Université du Québec

INRS-Eau

Estimation régionale des crues basée sur l'analyse canonique des corrélations

Par : Claude Girard

Baccalauréat en statistique

Maîtrise en mathématiques fondamentales

Mémoire présenté pour l'obtention du grade

Maître ès sciences (M.Sc.)

Jury d'évaluation

Examinateur externe	Younes Alila University of British Columbia
Examinateur interne	Marius Lachance INRS-Eau
Codirecteur	Bernard Bobée INRS-Eau
Directeur	Taha B.M.J. Ouarda INRS-Eau

Avril 2001

Remerciements

Ce mémoire n'aurait jamais vu le jour n'eut été de la contribution significative de plusieurs personnes que j'aimerais remercier ici.

Mes premiers remerciements vont à mon codirecteur M. Bernard Bobée qui m'a accueilli, il y a quelques années déjà, au sein de son équipe de recherche, ainsi qu'à M. Jean-Pierre Villeneuve et l'équipe de professeurs de l'INRS-Eau qui ont contribué de façon importante à ma formation tout au long de ma scolarité de maîtrise.

À mon ancien professeur de l'Université de Montréal M. Jean-François Angers qui, en bayésien dévoué et enthousiaste qu'il est, a répondu à mes mille et une questions sur l'approche bayésienne. Mille et un mercis Jean-François.

Sincères remerciements à mon directeur Taha Ouarda qui a su superviser mes travaux sans restreindre mes poussées d'originalité qui m'ont écarté à plus d'une reprise des sentiers battus; je reconnais en lui un grand scientifique dont je suis très fier de pouvoir me compter aujourd'hui comme en étant l'élève. Je remercie l'ami que j'ai trouvé en lui, pour son appui indéfectible, qui a de toujours cru en moi, même aux moments où de nous deux il était le seul.

À mon soleil, avec ces chauds rayons et son éternel optimisme en mes capacités, qui a fait d'un arbre vacillant aux quatre vents un arbre aux racines solides, qui se tient fièrement haut et droit, maintenant. Sans toi les fruits n'auraient jamais été ce qu'ils sont aujourd'hui. Merci à toi ma bonne et belle étoile ; sois-tu toujours dans mon ciel à me servir de repère.

Finalement, à mes parents qui ont toujours encouragé leurs fils, bon temps mauvais temps, à poursuivre des études pour leur permettre d'atteindre les sommets auxquels ils aspirent. À ma famille et mes amis qui m'ont toujours supporté; je dois à leurs efforts conjugués une grande part de ce que j'ai accompli ici.

Tables des matières

1. SYNTHÈSE

1.1 Introduction

L'information hydrologique disponible pour l'estimation statistique d'un quantile de crue en un point donné d'un cours d'eau prend plusieurs formes. Sans doute la plus directe est l'information locale qui découle des mesures prises depuis des années au site d'intérêt. Du point de vue statistique, la qualité de l'information locale dépend essentiellement de la longueur de la série d'observations faites au site. Si les mesures à un site ne sont disponibles que pour quelques années, mener une analyse statistique sur ces données en vue d'obtenir une estimation fiable d'un quantile de crue peut alors s'avérer difficile ou impossible. On présente généralement l'information locale sous la forme de maxima annuels de crue.

Dans le cas d'un bassin pour lequel nous disposons de peu ou d'aucune information locale, une estimation des quantiles de crue peut être basée sur l'utilisation de l'information disponible sur des bassins considérés hydrologiquement similaires. Une caractérisation de bassins similaires à un bassin donné, ainsi que de la façon dont l'information disponible pour ces bassins est exploitée pour produire l'estimation des débits d'événements hydrologiques rares, constitue une approche régionale.

Il existe donc plus d'une source pertinente d'information hydrologique à considérer en ce qui a trait à l'estimation d'un quantile de crue, par exemple, pour un bassin donné. La mise en commun de ces diverses sources apparaît ensuite comme une étape naturelle. Il ne s'agit donc pas de considérer l'information locale et l'information régionale de façon séparée, en se limitant par exemple aux estimations ponctuelles d'un quantile de crue auxquelles elles peuvent chacune donner lieu, mais plutôt de chercher à les intégrer en une source d'information maîtresse. Ainsi, l'information locale et l'information régionale disponibles pour un bassin donné seront combinées pour former une seule source d'information qui

résumera toute l'information contenue dans les modèles local et régional considérés pour ce bassin. L'estimation ponctuelle qui pourra ensuite en découler reposera alors sur de l'information plus complète au sujet du bassin donné que chacune des estimations locale et régionale prise séparément. Pour y arriver, cependant, on doit avant tout se donner un cadre statistique qui permette à l'analyste de combiner plusieurs sources d'information hydrologique à la fois.

En statistique, deux écoles de pensée coexistent : l'approche fréquentiste et l'approche bayésienne. Bien que la première soit de loin la plus courante, seule l'approche bayésienne permet de prendre en compte d'une façon rationnelle de l'information externe aux données observées. L'information produite par un modèle régional constitue un exemple important d'information externe aux données locales. C'est pourquoi la mise en commun de l'information hydrologique locale et régionale doit s'effectuer dans un cadre statistique bayésien plutôt que fréquentiste.

Cette étude, qui est un mémoire par articles présenté en vue d'obtenir le grade de maître *ès* sciences, porte sur le modèle régional obtenu de l'application de l'analyse des corrélations canoniques (ACC) aux données hydrologiques et physio-météorologiques disponibles pour un ensemble de bassins d'intérêt et comporte trois volets.

Dans un premier volet l'approche régionale existante basée sur l'analyse des corrélations canoniques (ACC) est décrite, puis révisée en profondeur pour donner lieu à une approche qui soit plus appropriée tant du point de vue théorique que pratique. En effet, l'approche régionale basée sur l'ACC est récente et il s'avère que certains de ses éléments sont encore en cours de développement. Ce mémoire décrit entre autres l'étude approfondie du cadre théorique que nous avons mené, ainsi que des révisions que cela a engendrées. Ce volet fait l'objet de deux articles Cavadias *et al.* (1999) et Ouarda *et al.* (2000), qui sont introduits aux sections 1.2.2 et 1.2.3 et présentés dans leur intégralité aux chapitres 2 et 3, respectivement, et du rapport de recherche Girard *et al.* (2000a) introduit à la section 1.2.4 et présenté au chapitre 4.

Un deuxième volet étudie la possibilité d'obtenir du modèle régional une estimation ponctuelle pour un quantile de crue d'un site d'intérêt donné. Bien que dans un cadre bayésien le modèle régional représente une source d'information parmi d'autres qu'il convient de considérer au sein d'un tout et non isolément, il demeure que plusieurs praticiens s'intéressent aux estimations ponctuelles, dites régionales, qu'il peuvent tirer du modèle. En effet, il est courant, en pratique, de produire une estimation régionale du quantile de crue de période de retour 100 ans, Q_{100}, d'un bassin donné en s'appuyant sur le voisinage homogène préalablement identifié pour ce bassin. En pratique, le modèle régional basé sur l'ACC est couplé à un modèle de régression log-linéaire afin de livrer une estimation régionale d'un bassin d'intérêt depuis un ensemble donné de bassins. Plus préciésment, le modèle log-linéaire considéré exprime le logarithme du quantile Q_{100} d'un bassin en fonction du logarithme de la superficie de ce bassin. Le modèle log-linéaire est d'abord ajusté à partir des bassins d'une région et est utilisé par la suite pour obtenir une prédiction pour un bassin d'intérêt qui n'intervient pas dans l'ajustement du modèle. Cette prédiction n'est pas effectuée pour le quantile Q_{100}, mais plutôt pour le logarithme de cette quantité. Une transformation exponentielle peut être appliquée à cette prédiction pour obtenir cette fois une prédiction de la quantité d'intérêt Q_{100}. Cependant, la prédiction ainsi obtenue pour le quantile Q_{100} ne possède pas les propriétés statistiques souhaitées. La transformation appliquée introduit notamment un biais dans la prédiction pour le quantile qui n'était pas présent dans la prédiction de départ. Cette problématique de biais dans la prédiction obtenue d'un modèle log-linéaire fait l'objet du rapport de recherche Girard *et al.* (2000b) introduit à la section 1.2.5 et présenté intégralement au chapitre 5.

Dans le dernier volet de cette étude, on considère la problématique de combinaison de l'information locale et régionale dans un cadre bayésien. Le défi de baser une approche statistique sur les idées bayésiennes repose sur le fait qu'elles sont encore peu connues et surtout peu documentées. Et paradoxalement, à maints égards, l'approche bayésienne est des deux approches statistiques, l'autre étant appelée fréquentiste ou classique, celle qui est la plus fondée pour ce type de problème. Alors que l'approche statistique fréquentiste fait

l'objet d'une multitude de livres académiques et de travaux de recherche, de tout genre et de tout niveau, l'approche statistique bayésienne, elle, n'est l'affaire que d'un nombre très restreint de statisticiens et d'utilisateurs. Il a donc été difficile de s'initier aux idées bayésiennes, d'autant plus que la littérature montre que les idées bayésiennes, lorsque appliquées, sont plus souvent qu'autrement justifiées de façon inappropriée, voire même erronée dans certains cas. Ces travaux sur l'approche bayésienne font l'objet du chapitre 6.

1.2 Modèle régional basé sur l'ACC

1.2.1 Survol du modèle régional basé sur l'ACC

Le modèle régional considéré utilise l'analyse des corrélations canoniques (ACC) appliquée aux données physiographiques, météorologiques et hydrologiques d'un ensemble de bassins pour lesquels ces données sont disponibles. Ce modèle régional est décrit en détails dans Ouarda *et al.* (2000). Brièvement, l'emploi de l'ACC sur un ensemble donné de bassins produit des variables qui sont fonction des variables originales considérées. Ces nouvelles variables, appelées variables canoniques, renferment l'essentiel de l'information contenue au départ au sein des variables originales tout en admettant des propriétés statistiques plus intéressantes. Les variables canoniques physiographiques remplaceront l'ensemble original de variables physiographiques considérées, alors que les variables hydrologiques donneront lieu aux variables canoniques hydrologiques.

En s'appuyant sur les propriétés statistiques des variables canoniques, il est possible d'identifier pour un bassin-cible donne un ensemble de bassins qui sont perçus comme lui étant hydrologiquement similaires. Cet ensemble de bassins est communément appelé voisinage homogène pour le bassin-cible. Ce voisinage homogène, qui prend forme dans l'espace engendré par les variables canoniques hydrologiques, dépend actuellement d'un paramètre auquel on doit donner une valeur afin que cette approche soit utilisable en pratique. Dans un espace à deux dimensions, c'est-à-dire lorsque l'ACC détermine deux variables canoniques hydrologiques, la frontière du voisinage homogène définit une ellipse.

La valeur du paramètre à fixer, appelée seuil de confiance du voisinage, détermine la taille du voisinage et influe donc sur le nombre de bassins qui formeront le voisinage.

1.2.2 Exploration de l'approche de voisinages homogènes basée sur l'ACC

Cavadias (1990) et Ribeiro-Corréa *et al.* (1995) ont envisagé pour la première fois l'application de l'analyse des corrélations canoniques (ACC) à la régionalisation. Le second article définit en détail un cadre théorique qui décrit comment l'analyse des corrélations canoniques peut être employée pour définir un voisinage homogène pour un bassin-cible donné. Un exemple d'application y est donné, basé sur un ensemble de 55 bassins hydrologiques de la province canadienne du Québec.

Cavadias *et al.* (1999) explorent plus à fond les aspects statistique et hydrologique de cette approche. Ils introduisent notamment des tests statistiques pour les corrélations canoniques, diverses analyses des variables canoniques dont l'analyse de redondance et considèrent des vecteurs d'erreurs pour comparer les estimations locale et régionale obtenues.

Cette étude a permis à l'auteur de s'initier activement à cette nouvelle approche régionale. L'auteur a fourni un support technique afin que les diverses analyses et explorations puissent être menées sur la base de bassins de la province de l'Ontario. L'article, accepté pour publication au *Journal of Hydrological Sciences*, est présenté dans son intégralité au chapitre 2.

1.2.3 Correctifs et améliorations apportés à l'approche régionale basée sur l'ACC

Une contribution importante à l'application de l'analyse des corrélations canoniques dans le contexte de la régionalisation consiste à rendre cette méthode plus accessible à la

communauté hydrologique. En effet, la reconnaissance de l'importance des résultats obtenus, la pertinence des travaux menés jusqu'à présent et ceux à venir dépendent tous directement d'une plus grande diffusion de la méthode dans le milieu hydrologique. Il est donc primordial de présenter un exposé clair et complet de la méthode pour faciliter la tâche d'éventuels utilisateurs. Une première contribution spécifique de l'auteur à l'étude décrite par Ouarda *et al.* (2000) a été de proposer deux algorithmes, un pour chacun des cas jaugés et non jaugés. Ces algorithmes décrivent les étapes successives à suivre pour utiliser la méthode en pratique. Pour la première fois avec cette approche un éventuel utilisateur dispose d'un protocole à suivre pour faire le pont entre la description théorique et l'implantation informatique de la méthode requise pour son application.

L'essentiel du cadre théorique développé jusqu'à présent à partir de l'ACC pour constituer des voisinages homogènes est décrit dans Ribeiro-Corréa *et al.* (1995). Cependant, le cadre théorique qui y est proposé présente quelques failles, certaines mineures et d'autres plus majeures. Les lacunes mineures se rapportent essentiellement aux justifications des hypothèses requises pour le développement de la méthode qui sont incomplètes, voire dans certains cas manquantes. À titre d'exemple de justifications insatisfaisantes dénotées dans Ribeiro-Corréa *et al.* (1995), nous pouvons citer la discussion qui y est faite de l'hypothèse de multinormalité de la distribution du vecteur $\begin{pmatrix} \mathbf{W} \\ \mathbf{V} \end{pmatrix}$. En fait, leur discussion se résume à énoncer l'hypothèse, qui est :

$$\begin{pmatrix} \mathbf{W} \\ \mathbf{V} \end{pmatrix} \approx N_{2p}(\mathbf{O}, L) \tag{1.1}$$

où $L = \begin{pmatrix} I_p & \Lambda \\ \Lambda & I_p \end{pmatrix}$ avec $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda p)$, les λ_i représentant les p corrélations canoniques, et I_p la matrice identité de dimension $p \times p$. En regard de cette hypothèse, il nous apparaît important de justifier le choix des paramètres de la multinormale, en particulier celui de la matrice de covariance, puisque l'hypothèse de multinormalité est centrale à la méthode. Ainsi, de la matrice identité I_p dans le bloc supérieur gauche de la matrice L, nous

déduisons que les variables canoniques W_i et W_j sont supposées non-corrélées si $i \neq j$. Or, nulle part dans Ribeiro-Corréa *et al.* (1995) trouve-t-on une justification ou une explication de cette hypothèse. De plus, et c'est là l'essentiel de notre critique à cet égard, aucune mention n'est faite des mesures que doit prendre l'éventuel praticien de la méthode sur les données qu'il compte employer afin de satisfaire au mieux à la condition de multinormalité. Une deuxième contribution spécifique de l'auteur à Ouarda *et al.* (2000) a été de justifier les diverses hypothèses nécessaires au développement de cette approche, ainsi que de donner de l'ACC une introduction théorique succincte et complète.

Une faille importante que nous ayons notée dans Ribeiro-Corréa *et al.* (1995) est que le développement qui y est proposé pour le cas d'un bassin jaugé n'est pas adéquat. Pour décrire aussi brièvement que possible ce dont il s'agit, il faut savoir que la méthode, lorsqu'elle est appliquée pour un bassin cible donné, produit une ellipse dont la taille dépend d'un paramètre auquel une valeur a été assignée au préalable par l'utilisateur. Cette ellipse est en fait une forme quadratique qui est une quantité aléatoire assujettie à une distribution de probabilités. On peut montrer que cette distribution s'obtient comme corollaire de l'hypothèse de multinormalité placée sur le vecteur (1.1). La faille dans le cadre théorique qui en découle survient lorsqu'on ne réalise pas que la forme quadratique, dont l'ellipse est la représentation graphique, n'admet pas la même distribution selon qu'on situe le bassin cible dans l'espace canonique hydrologique en $\mathbf{w}_0$ (en utilisant de l'information hydrologique partielle disponible) plutôt qu'en $\Lambda\mathbf{v}_0$ (lorsque aucune information hydrologique n'est disponible). En somme, il est possible de montrer que la distribution de la forme quadratique rattachée au cas d'un bassin jaugé n'a pas été correctement identifiée par Ribeiro-Corréa *et al.* (1995). L'identification de la faille et du correctif à apporter au cadre théorique constitue la troisième et dernière contribution spécifique de l'auteur à Ouarda *et al.* (2000).

En plus des points déjà énumérés dans Ouarda *et al.* (2000), qui constituent des contributions de l'auteur, l'article décrit notamment aussi les résultats d'une étude de la

robustesse de la méthode à l'égard de certains facteurs comme, par exemple, le type d'approche paramétrique/non paramétrique utilisée.

1.2.4 Approche par classification aux voisinages homogènes

Jusqu'à présent nos travaux sur le modèle régional basé sur l'analyse canonique des corrélations ont permis de rectifier la portion du cadre théorique développé qui s'applique au cas des bassins jaugés. Ils ont également servi à mieux justifier les hypothèses sous-jacentes à cette approche et à cerner les contraintes que cela amène pour l'utilisateur. Nous avons aussi donné de l'approche une description détaillée de son utilisation qui se révélera sans doute plus aisée à suivre pour le praticien que les comptes-rendus de la méthode jusqu'alors donnés.

Les corrections apportées, les justifications données et les clarifications faites ne changent rien au fait qu'il demeure difficile d'utiliser en pratique l'approche régionale basée sur l'analyse des corrélations canoniques telle que décrite par Ribeiro-Corréa *et al.* (1995). La principale difficulté rencontrée en pratique consiste à déterminer quelle valeur on assigne au paramètre qui détermine la taille de l'ellipse, ce qui influe directement sur le nombre de bassins qui seront inclus dans le voisinage homogène d'un bassin-cible. *A priori*, le cadre théorique développé ne fournit aucune indication sur la détermination de la valeur vers laquelle l'utilisateur doit tendre. À cet égard, le cadre théorique obtenu recèle un vide important.

Des travaux supplémentaires ont permis de révéler que, d'une certaine façon, le cadre théorique développé jusqu'à présent est incomplet; une portion utile de l'information disponible sous le cadre théorique initié n'a pas été utilisée. On réalise par le fait même que le rôle du paramètre qui définit actuellement le voisinage homogène est précisément de permettre à l'utilisateur de combler le vide dans la définition créée par l'information

disponible qui n'a pas été utilisée. Les contributions à ce sujet sont décrites dans le rapport de recherche Girard *et al.* (2000a).

Le rapport de Girard *et al.* (2000a) montre d'abord que l'information pertinente pour déterminer les voisinages homogènes est représentée par deux densités de probabilités au sein du cadre théorique développé, et comment une seule d'entre elles intervient dans la constitution des voisinages ellipsoïdaux. Naïvement, l'une des densités fournit des arguments en faveur de l'assignation d'un bassin donné au voisinage homogène d'un bassin cible, alors que l'autre permet des arguments contre l'assignation au voisinage de ce même bassin. L'assignation d'un bassin donné au voisinage homogène d'un bassin cible revient donc à évaluer « le pour et le contre » et trancher dans le sens qui se doit. Mais puisque l'information utile obtenue de l'une des deux densités n'est pas considérée, seuls les arguments pour l'assignation d'un bassin donné sont présentés. En somme, l'un des deux plateaux de la balance est vide. Et ce vide doit être comblé d'une façon ou d'une autre si une décision sur l'assignation du bassin donné au voisinage homogène doit être rendue. C'est là qu'intervient le paramètre dont la valeur doit être fixée par l'utilisateur : il sert de contrepoids décisionnel. Et l'importance du poids qu'il introduit dans la balance en défaveur de l'assignation du bassin à la densité conditionnelle est fonction directement de la valeur du paramètre à fixer.

Le rapport Girard *et al.* (2000a) montre comment on peut compléter de façon appropriée le cadre théorique développé en tenant compte des deux densités en présence par recours à la théorie de la classification. La définition résultante de voisinage homogène en est une qui utilise toute l'information pertinente disponible et qui, par conséquent, ne s'exprime pas en terme d'un paramètre auquel une valeur doit être fixée de quelque façon pour être utilisable en pratique.

1.2.5 Problématique du biais dans un modèle log-linéaire

Le modèle régional considéré pour représenter l'information régionale disponible est :

$$\ln(Q_{100}) = a \times \ln(\text{AIRE}) + b + \varepsilon \qquad (1.2)$$

où Q_{100} est le quantile de crue de période de retour 100 ans d'un bassin versant et AIRE sa superficie. Comme nous pouvons le voir, le modèle (1.2) décrit un lien linéaire non pas entre les variables originales Q_{100} et AIRE, mais entre les variables transformées $\ln(Q_{100})$ et ln(AIRE) ; c'est un modèle log-linéaire.

Le modèle régional peut être utilisé à deux fins. D'une part, il s'insère dans un cadre bayésien pour représenter l'information régionale disponible qu'on cherche à combiner à celle disponible au site afin que l'estimation résultante du quantile bénéficie de l'information la plus complète possible. Dans cette optique la question du biais n'est pas à proprement dit légitime (voir la section 6.4). D'autre part, hors de tout cadre bayésien de combinaison de l'information, le modèle régional est souvent employé pour obtenir une estimation pour le site d'intérêt, communément appelée estimation régionale. Dans ce contexte (fréquentiste) la question du biais devient très importante. Pour obtenir cette estimation régionale, la démarche la plus couramment utilisée consiste d'abord à obtenir du modèle une estimation de la quantité logarithmique pour le site cible. De là, l'estimation régionale est obtenue en prenant l'exponentielle de cette valeur, l'exponentielle étant la transformation inverse du logarithme naturel.

Miller (1984) met en garde les utilisateurs qui procèdent ainsi en raison du biais qu'ils introduisent alors sur l'estimation de Q_{100} du fait que la transformation n'est pas linéaire. Puisque rien dans Ribeiro-Corréa *et al.* (1995) ne suggère que cet avertissement ait été considéré dans l'utilisation du modèle (1.2), et qu'en général ce conseil n'est pas suivi en pratique en hydrologie, il convient d'aborder le sujet ici.

L'hypothèse sous-jacente au modèle est que le terme d'erreur représenté par ε dans (1.2) admet pour densité de probabilités la densité normale de moyenne nulle et de variance σ^2. Une fois l'estimation des paramètres obtenue par moindres carrés, l'équation du modèle prend la forme :

$$\ln(\hat{Q}_{100}) = \hat{a} \times \ln(\text{AIRE}) + \hat{b} \qquad (1.3)$$

où $\hat{a},\hat{b}$ sont les estimations par moindres carrés obtenues pour a et b, respectivement, et $\ln(\hat{Q}_{100})$ dénote la valeur prédite par le modèle pour une valeur de ln(AIRE) correspondante pour le bassin d'intérêt.

Pour une valeur de ln(AIRE) donnée, la valeur prédite associée $\ln(\hat{Q}_{100})$, obtenue par (1.3), est une estimation sans biais de la moyenne de la distribution de la variable aléatoire $\ln(Q_{100})$. Puisque cette distribution est par hypothèse normale, et par conséquent symétrique, il s'ensuit que $\ln(\hat{Q}_{100})$ est également une estimation sans biais de la médiane de cette distribution.

L'application de la transformation exponentielle à la valeur $\ln(\hat{Q}_{100})$ permet d'obtenir l'estimation suivante $\hat{Q}_{100}$ de Q_{100} pour la valeur de la variable AIRE correspondante :

$$\hat{Q}_{100} = AIRE^{\hat{a}} \times \exp(\hat{b}) \tag{1.4}$$

Cependant, il est possible de montrer Miller (1984) que cette estimation n'est pas sans biais pour la moyenne de la distribution de Q_{100}, bien qu'elle le soit pour sa médiane.

Puisque l'absence de biais est une des propriétés souhaitées des utilisateurs pour un estimateur, Miller (1984) suggère d'introduire dans l'équation (1.4) un facteur de correction qui retire la plus grande partie du biais de l'estimateur. Ainsi, cet auteur propose l'estimateur suivant pour la quantité $\ln(Q_{100})$:

$$\ln(\hat{Q}_{100}) = \hat{a} \times \ln(\text{AIRE}) + \hat{b} + \frac{1}{2}\hat{\sigma}^2 \tag{1.5}$$

où $\hat{\sigma}^2$ est l'estimation de la variance σ^2 obtenue par moindres carrés.

L'application de la transformation exponentielle de part et d'autre de (1.5) fournit l'estimation suivante pour la quantité Q_{100} :

$$\hat{Q}_{100} = AIRE^{\hat{a}} \times \exp(\hat{b}) \times \exp(\frac{1}{2}\hat{\sigma}^2) \qquad (1.6)$$

qui est moins biaisée pour Q_{100} que ne l'est l'estimation obtenue de (1.4). Le terme $\frac{1}{2}\hat{\sigma}^2$ de l'équation (1.5), ainsi que le terme $\exp(\frac{1}{2}\hat{\sigma}^2)$ de l'équation (1.6), sont communément appelés *facteurs de correction pour le biais*.

Pour faire de $\hat{Q}_{100}$ un estimateur sans biais, les modifications à lui apporter sont importantes Neyman et Scott (1960) ; c'est pourquoi la forme très simple (1.6) proposée par Miller (1984), qui réduit le biais, est celle communément employée en pratique.

Un point qui n'est pas discuté par Miller (1984) et qui apparaît néanmoins déterminant dans le choix entre les estimations issues de (1.4) et (1.6) pour obtenir en pratique des valeurs prédites est la comparaison de leurs écarts quadratiques moyens (EQM). Rappelons que l'EQM est la mesure de dispersion la plus adéquate pour juger de la performance d'un estimateur biaisé. Les variances d'estimateurs biaisés pour un même paramètre n'offrent pas un moyen sûr de comparaison entre les estimateurs considérés. En effet, considérons l'exemple suivant de l'estimateur qui donne la valeur 1 comme estimation quel que soit l'échantillon. Cet estimateur est biaisé (à moins que le paramètre à estimer ait pour valeur 1) et en dépit du fait que sa variance est nulle, il constitue clairement un choix absurde d'estimateur. Il en est ainsi car la variance ne mesure pas à la fois la dispersion des estimations et le biais, comme le fait lui l'EQM.

Miller (1984) laisse à entendre que l'introduction dans un modèle log-linéaire d'un facteur de correction doit être systématique maintenant qu'il a été montré qu'un biais existe dans l'estimation. En réalité, la situation n'est pas aussi simple. En effet, les simulations que nous avons effectuées montrent que l'introduction d'un facteur de correction pour le biais peut donner lieu à des prédictions moins bonnes que les prédictions obtenues par le même modèle, cette fois utilisé sans facteur de correction pour le biais. Cela s'explique par le fait que le facteur à introduire pour réduire le biais d'une estimation en accroît

considérablement la variance. En somme, c'est le problème fondamental qu'on rencontre en inférence statistique qui intervient ici : ce qu'on gagne sur le biais, on risque de le perdre en variance, et réciproquement. On ne peut pas généralement bien faire sur les deux plans en même temps. C'est pourquoi il importe d'évaluer l'EQM, qui prend en compte à la fois le biais et la variance, pour l'estimation avec et sans facteur de correction pour le biais pour connaître le « bénéfice net » sur la qualité de l'estimation de la réduction de biais.

Le rapport Girard *et al.* (2000b) propose une étude de l'impact de l'introduction d'un facteur de réduction de biais sur la qualité globale de l'estimation basée sur un modèle log-linéaire, une étude qui ne semble pas déjà avoir été menée en hydrologie. L'étude a permis de révéler qu'il existe des situations d'estimation où l'introduction du facteur de correction mis de l'avant par Miller (1984) mène à un écart quadratique moyen plus grand pour l'estimation corrigée pour le biais que pour l'estimation originale. Par conséquent, dans ces cas, il semble que le facteur de correction de biais cause globalement plus de problèmes qu'il n'en règle. En fait l'étude a permis d'énoncer une condition suffisante sur les paramètres du modèle sous laquelle l'introduction du facteur de réduction de biais de Miller accroît l'EQM de l'estimation résultante.

1.3 Liste des publications

Les travaux que nous avons menés et décrits dans ce mémoire ont donné lieu aux publications suivantes :

Cavadias, G.S., T.B.M.J. Ouarda, B. Bobée, C. Girard, 1999, A canonical correlation approach to the determination of homogeneous regions for regional flood estimation of ungauged basins, accepté pour publication au Journal of Hydrological Sciences.

Ouarda, T.B.M.J., C. Girard, G.S. Cavadias, B. Bobée, 2000, Regional flood frequency estimation with canonical correlation analysis, soumis au Journal of Hydrology.

Girard, C., T.B.M.J. Ouarda, B. Bobée, 2000a, Une approche par classification à la constitution des voisinages homogènes basés sur l'ACC, Rapport de recherche no R-576, INRS-Eau.

Girard, C., T.B.M.J. Ouarda, B. Bobée, 2000b, Étude du biais dans la prédiction depuis un modèle log-linéaire, Rapport de recherche no R-575, INRS-Eau.

2. EXPLORATION DE L'APPROCHE RÉGIONALE BASÉE SUR L'ACC

A canonical correlation approach to the determination of homogeneous regions for regional flood estimation of ungauged basins.

G.S. Cavadias[1]

T.B.M.J. Ouarda[2]

B. Bobée[2]

C. Girard[2]

Abstract

This paper describes a canonical correlation method for determining the homogeneous regions used for estimating flood characteristics of ungauged basins. The method emphasizes graphical and quantitative analysis of relations between the basin and the flood variables before the data of the gauged basins are used for estimating the flood variables of the ungauged basin. The method can be used for both homogeneous regions determined a priori by clustering algorithms in the space of the flood-related canonical variables as well as for «regions of influence» or «neighbourhoods» centered on the point representing the estimated location of the ungauged basin in that space.

---

(1) Visiting professor at INRS Eau.
(2) INRS Eau 2800 Einstein, C.P. 7500 Sainte-Foy, Quebec, GIV 4C7, Canada.

Détermination de régions homogènes pour l'estimation régionale de crues de bassins non jaugés par l'analyse canonique des corrélations

Résumé

Cet article décrit l'application de l'analyse canonique des corrélations à l'estimation régionale des crues annuelles maximales. La méthode projetée met l'accent sur l'étude des relations entre les variables de bassin et de crue des bassins jaugés avant leur utilisation pour l'estimation des crues de bassins non jaugés. Cette méthode peut être utilisée pour la détermination de régions homogènes obtenues par des algorithmes de classification ou des "voisinages hydrologiques" ou "régions d'influence" dont le centre est le point correspondant au bassin non jaugé.

Introduction

The flood characteristics of ungauged basins are estimated by regional methods i.e. by using relationships between physiographical and meteorological variables and characteristics of the maximum annual floods or the partial duration series of a set of gauged basins with hydrological regimes similar to those of the ungauged basin.

The usual steps of the estimation are:

(1) Determination of a set of similar basins («homogeneous region»).

(2) Regional estimation of the flood distribution of the ungauged basin.

Homogeneous regions may be defined in the space of geographical coordinates. This definition, however, has the disadvantage that it is not applicable to small areas and that contiguous basins may not be hydrologically similar (Linsley 1982, Cunnane 1986, Wiltshire 1986). To overcome this difficulty, some researchers have defined homogeneous regions in the space of flood-related variables (Mosley 1981, Gottschalk 1985, Wiltshire 1986). This definition has both advantages and disadvantages. The primary advantage is that it is based on variables directly related to the flood phenomenon; its disadvantages are firstly the difficulty of relating the characteristics of hydrologically defined homogeneous regions to the topographical, physiographical and meteorological conditions of the area and secondly the fact that homogeneous regions in this definition are usually determined by cluster analysis, the purpose of which is to discover «natural clusters» (Dillon and Goldstein 1984) based on the assumption that such clusters exist; however, the existence of such clusters cannot be taken for granted without prior testing (Rogers 1974, Dubes and Zeng 1987). The final set of homogeneous regions depends on the clustering method, the initial partitioning of the space and the metric used. For this reason, some researchers attempted to relate this type of homogeneous region to the geographical coordinates

empirically (Mosley 1981, Gottschalk 1985), or to introduce the concept of fractional membership of a basin to a homogeneous region (Wiltshire 1986).

An entirely different concept of homogeneous region is the «neighbourhood» or «region of influence». Here a homogeneous region is defined in the space of physiographical, meteorological and hydrological variables and centered on the basin under investigation (Acreman and Wiltshire 1989, Burn 1990ab, Zrinji and Burn, 1994, Ouarda et al., 1998). This type of region avoids the difficulties related to the existence of «real» clusters but, in contrast to a priori regions, it has to be determined specifically for each basin under investigation.

According to the region of influence method, the gauged basins enter the region of influence in the order of their weighted euclidean distances from the ungauged basin in the space of the physiographical and meteorological variables where the weights of the variables are selected by the user. At every step, a homogeneity test, based on the flood distributions of the gauged basins of the homogeneous region is used to determine whether the boundary of the homogeneous region has been reached. The «region of influence» approach has the following limitations:

(a) It requires a choice of an arbitrary weight for each basin variable.

(b) It uses weighted euclidean distances that do not take into account the correlations between the basin variables.

(c) The region of influence is determined using both the basin and the flood variables without taking into account the relations between these two sets of variables.

Another approach for determining basin-centered homogeneous regions, introduced by Cavadias (1989, 1990) and Ribeiro-Correa et al (1994), uses the multivariate method of canonical correlation analysis which takes into account the

relationships between the physiographical and meteorological variables and the characteristics of the distribution of the maximum annual floods. As a first approximation, linear relationships between the variables are assumed.

In a recent paper, Bates et al. (1998) classify a set of Australian basins in homogeneous regions on the basis of the L-moments of their flood characteristics. The results of this classification are verified by several multivariate techniques i.e. principal components, cluster analysis, canonical variate analysis, canonical correlation and tree based modelling of the meteorological and physiographical variables. The use of canonical correlation is restricted to a comparison of the canonical variable scores and loadings for the homogeneous regions, determined on the basis of L-moments. The paper does not address the problem of estimating the flood characteristics of an ungauged basin. More specifically, the authors do not examine the adequacy of the meteorological and physiographical basin variables for estimating the flood distribution of the ungauged basin, nor do they propose a method for classifying such a basin in one of the homogeneous regions.

The purpose of the present paper is to describe the use of canonical correlation for determining the basin-centered homogeneous region or neighbourhood of an ungauged basin. The proposed method is applied to the basins of the province of Ontario, one of which is considered ungauged. The emphasis is on the development of statistical methodology. A complete study of the flood hydrology of an ungauged basin requires, in addition to the statistical analysis, a detailed investigation of the climatology, meteorology, and geomorphology of the basin and its geographical environment.

In a recent intercomparison study (GREHYS, 1996a,b) the following methods of determining homogeneous regions for flood estimation were compared for the basins of the Canadian provinces of Ontario and Quebec: Correspondence analysis, hierarchical clustering, canonical correlation and L-moments. The study concludes

that "the specific use of the canonical correlation technique yielded the best results for the delineation of homogeneous regions.

The canonical correlation method

The method of canonical correlation was developed by Hotelling (1935) and introduced into hydrology by Torranin (1972) but, despite its theoretical interest, has not found many applications in data analysis mainly due to the difficulty of interpretation of the canonical variables (e.g. Kendall and Stuart 1968, vol. 3). The application of the method described in this paper emphasises the interpretation of the configurations of sample points in the spaces of uncorrelated basin and flood variables. In addition, a comparison of the scores and loadings of the canonical variables may be useful, particularly in the case of "a priori" defined homogeneous regions, as in the case of Bates et al. (1998). A comparison of a priori defined and basin-centered homogeneous regions is presented in Cavadias (1985).

The basic idea of the method is indicated in fig. 1: Starting from the data matrices X and Q of the basin-related and flood-related variables of a set of gauged basins, we compute the canonical correlation coefficients r_1, r_2, etc. and the two matrices V and W of the corresponding canonical variables. The next step is to represent the basins as points in the spaces of the pairs of uncorrelated canonical variables and examine the similarity of the point-patterns in these spaces, i.e. the capability of basin-related variables to represent flood-related variables. If the point-patterns are sufficiently similar, we proceed to identify homogeneous sub-regions in the space of the flood-related canonical variables. At this point, we have two alternatives: If there are well defined clusters, we delineate a number of fixed homogeneous sub-regions, classify the ungauged basin in one of the homogeneous sub-regions according to its coordinates in the space of the flood-related canonical variables computed from the corresponding coordinates in the space of the basin-

related canonical variables and use the basins of this sub-region to estimate its flood characteristics. In the absence of clusters we use the computed point in the space of the flood-related canonical variables as a center of a hydrological neighbourhood, the basins of which are used for the estimation of the flood characteristics of the ungauged basin.

The first of the above alternatives is discussed in Cavadias (1989) and the second in Cavadias (1990) and in Ribeiro-Correa et al (1994). The present paper describes in more detail the hydrological and statistical aspects of the second alternative i.e. the delineation of the hydrological neighbourhood of an ungauged basin.

A brief outline of canonical correlation in the context of regional flood estimation is given in Appendix A.

The method is applied in three steps:

(1) Analysis of gauged basins with the purpose of determining whether the chosen basin variables provide sufficient information for the estimation of the flood characteristics of the ungauged basin.

(2) Delineation of the homogeneous region (neighbourhood) of the ungauged basin Z.

(3) Estimation of the flood characteristics of the ungauged basin Z.

This paper deals mainly with the first two steps: Although the canonical correlation method can also be used for the third step, the intercomparison of methods of flood estimation carried out by the GREHYS group (GREHYS 1996 b) showed that the canonical correlation method is the most efficient for the first two steps whereas regression methods, which are equivalent to the canonical correlation method, are less efficient than the index flood method for the third step.

In the following paragraphs we describe in more detail the computational steps of the estimation of the flood distribution quantiles of the ungauged basin.

Step 1

1.1 Selection of the geographical, physiographical and meteorological basin variables (x_1, ..., x_p) and the flood-related variables (q_1, ..., q_m) (e.g. quantiles of the distribution of maximum floods) where usually $p \geq m$. The selection is based on hydrological considerations and data availability. The selected variables should be transformed to normality.

