliser à partir de $R = 4$ pour P1 et P4. Dans tous les cas, un palier s'esquisse à partir de $R = 4$ qui a donc été retenue comme valeur de compromis pour le problème à étudier.

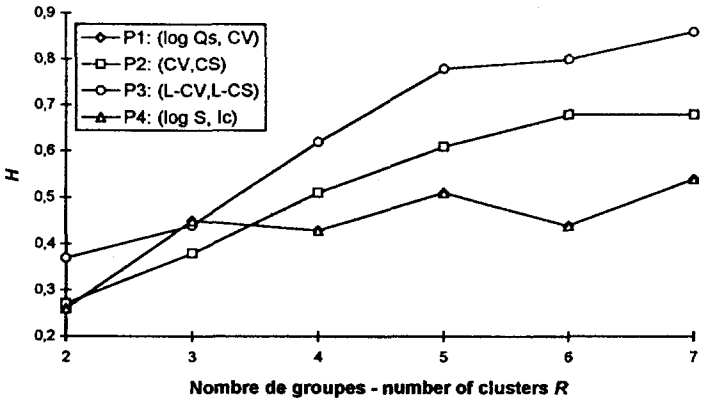

Figure 6b Variation du critère H en fonction de R.

Variation of the H criterion with R.

Matrices d'appartenance

L'examen des matrices d'appartenance montre qu'il y a peu de partage des stations entre les régions floues obtenues (figures 7a à 7d donnant sup{$\mu(i,k)$} ; i = 1,4, k = 1,39). À titre d'exemple, dans le plan P2 : (*CV,CS*) *(figure 7b)*, les stations que nous décrirons comme étant « à appartenance multiple » et que nous définirons comme vérifiant sup{$\mu(i,k)$} < 0,5 sont seulement au nombre de 5 sur 39. Sachant

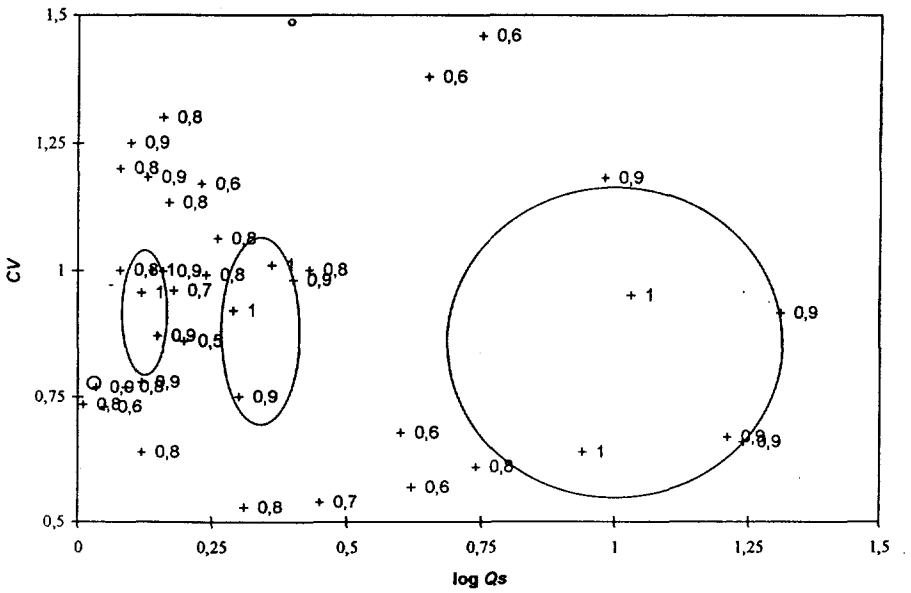

Figure 7a Courbes d'iso-appartenance m = 0,9 pour (log Qs, CV) obtenues par ISODATA.

Lines of iso-membership of level m = 0.9 in the (log Qs, CV) plane obtained using ISODATA.

