Record Number: 21170
Author, Monographic: Girard, C.//Ouarda, T. B. M. J.//Bobée, B.
Author Role:
Title, Monographic: Une approche par classification à la constitution de voisinages homogènes basés sur l'ACC
Translated Title:
Reprint Status:
Edition:
Author, Subsidiary:
Author Role:
Place of Publication: Québec
Publisher Name: INRS-Eau
Date of Publication: 2000
Original Publication Date: Novembre 2000
Volume Identification:
Extent of Work: iii, 22
Packaging Method: pages
Series Editor:
Series Editor Role:
Series Title: INRS-Eau, rapport de recherche
Series Volume ID: 576
Location/URL:
ISBN: 2-89146-444-3
Notes: Rapport annuel 2000-2001
Abstract:
Call Number: R000576
Keywords: rapport/ ok/ dl

Une approche par classification à la constitution de voisinages homogènes basés sur l'ACC

Rapport de recherche No R-576 *Novembre 2000*

Une approche par classification à la constitution de voisinages homogènes basés sur l'ACC

Rapport préparé par :

Claude Girard
Taha B.M.J. Ouarda
Bernard Bobée

Chaire industrielle en Hydrologie statistique
Institut national de la recherche scientifique, INRS-Eau
2800, rue Einstein, Case postale 7500, Sainte-Foy (Québec), G1V 4C7

Rapport de recherche No R-576

Novembre 2000

ISBN: 2-89146-444-3

Tables des matières

Tables des matières ____ *iii*

1. Introduction ____ *1*

2. définition actuelle d'un voisinage homogène basé sur l'ACC ____ *3*

3. Vers une révision de la définition actuelle de voisinage homogène ____ *9*

4. Théorie de la classification ____ *11*

4.1 Classification : le cas particulier de densités univariées ____ **13**

5. Définition révisée d'un voisinage homogène ____ *17*

6. Conclusion ____ *19*

7. Références ____ *21*

1. INTRODUCTION

La démarche de régionalisation des événements hydrologiques extrêmes nécessite de définir sur quelle base des bassins sont considérés comme hydrologiquement similaires. Une approche introduite récemment [Ribeiro-Corréa *et al.*, 1995] exploite l'analyse des corrélations canoniques (ACC) afin d'élaborer un cadre théorique permettant de définir un voisinage hydrologiquement homogène pour un bassin-cible.

L'emploi de l'ACC dans le cadre de la régionalisation a pour but de cerner plus nettement les différentes interactions qui existent entre les variables physiographiques et hydrologiques considérées pour un ensemble donné de bassins. Une fois les interactions mieux comprises, elles peuvent servir à une forme d'inférence sur des quantités hydrologiques à partir de quantités physiographiques qui sont généralement plus faciles à obtenir.

La définition de voisinage homogène extraite par du cadre théorique développé à partir de l'ACC comporte un paramètre α, auquel une valeur, *a priori* arbitraire, doit être assignée par le praticien afin d'être utilisable [Ribeiro-Corréa *et al.* 1995]. La détermination de la valeur à utiliser pour le paramètre α est une étape nécessaire pour la mise en œuvre de cette méthode.

Le présent rapport présente les résultats d'une étude approfondie que nous avons menée sur la dérivation par [Ribeiro-Corréa *et al.* 1995] de la définition actuelle de voisinages homogènes à partir de l'ACC. Il ressort de notre étude que la définition actuelle ne prend pas en compte toute l'information pertinente disponible pour identifier des voisinages homogènes sur la base de l'ACC; elle est donc incomplète. Nous verrons comment elle peut être révisée pour obtenir une définition de voisinage homogène qui renferme toute l'information utile disponible et comment cette définition révisée permet de s'affranchir des difficultés d'application de la définition de [Ribeiro-Corréa *et al.* 1995].