1.2 Calculation of the canonical correlation coefficients $r_1(v_1, w_1)$, $r_2(v_2, w_2)$ etc. and the two sets of canonical variables (v_1, ..., v_m) and (w_1, ..., w_m).

1.3 Examination of the corresponding point-patterns in the pairs of scatter diagrams $[(v_1, v_2), (w_1, w_2)]$, $[(v_1, v_3), (w_1, w_3)]$ etc. for determining whether there are clusters of points or outliers and whether the corresponding point-patterns are similar (Fig. 2). The similarity of the patterns is related to the significance of the canonical correlation coefficients (If $r_1 = r_2 = ... = r_m = 1$, the patterns are identical). Bartlett's test of significance of the canonical correlation coefficients is described in the appendix A.

There have been many attempts at developing «objective» indices of similarity of two multidimensional point-patterns (Andrews and Inglehart 1979, Leutner and Borg 1983, Borg and Leutner 1985, Borg and Lingoes 1987), based on the set of distances between all pairs of points in the two configurations. It must be noted, however, that the correlation coefficient between these distances is a misleading index because even if it is equal to one, the patterns may be

different [For example, if in the (v_1, v_2) diagram the distances of the points A, B, C are AB=1 BC=2 and AC=3, the three points A, B, C are on a straight line. If in the (w_1, w_2) diagram the distances of the corresponding points are AB=3, BC=4 and CA=5, the three points A, B, C form a right triangle. However, the correlation coefficient of the pairs (1,3), (2,4), (3,5) is equal to one].

To avoid this problem, Borg and Lingoes (1987) propose to use the correlation coefficient computed on the basis of a regression line through the origin of the distance space. This "congruence coefficient" c is computed from the formula:

$$c = \frac{\sum_{i=1}^{l} d_{iv}\, d_{iw}}{\left(\sum_{i=1}^{l} d_{iv}^2 \cdot \sum_{i=1}^{l} d_{iw}^2 \right)^{\frac{1}{2}}}$$

where:

d_{iv} = i^{th} distance of two points in the space (v_1, ..., v_m)

d_{iw} = i^{th} distance of two points in the space (w_1, ..., w_m)

$$l = \frac{n(n-1)}{2} = \quad \text{number of distances between pairs of points.}$$

The distribution of c is mathematically intractable. Leutner and Borg (1983) give graphs, based on simulations, of the 5% level of significance of this coefficient as a function of the dimension of the spaces of the scatter diagrams and the sample sizes. It must be noted, however, that the congruence coefficient attempts to condense the information of two scatter diagrams in a single number and therefore cannot show, like the scatter diagrams, where the

configurations match and where they do not (Borg and Lingoes 1987). Similarly, a correlation coefficient does not provide all the information contained in a scatter diagram, even in the case of linearly related variables.

The adequacy of the basin variables for estimating the flood variables can be tested more directly in the following way: (Cavadias 1995): In addition to the values of the flood-related canonical variables (w_1, ..., w_m) computed as linear combinations of the flood variables, we estimate the canonical variables $(\hat{w}_1, \cdots, \hat{w}_m)$ using the simple regressions on the corresponding variables (v_1, ..., v_m) and for each basin B_i we plot the «error vectors» $\hat{B}B_i$ (Fig. 3). The vector $\hat{B}B_i$ represents the difference between the local (B_i) and the regional ($\hat{B}$) estimation of the location of the i^{th} basin in the space (w_1, ..., w_m). A study of the lengths and directions of the error vectors for all basins B_i (i = 1, 2, ..., n) enables the user to discover outliers and local patterns and relate them to the causative factors of the annual floods.

Corresponding error vectors can also be computed in the space (q_1, ..., q_m) of the original flood related variables in which they can be interpreted directly (Fig. 3).

The error vectors are also useful in the case of a combination of local and regional estimates (Kuczera 1982, Bernier 1992). The combined estimate is represented by a point on the error vector that lies between the points $\hat{B}$ and B_i and its location depends on the respective variances of the regional and local estimates.

1.4 If the analysis of the previous paragraph provides a satisfactory estimation of the maximum floods of the gauged basins, it is useful to determine the proportions of the variances of the variables (q_1, ..., q_m) that can be explained

by the basin variables (x_1, ..., x_p) and the relative importance of various sub-groups of geographical, physiographical and meteorological variables. This is accomplished by an analysis of the structure correlation matrices i.e. the matrices of the correlation coefficients of the original and canonical variables and the study of the «redundancy indices» proposed by Stewart and Love (1968). The computation of these indices is described in appendix A.

It is also important to examine the stability of the canonical correlation coefficients and the coefficients of the canonical variables. This is achieved by either subdividing the group of gauged basins into two or more sub-groups and repeating the computations for each sub-group or by using the jackknife method (e.g. Mosteller and Tukey 1977) which has the advantage of providing approximate confidence intervals for the estimated coefficients. Another advantage of the jackknife method is that each gauged basin is considered in turn as ungauged and its flood characteristics are computed using the remaining basins.

Step 2

2.1 Computation of the basin-related canonical variables [$v_1(Z)$, ..., $v_m(Z)$] of the ungauged basin Z as linear combinations of the basin variables.

2.2 Estimation of the canonical variables $[\hat{w}_1(Z), \cdots, \hat{w}_m(Z)]$ using the simple regressions with the canonical variables $v_1(Z)$, ..., $v_m(Z)$.

2.3 **(a)** Calculation of the Mahalanobis distances M(i) of each gauged basin (i) from the estimated location of the ungauged basin. The Mahalanobis distance metric is a generalisation of the Euclidean distance adjusted to take into account the correlations between the variables. Other metrics could also be used.

(b) Sorting of the variable M(i) in descending order and determining a sequence of neighbourhoods of diminishing size by eliminating in turn the most distant basin.

These neighbourhoods become progressively more homogeneous and consist of basins more similar to the ungauged basin which is at the centre of the neighbourhoods.

A simple measure of homogeneity of a neighbourhood is the standard deviation of the basin and flood variables after a normalizing transformation. More elaborate homogeneity tests have been proposed in the literature (Wiltshire 1986, Chowdhuri et al 1991).

A set of tests based on the three L-moments (L-Cv, L-Cs and L-Kurtosis) of the distributions of the maximum annual floods of each basin B_i of the neighbourhood was proposed by Hosking and Wallis (1993). These tests were used in the inter-comparison study of flood frequency procedures (GREHYS 1996b) mentioned previously.

(c) The best neighbourhood is chosen on the basis of the minimum size of the prediction intervals for the flood variables of the ungauged basin, computed from the multiple regressions of the flood variables on the basin variables for each neighbourhood.

This statistical procedure for choosing the best neighbourhood must be supplemented by an examination of all relevant physical factors to ensure that the chosen neighbourhood makes hydrological sense and is not just an artifact of the statistical calculations.

Step 3

Several methods of estimation of the flood characteristics of an ungauged basin using the data of the basins of its neighbourhood have been reviewed and compared [GREHYS 1996a, b] and it was concluded that the following three methods called REM 1, REM 2 and REM 3 respectively, give the best results for the Canadian basins considered in the study:

REM 1. Generalized extreme value / PWM index flood procedure (Dupuis and Rasmussen 1993).

REM 2. Regional non-parametric analysis (Gingras and Adamowski 1992).

REM 3. Regional flood estimation by peaks-over threshold (POT) methods (Ouarda and Ashkar 1995).

It is indicated in appendix A that the estimation of the flood variables of the ungauged basin by canonical correlation is equivalent to the estimation by linear multiple regression on the basin variables. The regression method of estimation was compared to other current methods in GREHYS (1996b) for data from Quebec and Ontario, a subset of which was used in this paper. The conclusions of that study are:

(a) The best regression model for these data is the multiplicative model of the form

$$Q_T = a_o \cdot x1^{k1} \cdots x_p^{kp} \cdot e$$

where $x_1,..., x_p$ are basin characteristics and $k_1, ..., k_p$ are parameters estimated by nonlinear optimization.

(b) The above non-linear regression model performed less well than the three best estimation methods mentioned previously. Consequently, upon delineation of the best neighbourhood by canonical correlation, the selection of the most efficient estimation method must take into account the results of the GREHYS study or similar comparisons. Such a comparison of estimation methods is

beyond the scope of this paper and therefore, in the example given in a later section, we estimate the flood variables of the ungauged basin using the linear regressions on the basin variables of the best neighbourhood.

It is to be noted that the authors of the GREHYS paper (GREHYS, 1996b) state that the results of the regional analysis are affected more by the delineation of the homogeneous regions than by the method of flood estimation and conclude that «all computational methods associated with the canonical correlation method for the identification of the neighbourhood appear to give good results.»

A different approach for the estimation of the flood characteristics of ungauged basins on the basis of a given homogeneous region was proposed by Roy (1993). According to this method, a conceptual precipitation-runoff model is calibrated for the nearest gauged basin of the neighbourhood of the ungauged basin in the space of the canonical variables (w_1, ..., w_m) and used to simulate the daily discharges of the ungauged basin on the basis of its known meteorological inputs. The next step is to fit a probability distribution to the simulated maximum daily discharges for each year and estimate the floods of the required return periods. An important limitation of the method is that it deals with maximum daily and not instantaneous peaks. However, it is a useful complementary approach to the usual procedures needing further development.

Application

The canonical correlation method is applied to the estimation of the maximum annual floods of the Province of Ontario in Canada. The locations of the 106 basins are shown in figure 4. We consider basin (46) as ungauged and use the remaining 105 basins to estimate its flood variables.

Step 1

The following basin and flood variables are used:

Basin variables

LUAIRE = $\log_{10}$ [Drainage area (km^2)].

LUPCP = $\log_{10}$ [Slope of main river channel (m/km)].

LULCP = $\log_{10}$ [Length of main river channel (km)].

LUSLM = $\log_{10}$ (SLM + 0.01) where SLM= Area of drainage basin controlled by lakes and swamps (km^2).

LUPTMA = $\log_{10}$ [Mean annual precipitation (mm)].

Flood variables

LUQ2 = $\log_{10}$ [Annual maximum flood of 2-year return period (m^3/sec)].

LUQ1002 = $\log_{10}$ [Ratio of the 100-year maximum flood to the 2-year maximum flood).

The Kolmogorov-Smirnov test was used to examine the normality of the transformed variables. The results showed that the normality assumption cannot be rejected, except for the variable LUSLM which has many equal values corresponding to SLM=0.

We form the (105x5) matrix of the basin variables and the (105x2) - matrix of the flood variables and carry out the canonical correlation computations based on the 105 gauged basins. The resulting canonical correlation coefficients $r_1 = 0.96$ and $r_2 = 0.33$ are both significant at the 5% level according to Bartlett's test.

The scatter diagrams of the two pairs of canonical variables (v_1, v_2) and (w_1, w_2) are shown in figures 5 and 6 respectively. An examination of these diagrams shows that:

(a) The locations of points corresponding to the same basin in the two diagrams are approximately similar.

(b) These are no clearly defined clusters of points and consequently the basin-centered homogeneous region (neighbourhood) approach is more appropriate for this problem.

An examination of the structure correlation matrices (Table 1) shows that, as expected, the correlations of the canonical variables v_1 and w_1 with the basin and flood variables are higher than those of the canonical variables v_2 and w_2. The squares of the structure correlation coefficients are used for the computation of the redundancy indices. The overall index $R_{q/v} = 0.548$ is the sum of the contributions of the two sets of canonical variables which are respectively 0.497 and 0.051.

The diagrams of error vectors using the initial number of 105 basins are not shown because the large number of points makes the interpretation difficult.

The estimated coordinates of the point corresponding to the ungauged basin are $\hat{w}_1(46) = -0.19$ and $\hat{w}_2(46) = -0.27$. We proceed to compute the Mahalanobis distances M(i) of each gauged basin (i) from this point and to arrange them in descending order (Fig. 7). These distances are used to form a sequence of neighbourhoods of diminishing sizes by consecutive omission of the basin with maximum M(i). These neighbourhoods have increasing homogeneity as indicated by the decreasing standard deviations of the flood variables LUQ2 and LUQ1002 (Figure 8). The estimated values of these variables as well as the corresponding 95% prediction intervals are plotted against the size of the neighbourhood in figures 9 and 10 which indicate that the minimum value of these intervals corresponds to a neighbourhood of 20 basins, which is then used for the estimation of the flood variables of the ungauged basin (46). The multiple correlation coefficients for neighbourhoods with 17 or fewer basins are not significant.

The canonical correlations for the 20-basin neighbourhood are $r_1 = 0.87$ and $r_2 = 0.65$ and their significance levels are respectively 0.001 and 0.1. The structure correlation matrices for 20 basins are shown in Table 2. The overall redundancy index is $R_{q/v} = 0.555$ i.e. slightly higher than the redundancy index corresponding to 105 basins. Since $x^2 = 2.53$ for the most distant basin, the 20-basin neighbourhood corresponds to a 63% confidence region. The F-test which is appropriate in the case of estimated covariance matrix results in a confidence region of 65%.

Tables 3 and 4 present respectively a comparison of the means and standard deviations of all basin and flood variables and the results of the multiple regressions of the flood - on the basin variables, for the neighbourhoods of 105 and 20 basins. An examination of these statistics shows that the standard deviations of all basin and flood variables (except for the basin variable (PTMA) are smaller for the 20-basin neighbourhood than for the initial 105 basin neighbourhood. Although the adjusted squared correlation coefficient $\overline{R}^2$ of the basin variable LUQ2 is smaller for 20 basins than for 105 basins, the corresponding prediction interval is smaller because of the smaller standard deviation of LUQ2 for this neighbourhood.

Thus, the 20-basin neighbourhood is more satisfactory for estimating the flood variables of the ungauged basin (46) because it is more homogeneous.

Figures 12 and 13 show the error vectors for the 20-basin neighbourhood in the spaces (w_1, w_2) and (Q_2, Q1002) respectively. The latter diagram whose axes are in the original units of the flood variables gives an intuitive picture of the performance of the canonical correlation method.

These two diagrams must be studied in detail to discover the reasons for the different lengths and directions of the error vectors for the different gauged basins of the neighbourhood.

Figure 11 shows the geographical locations of the basins of the 20-basin neighbourhood. The fact that four basins of north-western Ontario are grouped with basins of south-western Ontario requires further hydrological analysis.

Concluding remarks

The most important features of the canonical correlation method are:

(1) Before proceeding to the estimation of the flood characteristics of the ungauged basin, it includes a detailed study of the relationships between the basin and the flood variables of the gauged basins to ensure that the latter variables can be estimated from the former.

(2) The homogeneous region used for the estimation is determined in the canonical space of the flood variables which is based on the relationships between the basin and the flood variables.

The method of canonical correlation was compared to other current methods of delineation of homogeneous regions such as regions of influence (Zrinji and Burn 1994), correspondence analysis (Birikundavyi et al 1993), and L-moments (Gingras et al 1994), and was found to give better results for the basins considered in the study (GREHYS 1996b)].

Acknowledgments

The authors wish to thank the anonymous reviewers whose comments were helpful and very appreciated.

Appendix A

Given n basins, p standardized basin-related variables x_j and m standardized flood-related variables q_j (e.g. quantiles of a fitted probability distribution), where usually $p \geq m$, we compute the m canonical correlation coefficients $r_1 \geq r_2, \ldots, \geq r_m$ and the m pairs of standardized basin - and flood-related canonical variables v_j and w_j respectively.

The flood-related canonical variables w_j can be estimated from the corresponding basin-related canonical variables v_j using the simple regression equations:

$$\hat{w}_j = r_j v_j \left(j = 1, 2, \cdots, m\right)$$

It must be noted that the flood variables ($q_1, \ldots, q_m$) can be estimated using the multiple regressions $\hat{q}_j = f_j\left(x_1, \cdots, x_p\right)$ or equivalently the multiple regressions $\hat{q}_j = f_j\left(v_1, \cdots, v_m\right)$. Thus the use of canonical variables results in a reduction of the dimensionality of the space of basin variables from p to m in a way that takes into account their relations with the flood variables. As a result, the number of flood variables that can be estimated is equal to the number of significant canonical correlation coefficients.

Statistical packages usually plot diagrams of (v_1, w_1), (v_2, w_2) etc., i.e. the pairs of canonical variables having maximum correlation coefficients. Given the difficulties in interpreting the canonical variables (e.g. Kendall and Stuart, 1968), it is preferable to plot the uncorrelated pairs of canonical variables (v_1, v_2), (v_1, v_3) ... (v_j, v_k) etc., where $j \neq k$ along with the corresponding scatter diagrams (w_1, w_2) ... (w_j, w_k) of uncorrelated flood-related canonical variables. The pairs of canonical variables (v_1, v_2) (w_1, w_2) etc. respectively define the spaces of linearly transformed basin- and flood-related variables in which the points represent individual basins. If

the basin variables are good predictors of the flood-related variables, the patterns of points in the corresponding scatter diagrams are similar.

It is useful to compute the structure correlation coefficients i.e. the correlation coefficients between the original and the canonical variables which help to determine the contribution of each of the basin variables to the flood variables.

For large samples, the statistical significance of the set of m canonical correlation coefficients can be tested by Bartlett's statistic:

$$v = -(n-1-(p+1+m)/2)\sum_{j=1}^{m}\ln(1-r_j^2)$$

which, under the normality assumption, is distributed as chi-square with (pxm) degrees of freedom. If successive pairs of canonical variables are to be tested, the degrees of freedom must be modified. For example, the degrees of freedom associated with the second pair of canonical variables are (p-1) • (m-1) and so on.

The square of the canonical correlation, coefficient r_j^2 is a measure of the shared variance between the canonical variance v_j and w_j. Our main interest, however, is the proportion of the variance of the flood variables accounted for by the basin variables. An appropriate measure of this proportion is the redundancy index $Rd_{q/v}$ which is equal to the mean of the proportions of the variances of the flood variables ($q_1,...q_m$) accounted for by the canonical variables ($v_1,..., v_m$) or equivalently by the basin variables ($x_1,..., x_p$). The redundancy index is given by the following equations (Cooley and Lohnes 1971 p. 173):

$$Rd_{q/v} = \sum_{k=1}^{m} Rd_{q/v_k} = \sum_{k=1}^{m} Rd_{q_j/v}$$

where:

Rd_{q/v_k} = contribution of the kth pair of canonical variables to the redundancy index $Rd_{q/v}$

$$= \sum_{j=1}^{m} \frac{r^2(q_i, v_k)}{m} = r_k^2 \sum_{j=1}^{m} \frac{r^2(q_i, w_k)}{m}$$

$Rd_{q_j/v}$ = contribution of the jth basin variable to the redundancy index $Rd_{q/v}$

$$= \sum_{k=1}^{m} \frac{r^2(q_j, v_k)}{m}$$

Given the estimated point $\underline{\hat{w}(Z)} = \hat{w}_1(Z) \ldots \hat{w}_m(Z)$ in the m-dimensional space of the canonical variables and under the normality assumption, the $(1-\alpha)$ per cent confidence region for the point is given by the equation:

$M(i) = M[\underline{w(i)} - \underline{\hat{w}(Z)}] = [\underline{w(i)} - \underline{\hat{w}(Z)}]' \, (I_m - \Lambda)^{-1} \, [\underline{w(i)} - \underline{\hat{w}(Z)}] \leq x^2(\alpha, m)$

where Λ is the (m x m) diagonal matrix of the squared canonical correlation coefficients ($r_1^2, \ldots, r_m^2$) and $M\left[\underline{w(i)} - \underline{\hat{w}(Z)}\right]$ is the Mahalanobis distance of the point $\underline{w(i)}$ from the estimated point $\underline{\hat{w}(Z)}$.

This confidence region can be interpreted as the $(1-\alpha)$ per cent neighbourhood of the point $\underline{\hat{w}(Z)}$ (Ribeiro-Correa et al 1994). In the special case of m=2 the above equation is simplified:

$$M\left[\underline{w(i)}, \underline{\hat{w}(Z)}\right] = \frac{[w_1(i) - w_1(Z)}{1 - r_1^2} + \frac{[w_2(i) - w_2(Z)}{1 - r_2^2} \leq x^2(\alpha, 2)$$

If the normality assumption is not valid, the Mahalanobis distance is a weighted distance of the basin $\underline{w(i)}$ from the estimated location $\hat{\underline{w}}(Z)$ of the ungauged basin, with weights depending on the canonical correlation coefficients.

Appendix B - Notation

α	Significance level.
k_1, k_p	Regression parameters.
d_{iv}	i^{th} distance of two points in the space $(v_1,..., v_m)$.
d_{iw}	i^{th} distance of two points in the space $(w_1,..., w_m)$.
e	Regression error.
LUAIRE	$\log_{10}$ [Drainage area (km^2)].
LULCP	$\log_{10}$ [Length of main river channel (km)].
LUPCP	$\log_{10}$ [Slope of main river channel (m/km)].
LUSLM	$\log_{10}$ (SLM + 0.01) where SLM = area of drainage basin controlled by lakes and swamps (km^2)
LUPTMA	$\log_{10}$ [Mean annual precipitation (mm)].
LUQ2	$\log_{10}$ [Annual maximum flood of 2-year return period (m^3/sec)].
LUQ1002	$\log_{10}$ [Ratio of the 100-year maximum flood to the 2-year maximum flood).
M(i)	Mahalanobis distance of gauged basin (i) from the estimated point [$\underline{w(i)}$, $\hat{\underline{w}}(Z)$] of the ungauged basin Z.
m	Number of flood variables.
n	Number of basins.
p	Number of basin variables.
$q_1,..., q_m$	Flood variables.

$\mathbf{r_1,\ldots, r_m}$	Canonical correlation coefficients.
$\overline{R}$	Adjusted multiple correlation coefficient.
$\mathbf{Rd_{q/v}}$	Redundancy index.
$\mathbf{v_1,\ldots, v_m}$	Canonical variables of the basin variables.
$\mathbf{V}$	Bartlett's statistic.
$\mathbf{w_1,\ldots, w_m}$	Canonical variables of the flood variables.
$\mathbf{x_1,\ldots, x_p}$	Basin variables.

References

Acreman, M.C. and Wiltshire, S.E. (1989). The regions are dead: long live the regions. Methods of identifying and dispensing with regions for flood frequency analysis. IAHS Publ. no. 187, 175-1988.

Andrews, F.M. and Inglehart, R.F. (1979). The structure of subjective well being in nine Western Societies. Social Indicators Research 6, 73-90.

Bates, B.C., Rahman, A., Mein, R.G. and Weinmann P.E. (1998). Climatic and physical factors that influence the homogeneity of regional floods in southeastern Australia, Water Resources Research 34 (12) 3369-3381.

Bernier, J. (1992). Modèle regional à deux niveaux d'aléas. Interim Report NSERC Strategic Grant No STR 0118482, 11 p.p.

Birikundavyi, S., Rousselle, J. and Nguyen, V.T.B. (1993). Determination des régions homogènes pour le Québec et l'Ontario: Une approche par l'analyse des corréspondances et la classification ascendante hierarchique. Rapport final, Subvention stratégique CRSNG, No STR 0118482, École Polytechnique de Montréal, p. 44.

Borg, I. and Lingoes, D. (1987). Multidimensional similarity structure analysis. Springer-Verlag.

Burn, D.H. (1990a). An appraisal of the 'region of influence' approach to flood frequency analysis. Hydrological Sciences, Journal, 35 (2) 149-165.

Burn, D.H. (1990b). Evaluation of regional flood frequency analysis with a region of influence approach. Water Resources Research 26 (10) 2257-2265.

Cavadias, G.S. (1989). Regional flood estimation by canonical correlation. Paper presented to the 1989 Annual Conference of the Canadian Society of Civil Engineering, St. John's Newfoundland.

Cavadias, G.S. (1990). The canonical correlation approach to regional flood estimation. Regionalization in Hydrology. Proc. of the Ljubljana Symposium, IAHS. Publ. No. 191:171-178.

Cavadias, G.S. (1995). Regionalization and multivariate analysis: The canonical correlation approach. In "Statistical and Bayesian methods in hydrological sciences". IHP-V, Technical Documents in Hydrology No. 20, UNESCO, Paris.

Chowdhury, J.V., Stedinger J.R. and Lu, L.H. (1991). Goodness of fit test for regional generalized extreme value distributions. Water Resour. Res. 27(7), 1765-1776.

Cooley, W.W. and Lohnes, P.R. Multivariate data analysis. Wiley 1971.

Cunnane, C. (1986). Review of statistical models for flood frequency estimation. Keynote paper in: International Symposium on Flood Frequency and Risk Analysis (Baton Rouge, May 1986). Reidel.

Dalrymple, T. (1960). Flood frequency analysis. USGS Water Supply Paper 1534 A., p. 60.

Dillon, W.E. and Goldstein, M. (1984). Multivariate Analysis, p. 139. John Wiley.

Dubes, R. and Zeng, G. (1987). A test for spatial homogeneity in cluster analysis. Classification 4, 33-56.

Dupuis, L. and Rasmussen, R.F. (1993). Évaluation de méthodes «indice de crue» pour l'estimation d'une distribution régionale. Rapport interne INRS-Eau, No 1-125, pp. 26.

Gingras D., Adamowski, K. and Pilon, P.J. (1994). Regional flood equations for the Provinces of Ontario and Quebec. Water Resour. Bull. 30(1), 55-67.

Gingras, D. and Adamowski, K. (1992). Coupling of nonparametric frequency and L-moment analysis for mixed distribution identification. Water Resour. Bulletin: 28(2), 263-272.

Gottschalk, L. (1985). Hydrological regionalization in Sweden. Hydrol. Sci. J. (30) (1).

Green, P.E. (1978). Analyzing multivariate data. The Dryden Press, Hinsdale, Illinois.

GREHYS (1996a). Presentation and review of some methods of regional flood frequency analysis. J. Hydrol. 186 (1-4), 63-84.

GREHYS (1996b). Intercomparison of flood frequency procedures for Canadian rivers. J. Hydrol. 186 (1-4), 85-103.

Hosking, J.R.M. and Wallis, J.R. (1993). Some statistics useful in regional frequency analysis. Water Resour. Res. 29(2), 271-281.

Hotelling, H. (1936). Relations between two sets of variates. Biometrica 28:321-377.

Kendall, M.G. and Stuart, A. (1968). The advanced Theory of Statistics, Vol 3. 2nd ed. Charles Griffin & Co. London.

Kuczera, G. (1982). Combining site-specific and regional information: an empirical Bayes' approach. Water Resour. Res. Vol. 8, No. 2, pp. 306-314.

Leutner, D. and Borg, I. (1983). Zufallskritische Beurteilung der Übereinstimmung von Factor und MDS-Konfigurationen. Diagnostica 29, 320-335.

Leutner, D. and Borg, I. (1985). Zur Messung der Übereinstimmung von multidimensionalen Konfigurationen mit Indizes. Zeischrift für Sozialpsychologie, 16, 29-35.

Linsley, R.K. (1982). Flood estimates. How good are they? Wat. Resour. Res. 22 (9).

Mosley, M.P. (1981). Delimitation of New Zealand hydrological regions. J. Hydrol. 49, 173-192.

Mosteller, F. and Tukey, J.W. (1977). Data Analysis and Regression. Addison-Wesley.

Ouarda, T.B.M.J., and Ashkar, F. (1995). The peaks-over-threshold (POT) method for regional flood frequency estimation. Proceedings of the 48th Annual Conference of the Canadian Water Resources Association, Fredericton, N.B., Canada, 641-659.

Ouarda, T.B.M.J., Haché, M. and Bobée, B. (1998). Regional estimation of extreme hydrological events, INRS-Eau, Research Report No. R-534 (In French), Quebec, Canada, p. 181.

Panu, V.S., Smith, D.A. and Ambler, D.C. (1984). Regional flood frequency Analysis for the island of Newfoundland. Environment Canada and Department of Environment, Province of Newfoundland.

Ribeiro-Correa, B., Cavadias, G.S., Clement, B. and Rouselle, J. (1994). Identification of hydrological neighborhoods using canonical correlation analysis. Journal of Hydrology 173 (1995) 71-89.

Rogers, A. (1974). Statistical Analysis of Spatial Dispersion. Pion Ltd.

Roy, R. (1993). Regionalisation de Caractéristiques de Crue, Utilisation d'une Méthode Combinant les Approches Déterministes et Stochastiques, Ph.D. Thesis, INRS-Eau.

Stewart, D.K. and Love, W.A. (1968). A general canonical correlation index. Psychological Bulletin 70 (1968) 160-163.

Torranin, P. (1972). Applicability of canonical correlation in hydrology. H.P. 58, Colorado State University, p. 30.

United States Water Resources Council (1977). Guidelines for Determining Flow Frequency. USWRC, 2120 Long Island NW, Washington, DC.

Wiltshire, S.E. (1986). Regional flood analysis II: multivariate classification of drainage basins in Britain. Hydrol. Sci. J. 31 (3).

Zrinji, Z. and Burn, D.H. (1994). Flood frequency analysis for ungauged sites using a region of influence approach. Journal of Hydrology 153, 1-21.

List of tables

List of figures

	V1	V2	W1	W2
LUAIRE	.9468	.0702	.9069	.0235
	(105)	(105)	(105)	(105)
	.0000	.4767	.0000	.8120
LUPCP	-.6049	.1822	-.5794	.0609
	(105)	(105)	(105)	(105)
	.0000	.0629	.0000	.5368
LULCP	.9188	-.1477	.8801	-.0494
	(105)	(105)	(105)	(105)
	.0000	.1326	.0000	.6166
LUSLM	.4482	-.2652	.4293	-.0887
	(105)	(105)	(105)	(105)
	.0000	.0062	.0000	.3681
LUPTMA	-.3417	-.5216	-.3273	-.1745
	(105)	(105)	(105)	(105)
	.0004	.0000	.0007	.0751
LUQ2	.9549	-.0264	.9969	-.0788
	(105)	(105)	(105)	(105)
	.0000	.7895	.0000	.4242
LUQ1002	-.2863	.3192	-.2989	.9543
	(105)	(105)	(105)	(105)
	.0031	.0009	.0019	.0000

Table 1: Structure correlation matrix (105 basins).

	V1	V2	W1	W2
LUAIRE	.2355	.0151	.2049	.0096
	(20)	(20)	(20)	(20)
	.3175	.9496	.3861	.9678
LUPCP	-.0486	.0734	-.0423	.0468
	(20)	(20)	(20)	(20)
	.8388	.7586	.8596	.8446
LULCP	.1268	.4252	.1103	.2715
	(20)	(20)	(20)	(20)
	.5944	.0616	.6435	.2470
LUSLM	-.5520	.4394	-.4803	.2806
	(20)	(20)	(20)	(20)
	.0116	.0525	.0321	.2308
LUPTMA	.2021	.4849	.1759	.3096
	(20)	(20)	(20)	(20)
	.3927	.0302	.4582	.1840
LUQ2	.7894	.2586	.9072	.4206
	(20)	(20)	(20)	(20)
	.0000	.2522	.0000	.0648
LUQ1002	.1289	-.6314	.1482	-.9890
	(20)	(20)	(20)	(20)
	.5879	.0028	.5329	.0000

Table 2: Structure correlation matrix (20 basins).

Table 3: Comparative statistics of the basin and flood variables for 105 and 20 basins.

VARIABLES	OBSERVED VALUE	MEAN		STANDARD DEVIATION	
	BASIN (46)	105 BASINS	20 BASINS	105 BASINS	20 BASINS
LUAIRE	2.467	2.759	2.573	0.769	0.367
LUPCP	-0.187	0.004	-0.039	0.511	0.449
LULCP	1.903	1.758	1.633	0.430	0.171
LUSLM	-2.000	0.759	0.559	2.353	2.208
LUPTMA	2.923	2.919	2.936	0.072	0.079
LUQ2	1.777	1.758	1.617	0.556	0.149
LUQ1002	0.362	0.422	0.401	0.126	0.083

Table 4: Multiple regression results for 105 and 20 basins.

MULTIPLE REGRESSION RESULTS						
DEPENDENT VARIABLES	ADJUSTED R^2		ESTIMATED VALUE		95% PREDICTION INTERVAL	
	105 BASINS	20 BASINS	105 BASINS	20 BASINS	105 BASINS	20 BASINS
LUQ2	0.908	0.586	1.664	1.651	0.672	0.418
LUQ1002	0.143	0.206	0.397	0.349	0.464	0.324

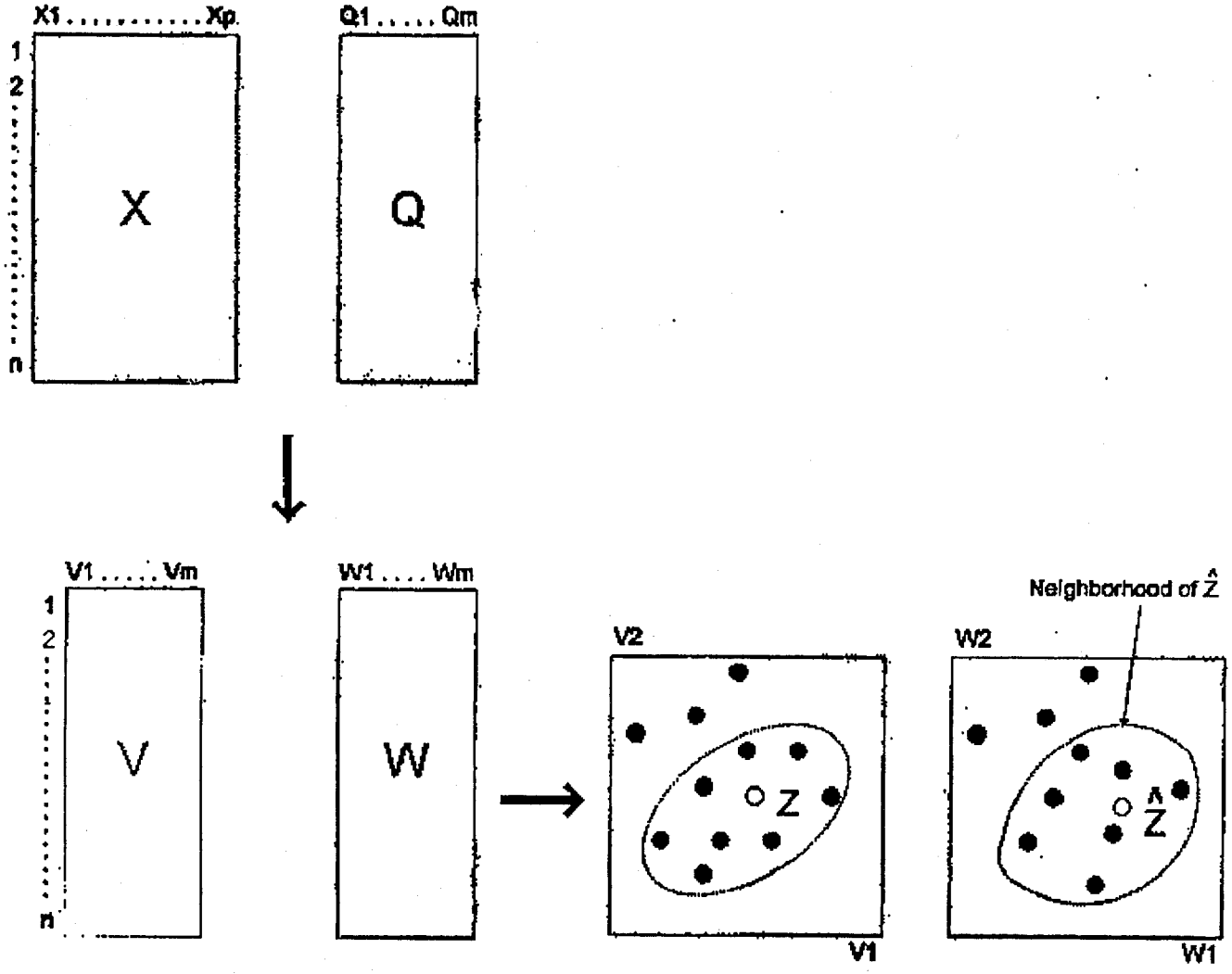

Figure 1. The basic concept of canonical correlation.

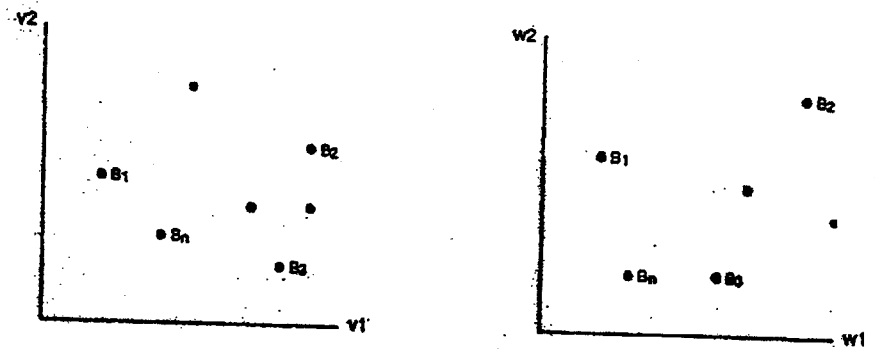

Figure 2. Scatter diagrams (v_1, v_2) and (w_1, w_2).

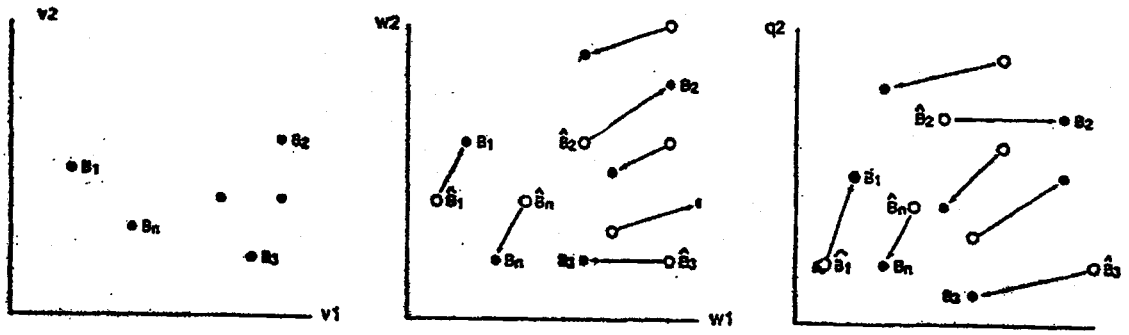

Figure 3. Error vector diagrams in the spaces (m_1, w_2) and (q_1, q_2).