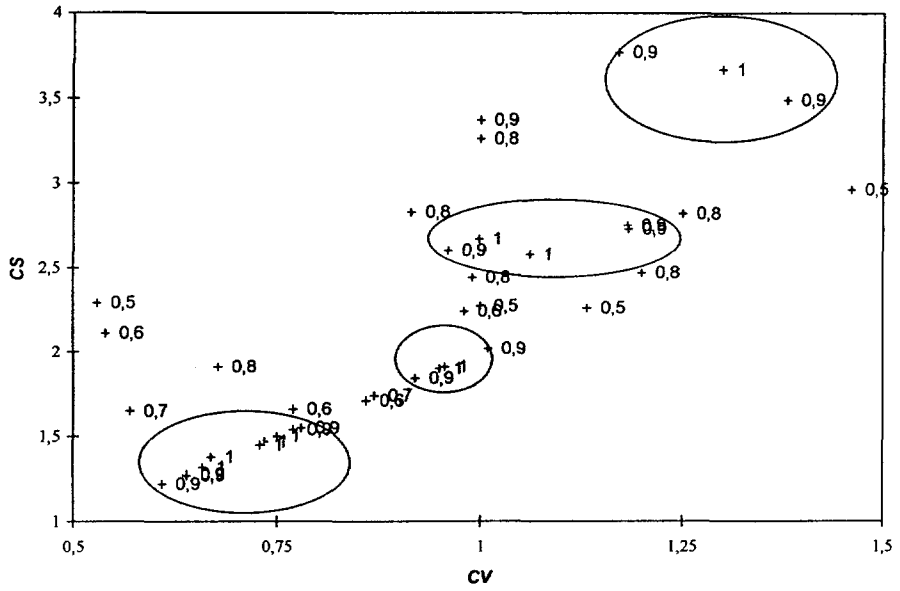

Figure 7b Courbes d'iso-appartenance m = 0,9 pour (CS, CV) obtenues par ISODATA.

Lines of iso-membership of level m = 0.9 in the (CS, CV) plane obtained using ISODATA.

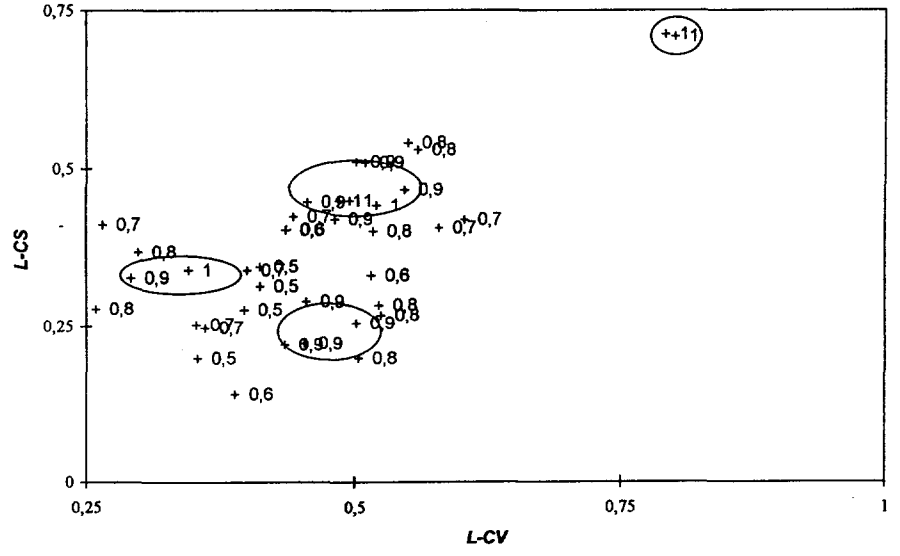

Figure 7c Courbes d'iso-appartenance m = 0,9 pour (L-CV, L-CS) obtenues par ISODATA.

Lines of iso-membership of level m = 0.9 in the (L-CV, L-CS) plane obtained using ISODATA.