2. DÉFINITION ACTUELLE D'UN VOISINAGE HOMOGÈNE BASÉ SUR L'ACC

Nous présentons ici que les éléments principaux de l'approche d'analyse des corrélations canoniques (ACC) nécessaires à la définition de voisinages homogènes telle qu'obtenue par [Ribeiro-Corréa *et al.*, 1995] puisque [Ouarda *et al.*, 1999] en propose déjà un exposé détaillé et complet.

Nous disposons au préalable d'un bassin-cible et de N bassins pour lesquels on connaît les valeurs pour deux ensembles donnés de variables, l'un décrivant des caractéristiques physiographiques et météorologiques et l'autre des caractéristiques hydrologiques. Dans le contexte de régionalisation des valeurs extrêmes de crue, une partie importante de l'information véhiculée par les variables utilisées est contenue dans les interactions qui existent entre ces deux ensembles de variables. Cependant, l'information contenue dans ces interactions, qui s'expriment concrètement par les corrélations entre les diverses variables utilisées, n'est pas aisément disponible. En effet, le portrait des corrélations entre les deux ensembles de variables est brouillé par les corrélations qui existent entre les variables d'un même ensemble.

Dans ce contexte, l'emploi de l'ACC permet de mieux cibler les sources distinctes de corrélations entre les 2 ensembles. Pour y arriver, l'ACC remplace les ensembles originaux par 2 ensembles de variables, dites canoniques, sans perte significative de l'information contenue dans les ensembles originaux. Les variables canoniques, au nombre de p, sont obtenues de telle sorte qu'aucune interaction (corrélation) ne subsiste entre les variables (canoniques) d'un même ensemble. De plus, une variable canonique est corrélée avec une, et une seule, variable canonique de l'autre ensemble. Nous pouvons voir les variables canoniques d'un même ensemble comme étant autant de sources distinctes d'interactions présentes dans cet ensemble.

Appliquée aux données disponibles, l'ACC produit alors N doublets $(\mathbf{w}_i, \mathbf{v}_i)$, i=1,...,N, qui sont des vecteurs de scores des N bassins pour les p-vecteurs $\mathbf{W}$ et $\mathbf{V}$ constitués des p variables canoniques hydrologiques et des p variables canoniques physio-météorologiques, respectivement.

L'hypothèse au centre du développement théorique de l'approche proposée dans [Ribeiro-Corréa *et al.*, 1995] consiste à supposer que les N doublets de scores canoniques sont toutes des réalisations de la densité multinormale

$$\mathrm{N}_{2p}(\mathbf{0}, \mathbf{L}) \tag{1}$$

où

$$\mathbf{L} = \mathrm{Cov}\begin{pmatrix} \mathbf{W} \\ \mathbf{V} \end{pmatrix} = \begin{pmatrix} \mathbf{I}_p & \mathbf{\Lambda} \\ \mathbf{\Lambda}' & \mathbf{I}_p \end{pmatrix} \tag{2}$$

avec

$$\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, ..., \lambda_p) \tag{3}$$

et où λ_i est la corrélation canonique associée à la paire de scores canoniques $(\mathrm{w}_i, \mathrm{v}_i)$.

La matrice $\mathbf{L}$ est donc fonction des corrélations canoniques qui sont les corrélations liant les p paires de scores canoniques $(\mathrm{w}_k, \mathrm{v}_k), \mathrm{k} = 1,2,...,\mathrm{p}$ produites par l'ACC à partir des données. Une justification détaillée de cette hypothèse est présentée dans [Ouarda *et al.*, 1999] ainsi qu'une description des mesures prises en pratique pour viser à la satisfaire.

Pour définir un voisinage hydrologiquement homogène à partir de l'ACC, la stratégie suivie jusqu'à présent consiste à obtenir de l'information sur le vecteur $\mathbf{W}$ de variables canoniques hydrologiques lorsque son vecteur associé $\mathbf{V}$ de variables canoniques

physiographiques s'est réalisé en $\mathbf{v}_0$, le vecteur de scores canoniques du bassin-cible. Convenons de noter ces vecteurs $\mathbf{W}$ par l'expression $\mathbf{W}|\mathbf{V}=\mathbf{v}_0$. En d'autres mots, on considère étudier la variabilité des réponses hydrologiques canoniques de bassins qui sont physiographiquement similaires au bassin-cible. Évidemment, la similitude physiographique notée entre des bassins et un un bassin-cible est limitée par la fidélité de la représentation de la réalité physique de ces bassins que permet le nombre fini de variables physio-météorologiques employées pour la décrire.