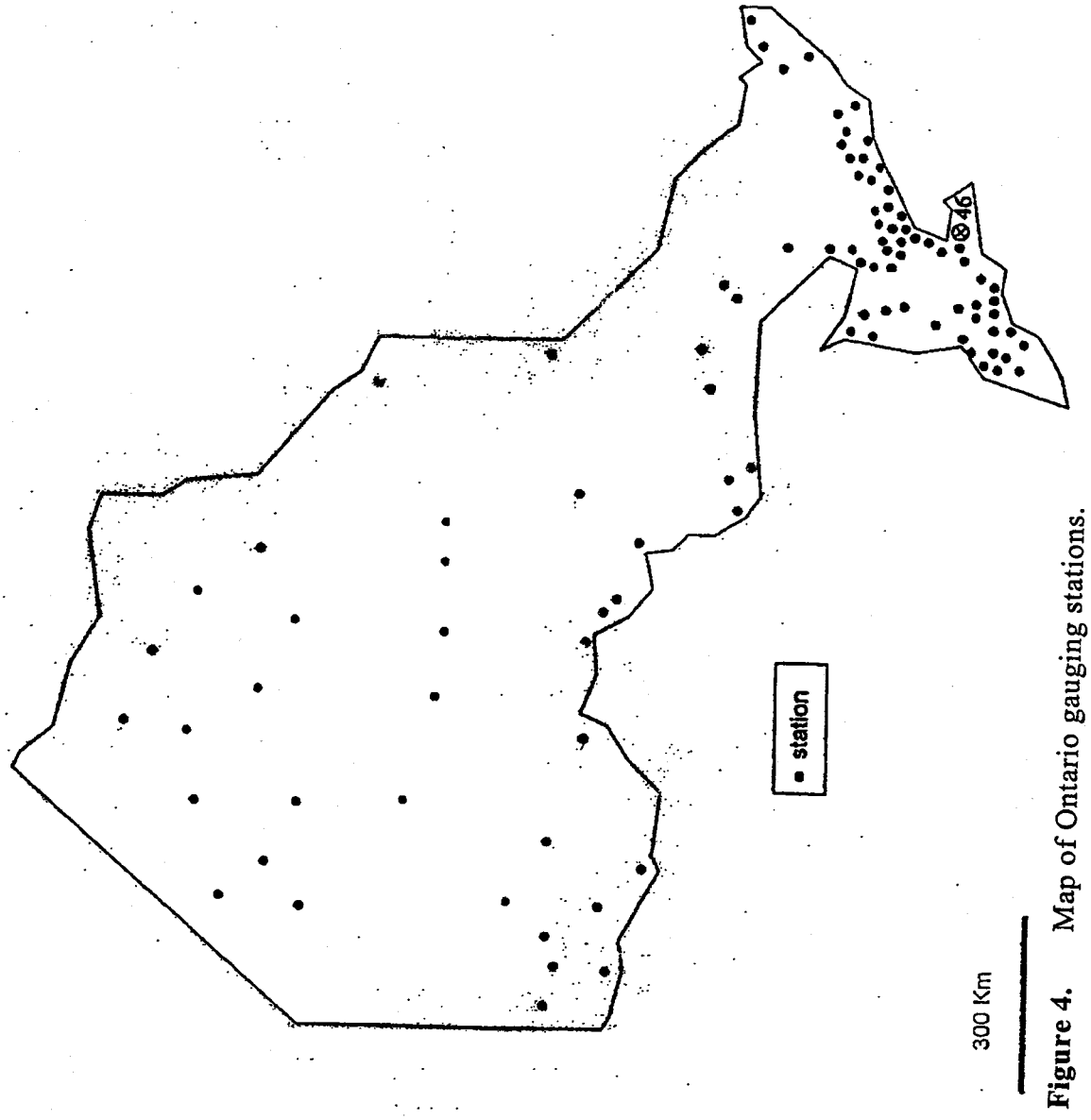

Figure 4. Map of Ontario gauging stations.

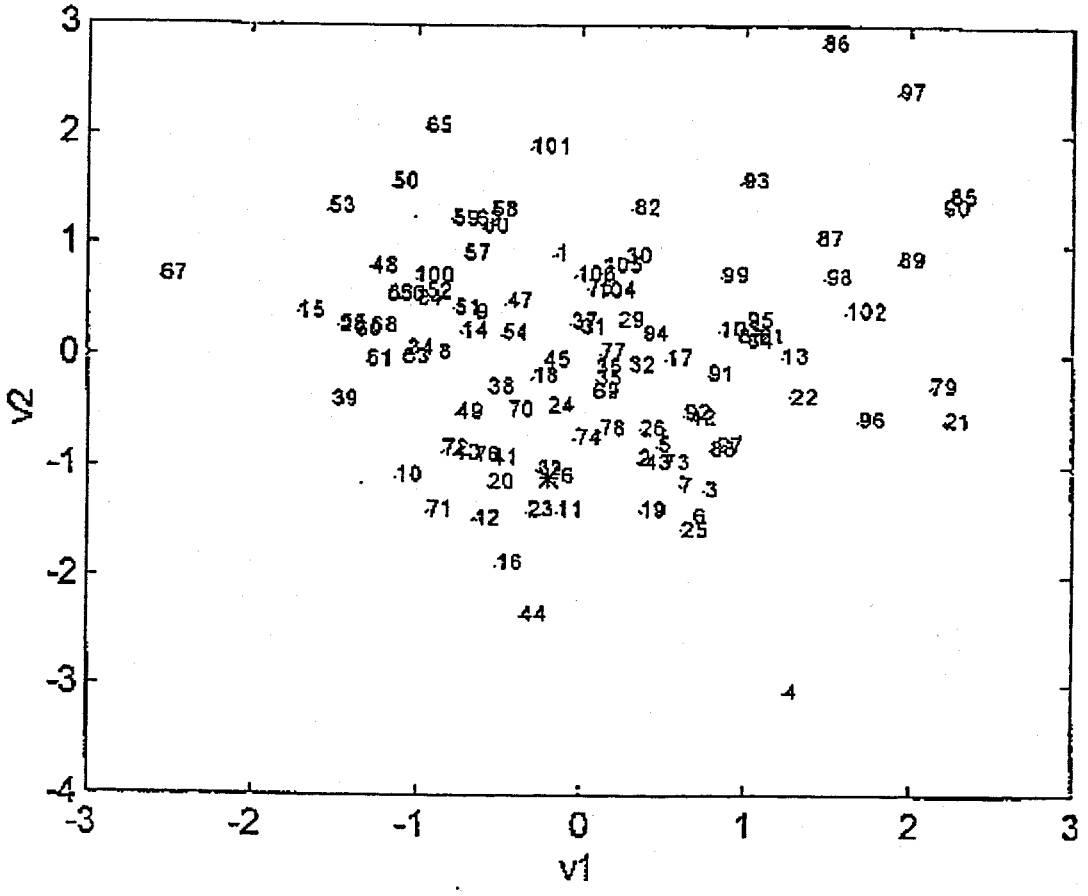

Figure 5. Scatter diagram (v_1, v_2) for 105 basins.

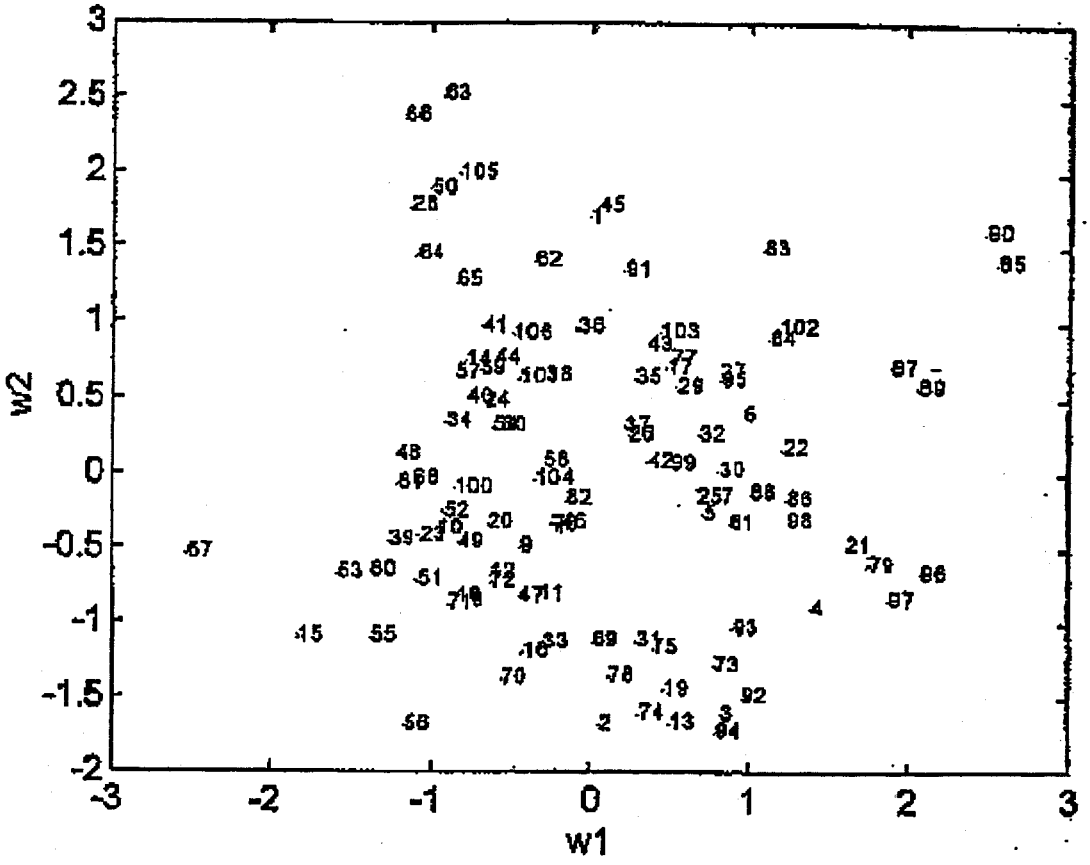

Figure 6. Scatter diagram (w_1, w_2) for 105 basins.

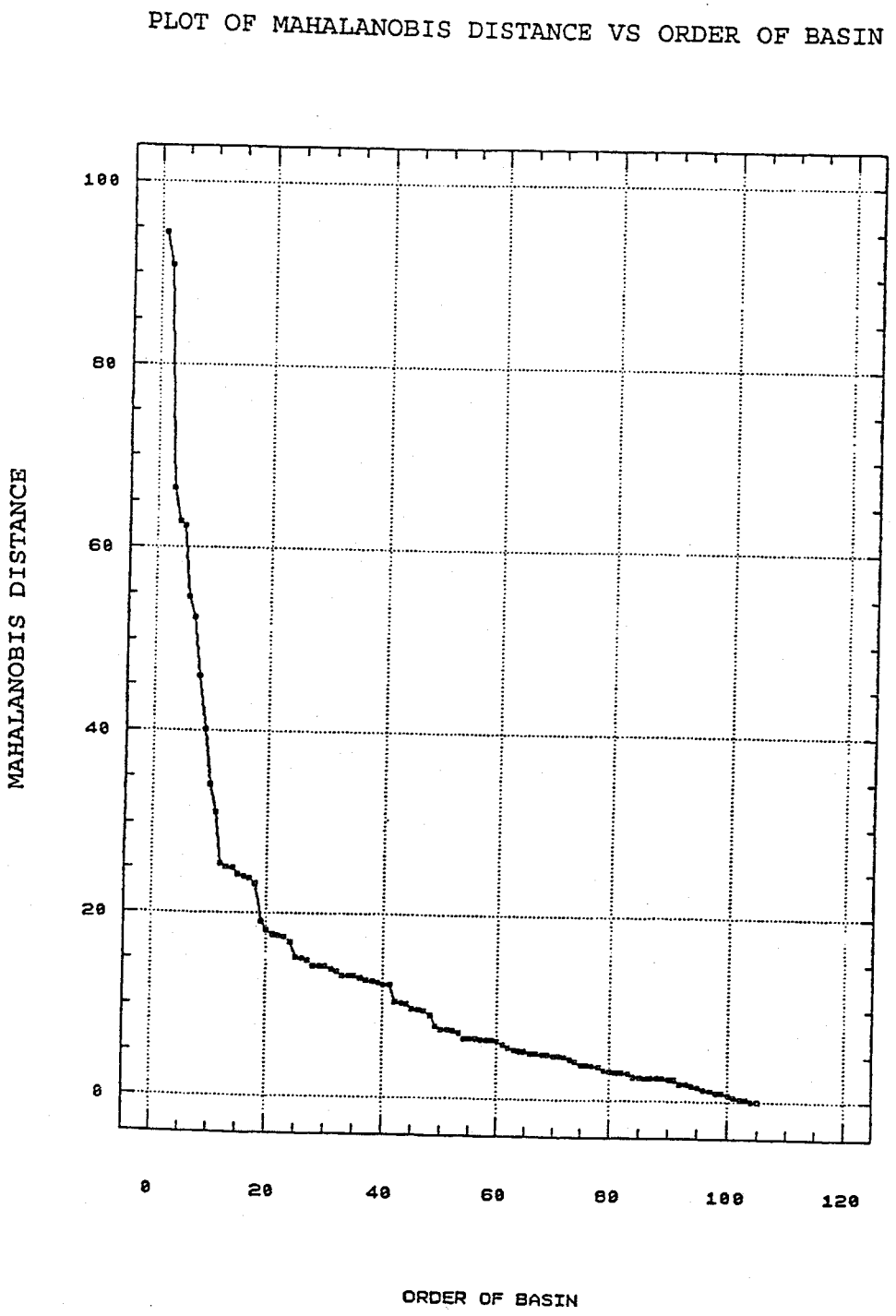

Figure 7. Diagram of ordered Mahalanobis distances.

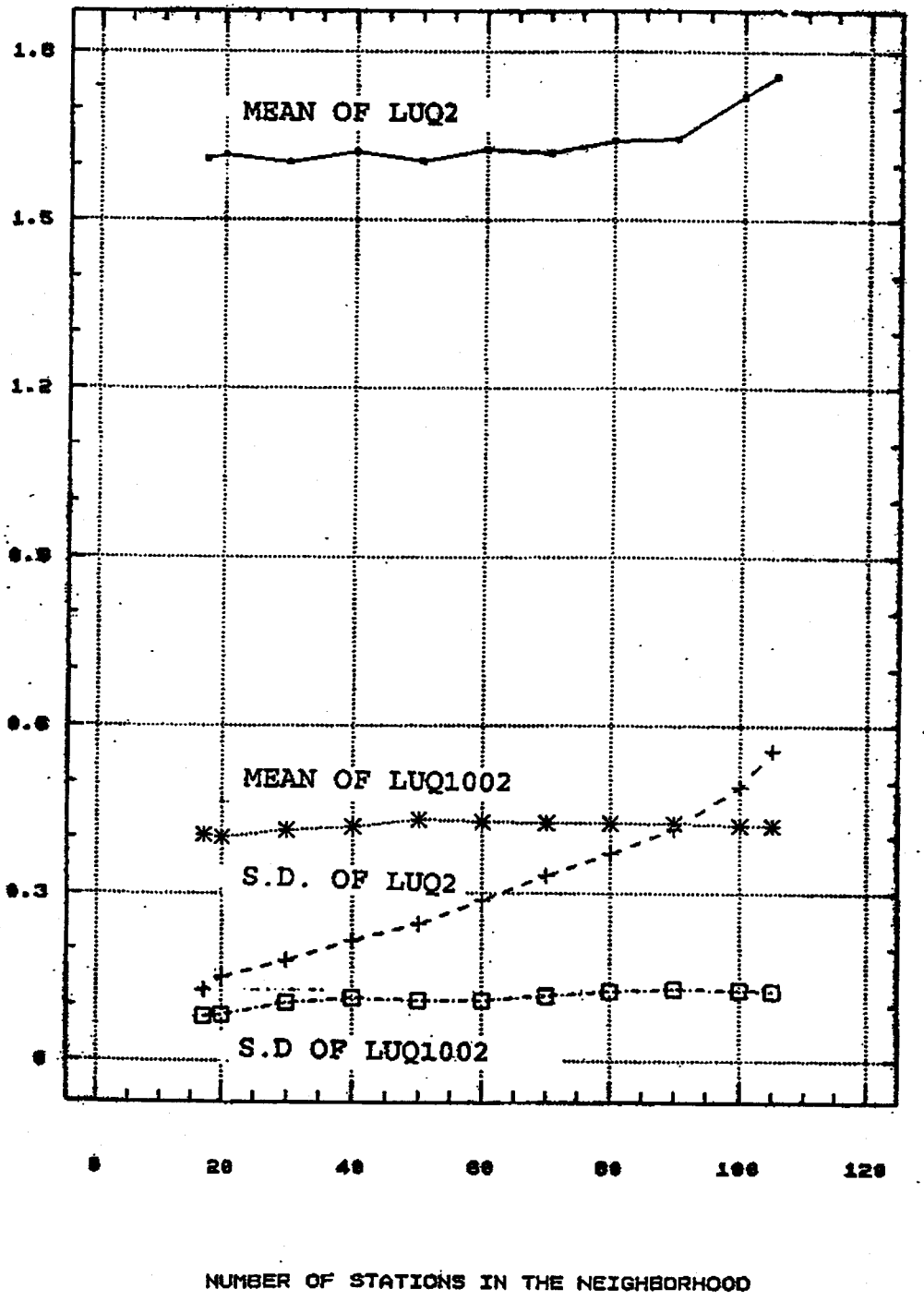

Figure 8. Means and standard deviations of the variables LUQ2 and LUQ1002 vs the number of stations in the neighborhood.

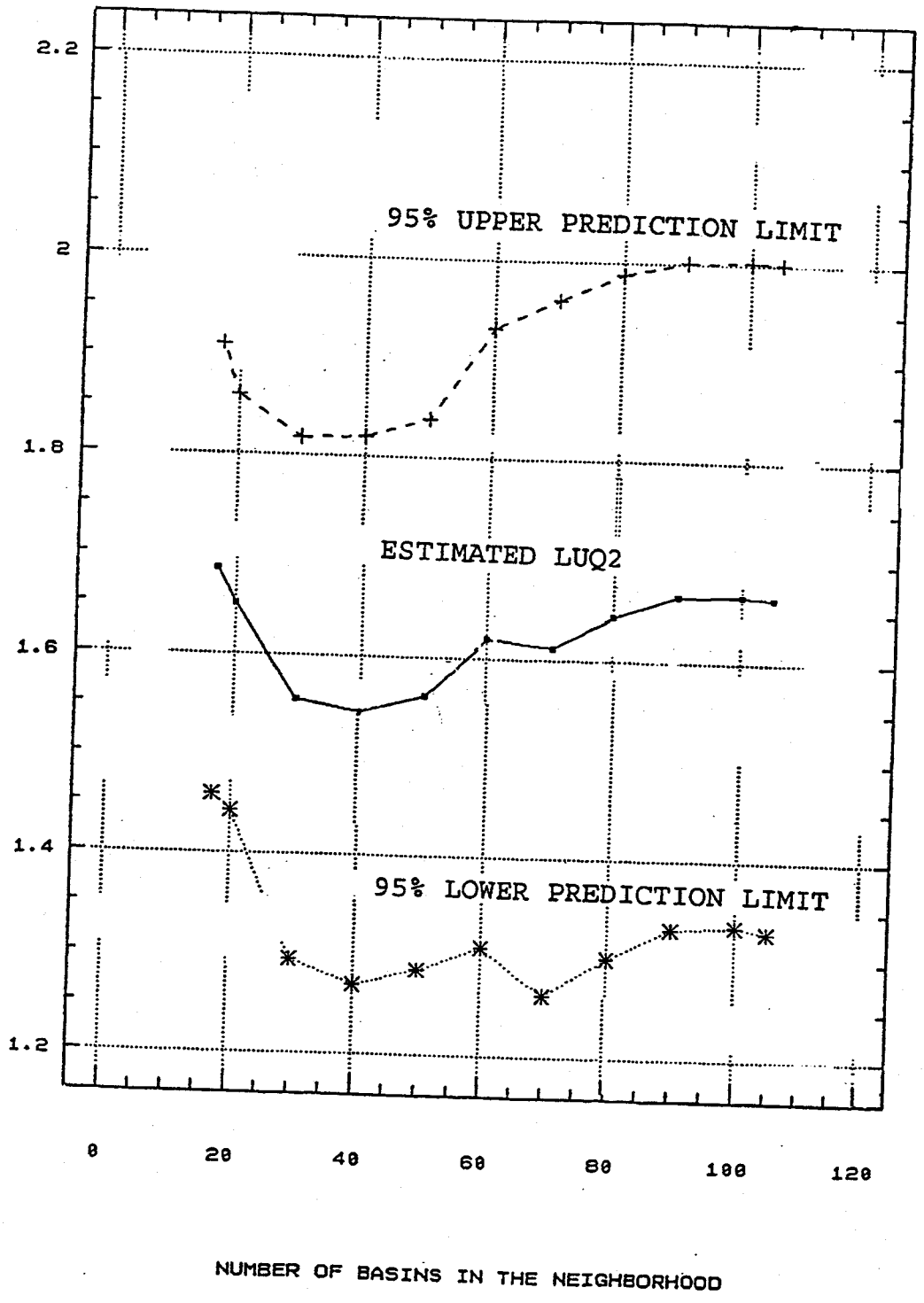

Figure 9. Estimated LUQ2 of the ungauged basin (46) with 95% prediction intervals vs the number of stations in the neighborhood.

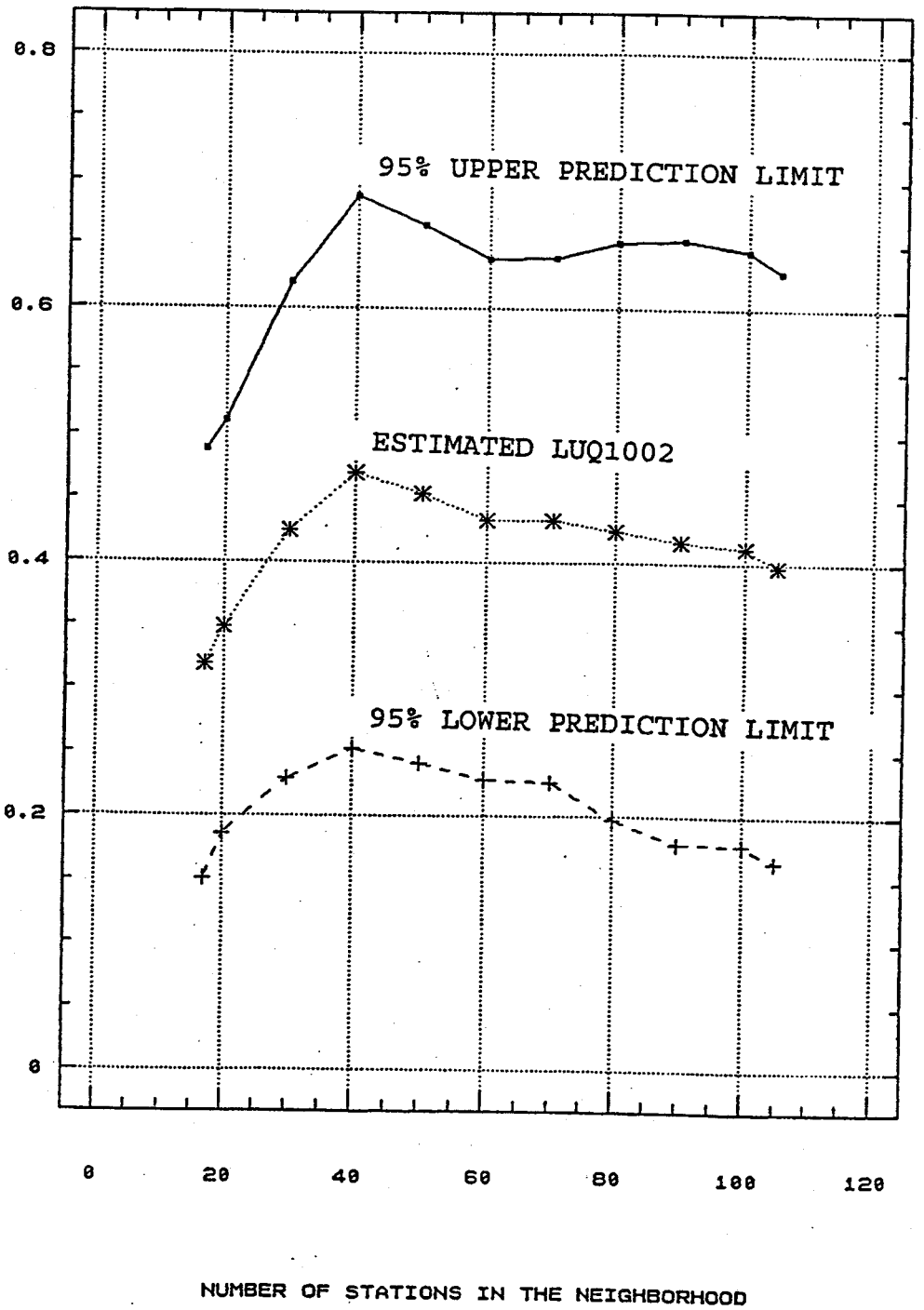

Figure 10. Estimated LUQ1002 of the ungauged basin (46) with 95% prediction intervals vs the number of stations in the neighborhood.

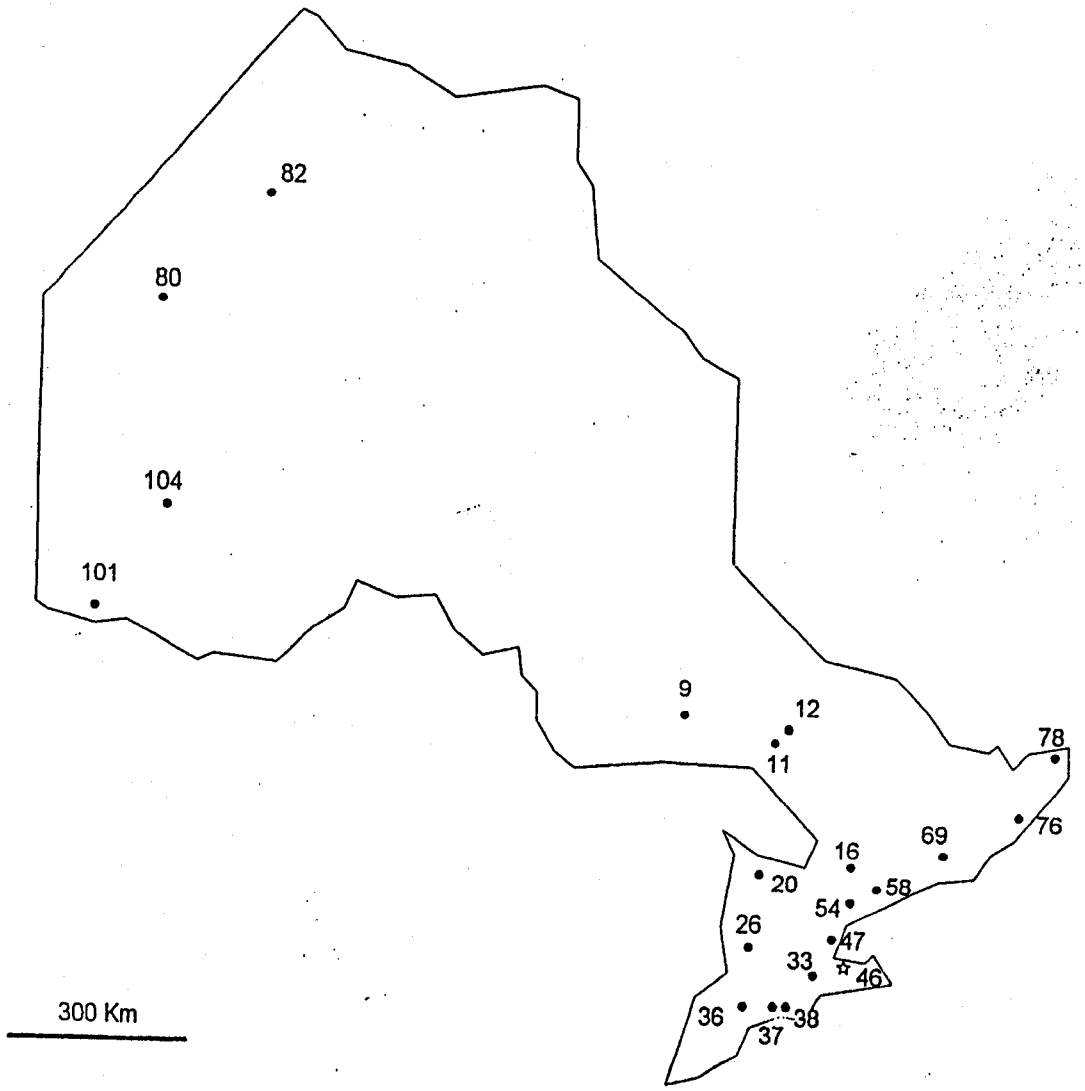

Figure 11. Map of Ontario with the gauging stations of the 20-basin neighborhood.

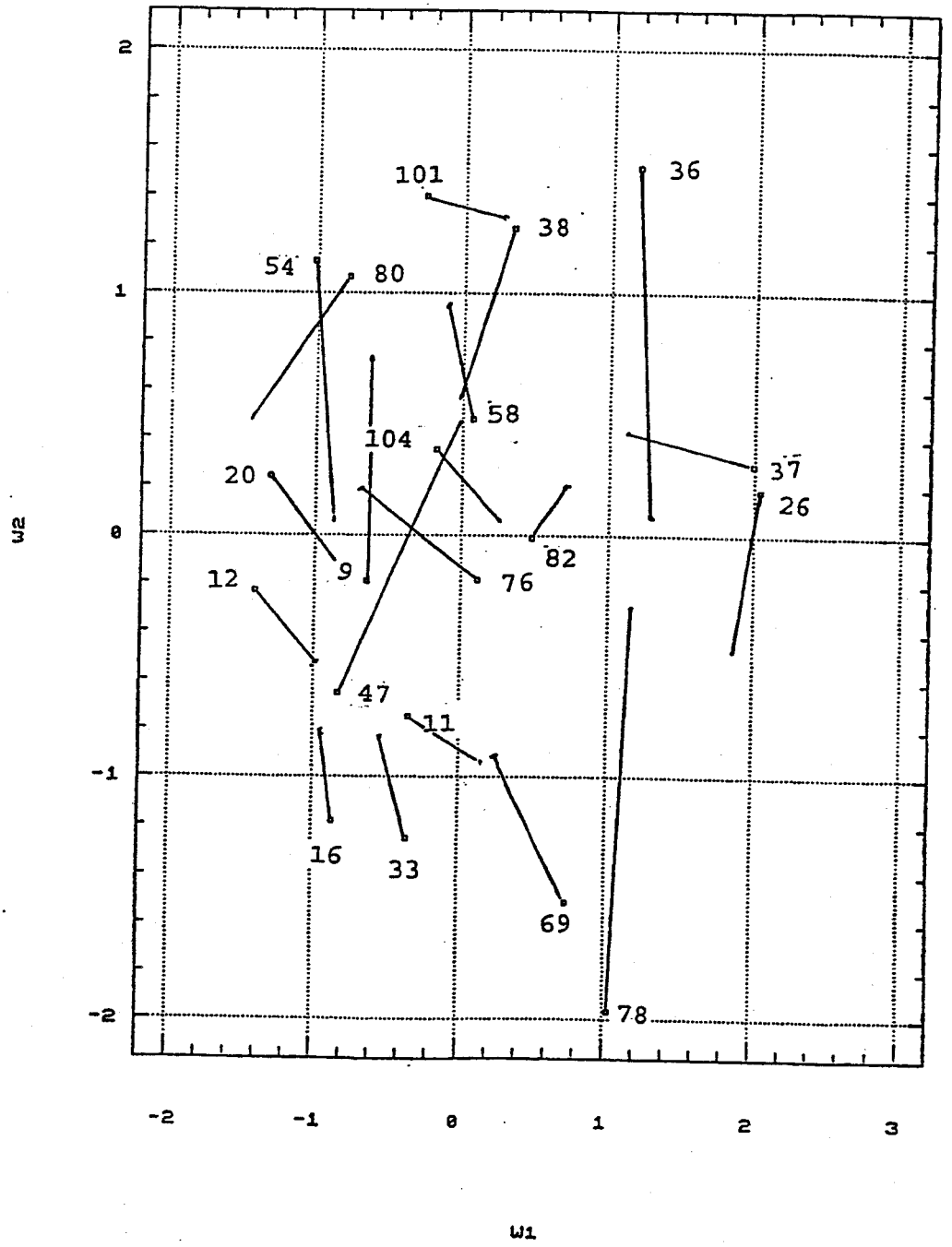

Figure 12. Error vector diagram in the space (w_1, w_2) for the 20-basin neighborhood.

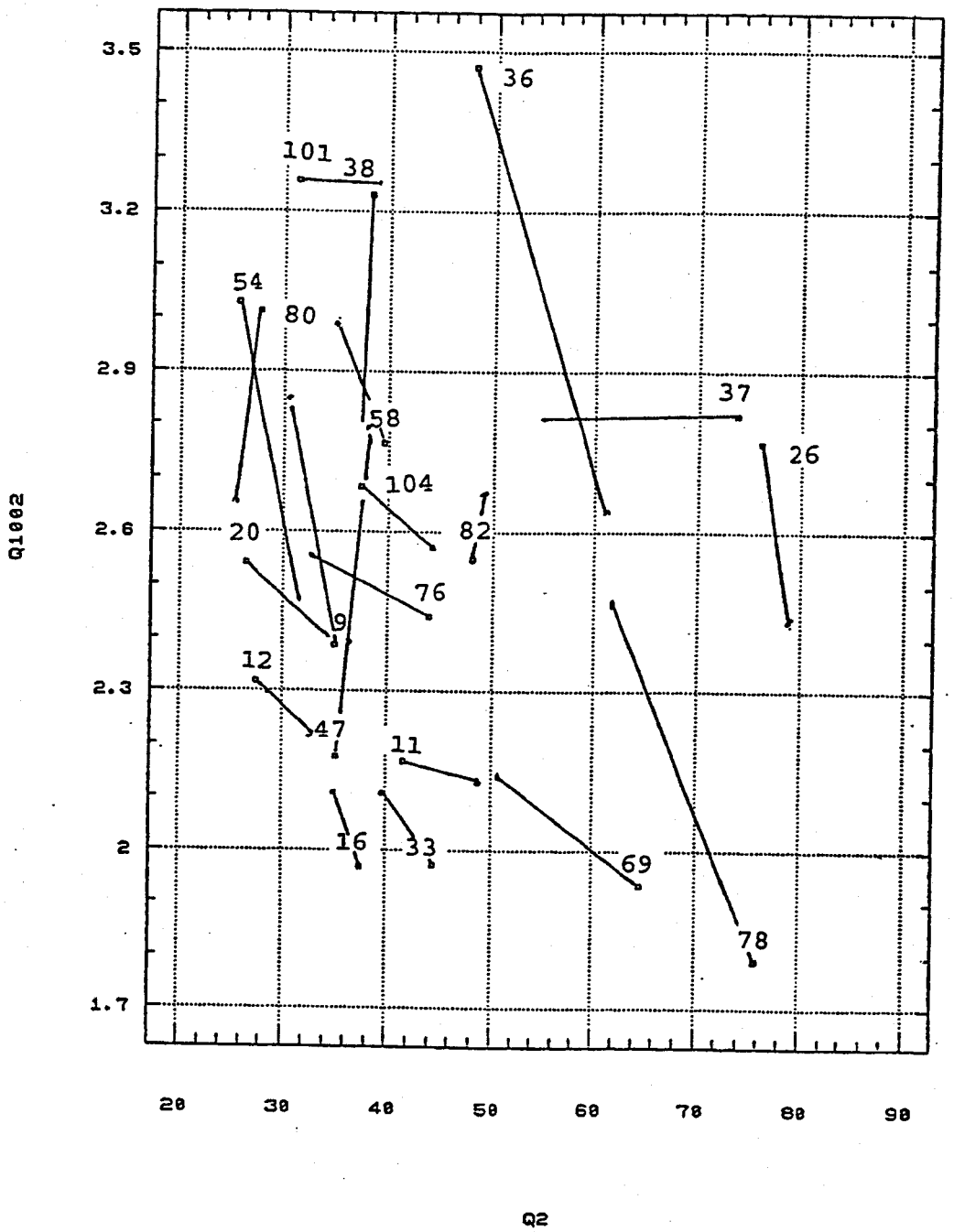

Figure 13. Error vector diagram in the space (Q2, Q1002) for the 20-basin neighborhood.

3. CORRECTIFS ET AMÉLIORATIONS APPORTÉS À L'APPROCHE RÉGIONALE BASÉE SUR L'ACC

REGIONAL FLOOD FREQUENCY ESTIMATION WITH CANONICAL CORRELATION ANALYSIS

Taha B.M.J. Ouarda, Claude Girard, George S. Cavadias, and Bernard Bobée

Chair in Statistical Hydrology,
INRS-Eau, University of Quebec
2800 Einstein, C.P. 7500,
Sainte-Foy, Qc,
Canada, G1V 4C7
Tel: (418) 654-3842,
Fax: (418) 654-2600,
E-mail: ouardata@inrs-eau.uquebec.ca

Submitted to
Journal of Hydrology

ABSTRACT: Despite its potential advantages, canonical correlation analysis has been little used in the fields of hydrology and water resources. In a regional flood frequency analysis, canonical correlations can be used to investigate the correlation structure between the two sets of variables represented by watershed characteristics and flood peaks. This paper presents a clear theoretical framework for the use of canonical correlations in regional flood frequency analysis. Some additional results are also presented for the case of gauged target-basins. The approach described in this paper allows to carry out the determination of homogeneous hydrologic neighborhoods and identifies the variables to use during the step of regional estimation. A data set of 106 stations from the province of Ontario (Canada) is used to demonstrate the advantages of this method and investigate various aspects in relation with its robustness. Results indicate that the method is robust to such factors as the number of stations and the type of parametric distribution being used. Step-by-step algorithms for the delineation of hydrologic neighborhoods in the cases of gauged and ungauged basins are also presented.

Key words :

Canonical correlation analysis, regional estimation, at-site estimation, multivariate regression, hydrologic neighborhood, non-central Chi-squared distribution, error vector.