que le noyau d'un sous-ensemble flou est défini comme l'ensemble des éléments dont le degré d'appartenance est égal à 1 (YAGER et FILEV, 1994), nous observons que les noyaux des régions floues obtenues peuvent être au nombre de deux, trois ou quatre. Soit $\alpha \in [0,1]$, l'ensemble des α-coupes A_a est défini (YAGER et FILEV, 1994) comme l'ensemble dont les éléments ont un degré d'appartenance à A supérieur ou égal à α. Pour $\alpha = 0,9$, pour chacun des quatre plans et pour $R = 4$, les figures 7a, 7b, 7c et 7d montrent le domaine couvert par les stations dont $\sup\{\mu(i,k)\} \geq ,9$. Leur nombre peut aller jusqu'à 6, montrant jusqu'à quel point ces régions sont strictes. La configuration des noyaux et des ensembles des $\alpha = 0,9$-coupes, permet de dégager les séparateurs des régions homogènes. Pour P1 *(figure 7a)*, le principal critère de séparation des groupes est $Q_s = \overline{Q}/S$, alors que pour P4 *(figure 7d)* il s'agit de *S*. La surface des bassins a donc une grande importance dans la formation des groupes. Pour P2 *(figure 7b)*, c'est *CS* qui constitue le séparateur entre groupes, confirmant les résultats obtenus par la méthode Iphigénie. C'est dans ce plan que l'appartenance partielle des stations à un groupe est la plus prononcée. Dans le plan P3 *(figure 7c)*, les stations 36 et 37 forment un groupe distinct que l'on aperçoit en haut à droite. On remarquera sur la carte géographique de la Tunisie *(figure 1)* qu'il s'agit des deux stations les plus au sud du territoire : étant donné les variations de climat en Tunisie avec la latitude, ce groupe distinct correspond donc à une réalité physique. Les trois autres groupes présentent une valeur minimale pour μ de 0,56. Ces figures confirment ainsi le caractère pratiquement strict des régions.

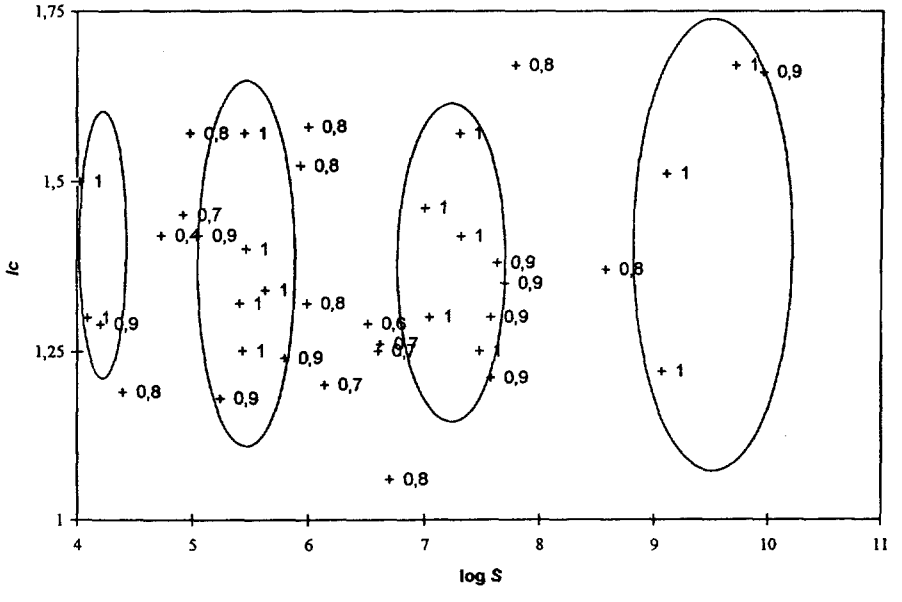

Figure 7d Courbes d'iso-appartenance m = 0,9 pour (log S, Ic) obtenues par ISODATA.

Lines of iso-membership of level m = 0.9 in the (log S, Ic) plane obtained using ISODATA.