L'information cherchée sur $\mathbf{W}|\mathbf{V}=\mathbf{v}_0$ peut s'obtenir sous la forme d'une densité de probabilités. En effet, sous l'hypothèse initiale, donnée par l'équation (1), de multinormalité du couple de vecteurs $(\mathbf{W},\mathbf{V})$, il est possible de montrer [Muirhead, 1982] que la densité (conditionnelle) de $\mathbf{W}|\mathbf{V}=\mathbf{v}_0$ est alors

$$N_p(\boldsymbol{\Lambda}\mathbf{v}_0, \mathbf{I}_p - \boldsymbol{\Lambda}\boldsymbol{\Lambda}') \tag{4}$$

La densité donnée par (4) permet d'identifier dans l'espace canonique, c'est-à-dire l'endroit où les réalisations de $\mathbf{W}$ se font, la région où les scores canoniques hydrologiques correspondant à des bassins physiographiquement similaires au bassin-cible sont susceptibles d'être trouvés. Et c'est avec les réalisations de cette densité qu'on entend former le voisinage homogène pour le bassin-cible.

À ce stade du développement théorique, une difficulté survient: les scores canoniques dont nous disposons ne sont pas des réalisations de la densité conditionnelle donnée par (4). En effet, les scores hydrologiques canoniques que nous avons à partir des bassins n'ont pas été obtenus en supposant que tous les scores canoniques physiographiques correspondants étaient égaux à $\mathbf{v}_0$. Ce sont plutôt, par hypothèse, des réalisations du couple $(\mathbf{W},\mathbf{V})$ sans condition *a priori* sur $\mathbf{V}$, dont la densité conjointe est donnée par (1). Pour décrire les réalisations de $\mathbf{W}$ uniquement, il suffit de considérer la marginale de $\mathbf{W}$ au sein de (1) qui est (voir [Muirhead, 1982, p.7] par exemple) :

$$N_p(\mathbf{0},\mathbf{I}) \tag{5}$$

C'est la densité qui décrit les scores canoniques hydrologiques obtenus des N bassins considérés.

Puisqu'on ne dispose pas de réalisations de (4), on se propose d'identifier parmi les *N* réalisations de (5), celles qui pourraient correspondre à des réalisations de la densité conditionnelle donnée par (4). Ainsi, si en vertu de l'application d'un certain critère de sélection (qui demeure à être énoncé), $n<N$ points peuvent être raisonnablement considérés comme s'ils étaient au départ des réalisations de (4), alors ils seront utilisés comme tels pour constituer le voisinage homogène pour le bassin-cible.

La problématique est donc de déterminer sur quelle base on juge raisonnable, pour un score canonique donné, de considérer la densité (4) comme si elle l'avait originalement produit. [Ribeiro-Corréa *et al.*, 1995] ont suggéré d'extraire la forme quadratique suivante de l'expression de la densité conditionnelle (4):

$$Q(\mathbf{W}) = (\mathbf{W} - \mathbf{\Lambda v}_0)'(\mathbf{I}_p - \mathbf{\Lambda\Lambda}')^{-1}(\mathbf{W} - \mathbf{\Lambda v}_0) \tag{6}$$

Cette quantité est une distance de Mahalanobis; c'est une mesure standardisée de la distance qui sépare le point **W** de la moyenne $\mathbf{\Lambda v}_0$ de la densité (4). Mais comment décide-t-on si un point est trop éloigné du centre de la densité (4) pour qu'il soit raisonnable de le considérer comme étant une réalisation de (4)? Puisque les points éloignés forment la queue de la distribution de Q, que [Ribeiro-Corréa *et al.* 1995] reconnaissent correctement comme étant un khi-deux, cela revient à se demander à partir de quel quantile χ^2_α de la densité du khi-deux, défini par (7), tronque-t-on la queue de la distribution de Q?