1 INTRODUCTION AND BRIEF REVIEW

Canonical correlation analysis (CCA) plays an important role in multivariate statistics by providing the general theoretical framework for the techniques of factorial discriminant analysis, multivariate regression and correspondence analysis, and by establishing the link between their theories. In fact, the techniques of factorial discriminant analysis and multivariate regression represent special cases of the method of CCA. CCA permits to establish the interrelations that may exist between two groups of variables, by identifying the linear combinations of the variables of the first group that are the most correlated to some linear combinations of the variables of the second group. Hence, it can be easily seen that this technique degenerates to the multiple regression technique when one of the groups is reduced to a single variable (Lebart et al., 1977).

The first applications of multivariate analysis tools for hydrological analysis were made by Snyder (1962) and Wong (1963). Factorial analysis applications in hydrology were first discussed by Matalas & Reiher (1967) and Wallis (1967). Torranin (1972), Rice (1972), and Decoursey (1973) presented the first applications of canonical correlations in hydrology. Torranin (1972) demonstrated the potential for application of canonical correlations to hydrologic problems through the study of the forecasts of coastal monthly precipitation and seasonal snowmelt runoff, while Rice (1972) applied CCA to hydrological prediction. More recently CCA-based linear operational models for long term rainfall forecasting using El-Nino-Southern Oscillation (ENSO) indices have been proposed

(Barnston et al., 1994, 1996; Shabbar and Barnston, 1996; Aceituno and Montecinos, 1996; He and Barnston, 1996; Thiao and Barnston, 1997).

Regional frequency analysis is commonly used for the estimation of extreme hydrological or meteorological events (such as floods) at sites where little or no data are available (Stedinger and Tasker, 1986 ; Burn, 1990a; Rosbjerg and Madsen, 1994; Durrans and Tomic, 1996; Nguyen and Pandey, 1996; Pandey and Nguyen, 1999; Alila, 1999, 2000). The two main steps are the identification of groups of hydrologically homogeneous basins (or "homogeneous regions") and the application of a regional estimation method within each delineated region. Homogeneous regions can be defined as geographically contiguous regions, geographically non-contiguous regions, or as hydrological neighbourhoods. In the "neighborhood" approach (or "region of influence approach" as called in the terminology by Burn, 1990a) each target-site is assumed to have its own homogeneous region. Figure 1 illustrates the various approaches for the determination of homogeneous regions.

In a regional flood frequency study, DeCoursey (1973) investigated the correlation structure between the two sets of variables represented by watershed characteristics and flood peaks associated with four different return periods. Cavadias (1989, 1990) pioneered the use the CCA approach to estimate the quantiles of the maximum annual flood distribution in the province of Newfoundland, Canada. The approach proposed by Cavadias (1989, 1990) is based on a visual judgement of clustering patterns that may be available. Ribeiro-Corréa et al. (1991) and Cavadias (1995), extended the approach to determining homogeneous hydrological "neighborhoods" and applied it to the regionalization of flood

flows. Ribeiro-Corréa et al. (1995) proposed the approach of confidence level ellipsoids for the identification of hydrological neighborhoods and indicated that results of regression analysis for estimating flood quantiles could be enhanced by using hydrological neighborhoods defined with the CCA approach. Bates et al. (1998) classified a set of Australian basins in homogeneous regions, then verified the results of this classification using several multivariate techniques including canonical correlations. In Bates et al. (1998), the use of CCA was restricted to a comparison of the canonical variable scores and loadings for the homogeneous regions, and does not address the problem of estimating flood characteristics at ungauged basins. Ouarda et al. (2000) applied a CCA-based procedure for the joint regional estimation of flood peaks and volumes in the Northern part of the province of Quebec, Canada.

GREHYS (1996a, b) presented the results of an inter-comparison of various regional flood estimation procedures obtained by coupling four methods for delineating homogeneous regions (or neighborhoods) and seven regional estimation methods. GREHYS (1996b) concluded that CCA yields good results when combined with any of the regional estimation methods that were considered. The results of GREHYS indicated clearly that the "neighborhood" approach (Ribeiro-Corréa et al., 1991; Burn, 1990a, 1990b) for the delineation of groups of hydrologically homogeneous basins is superior to the « fixed set of regions » approach.

In this paper we present additional results concerning the use of the method of CCA for the identification of hydrological neighborhoods. The present work attempts to provide an answer to such questions as the determination of the optimal number of stations to

include in the neighborhood and the identification of the most appropriate physical and meteorological variables to use. The sensitivity of the technique to such factors as the distribution/estimation procedure used for at-site estimation and the changes in the configuration of hydrometric networks are also investigated. A brief presentation of the theoretical foundations of CCA is presented in the next section. A detailed description of the mathematical background of the method of CCA can be found in reference textbooks on multivariate statistical analysis techniques (Lebart et al., 1977; Muirhead, 1982; Krzanowski, 1988). A complete description of the approach used to apply CCA to the problem of regionalization in hydrology is presented in section 3.

2 THEORETICAL BACKGROUND

The multivariate approach of CCA is most commonly used in the context where we have two sets of random variables $\underline{X} = \{X_1, X_2, ..., X_n\}$ and $\underline{Y} = \{Y_1, Y_2, ..., Y_r\}$, $n \geq r$. For instance, the set $\underline{X}$ could contain basin physiographical and meteorological variables (e.g. watershed area, slope of main channel, mean annual precipitation) and the set $\underline{Y}$ could represent hydrologic variables such as flood quantiles. Suppose we are interested in the information content of these sets concerning the correlations among their variables. The basic approach consists in computing the correlation matrix for the variables of sets $\underline{X}$ and $\underline{Y}$. This leads to a numerical description of the interaction between any two variables. However, in order to get a clear view of the correlation structure, we have to identify the sources of interaction among these variables ; an effective way to do so is to employ CCA.

CCA allows to identify the dominant linear modes of covariability between the sets $\underline{X}$ and $\underline{Y}$.

Consider any linear combination V and W of the variables $X_1, X_2, ..., X_n$ and $Y_1, Y_2, ..., Y_r$, respectively,

$$\begin{aligned} V &= a_1X_1 + a_2X_2 + ... + a_nX_n = \mathbf{a'X} \\ W &= b_1Y_1 + b_2Y_2 + ... + b_rY_r = \mathbf{b'Y} \end{aligned} \tag{3.1}$$

where vectors are denoted in bold, and $\mathbf{c}'$ denotes the transpose of vector $\mathbf{c}$. Let $\mathbf{C}$ be the covariance matrix of the variables $X_1,...,X_n,Y_1,...,Y_r$ partitioned with respect to the sets to which they belong

$$\mathbf{C} = \begin{pmatrix} \mathbf{C}_{XX} & \mathbf{C}_{XY} \\ \mathbf{C}'_{XY} & \mathbf{C}_{YY} \end{pmatrix} \tag{3.2}$$

The correlation between the random variables V and W is given by

$$\text{corr}(V, W) = \frac{\text{cov}(V, W)}{\sqrt{\text{Var}(V)}\sqrt{\text{Var}(W)}} = \frac{\mathbf{a'C}_{XY}\mathbf{b}}{\sqrt{\mathbf{a'C}_{XX}\mathbf{a}}\sqrt{\mathbf{b'C}_{YY}\mathbf{b}}} \tag{3.3}$$

CCA allows to identify vectors $\mathbf{a}$ and $\mathbf{b}$ for which corr(V,W) is maximal. Furthermore, when identifying the maximal value for corr(V,W) we only need to consider vectors $\mathbf{a}$ and $\mathbf{b}$ for which V and W have unit variances. Indeed, for any real numbers θ and γ, the correlation function possesses the property

$$\text{corr}(\theta V, \gamma W) = \text{corr}(V, W) \tag{3.4}$$

Hence, by choosing $\theta = \frac{1}{\sqrt{Var(\mathrm{V})}}$ and $\gamma = \frac{1}{\sqrt{Var(\mathrm{W})}}$ we can restrict the search to the domain of V and W with unit variances. We use the Lagrange multipliers technique to maximize corr(V,W) subject to the constraints that Var(V)=Var(W)=1. Note that with these constraints we have

$$\mathrm{corr(V,W)=cov(V,W)=}\ \mathbf{a}'\mathrm{C}_{XY}\mathbf{b} \tag{3.5}$$

The lagrangian function L is defined as

$$\mathbf{L} = \mathbf{a}'\mathrm{C}_{XY}\mathbf{b} - \tau_1(\mathbf{a}'\mathrm{C}_{XX}\mathbf{a} - 1) - \tau_2(\mathbf{b}'\mathrm{C}_{YY}\mathbf{b} - 1) \tag{3.6}$$

where τ_1 and τ_2 are the lagrangian parameters. The partial derivatives of the lagrangian function with respect to **a** and **b** are given by

$$\frac{\partial \mathbf{L}}{\partial \mathbf{a}} = \mathrm{C}_{XY}\mathbf{b} - 2\tau_1\mathrm{C}_{XX}\mathbf{a} \tag{3.7}$$

$$\frac{\partial \mathbf{L}}{\partial \mathbf{b}} = \mathrm{C}'_{XY}\mathbf{a} - 2\tau_2\mathrm{C}_{YY}\mathbf{b} \tag{3.8}$$

Equaling equations (3.7) and (3.8) to zero to obtain the maximum of **L** we have

$$\mathrm{C}_{XY}\mathbf{b} = 2\tau_1\mathrm{C}_{XX}\mathbf{a} \tag{3.9}$$

$$\mathrm{C}'_{XY}\mathbf{a} = 2\tau_2\mathrm{C}_{YY}\mathbf{b} \tag{3.10}$$

By pre-multiplying (3.9) by $\mathbf{a}'$ and (3.10) by $\mathbf{b}'$, and given that $\mathbf{a}'C_{XX}\mathbf{a} = \mathbf{b}'C_{YY}\mathbf{b} = 1$, we have

$$2\tau_1 = \mathbf{a}'C_{XY}\mathbf{b} \tag{3.11}$$

$$2\tau_2 = \mathbf{b}'C'_{XY}\mathbf{a} \tag{3.12}$$

Since $\mathbf{a}'C_{XY}\mathbf{b}$ is a scalar and given that $(\mathbf{a}'C_{XY}\mathbf{b})' = \mathbf{b}'C'_{XY}\mathbf{a}$ we conclude that expressions (3.11) and (3.12) are equivalent to

$$\lambda = \mathbf{a}'C_{XY}\mathbf{b} \tag{3.13}$$

where $\lambda = 2\tau_1 = 2\tau_2$. Isolating $\mathbf{a}$ in (3.9) and substituting it into (3.10) we have

$$\frac{1}{\lambda}C'_{XY}C_{XX}^{-1}C_{XY}\mathbf{b} = \lambda C_{YY}\mathbf{b} \tag{3.14}$$

Expression (3.14) can also be written as

$$C_{YY}^{-1}C'_{XY}C_{XX}^{-1}C_{XY}\mathbf{b} - \lambda^2\mathbf{b} = 0 \tag{3.15}$$

In a similar way, starting with vector $\mathbf{b}$ we obtain

$$C_{XX}^{-1}C_{XY}C_{YY}^{-1}C'_{XY}\mathbf{a} - \lambda^2\mathbf{a} = 0 \tag{3.16}$$

Thus, the vectors $\mathbf{a}$ and $\mathbf{b}$ we seek are eigenvectors of $C_{XX}^{-1}C_{XY}C_{YY}^{-1}C'_{XY}$ and $C_{YY}^{-1}C'_{XY}C_{XX}^{-1}C_{XY}$, respectively. By substituting (3.13) into (3.5) we deduce that for the

optimal vectors **a** and **b** we have corr(V,W)=λ, the square-root of the corresponding eigenvalue. If p is the rank of C_{XY}, then solving equations (3.15) and (3.16) leads to p solution-triplets $(\lambda_i, \mathbf{a}_i, \mathbf{b}_i)$, each giving rise to a triplet (λ_i, V_i, W_i), according to (3.1). Note that

$$\lambda_i = \text{corr}(V_i, W_i) \tag{3.17}$$

Two important properties of these solutions are (see appendix A):

(P1) : distinct W_i, as well as distinct V_i, are uncorrelated,

(P2) : W_i and V_j, $i \neq j$, are uncorrelated.

In retrospect, CCA provides us with new variables, known as canonical variables, $V_1,V_2,...,V_p$ and $W_1,W_2,...,W_p$ (to replace $X_1,X_2,...,X_n$ and $Y_1,Y_2,...,Y_r$, respectively), along with the correlation coefficients $\lambda_1,\lambda_2,...,\lambda_p$. Note that, in regard to properties (P1) and (P2), each of the V_i's and W_i's can be seen as a distinct source of correlation, thus revealing the correlation structure within and among the sets $\underline{X}$ and $\underline{Y}$.

3 APPLICATION OF CCA TO REGIONAL ESTIMATION

In regional hydrologic estimation, the set $\underline{X}$ contains physiographic and meteorologic variables for a given set $\underline{B}$ of basins and the set $\underline{Y}$ represents basin hydrologic variables, such as flood quantiles corresponding to different return periods. Thus, every basin in $\underline{B}$ acts as a realization of all these random variables. We require all variables to be transformed for normality, and standardized, the reason for which will presented thereafter.

CCA can then be performed to obtain canonical variables V_i and W_i. For each basin B_k, k=1,...,K, of $\underline{B}$ correspond values for V_i and W_i denoted as $v_{i,k}$ and $w_{i,k}$. The vectors $\mathbf{v}_k$ and $\mathbf{w}_k$ are formed by assembling all values $v_{i,k}$ and $w_{i,k}$ for a given basin B_k.

We assume that the vector $\begin{pmatrix}\mathbf{W}\\ \mathbf{V}\end{pmatrix}$, of which all $\begin{pmatrix}\mathbf{w}_k\\ \mathbf{v}_k\end{pmatrix}$ are realizations, is 2p-normally distributed with mean-vector $\mathbf{O}_{2p\times 2p}$ and covariance matrix

$$L=\begin{pmatrix}I_p & \Lambda\\ \Lambda & I_p\end{pmatrix} \tag{3.18}$$

where $\Lambda=\mathrm{diag}(\lambda_1,...,\lambda_p)$ and I_p is the $p\times p$ identity matrix. We note that covariance matrix L is simply a compact way of describing properties (P1) and (P2) along with equation (3.17). The multinormality assumption cannot be easily tested. Note that having normal marginal distributions for W and V is a necessary, but not sufficient condition to multinormality. In practice, we transform the original variables to ensure normality. In CCA, standardizing the data prior to the analysis is usually done to obtain canonical variables that are free of scaling effects; which makes it more difficult to interpret canonical variables in terms of the original variables. Standardization is also carried out in order to obtain a simple expression for the mean-vector of $\begin{pmatrix}\mathbf{W}\\ \mathbf{V}\end{pmatrix}$.

We denote $\mathbf{v_0}$ the canonical score corresponding to the physiographic values of a target-basin, that is, the corresponding values of the canonical variables for the target-basin with respect to its physiographic scores. We are then interested in the inferences that can be

made on the corresponding hydrological canonical score $\mathbf{w}_0$. It is then possible to consider the distribution of all **W** having an associated canonical variable **V** that has realized into $\mathbf{v}_0$, i.e. $\mathbf{V}=\mathbf{v}_0$. According to Muirhead (1982, p.12) the conditional density of **W** given that $\mathbf{V}=\mathbf{v}_0$ is

$$(\mathbf{W}|\mathbf{V}=\mathbf{v}_0)\approx N_p(\Lambda\mathbf{v}_0, I_p-\Lambda^2) \tag{3.19}$$

More explicitly, these **W** would be scattered around a mean position $\Lambda\mathbf{v}_0$ according to

$$f_{\mathbf{W}|\mathbf{V}=\mathbf{v}_0}(\mathbf{w}|\mathbf{v}_0)=(2\pi)^{-p}\left|I_p-\Lambda^2\right|^{-1/2}\exp\left[\frac{-1}{2}\underbrace{(\mathbf{w}-\Lambda\mathbf{v}_0)'(I_p-\Lambda^2)^{-1}(\mathbf{w}-\Lambda\mathbf{v}_0)}_{D^2}\right] \tag{3.20}$$

In these conditions it can be shown (Muirhead, 1982, p.26) that D^2, being a Mahalanobis distance, has a khi-square χ^2 distribution with p degrees of freedom. A definition of a homogeneous neighborhood for the target-basin can then be deduced. Indeed, D^2 can serve to indicate where, in the canonical space ***W*** that encompasses all possible realizations of the random variable **W**, would be found the realizations **w** of **W** for which **V** has realized in $\mathbf{v}_0$. In practice, two cases can arise : the target-basin can be ungauged or gauged.

3.1. Ungauged target-basin

In this case, no hydrologic data that would translate into information about $\mathbf{w}_0$ is available for the target-basin. We suppose that physiographic information is available and, consequently, that the score $\mathbf{v}_0$ of the target-basin on the canonical vector **V** is known.

We can now proceed to identify and exclude the realizations **w** that are located far from the mean position $\Lambda\mathbf{v}_0$ of the target-basin. The definition of the homogeneous neighborhood excludes then the tail of the distribution of D^2 in the following manner : A realization **w** of **W** is said to be part of the 100(1-α)% confidence level neighborhood for $\Lambda\mathbf{v}_0$ if, and only if, it satisfies :

$$(\mathbf{w}-\Lambda\mathbf{v}_0)'(\mathrm{I}_p-\Lambda^2)^{-1}(\mathbf{w}-\Lambda\mathbf{v}_0)\leq\chi^2_{\alpha,p} \qquad (3.21)$$

where $\chi^2_{\alpha,p}$ is such that $P(\chi^2\leq\chi^2_{\alpha,p})=1-\alpha$. In the canonical space *W*, the neighborhood defined by expression (3.21) describes the interior of an ellipsoidal region. Although the neighborhood is defined in the canonical space *W*, a basin-related physical interpretation is possible. Indeed, **w** is simply the score of a basin on the canonical variables which are linear combinations of the hydrologic variables. In appendix B we present a step-by-step algorithm for the delineation of the hydrologic neighborhood in the case of an ungauged basin.

3.2. Gauged target-basin

We now consider the case where a few years of hydrologic data are available at the target-basin, though this limited amount of information may not be sufficient for at-site

flood frequency estimation. Nonetheless we are interested in using this information to acquire additional knowledge about $\mathbf{W}_0$ and to improve the regional estimates.

Ribeiro-Corréa et al. (1995) propose to use $\mathbf{w}_0$ instead of $\Lambda\mathbf{v}_0$ as an indicator of the real location of $\mathbf{W}_0$. Although this seems adequate, the probability statement that results [equation (10) in Ribeiro-Corréa et al.] is flawed:

$$P\left((\mathbf{W}-\mathbf{W}_0)'\left(I_p-\Lambda^2\right)^{-1}(\mathbf{W}-\mathbf{W}_0)\le\chi^2_{\alpha,p}\right)=1-\alpha \qquad (3.22)$$

Indeed, for equation (3.22) to be true we would need the quadratic form $(\mathbf{W}-\mathbf{W}_0)'\left(I_p-\Lambda^2\right)^{-1}(\mathbf{W}-\mathbf{W}_0)$ to have a central chi-squared χ^2 distribution. However, this is not true. To illustrate this we denote $\mathbf{Z}=\mathbf{W}-\mathbf{W}_0$ and recall that $(\mathbf{W}|\mathbf{V}=\mathbf{v}_0)\approx N_p\left(\Lambda\mathbf{v}_0, I_p-\Lambda^2\right)$ from equation (3.19). We now have

$$\mathbf{Z}\approx N_p\left(\Lambda\mathbf{v}_0-\mathbf{W}_0, I_p-\Lambda^2\right) \qquad (3.23)$$

The quadratic expression $(\mathbf{W}-\mathbf{W}_0)'\left(I_p-\Lambda^2\right)^{-1}(\mathbf{W}-\mathbf{W}_0)$ becomes then $\mathbf{Z}'\left(I_p-\Lambda^2\right)^{-1}\mathbf{Z}$, which has a non-central chi-squared distribution $\chi^2_p(\lambda)$ with p degrees of freedom and non-centrality parameter λ (Muirhead, 1982, p.26) :

$$\lambda=(\Lambda\mathbf{v}_0-\mathbf{W}_0)'\left(I_p-\Lambda^2\right)^{-1}(\Lambda\mathbf{v}_0-\mathbf{W}_0) \qquad (3.24)$$

Therefore the correct probability statement involving the quadratic form $(\mathbf{W}-\mathbf{W}_0)'(\mathrm{I}_\mathrm{p}-\Lambda^2)^{-1}(\mathbf{W}-\mathbf{W}_0)$ should read

$$\mathrm{P}\left((\mathbf{W}-\mathbf{W}_0)'(\mathrm{I}_\mathrm{p}-\Lambda^2)^{-1}(\mathbf{W}-\mathbf{W}_0)\leq \chi^2_{\alpha,\mathrm{p}}(\lambda)\right)=1-\alpha \tag{3.25}$$

Using this expression requires the evaluation of the term $\chi^2_{\alpha,\mathrm{p}}(\lambda)$. We can then use the following approximation due to Pearson (Johnson et al., 1995, chap. 29) linking the non-central chi-squared distribution to the central chi-squared distribution

$$\mathrm{P}\left(\chi^2_\mathrm{p}(\lambda)\leq x\right)\cong \mathrm{P}\left(\mathrm{c}\chi^2_\mathrm{f}+b\leq x\right) \tag{3.26}$$

where the appropriate values of *b*, *c* and *f* are

$$b=-\frac{\lambda^2}{\mathrm{p}+3\lambda},\quad c=\frac{\mathrm{p}+3\lambda}{\mathrm{p}+2\lambda},\quad f=\frac{(\mathrm{p}+2\lambda)^3}{(\mathrm{p}+3\lambda)^2} \tag{3.27}$$

It was verified that this approximation is very precise for the purpose of this application. Since the number of degrees of freedom *f* is in most cases fractional, expression (3.26) needs to be rewritten in terms of the gamma distribution, of which the chi-squared distribution is a special case. This leads to

$$\mathrm{P}\left(\chi^2_\mathrm{p}(\lambda)\leq x\right)\cong \mathrm{P}\left(\mathrm{c}\gamma\left(\frac{f}{2},\mathrm{p}\right)+b\leq x\right)=\mathrm{P}\left(\gamma\left(\frac{f}{2},\mathrm{p}\right)\leq\frac{x-b}{c}\right) \tag{3.28}$$

where γ denotes the gamma distribution. Since the tabulated value $\gamma_\alpha\left(\frac{f}{2},\mathrm{p}\right)$ is the only number such that

$$\mathrm{P}\left(\gamma\left(\frac{f}{2},\mathrm{p}\right)\le\gamma_\alpha\left(\frac{f}{2},\mathrm{p}\right)\right)=1-\alpha \tag{3.29}$$

then combining (3.29), (3.28) and (3.25) leads to the following approximation of $\chi^2_{\alpha,\mathrm{p}}(\lambda)$

$$\chi^2_{\alpha,\mathrm{p}}(\lambda)\approx c\gamma_\alpha\left(\frac{f}{2},\mathrm{p}\right)+b \tag{3.30}$$

where *b*, *c* and *f* are given by (3.27). Therefore, from equations (3.25) and (3.30) it can be deduced that a 100(1-α)% confidence level neighborhood about $\mathbf{w}_0$ is constituted of all realizations **w** of **W** satisfying

$$(\mathbf{w}-\mathbf{w}_0)'\left(\mathrm{I}_\mathrm{p}-\Lambda^2\right)^{-1}(\mathbf{w}-\mathbf{w}_0)\le c\gamma_\alpha\left(\frac{f}{2},\mathrm{p}\right)+b \tag{3.31}$$

where *b*, *c* and *f* are defined in (3.27).

Instead of using the neighborhood defined by (3.31) it is possible to define a neighborhood about $\Lambda\mathbf{v}_0$ in the same manner than was done for the ungauged case : A 100(1-α)% confidence level neighborhood about $\Lambda\mathbf{v}_0$ is constituted of all realizations **w** of **W** meeting the condition

$$(\mathbf{w}-\Lambda\mathbf{v}_0)'\left(\mathrm{I}_\mathrm{p}-\Lambda^2\right)^{-1}(\mathbf{w}-\Lambda\mathbf{v}_0)\le\chi^2_{\alpha,\mathrm{p}} \tag{3.32}$$

where $\chi^2_{\alpha,p}$ is such that $P\left(\chi^2 \leq \chi^2_{\alpha,p}\right) = 1-\alpha$.

Neighborhood definitions (3.31) and (3.32) differ in two essential ways. First, definition (3.31) makes use of the local hydrologic information that is available, whereas (3.32) rests entirely upon the estimate Λv_0. Secondly, although they both describe the interior of an ellipsoidal region in the canonical space W, they are not centered at the same point. The second difference is clearly a result of the first one. It is reasonable to prefer neighborhood definition (3.31) since it uses available at-site information. It is also possible to use a weighted combination of estimates w_0 and Λv_0. The weights would then reflect the amount of data that is available at the target-basin and the level of confidence that can be allocated to this limited at-site information. Further efforts could be devoted in the future to address this point. We list in appendix C a complete step-by-step algorithm, based on equation (3.31) for the delineation of the hydrologic neighborhood in the case of a gauged basin. The algorithms presented in appendices B and C have been easily programmed in the Matlab environment.

4 DATA SET FOR CASE STUDY

A data set of 106 drainage basins from the province of Ontario (Canada) is used. Figure 2 illustrates the location of gauging stations across the province of Ontario. Catchments are characterized by the following physiographic and meteorological variables :

- AREA, drainage area (km^2) ;

- LCP, length of the main channel (km) ;
- PCP, slope of the main channel (m/km) ;
- SLM, basin area controlled by lakes (km^2) ;
- LAT, gauging station latitude ;
- LONG, gauging station longitude ;
- PTMA, mean total annual precipitation (mm)

The 2-year (Q_2) and 100-year (Q_{100}) flood quantiles are computed, and the flood regime in the province is characterized by the two variables Q_2 and Q_{100}/Q_2. These two flood distribution characteristics are respectively representative of the median annual flood and the dispersion of the distribution. The various models that are used for the computation of these flood quantiles are discussed in a subsequent section of the paper.

5 OPTIMAL NUMBER OF STATIONS IN THE NEIGHBORHOOD

The use of geographically defined homogeneous regions is convenient for practical purposes. However, geographical proximity is not a guarantee of hydrological similarity. The CCA method discussed in this paper adopts the neighborhood approach to the definition of homogeneous regions. This approach considers each basin as having its own homogeneous region, or neighborhood. The notion of degree of homogeneity is directly related to the size of the neighborhood considered. Indeed, the consideration of a small neighborhood guarantees that only the nearest basins in the canonical hydrological space are included, ensuring therefore a high degree of homogeneity within the neighborhood.

However, when the number of basins in the neighborhood is too small it is difficult or even impossible to carry out an appropriate statistical estimation within the neighborhood (such as the estimation of regression coefficients). On the other side, the adoption of large neighborhoods ensures the availability of an adequate number of observations (basins) to perform the regional analysis, but may introduce in the neighborhood basins that are not quite similar to the target one. To address this trade-off between homogeneity and amount of observations, rational criteria for the identification of the optimal number of observations need to be developed.

CCA was originally carried out using the whole 106-basin data set. With n = 7 (number of physical/meteorological variables) and r = 2 (number of hydrological variables), two pairs of canonical variables can be defined. The physiographical and meteorological variables that were found to be most significant in the interpretation of the canonical variables are AREA, LCP, PCP, SLM, and PTMA. This confirms that geographical vicinity is not a guaranty of hydrological homogeneity, as the latitude and longitude are not among these variables. The hydrological and physiographical canonical variables are given by

$$V_1 = 1.22Ln(AREA) + 0.12Ln(LCP) + 0.17Ln(PCP) - 0.26Ln(SLM) + 0.27Ln(PTMA) \quad (3.33)$$

$$V_2 = 1.99Ln(AREA) + 0.30Ln(LCP) - 1.84Ln(PCP) - 0.49Ln(SLM) - 0.76Ln(PTMA) \quad (3.34)$$

$$W_1 = 1.02Ln(Q_2) + 0.08Ln(Q_{100}/Q_2) \quad (3.35)$$

$$W_2 = 0.31Ln(Q_2) + 1.07Ln(Q_{100}/Q_2) \quad (3.36)$$

A jack-knife resampling procedure was then used to address the compromise between the conflicting objectives of increasing the degree of similarity between the basins of the neighborhood and increasing the size of the neighborhood itself. Each one of the 106 basins of the province was in turn considered as an ungauged target-basin and removed from the data base. A pilot study was then carried out to investigate the sensitivity of the results to the estimation method. Two estimation methods were tested : the index-flood procedure and the multiple regression model. The results of the pilot study confirmed the conclusions of GREHYS (1996b) indicating that CCA-based regional flood estimates were very insensitive to the estimation method used. Final estimates of quantiles were then obtained in each target-basin using the canonical correlation analysis approach coupled with an index-flood estimation procedure.

Various confidence levels 100(1-α) % were considered by varying the neighborhood parameter α. This allows to study the impact of considering a large range of neighborhood sizes with all available target-basins. Two performance indices were considered

$$BIAS(\alpha) = \frac{rBIAS_{\alpha}}{rBIAS_{0}} \tag{3.37}$$

$$MSE(\alpha) = \frac{rMSE_{\alpha}}{rMSE_{0}} \tag{3.38}$$

where *rBIAS* and *rMSE* are successively the quantile estimation relative bias and relative mean square error corresponding to a confidence level of 100(1-α) %. The case α=0 corresponds to the inclusion of all basins in the neighborhood of the target-site. The results are illustrated by figure 3 which indicates, in the case of the 100-year quantile, that the optimal number of stations to include in the neighborhood corresponds to a value of α ranging between 0.10 and 0.20. The corresponding mean number of stations is in the range of 23 to 30 stations. Similar results are obtained with the 2-year quantile. These results are confirmed by figure 4, in which the same performance indices are presented as a function of the mean number of stations in the neighborhood. The threshold level that was adopted for the remainder of the study is α=0.20 corresponding to a mean number of neighboring stations of 23.

6 SENSITIVITY TO AT-SITE ESTIMATION

The application of any regionalization technique requires the availability of at-site estimates of quantiles in all gauged basins. The computation of these local quantile estimates is usually based on the adoption of a regional distribution. The choice of distribution and estimation method may have a high impact on the regionalization results. In order to test the sensitivity of the CCA technique to the choice of regional distribution/estimation (D/E) procedure, five D/E procedures are selected : 1) General Extreme Value distribution in association with the Probability Weighted Moments estimation method (GEV-PWM), 2) General Extreme Value distribution in association with

the Maximum likelihood estimation method (GEV-ML), 3) the Log-Normal distribution with the Maximum likelihood estimation method (LN-ML), 4) the Log-Pearson type III distribution with the direct method of moments (LPIII-BOB), and 5) the non-parametric approach using an Epanechnikov kernel (NP). The use of non-parametric methods for the estimation of quantiles allows to eliminate the effect of the subjectivity due to the choice of distribution.

The same jack-knife resampling procedure is used in this section. Basins are in turn considered ungaged and estimates of quantiles are obtained using the remaining basins. Results are presented in Table 1 for the canonical correlation coefficients and components of the physical canonical vectors V_i and the hydrological canonical vectors W_i. These results indicate that the components of both pairs of canonical vectors are very insensitive to the D/E procedure being used. This is even more true for the first vector of each pair where the difference occurs often at the level of the third decimal. Results of quantile estimation relative bias and relative root mean square error were also produced and confirmed these results.

7 SENSITIVITY TO CHANGES IN THE CONFIGURATION OF HYDROMETRIC NETWORKS

Significant changes in the configuration of hydrometric networks are presently taking place in Canada and some other countries. In many regions, large number of hydrometeorologic stations are being relocated or removed to eliminate "information

redundancy", while in other regions new stations are being added to the hudrometric network. Therefore it is important to study the sensitivity of the CCA approach to these changes in the hydrologic monitoring networks.

A bootstrap resampling procedure (Efron, 1979, 1987) was developed in which 10 basins are randomly selected and assumed ungaged. At each stage of the procedure, these 10 basins are, in turn, removed from the data base and considered as target-basins. Estimates of flood quantiles are then obtained in these sites by considering various network rationalization scenarios. Rationalization schemes considering the elimination of 1, 2, 3, and up to 53 stations (50% of the network of the province) were considered. Some of these schemes are clearly extreme and are just intended to study the sensitivity of the CCA to changes in the configuration of the monitoring network. The relocation of a station can be considered as the removal of a station and then the installation of a new one, and can consequently be considered as a special case of the scenarios modeled. Note that the number of possible combinations of stations to be removed increases quickly and can become very important. For instance, in the case of the elimination of 5 stations, there are around 100 million possible combinations (C_5^{105} =96 560 646). At each stage, 1000 combinations of station elimination are bootstrapped and estimates of quantiles are obtained with the remaining network. Elimination schemes in which more than 53 stations are removed are not considered as they are non-realistic and they sometimes lead to small neighborhood sizes. The relative estimation root mean square error (denoted *RMSE*) is adopted as an index to study the sensitivity of the canonical correlation analysis approach to these network rationalization scenarios.

$$RMSE = 100\left\{\frac{1}{10}\sum_{j=1}^{10}\left[\sqrt{\frac{1}{1000}\sum_{i=1}^{1000}\frac{(\tilde{q}(i,j)-q(j))^2}{q(j)^2}}\right]\right\}\% \qquad (3.39)$$

where i is the index of quantile estimates (or station elimination combination), j is the index of the target-station, $q(j)$ is the at-site quantile estimate, and $\tilde{q}(i,j)$ is the regional flood quantile estimate at site j with station elimination combination i.

Figure 5 illustrates the results for the 2-year and 100-year quantiles. The values of the RMSE index start increasing with the first reductions in the number of stations. However, these results suggest that the method is robust and relatively insensitive to small and medium changes in network configurations. The relative root mean square error is consistently higher for the two-year return period flood than the hundred-year return period flood. However *RMSE* values for the latter seem to be more sensitive to these network changes. Finally, it is important to underline the long-term role and the importance of hydrometric data collection networks, as well as their benefits to society and their usefulness for the safety of citizens. The exercise presented herein is simply an investigation to the impact of changes in the hydrologic monitoring networks on the performance of the CCA-based regional model. The results of this exercise indicate clearly that, even with the most robust estimation procedures, any reduction in the size of hyrometric networks can be translated by a loss of information and a reduction in the quality of design flood estimates.

8 ANALYSIS OF ERROR VECTORS

The analysis of estimation error vectors for canonical variables can be achieved by plotting the points $W=(W_1, W_2)$ and $\hat{W} = (\hat{W}_1, \hat{W}_2)$ (where $\hat{W}_i$ is the estimate of W_i), and then identifying the vector $\vec{E} = \hat{W}W$. The two components of this error vector are uncorrelated (due to the properties of canonical variables) and can be determined for each site of the study. Similar error vectors $\vec{E}'$ can be identified in the space of the original hydrological variables Q_2 and Q_{100}/Q_2. These error vectors are readily interpretable and can be useful in analyzing the performance of the method. Figure 6 presents the scatter diagram of the error vectors $\vec{E}'$ for the first 11 basins of the study.

The error vector $\vec{E}'$ scatter plot for the whole province is examined to study any pattern that may be identifiable concerning the components of these vectors. The angle between the error vector and the horizontal axis is denoted θ. It is noticed that the vertical component is generally larger than the horizontal one. No specific patterns were identified for most physiographic and meteorological variables. However, it is observed that large main channel slopes are usually associated with large values of the angle θ (the vertical component of the error vector is larger than the horizontal one) and small angles are associated with small channel slopes (figure 7). On the other side, large main channel lengths correspond to small angles and large angles correspond to small main channel lengths (figure 8). This type of results can be useful for the analysis of the relative importance of various groups of explanatory variables (physiographical, geographical,

meteorological) and to provide a clearer image of the physical factors influencing the flood phenomena.

9 CONCLUSIONS

A complete theoretical framework for the use of canonical correlation analysis in regional flood estimation, as well as some additional results for the case of gauged target-basins are presented. Step-by-step algorithms for the delineation of hydrologic neighborhoods in the cases of gauged and ungauged basins are also presented to assist practitioners with this method. Simulations based on the jack-knife and bootstrap resampling techniques, associated with a data set of 106 stations from the province of Ontario (Canada) were used to illustrate the advantages of the method of canonical correlation analysis, and its usefulness in regional flood frequency analysis. Results are presented concerning the determination of the optimal number of stations to include in the neighborhood and the identification of the most appropriate physiographical and meteorological variables to use with the method. Sensitivity analysis demonstrates that the method is stable and robust to such factors as the change in the configuration of hydrometric networks and the type of parametric/non-parametric approach used for at-site quantile estimation. Further work may be directed towards the application of the technique for the regionalization of other hydrologic variables such as precipitations, or low-flows, and may serve to familiarize the hydrologic community with the potential of this method.

ACKNOWLEDGEMENTS

The financial support provided by the Natural Sciences and Engineering Research Council of Canada (NSERC) is gratefully acknowledged. The authors wish also to thank Geneviève Boucher for her assistance.

REFERENCES

Aceituno, P., and A. Montecinos. 1996. Assessing upper limits of seasonal predictability of rainfall in central Chile based on SST in the equatorial Pacific, *Exp. Long-Lead Forecast Bull.*, 5, 5-4.

Alila, Y. 1999. A hierarchical approach for the regionalization of precipitation annual maxima in Canada, *J. Geophys. Res.* 104(D24): 31645-31655.

Alila, Y. 2000. Regional rainfall depth-duration-frequency equations for Canada. *Water Resour. Res.* 36(7): 1767-1778.

Barnston, A. G. 1994. Linear statistical short-term climate predictive skill in the northern hemisphere, *J. Clim.*, 7, 1513-1564.

Barnston, A. G., W. Thiao, and V. Kumar. 1996. Long-Lead forecasts of seasonal precipitation in Africa using CCA, *Weather Forecasting*, 11, 506-520.

Bates, B. C., A. Rahman, R. G. Mein, and P. E. Weinmann. 1998. Climatic and physical factors that influence the homogeneity of regional floods in southern Australia, *Water Resour. Res.* 34(12): 3369-3381.

Burn, D. H. 1990a. Evaluation of regional flood frequency analysis with a region of influence approach. *Water Resour. Res.* 26(10): 2257-2265.