Sensibilité des résultats à la métrique choisie

Un aspect arbitraire relié à l'application d'une méthode d'agrégation dans un espace non géographique est le choix de la métrique. Nous avons choisi d'utiliser une distance euclidienne, accordant ainsi le même poids à chaque variable de l'analyse. Comme nous n'avons considéré que des espaces à deux variables, la distance $\mathbf{D}_{i,j}$ entre deux stations $s_i = (x_i,y_i)$ et $s_j = (x_j,y_j)$ peut donc être obtenue par l'équation $\mathbf{D}_{i,j} = (x_i-x_j)^2 + (y_i-y_j)^2$. Lorsque l'on utilise une telle distance, le choix de l'unité de mesure des variables influence évidemment la mesure de la distance. C'est pourquoi nous avons appliqué une transformation logarithmique à certaines variables (le débit spécifique *Qs* et la surface *S*) ; ainsi, l'échelle de mesure des variables étudiées est similaire. Il est possible de modifier la métrique utilisée pour éliminer l'effet du choix de l'unité de mesure en divisant chaque variable par sa variance. On obtient ainsi $\mathbf{D}_{i,j} = (x_i-x_j)^2/\sigma_x^2+(y_i-y_j)^2/\sigma_y^2$. Les groupements présentés précédemment ont été refaits avec cette métrique différente sans que n'apparaisse de différences importantes dans l'analyse des résultats : l'importance de la surface dans la formation des groupes est ainsi confirmée.

CONSTRUCTION DES COURBES RÉGIONALES D'INDICE DE CRUE

La méthode de l'indice de crues proposée par DALRYMPLE (1960) permet une estimation régionale des crues en un site. Cette méthode suppose que la région dont est extraite l'information est homogène : c'est-à-dire que les fonctions de répartition des débits en différents sites ne s'y distinguent que par un facteur d'échelle qui est l'indice de crues, Q_m. On appelle variable standardisée le rapport du débit à l'indice de crue. L'intérêt de choisir la moyenne de la distribution en chaque site comme indice de crue réside dans le fait que la variable standardisée a alors pour moyenne l'unité. Soit $Q_r(i)$ la variable standardisée régionale pour chaque région $i = 1,\ldots,R$. Lorsque la distribution de cette variable possède uniquement deux paramètres (lois Gumbel ou Gamma par exemple), en appliquant la méthode des moments, on peut écrire une première équation avec :

$$E(Q_{r(i)}) = 1 \qquad (15)$$

On peut écrire la seconde équation en utilisant la statistique régionale $CV_{r(i)}$, c'est-à-dire le coefficient de variation pour la région *i*. Lorsque la distribution de la variable standardisée possède trois paramètres (lois GEV ou Pearson de type 3 par exemple), une des méthodes régionales proposées consiste à fixer le paramètre de forme régional (RASMUSSEN *et al.*, 1994), de façon à se retrouver dans des conditions analogues au cas précédent. Deux distributions ont été considérées dans cette étude : une distribution à 2 paramètres, la distribution gamma, et une distribution à 3 paramètres, la distribution Pareto. Pour la distribution Pareto, de plus en plus utilisée en hydrologie, nous avons appliqué la méthode d'estimation des L-moments (HOSKING, 1990) en utilisant les estimateurs non biaisés plutôt que ceux issus des probabilités empiriques (*plotting position*). Pour la distribution gamma nous avons procédé ainsi :

1) Pondération des estimateurs locaux du coefficient de variation pour obtenir un estimateur régional $CV_{r(i)}$ pour chaque région $i = 1,2,\ldots,R$: chaque estimateur

local est pondéré par la taille d'échantillon. Seules les stations ayant un degré d'appartenance à une région supérieur ou égal à 0.5 sont considérées dans le calcul de $CV_{r(i)}$:

$$CV_{r(i)} = \sum_{k=1}^{N} I(\mu_i(k) \geq 0.5) CV_k \cdot n_k \Big/ \sum_{k=1}^{N} I(\mu_i(k) \geq 0.5) \cdot n_k \qquad (16)$$

où CV_k est la valeur du coefficient de variation au site k et n_k est le nombre d'observations disponibles à ce site, et $I(P)$ est une fonction valant 1 si la proposition P est vraie et 0 sinon.

2) Calcul des quantiles de la variable standardisée régionale $Q_{r(i)}(T)$: on utilise la relation $CS = 2.CV$ qui relie la valeur théorique du coefficient d'asymétrie au coefficient de variation pour estimer le coefficient d'asymétrie régional $CS_{r(i)}$, ce qui revient à utiliser une méthode des moments régionale, et l'on estime par la suite les quantiles à partir du facteur de fréquence $K_T(CS)$ de la distribution Pearson type 3 standardisée (BOBÉE et ASHKAR, 1991) :