$$\text{Prob}\left(Q(\mathbf{W}) \geq \chi^2_\alpha\right) = 1 - \alpha \tag{7}$$

On voit que fixer le niveau α dans (7) revient à déterminer la distance critique au-delà de laquelle un point est considéré trop éloigné de la densité (4) pour raisonnablement lui être assigné. [Ribeiro-Corréa *et al.* 1995] définissent alors un *voisinage hydrologique*

homogène de niveau de confiance 1-α en prenant tous les scores canoniques **W** qui satisfont:

$$(\mathbf{W}-\mathbf{\Lambda v}_0)'(\mathbf{I}_p-\mathbf{\Lambda\Lambda}')^{-1}(\mathbf{W}-\mathbf{\Lambda v}_0)\leq \chi^2_\alpha \qquad (8)$$

C'est la définition actuelle d'un voisinage hydrologiquement homogène pour le bassin-cible considéré.

3. VERS UNE RÉVISION DE LA DÉFINITION ACTUELLE DE VOISINAGE HOMOGÈNE

Le problème dans l'application de la définition (8) provient du choix *a priori* arbitraire d'une valeur du paramètre α. De plus, aucun élément du cadre théorique développé jusqu'à maintenant ne semble pouvoir permettre de déterminer la valeur α qui précise la distance critique, évaluée par le membre de gauche de (8), au-delà de laquelle il n'est pas raisonnable d'assimiler un point à une réalisation de (4). Les méthodes qui ont été proposées pour cibler un niveau de confiance approprié, telle que l'emploi d'une approche de ré-échantillonnage du type jackknife par [Ribeiro-Corréa *et al.*, 1995], ne sont guères satisfaisantes puisqu'elles font intervenir de l'information externe dont on ne dispose pas en pratique. Le problème que pose la détermination d'une valeur pour le paramètre α est dû au fait que cette définition est incomplète.

Une partie importante de l'information pertinente pour opérer la sélection des bassins pour composer le voisinage homogène n'a pas été considérée. En effet, il faut reconnaître que dans le cadre théorique développé il n'y a pas que la densité (4) qui joue un rôle important, mais aussi la densité (5) . Rappelons que la densité (5) est celle dont sont issus tous les scores canoniques obtenus à partir des bassins considérés, et qu'on cherche à identifier parmi ces scores ceux qui pourraient passer plutôt pour être des réalisations de (4). Ces scores formeraient alors un voisinage homogène. Or, assimiler des bassins à des réalisations de (4), via leur score canonique, revient à déterminer les scores qui sont suffisamment près de son centre pour paraître avoir été engendrés plutôt par elle que par (5). Il faut donc établir une mesure convenable de la proximité d'un bassin donné par rapport aux deux densités; on assimilera ensuite le bassin à une réalisation de (4) si pour cette densité la mesure de proximité obtenue pour le bassin est la plus marquée.

L'approche suivie par [Ribeiro-Corréa *et al.* 1995] qui induit (8) incorpore une mesure de proximité par rapport au centre de la densité (4) seulement. Par conséquent, aucun élément de la définition actuelle ne prend en compte la densité (5) comme il se doit. On constate *a posteriori* que le rôle du quantile introduit dans (8) est de permettre à l'utilisateur de compenser lui-même pour l'information manquante en fixant une valeur pour α. La constitution d'un voisinage homogène est donc en réalité une question d'identification des bassins qui passent davantage pour être des réalisations de (4) que de (5). Cette problématique s'avère être à la base du développement de la théorie statistique de la classification. Il est donc naturel d'étudier maintenant la question sous l'angle de la classification.

4. THÉORIE DE LA CLASSIFICATION

Puisque la problématique qui nous occupe est d'assigner des points à l'une de deux densités, il est très instructif de considérer ce qui a été fait en théorie statistique de classification. En effet, la théorie de classification a précisément pour objet l'étude de problématiques de cette nature. C'est donc en étudiant ce qui a été fait en théorie statistique de la classification que nous verrons comment on peut tirer le maximum de l'information disponible pour identifier au mieux les bassins, expliqués par la densité (5), qui peuvent passer pour être des réalisations de (4).