Burn, D. H. 1990b. An appraisal of the region of influence approach to flood frequency analysis. *Hydrol. Sciences J.* 35(2): 149-165.

Cavadias, G. S. 1989. Regional flood estimation by canonical correlation. *Proceedings of the 1989 Conference of the Canadian Society of Civil Engineers*: 11A:212-231. St-John's, Newfoundland.

Cavadias, G. S. 1990. The canonical correlation approach to regional flood estimation. In M. A. Beran, M. Brilly, A. Becker, & O. Bonacci (eds), *Proceedings of the Ljubljana Symposium on Regionalization in Hydrology, IAHS Publ. No. 191*: 171-178. Wallingford, England.

Cavadias, G. S. 1995. Regionalisation and multivariate analysis: the canonical correlation approach. *Proceedings of the UNESCO International Conference on « Statistical and Bayesian Methods in Hydrological Sciences », Paris*, 19 pp.

DeCoursey, D. G. 1973. Objective regionalization of peak flow rates. Floods and Droughts. *Proc. Second International Symposium in Hydrology*. Fort-Collins, CO, Sept. 11-13, 1972. Water Resour. Publ., Fort-Collins, CO.

Durrans, S. R. and S. Tomic. 1996. Regionlisation of low-flow frequency estimates: an Alabama case study. *Water Resour. Bulletin*. 32(1): 23-37.

Efron, B. 1979. Computers and the theory of statistics : Thinking the unthinkable. *Society for Industrial and Applied Mathemtics*, 21(4) : 460-480.

Efron, B. 1987. Better Bootstrap Confidence Intervals. *Journal of the American Statistical Association*, 82(397) :171-200.

GREHYS 1996a. Presentation and review of some methods for regional flood frequency analysis. *J. Hydrol.* 186(1-4): 63-84.

GREHYS 1996b. Intercomparison of flood frequency procedures for Canadian rivers. *J. Hydrol.* 186(1-4): 85-103.

He, Y.-X., and A. G. Barnston. 1996. Long-Lead forecasts of seasonal precipitation in the tropical Pacific islands using CCA, *J. Clim.*, 9, 2020-2035.

Johnson, N., L. S. Kotz, and N. Balakrishnan. 1995. Continuous univariate distributions, vol.2. *Wiley series in probability and mathematical statistics.*

Krzanowski, W. J. 1988. Principles of multivariate analysis : a user's perspective. Clarendon Press, Oxford.

Lebart, L., A. Morineau and N. Tabard. 1977. Techniques de la description statistique, méthodes et logiciels pour l'analyse des grands tableaux. Dunod, Paris, 351 pp.

Matalas, N. C. and B. J. Reiher 1967. Some comments on the use of factor analysis. *Water Resour. Res.* 3(1): 213-224.

Muirhead, R. J. 1982. Aspects of multivariate statistical theory. J. Wiley, 673 pp.

Nguyen, V.-T.-V., and G. Pandey. 1996. A new approach to regional estimation of floods in Quebec, Delisle C.E. and Bouchard M.A. Eds, *Proceedings of the 49th Annual Conference of the CWRA*, 26-28 June, Quebec City. Collection Environnement de l'U. de M., 587-596.

Ouarda, T.B.M.J., M. Haché, P. Bruneau, and B. Bobée. 2000. Regional flood peak and volume estimation in northern canadian basin. *J. Cold Regions Eng. ASCE.* 14(4): 176-191.

Pandey, G. R., and V.-T.-V. Nguyen. 1999. A comparative study of regression based methods in regional flood frequency analysis. *J. Hydrol.* 225: 92-101.

Ribeiro-Corréa, J., G. S. Cavadias, and J. Rousselle 1991. An application of canonical correlation analysis to flood regionalisation. Presented at *the NATO-ASI on Risks and Reliability in Hydrology and Hydraulics*, Porto Carras, Greece.

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

Rice, R. M. 1972. Using canonical correlation for hydrological predictions. *Bull. Int. Assoc. Hydrol. Sci.*, XVII, 3(10): 315-321.

Rosbjerg, D., and H. Madsen. 1994. Uncertainty measures of regional flood frequency estimators. *J. Hydrol.* 167: 209-224.

Shabbar, A., and A. G. Barnston. 1996. Skill of seasonal climatic forecasts in Canada using canonical correlation analysis, *Mon. Weather Rev.*, 124, 2370-2385.

Snyder, W. M., 1962: Some possibilities for multivariate analysis in hydrologic studies, *J. of Geophysical Research*, 67(2): 721-729.

Stedinger, J. R., and G. Tasker. 1986. Regional hydrologic analysis, 2, Model-error estimators, estimation of sigma and log Peraron type 3 distributions, *Water Resour. Res.* 22(10): 1487-1499.

Thiao, W., and A. G. Barsnston. 1997. CCA forecast for Sahel rainfall in Jul-Aug-Sep 1997, Exp. *Long-Lead Forecast Bull.*, 6, 13-14.

Torranin, P. 1972. Applicability of cononical correlation in hydrology. *Hydrol. Pap. 58.* Colo. State Univ., Fort-Collins, CO, 30 pp.

Wallis, J. R. 1967. When is it safe to extend a prediction eqaution? - An answer based on factor and discriminant function analysis. *Water Resour. Res.* 3(2): 375-384.

Wong, S. T. 1963. A multivariate statistical model for predicting mean annual flood in New England. *Annals of Association of American Geographers* 53(3): 298-311.

Appendix A : Proofs of properties (P1) and (P2)

We provide in this appendix the proofs for the following properties :

(P1) : distinct W_i, as well as distinct V_i, are uncorrelated

(P2) : W_i and V_j, $i \neq j$, are uncorrelated

Proof of (P1) :

Suppose we have two distinct pairs of canonical variables (V_i,W_i) and (V_j,W_j). Then

$$\text{corr}(V_i, W_i) = \lambda_i \neq \lambda_j = \text{corr}(V_j, W_j) \qquad \text{(A1)}$$

We prove herein that V_i and V_j are uncorrelated. The same procedure can be used for W_i and W_j.

By applying equation (3.16) to V_i we have

$$C_{XY}C_{YY}^{-1}C'_{XY}\mathbf{a}_i = \lambda_i^2 C_{XX}\mathbf{a}_i \qquad \text{(A2)}$$

In the same manner, we obtain for V_j

$$C_{XY}C_{YY}^{-1}C'_{XY}\mathbf{a}_j = \lambda_j^2 C_{XX}\mathbf{a}_j \qquad \text{(A3)}$$

Pre-multiplying (A2) by $\mathbf{a}_j'$ we have

$$\mathbf{a}_j' C_{XY}C_{YY}^{-1}C'_{XY}\mathbf{a}_i = \lambda_i^2 \mathbf{a}_j' C_{XX}\mathbf{a}_i \qquad \text{(A4)}$$

and by pre-multiplying (A3) by $\mathbf{a}_i$ and then transposing we obtain

$$\mathbf{a}_j' C_{XY} C_{YY}^{-1} C_{XY}' \mathbf{a}_i = \lambda_j^2 \mathbf{a}_j' C_{XX} \mathbf{a}_i \qquad \text{(A5)}$$

Combining (A4) and (A5) we obtain

$$\lambda_i^2 \mathbf{a}_j' C_{XX} \mathbf{a}_i = \lambda_j^2 \mathbf{a}_j' C_{XX} \mathbf{a}_i \qquad \text{(A6)}$$

Since $\lambda_i \neq \lambda_j$ we must have $\mathbf{a}_j' C_{XX} \mathbf{a}_i = \mathrm{Cov}(V_i, V_j) = 0$. This completes the proof.

Proof of (P2) :

By applying equation (3.15) to W_j we have

$$C_{YY}^{-1} C_{XY}' C_{XX}^{-1} C_{XY} \mathbf{b}_j = \lambda_j^2 \mathbf{b}_j \qquad \text{(A7)}$$

and by applying equation (3.16) to V_i we obtain

$$C_{XX}^{-1} C_{XY} C_{YY}^{-1} C_{XY}' \mathbf{a}_i = \lambda_i^2 \mathbf{a}_i \qquad \text{(A8)}$$

Pre-multiplying (A7) by $\mathbf{a}_i' C_{XY}$ leads to

$$\mathbf{a}_i' C_{XY} C_{YY}^{-1} C_{XY}' C_{XX}^{-1} C_{XY} \mathbf{b}_j = \lambda_j^2 \mathbf{a}_i' C_{XY} \mathbf{b}_j \qquad \text{(A9)}$$

and pre-multiplying (A8) by $\mathbf{b}_j' C'_{XY}$ we obtain

$$\mathbf{b}_j' C'_{XY} C_{XX}^{-1} C_{XY} C_{YY}^{-1} C_{XY} \mathbf{a}_i = \lambda_i^2 \mathbf{b}_j' C'_{XY} \mathbf{a}_i \qquad \text{(A10)}$$

By transposing both sides of (A10) and combining with (A9) we have

$$\lambda_j^2 \mathbf{a}_i' C_{XY} \mathbf{b}_j = \lambda_i^2 \mathbf{a}_i' C_{XY} \mathbf{b}_j \qquad \textbf{(A11)}$$

Since $\lambda_i \neq \lambda_j$ we must have $\mathbf{a}_i' C_{XY} \mathbf{b}_j = \mathrm{Cov}(V_i, W_j) = 0$. This completes the proof.

Appendix B : CCA algorithm for an ungauged basin

Step 1

- identify the target-basin and determine a set B of basins of interest ;
- determine the set of physiographic and meteorologic variables to use ;
- determine also the hydrologic variables to include in the analysis ;

Step 2

- normalise all variables, and then standardize the data (target-basin data is not used) ;
- denote the transformed and standardized physiographic and meteorologic variables X_1, X_2, ..., X_n ; and denote the transformed and standardized hydrologic variables Y_1, Y_2, ..., Y_r ;

- compute the following matrices on the basis of all basins (except the target-basin) : C_{XY}, C_{XX} and C_{YY} (refer to (2) for a description of these matrices if needed) ;

Step 3

- determine all eigenvalues of $C_{YY}^{-1}C'_{XY}C_{XX}^{-1}C_{XY}$. Denote them as $\lambda_1^2, \lambda_2^2, ..., \lambda_p^2$;
- for each eigenvalue λ_i^2 determine its corresponding eigenvector from $C_{YY}^{-1}C'_{XY}C_{XX}^{-1}C_{XY}$ (denoted $\mathbf{b}_i$) ; then determine its corresponding eigenvector from $C_{XX}^{-1}C_{XY}C_{YY}^{-1}C'_{XY}$ (denoted $\mathbf{a}_i$) ;

Step 4

- consider the following $(n \times 1)$ vector $\mathbf{X} = (X_1, X_2, ..., X_n)$. Replace in $\mathbf{X}$ each variable by the corresponding value of the target-basin ; denote the corresponding vector $\mathbf{X_{t\text{-}b}}$;
- determine : $\mathbf{v}_0(i) = \mathbf{X}'_{t\text{-}b} \bullet \mathbf{a}_i$, where $\mathbf{v}_0(i)$ is the i^{th} component of vector $\mathbf{v}_0$;

Step 5

- consider the following $(r \times 1)$ vector $\mathbf{Y} = (Y_1, Y_2, ..., Y_r)$. For each basin B_k of $\underline{B}$, replace in $\mathbf{Y}$ each variable by its corresponding value ; denote the corresponding vector $\mathbf{Y}_k$;
- for each $\mathbf{Y}_k$, determine : $\mathbf{w}_k(i) = \mathbf{Y}'_k \bullet \mathbf{b}_i$, where $\mathbf{w}_k(i)$ is the i^{th} component of vector $\mathbf{w}_k$;

Step 6

- fix a confidence level α and obtain $\chi^2_{\alpha,p}$;
- In turn, replace in (3.19) **w** by $\mathbf{w}_k$ and record the $\mathbf{w}_k$ that satisfy the inequality. The corresponding basins belong to the 100(1-α)% confidence level neighborhood for the target-basin.

Appendix C : CCA algorithm for a gauged basin

Step 1

- identify the target-basin and determine a set B of basins of interest ;
- determine the set of physiographic and meteorologic variables to use ;
- determine also the hydrologic variables to include in the analysis ;

Remark : Information for variables used in the analysis must be available for all basins (including the target-basin).

Step 2

- normalise all variables, and then standardize the data (for all basins including the target-basin) ;

- denote the transformed and standardized physiographic and meteorologic variables X_1, X_2, ..., X_n ; and denote the transformed and standardized hydrologic variables Y_1, Y_2, ..., Y_r ;
- compute the following matrices on the basis of all basins (target-basin included) : C_{XY}, C_{XX} and C_{YY} (refer to (2) for a description of these matrices if needed) ;

Step 3

- determine all eigenvalues of $C_{YY}^{-1}C'_{XY}C_{XX}^{-1}C_{XY}$. Denote them as $\lambda_1^2, \lambda_2^2, ..., \lambda_p^2$;
- for each eigenvalue λ_i^2 determine its corresponding eigenvector from $C_{YY}^{-1}C'_{XY}C_{XX}^{-1}C_{XY}$ (denoted $\mathbf{b}_i$) ;

Step 4

- consider the following $(r \times 1)$ vector $\mathbf{Y} = (Y_1, Y_2, ..., Y_r)$. Replace in $\mathbf{Y}$ each variable by the corresponding value of the target-basin ; denote the corresponding vector $\mathbf{Y_{t\text{-}b}}$;
- determine : $\mathbf{w}_0(i) = \mathbf{Y}'_{t\text{-}b} \bullet \mathbf{b}_i$, where $\mathbf{w}_0(i)$ is the i^{th} component of vector $\mathbf{w}_0$;

Step 5

- consider the following $(r \times 1)$ vector $\mathbf{Y} = (Y_1, Y_2, ..., Y_r)$. For each basin B_k of $\underline{B}$, replace in $\mathbf{Y}$ each variable by its corresponding value ; denote the corresponding vector $\mathbf{Y}_k$;
- for each $\mathbf{Y}_k$, determine : $\mathbf{w}_k(i) = \mathbf{Y}'_k \bullet \mathbf{b}_i$, where $\mathbf{w}_k(i)$ is the i^{th} component of vector $\mathbf{w}_k$;

Step 6

- fix a confidence level α_0 and obtain $\chi^2_{\alpha,p}$ using (3.30) ;

In turn, replace in (3.31) **w** by $\mathbf{w}_k$ and record the $\mathbf{w}_k$ that satisfy the inequality. The corresponding basins belong to the $100(1-\alpha_0)\%$ confidence level neighborhood for the target-basin.

Table 1. Results of the sensitivity to the at-site D/E procedure.

			Components of canonical vectors						
			V_i					W_i	
Distribution	Vector i	Canonical corr.coef.	Area	LCP	PCP	SLM	PTMA	Q2	Q_{100}/Q_2
GEV-PWM	1	0.9592	1.2227	0.1221	0.1693	-0.2614	0.2688	1.0247	0.0765
GEV-ML		0.9590	1.2149	0.1207	0.1770	-0.2601	0.2728	1.0209	0.0709
LN-ML		0.9589	1.2183	0.1216	0.1749	-0.2617	0.2725	1.0169	0.0577
LP III-BOB		0.9594	1.2254	0.1230	0.1671	-0.2608	0.2698	1.0262	0.0800
NP		0.9596	1.2245	0.1215	0.1678	-0.2597	0.2737	1.0251	0.0832
GEV-PWM	2	0.3093	1.9988	0.3050	-1.8421	-0.4856	-0.7615	0.3093	1.0677
GEV-ML		0.3077	2.0443	0.1140	-1.8736	-0.6966	-0.7452	0.2777	1.0556
LN-ML		0.4222	2.2948	0.2691	-1.9121	-0.8032	-0.5946	0.2821	1.0538
LP III-BOB		0.2790	2.2563	0.3610	-2.0447	-0.4584	-0.6438	0.3129	1.0698
NP		0.2941	2.2999	0.4762	-2.0474	-0.3694	-0.5923	0.2798	1.0593

FIGURE CAPTIONS

Figure 1. Approaches for the delineation of homogeneous regions

Figure 2. Location of gauging stations across the province of Ontario (Canada)

Figure 3. Optimal value of α for the 100-year quantile

Figure 4. Optimal value of the mean number of stations in the neighborhood for the 100-year quantile

Figure 5. Sensitivity of the technique to the reduction of hydrometric monitoring networks

Figure 6. Illustration of error vectors

Figure 7. Effect of main channel slope on error vectors

Figure 8. Effect of main channel length on error vectors

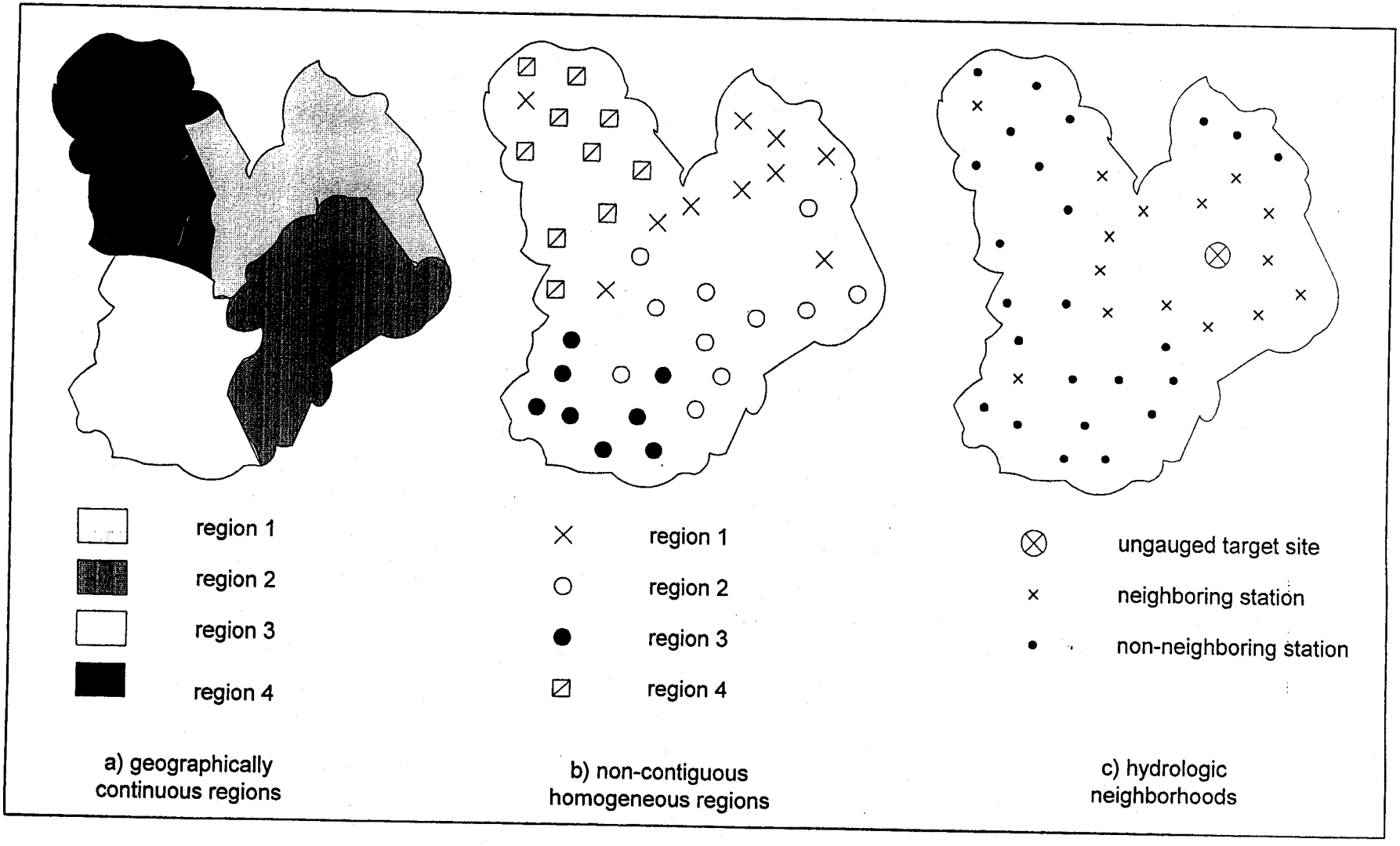

Figure 1. Approaches for the delineation of homogeneous regions

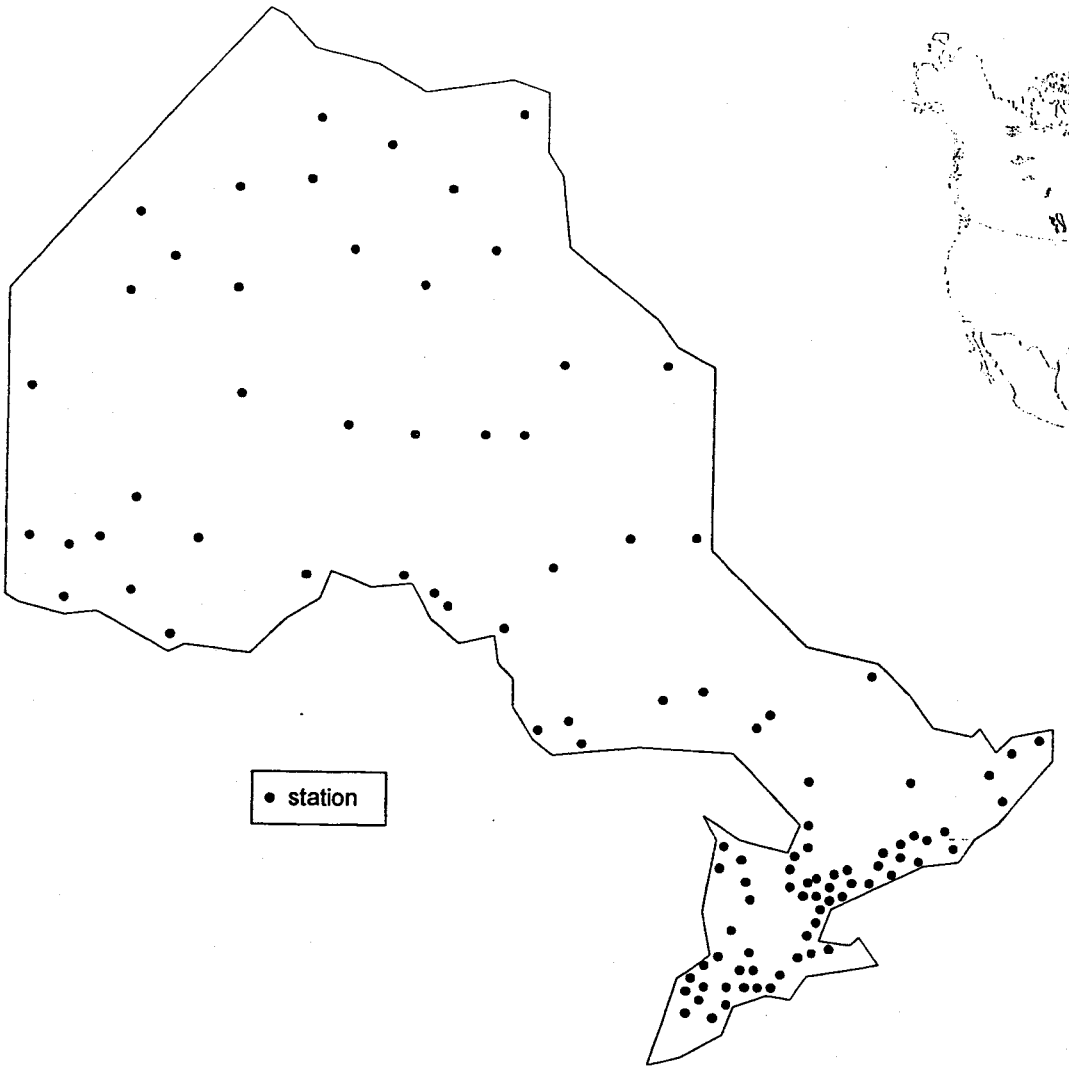

Figure 2. Location of gauging stations across the province of Ontario (Canada)

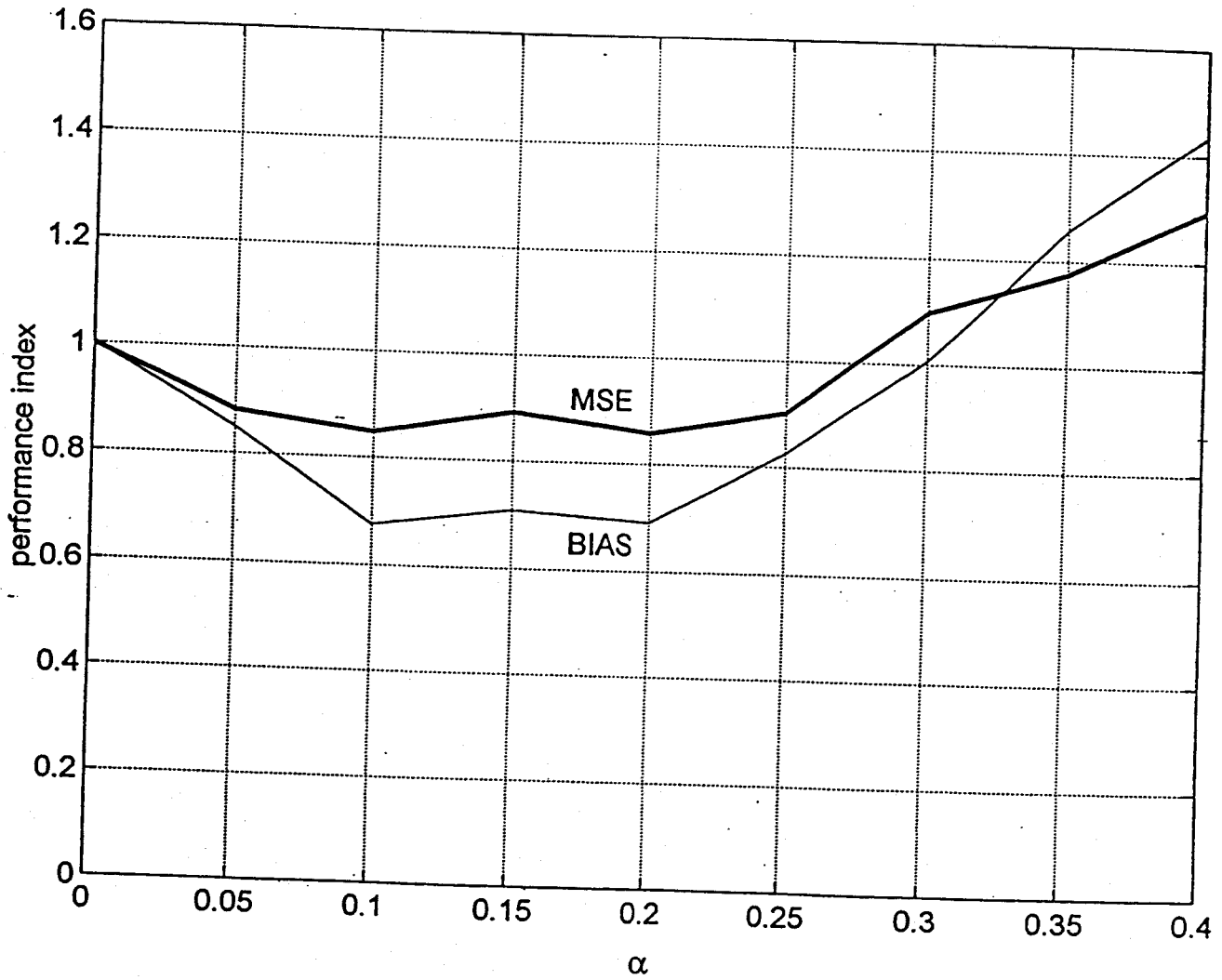

Figure 3. Optimal value of α for the 100-year quantile

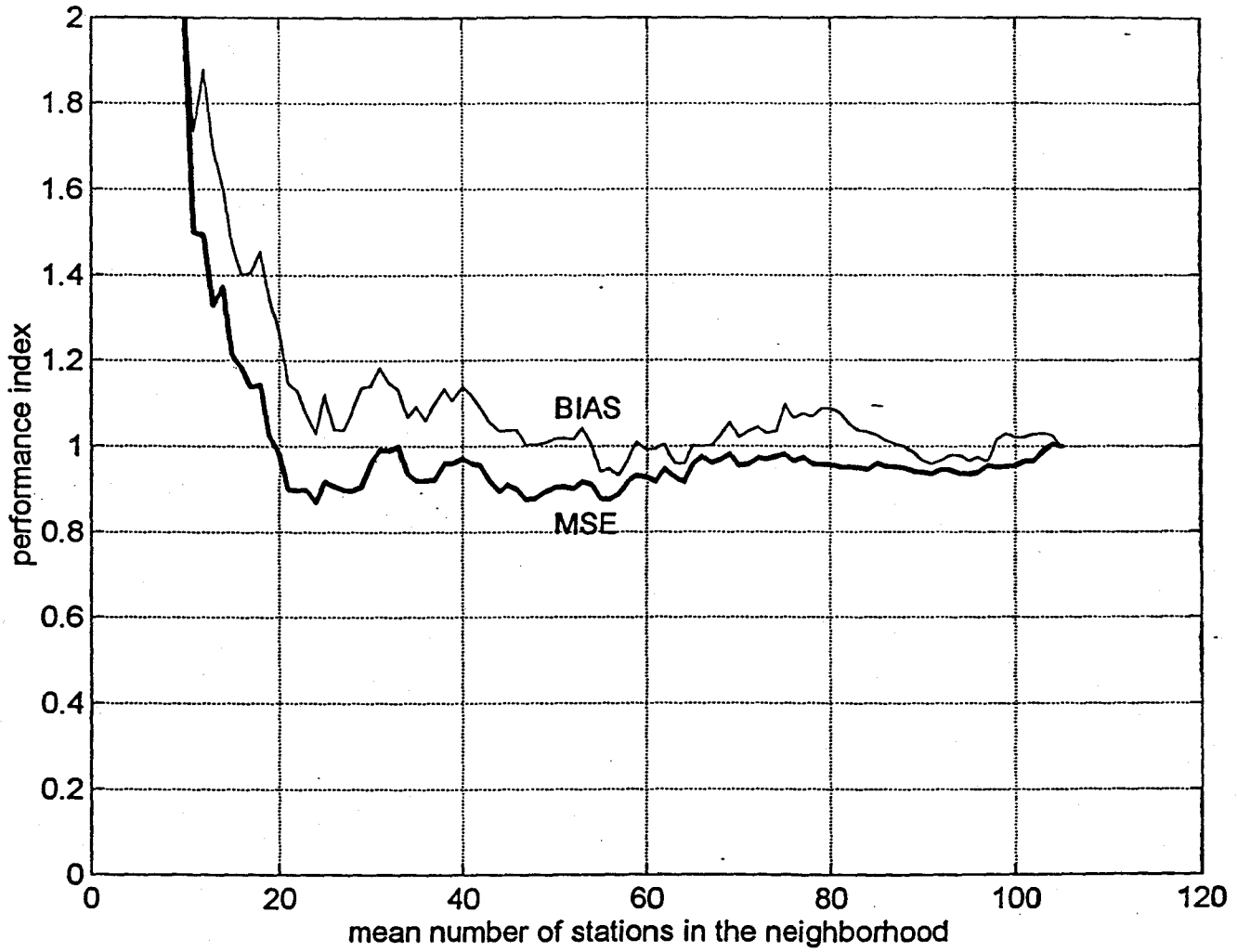

Figure 4. Optimal value of the mean number of stations in the neighborhood for the 100-year quantile

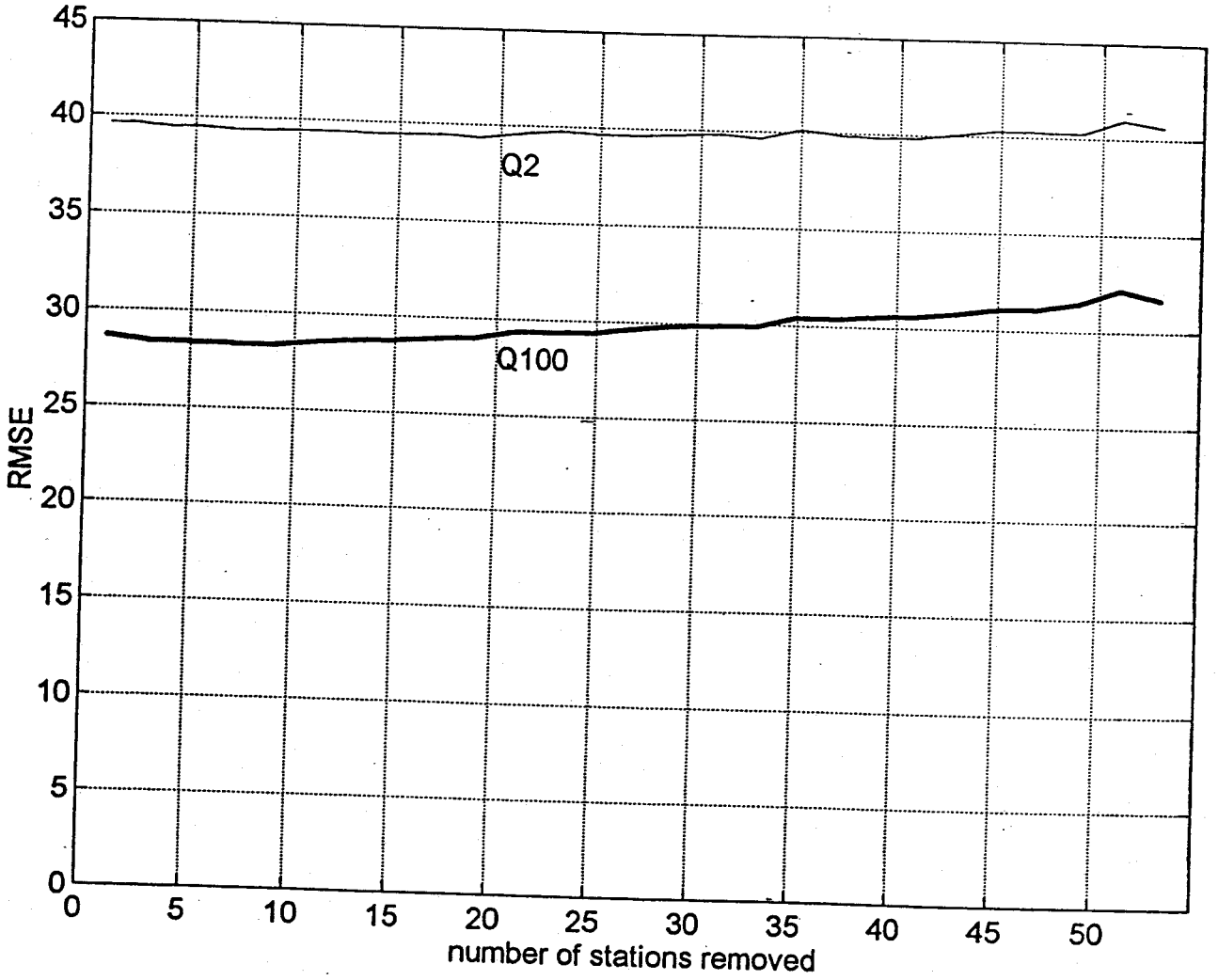

Figure 5. Sensitivity of the technique to the reduction of hydrometric monitoring networks