$$Q_{r(i)}(T) = 1 + K_T(CS_{r(i)}).CV_{r(i)} \qquad (17)$$

COMPARAISON DE L'ESTIMATION LOCALE ET RÉGIONALE

Pour le quantile correspondant à une période de retour $T = 100$ ans, il est intéressant de comparer différents estimateurs régionaux entre eux, et avec l'estimation locale. Les trois estimateurs régionaux comparés ici sont ceux obtenus à l'aide des procédures Iphigénie et ISODATA ainsi qu'avec la méthode de la région d'influence de BURN (1990), en donnant un poids égal à toutes les stations constituant la région d'influence. La distance critique pour la méthode de Burn est celle issue de l'application de la procédure Iphigénie. L'estimation du quantile à un site k élément d'une région i est donné par (RASMUSSEN *et al.*, 1994) :

$$Q_k(T) = Q_{m(k)} \sum_{i=1}^{R} \alpha_{ik} Q_{r(i)}(T) \qquad (18)$$

où $Q_{m(k)}$ est l'indice de crues de la série associée à la station k, et a_{ik} est le poids donné à la station dans sa région. Étant donné le caractère strict des régions, il ne nous a pas semblé opportun de proposer, à l'instar de FORTIN *et al.* (1995) de prendre pour a_{ik} la valeur de la fonction d'appartenance de la station k à la région i. Nous avons donc considéré dans (18) une fonction indicatrice I dépendant du degré d'appartenance :

$$\alpha_{ik} = I(\mu_i(k) \geq 0.5) = \begin{cases} 1 & \text{si la station i appartient à la région k (i.e. } \mu_i(k) \geq 0.5) \\ 0 & \text{sinon} \end{cases} \qquad (19)$$

Estimation de l'indice de crues à des sites non jaugés

Dans ce type d'approche, la qualité d'estimation du quantile régional dépendra en grande partie de la qualité de l'estimation de l'indice de crues, obtenu généralement à des sites non jaugés par une régression entre l'indice de crues et les caractéristiques physiographiques. Du fait de la structure de la matrice de corrélation entre ces dernières *(tableau 1)*, les variables indépendantes à inclure

dans une régression multiple seraient la surface S, l'indice de compacité I_c et la pente moyenne du cours d'eau I_g qui sont les variables les moins corrélées entre elles. Une alternative pour prendre en compte la colinéarité au lieu d'éliminer des variables serait d'utiliser la régression Ridge (HOERL et KENNARD, 1970). Les meilleures reconstitutions d'indice de crues ont été obtenues par un regroupement géographique préalable de stations, les stations 5 et 26 n'ayant pu toutefois être convenablement incluses dans aucun de ces regroupements. L'existence dans leur bassin d'une importante zone d'infiltration pourrait en être la cause. La figure 8 reproduit l'indice de crues calculé en fonction de celui observé. Le coefficient de corrélation est de $r = 0{,}93$.

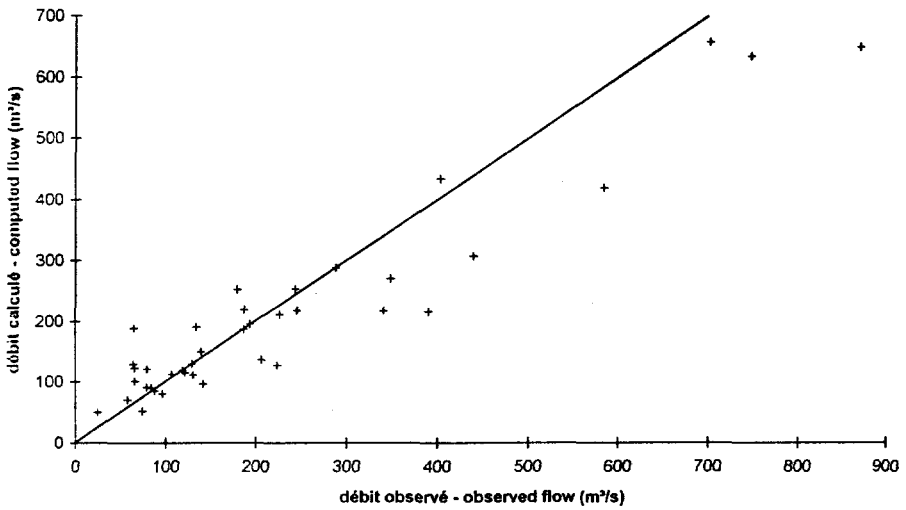

Figure 8 Débit moyen annuel calculé et observé.
Computed versus observed average annual flow.