L'approche de classification, tout comme l'analyse des corrélations canoniques, appartient aux statistiques multivariées; c'est un champ d'activités très vaste duquel nous ne couvrirons qu'un cas particulier ici, à savoir celui de deux populations multi-normales.

Dans le problème-type en classification on dispose d'une observation et on ignore à laquelle de deux populations-cibles sa réalisation est attribuable. L'approche de classification fournit des outils et des règles de classification qui permettent d'assigner l'observation à celle des deux populations en présence davantage susceptible de l'avoir produite.

Considérons la situation où nous avons deux populations π_1 et π_2 qui admettent respectivement les densités $f_1 \approx N_q(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ et $f_2 \approx N_q(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$. La règle de classification classique, attribuable à Wald et Fisher [Anderson, 1984, pp. 204-209], s'énonce comme suit:

Assigner l'observation $\mathbf{x}$ *à la population* π_1 *si, et seulement si,*

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq 1 \tag{9}$$

Il est entendu ici que si l'observation n'est pas assignée à la population π_1, à savoir lorsque (9) n'est pas vérifiée, alors elle est assignée à la population π_2. Il est instructif de savoir qu'une formulation plus générale de cette règle permet de prendre en compte des éléments supplémentaires absents du cadre théorique développé ici comme des probabilités *a priori* pour les densités en présence, ainsi que les coûts liés à une mauvaise classification. De plus, la règle de Wald-Fisher est optimale à bien des égards [Anderson, 1984].

En somme, la règle de Wald-Fisher consiste à assigner l'observation **x** à la population qui l'a le plus vraisemblablement produite. On peut facilement remanier l'expression (9) de cette règle pour l'exprimer en fonction des distances standardisées Q données par (6). En effet, (9) est équivalente à

$$\ln\left(f_1(\mathbf{x}) \Big/ f_2(\mathbf{x})\right) \geq 0 \tag{10}$$

où ln est le logarithme népérien.

Considérons la notation

$$Q_i(\mathbf{x}) = (\mathbf{x} - \boldsymbol{\mu}_i)' \boldsymbol{\Sigma}_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i) \quad i=1,2 \tag{11}$$

Après quelques simplifications, l'expression (10) devient:

$$\ln(|\boldsymbol{\Sigma}_2|) - \ln(|\boldsymbol{\Sigma}_1|) - \frac{Q_1(\mathbf{x})}{2} + \frac{Q_2(\mathbf{x})}{2} \geq 0 \tag{12}$$

où $|\boldsymbol{\Sigma}|$ dénote le déterminant de la matrice $\boldsymbol{\Sigma}$.

Pour résumer, la règle de Wald-Fisher peut être donnée à partir de (12) de la maière suivante:

Assigner l'observation **x** *à la population* π_1 *si, et seulement si,*

$$Q_1(\mathbf{x}) + \ln\left(\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|}\right)^2 \leq Q_2(\mathbf{x}) \tag{13}$$

L'expression (13) montre que la règle simple et raisonnable de Wald-Fisher, initialement donnée par la relation (9) en termes de densités de probabilités, ne s'écrit pas comme on aurait pu l'anticiper en termes de distances standardisées seulement :

Assigner l'observation **x** *à la population* π_1 *si, et seulement si,*

$$Q_1(\mathbf{x}) \leq Q_2(\mathbf{x}) \tag{14}$$

La règle (14) attribue l'observation x à la population dont elle est la plus rapprochée en terme de distance standardisée. Bien qu'*a priori* intuitive, cette règle ne constitue pas un choix adéquat pour effectuer la classification d'une nouvelle observation **x**; il est très instructif de s'en convaincre en considérant en détail un cas particulier très simple qui nous permet de mieux voir les processus de sélection en présence.