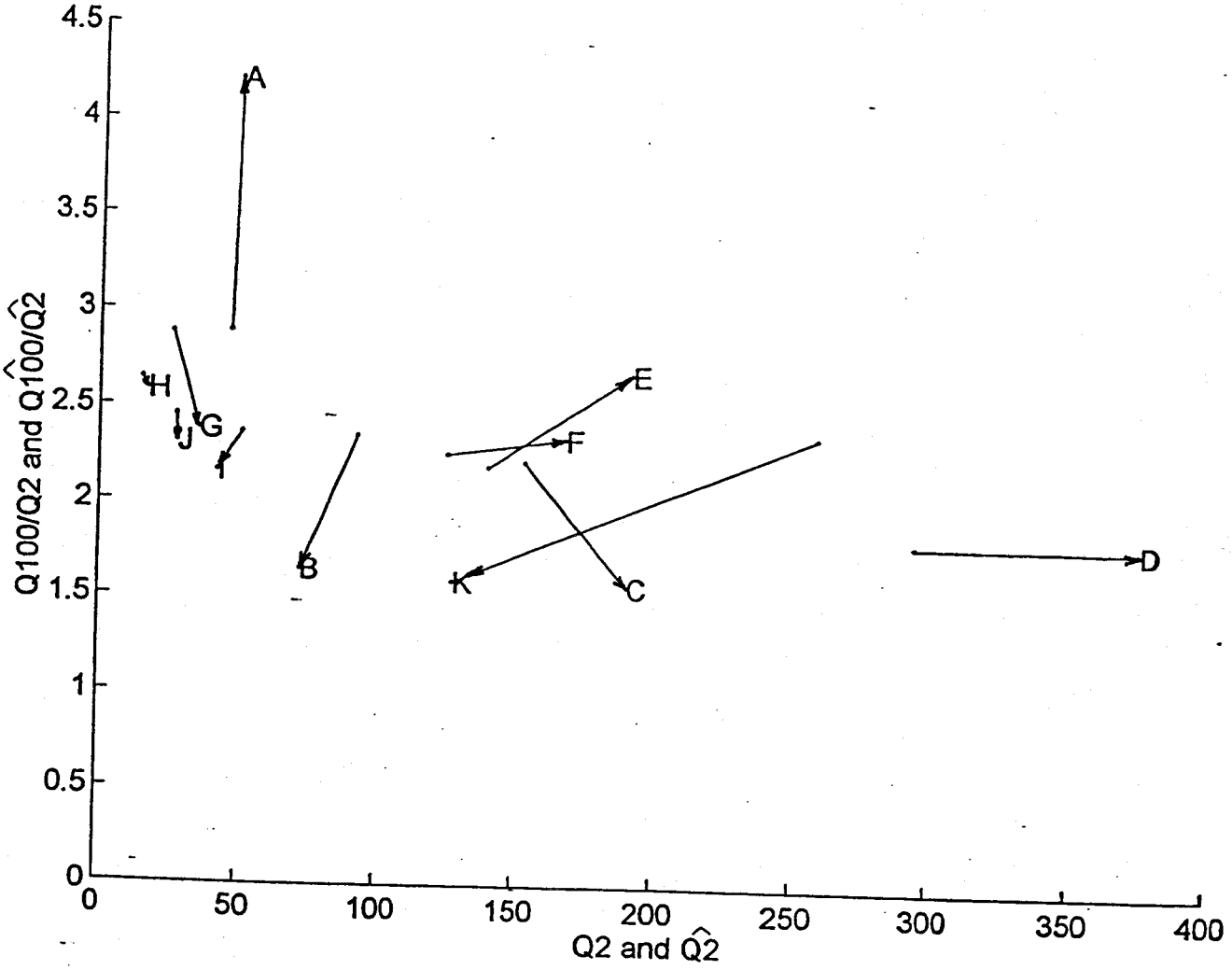

Figure 6. Illustration of error vectors

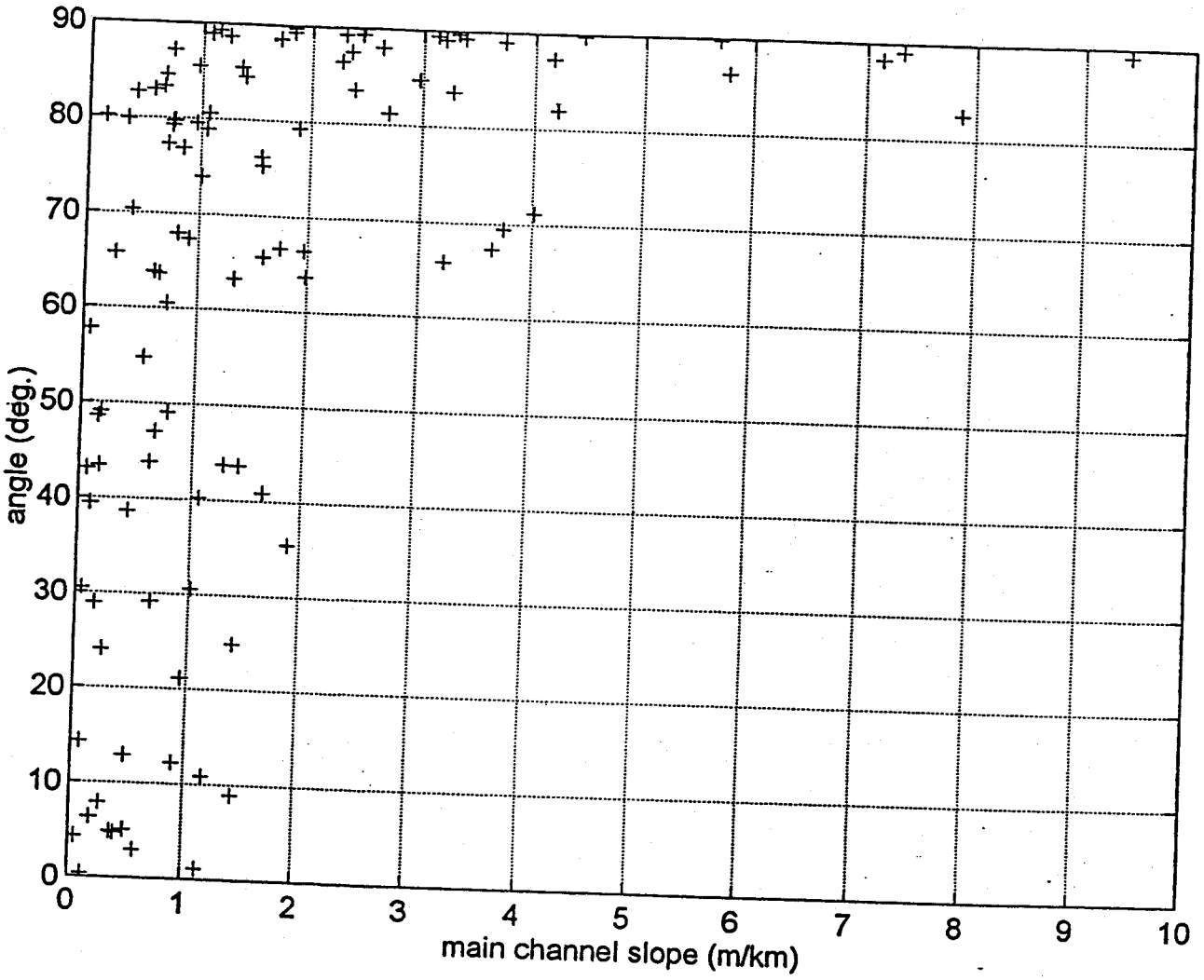

Figure 7. Effect of main channel slope on error vectors

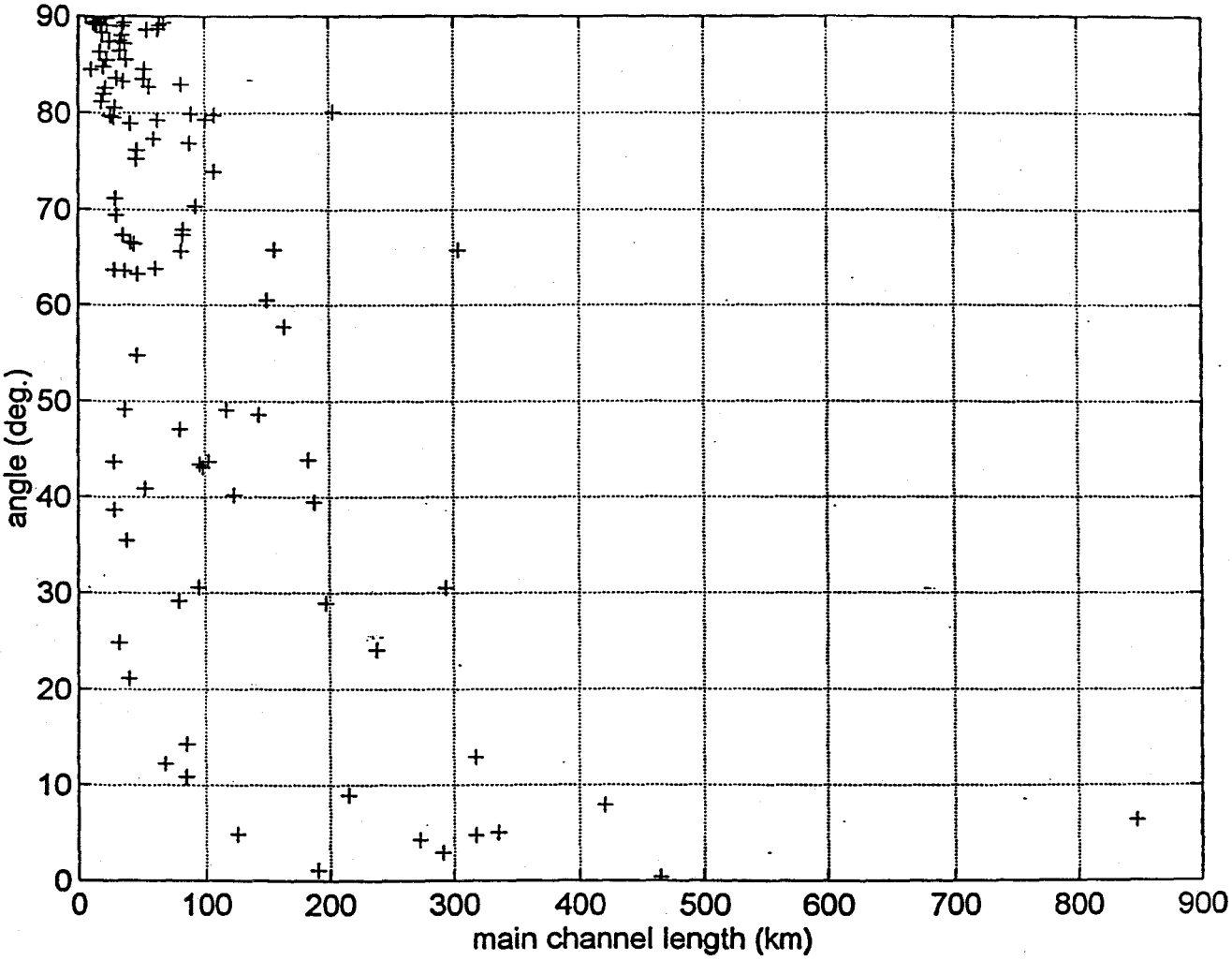

Figure 8. Effect of main channel length on error vectors

4. APPROCHE PAR CLASSIFICATION (RAPPORT 1)

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

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

4.1 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 par 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 Ribeiro-Corréa *et al.* (1995) 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. La détermination de la valeur à utiliser pour le paramètre α est une étape nécessaire pour la mise en œuvre de cette méthode.

Ce 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).

4.2 Définition actuelle d'un voisinage homogène

Nous ne 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.* (2000) 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 deux ensembles. Pour y arriver, l'ACC remplace les ensembles originaux par deux ensembles de variables, dites canoniques, sans perte significative de l'information contenue dans les ensembles originaux. Les variables canoniques, qui forment p paires, 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

$$\mathbf{N}_{2p}(\mathbf{0}, \mathbf{L}) \tag{4.1}$$

où

$$\mathbf{L} = \mathrm{Cov}\begin{pmatrix}\mathbf{W}\\ \mathbf{V}\end{pmatrix} = \begin{pmatrix}\mathbf{I}_p & \mathbf{\Lambda}\\ \mathbf{\Lambda}' & \mathbf{I}_p\end{pmatrix} \tag{4.2}$$

avec

$$\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, ..., \lambda_p) \tag{4.3}$$

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.*(2000) 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 **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 **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 de multinormalité du couple de vecteurs $(\mathbf{W},\mathbf{V})$ donnée par (4.1), il est possible de montrer (Muirhead, 1982) que la densité (conditionnelle) de $\mathbf{W}|\mathbf{V}=\mathbf{v}_0$ est alors

$$N_p\left(\Lambda\mathbf{v}_0, \mathbf{I}_p - \Lambda\Lambda'\right) \qquad (4.4)$$

La densité donnée par (4.4) permet d'identifier dans l'espace canonique, c'est-à-dire l'endroit où les réalisations de **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 (4.4). 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 **V**, dont la densité conjointe est donnée par (4.1). Pour décrire les

réalisations de **W** uniquement, il suffit de considérer la marginale de **W** au sein de (4.1) qui est (voir Muirhead (1982) p.7, par exemple) :

$$N_p(\mathbf{0}, \mathbf{I}) \tag{4.5}$$

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

Puisqu'on ne dispose pas de réalisations de (4.4), on se propose d'identifier parmi les *N* réalisations de (4.5), celles qui pourraient correspondre à des réalisations de la densité conditionnelle donnée par (4.4). 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 (4.4), 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 (4.4) 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 (4.4):

$$Q(\mathbf{W}) = (\mathbf{W} - \mathbf{\Lambda v}_0)'(\mathbf{I}_p - \mathbf{\Lambda\Lambda}')^{-1}(\mathbf{W} - \mathbf{\Lambda v}_0) \tag{4.6}$$

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 (4.4). Mais comment décide-t-on si un point est trop éloigné du centre de la densité (4.4) pour qu'il soit raisonnable de le considérer comme étant une réalisation de (4.4)? 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 (4.7), tronque-t-on la queue de la distribution de Q?

$$\text{Prob}\left(Q(\mathbf{W}) \geq \chi_\alpha^2\right) = 1 - \alpha \qquad (4.7)$$

On voit que fixer le niveau α dans (4.7) revient à déterminer la distance critique au-delà de laquelle un point est considéré trop éloigné de la densité (4.4) 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:

$$(\mathbf{W} - \Lambda \mathbf{v}_0)' \left(\mathbf{I}_p - \Lambda\Lambda'\right)^{-1} (\mathbf{W} - \Lambda \mathbf{v}_0) \leq \chi_\alpha^2 \qquad (4.8)$$

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

4.3 Vers une révision de la définition actuelle de voisinage

Le problème dans l'application de la définition (4.8) 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 définit la distance critique au-delà de laquelle il n'est pas raisonnable d'assimiler un point à une réalisation de (4.4). 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 (4.4) qui joue un rôle

important, mais aussi la densité (4.5) . Rappelons que la densité (4.5) 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 (4.4). Ces scores formeraient alors un voisinage homogène. Or, assimiler des bassins à des réalisations de (4.4), 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 (4.5). 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 (4.4) 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 (4.8) incorpore une mesure de proximité par rapport au centre de la densité (4.4) seulement. Par conséquent, aucun élément de la définition actuelle ne prend en compte la densité (4.5) comme il se doit. On constate *a posteriori* que le rôle du quantile introduit dans (4.8) 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 (4.4) que de (4.5). 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.

4.4 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 (4.5), qui peuvent passer pour être des réalisations de (4.4).

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,*

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq 1 \qquad (4.9)$$

Il est entendu ici que si l'observation n'est pas assignée à la population π_1, à savoir lorsque (4.9) 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 $\mathbf{x}$ à la population qui l'a le plus vraisemblablement produite. On peut facilement remanier l'expression (4.9) de cette règle pour l'exprimer en fonction des distances standardisées Q données par (4.6). En effet, (4.9) est équivalente à

$$\ln\left(f_1(\mathbf{x}) / f_2(\mathbf{x})\right) \geq 0 \tag{4.10}$$

où ln est le logarithme népérien.

Considérons la notation

$$Q_i(\mathbf{x}) = (\mathbf{x} - \boldsymbol{\mu}_i)' \boldsymbol{\Sigma}_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i) \; i=1,2 \tag{4.11}$$

Après quelques simplifications, l'expression (4.10) devient:

$$\ln(|\boldsymbol{\Sigma}_2|) - \ln(|\boldsymbol{\Sigma}_1|) - \frac{Q_1(\mathbf{x})}{2} + \frac{Q_2(\mathbf{x})}{2} \geq 0 \tag{4.12}$$

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

Pour résumer, la règle de Wald-Fisher peut être donnée à partir de (4.12) de la manière suivante:

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

$$Q_1(\mathbf{x}) + \ln\left(\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|}\right)^2 \leq Q_2(\mathbf{x}) \tag{4.13}$$

L'expression (4.13) montre que la règle simple et raisonnable de Wald-Fisher, initialement donnée par la relation (4.9) en termes de densités de probabilités, ne s'écrit pas comme on aurait pu l'anticiper en termes de distances standardisées :

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

$$Q_1(\mathbf{x}) \leq Q_2(\mathbf{x}) \tag{4.14}$$

La règle (4.14) attribue l'observation $\mathbf{x}$ 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 $\mathbf{x}$; 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.

4.4.1 Classification : le cas particulier de densités univariées

Afin d'illustrer l'inadéquation de la règle de classification simpliste donnée en (4.14), 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 (4.14) se simplifie ici pour donner:

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

$$\frac{1}{\sigma_1^2} \leq \frac{1}{\sigma_2^2} \tag{4.15}$$

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 (4.13) sous les mêmes conditions:

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

$$\left(\frac{x - \mu}{\sigma_1}\right)^2 + \ln\left(\frac{\sigma_1^2}{\sigma_2^2}\right) \leq \left(\frac{x - \mu}{\sigma_2}\right)^2 \tag{4.16}$$

Il est aisé de montrer que cette règle équivaut à :

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

$$(x - \mu)^2 \geq \ln\left(\frac{\sigma_1^2}{\sigma_2^2}\right)\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2 - \sigma_2^2} \tag{4.17}$$

En d'autres mots, avec la règle de Wald-Fisher, contrairement à la règle (4.14), 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 (4.14) et la règle de Wald-Fisher donnée par (4.13), 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 (4.8) présente d'importantes lacunes.

Premièrement, la règle de sélection (4.8) ne fait intervenir dans la balance qu'une des deux populations en jeu, à savoir la densité conditionnelle (4.4). Ainsi, d'un côté de la balance il y a la distance de l'observation **W** par rapport à la densité conditionnelle (4.4), mais de l'autre, comme contrepoids, il n'y a rien puisque l'autre densité en jeu, à savoir (4.5), n'est pas considérée. D'une certaine façon la règle simpliste (4.14), 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 (4.5) comme contrepoids à la distance séparant **W** de la densité conditionnelle (4.4), 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 (4.8) ne s'exprime qu'en fonction de la distance standardisée. Or, nous avons vu à la section 4.4.1 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 (4.4) et (4.5) à produire des observations dans une région donnée.

4.5 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 (4.4) 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 (4.4) et par π_2 les scores canoniques obtenus de la densité (4.1). Rappelons que π_2 est en fait la marginale de **W** par rapport à (4.1) et elle est donnée par (4.5).

Puisque nous sommes en présence de deux populations multi-normales, la règle de Wald-Fisher donnée par (4.13) devient alors à partir de (4.6):

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

$$(\mathbf{W}-\Lambda\mathbf{v}_0)'(\mathbf{I}_p-\Lambda\Lambda')^{-1}(\mathbf{W}-\Lambda\mathbf{v}_0)\leq(\mathbf{W}-\mathbf{0})'\mathbf{I}_p^{-1}(\mathbf{W}-\mathbf{0})+\ln\left(\frac{|\mathbf{I}_p|}{|\mathbf{I}_p-\Lambda\Lambda'|}\right)^2 \quad (4.18)$$

En comparant la définition (4.18) à la définition existante (4.8), 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 (4.5) alors que le second prend en compte la dispersion relative des deux densités.

4.6 Conclusion

En apparence nous nous retrouvons donc avec 2 définitions de voisinages homogènes obtenues à partir de l'ACC qui pourraient être vues comme des concurrentes. En réalité, cependant, la définition (4.18) est la version complétée de la définition existante (4.8). En effet, nous avons montré que des deux densités en présence (4.4) et (4.5), seule (4.4) 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 (4.4), alors qu'ils sont tous des réalisations de (4.5), au départ, par hypothèse.

Nous avons montré comment la densité (4.5) 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 des 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.

4.7 Références

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

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, 2000, 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.

5. PROBLÉMATIQUE DU BIAIS DANS UN MODÈLE LOG-LINÉAIRE (RAPPORT 2)

Étude du biais dans la prédiction depuis un modèle log-linéaire

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-575

Novembre 2000

5.1 Introduction

En hydrologie, comme dans plusieurs domaines où la statistique est appliquée, l'utilisation d'un modèle log-linéaire est courante. Dans le cadre plus spécifique de l'estimation régionale des quantiles de crues, l'emploi d'un modèle log-linéaire, liant un quantile de crue d'un bassin à certaines de ses caractéristiques physiographiques, est très répandu. Un exemple d'application de ce modèle dans le cadre de l'estimation régionale est présenté par Ribeiro-Corréa *et al.*(1995) Rappelons que sous sa forme générale le modèle log-linéaire est décrit par :

$$\ln(Y) = \beta_0 + \beta_1 X + \varepsilon \qquad (5.1)$$

où Y est la variable aléatoire d'intérêt, X est la variable (aléatoire) explicative, β_0 et β_1 sont des paramètres du modèle et ε est un terme d'erreur. Nous faisons l'hypothèse que le terme d'erreur, ε, est une variable aléatoire de densité normale $N(0,\sigma^2)$. Lorsque $X=x_0$, le terme d'erreur ε et la variable $\ln(Y)$ sont les seules variables aléatoires dans le modèle (5.1); il en résulte que $\ln(Y)$ admet une densité (conditionnelle à $X=x_0$) normale.

Le modèle log-linéaire est souvent utilisé pour établir une prédiction $\hat{\ln}(Y_0)$ de la valeur de la variable $\ln(Y)$ pour une valeur donnée x_0 de la variable X. À cette fin, la procédure généralement suivie consiste d'abord à obtenir les estimations par moindres carrés $\hat{\beta}_0$ et $\hat{\beta}_1$ des paramètres β_0 et β_1, respectivement, pour établir l'équation de prédiction en fonction de X :

$$\hat{\ln}(Y) = \hat{\beta}_0 + \hat{\beta}_1 X \qquad (5.2)$$

L'équation (5.2) représente le modèle (5.1) ajusté aux données utilisées. Ensuite, en substituant à X dans (5.2) une valeur donnée x_0, une valeur correspondante est obtenue pour $\hat{\ln}(Y_0)$; c'est la valeur prédite par le modèle (5.1) pour la valeur donnée x_0. Puisque Y est

la véritable variable d'intérêt, et non $\ln(Y)$, il est courant d'obtenir de $\hat{\ln}(Y_0)$ une estimation $\hat{Y}_0$ de Y en lui appliquant la transformation exponentielle.

Cette façon de procéder pour obtenir une prédiction $\hat{Y}_0$ pour Y (étant donné $X=x_0$), analogue à la façon de faire pour un modèle linéaire simple, fait l'objet de sérieuses critiques. La principale critique se rapporte au biais qui est introduit en passant d'une valeur prédite $\hat{\ln}(Y_0)$ pour $\ln(Y)$ à la valeur prédite induite $\hat{Y}_0$ en lui appliquant la transformation exponentielle, l'inverse du logarithme. En effet, la valeur prédite $\hat{\ln}(Y_0)$ obtenue par (5.2) est une estimation sans biais pour la moyenne de la distribution conditionnelle de la variable aléatoire ln(Y) étant donnée $X=x_0$. Puisque cette distribution est par hypothèse normale, et par conséquent symétrique, il s'ensuit que la valeur prédite $\hat{\ln}(Y_0)$ est également une estimation sans biais pour la médiane de cette distribution. L'application de la transformation exponentielle à la valeur prédite $\hat{\ln}(Y_0)$ obtenue de (5.2) fournit l'estimation suivante $\hat{Y}_0$ de Y pour la valeur donnée x_0 de X :

$$\hat{Y}_0 = \exp(\hat{\beta}_0) \times \exp(\hat{\beta}_1 x_0) \qquad (5.3)$$

Il est possible de montrer (Miller, 1984) que cette estimation est biaisée pour la moyenne de la distribution de Y pour la valeur donnée x_0 de X, bien qu'elle ne le soit pas pour sa médiane. Puisque c'est la moyenne de la distribution conditionnelle de Y étant donnée $X = x_0$ qui nous intéresse généralement plutôt que sa médiane, l'importance de ce biais doit faire l'objet d'une étude approfondie.

La présente étude a pour but d'investiguer plus à fond le problème de biais dans un modèle log-linéaire et de cerner l'impact que cette problématique a en pratique pour l'utilisateur d'un tel modèle.

5.2 Approches de réduction/correction du biais

Il existe dans la littérature deux grandes approches pour apporter une solution considérée satisfaisante à ce problème de biais. La première approche consiste à substituer à l'estimateur défini par (5.3) un estimateur sans biais. Une possibilité consiste à introduire un facteur multiplicatif dans l'équation (5.3) afin de rendre l'estimation initiale $\hat{Y}_0$ non biaisée pour la moyenne de la distribution conditionnelle de Y étant donné $X = x_0$. Cette méthode de correction du biais est décrite par Neyman et Scott (1960). Cependant, Miller (1984) décrit comme complexes les formulations qui y sont présentées pour obtenir des estimations sans biais pour la moyenne, un avis partagé par Seber et Wild (1989).

La deuxième approche consiste à introduire un facteur multiplicatif dans l'équation (5.3) afin d'en atténuer le biais, et non plus de l'éliminer complètement. L'avantage premier de cette approche réside dans la simplicité du facteur multiplicatif employé. Un premier facteur de correction est présenté en détail par Miller (1984), et mentionné par Duan (1983). C'est d'ailleurs celui que nous nous proposons d'étudier ici puisqu'il est possible d'en établir théoriquement les propriété. Duan (1983) présente également un facteur de correction dans un cadre non-paramétrique, c'est-à-dire qu'aucune densité n'est assignée au terme d'erreur. Puisque nous ne considérons ici que le cas où les erreurs sont normales, nous limitons cette étude au facteur de correction proposé par Duan-Miller. Il faut cependant noter que l'approche non-paramétrique s'avère très intéressante en pratique dans le cas où l'hypothèse de normalité des erreurs ne semble pas être adéquatement remplie. Mentionnons que Hoos (1996) fait exclusivement usage de cette approche.

L'approche de réduction de biais de Duan-Miller consiste à remplacer l'équation (5.3) par l'équation de prédiction suivante, pour une valeur donnée x_0 de X :

$$\hat{Y}_0 = \exp(\hat{\beta}_0) \times \exp(\hat{\beta}_1 x_0) \times \exp\left(\frac{\hat{\sigma}^2}{2}\right) \qquad (5.4)$$

où $\hat{\sigma}^2$ est l'estimation sans biais de la variance du terme d'erreur du modèle, σ^2, induite de l'estimation des paramètres du modèle par la méthode des moindres carrés.

Pour justifier l'introduction de ce facteur, Miller (1984) fait l'observation que la moyenne de la distribution conditionnelle de Y étant donné $X=x_0$ est :

$$E(Y|X = x_0) = \exp(\beta_0) \times \exp(\beta_1 x_0) \times \exp\left(\frac{\sigma^2}{2}\right) \qquad (5.5)$$

La démonstration de cette relation est présentée à l'annexe A. En comparant (5.3) et (5.5), il apparaît effectivement naturel d'introduire le facteur $\exp\left(\frac{\hat{\sigma}^2}{2}\right)$ qui manque en quelque sorte dans (5.3) pour être la contrepartie estimée de (5.5).

Il faut cependant prendre note que l'estimation (5.4), tout comme (5.3), admet un biais pour la moyenne de la distribution conditionnelle de Y étant donné $X=x_0$. Cela est essentiellement dû au fait que chacune des estimations $\exp(\hat{\beta}_0)$, $\exp(\hat{\beta}_1 x_0)$ et $\exp\left(\frac{\hat{\sigma}^2}{2}\right)$ est biaisée. Toutefois, le biais est moins important dans le cas de (5.4) que dans (5.3) Miller (1984).

5.3 Considérations générales sur le biais en estimation

Utiliser en remplacement de (5.3) un estimateur qui présente un biais moindre, voire un biais nul, c'est admettre que la présence de biais est inacceptable dans la prédiction à partir d'un modèle log-linéaire. Cela semble être une attitude naturelle et justifiée en estimation. Mais qu'en est-il vraiment? Puisque la notion de biais est au centre de la présente étude, il convient d'en discuter de manière plus approfondie.

Depuis longtemps en statistique, du moins en statistique fréquentiste (par opposition à la statistique bayésienne), l'absence de biais constitue un principe clé sous-jacent à l'inférence. Il est pratique courante dans une problématique d'estimation d'un paramètre de limiter l'étude aux seuls estimateurs sans biais pour le paramètre en question. Par la suite, s'il faut choisir parmi plusieurs estimateurs sans biais, seulement alors étudie-t-on d'autres propriétés des estimateurs. C'est ainsi, par exemple, qu'on est amené à s'intéresser en statistique-mathématique (fréquentiste) à l'estimateur sans biais à variance minimale. Ce concept est cependant trompeur puisque cet estimateur, bien qu'à variance minimale parmi les estimateurs sans biais, peut en fait présenter une variance très grande lorsque comparée (adéquatement) à la "variance" de tous les autres estimateurs possibles.

Efron (1975) présente une critique très intéressante du principe d'absence de biais en estimation. La critique est essentiellement fondée sur l'approche des fonctions de risque, une ouverture en quelque sorte vers une vision bayésienne de la question. C'est d'ailleurs à l'aide des fonctions de risque que les premières perçées pour comprendre le paradoxe de Stein ont été obtenues Efron et Morris (1977).

Les bayésiens, quant à eux, considèrent tout simplement non pertinent le fait de considérer le biais d'un estimateur puisqu'il est basé sur un concept contradictoire avec le principe de vraisemblance. À ce sujet Berger (1986) est très clair :

"[...] frequentist measures such as error probabilities, bias, coverage probability, and p-values, which involve averages over unobserved x, are not of this form, and can hence not be of basic interest from the conditional viewpoint."

En un mot, les bayésiens n'admettent pas les arguments qui font appel à des observations potentielles, c'est-à-dire à des résultats d'échantillon possibles autres que ceux obtenus. Par exemple, la notion de biais d'un estimateur consiste à faire la moyenne de tous les résultats potentiellement observables. Les bayésiens considèrent que toute l'information pertinente

au niveau de l'échantillon, d'une réalisation, réside dans ce qui est réellement observé et non dans tout ce qui aurait pu être observé mais qui ne l'a pas été.

5.4 Mesure de dispersion en présence d'estimations biaisées

L'estimation (5.4) de Duan-Miller, avec un facteur de correction du biais, présente un biais moins important pour la moyenne de la distribution conditionnelle de Y étant donné X=x que celui présenté par l'estimation directe (5.3). Sur la base de la réduction de biais, Miller (1984) suggère que (5.4) soit préféré à (5.3), un avis repris par Seber et Wild (1989). Nous ne partageons cependant pas cet avis. En effet, il est fort possible, et nous montrerons dans les prochaines sections que c'est effectivement le cas ici, que les mesures prises pour corriger le biais des estimations en accroissent la dispersion. En d'autres termes, ce qui est gagné sur la précision moyenne des estimations est en quelque sorte perdu par l'étalement de ces dernières. Par conséquent, le fait que (5.4) admette un biais moins important que (5.3) ne constitue pas à lui seul, comme nous allons le montrer, un argument solide pour justifier l'emploi du premier au lieu du second pour fournir des estimations de qualité.

En présence de deux estimations biaisées, la mesure traditionnelle de dispersion qu'est la variance n'est pas un outil fiable pour juger de la qualité des estimations. Pour s'en convaincre, considérons une population de densité normale $N(\mu,1)$ de moyenne inconnue et de variance unité. Dans ce contexte considérons l'estimateur qui donne la valeur (arbitraire) 100, disons, comme estimation ponctuelle quelles que soient les données observées de cette population. Cet estimateur est clairement un choix d'estimateur absurde, et pourtant sa variance est nulle car l'estimation donnée est toujours la même valeur, à savoir ici 100.

Une mesure plus adéquate de dispersion en présence de biais est fournie par l'Écart Quadratique Moyen, l'EQM. Formellement, l'EQM d'un estimateur $\hat{\theta}$ pour le paramètre θ est défini par :

$$\mathrm{EQM}(\theta) = \mathrm{E}\left(\hat{\theta} - \theta\right)^2 \tag{5.6}$$

Une autre mesure possible est l'écart moyen absolu:

$$\mathrm{EQM}(\theta) = \mathrm{E}\left|\hat{\theta} - \theta\right| \tag{5.7}$$

Cependant, l'EQM lui est préféré en pratique pour deux raisons: 1) ses propriétés analytiques sont plus aisément établies; 2) l'EQM peut s'écrire facilement en fonction de la variance et du biais des estimations ; on a en effet:

$$\mathrm{EQM}(\hat{\theta}) = \mathrm{Var}(\hat{\theta}) + \left(\mathrm{biais}(\hat{\theta})\right)^2 \tag{5.8}$$

où $\mathrm{Var}(\hat{\theta})$ est la variance de l'estimateur et

$$\mathrm{biais}(\hat{\theta}) = \mathrm{E}(\hat{\theta}) - \theta \tag{5.9}$$

Une forme dérivée de l'équation (5.8), qui nous sera utile par la suite, permet d'exprimer explicitement l'EQM en fonction de la variance et de l'espérance de l'estimateur de la façon suivante :

$$\mathrm{EQM}\left(\hat{\theta}\right) = \mathrm{Var}\left(\hat{\theta}\right) + \mathrm{E}^2\left(\hat{\theta}\right) - 2\theta\mathrm{E}\left(\hat{\theta}\right) + \theta^2 \tag{5.10}$$

Considérons $\hat{\theta}_1$ et $\hat{\theta}_2$, deux estimateurs d'un même paramètre θ de la distribution d'un phénomène aléatoire Y. En utilisant la forme (5.10) pour l'EQM, la différence des EQM, $\Delta\mathrm{EQM} = \mathrm{EQM}\left(\hat{\theta}_1\right) - \mathrm{EQM}\left(\hat{\theta}_2\right)$, pour les estimateurs $\hat{\theta}_1$ et $\hat{\theta}_2$ s'écrit :

$$\begin{aligned}\Delta\mathrm{EQM} &= \mathrm{EQM}\left(\hat{\theta}_1\right) - \mathrm{EQM}\left(\hat{\theta}_2\right) \\ &= \mathrm{Var}\left(\hat{\theta}_1\right) - \mathrm{Var}\left(\hat{\theta}_2\right) + \mathrm{E}^2\left(\hat{\theta}_1\right) - \mathrm{E}^2\left(\hat{\theta}_2\right) + 2\theta\left(\mathrm{E}\left(\hat{\theta}_2\right) - \mathrm{E}\left(\hat{\theta}_1\right)\right)\end{aligned} \tag{5.11}$$

5.5 Expressions théoriques pour l'EQM des estimations biaisées et non biaisées

Convenons de noter par $\hat{\theta}_1$ et $\hat{\theta}_2$ les estimations obtenues, respectivement, avec (équation (5.4)) et sans (équation (5.3)) l'ajout du facteur de correction de Duan-Miller. Pour exploiter la forme (5.11) pour comparer les EQM des estimations (5.3) et (5.4) pour la moyenne de la distribution conditionnelle de Y étant donné $X=x_0$, nous devons disposer de leurs moyennes et de leurs variances, ainsi que d'une expression pour le paramètre à estimer.

À ce propos, il est possible de montrer les identités suivantes (voir Annexe A pour (5.12), Annexe C pour (5.14) et (5.16) et Annexe D pour (5.13) et (5.15)) :

- $$\theta = \mathrm{E}(Y|X=x_0) = \exp\left(\beta_0 + \beta_1 x_0 + \frac{\sigma^2}{2}\right) \tag{5.12}$$

- $$\mathrm{E}(\hat{\theta}_1) = \mathrm{E}(\hat{Y}_M|X=x_0) = \exp\left(\beta_0 + \beta_1 x_0 + \frac{\sigma^2}{2}\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}} + \frac{\sigma^2}{2(n-2)}\right)\right) \tag{5.13}$$

- $$\mathrm{E}(\hat{\theta}_2) = \mathrm{E}(\hat{Y}|X=x_0) = \exp\left(\beta_0 + \beta_1 x_0 + \frac{\sigma^2}{2}\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}}\right)\right) \tag{5.14}$$

- $$\mathrm{Var}(\hat{\theta}_1) = \mathrm{Var}(\hat{Y}_M|X=x_0) = \exp\left(2\beta_0 + 2\beta_1 x_0 + \sigma^2\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}} + \frac{\sigma^2}{2(n-2)}\right)\right) \times \left[\exp\left(\sigma^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}} + \frac{\sigma^2}{2(n-2)}\right)\right) - 1\right] \tag{5.15}$$

- $$\mathrm{Var}(\hat{\theta}_2) = \mathrm{Var}(\hat{Y}|X=x_0) = \exp\left(2\beta_0 + 2\beta_1 x_0 + \sigma^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}}\right)\right) \times$$

$$\left[\exp\left(\sigma^2\left(\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)\right)-1\right] \tag{5.16}$$

où

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i \tag{5.17}$$

et

$$S_{xx}=\sum_{i=1}^{n}(x_i-\bar{x})^2 \tag{5.18}$$

Pour alléger la présentation qui suit, réécrivons les identités (5.12) à (5.16) avec :

- $\beta_0+\beta_1 x_0=A$ (5.19)
- $\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}=B$ (5.20)
- $\frac{\sigma^2}{2(n-2)}=C$ (5.21)

pour obtenir :

- $\theta=E(Y|X=x_0)=\exp\left(A+\frac{\sigma^2}{2}\right)$ (5.22)
- $E(\hat{\theta}_1)=E(\hat{Y}_M|X=x_0)=\exp\left(A+\frac{\sigma^2}{2}(1+B+C)\right)$ (5.23)
- $E(\hat{\theta}_2)=E(\hat{Y}|X=x_0)=\exp\left(A+\frac{\sigma^2}{2}B\right)$ (5.24)
- $Var(\hat{\theta}_1)=Var(\hat{Y}_M|X=x_0)=\exp(2A+\sigma^2(1+B+C))\times[\exp(\sigma^2(B+C))-1]$ (5.25)
- $Var(\hat{\theta}_2)=Var(\hat{Y}|X=x_0)=\exp(2A+\sigma^2B)\times[\exp(\sigma^2B)-1]$ (5.26)

Après quelques manipulations algébriques, l'introduction des quantités (5.22) à (5.26) dans l'équation (5.11) permet de la récrire sous la forme suivante :

$$\Delta EQM = EQM\left(\hat{Y}_M \middle| X = x_0\right) - EQM\left(\hat{Y} \middle| X = x_0\right)$$

$$= \exp(2A) \times \left[\exp\left(\sigma^2(2B + 2C + 1)\right) - \exp\left(2B\sigma^2\right) + 2\exp\left(\frac{\sigma^2}{2}(1 + B)\right) - 2\exp\left(\sigma^2\left(1 + \frac{B + C}{2}\right)\right)\right] \tag{5.27}$$

Ce qu'il importe de connaître au sujet de ΔEQM donné par (5.27) c'est son signe. En effet, si (5.27) est toujours de signe négatif, par exemple, alors nous pourrons en conclure que l'estimateur (5.4) est meilleur que (5.3) en ce sens qu'il donne lieu à des estimations moins dispersées. Puisque le terme exp(2A) de (5.27) est toujours positif, le signe de l'expression (5.27) est le signe que prend l'expression entre crochets dans (5.27) qui dépend des quantités B et C seulement. Le prochain chapitre donne des conditions suffisantes fonction des quantités B et C afin que (5.27) soit positive.