DISCUSSION

Plusieurs critères de comparaison d'une méthode régionale (associée à un procédé de délimitation de régions homogènes) avec l'estimation locale peuvent être considérés. Parmi eux ont été considérés :

– le coefficient de corrélation r entre le quantile estimé au site et avec la méthode régionale ;
– le biais relatif (*i.e.* l'écart relatif moyen) ;
– l'écart-type de l'erreur.

Pour la méthode Iphigénie (associée à la procédure régionale présentée précédemment et sous l'hypothèse d'une loi Gamma), les résultats *(tableau 3)* montrent que le plan P2 : (*CV,CS*) semble le plus approprié pour reconstituer les débits centennaux (biais et écart-type d'estimation les plus faibles et coefficient de corrélation le plus élevé). De plus, les trois plans hydrologiques ont un coeffi-

Tableau 3 Comparaison de l'estimation locale et régionale pour Iphigénie (T = 100 ans).

Table 3 *Comparison of local and regional estimation for Iphigénie (T = 100 years).*

Critère	P1 : (log Q_s, CV)	P2 : (CV,CS)	P3 : (L-CV, L-CS)	P4 : (log S, I_c)
Biais relatif	− 0,14	− 0,19	− 0,16	− 0,10
Écart-type	0,48	0,54	0,58	0,43
r	0,91	0,87	0,91	0,92

cient de corrélation légèrement plus important que le plan physiographique P4 : (log *S*, I_c). Tout de même, il est intéressant de constater que la surface du bassin versant *S* et le coefficient de compacité I_c sont des descripteurs puissants, expliquant environ 75 % de la variance d'estimation des débits centennaux de crue.

Il semble cependant qu'un biais négatif persiste pour les 4 plans. Toutefois, selon la figure 9 où sont comparés les trois procédés de régionalisation (Iphigénie, ISODATA, méthode de Burn) pour le plan P1 : (log Q_s, *CV*) ce biais est beaucoup plus faible lorsque la taille d'échantillon dépasse 20. Cette figure ne permet cependant pas de donner une nette préférence à l'un des trois procédés. Par contre, le coefficient de corrélation entre le quantile local et le quantile régional est de ,91 pour Iphigénie et ,85 à la fois pour ISODATA et la méthode de Burn, ce qui place Iphigénie au premier rang.

L'hypothèse d'une distribution parente de type Pareto pour les débits aux stations conduit à une réduction notable du biais *(figure 10)* pour les petites tailles d'échantillon et confirme l'avantage de la méthode Iphigénie, tout en suggérant l'utilisation d'une distribution à 3 paramètres pour l'extrapolation (*T*>*n*). Cependant, le choix d'une distribution reste un problème épineux qu'il faudrait étudier plus en détail. Pour des stations de l'Ontario et du Québec (Canada), FORTIN

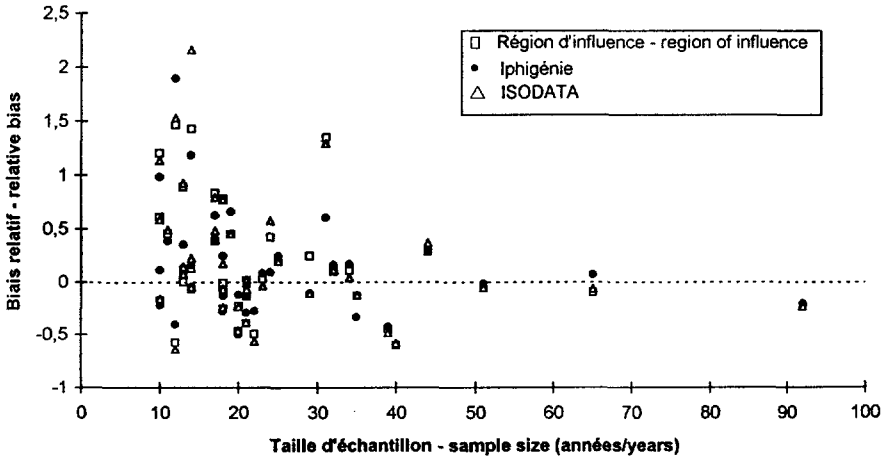

Figure 9 Biais relatif de l'estimation régionale du débit de période de retour T = 100 ans en fonction de la taille d'échantillon (distribution Gamma).