4.1 Classification : le cas particulier de densités univariées

Afin d'illustrer l'inadéquation de la règle de classification simpliste donnée en (14), considérons le cas où les populations normales en présence sont univariées et de moyennes égales $\mu_1 = \mu_2$ et de variances connues. De plus, supposons aussi que $\Sigma_1 = \sigma_1^2 > \sigma_2^2 = \Sigma_2$, c'est-à-dire que la population π_1 est davantage étalée que la population π_2. En raison de l'égalité des moyennes, la règle (14) se simplifie ici pour donner:

Assigner l'observation x à π_1 *si, et seulement si:*

$$\frac{1}{\sigma_1^2} \leq \frac{1}{\sigma_2^2} \tag{15}$$

Sous les hypothèses de cet exemple, cette condition est toujours satisfaite; par conséquent toute observation à classifier sera assignée à la population π_1. Cette assignation systématique des observations à la population π_1 est une erreur qui est révélée en comparant cette décision à celle qu'on obtient à partir de la règle (13) sous les mêmes conditions:

Assigner x à la population π_1 si, et seulement si:

$$\left(\frac{\mathrm{x} - \mu}{\sigma_1}\right)^2 + \ln\left(\frac{\sigma_1^2}{\sigma_2^2}\right) \leq \left(\frac{\mathrm{x} - \mu}{\sigma_2}\right)^2 \tag{16}$$

Il est aisé de montrer que cette règle équivaut à

Assigner x à la population π_1 si, et seulement si:

$$(\mathrm{x} - \mu)^2 \geq \ln\left(\frac{\sigma_1^2}{\sigma_2^2}\right) \frac{\sigma_1^2 \sigma_2^2}{\sigma_1^2 - \sigma_2^2} \tag{17}$$

En d'autres mots, avec la règle de Wald-Fisher, contrairement à la règle (14), seules les observations suffisamment éloignées de la moyenne (commune) se retrouvent assignées à la population π_1. La règle de Fisher tient compte du fait, comme il se doit, que la population π_2 est davantage concentrée autour de sa moyenne que ne l'est la population π_1. Cette concentration plus marquée de la population π_2 autour de la moyenne (commune) l'amène à générer plus massivement que π_1 des réalisations voisines de la moyenne. Ainsi, même si une observation x peut être plus proche de la population π_1 en raison de son fort étalement lorsque comparée à π_2, il demeure que π_1 produit plus

rarement que π_2 des réalisations dans la région immédiate de la moyenne. De façon équivalente, puisque les queues de la population π_2 sont moins importantes que celles de la population π_1, la règle de Wald-Fisher reconnaît en π_1 la population la plus susceptible de produire les réalisations extrêmes observées.

L'étude de ce cas révèle qu'il y a deux dimensions importantes à la classification d'une observation: la distance standardisée qui sépare l'observation des deux populations et la propension de chacune des deux populations à créer des réalisations dans la région occupée par l'observation. Et des deux règles considérées, la règle simpliste donnée par (14) et la règle de Wald-Fisher donnée par (13), seule cette dernière prend en compte ces deux aspects.

D'un point de vue de classification, il devient clair maintenant que la règle de sélection à la base de la définition d'un voisinage homogène donnée par (8) présente d'importantes lacunes.

Premièrement, la règle de sélection (8) ne fait intervenir dans la balance qu'une des deux populations en jeu, à savoir la densité conditionnelle (4). Ainsi, d'un côté de la balance il y a la distance de l'observation **W** par rapport à la densité conditionnelle (4), mais de l'autre, comme contrepoids, il n'y a rien puisque l'autre densité en jeu, à savoir (5), n'est pas considérée. D'une certaine façon la règle simpliste (14), bien qu'inadéquate comme nous venons de le voir, lui est supérieure puisqu'elle prend au moins en compte les deux populations en présence.

On voit qu'à défaut d'avoir considéré, par exemple, la distance qui sépare l'observation **W** à la population donnée par (5) comme contrepoids à la distance séparant **W** de la densité conditionnelle (4), on est contraint d'introduire un contrepoids artificiel dans la balance: le quantile χ^2_α déterminé par le niveau α.