5.6 Comparaison des EQM

Nous avons vu à la fin de la section 5.5 que le signe de l'expression (5.27) dépend seulement des quantités B et C. Or, des expressions (5.20) et (5.21) pour B et C, respectivement, il est clair qu'une fois le modèle ajusté, la quantité B est l'unique quantité dans (5.27) qui peut varier dépendant du x_0 qui est fixé. Pour que l'expression (5.27) soit de signe positif il suffit d'avoir simultanément :

$$\exp\left(\sigma^2(2B + 2C + 1)\right) - 2\exp\left(\sigma^2\left(1 + \frac{B + C}{2}\right)\right) > 0 \tag{5.28}$$

et

$$2\exp\left(\frac{\sigma^2}{2}(1 + B)\right) - \exp\left(2B\sigma^2\right) > 0 \tag{5.29}$$

La condition (5.28) est équivalente à :

$$\exp\left(\sigma^2(2B+2C+1)\right) > \exp\left(\sigma^2\left(1+\frac{B+C}{2}\right)+\ln(2)\right) \tag{5.30}$$

$$\Leftrightarrow 2B+2C+1 > 1+\frac{B+C}{2}+\frac{\ln(2)}{\sigma^2} \tag{5.31}$$

$$\Leftrightarrow B > \frac{2}{3}\frac{\ln(2)}{\sigma^2} - C \tag{5.32}$$

La condition (5.29), quant à elle, est équivalente à :

$$\exp\left(\frac{\sigma^2}{2}(1+B)+\ln(2)\right) > \exp\left(2B\sigma^2\right) \tag{5.33}$$

$$\Leftrightarrow 1+B+\frac{2}{\sigma^2}\ln(2) > 4B \tag{5.34}$$

$$\Leftrightarrow B < \frac{1}{3}+\frac{2}{3}\frac{\ln(2)}{\sigma^2} \tag{5.35}$$

En combinant (5.32) et (5.35), nous avons alors que l'expression (5.27) est de signe positif si la double condition suivante sur B est satisfaite :

$$\frac{2}{3}\frac{\ln(2)}{\sigma^2} - C < B < \frac{1}{3}+\frac{2}{3}\frac{\ln(2)}{\sigma^2} \tag{5.36}$$

ou, de façon équivalente, si la double condition suivante sur x est satisfaite :

$$\frac{2}{3}\frac{\ln(2)}{\sigma^2} - C < \frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}} < \frac{1}{3}+\frac{2}{3}\frac{\ln(2)}{\sigma^2} \tag{5.37}$$

Puisque $C = \frac{\sigma^2}{2(n-2)} > 0$, la condition (5.37) peut toujours être satisfaite, pourvu qu'une valeur appropriée pour x_0 soit choisie. Par conséquent, nous avons montré qu'il existe des situations où l'emploi du facteur de réduction de biais proposé par Miller (1984) peut mener

à des estimations de moindre qualité (au sens de l'EQM) que celles obtenues sans qu'une réduction du biais n'ait été appliquée.

5.7 Discussion et conclusion

Nous avons montré que l'utilisation du facteur de correction de Duan-Miller pour atténuer le biais dans l'estimation de la moyenne de la distribution conditionnelle étant donné X=x ne mène pas nécessairement à des estimations de meilleure qualité au sens de l'EQM. En effet, nous avons montré que pour certaines valeurs de x, données par l'équation (5.37), l'EQM pour l'estimation sans facteur de réduction de biais est inférieur à l'EQM pour les estimations rectifiées.

Miller (1984) fait la remarque que le facteur de réduction de biais $\exp\left(\frac{\hat{\sigma}^2}{2}\right)$ devient de plus en plus important à mesure que la variance de Y s'accroît. Cela semble suggérer que le besoin d'introduire un facteur de réduction de biais est d'autant plus important que la variance de Y est grande, puisqu'alors le biais dans l'estimation est plus considérable. À ce sujet, il est intéressant de noter que le domaine de valeurs de x décrit par (5.37) pour lesquelles l'expression ΔEQM, donnée par (5.27), est positive ne rétrécit pas de façon appréciable à mesure que la variance de Y augmente. Cela signifie que l'utilité d'introduire le facteur de réduction $\exp\left(\frac{\hat{\sigma}^2}{2}\right)$ en raison d'un biais grandissant est annihilée par la forte dispersion dans les estimations que provoque son insertion dans le modèle.

En pratique, il ne semble pas facile de déterminer pour un x donné laquelle des prédictions (5.3) ou (5.4) il est préférable d'utiliser sur la base de (5.27). En effet, les quantités A et C qui apparaissent dans (5.27) dépendent des paramètres β_0, β_1 et σ^2 du modèle considéré et qui sont généralement inconnus. Certes, nous pouvons remplacer A et C en substituant aux paramètres que ces expressions contiennent leurs estimations sans biais

correspondantes. Cependant, cette façon de faire n'apparaît pas très sûre. En effet, nous croyons que les estimations des paramètres β_0, β_1 et σ^2 ne seront pas suffisamment précises pour assurer que le signe de (5.27) établi à partir des estimations soit vraiment le signe que nous aurions obtenu avec les véritables valeurs des paramètres.

Nous croyons avoir montré clairement que la seule considération du biais pour départager deux estimateurs ne donne pas un aperçu fiable de la qualité des estimations à être produites. Même si dans l'expression de l'EQM le biais apparaît au carré, cela ne signifie pas pour autant qu'il soit un facteur prépondérant quant à la qualité des estimations. En fait, la formule suggère plutôt, et c'est ce que nous avons montré dans cette étude, que la qualité des estimations dépend d'au moins deux facteurs: le biais et la variance. Et plus encore, là où le biais parvient à être diminué, la variance s'en trouve accrue.

5.8 Références

Berger, J.O. (1986), *Bayesian Inference and Decision Techniques, edited by P. Goel and A. Zellner,* Elsevier Science Publishers, pp. 473-487.

Casella G., R.L. Berger (1990), Statistical Inference, Duxbury Press.

Duan, N. (1983), Smearing estimate : a non parametric retransformation method, *Journal of the American Statistical Association,* 78 : 605-610.

Dudewicz E.J. (1988), Modern Mathematical Statistics, John Wiley & Sons.

Efron, B. (1975), Biased Versus Unbiased Estimation, *Advances in Mathematics*, 16 :259-277.

Efron, B. et C. Morris, 1977, Le paradoxe de Stein, *Pour La Science,* pp. 28-37.

Hoos, A.B. (1996), Improving regional-model estimates of urban-runoff quality using local data, *Water Resources Bulletin,* 32 : 855-863.

Miller, D. M., 1984, Reducing transformation bias in curve fitting, *The American Statistician*, vol.38, no.2, 124-126.

Neyman, J., E. L. Scott, Correction for bias introduced by a transformation of variables, Ann. Math. Statist., 1960, 31, pp. 643-655.

Seber, G.A.F. et C.J. Wild, 1989, *Nonlinear regression*, John Wiley & Sons

5.9 Annexe A :Moyenne de la distribution conditionnelle de Y étant donné X=x

Cette annexe fournit les détails qui justifient l'équation (5.12).

Rappelons (5.1) qui décrit la forme générale du modèle log-linéaire :

$$\ln(Y) = \beta_0 + \beta_1 X + \varepsilon \qquad \text{(A1)}$$

Par hypothèse, ln(Y) est distribuée normalement et a pour moyenne :

$$E(\ln(Y)) = E(\beta_0 + \beta_1 X) + E(\varepsilon)$$

$$= \beta_0 + \beta_1 X \qquad \text{(A2)}$$

puisque $E(\varepsilon) = 0$. De plus,

$$\text{Var}(\ln(Y)) = \text{Var}(\varepsilon) = \sigma^2 \qquad \text{(A3)}$$

Par conséquent, la moyenne de la distribution conditionnelle de Y étant donné $X=x_0$, $E(Y|X=x_0)$, peut être obtenue de (B4) de l'annexe B (section 5.10) en utilisant (A2) et (A3) :

$$E(Y|X=x_0)=\exp\left(\beta_0+\beta_1 x_0+\frac{\sigma^2}{2}\right) \qquad (A4)$$

5.10 Annexe B : Expressions pour la moyenne et la variance d'une variable log-normale

Si X est une variable aléatoire de densité normale $N(\mu,\sigma^2)$, alors la variable aléatoire $Y = e^X$ est dite log-normale. Nous sommes intéressés à exprimer la moyenne et la variance de Y en fonction de μ et σ^2, respectivement la moyenne et la variance de X.

Il est facile d'établir que la densité de la variable aléatoire Y, $f_Y(y)$ est donnée par :

$$f_Y(y)=\frac{1}{\sqrt{2\pi\sigma}}\frac{1}{y}\exp\left(\frac{-(\ln(y)-\mu)}{2\sigma^2}\right),\ y>0 \qquad (B1)$$

Pour obtenir l'espérance et la variance de Y il est utile de rappeler que la fonction génératrice des moments pour la variable X, $M_X(t)$, est :

$$M_X(t)=E(e^{tx})=\int_{-\infty}^{\infty}e^{tx}\frac{1}{\sqrt{2\pi\sigma}}\exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)dx=\exp\left(\mu t+\frac{\sigma^2}{2}\right) \qquad (B2)$$

En effet,

$$E(Y)=\int_0^{\infty}\frac{1}{\sqrt{2\pi\sigma}}\exp\left(\frac{-(\ln(y)-\mu)^2}{2\sigma^2}\right)dy \qquad (B3)$$

devient, en posant x=ln(y),

$$E(Y) = \int_{-\infty}^{\infty} e^{x} \frac{1}{\sqrt{2\pi\sigma}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right) dx$$
$$= M_X(1)$$

$$= \exp\left(\mu + \frac{\sigma^2}{2}\right) \quad \text{(B4)}$$

Pour établir l'expression pour la variance de Y, nous avons besoin d'établir $E(Y^2)$ puisque

$$Var(Y) = E(Y^2) - E^2(Y) \quad \text{(B5)}$$

Nous avons,

$$E(Y^2) = \int_{0}^{\infty} \frac{y}{\sqrt{2\pi\sigma}} \exp\left(\frac{-(\ln(y)-\mu)^2}{2\sigma^2}\right) dy \quad \text{(B6)}$$

En posant à nouveau x=ln(y), nous obtenons

$$E(Y^2) = \int_{-\infty}^{\infty} \frac{e^{2x}}{\sqrt{2\pi\sigma}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right) dx$$
$$= M_X(2)$$

$$= \exp(2\mu + 2\sigma^2) \quad \text{(B7)}$$

Par (B5) nous obtenons :

$$Var(Y) = e^{2\mu}\left(e^{2\sigma^2} - e^{\sigma^2}\right) \quad \text{(B8)}$$

5.11 Annexe C : Espérance et variance de la prédiction

Cette annexe présente le détail des calculs pour l'obtention des expressions pour $E(\hat{Y})$ et $Var(\hat{Y})$ présentées en (5.14) et (5.16), respectivement.

Rappelons que le modèle log-linéaire ajusté fournit, pour x donné, l'estimation suivante pour la moyenne de la distribution conditionnelle de Y étant donné $X=x_0$:

$$\ln(\hat{Y}) = \hat{\beta}_0 + \hat{\beta}_1 x_0 \qquad \text{(C1)}$$

Puisque $\hat{\beta}_0 = \overline{\ln}(Y) - \hat{\beta}_1 \overline{x}$, où $\overline{\ln}(Y) = \frac{\sum_{i=1}^{n} \ln(y_i)}{n}$, et $\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}}$ sont des estimations sans biais pour β_0 et β_1, respectivement, il s'ensuit que :

$$E(\ln(\hat{Y})) = \beta_0 + \beta_1 x_0 \qquad \text{(C2)}$$

Pour obtenir la variance $\mathrm{Var}(\hat{Y})$ il vaut mieux récrire l'équation (C1) sous la forme équivalente suivante :

$$\ln(\hat{Y}) = \overline{\ln}(Y) + \hat{\beta}_1 (x_0 - \overline{x}) \qquad \text{(C3)}$$

puisque sous cette forme, contrairement à la forme (C1), l'ordonnée à l'origine et la pente, en raison de l'hypothèse de normalité sur le modèle, sont indépendantes. En effet,

$$\begin{aligned} \mathrm{Cov}(\overline{\ln}(Y), \hat{\beta}_1) &= \frac{\sum_i \sum_j \mathrm{Cov}(\ln(y_i), \ln(y_j)(x_j - \overline{x}))}{nS_{XX}} \\ &= \frac{\sum_i \sum_{j \neq i} \mathrm{Cov}(\ln(y_i), \ln(y_j)(x_j - \overline{x})) + \sum_i (x_i - \overline{x})\sigma^2}{nS_{XX}} \\ &= 0 \end{aligned} \qquad \text{(C4)}$$

puisque $\mathrm{Cov}(\ln(y_i), \ln(y_j)) = 0$ (indépendance de $\ln(y_i)$et $\ln(y_j)$pour $i \neq j$) et $\sum_i (x_i - \overline{x}) = 0$.

Par conséquent,

$$\mathrm{Var}(\ln(\hat{Y})) = \mathrm{Var}(\overline{\ln}(Y)) + (x_0 - \overline{x})^2 \mathrm{Var}(\hat{\beta}_1)$$

$$= \frac{1}{n^2} \sum_i \mathrm{Var}(\ln(y_i)) + \frac{(x_0 - \overline{x})^2}{S_{XX}^2} \sum_i (x_i - \overline{x})^2 \mathrm{Var}(\ln(y_i))$$

$$= \frac{\sigma^2}{n} + \frac{(x_0 - \overline{x})^2 \sigma^2}{S_{XX}} \qquad \text{(C5)}$$

Nous avons donc que $\ln(\hat{Y})$ est normalement distribuée dont la moyenne et la variance sont donnés par (C2) et (C5), respectivement. Il s'ensuit que $\hat{Y}$ est de densité log-normale. La moyenne de $\hat{Y}$, $E(\hat{Y})$, est alors, d'après (B4) :

$$E(\hat{Y}) = \exp\left(\beta_0 + \beta_1 x_0 + \frac{\sigma^2}{2}\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}}\right)\right) \qquad \text{(C6)}$$

et la variance de $\hat{Y}$, $\mathrm{Var}(\hat{Y})$, est donnée par (B8) :

$$\mathrm{Var}(\hat{Y}) = \exp\left(2\beta_0 + 2\beta_1 x_0 + \sigma^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}}\right)\right) \times \left[\exp\left(\sigma^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}}\right)\right) - 1\right] \qquad \text{(C7)}$$

5.12 Annexe D : Espérance et variance de la prédiction corrigée pour le biais

Cette annexe présente le détail des calculs pour l'obtention des expressions pour $E(\hat{Y}_M | X = x_0)$ et $\mathrm{Var}(\hat{Y}_M | X = x_0)$ présentées en (5.13) et (5.15), respectivement.

Rappelons que le modèle log-linéaire ajusté, puis corrigé pour la réduction du biais, fournit, pour x donné, l'estimation suivante pour la moyenne de la distribution conditionnelle de Y étant donné $X=x_0$:

$$\ln(\hat{Y}_M | X = x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0 + \exp\left(\frac{\hat{\sigma}^2}{2}\right) \qquad \text{(D1)}$$

Puisque $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$, $\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}}$ et $\hat{\sigma}^2$ sont des estimations sans biais pour β_0, β_1 et σ^2, respectivement, il s'ensuit que :

$$E(\ln(\hat{Y}_M | X = x_0)) = \beta_0 + \beta_1 x_0 + \frac{\sigma^2}{2} \qquad \text{(D2)}$$

Pour obtenir la variance $\mathrm{Var}\left(\hat{Y}_M | X = x_0\right)$ il est préférable de récrire l'équation (D1) sous la forme équivalente suivante :

$$\ln\left(\hat{Y}_M | X = x_0\right) = \overline{\ln}(Y) + \hat{\beta}_1 (x_0 - \bar{x}) + \frac{\hat{\sigma}^2}{2} \qquad \text{(D3)}$$

puisque sous cette forme, contrairement à la forme (D1), l'ordonnée à l'origine et la pente, en raison de l'hypothèse de normalité sur le modèle, sont indépendantes. La démonstration de ce fait est présentée dans l'annexe C après l'équation (C3). De plus, il est possible de montrer que $\hat{\beta}_1$ et $\hat{\sigma}^2$ sont indépendants (Casella et Berger, 1990, p. 569), comme le sont également $\overline{\ln}(Y)$ et $\hat{\sigma}^2$. L'indépendance $\overline{\ln}(Y)$ et $\hat{\sigma}^2$ est un fait bien connu dont une démonstration élémentaire est présentée dans Dudewicz (1988).

L'indépendance mutuelle des termes de (D3) justifie alors :

$$\mathrm{Var}\left(\ln\left(\hat{Y}_M | X = x_0\right)\right) = \mathrm{Var}\left(\overline{\ln}(Y)\right) + (x_0 - \bar{x})^2 \mathrm{Var}\left(\hat{\beta}_1\right) + \frac{\mathrm{Var}\left(\hat{\sigma}^2\right)}{4} \qquad \text{(D4)}$$

$$= \frac{1}{n^2} \sum_i \mathrm{Var}\left(\ln(y_i)\right) + \frac{(x_0 - \bar{x})^2}{S_{XX}^2} \sum_i (x_i - \bar{x})^2 \mathrm{Var}\left(\ln(y_i)\right) + \frac{\mathrm{Var}\left(\hat{\sigma}^2\right)}{4} \qquad \text{(D5)}$$

La variance de l'estimation $\hat{\sigma}^2$, $\mathrm{Var}\left(\hat{\sigma}^2\right)$, s'obtient en établissant la distribution de $\hat{\sigma}^2$. En effet, de Casella et Berger (1990) p.569, nous avons que la variable aléatoire $\frac{(n-2)\hat{\sigma}^2}{\sigma^2}$ est distribuée selon un khi-deux χ^2_{n-2} à (n-2) degrés de liberté. Par conséquent,

$$\mathrm{Var}\left(\hat{\sigma}^2\right) = \mathrm{Var}\left(\frac{\sigma^2}{n-2} \frac{(n-2)\hat{\sigma}^2}{\sigma^2}\right) = \frac{\sigma^4}{(n-2)^2} \mathrm{Var}\left(\chi^2_{n-2}\right) = \frac{\sigma^4 2(n-2)}{(n-2)^2} = \frac{2\sigma^4}{n-2} \qquad \text{(D6)}$$

Puisque $\ln\left(\hat{Y}_M\right)$ est normalement distribuée dont la moyenne et la variance sont donnés par (D2) et (D6), respectivement. Il s'ensuit que $\hat{Y}_M$ est de densité log-normale. La moyenne de $\hat{Y}_M$, $E\left(\hat{Y}_M | X = x_0\right)$, est alors, d'après (A4) :

$$E\left(\hat{\theta}_1\right) = E\left(\hat{Y}_M \middle| X = x_0\right) = \exp\left(\beta_0 + \beta_1 x_0 + \frac{\sigma^2}{2}\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}} + \frac{\sigma^2}{2(n-2)}\right)\right) \quad \text{(D7)}$$

et la variance de $\hat{Y}_M$, $\mathrm{Var}\left(\hat{Y}_M \middle| X = x_0\right)$, est donnée par (A8) :

$$\mathrm{Var}\left(\hat{Y}_{corr}\right) = \exp\left(2\beta_0 + 2\beta_1 x + \sigma^2\left(1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{XX}} + \frac{\sigma^2}{2(n-2)}\right)\right) \times \left[\exp\left(\sigma^2\left(\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{XX}} + \frac{\sigma^2}{2(n-2)}\right)\right) - 1\right]$$

(D8)

6. COMBINAISON DE L'INFORMATION HYDROLOGIQUE DANS UN CADRE BAYÉSIEN

6.1 Introduction

Le but principal de ce chapitre est de proposer une approche bayésienne qui permette de combiner les informations locale et régionale telles qu'elles nous sont disponibles dans le cadre de travail que nous nous sommes donné. Cela signifie en particulier que l'approche bayésienne doit intégrer l'information régionale qui se présente pour nous sous la forme d'un modèle log-linéaire.

La première étape de cette étude a été d'investiguer les approches bayésiennes qui ont été proposées dans des cadres similaires à celui qui nous occupe. Les travaux que nous considérons sont ceux mis de l'avant par Bernier(1993), Kuczera(1982), Fill et Stedinger (1995) ainsi que Rosbjerg et Madsen(1995).

Il est devenu apparent au cours de cette documentation que les principes mêmes de la démarche bayésienne ne semblent pas avoir été bien compris par ses utilisateurs. Si certaines interprétations ou certains développements sont hésitants, d'autres conceptions sont simplement erronées. Les justifications présentées dans les applications courantes font souvent intervenir de façon interchangeable des concepts fréquentiste et bayésien. Or, dans bien des cas le cadre bayésien initié dans les applications ne présente pas le support théorique requis pour y exploiter simultanément une approche fréquentiste! Nous avons donc consacré beaucoup d'énergie et de temps à étudier puis à décrire les fondements d'une démarche bayésienne acceptable afin de contribuer, espérons-le, à rendre les applications éventuelles moins confuses et moins controversées.

En hydrologie statistique, comme dans plusieurs autres domaines d'application de la statistique, on constate que les idées bayésiennes sont peu appliquées. Et lorsqu'on trouve des applications, bien souvent c'est pour réaliser que les préceptes bayésiens n'ont pas été bien appliqués.

En statistique il existe deux grandes familles de pensées: l'école classique, aussi dite fréquentiste, et l'école bayésienne. La quasi-totalité des programmes universitaires en statistique et des ouvrages didactiques sont consacrés aux idées classiques; c'est le constat de la situation que fait Lindley (1990) :

« There are very few universities in the world with statistics departments that provide a good course on the subject. Only exceptional graduate students leave the field of their advisor and read for themselves. »

Cependant, les idées bayésiennes persistent et ces dernières années avec l'avènement du calcul numérique les idées bayésiennes gagnent en popularité. Il demeure cependant que les ouvrages didactiques qui abordent la statistique du point de vue bayésien ne font qu'émerger. À ce chapitre le livre élémentaire et entièrement voué aux idées bayésiennes de Berry (1996) est une rare exception car les quelques ouvrages bayésiens sont de niveau mathématique très relevé (e.g. Berger (1982)). Par conséquent, l'initiation aux idées bayésiennes pour un éventuel utilisateur doit se faire à l'heure actuelle via la littérature de recherche ou à travers des livres d'un niveau mathématique relativement avancé. Il importe donc de donner ici une introduction à la démarche bayésienne qui soit aussi accessible, transparente et complète que possible.

6.2 Démarches fréquentiste et bayésienne à l'inférence

Un problème statistique type qui survient en hydrologie, comme ailleurs, est de déterminer la valeur d'un paramètre d'une population sur la base d'information (partielle) contenue dans un échantillon de données; c'est le problème de l'inférence statistique. Un exemple important en hydrologie, et c'est celui qui nous occupe dans cette étude, est le quantile de crue de période de retour de T ans, Q_T (par exemple le débit centenaire Q_{100}) d'un bassin donné pour lequel nous cherchons à établir une estimation à partir des données hydrologiques disponibles au site. Cette information se présente souvent sous la forme d'une série de maxima annuels observés au site. En statistique il y a deux approches à un problème de cette nature: l'approche fréquentiste et l'approche bayésienne.

6.2.1 Approche fréquentiste

L'inférence à partir de l'approche fréquentiste exige de l'analyste qu'il décrive le lien qui existe entre les données et la quantité d'intérêt à l'aide d'un modèle. Sous l'hypothèse que les maxima annuels observés $x_1, x_2, ..., x_n$ pour n années sont autant de réalisations d'une densité log-normale $\mathrm{LN}(\mu, \sigma^2)$[1] de paramètres μ et σ^2, un premier modèle s'écrit:

$$\ln(x_1), \ln(x_2), ..., \ln(x_n) \approx N(\mu, \sigma^2) \tag{6.1}$$

Le modèle choisi ne lie pas directement les données observées au paramètre d'intérêt Q_{100}, c'est pourquoi il faut un deuxième modèle qui exprime le quantile en fonction des paramètres du premier modèle :

$$Q_{100} = \exp(\mu + k\sigma) \tag{6.2}$$

avec k tel que $P(N(\mu, \sigma^2) > k) = 1/100$ ($k \cong 2.326$).

[1] Contrairement à la loi Normale, les paramètres μ et σ^2 de la loi Log-Normale ne représentent pas sa moyenne et sa variance, respectivement.

L'analyste fréquentiste obtient d'abord des estimations ponctuelles $\hat{\mu}$ et $\hat{\sigma}^2$ pour μ et σ^2, respectivement, par méthode du maximum de vraisemblance. La méthode du maximum de vraisemblance est préférée ici à la méthode des moments en raison de sa propriété d'invariance[1] qui permet d'obtenir directement une estimation de Q_{100} en substituant aux paramètres de (6.2) leurs estimations par maximum de vraisemblance

$$\hat{Q}_{100} = \exp\left(\hat{\mu} + k\sqrt{\hat{\sigma}^2}\right) \tag{6.3}$$

Sous une approche fréquentiste, la qualité de l'estimation obtenue pour Q_{100} est fonction des propriétés que possède l'estimateur. Or, les propriétés théoriques d'un estimateur sont établies en considérant le comportement des estimations sous des répétitions hypothétiques de la même expérience. Cela signifie que la qualité de l'estimation n'est pas fonction de l'estimation obtenue en pratique, mais des estimations obtenues en considérant tous les résultats possibles de l'expérience. En particulier, cela implique que la mesure de qualité fréquentiste est établie *a priori* aux données, c'est-à-dire avant que les données aient été obtenues. Une façon équivalente d'exprimer cela est de dire que l'approche fréquentiste prend en compte tout l'univers réalisable comme ensemble pertinent d'observations et non l'univers (plus restreint) de ce qui s'est réalisé en réalité. C'est cette façon de mesurer la qualité d'une estimation faite (et non tant la façon dont l'estimation a été obtenue) qui distingue l'approche fréquentiste de l'approche bayésienne. Certes, la mesure de qualité pourra influer sur le choix de la méthode d'estimation, mais il demeure qu'avant tout c'est la manière avec laquelle on établit la qualité d'une estimation qui est à la base des différences entre les deux approches.

Un exemple de propriété fréquentiste est l'absence de biais d'un estimateur. On rappelle qu'un estimateur est sans biais si l'espérance des résultats qu'il peut donner a pour valeur le

[1] Si $\hat{\theta}$ est l'estimateur de maximum de vraisemblance pour θ, alors $g(\hat{\theta})$ est l'estimateur correspondant par la méthode de maximum de vraisemblance de la quantité $g(\theta)$.

paramètre d'intérêt. Cette propriété, comme tout autre propriété fréquentiste, s'établit en considérant non pas seulement ce qui a été observé (i.e. ce qui a généré l'estimation obtenue) mais aussi ce qui est observable et qui n'a pas été observé. Toute propriété fréquentiste décrit comment l'estimateur répond « en moyenne » quand on considère toutes les éventualités possibles.

6.2.2 Approche bayésienne

L'inférence à partir de la démarche bayésienne pour un paramètre d'intérêt se décrit mieux par étapes successives:

1- faire un sommaire, avant que l'échantillon ne soit obtenu ou pris en compte dans l'analyse, de ce qui est connu sur le paramètre d'intérêt. L'analyse bayésienne typique amène l'analyste à formuler cette information sous la forme d'une densité de probabilités associée au paramètre, et ce avant que les données de l'échantillon ne soient obtenues; c'est la densité de probabilités du paramètre *a priori* (aux données). La détermination de la densité *a priori* fait souvent appel à l'expérience qu'a de problèmes du même genre ou à toute information indépendante des données observées.

C'est un premier contraste avec l'approche fréquentiste car ici on considère de l'information qui ne provient pas exclusivement des données réalisables, réalisées ou non. En effet, dans un cadre bayésien les données ne sont qu'une source partielle d'information, parmi d'autres, qui se rapportent au paramètre d'intérêt.

2- Décrire l'information qui est contenue dans les données observées en ce qui a trait au paramètre d'intérêt. Cette information s'exprime aussi sous forme d'une densité de probabilité appelée fonction de vraisemblance; elle est fonction du paramètre d'intérêt inconnu.

La fonction de vraisemblance joue un rôle clé dans l'approche bayésienne puisqu'elle est censée receler toute l'information que contient l'espace échantionnal (i.e. l'ensemble des réalisations qui pourraient être observées) sur le paramètre d'intérêt. Et puisque la fonction de vraisemblance ne dépend que des données observées, cela revient à dire que pour les bayésiens les données réalisables, mais qui ne se sont pas réalisées, sont d'aucune importance dans l'analyse statistique. Cela contraste avec le point de vue fréquentiste où, comme on l'a vu, les données réalisables servent à déterminer la qualité de l'estimation à obtenir.

3- À l'aide du théorème de Bayes sous quelque forme, on combine ces deux densités, ces deux sources d'information sur le paramètre d'intérêt dont la valeur est inconnue, pour obtenir ce qui est appelé la densité de probabilités *a posteriori* (aux données) pour le paramètre.

On voit donc que la démarche bayésienne est vraiment une démarche de combinaison de diverses sources d'information sur un paramètre d'intérêt. La densité *a posteriori* est perçue comme étant la description la plus complète de l'information disponible pour faire l'inférence sur un paramètre d'intérêt. La densité *a posteriori* nous informe sur ce qui est le plus crédible de voir comme valeur du paramètre d'intérêt sur la base de l'information recueillie auprès des données et auprès de sources externes. Dans notre contexte, l'estimation a posteriori n'est alors rien d'autre que la valeur (ou plutôt la distribution) la plus crédible pour le paramètre.

Si on reprend l'exemple précédent qui se rapporte au quantile de crue mais cette fois sous un angle bayésien, on obtient de façon générique:

1- Exprimer l'information régionale sur le quantile Q_{100} pour le site d'intérêt sous la forme d'une densité de probabilités, la densité *a priori*;

2- L'information que recèlent les données obtenues au site, par exemple sous la forme d'une série de maxima annuels de crue, au sujet du quantile Q_{100} est exprimée par la fonction de vraisemblance;
3- À partir du théorème de Bayes on obtient la densité *a posteriori*, basée sur la densité a priori et la vraisemblance déterminées à l'étape 1 et 2, respectivement, qui combine donc l'information locale et l'information régionale disponible à propos du quantile Q_{100}.

La détermination de la densité *a priori* se fait en pratique de deux façons, qui feront de l'analyse subséquente à ce choix une analyse bayésienne hiérarchique ou bayésienne empirique. La prochaine section décrit les deux types d'analyses bayésiennes possibles.

6.3 Analyses bayésiennes hiérarchique et empirique

Toute démarche bayésienne à l'inférence peut se réduire au schéma donné à la section précédente. Le schéma donné n'est cependant pas assez précis pour décrire les applications qui surviennent en pratique. En effet, deux raffinements possibles de ce schéma doivent être faits pour correspondre à la pratique. Ainsi, selon la façon dont on caractérise complètement la densité *a priori* on obtient soit une analyse hiérarchique ou une analyse empirique.

Le schéma de la section 6.2.2 montre qu'une analyse bayésienne consiste d'abord à exprimer à l'aide d'une densité a priori l'état des connaissances sur les paramètres impliqués dans l'étude menée avant que les données n'aient été obtenues. Une partie importante de l'étude bayésienne consiste donc à déterminer complètement la densité *a priori* appropriée à utiliser. Cela signifie qu'il faut déterminer à la fois la forme paramétrique qu'elle prendra et les valeurs des paramètres impliqués.

Idéalement, l'information *a priori* permet de caractériser complètement la densité *a priori*. Mais dans plusieurs cas, l'information disponible ne permet d'identifier que la forme paramétrique de la densité *a priori* appropriée mais pas les paramètres eux-mêmes. Ainsi, dans un contexte donné l'information disponible peut permettre de supposer qu'une densité normale est appropriée. L'information disponible, cependant, peut ne pas permettre de se prononcer sur les valeurs à utiliser pour la moyenne et la variance de la densité normale. On peut alors modéliser l'information dont on dispose sur les paramètres inconnus à l'aide d'une autre densité de probabilités dont les paramètres, les hyper-paramètres, sont eux connus. On se retrouve alors avec deux niveaux de modélisation de l'information. En effet, la densité *a priori* elle-même constitue un premier niveau de modélisation, alors que la densité de probabilités qui représente l'information sur ses paramètres, qui elle dépend d'hyper-paramètres, constitue un deuxième niveau. Il n'y a pas de raison conceptuelle pour se limiter à deux niveaux; il est possible de représenter l'information disponible sur les hyper-paramètres dont les valeurs à utiliser sont inconnues avec une densité de probabilités et ainsi constituer un troisième niveau, et ainsi de suite (e.g. Perreault, 2000). En pratique, cependant, on utilise rarement plus de 2 niveaux. L'important est qu'au niveau le plus élevé le modèle utilisé est entièrement déterminé, i.e. les hyper-paramètres de ce modèle sont connus. Une telle modélisation par niveaux où les paramètres du modèle du plus haut niveau sont connus est une analyse bayésienne hiérarchique.

Dans divers contextes l'analyse hiérarchique n'est que d'intérêt théorique puisque les paramètres ne peuvent pas être entièrement déterminés, quel que soit le nombre de niveaux utilisés. En effet, il arrive souvent en pratique qu'on sache que l'information *a priori* correspond bien à une densité de type normale, par exemple, sans pour autant que les paramètres (et éventuels hyper-paramètres) à utiliser puissent être déterminés. Dans de tels cas l'analyse hiérarchique ne peut pas être utilisée en pratique. L'analyse empirique a été développée pour permettre de suivre un raisonnement inférentiel aussi proche de l'esprit bayésien que possible dans des situations pratiques où l'approche hiérarchique est inapplicable. Tout comme l'analyse hiérarchique, on suppose que l'information issue des

données sur le paramètre d'intérêt est modélisée par la fonction de vraisemblance et que l'information *a priori* sur le paramètre est représentée par la densité *a priori*. Dans le cas de l'analyse empirique, la modélisation de l'information *a priori* peut aussi être faite en utilisant plusieurs niveaux, bien qu'en pratique un seul niveau soit employé. La modélisation a ceci de particulier cependant: les paramètres de la *densité a priori* sont inconnus. En effet, l'information disponible n'a pas pu être utilisée, pour une raison ou une autre, pour déterminer la valeur des paramètres impliqués. L'analyse empirique consiste à utiliser les données pour obtenir une estimation de la valeur des paramètres *a priori* à utiliser. Cette estimation obtenue, on procède par la suite comme avec l'analyse hiérarchique comme si la valeur des paramètres de la densité *a priori* avait été connue dès le départ. Cela sous-entend qu'on néglige de considérer l'erreur d'estimation qui entre dans la détermination de la valeur des paramètres à utiliser.

Par conséquent, dans l'analyse empirique les données fournissent de l'information au modèle à la fois par le biais de la fonction de vraisemblance que par le biais de la densité *a priori*. Il s'ensuit que la modélisation bayésienne empirique de l'information est une approche intégrée: les modèles de vraisemblance et *a priori* ne peuvent pas être perçus comme indépendants. Seule une démarche intégrée d'estimation des paramètres de la densité *a priori* à partir des données peut s'avérer adéquate dans ce contexte.

Une approche intégrée d'estimation des paramètres de la densité *a priori* à partir des données est fournie en établissant au départ la densité marginale des données. Dans un modèle empirique, l'information obtenue des données est éparpillée dans la fonction de vraisemblance et dans la densité *a priori*; une mise en commun de cette information sous une seule densité se fait en établissant la densité marginale. Concrètement, si x dénote les données, θ le paramètre d'intérêt, $f(x|\theta)$ dénote la fonction de vraisemblance et $\pi(\theta)$ la densité *a priori*, alors la densité marginale des données $m(x)$ est

$$m(x) = \int_{\theta \in \Theta} f(x|\theta)\pi(\theta)d\theta \qquad (6.4)$$

La densité marginale décrit en une densité la totalité de l'information fournie au modèle global par les données. Une fois obtenue, l'estimation des paramètres de la densité marginale, qui est à la fois fonction des paramètres de la fonction de vraisemblance et de la densité *a priori*, est effectuée à partir des données par la méthode du maximum de vraisemblance. En d'autres mots, on cherche à déterminer les valeurs des paramètres les plus susceptibles d'expliquer les données qui ont été observées.

Dans un contexte bayésien hiérarchique, où la densité *a priori* est entièrement déterminée, la marginale est utilisée par certains statisticiens pour valider le modèle obtenu. En effet, ils regardent si la densité marginale "prédit" raisonnablement bien les données actuellement observées. En d'autres termes, si à la lumière de la densité marginale les observations obtenues paraissent extrêmes, cela peut amener l'analyste à réviser la modélisation employée pour représenter l'information disponible à la recherche d'erreurs qui auraient pu se glisser.

L'estimation des paramètres de la densité *a priori* qui se fait d'une façon qui n'est pas équivalente à l'estimation par le biais de la densité marginale est dite *ad hoc*. Une estimation *ad hoc* typique dans une approche empirique consiste à estimer les paramètres de la densité *a priori* à partir directement des données observées.

Il n'est pas aisé de mesurer l'impact sur l'analyse finale de recourir à une estimation *ad hoc* des paramètres de la densité *a priori* puisque la plupart de ces démarches sont très complexes. En fait, dans la plupart des applications le recours à la densité marginale n'est qu'implicite. Par conséquent, il n'apparaît pas toujours clairement si une analyse donnée repose, ou non, sur une estimation *ad hoc* des paramètres de la densité *a priori*. C'est le cas par exemple de l'analyse proposée par Fill et Stedinger (1995) qui paraît être à la frontière d'une estimation *ad hoc* des paramètres impliqués. Il est cependant possible d'illustrer les problèmes éventuels qu'on peut rencontrer en ayant recours à l'estimation *ad hoc* en

considérant un cas plus simple où l'impact sur l'estimation de la variance *a priori* est notable.

Exemple (Surestimation de la variance *a priori*)

Dans cet exemple illustratif, on considère p observations $x_i, i = 1,..., p$ tirées indépendamment de $N(\theta, \sigma^2)$ avec σ^2 connue. L'information a priori sur le paramètre d'intérêt θ est modélisée par une autre densité normale:

$$\theta \approx N(\mu_\pi, \sigma_\pi^2) \tag{6.5}$$

Une analyse *ad hoc* typique serait ici d'obtenir des estimations pour μ_π et σ_π^2 sur la base des données en utilisant les estimateurs sans biais que sont $\bar{x} = \sum_{i=1}^{p} x_i$ et $s^2 = \frac{\sum_{i=1}^{p} (x_i - \bar{x})^2}{p-1}$ pour la moyenne et la variance d'une densité normale, respectivement.

Une analyse empirique bayésienne adéquate passe par l'obtention au préalable de la densité marginale $m(x)$ qui est ici Berger (1982) p. 128:

$$m(x) \approx N(\mu_\pi, \sigma_\pi^2 + \sigma^2) \tag{6.6}$$

Dans un contexte bayésien recourir à des estimateurs parce qu'ils sont sans biais, comme c'est le cas pour l'analyse *ad hoc* ici, n'est pas un critère de choix d'estimateur valable; nous reviendrons sur ce point à la section suivante. Une approche bayésienne favorise plutôt les estimateurs respectifs suivants obtenus par maximum de vraisemblance pour μ_π et σ_π^2 :

$$\bar{x} \tag{6.7}$$

$$\hat{\sigma}_{\pi}^{2} = \max\left\{0, \frac{p-1}{p} s^{2} - \sigma^{2}\right\} \tag{6.8}$$

Dans les estimations par maximum de vraisemblance, seules les données observées sont considérées dans le processus d'estimation. Il est intéressant de noter ici que la démarche bayésienne acceptable mène à un estimateur de variance qui ne serait pas optimal du point de vue fréquentiste puisque (légèrement) biaisé pour la variance à estimer, σ_{π}^{2}.