Relative bias of regional estimate of the 100-year flood against sample size (Gamma distribution).

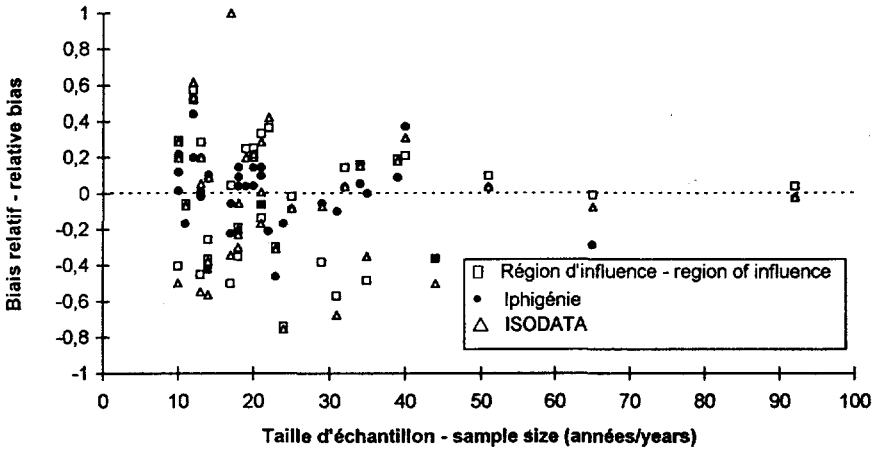

Figure 10 Biais relatif de l'estimation régionale du débit de période de retour T = 100 ans en fonction de la taille d'échantillon (distribution Pareto).

Relative bias of regional estimate of the 100-year flood against sample size (Pareto distribution).

et al. (1997) ont montré que certaines distributions à 3 paramètres, lorsque utilisées de façon parcimonieuse, permettent de réduire l'erreur quadratique moyenne pour l'extrapolation, mais que leur utilisation systématique mène à des résultats moins satisfaisants que ceux des distributions Gumbel et gamma, qui n'ont que 2 paramètres.

CONCLUSION

Deux méthodes de délimitation de régions hydrologiques homogènes basées sur la théorie des sous-ensembles flous ont été appliquées (Iphigénie et ISODATA). La première de ces méthodes utilise un procédé d'agrégation par cohérence et conduit, par construction, à des classes strictes. La seconde est une extension de la méthode de centres mobiles et conduit à la définition de classes floues grâce à la détermination des fonctions d'appartenance des sites à un nombre préfixé de classes. Ces méthodes ont été mises en œuvre dans deux plans hydrologiques, (*CV,CS*) et (*L-CV,L-CS*), un plan physiographique (log *S*, I_c) et un plan hybride (Q_s,*CV*). En considérant les crues maximales annuelles de Tunisie, il est apparu que les classes floues ont un caractère assez strict, l'appartenance multiple de stations à plusieurs régions étant plutôt une exception. La comparaison des estimateurs régionaux obtenus en utilisant les classes ainsi identifiées avec les estimateurs locaux conduit à une légère préférence pour la méthode d'agrégation par cohérence (Iphigénie), par rapport à l'approche ISODATA et la méthode de la région d'influence de Burn (1990). Le fait que la surface des bassins versants ait été identifiée, dans les espaces physiographique et hybride par les deux méthodes comme un facteur important pour la séparation de stations

hydrométriques en groupes homogènes présente des conséquences pratiques dans la conception des réseaux : il sera important de s'assurer que des bassins de tailles variées fassent partie du réseau de mesure utilisé pour l'estimation régionale.

REMERCIEMENTS

Ce travail a été réalisé grâce à l'aide financière du Fond Universitaire de Coopération universitaire de l'Aupelf 93/PAS/31(121). Z. Bargaoui voudrait remercier H. Muster de l'IHW (Karlsrhue) pour son aide précieuse dans la recherche bibliographique.