Deuxièmement, même si la première difficulté était surmontée de quelque façon, il demeure que la règle de sélection (8) ne s'exprime qu'en fonction de la distance standardisée. Or, nous avons vu à la section 4.1 que les distances standardisées ne sont pas à elles seules des éléments suffisants pour constituer une règle de sélection adéquate; il faut prendre en compte la propension des deux densités (4) et (5) à produire des observations dans une région donnée.

5. DÉFINITION RÉVISÉE D'UN VOISINAGE HOMOGÈNE

Nous venons de voir que l'identification adéquate de bassins qui peuvent passer pour être des réalisations de la densité (4) doit se faire en utilisant la règle de Wald-Fisher. Il ne reste plus qu'à rappeler les éléments importants de notre cadre théorique qui sont impliqués dans cette règle.

Convenons de noter par π_1 la population qui représente les scores canoniques **W** issus de la densité conditionnelle (4) et par π_2 les scores canoniques obtenus de la densité (1). Rappelons que π_2 est en fait la marginale de **W** par rapport à (1) et elle est donnée par (5).

Puisque nous sommes en présence de deux populations multi-normales, la règle de Wald-Fisher donnée par (13) devient alors à partir de (6):

Assigner une observation **W** *à la population* π_1 *si, et seulement si*:

$$(\mathbf{W}-\Lambda\mathbf{v}_0)'(\mathbf{I}_p-\Lambda\Lambda')^{-1}(\mathbf{W}-\Lambda\mathbf{v}_0)\leq(\mathbf{W}-\mathbf{0})'\mathbf{I}_p^{-1}(\mathbf{W}-\mathbf{0})+\ln\left(\frac{|\mathbf{I}_p|}{|\mathbf{I}_p-\Lambda\Lambda'|}\right)^2 \qquad (18)$$

En comparant la définition (18) à la définition existante (8), on voit que le quantile a été remplacé par deux termes : le premier tient compte de la distance qui sépare **W** du centre de la densité (5) alors que le second prend en compte la dispersion relative des deux densités.

6. CONCLUSION

En apparence nous nous retrouvons donc avec 2 définitions de voisinages homogènes obtenues à partir de l'ACC qui pourraient être vues comme concurrentes. En réalité, cependant, la définition (18) est la version complétée de la définition existante (8). En effet, nous avons montré que des deux densités en présence (4) et (5), seule (4) intervient dans la définition actuelle d'un voisinage homogène. Il faut reconnaître que dans le cadre théorique développé à partir de l'ACC, un voisinage homogène est un ensemble de points à déterminer qui peuvent passer pour des réalisations de la densité (4), alors qu'ils sont tous des réalisations de (5), au départ, par hypothèse.

Nous avons montré comment la densité (5) devait être utilisée afin d'obtenir une définition de voisinage homogène qui soit complète. Pour obtenir cette définition nous avons eu recours à la théorie de la classification en reconnaissant que le problème en était un d'assignation d'observations à l'une de deux densités en présence. L'approche par classification permet de prendre en compte toute l'information disponible dans le cadre théorique initial qui est pertinente à la formation des voisinages homogènes. Il en résulte une définition révisée de voisinages homogènes qui prend en compte toute l'information disponible et qui permet de s'affranchir de la difficulté associée à la valeur du paramètre α que l'on doit fixer.

7. RÉFÉRENCES

Anderson, T.W., An introduction to multivariate statistical analysis, John Wiley & Sons, 1984.

Muirhead, R.J., Aspects of multivariate statistical theory, John Wiley & Sons, 1982.

Ouarda, T. B.M.J., C. Girard, G. S. Cavadias, et B. Bobée, Regional flood frequency estimation with canonical correlation analysis, soumis au Journal of Hydrology.

Ribeiro-Corréa, J., G. S. Cavadias, B. Clément et J. Rousselle, 1995. Identification of hydrological neighborhoods using canonical correlation analysis, J. Hydrol., 173 : 71-89.