On constate que l'analyse *ad hoc* ne reconnaît pas que sous le modèle global deux sources de variabilité expliquent la variabilité qu'on observe dans les données. En effet, il y a la variabilité expliquée par la fonction de vraisemblance et aussi la variabilité expliquée par la densité *a priori*. L'analyse *ad hoc* ne considère que la seconde source de variabilité. Cela a pour effet d'associer toute la variabilité observée dans les données à la variabilité du modèle qui décrit l'information *a priori*. Il en résulte donc une surestimation de la variabilité mesurée dans les données qui peut être attribuée à la variabilité contenue dans le modèle *a priori*. Cette surestimation sera d'autant plus grande que la variabilité dans la fonction de vraisemblance est grande. L'analyse bayésienne empirique décrite ici ne présente pas ce problème puisque la marginale fait intervenir simultanément les deux sources de variabilité qui expliquent la variabilité des données. Reconnaissant les deux composantes de variabilité, l'analyse empirique attribue comme variabilité due au modèle *a priori* la variabilité totale au sein des données moins la variabilité explicable par le modèle de vraisemblance.

Les conséquences de la surestimation de la variance dans ce contexte, contrairement à d'autres situations d'inférence, ne peuvent pas être contrebalancés par l'argument que l'inférence résultante est bonne, bien que trop conservatrice. La surestimation de la variance qu'amène ici l'analyse *ad hoc* donne lieu à une représentation de *l'information a priori* moins précise qu'elle l'est autrement. En d'autres mots, une surestimation importante signifie que l'information *a priori* disponible n'est pas utilisée à sa juste mesure.

En effet, la représentation qu'on en donne alors par densité de probabilités est si floue (i.e. variance grande) qu'elle correspond à l'extrême à une situation où aucune information *a priori* n'est disponible. En somme, on retombe alors sur une approche purement fréquentiste, puisque cette dernière, par définition, ne considère aucune information *a priori*. Avec l'analyse *ad hoc* l'information a priori disponible est sous-utilisée.

6.4 Considérations sur les mesures fréquentistes de performance dans un cadre bayésien

Plusieurs applications bayésiennes sont en fait un mélange de considérations bayésiennes et fréquentistes. Comme on l'a vu à la section précédente, c'est souvent le cas des analyses bayésiennes empiriques développées à partir d'estimations *ad hoc,* basée sur les données, des paramètres de la densité *a priori*. Cette courte section discute de la viabilité d'un tel mélange de concepts dans un cadre bayésien.

Puisqu'une démarche bayésienne ne prend pas en compte l'information sur les données observables non observées, les propriétés fréquentistes comme l'absence de biais, par exemple, n'ont ni d'attrait ni de sens dans un tel contexte. Un estimateur sans biais, sur la seule base de son absence de biais, n'est pas considéré meilleur par le bayésien que n'importe quel autre estimateur. Berger (1986) est très explicite à ce sujet (des précisions sur le contenu de la citation ont été apportées par l'auteur et apparaissent au fur et à mesure entre crochets):

"A statistician seeking to follow the Likelihood Principle [i.e. a statistician with a bayesian view] would only trust procedures or measures that depend on the experiment soley through the observed likelihood function. Classical maximum likelihood estimates are of this form (as are all Bayesian procedures and measures based on the posterior distribution), but frequentist measures such as error probabilities [i.e. Type I and II errors

in hypothesis testing], bias, coverage probability [confidence intervals], and p-values, which involve averages over unobserved x, are not of this form, and can hence not be of basic interest from the conditional viewpoint [such as a bayesian one]."

En pratique, il s'avère souvent que l'estimateur préconisé par le fréquentiste et le bayésien pour un même contexte est le même, bien que les arguments en faveur du choix de cet estimateur différent entre eux. La raison pour laquelle il est difficile d'admettre des propriétés fréquentistes dans un cadre bayésien est que ce dernier ne s'appuie pas sur la totalité de l'espace échantillonal (i.e. ce qui est observé et ce qui aurait pu l'être) mais seulement sur ce qui est observé. L'approche fréquentiste prend appui sur la possibilité de répétition de l'expérience qui a engendré les données observées, ce qui implique le recours à ce qui aurait pu être observé (i.e. au reste de l'espace échantillonal) pour établir des propriétés telles que l'absence de biais. Or, dans un cadre bayésien, la seule portion de l'espace échantillonal considérée est celle qui se rapporte strictement aux données observées. Il n'y a donc pas dans un cadre bayésien l'appui nécessaire pour considérer des propriétés fréquentistes telle que le biais.

Un autre exemple qui illustre bien la différence entre les deux approches en ce qui concerne l'inférence est donné par l'estimateur de James-Stein. L'optimalité de cet estimateur s'observe avec des mesures fréquentistes bien que d'un point de vue fréquentiste lui-même l'estimateur n'ait pas de sens. Cet estimateur est demeuré un paradoxe jusqu'au développement des idées bayésiennes où l'estimateur de James-Stein prend un sens.

Une description accessible de cet estimateur est donnée par Efron et Morris (1977). Pour nos besoins, il suffit de savoir que l'estimateur de James-Stein se présente dans un contexte d'estimation simultanée de plusieurs moyennes. L'exemple illustratif fournit par Efron et Morris (1977) porte sur les moyennes au bâton avec lesquelles termineront les joueurs de baseball d'une ligue professionnelle pour une année donnée. Les présences au bâton pour chaque joueur du début jusqu'à la mi-saison constituent les données observées. D'un point

de vue fréquentiste, le choix de l'estimateur pour les moyennes est clair: la moyenne de frappe de fin de saison d'un joueur s'estime par la moyenne correspondante de ce joueur sur la base des données observées, à savoir sur ses présences au bâton pour la demi-saison. La justification est intuitive: pour chaque moyenne prise séparément il est connu que la moyenne arithmétique est le choix (fréquentiste) optimal. Par conséquent, le vecteur de moyennes arithmétiques doit être optimal pour le vecteur de moyennes inconnues.

Pourtant ce n'est pas le cas. Il s'avère que l'estimateur de James-Stein, qui dans l'estimation d'une moyenne donnée prend en compte aussi les autres moyennes, s'avère être plus performant selon des critères fréquentistes. Ce résultat, perçu comme un vrai paradoxe, ne peut pas s'expliquer dans un cadre fréquentiste: comment se peut-il que considérer d'autres populations, à savoir ici la performance au bâton des autres joueurs, que celle pour laquelle on cherche à obtenir une estimation de la moyenne puisse améliorer l'estimation de leurs moyennes?

Dans un cadre bayésien, le paradoxe de James-Stein n'en est pas un. L'estimateur de James-Stein est équivalent à l'estimateur *a posteriori* obtenu en considérant une certaine information *a priori* sur les moyennes des populations en plus de l'information au site. En d'autres mots, l'estimation de James-Stein fait appel implicitement à un modèle "régional" qui lie les populations entre elles et en retire l'information dégagée de cette mise en commun. S'il n'y a pas d'information supplémentaire à puiser en considérant simultanément les populations, la densité *a priori* sera tout simplement plate (on dit aussi non informative) et l'estimateur de James-Stein coïncidera alors avec le vecteur des moyennes individuelles. Mais si le modèle *a priori* postulé décrit de l'information utile que recèle la mise en commun de populations, alors l'estimation *a posteriori* obtenue sera plus performante que l'estimation par moyennes individuelles. Elle est plus performante car elle fait intervenir dans l'estimation du paramètre d'une population plus que l'information qui est contenue dans les données en rapport strictement avec cette population. La différence importante à noter ici est la possibilité de tenir compte par l'approche bayésienne

d'information utile qui n'a rien à voir avec l'information que peuvent livrer les données observées elles-mêmes.

6.5 Considérations sur certaines fausses conceptions à propos des idées bayésiennes

La description de l'approche bayésienne que nous avons donnée met l'emphase sur la recherche et la combinaison de diverses sources d'information à l'égard d'un paramètre d'intérêt. Comme on l'a vu, c'est le recours à de l'information externe à celle que peut livrer les données qui différencie l'approche bayésienne de l'approche fréquentiste. Cette distinction pourtant fondamentale entre les deux approches ne semble pas avoir été bien comprise par les utilisateurs et il paraît important de dissiper la confusion qui est nettement perceptible dans les applications bayésiennes faites jusqu'à présent.

L'explication erronée qui est la plus fréquemment donnée pour expliquer en quoi l'approche bayésienne diffère de l'approche fréquentiste consiste à affirmer que dans l'optique bayésienne le paramètre d'intérêt est considéré aléatoire. Cette explication décrit la façon par laquelle les fréquentistes se font superficiellement un sens de l'approche bayésienne. Mais cette explication, tant du point de vue bayésien que fréquentiste, n'a pas de sens quand on s'y attarde davantage. Et c'est en bonne partie cette rationalisation boiteuse des fréquentistes qui explique pourquoi les idées bayésiennes sont si peu diffusées et pourquoi d'éventuels utilisateurs des idées bayésiennes se font hésitants. Lindley (1990), un fervent bayésien, résume bien l'état des connaissances qu'ont les statisticiens en général sur l'approche bayésienne:

« What most statisticians have is a parody of the Bayesian argument, a simplistic view that just adds a woolly prior to the sampling theory paraphernalia. They look at the parody, see how absurd it is, and thus dismiss the coherent approach as well. »

Dans la littérature, le recours à l'explication du paramètre qui doit être considéré aléatoire pour que le cadre de l'analyse soit bayésien est très courant. Un exemple tiré de la littérature récente en hydrologie est fourni par Rosbjerg et Madsen(1995):

"*In a Bayesian framework parameters are treated as stochastic variables, to account for the imperfect knowledge of their exact values.*"

Même les livres d'introduction à la statistique basent la différence entre les vues bayésienne et fréquentiste sur cette "observation"; Chou(1989) est un bon exemple:

"*The classical [frequentist] conception is that of a fixed but unknown constant [...]. The Bayesian conception is that a population parameter is itself a random variable!*"

Même si certains auteurs bayésiens adoptent cette façon de décrire l'approche bayésienne, Berger (1982) par exemple (probablement pour s'attirer plus facilement un public fréquentiste au départ), la réplique bayésienne à une telle façon de présenter l'approche bayésienne est pourtant claire à ce sujet avec Jaynes (1986):

"*For decades Bayesians have been accused of 'supposing that an unknown parameter is a random variable'; and we have denied hundreds of times, with increasing vehemence, that we are making any such assumption. We have been unable to comprehend why our denials have no effect, and that charge continues to be made.*"

Qu'une analyse donnée s'inscrive dans un cadre bayésien ou non n'a rien à voir avec la nature du paramètre. En réalité, les bayésiens le considèrent tout aussi fixé, i.e. non-aléatoire, et inconnu que les fréquentistes. Mais qu'est alors la densité (*a priori*) rattachée au paramètre si ce n'est que d'admettre qu'il est de nature aléatoire? La confusion provient du fait que la distribution ne porte pas sur le paramètre mais sur les probabilités par

lesquelles on représente l'information qu'on peut avoir sur la valeur (fixée) inconnue du paramètre. Cette conception erronée des bases bayésiennes provient de l'abus de langage "distribution du paramètre" que font souvent les fréquentistes pour décrire ce qui est en réalité la "distribution de probabilités pour le paramètre" comme il se doit. Que les probabilités reflètent des fréquences (approche fréquentiste) ou traduise l'information dont on dispose (approche bayésienne), il demeure que ce sont les probabilités qui sont distribuées et non le paramètre lui-même! Cet abus de langage, bien en vogue, a mené à l'interprétation que la densité *a priori* du paramètre démontrait que le paramètre est considéré aléatoire, puisque seule une quantité aléatoire peut être "distribuée".

6.6 Une approche hiérarchique à deux niveaux

Le modèle mis de l'avant par Bernier (1993) repose sur une fonction de vraisemblance qui lie le paramètre d'intérêt aux données, qui entrent dans le modèle sous la forme d'une estimation locale du paramètre d'intérêt, par le biais de la densité normale:

$$\hat{Q}_{100} \approx N\left(Q_{100}, \sigma^2\right) \tag{6.9}$$

La densité *a priori* choisie pour représenter l'information disponible sur le paramètre d'intérêt est aussi une densité normale:

$$Q_{100} \approx N\left(\alpha_0 + \alpha_1 Q_{100}^R, \lambda^2 V^2\right) \tag{6.10}$$

où Q_{100}^R est une estimation régionale du quantile Q_{100}, V^2 est la variabilité régionale et $\alpha_0, \alpha_1, \lambda^2$ sont des facteurs d'échelle utilisés pour traduire les estimations régionales en estimations locales pour le site considéré.

La densité *a priori* fait intervenir des hyper-paramètres dont la valeur est estimée à partir des données des autres bassins de la région homogène considérée et d'une estimation régionale obtenue du modèle régional sous-jacent. Le modèle décrit par Bernier (1993) est

donc un modèle hiérarchique à deux niveaux. Bien que les hyper-paramètres doivent être estimés, il ne s'agit pas pour autant d'un modèle empirique. En effet, un modèle bayésien est empirique si l'estimation des hyper-paramètres se fait à partir des données utilisées dans la fonction de vraisemblance. Dans le cas de Bernier (1993), la fonction de vraisemblance s'exprime en terme des données hydrologiques cueillies au site alors que l'estimation des hyper-paramètres fait appel aux données hydrologiques des autres bassins de la région homogène.

La difficulté importante avec cette approche est de parvenir à calculer la variance régionale V^2. Les efforts pour la déterminer dans le cas de l'approche par voisinages homogènes obtenus de l'application de l'ACC sont restés infructueux. On remarque que le modèle décrit par Bernier (1993) ne combine pas l'information brute, mais qu'on a plutôt choisi de résumer chaque portion d'information sous la forme d'une estimation. Par exemple, les maxima annuels n'interviennent pas directement dans la fonction de vraisemblance, mais apparaissent sous la forme d'un résumé d'information que constitue l'estimation locale obtenue en déterminant le quantile de la distribution ajustée aux données (e.g. log-normale, GEV, etc.).

6.7 Approche basée sur le concept de superpopulation

Kuzcera (1982) a développé un cadre théorique qui repose sur le concept de paramètre aléatoire. Pour rendre légitime ce concept qu'on sait maintenant être erroné, beaucoup d'efforts sont investis au début de l'article pour fournir un cadre auquel le rattacher; c'est la superpopulation:

"*According to the Empirical Bayes model, there exists a superpopulation [...] from which nature randomly and independently samples to assign a site a particular set of parameters*".

Les problèmes que génère cette description sont réels et sont particulièrement mis en évidence dans le cadre développé par la suite par Kuczera (1982). En effet, on trouve dans le développement du cadre bayésien de Kuczera (1982) que les paramètres de la superpopulation, les hyperparamètres, dépendent des sites. En d'autres mots, chaque site possède ses valeurs d'hyperparamètres ce qui n'a pas vraiment de sens. En d'autres mots, la densité de la superpopulation n'a pas les mêmes hyper-paramètres, disons (α,β) pour employer une notation proche de celle de Kuczera (1982), pour générer toutes les valeurs du paramètre *a priori*, mais possède un ensemble de paramètres (α_j, β_j)pour chaque site *j*. L'équation (20) de Kuczera (1982) est particulièrement explicite à ce niveau. L'ambiguïté dans la notation employée rend difficile l'interprétation claire du développement avancé dans Kuczera(1982).

De plus, le développement proposé par Kuczera (1982) repose fortement sur l'hypothèse que l'écart-type régional *s* est décrit par une densité gamma-inverse et sur le "corollaire" que la densité correspondante pour la variabilité régionale s^2 est alors gamma. Ce résultat est faux: par définition c'est 1/s, et non s^2, qui admet pour densité une gamma si *s* est gamma-inverse.

La difficulté d'un développement qui repose sur le concept de superpopulation est d'en donner une représentation équivalente mais mieux justifiée, soit sous la forme d'une approche empirique ou hiérarchique, selon le cas. Cela est possible quand le concept de superpopulation ne sert qu'à expliquer des étapes qui sont autrement justifiées. En d'autres mots, l'approche peut être défendable mais pas pour les raisons qui en sont données. L'analyse proposée par Rosjberg et Madsen (1995) est un bon exemple. De plus, comme leur analyse suit étroitement le développement initié par Kuczera (1982) tout en dépendant moins du concept de superpopulation, on se propose d'étudier leur développement plutôt que celui de Kuczera.(1982).

Le développement théorique proposé par Rosbjerg et Madsen (1995) peut être décrit par étapes successives:

1) l'hypothèse est faite que le logarithme naturel des crues annuelles est représenté par une densité normale de moyenne μ_j et de variance s_j^2, pour les bassins indexés par j=1, 2,...,k.
2) la quantité d'intérêt est le quantile de période de retour 100 ans au site *j*, $Q_{100}^{(j)}$:

$$Q_{100}^{(j)} = \exp\left(\mu_j + 2.326 s_j\right) \tag{6.11}$$

3) Les paramètres d'intérêts sont donc la moyenne μ_j et la variance s_j^2 pour chacun des sites *j*.
4) Puisque les estimations ponctuelles des moyennes sont considérées plus stables que celles de s_j, il est proposé d'utiliser simplement les estimations pour μ_j obtenues par maximum de vraisemblance:

$$\overline{q}_j = \left[\sum_{i=1}^{n_j} q_{ij}\right] n_{ij}^{-1} \tag{6.12}$$

Cela revient à dire qu'on choisit de ne pas poser de modèle a priori pour décrire la variabilité régionale dans les moyennes. On n'a recours à l'information régionale que pour Ks_j. Dans une approche bayésienne, tous les paramètres d'intérêt doivent avoir une densité *a priori*, non-informative au besoin s'il y a absence d'information. En ne considérant de l'information *a priori* que pour l'un des deux paramètres, le cadre est ni fréquentiste ni bayésien! La façon bayésienne d'exprimer le fait qu'on ne possède pas d'information *a priori* sur un paramètre est de considérer une loi *a priori* non-informative pour ce paramètre. Évidemment, dans le présent contexte, le développement s'en trouverait compliqué parce qu'il faudrait trouver une loi *a priori* "conjointe" qui marginalement pour *m* est non-informative et que marginalement pour *s* soit gamma-inverse.

Deux modèles régionaux sont ensuite proposés pour exprimer l'information régionale disponible au niveau de la variabilité régionale. Un premier modèle est essentiellement la méthode de l'indice de crue où la variabilité aux sites est la même pour les bassins d'une région, à des fluctuations aléatoires près.

$$\text{variabilité hydrologique des bassins } j = \gamma_0 + \varepsilon \tag{6.13}$$

Le second modèle inclut le précédent; il décrit la variabilité hydrologique comme différente d'un bassin à l'autre, fonction qu'elle est des caractéristiques physiographiques de chacun des bassins:

$$\text{variabilité hydrologique du bassin } j = \gamma_0 + \sum \gamma_i X_{ij} + \varepsilon \tag{6.14}$$

Le modèle physiographique est sans doute plus réaliste que le modèle d'indice de crue; il comporte cependant le même problème potentiel de multicollinéarité qui est décrit à la section qui traite de l'approche mise de l'avant par Fill et Stedinger (1998).

Cette façon de représenter l'information régionale n'est cependant pas compatible avec le cadre qu'on s'est donné. En effet, dans notre cas l'information régionale s'exprime naturellement en terme des quantiles de crue pour chacun des sites et non en terme de variabilité hydrologique au site.

6.8 Approche régionale par quantiles normalisés

Fill et Stedinger (1998) rapportent que la dépendance du modèle sur la variabilité des crues logarithmiques de l'approche de Kuczera (1982) et Rojsberg et Madsen (1995) n'est pas robuste. Ils suggèrent de recourir aux quantiles normalisés, c'est-à-dire les quantiles de crues divisés par la moyenne des crues annuelles du site. La difficulté principale de cette approche est l'accroissement en complexité de l'analyse bayésienne subséquente.

Fill et Stedinger (1998) utilisent des quantiles normalisées pour former leur modèle régional; ils soutiennent que ce modèle est plus robuste que celui qui s'exprime en termes des quantiles. Le modèle régional est une régression multiple qui exprime le logarithme des crues annuelles standardisées comme fonction de caractéristiques physiographiques:

$$\ln\left(\frac{Q_{100}}{\mu}\right) = \mathbf{A}\beta + \varepsilon \qquad (6.15)$$

où μ est la moyenne des crues annuelles observées au site d'intérêt.

En pratique le modèle régional suggéré peut poser des problèmes. En effet, la régression multiple s'exprime en fonction de caractéristiques physiographiques qui, habituellement, sont fortement corrélées entre elles. Sous ces conditions de corrélation entre les variables, un problème de multicollinéarité peut survenir. On peut trouver dans Montgomery et Peck (1992) une discussion complète du problème, des diagnostics disponibles et des solutions possibles.

Pour les besoins de cette étude, cependant, il suffit de savoir que la multicollinéarité survient lorsque au moins deux des variables utilisées comme variables indépendantes d'un modèle de régression multiples sont fortement corrélées. Et dans le cas qui nous occupe, la variable « aire du bassin versant » est souvent fortement corrélée avec d'autres variables de taille utilisées comme, par exemple, la variable « superficie du bassin versant contrôlée par les lacs » (utilisée par exemple dans GREHYS, 1996b). Lorsqu'il y a une forte corrélation entre 2 variables indépendantes (ou plus), les colonnes correspondantes dans la matrice de design, notée X, et par conséquent X'X, seront pratiquement égales à un facteur multiplicatif près, de sorte que le déterminant de X'X sera près de 0. Et la façon non-linéaire avec laquelle le déterminant approche zéro à mesure que la colinéarité présente s'accentue fait en sorte que, la valeur que prend successivement le déterminant d'une réalisation à une autre varie considérablement. Et puisque les estimations des paramètres

de la régression sont fonction de ce déterminant, il en résulte des estimations très variables (on dit souvent qu'elles sont instables) d'une réalisation à une autre. Cela signifie que pour une réalisation donnée, les mesures fréquentistes de performance telles que les intervalles de confiance sur les estimations des paramètres de la régression seront très grands. En d'autres mots, en présence de multicollinéarité l'estimation des paramètres de la régression n'est pas fiable.

Il existe des techniques qui rendent le processus d'estimation des paramètres d'une régression multiple plus fiable lorsqu'en présence de multicollinéarité. Une telle technique est la « ridge regression » (Hoerl et Kennard, 1970). Ce type de solution, cependant, ne fait qu'accroître la complexité du modèle de régression qu'on a au départ. Notre optique est davantage d'éliminer le problème de multicollinéarité à la source en considérant un modèle régional simple (i.e. à une seule variable indépendante) plutôt que multiple.

Le modèle régional préconisé par Fill et Stedinger (1998) donne lieu à une analyse bayésienne passablement complexe. Dans le modele bayésien proposé par Fill et Stedinger (1998) l'information locale est exprimée dans la fonction de vraisemblance de façon condensée sous la forme d'une estimation de l'estimateur 2P, qui est fonction du L-CV au site. Cette estimation est rattachée au quantile normalisé du site par une densité normale. L'information *a priori* sur le quantile normalisé est exprimée par une densité *a priori* normale ou la valeur des paramètres de moyenne et de variance prennent en compte le « comportement régional » du quantile normalisé.

Le comportement régional du quantile normalisé est décrit par un modèle de régression (multiple) log-linéaire qui l'exprime comme fonction de caractéristiques physiographiques des bassins de la région considérée. Plusieurs transformations et hypothèses sur les quantiles normalisés de la région sont ensuite nécessaires pour dégager du modèle log-linéaire une estimation de la moyenne et de la variance régionales du quantile normalisé. La moyenne et la variance régionales ainsi déterminées constituent les paramètres de la

densité *a priori*. Les transformations utilisées et la présence de multicollinéarité rendent difficile d'évaluer concrètement la validité des estimations régionales obtenues; c'est là le principal inconvénient de cette approche qui est dû en majeur partie à la présence du modèle log-linéaire.

6.9 Ébauche de démarche bayésienne englobant le modèle régional basé sur l'ACC

Des approches considérées jusqu'à présent, aucune ne semble pouvoir s'adapter au contexte régional basé sur l'ACC considéré ici. Il est possible de proposer une approche bayésienne de combinaison de l'information hydrologique disponible qui soit simple et bien adaptée au contexte régional considéré ici. Mais comme pour les autres approches étudiées, l'approche bayésienne avancée ici n'est pas pleinement satisfaisante.

Sous l'hypothèse que les crues annuelles sont décrites par une densité log-normale on peut écrire la fonction de vraisemblance comme suit :

$$\ln(x_1), \ln(x_2), ..., \ln(x_n) \approx N(\mu, \sigma^2) \tag{6.16}$$

La difficulté dans le présent contexte est que l'information régionale disponible ne se traduit pas directement sous forme de modèles séparés pour les paramètres de moyenne et de variance. En effet, le modèle régional employé lie le quantile de crue à la variable physiographique de superficie du bassin versant par un modèle log-linéaire :

$$\ln(Q_{100}) = \beta_0 + \beta_1 \ln(AIRE) + \varepsilon \tag{6.17}$$

Or, le quantile de période de retour 100 ans, Q_{100}, est fonction de la moyenne et de la variance de la densité log-normale utilisée comme fonction de vraisemblance :

$$Q_{100} = \exp(\mu + k\sigma) \tag{6.18}$$

avec k tel que $P(N(\mu, \sigma^2) > k) = 1/100$ ($k \cong 2.326$).

On peut combiner (6.17) et (6.18) pour exprimer le modèle régional en terme directement de μ et σ :

$$\mu + k\sigma = \beta_0 + \beta_1 \ln(AIRE) + \varepsilon \tag{6.19}$$

De façon équivalente on a :

$$\mu + k\sigma \approx N(\beta_0 + \beta_1 \ln(AIRE), \tau^2) \tag{6.20}$$

puisque $\varepsilon \approx N(0, \tau^2)$.

Le modèle régional considéré est un modèle de régression log-linéaire à une seule variable (i.e. simple) qui relie le quantile Q_{100} d'un bassin à la superficie de son bassin versant. Ce modèle de régression comporte de l'information sur le quantile Q_T du bassin cible dans la mesure où le modèle est adéquat dans sa prédiction pour le bassin cible. Le modèle a été ajusté à partir des bassins du voisinage homogène du bassin cible tel qu'obtenu de l'application de l'ACC. Par conséquent, le modèle est entièrement déterminé et ce indépendamment des données que sont les maxima annuels à la station cible. Le modèle de régression contribue donc au modèle avec de l'information *a priori*.

L'information régionale fournit donc de l'information que sur la somme des deux paramètres de la fonction de vraisemblance et non pas sur les paramètres séparément. En pratique, les paramètres inconnus β_0, β_1 et τ^2 de (6.20) sont remplacés par leurs estimations usuelles obtenues à partir des bassins du voisinage homogène autour du bassin d'intérêt. On peut procéder ainsi car l'information obtenue des données, et qui entre dans la fonction de vraisemblance, se rapporte au bassin d'intérêt et non pas aux bassins du voisinage homogène.

La difficulté évidente à ce point est de combiner l'information *a priori* sous la forme où elle se présente, i.e. (6.20), avec la fonction de vraisemblance (6.16). Et c'est une difficulté qu'on rencontre souvent en pratique avec l'approche bayésienne : quelle est la densité *a*

posteriori qui résume l'information disponible? Jusqu'à présent la caractérisation complète de la densité *a posteriori* se limite essentiellement aux cas simples où la densité *a priori* et la fonction de vraisemblance se rapportent à un seul paramètre et sont de la même famille. C'est pourquoi Kuczera (1982) et Rosbjerg et Madsen (1995) n'ont pas considéré dans leur analyse bayésienne les paramètres de moyenne et de variance simultanément mais ont plutôt choisi de réduire le problème à la caractérisation bayésienne d'un seul paramètre, à savoir la variance. En effet, sur la base que l'estimation de la moyenne régionale est plus stable que la variance, seul le paramètre de variance a fait l'objet d'une approche bayésienne. Comme on l'a vu à la section 6.7 cette approche n'est pas valable d'un point de vue bayésien, mais elle présente l'énorme avantage de donner lieu à une analyse subséquente qui se concentre sur un seul paramètre. Dans un deuxième temps, l'analyse qu'ils ont développée fait appel à la combinaison gamma-inverse comme densité *a priori* et gamma pour la vraisemblance puisque ces deux densités appartiennent à la même famille. On dit que ce sont des densités conjuguées. Les densités conjuguées ont l'avantage de donner lieu à une densité *a posteriori* de la même famille qui est, par conséquent, complètement caractérisée dès que les densités *a priori* et de vraisemblance le sont.

Il est possible de présenter une solution théorique, mais en pratique elle ne s'avère pas très satisfaisante. Elle consiste à supposer que les densités séparées pour μ et σ^2 sont normales. Ainsi on s'assure que leur somme, décrite par (6.20), est normale comme il se doit. Mais pour y arriver il faut déterminer quelle portion de la moyenne et de la variance de (6.20) doit être assignée à chacune des densités séparées. Considérons les densités séparées pour μ et σ^2:

$$\mu \approx N(\lambda(\beta_0 + \beta_1 \ln(AIRE)), \sqrt{\lambda}\tau^2) \tag{6.21}$$

$$\sigma^2 \approx N((1-\lambda)(\beta_0 + \beta_1 \ln(AIRE)), \sqrt{1-\lambda}\tau^2) \tag{6.22}$$

Le paramètre supplémentaire, auquel il faut trouver une valeur, délimite la part de la moyenne et de la variance de (6.20) qui doit revenir à chacune des densités séparées. Ce

paramètre peut être déterminé en considérant la densité marginale afin d'utiliser les données pour déterminer la valeur du paramètre impliqué qui maximise l'occurrence des données observées. Cette modélisation présente cependant plusieurs désavantages. D'une part, la complexité du modèle s'accroît passablement avec cette caractérisation additionnelle des densités séparées. Aussi, la pertinence de modéliser la variance par une densité normale est discutable. En effet, tout dépendant des paramètres résultant de la normale, il se peut que la densité normale ait une queue négative trop importante ce qui n'est pas une modélisation appropriée pour une variable qui prend des valeurs positives seulement. De plus, une répartition égale de la moyenne et de la variance de (6.20) pour les densités séparées est une contrainte (opérationnelle) forte qui peut ne pas s'avérer justifiable.

Il apparaît donc qu'une analyse bayésienne en bonne et due forme, opérationnelle avec les moyens actuels, ne peut pas être menée dans ce contexte. C'est un constat similaire à celui qui résulte des approches considérées dans d'autres contextes, e.g. Rosbjerg et Madsen (1995), Kuczera (1982), Bernier (1993), etc. Tout récemment, cependant, bon nombre de méthodes numériques, dont l'algorithme de Metropolis-Hastings (Chib et Greenberg, 1995) et l'échantillonneur de Gibbs (Casella et George, 1992), laissent entrevoir qu'il est possible de résoudre le problème fondamental de caractérisation complète de la densité a posteriori dans un contexte donné sans devoir se limiter aux cas élémentaires, et généralement peu appropriés, des familles de densités conjuguées. Si ces outils s'avèrent efficaces, ils rendront l'approche statistique bayésienne beaucoup plus souple à l'application. Les contraintes que requièrent présentement le développement théorique complet d'une approche bayésienne pour un contexte comme celui-ci pourront alors être levées par l'apport de méthodes numériques efficaces. C'est à souhaiter car la détermination théorique explicite de la densité a posteriori peut être passablement difficile, voire même impossible dans plusieurs cas.

7. DISCUSSION ET CONCLUSION

Les travaux menés sur l'approche de régionalisation des valeurs extrêmes de crue basée sur l'ACC ont donnés lieu à plusieurs contributions significatives sur ce sujet. D'une part, avec les travaux contenus dans Ouarda *et al.* (2000), une justification détaillée des hypothèses sous-jacentes au développement théorique de l'approche est maintenant disponible. Aussi, pour la première fois, une description détaillée et succinte des étapes à suivre pour utiliser l'approche basée sur l'ACC en pratique est donnée.

D'autre part, les travaux contenus dans Girard *et al.* (2000a) ont permis de révéler que l'approche actuelle n'est pas complète, d'où les difficultés à l'application qu'occasionne la détermination du paramètre impliqué dans la définition actuelle des voisinages homogènes. Dans ce rapport scientifique nous avons également montré comment le cadre théorique initié pouvait être dûment complété, menant à une nouvelle définition de voisinage homogène qui ne nécessite pas de paramètre à fixer. La nouvelle définition ne comporte donc pas les difficultés à l'application que présentait la précédente. Maintenant que le cadre théorique qui sous-tend l'application de l'ACC dans un contexte de régionalisation est complet, il reste à donner de l'approche plusieurs exemples explicites d'applications. D'une part, l'approche pourrait être appliquée à un ensemble donné de bassins autre que celui des bassins de l'Ontario ou du Québec, déjà utilisés pour l'illustrer. Ainsi, l'approche de régionalisation basée sur l'ACC serait mise à l'épreuve dans des contextes physiographiques et hydrologiques différents de ceux exploités jusqu'à présent pour en illustrer le fonctionnement.

Les résultats présentés dans Girard *et al.* (2000b) sur le biais dans la prédiction depuis un modèle log-linéaire mettent clairement en garde les utilisateurs sur les conséquences de réduire le biais en introduisant un facteur de correction approprié. En effet, nous avons

montré que l'introduction d'un facteur de correction du biais dans un modèle log-linéaire a pour effet, largement insoupçonné, d'accroître la variance des prédictions obtenues du modèle. En somme, les efforts investis à réduire le biais ne se transforment pas en gains nets sur l'estimation puisqu'ils provoquent des pertes au niveau de la variance en contrebalance.

Il reste clairement beaucoup de travail à faire pour développer un cadre bayésien qui soit adéquat et satisfaisant pour combiner l'information hydrologique quelle que soit la forme sous laquelle elle peut se présenter. Cependant, notre travail a permis de cerner les fondements de l'approche bayésienne et d'en donner une couverture qui soit plus adéquate et substantielle que ce que la littérature hydrologique à ce sujet avait à proposer au début de cette étude. L'essor récent de techniques numériques poussées contribuera sans nul doute à rendre les applications bayésiennes plus nombreuses. En effet, ces techniques promettent de pouvoir établir la densité *a posteriori* quelles que soient la fonction de vraisemblance et la densité *a priori* en jeu, alors que présentement seuls les cas les plus simples peuvent être entièrement caractérisés.

8. RÉFÉRENCES

Berger, J.O., Bayesian inference and decision techniques, P. Goel and A. Zellner editors, Elsevier Science publishers, 1986, pp. 473-487.

Berger, J.O., Statistical decision theory and Bayesian analysis, Springer, 1982.

Berry, D.A., Statistics: A Bayesian perspective, 1996.

Bernier, J., Combinaison des informations locales et spatiales dans les modèles de régionalisation des crues, Note interne, 1993.

Casella, G., E.I. George, Explaining the Gibbs sampler, The American Statistician, 46, pp. 167-174.

Chib, S., E. Greenberg, Understanding the Metropolis-Hastings algorithm, The American Statistician, 49, pp.327-335.

Cavadias, G., T.M.B.J. Ouarda, B. Bobée, C. Girard, A canonical correlation approach to the determination of homogeneous regions for regional flood estimation of ungauged basins, accepté pour publication au Journal of Hydrologic Sciences, 2000.

Chou, Y., Statistical analysis for business and economics, Elsevier Science Publishing, 1989.

Efron, B., C. Morris, Le paradoxe de Stein, Pour la Science, pp. 28-37, 1977.

Fill, D.F., J.R. Stedinger, Using regional regression within index flood procedures and an empirical Bayesian estimator, Journal of Hydrology, 1998, 210, pp.128-145.

Girard, C., T.B.M.J. Ouarda, B. Bobée, Étude du biais dans la prédiction depuis un modèle log-linéaire, Rapport de recherche No. R-575 INRS-Eau, 2000b.

Girard, C., T.B.M.J. Ouarda, B. Bobée, Une approche par classification à la constitution de voisinages homogènes basés sur l'ACC, Rapport de recherche No. R-576 INRS-Eau, 2000a.

GREHYS 1996b. Intercomparison of flood frequency procedures for Canadian rivers. *J. Hydrol.* 186(1-4): 85-103.

Hoerl, A.E., R.W. Kennard, Ridge regression: biased estimation for non-orthogonal problems, Technometrics, 1970, 12, pp. 55-67

Hoos, A.B., Improving regional-model estimates of urban-runoff quality using local data, Water Resources Bulletin, 1996, Vol. 32, no.4, pp. 855-863.

Jaynes, E.T., Maximum entropy and Bayesian Methods in Applied Statistics, J. H. Justice editor, 1986, pp. 1-25.

Kuczera, G., Combining site-specific and regional information : an empirical bayes approach, Water Resources Research, 1982, Vol. 18, no. 2, pp. 306-314.

Lindley, D.V., Readings in uncertain reasoning, Glenn Shafer editor, p.26, 1990.

Miller, D.M., Reducing transformation bias in curve fitting, The American Statistician, 1984,Vol. 38, no.2, pp-124-126.

Montgomery, D.C., E. A. Peck, Introduction to linear regression analysis, 1992, John Wiley & Sons, Inc.

Neyman, J., E. L. Scott, Correction for bias introduced by a transformation of variables, Ann. Math. Statist., 1960, 31, pp. 643-655.

Ouarda, T.B.M.J., G. Boucher, P.F. Rasmussen, B. Bobée, Regionalization of floods by canonical correlation analysis, Proceedings of the European water resources association conference, 1997, pp. 297-302.

Ouarda, T.B.M.J., C. Girard, G.S. Cavadias, B. Bobée, Regional flood frequency estimation with canonical correlation analysis, soumis au Journal of Hydrology, 2000.

Perreault, L., Analyse bayésienne rétrospective d'une rupture dans les séquences de variables aléatoires hydrologiques, Thèse de doctorat, INRS-Eau, Nov. 2000

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

Rosbjerg, D., H. Madsen, Uncertainty measures of flood frequency estimators, Journal of Hydrology, 1995, 167, pp. 209-224.