RÉFÉRENCES BIBLIOGRAPHIQUES

ACREMAN M.C., SINCLAIR C.D., 1986. Classification of drainage basins according to their physical characteristics; an application for flood frequency analysis in Scotland. *Journal of Hydrology*, 84: 365-380.

ACREMAN M.C., WILTSHIRE S.E., 1989. The regions are dead: long live the regions. Methods of identifying and dispensing with regions for flood frequency analysis, Dans: *Friends in Hydrology*, IAHS Publication 187: 175-188.

BERNARD G., BESSON M.L., 1971. Douze méthodes d'analyse multicritère. R.I.R.O., no V-3.

BEZDEK J.C., 1981. *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, New York.

BOBÉE B., ASHKAR F., 1991. *The Gamma Family and Derived Distributions Applied in Hydrology*, Water Resources Publications, Littleton, Colorado, 203 p.

BURN D.H., 1988. Delineation of groups for regional flood frequency analysis. *Journal of Hydrology*, 104 (3/4): 345-361.

BURN D.H., 1989. Cluster Analysis as Applied to Regional Flood Frequency. *Journal of Water Resources Planning and Management*, 115 (5): 567-582.

BURN D.H., 1990. Evaluation of Regional Flood Frequency Analysis with a Region of Influence Approach, *Water Resources Research* 26 (10): 2257-2265.

CAVADIAS G., 1990. The canonical correlation approach to regional flood estimation. *Regionalization in hydrology*, IAHS publication 191, pp. 171-178.

DALRYMPLE T., 1960. *Flood Frequency Analysis*. U.S. Geological Survey, Water Supply Paper 1543-A.

DE LUCA A., TERMINI S.A., 1972. A definition of a non probabilistic Entropy in the setting of Fuzzy sets Theory, *Inf. Control*, (20): 301-312.

DIDAY É, LEBART L., 1977. L'analyse des données. *La Recherche*, 74 : 15-25.

DUDA R., HART P., 1973. *Pattern Classification and Scene Analysis*, Wiley, New York.

FORTIN V., BOBÉE B., BERNIER J., 1997. A rational approach to the comparison of flood distributions by simulation. *ASCE Journal of Hydrologic Engineering*, 2 (3): 95-103.

FORTIN V., BOBÉE B., DUCKSTEIN L., BARGAOUI Z., 1995. Détermination floue des zones hydrologiques homogènes. In: *Modeling and Management of Sustainable Basin-scale Water Resource* Systems, IAHS publ. 231, pp. 367-375.

HOERL, KENNARD, 1970. Ridge Regression: Applications to Nonorthogonal Problems. *Technometrics*, 12 (1): 55-67.

HOSKING J.R.M., 1990. L-moments : analysis and estimation of distributions using linear combinations of order statistics. *Journal of the Royal Statistical Society B*, 52: 105-124.

KAUFMANN A., 1975. *Introduction à la théorie des sous-ensembles flous à l'usage des ingénieurs (Fuzzy Sets Theory) 3. Applications à la classification et à la reconnaissance des formes, aux automates et aux systèmes, au choix des critères*, Masson, Paris.

KIRBY W., 1974. Algebraic Boundness of Sample Statistics. *Water Resources Research*, 10(2): 220-222.

MOSLEY M.P., 1981. Delimitation of New Zealand hydrologic regions. *Journal of Hydrology*, 49: 173-192.

RASMUSSEN P.F., BOBÉE B., BERNIER J., 1994. Une méthodologie générale de comparaison de modèles d'estimation régionale de crue. *Revue des sciences de l'eau*, 7 : 23-41.

RUSPINI E.H., 1970. Numerical Methods for Fuzzy Clustering. *Information Sciences*, 2: 319-350.

TERANO T., ASAI K., SUGENO M., 1992. *Fuzzy Systems Theory and its Applications*, Academic Press.

WILTSHIRE S.E., 1986. Regional flood frequency analysis, II: Multivariate classification of drainage basins in Britain, *Hydrol. Sci. J.* 31: 334-346.

YAGER R.R., FILEV D.P., 1994. *Essentials of fuzzy modeling and control*, John Wiley and Sons. N.Y.